Supplemental Information

Supplemental Figure Legends

Figure S1. Identification of mouse- and human-specific exons, Related to Figure 1.
(A) As in Figure 1D, but shows a portion of the mouse tumor protein D52 (TPD52) gene (Ensembl ID ENSMUSG00000027506 on mouse chr3) together with homologous sequences from the rat TPD52 gene on rat chr2. This region contains a mouse-specific exon that is predicted to encode transmembrane domains by TMHMM (Kall et al., 2004).
(B) As in Figure 1C. Top: a phylogenetic tree presenting the main species used for dating exons and the branch lengths in millions of years. Bottom: exons of increasing evolutionary splicing age, their pattern of presence or absence in various species, and the number of each class of exons identified.

Figure S2. Genomic origins of new exons, Related to Figure 3.
As in Figure 3B,C.
(A) Proportion of human-specific exons that match various categories of repeats.
(B) Proportion of human genome belonging to various repeat categories.
(C) Splice site selection in mouse B1 derived exons and human Alu derived exons.
(D) Distribution of PSI values of exons with canonical (GYAG) and non-canonical (non-GYAG) splice site dinucleotides.
(E) Splice site strength changes associated with exons already containing minimal splice sites, measured by the MaxEnt method (Yeo and Burge, 2004).

Figure S3. Sequence substitution rates and intron length changes associated with rat- and mouse-specific exons, Related to Figures 2, 3.
(A) Substitution rates per base were computed from alignments of mouse exons to orthologous rat sequences, pegged to the mouse splice junctions. Mean substitutions per base are plotted for the 80 bases upstream and downstream of exons, as well as the first and last 30 bases of the exons (for all exons $\geq$ 60 bp in length). Mouse exons were grouped by evolutionary age. Mouse-

specific exons (age 0) were further subdivided by their homology to intronic or intergenic sequences in rat and whether they derived from unique or repetitive sequences.
(B) The change in length of the entire intron region between rat and mouse in regions associated with a rat-specific exon. The length in mouse is plotted as a percentage of the length in rat (mean ± SEM). As in Figure 3E-F.
(C) The relative length of the downstream intron as a percentage of the upstream intron (rat) or the downstream aligned intron/region as a percentage of the upstream aligned intron/region (mouse) (mean ± SEM) in regions associated with a rat-specific exon. The mouse bar in the -R-- class is hatched to represent the fact that it is not an exon.
(D) The change in length of the entire intron region between macaque and mouse in regions associated with a mouse-specific exon. The length in macaque is plotted as a percentage of the length in mouse (mean ± SEM).
(E) The relative length of the downstream intron as a percentage of the upstream intron (mouse) or the downstream aligned intron/region as a percentage of the upstream aligned intron/region (macaque) (mean ± SEM) in regions associated with a mouse-specific exon. The macaque bars in the M---- and MR--- classes are hatched to represent the fact that they are not an exon.

Figure S4. New exons with upstream intronic shortening lack the increase in ESEs observed in new exons overall, Related to Figures 3, 4.
As in Figures 1D and 3D.
(A) Proportion of new mouse exons with specific splice site dinucleotide sequences in mouse and rat for new exons with upstream intronic shortening (>= 20 nucleotides).
(B) Change in SRE number in various regions in and around new exons with upstream shortening.
(C) Proportion of new mouse exons with specific splice site dinucleotide sequences in mouse and rat for new exons without upstream intronic shortening.
(D) Change in SRE number in various regions in and around new exons without upstream shortening.

Figure S5. Nucleosome positioning near human exons with upstream structural sQTLs, Related to Figure 4.

Nucleosome positioning (measured by protection from MNase treatment) around exons with a structural sQTL in the upstream intron grouped by sQTL genotype and binned by distance from the associated exon:
(A) The subset of sQTLs with variant located < 2 kb from the exon.
(B) The subset of sQTLs with variant located ≥ 2 kb from the exon.
(C) Base composition of ancient exons and new exons with and without upstream shortening.

Figure S6. Species-specific expression changes are associated with species-specific exons, Related to Figure 5.

(A) To evaluate the possibility that the increase in expression associated with species-specific exon splicing is due to species-specific promoters, the ratio of junction reads overlapping the 3' splice site was compared to the ratio of junction reads overlapping the 5' splice site in exons of different ages.
(B) The fraction of genes containing a new exon in the subset of genes containing a species-specific increase in expression is compared to the fraction in the background of genes without a species-specific exon for `M----` (beige) or `-R---` (brown), showing that species-specific exons are enriched within genes that have species-specific expression changes.

Supplemental Table Legends

Table S1. Summary of datasets and literature related to mammalian exon evolution, Related to Figures 1-6.

Table S2. Mouse, rat, macaque and human exons and their splice sites, coding status, evolutionary ages, and presence in other annotations, Related to Figures 1-6.
In column F, AE and CE are the number of individuals in which the exon is alternative or constitutive respectively, not_expressed is the number of individuals in which the exon did not meet the expression filter, and CE_NET is the number of individuals in which the exon was detected as constitutive (psi = 1) in all tissues expressed but did not meet the required filters to be classified as such (psi = 1 in at least 3 tissues). Genomic and splicing ages are as defined in the text (-1 means the age could not be determined). SplicingAge_ESC is the splicing age after correction with human embryonic stem cell (ESC) data (splicing ages of exons detected in human ES cells were adjusted accordingly). In columns M through P, 0 means the exon is not annotated in that source, 1 means it is annotated with the same start and stop, 2 means it is annotated with the same start or stop, and a 3 means it is annotated with a different start and stop. Separate sheets for mouse, rat, macaque and human exons.

Table S3. Phylogenetic patterns of mouse exons, Related to Figure 1.

Table S4. Gene Ontology enrichment analysis for genes containing species-specific coding exons, Related to Figure 2.

Supplemental Experimental Procedures

RNA-seq and genome builds

Data from from mouse, rat, rhesus, cow, and chicken were processed as in(Merkin et al., 2012) using TopHat v1.1.4 (Trapnell et al., 2009) and Cufflinks v1.0.2 (Trapnell et al., 2012). Mouse data were mapped to mm9, rat data to rn4, rhesus data to rhemac2, cow data to bostau4, and chicken data to galgal3.

Assignment of ages to exons

Exons from each species (mouse, rat, rhesus, cow, chicken) from (Merkin et al., 2012) were used in this analysis. In that study, to define an exon we required a certain minimum gene expression level (FPKM $\geq$ 2) in the tissue, corresponding to roughly 40 reads in the region of the exon. Additional splice site junction read requirements were implicit in the TopHat mapping, including an estimated minimum proportion of junction reads supporting the exon of 15% in at least one sample and multiple unique junction reads with at least 6 nt of overhang for the junction to be reported.

As in our 2012 study, we only considered single-copy genes. We flagged and removed initial and terminal exons and focused only on internal exons from these genes. We filtered internal exon duplications by aligning each exon to other exons in the same gene. Aligned regions in other species for each query exon were collected based on whole genome alignments generated by PECAN and EPO (Paten et al., 2008), and pairwise alignments from BLASTZ (Schwartz et al., 2003). In addition, in order to reduce the exon splicing age bias resulted from the missing of aligned genomic regions, we applied BLAT (Kent, 2002) on exons which do not have a genomic aligned region expressed in chicken. The best aligned regions were selected using a minimum threshold of 80% identity for alignment to rat, 66% identity for alignment to rhesus, 65% identity for alignment to cow, and 54% identity for alignment to chicken. These thresholds were calculated by taking values 3 standard deviations below the average percentage identity of exons between the query species (mouse) and the other species in question.

An exon's genomic age was defined based solely on the pattern of species with genomic regions aligned to the query exon. We interpreted this pattern using parsimony, considering the minimum number of changes that can explain the pattern of aligned regions, and mapping these

onto a precomputed species tree (Alekseyenko et al., 2007; Roy et al., 2008). We only considered unambiguous age assignments (i.e. if there were multiple equally parsimonious assignments that would yield different ages, then the exon was excluded from analysis). An exon's splicing age was assigned in a similar manner to the genomic age, only it was based the pattern of presence or absence of an expressed region in the orthologous gene overlapping the genomic aligned region.

For example, a mouse exon's genomic age was assigned to 25, 90, 110 or 300 My if there were aligned regions in rat, rat/rhesus, rat/rhesus/cow and rat/rhesus/cow/chicken, respectively, or 0 (indicating < 25 My) if no aligned region in other species was detected. Similarly, its splicing age was assigned to similar categories if there were aligned regions expressed or annotated as exons in rat, rat/rhesus, rat/rhesus/cow and rat/rhesus/cow/chicken (Figure 1C). All of the RNA-seq data were derived from adults of each species, so isoforms expressed exclusively at earlier developmental stages would be missed. To assess the magnitude of this effect on our classifications, we mapped RNA-seq reads from human embryonic stem cells (ESCs) (Consortium, 2012) to the human genome, and reclassified the evolutionary ages of exons detected in ESC but not adult tissue data (Supplemental Table 2). This analysis resulted in reclassification of 0.13% of mouse and rat exons and fewer than 0.01% of macaque exons, suggesting that misclassification on the basis of exclusive early developmental expression in some lineages is not frequent in this dataset.

We estimated the rate of false identification of exons as species-specific to be ~1.5%. For this purpose, we used a modification of the approach taken by (Nielsen et al., 2004), in which this quantity is estimated by the probability that the distribution of presence-absence of the exon resulted from multiple losses rather than species-specific gain.

The numbers of genes containing new exons in mouse, rat, macaque and human were 795, 945, 941 and 1412. Of these, the numbers with >1 novel exon were 159, 276, 275 and 380, respectively, indicating that most genes that contained new exons had only one. Gene ontology analysis of genes containing novel coding exons yielded only the category "splice variant" as significant after correcting for multiple comparisons, suggesting that these genes have diverse functions. We only considered exons detected in the previous RNA-seq study (Merkin et al., 2012) in order to mitigate the effects of prior transcript annotation quality on our subsequent results since, for instance, rhesus annotations (largely derived from human annotations) would be

expected to be much better than cow. This approach will miss annotated exons only included in embryonic tissues, for instance (as discussed above), but those would likely have been incorrectly assigned to the novel, recently created exon category due to the possibility of their not being found in other species because comparable data are not available.

Basic exon properties

Exons with PSI > 0% and PSI < 97% (where PSI represents the Percent Spliced In, or the percentage of transcripts in a particular tissue estimated to include the exon in question) in at least 1 tissue and at least 2 individuals were categorized as skipped exons (SE). Exons with PSI > 97% in all tissues with expression were defined as constitutive exons (CE), provided that the gene was expressed in 3 or more tissues in at least 2 individuals (since the probability of detecting exon skipping increases with the number of tissues considered).

Transcripts' open reading frames (ORFs) were annotated as in (Merkin et al., 2012). Briefly, if a transcript contained an annotated translation start site, then the longest ORF originating from that site was used. If no such site was contained in the transcript, then the longest ORF 100 amino acids or longer was used. If none existed, then the transcript was considered non-coding. Exons that can map to transcripts' ORF region, upstream and downstream region of transcript ORF, and regions in transcripts without ORF were categorized as coding exons, 5' UTR and 3' UTR exons and non-coding exons, respectively. In Figure 2C, the proportion of coding exons were calculated by coding exons / total, where coding is the number of coding exons and total is total counts of exons at each age.

Genomic sources of new exons

We traced the origins of new exons by allocating the genomic locations of aligned regions in the closest species (for example, in mouse, we used rat as its closest species). In Figure 3A, exons were categorized into intronic, proximal intergenic, non-proximal intergenic, other coding gene, other intron and other ncRNA gene if their aligned regions in the closest species are located in the intronic regions of the same gene, intergenic regions but closer to the orthologous gene than any other gene, other intergenic regions, exonic regions of other genes, intronic regions of other genes and other regions of ncRNA, respectively.

The origins of new exons were also categorized based on overlap with repeated sequences. The RepeatMasker [http://www.bioinfo.org.cn/relative/RepeatMasker1.htm] track was downloaded from the UCSC browser and used to identify repeats overlapping each exon. Exons were categorized as containing SINEs, LINEs, LTRs, or other repeats (rarer categories). Exons not overlapping any repeat class were assigned to the “unique" group in Figure 3B. The whole genomic unique sequence class, SINEs, LINEs, LTRs and other repeat classes were used as genomic background (Figure 3C).

Splice site and splicing regulatory element analysis

The dinucleotide frequencies of the intronic 5' and 3' splice sites of mouse new exons and their aligned regions in rat were compared in Figure 1D. In Figure 3D, exonic splicing enhancers (ESEs) from (Fairbrother et al., 2002), exonic splicing silencers (ESSs) from (Wang et al., 2004), intronic splicing enhancers (ISEs) from (Wang et al., 2012), and intronic splicing silencers (ISSs) from (Wang and Wang, 2014) were used. The 100 nt of intronic sequence upstream and downstream of each exon in mouse or the aligned region in rat was considered for searching for intronic splicing regulatory elements. The entire exon was searched for exonic splicing regulatory elements. To control for differences in exon length, the average frequency of such changes were multiplied by the average new exon length to arrive at the average change per exon.

Intron length analyses

For each exon age, the sum of lengths of each mouse exon and its upstream and downstream introns were compared to the corresponding sum in rat by summing the lengths of the rat exon (or aligned region for mouse-specific exons) and the surrounding introns (Figure 3E). For Figure 3F, the length of the downstream mouse intron was divided by the length of the upstream mouse intron. A similar ratio was calculated in rat, dividing the length of the intronic region downstream of the rat exon or proto-exon (in case of mouse-specific exons) divided by the length of the upstream intronic region.

Z-score conversion for comparisons

For each change considered (changes in intronic or exonic splicing enhancers or silencers, or deletions), the empirical distribution of such changes in the ancient set of exons (MRQCG) was determined. The mean and standard deviation of this distribution was calculated. Each change was then calculated for each new exon and converted to a z-score using the values calculated in the ancient group (Figure 3G).

Nucleosome localization and GRO-seq analyses

We downloaded the MNase-seq data from (Gaffney et al., 2012) from GEO (accession no. GSE36979). We mapped the reads with Bowtie v0.12.7(Langmead et al., 2009) to mm9. We considered ancient (MRQCG) exons, new mouse exons with no upstream intron deletion, new mouse exons with an upstream intron deletion, and the orthologous region of new rat exons. We used pysam v0.7.7 and samtools v0.1.16 (Li et al., 2009) to count the number of reads in a 1 kb window of each exon. Each exon's profile was internally normalized, and the average profile of each set of exons was smoothed with a sliding window and plotted, centered on the exon midpoint.

We downloaded the GRO-seq data from (Kaikkonen et al., 2013) from GEO (accession no. GSE48759). We combined the various samples to increase statistical power. While the transcriptional level of a particular gene in each condition may be different, since we focused on internal exons and internally normalized each region, this should not affect our results. These data were then processed in the same manner as the MNase-seq data. Some of the data types shown in Figure 4, particularly GRO-seq data, can be biased by GC content. However, this possibility is unlikely to have impacted the results in Figure 4 as new exons with shortened upstream introns had similar GC content to other new exons (both slightly lower than for ancient exons; Figure S7C).

To investigate the impact of intronic structural variants on nucleosome localization (Figure 4C), we downloaded the following files:

- the sQTL table EUR373.exon.cis.FDR5.all.rs137.txt.gz from the Geuvadis consortium (Lappalainen et al., 2013),
- Gencode v12 (Derrien et al., 2012), matching the annotations used in the Geuvadis study from http://www.gencodegenes.org,

- MNase-seq data from individuals included in the Geuvadis study from (Gaffney et al., 2012),
- Genotype data for these individuals from the 1000 Genomes Project (Abecasis et al., 2012) tables ALL.chr**.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz, where ** represents all chromosome numbers,
- GRCh37.remap.all.germline.gvf from ref (Lappalainen et al., 2013) for determining variant lengths.

The MNase-seq data were processed as described above. We filtered out all SNV sQTLs, as well as any indel or structural variant that was smaller than 5 bp. We further filtered this list such that the sQTL was wholly contained within the upstream or downstream intron. We then further filtered the sQTLs considered such that all individuals analyzed did not contain the same genotype for that particular variant. We then compiled the MNase profiles of individuals with genotypes representing shorter upstream introns (reference allele for upstream insertions and variant allele for upstream deletions) and longer upstream introns (reference allele for upstream deletions and variant allele for upstream insertions) and processed and plotted as done previously.

We used the Kolmogorov-Smirnov test to compare empirical cumulative distribution functions in several places, as described (Mootha et al., 2003).

New exon inclusion and species-specific expression changes

Gene expression in mouse was compared to gene expression in rat by taking the ratio of mouse to rat expression using RNA-seq estimates of gene expression from (Merkin et al., 2012). We considered the following cases in Figure 5A: 1) genes with a new exon where the new exon is included in the tissue in question, 2) genes with a new exon where the new exon is not included in the tissue in question, and 3) genes with no new exon in either mouse or rat.

The intra-species expression ratio (Figure 5B) is calculated by averaging a gene's expression in mouse in the tissues where the exon is included and dividing that by the mean expression in tissues where the exon is not included. This ratio was then calculated in rat, matching the tissues in the fore- and background, and the ratio of these two values was analyzed. As a control, the tissue labels were shuffled and the statistic was recalculated.

The analysis detecting enrichment for new exons in genes containing expression changes (Figure S8B) was conducted as follows. For each gene, we constructed a set of constitutive exons in each species containing no alternatively spliced segments. For each tissue in mouse and rat, we counted the number of reads overlapping each region using pysam and adjusted the raw counts for differences in length considered between species, down-sampling to match the shorter length. We then applied DEseq (Anders and Huber, 2010) and identified genes with higher expression within the species being studied with an adjusted FDR of 0.0001, or approximately 0.001 when adjusting for additional tests across tissues. We then divided the fraction of genes with significantly elevated expression that contain a novel exon to the overall fraction of genes that contain a novel exon.

We also compared the gene expression ratio: 1) for ancient exons included in rat but skipped in mouse (Figure 5D); 2) for exons alternatively spliced in mouse (grouped by inclusion level, PSI) but constitutive in rat (Figure 5E); 3) for 5'-most and 3'-most internal exons alternatively spliced in mouse but constitutive in rat (Figure 5F); and 4) for mouse- and rat-specific exons (grouped into tertiles by inclusion level) but excluded in the closest aligned species (rat and mouse) (Figure 5C). For case 1) we used ancient exons included in both species as control; for case 2) and 3), we used constitutive exons included in both mouse and rat as control; for case 4) we used species-specific exons excluded in tissues as control.

The incomplete splicing ratio (ISR) was determined for ancient exons constitutively included in rat but alternatively spliced in mouse (same exons as in Figure 5D). The ISR measures the ratio of intronic to exonic reads in RNA-seq data, as intronic reads found in these data originate from unspliced pre-mRNA. The ISR for a particular rat exon was calculated by dividing the mean read density in its flanking introns by that of its flanking exons, averaged across all tissues with a minimum of 10 reads per base pair on average over the exon. ISR values from the species with constitutive splicing (i.e. rat) were used to avoid the complications of correcting the ISR for alternative splicing level.

Analysis of splice junction data to confirm internal exon classifications.

To confirm our classification of these mouse-specific exons as internal exons (Fig. 1A), we counted RNA-seq reads supporting their 3' and 5' splice junctions. We observed similar read

densities at both junctions, consistent with their classification as internal exons and inconsistent with models in which changes in expression result from new internal promoters (Fig. S8A).

Software versions

The analyses were conducted in Python v2.7.2 (www.python.org) using Scipy v0.13.2 (http://www.scipy.org), Numpy v1.8.0 (http://dx.doi.org/10.1109/MCSE.2007.58), Matplotlib v1.3.1, pycogent v1.5.1 and pandas v0.10.0.

Supplemental References

Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56-65.

Alekseyenko, A.V., Kim, N., and Lee, C.J. (2007). Global analysis of exon creation versus loss and the role of alternative splicing in 17 vertebrate genomes. RNA *13*, 661-670.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome biology *11*, R106.

Consortium, E.P. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G., *et al*. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome research *22*, 1775-1789.

Fairbrother, W.G., Yeh, R.F., Sharp, P.A., and Burge, C.B. (2002). Predictive identification of exonic splicing enhancers in human genes. Science *297*, 1007-1013.

Gaffney, D.J., McVicker, G., Pai, A.A., Fondufe-Mittendorf, Y.N., Lewellen, N., Michelini, K., Widom, J., Gilad, Y., and Pritchard, J.K. (2012). Controls of nucleosome positioning in the human genome. PLoS genetics *8*, e1003036.

Kaikkonen, M.U., Spann, N.J., Heinz, S., Romanoski, C.E., Allison, K.A., Stender, J.D., Chun, H.B., Tough, D.F., Prinjha, R.K., Benner, C., *et al*. (2013). Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription. Molecular cell *51*, 310-325.

Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. Genome research *12*, 656-664.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol *10*, R25.

Lappalainen, T., Sammeth, M., Friedlander, M.R., t Hoen, P.A., Monlong, J., Rivas, M.A., Gonzalez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., *et al*. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506-511.

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

Merkin, J., Russell, C., Chen, P., and Burge, C.B. (2012). Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. Science (New York, NY) *338*, 1593-1599.

Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., *et al*. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet *34*, 267-273.

Nielsen, C.B., Friedman, B., Birren, B., Burge, C.B., and Galagan, J.E. (2004). Patterns of intron gain and loss in fungi. PLoS biology *2*, e422.

Paten, B., Herrero, J., Beal, K., Fitzgerald, S., and Birney, E. (2008). Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. Genome research *18*, 1814-1828.

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

Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D., and Miller, W. (2003). Human-mouse alignments with BLASTZ. Genome research *13*, 103-107.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111.

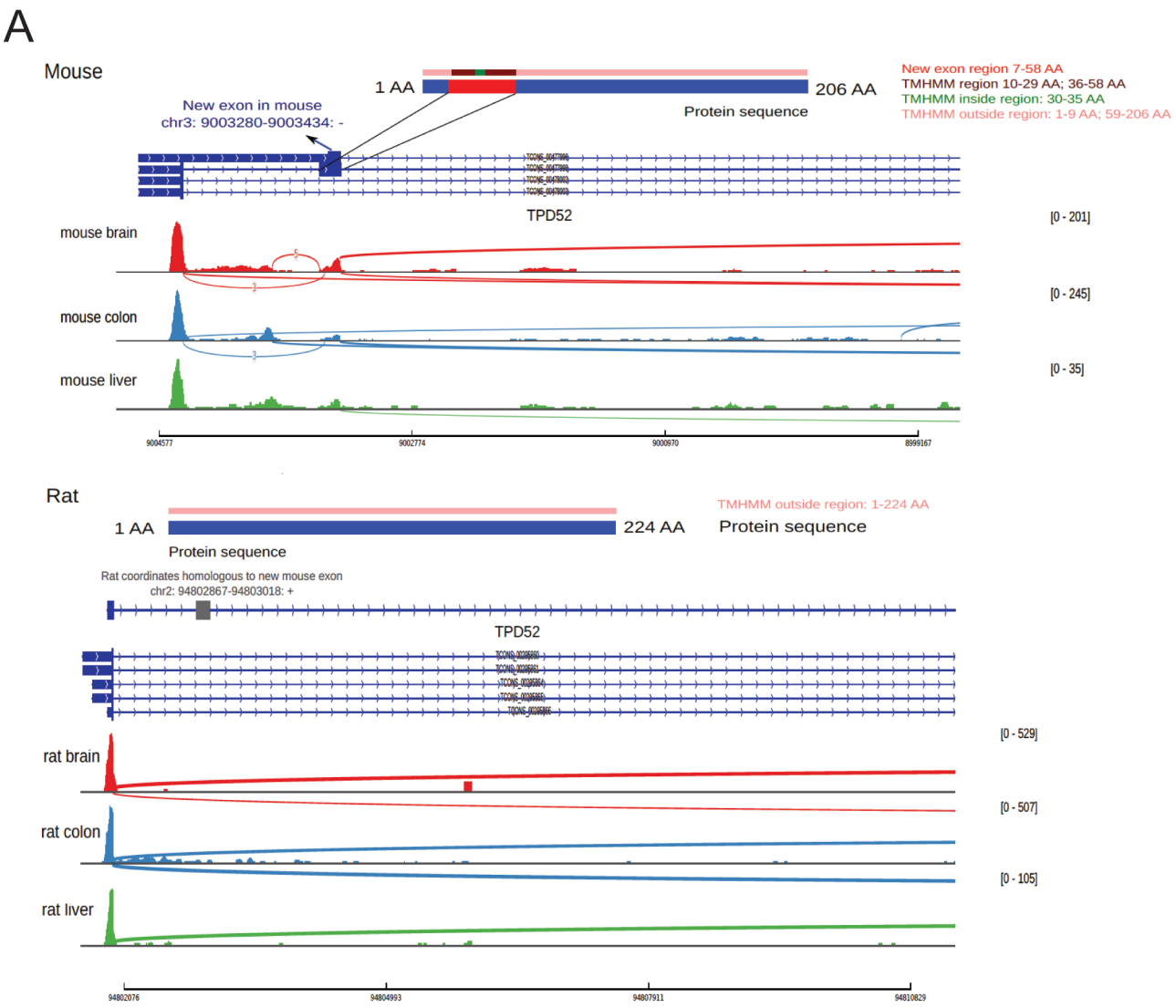
Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols *7*, 562-578.

Wang, Y., Ma, M., Xiao, X., and Wang, Z. (2012). Intronic splicing enhancers, cognate splicing factors and context-dependent regulation rules. Nature structural & molecular biology *19*, 1044-1052.

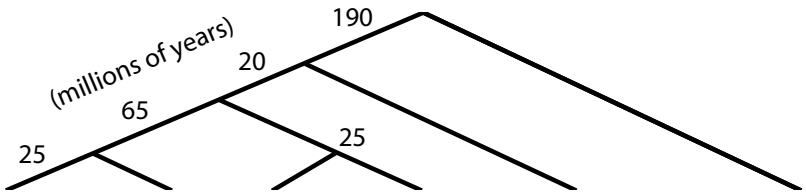
Wang, Y., and Wang, Z. (2014). Systematical identification of splicing regulatory cis-elements and cognate trans-factors. Methods *65*, 350-358.

Wang, Z., Rolish, M.E., Yeo, G., Tung, V., Mawson, M., and Burge, C.B. (2004). Systematic identification and analysis of exonic splicing silencers. Cell *119*, 831-845.

Figure s1

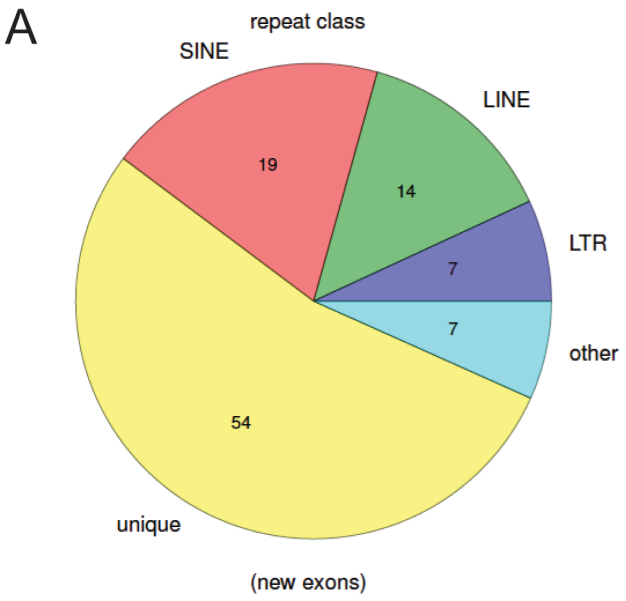


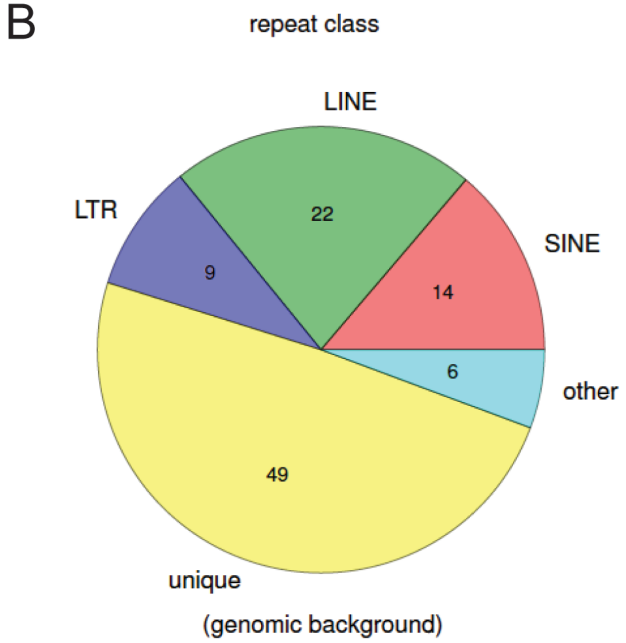
B

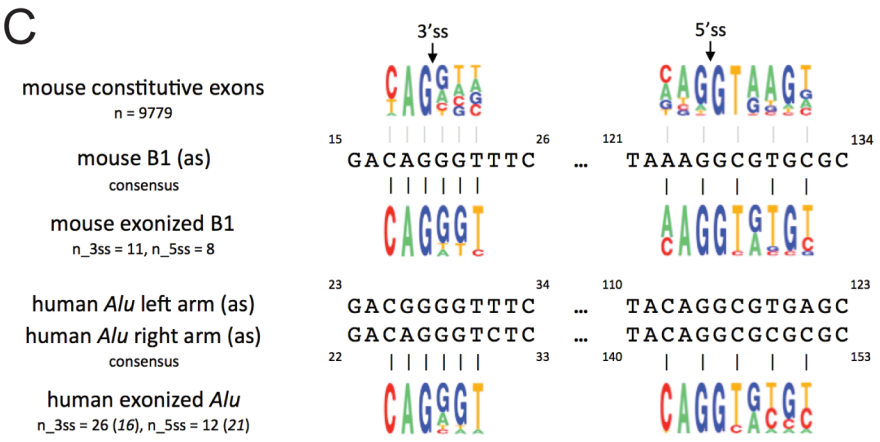


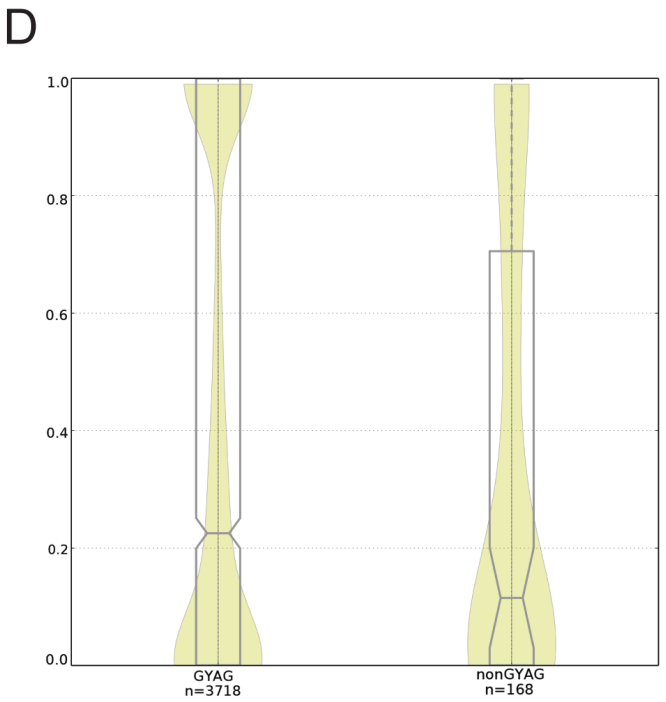
splicing age	abbr	human	macaque	mouse	rat	cow	chicken	no. of exons
	Present	H	Q	M	R	C	G	
Mya	absent	–	–	–	–	–	–	
0-25	H-----							2073
25-90	HQ----							231
90-110	HQMR--							636
110-300	HQMRC-							4553
300+	HQMRCG							43613

Figure s2


A
repeat class
SINE
LINE
LTR
other
unique
19
14
7
7
54
(new exons)


B
repeat class
LINE
LTR
SINE
other
unique
22
9
14
6
49
(genomic background)


C
3'ss
5'ss
mouse constitutive exons
n = 9779
mouse B1 (as)
consensus
mouse exonized B1
n_3ss = 11, n_5ss = 8
GACAGGGTTTC ... TAAAGGCGTGCGC
human Alu left arm (as)
human Alu right arm (as)
consensus
human exonized Alu
n_3ss = 26 (16), n_5ss = 12 (21)
GACGGGGTTTC ... TACAGGCGTGAGC
GACAGGGTCTC ... TACAGGCGCGCGC


D
1.0
0.8
0.6
0.4
0.2
0.0
GYAG
n=3718
nonGYAG
n=168

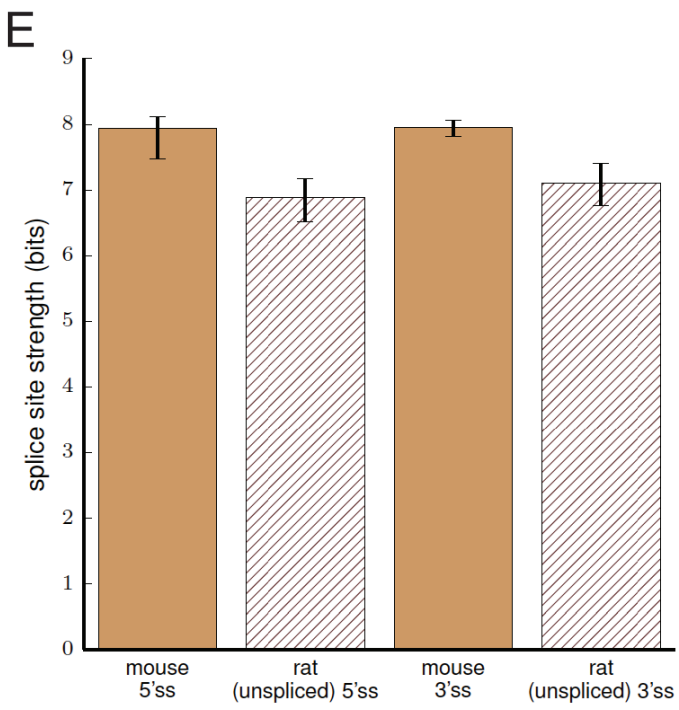

E
splice site strength (bits)
0
1
2
3
4
5
6
7
8
9
mouse
5'ss
rat
(unspliced) 5'ss
mouse
3'ss
rat
(unspliced) 3'ss

Figure s3

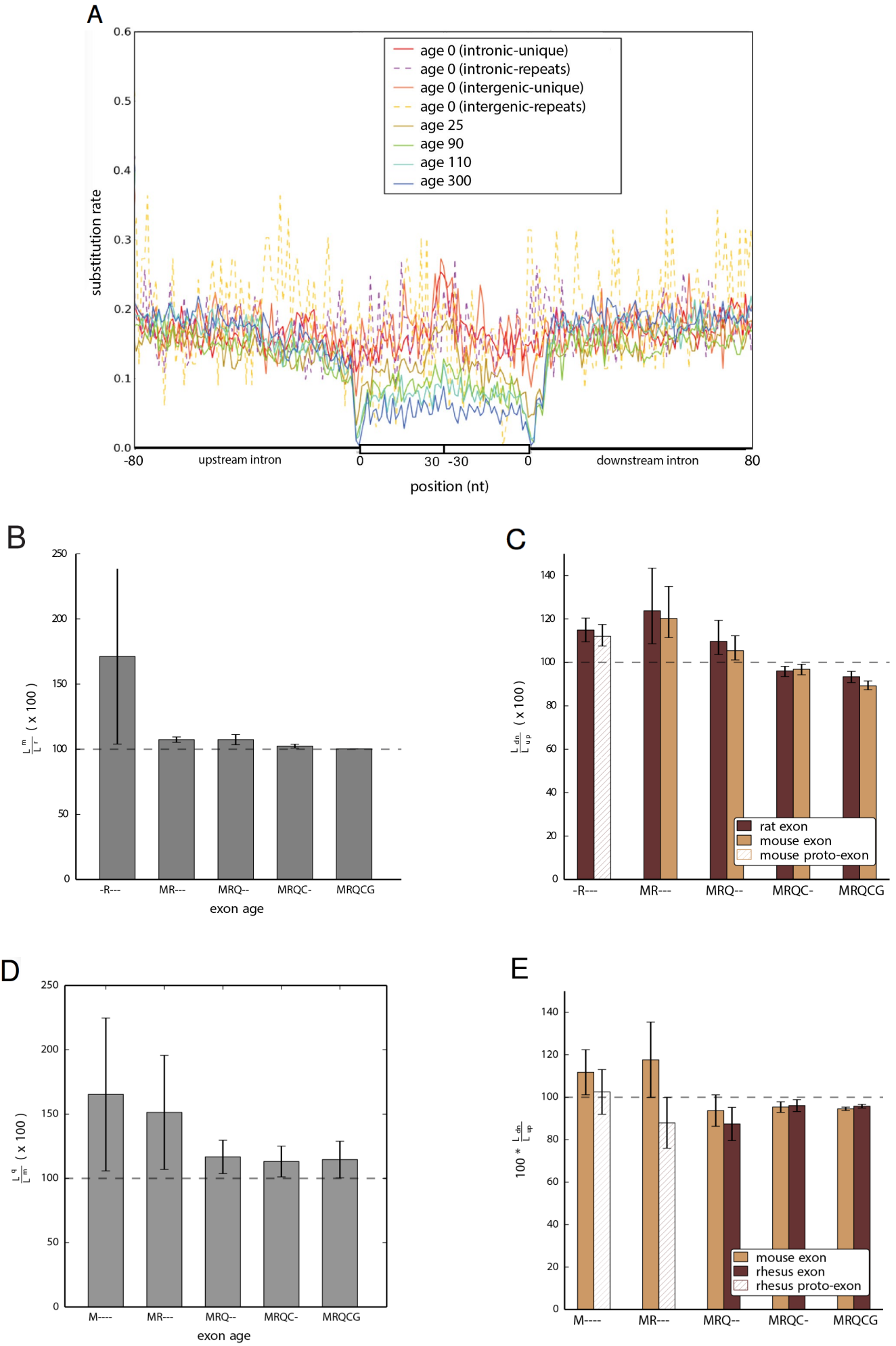

A
age 0 (intronic-unique)
age 0 (intronic-repeats)
age 0 (intergenic-unique)
age 0 (intergenic-repeats)
age 25
age 90
age 110
age 300
substitution rate
upstream intron
downstream intron
position (nt)
B
exon age
-R---
MR---
MRQ--
MRQC-
MRQCG
C
rat exon
mouse exon
mouse proto-exon
D
exon age
M----
E
mouse exon
rhesus exon
rhesus proto-exon

Figure s4

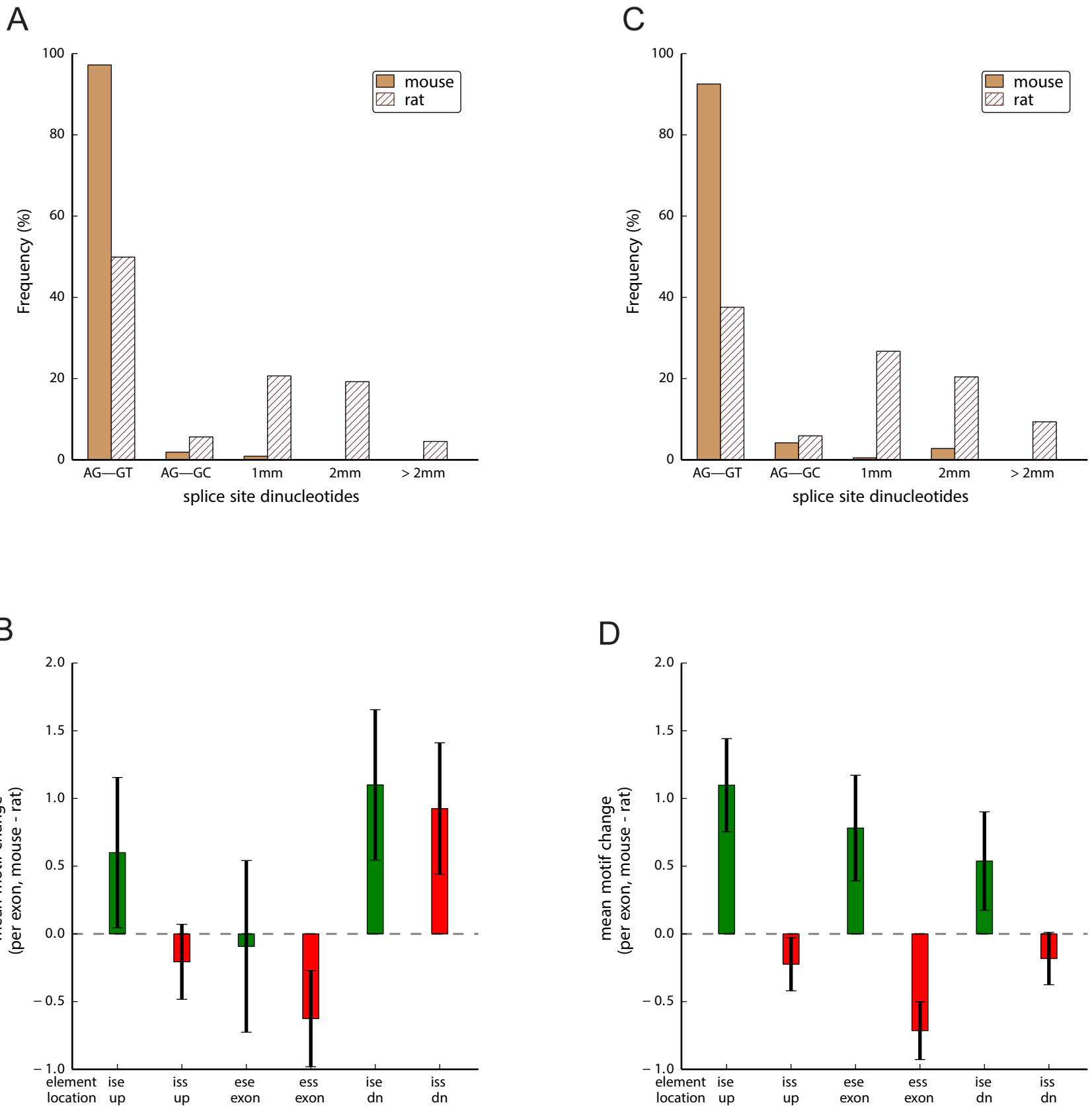
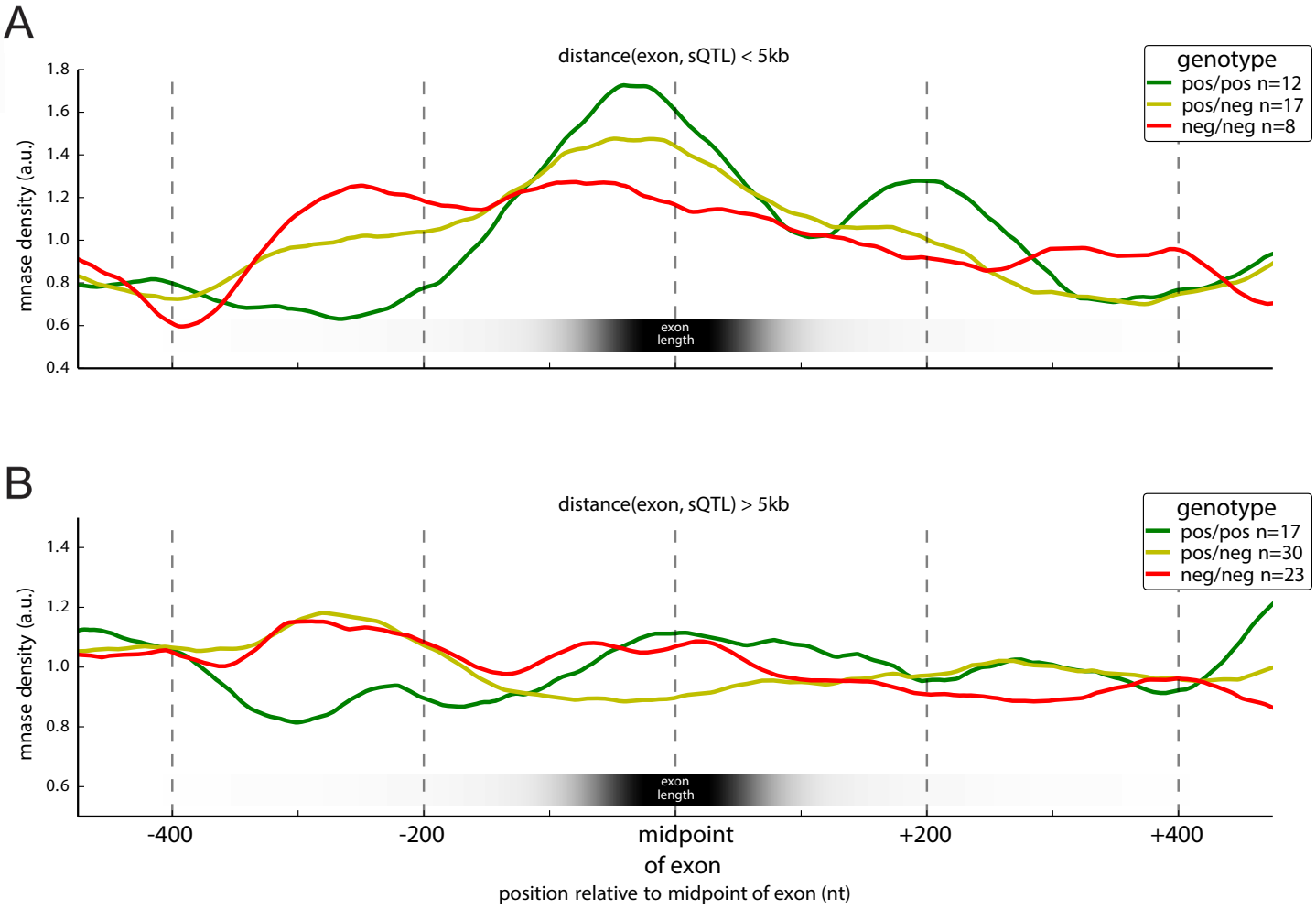

A
C
B
D
mouse
rat
Frequency (%)
AG—GT
AG—GC
1mm
2mm
> 2mm
splice site dinucleotides
mean motif change
(per exon, mouse - rat)
element
location
ise
up
iss
up
ese
exon
ess
exon
ise
dn
iss
dn

Figure s5


A
distance(exon, sQTL) < 5kb
genotype
pos/pos n=12
pos/neg n=17
neg/neg n=8
mnase density (a.u.)
exon length
B
distance(exon, sQTL) > 5kb
genotype
pos/pos n=17
pos/neg n=30
neg/neg n=23
mnase density (a.u.)
exon length
-400
-200
midpoint of exon
+200
+400
position relative to midpoint of exon (nt)

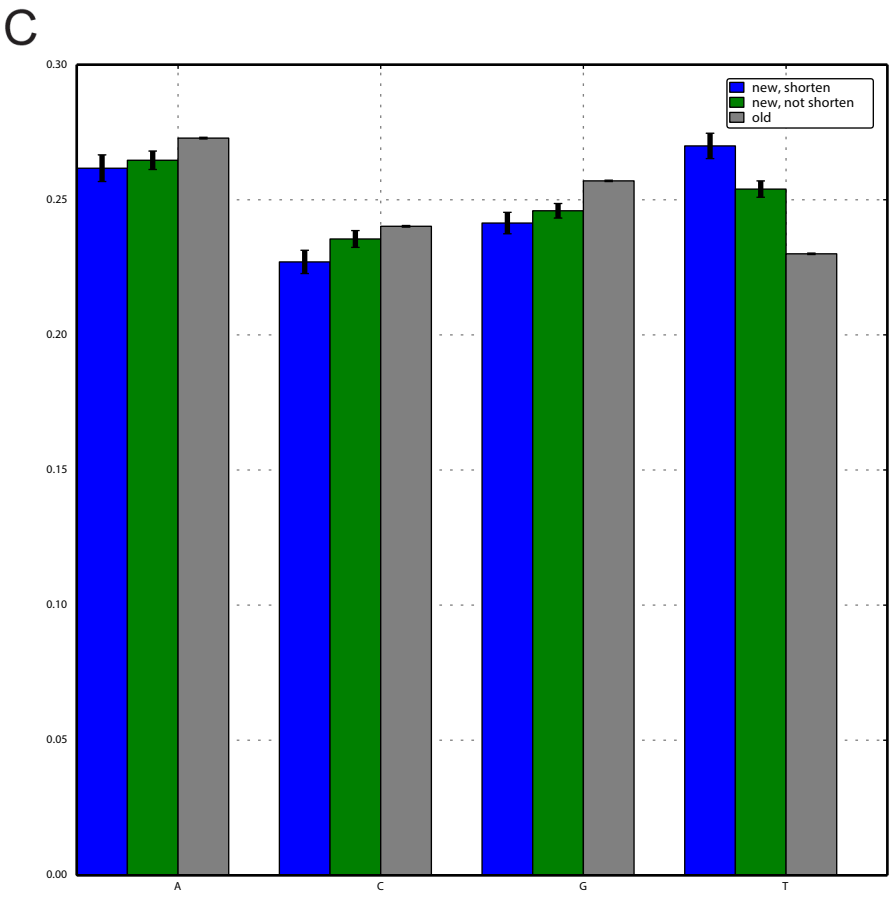
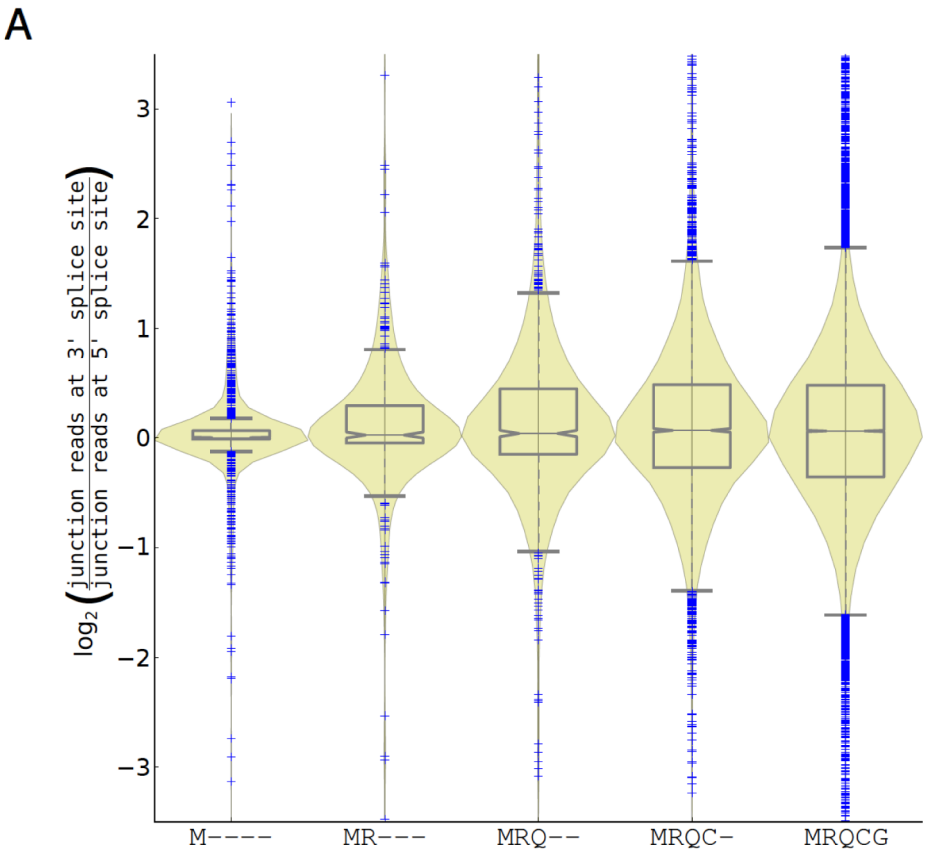

C
new, shorten
new, not shorten
old
A
C
G
T

Figure s6


A
log2 (junction reads at 3' splice site / junction reads at 5' splice site)
3
2
1
0
−1
−2
−3
M----
MR---
MRQ--
MRQC-
MRQCG

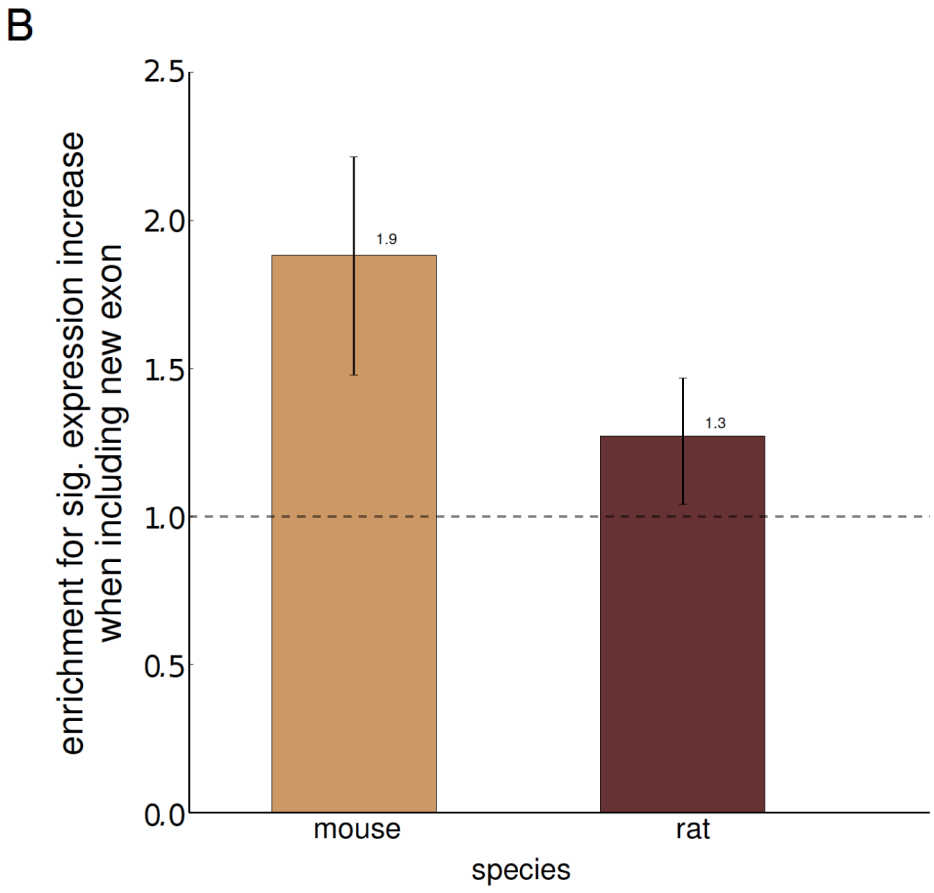

B
enrichment for sig. expression increase when including new exon
2.5
2.0
1.5
1.0
0.5
0.0
1.9
1.3
mouse
rat
species

Table S1. Summary of previous literature on exon evolution, Related to Figures 1-6.

Study	Year	Type of evidence	Species/lineages studied	Transcriptome coverage	Number of tissues	Number of exons found (from transcriptomic analysis)	Major findings
Modrek & Lee	2003	ESTs & cDNAs	human, mouse, rat	5 M human sequences 3.25 M mouse sequences 500 K rat sequences	NA	6,472 AS human exons 2,506 AS mouse exons 258 AS rat exons	- Alternatively spliced (AS) exons are mostly non-conserved, due to recent exon gain and/or loss
Wang et al.	2005	ESTs & cDNAs	mouse, rat (genome only), human (outgroup), pig (double-outgroup)	750 K pig sequences 4.5 M mouse sequences 7 M human sequences	NA	2,695 rodent new exons	- New exons tend to be alternative, lowly included, and rapidly evolving - Most new exons are derived from unique intronic sequences rather than repeats
Zhang & Chasin	2006	ESTs & cDNAs, genomic alignments	2 species used for transcriptomic analyses (human, mouse), 8 for genomic analyses (human, chimpanzee, dog, mouse, rat, chicken, zebrafish, fugu)	7.5 M human sequences 4.5 M mouse sequences	NA	2179 primate-specific exons 1249 rodent-specific exons	- 40% of new human exons are alternatively spliced and AS levels are anticorrelated with exon age - > 90% of primate-specific exons overlap repeats - Recent exons are preferentially located in the 5'UTR
Alekseyenko et al.	2007	ESTs & cDNAs, genomic alignments	5 species used for transcriptomic analyses (human, mouse, cow, dog, zebrafish), 17 for genomic analyses (across mammals, fish, and birds)	6.6 M human sequences 4.2 M mouse sequences 270 K dog sequences 722 K cow sequences 640 K zebrafish sequences	NA	167 alternative and 1999 constitutive human exons 207 alternative and 4546 constitutive mouse exons	- Exon creation rate is inversely correlated with exon inclusion level - Most non-repeat derived exons come from exaptation of unique intronic sequence
Calarco et al.	2007	custom microarray (designed to detect AS events selected by mining human ESTs & cDNAs)	human, chimpanzee, mouse	NA	2 tissues: frontal cortex, heart	~5000 AS human exons surveyed by microarray, 1700 of which were kept after filtering	- 6-8% of examined exons show differential splicing between human and chimp - Genes showing differential splicing are mostly orthogonal to those with differential expression
Sela et al.	2007	ESTs & cDNAs	human, mouse	7 M human sequences 4.5 M mouse sequences	NA	3477 human TE (transposable element) exonizations 1228 mouse TE exonizations	- TE families present in human and mouse have similar exonization levels - Human *Alu* exonization level is significantly greater than that of any other TE
Shen et al.	2011	RNA-seq (and RT-PCR)	human	123 M reads (dataset 1) + 90 M reads (dataset 2) single-end reads, length=32-36bp	2 tissues: cerebellum (dataset 1), liver (dataset 2)	287 *Alu*-derived exons (85 high-confidence)	- Highly included *Alu*-derived exons exhibit a 5'UTR bias and tend to decrease mRNA translational efficiency through the creation/elongation of uORFs - *Alu*-exonization is enriched in Zinc Finger Transcription Factors (ZNFs)
Merkin et al. (this study)	2015	RNA-seq, genomic alignments	6 species (mouse, rat, human macaque, cow, chicken)	3.2 B reads in macaque 4.4 B reads in mouse 3.9 B reads in cow 3.1 B reads in rat 3.2 B reads in chicken > 2.0 B reads in human (BodyMap 2.0) paired-end reads, length=36-80 bp (mostly 50 bp)	9 (matched) tissues	1089 mouse new exons 1571 rat new exons 1417 macaque new exons 2073 human new exons	Detailed in this paper

Table S3. Phylogenetic patterns of mouse exons, Related to Figure 1.
(1=presence, 0=absence)

M	R	Q	C	G	Number of exons
1	0	0	0	0	1134
1	0	0	1	0	65
1	0	0	1	1	248
1	0	1	0	0	75
1	0	1	0	1	120
1	0	1	1	0	293
1	0	1	1	1	2669
1	1	0	0	0	583
1	1	0	0	1	206
1	1	0	1	0	395
1	1	0	1	1	2386
1	1	1	0	0	450
1	1	1	0	1	1820
1	1	1	1	0	4115
1	1	1	1	1	40301

Table S4. Gene Ontology encrichment analysis for genes containing species-specific coding exons, Related to Figure 2.

term	p-value	benjamini	fdr	fold_enrich	genes
splice variant.UP_SEQ_FEATURE	3.19E-05	3.19E-05	3.85E-05	1.22536421	ENSMUSG00000024286:cyclin Y; similar to cyclin fold protein 11,…
alternative splicing.SP_PIR_KEYWORDS	1.51E-05	1.51E-05	1.93E-05	1.23569695	ENSMUSG00000024286:cyclin Y; similar to cyclin fold protein 11,…