Origins and impacts of new exons

April 22, 2014

Jason Merkin[1,*], Ping Chen[2,*], Sampsa Hautaniemi[2] and Christopher B. Burge[1,3]

[1]Department of Biology

Massachusetts Institute of Technology

Cambridge, MA 02142 USA

[2] Research Programs Unit

Genome-Scale Biology and Institute of Biomedicine

University of Helsinki

Helsinki, Finland

[3]Address correspondence to: cburge@mit.edu / 617 258-5997

* These authors contributed equally

Abstract

Mammalian genes are typically broken into several protein-coding and non-coding exons, but the evolutionary origins and functions of new exons are not well understood. Here, we analyzed patterns of exon gain using deep cDNA sequencing data from several mammals and one bird, identifying thousands of species- and lineage-specific exons. While exons conserved across mammals are mostly protein-coding and constitutively spliced, species-specific exons were mostly located in 5' untranslated regions and alternatively spliced. New exons most often derived from unique intronic sequence rather than repetitive elements, and were associated with upstream intronic deletions, increased nucleosome occupancy and RNA polymerase II pausing. Surprisingly, exon gain was associated with increased gene expression, but only in tissues where the exon was included, suggesting that splicing enhances steady-state mRNA levels and that changes in splicing represent a major contributor to the evolution of gene expression.

We recently studied alternative splicing of exons conserved across mammals using RNA-seq analysis of 9 diverse organs from 4 mammals and one bird, in biological triplicate(*1*). Here, we combined genomic mappings of these data(*2*) with whole-genome alignments(*3-5*) to classify exons as species-specific, lineage-specific (e.g., unique to rodents, to primates or to mammals), or ancient (present in both mammals and birds). These classifications were applied at both the genomic sequence level ("genomic age") and at the transcript level ("splicing age") (Fig. 1A). Using the principle of parsimony(*6-8*), we assigned both a genomic age and a splicing age to ~60,000 internal exons, restricting our analysis to unduplicated protein-coding genes in these species to facilitate accurate read mapping and assignment of orthology. Here, genomic age estimates the duration over which sequences similar to the exon were present in ancestral genomes, while splicing age estimates the duration over which these sequences were spliced in to mRNAs, based on the RNA-seq data.

Most exons in the analyzed genes (~85%) predated the split between birds and mammals (~300 million years ago, Mya) in their splicing (and genomic) age; such exons are designated `MRQCG` using a one-letter code for the five organisms (Fig. 1B). However, we found that creation of novel exons has occurred fairly often during mammalian evolution. For example, we classified 1089 mouse exons as mouse-specific (designated `M----`), as they were detected in RNA-seq data from mouse but not from any other species (Fig. 1B; estimated false discovery rate ~1.5% using the approach of(*9*)). These exons were assigned an age of < 25 My, corresponding to the time of divergence between mouse and rat. We also identified ~7000 mouse exons whose splicing was restricted to particular mammalian lineages (Fig. 1). Overall, presence of one or more mouse- or rodent-specific exons was detected in 17% of the ~6300 genes analyzed (Methods). To ask whether species-specific exons occurred at a similar

frequency in human, we compared our data to corresponding tissue RNA-seq data from the Illumina Human Body Map 2.0 dataset(*10*). Although the human tissues were sequenced at somewhat lower depth, we identified similar numbers of exons at each splicing age (Fig. S1), including ~2000 human-specific exons (not observed even in macaque), occurring in 25% of analyzed human genes. Together, these observations indicate that, when comparing a human gene to its ortholog in the most commonly used mammalian model, the mouse, 35% of ortholog pairs will differ by presence/absence of a species-specific exon. The prevalence of such species-specific exons could contribute to widespread functional differences between human and mouse orthologs, complicating extrapolations from mouse models to human phenotypes.

To assess whether species-specific exons have defined tissue-specific splicing patterns, we performed clustering of exons and tissue samples based on the tissue-specific splicing patterns of mouse-specific alternative exons. This analysis revealed robust clustering by tissue of origin across the three mouse strains analyzed (Fig. 2F). The only deviation from this pattern was some overlap between cardiac and skeletal muscle, which was also seen when considering ancient (MRQCG) alternative exons(*1*), consistent with similarity between the splicing programs of these developmentally related tissues. A substantial fraction of novel exons showed predominant inclusion in testis (Fig. 2F), similar to the pattern of testis-biased expression observed for novel (species-specific) genes(*11*). We also found that genes containing species-specific constitutive exons are enriched for testis expression (Fig. S2). These observations suggest a role for germ cell transcription in exon creation and are consistent with previous studies indicating that germ cell transcription can increase the frequency of mutations, presumably including those that give rise to novel exons(*12*).

Species-specific exons are mostly alternatively spliced and located in 5' UTRs

In many respects, exons of different evolutionary ages had dramatically different properties. While constitutive splicing was the norm for ancient (MRQCG) exons in these data, the vast majority of species-specific exons were alternatively excluded (skipped) in at least one tissue (Fig. 2A)(*7*). Similarly, ancient exons were mostly located within the open reading frame (ORF), while most species-specific exons were located in non-coding regions (Fig. 2B). New exons occurred with much higher frequency in 5' untranslated regions (UTRs) than in 3' UTRs (Fig. 2C). Various factors may contribute to the bias for 5' UTRs, including the greater length of first introns relative to later introns(*13*), the low frequency of 3' UTR introns(*14*), and the potential for some new 3' UTR exons to destabilize messages via the nonsense-mediated mRNA decay (NMD) pathway. By contrast, non-coding ancient exons were located predominantly in non-coding transcripts deriving from the set of protein coding genes analyzed (Fig. 2C). The set of species-specific coding exons was too small to analyzed systematically, but we did note some examples in which new exons encode predicted signal peptides or transmembrane domains (Fig. S3) – two types of protein motifs that have relatively relaxed sequence requirements.

Most species-specific exons arise by exaptation of intronic sequences

Our classification pipeline used sequence similarity filters to exclude ~6800 mouse exons that likely arose from intra-genic duplications, a class that has been well studied previously(*15*). Therefore, the novel exons studied here must have arisen by insertions into introns(*16*) or by exaptation of pre-existing intronic sequence(*17*). To compare the relative contributions of these two mechanisms, we analyzed the genomic age of each recently created exon. Approximately 1% of mouse-specific exons arose in sequence detected only in mouse while ~¾ of mouse-

specific exons arose in sequence that predates the rodent-primate split, despite being spliced exclusively as intron in the other species studied, and the remainder were alignable to rat only (Fig. 2D). Exons with rodent-specific or rodent/primate-specific splicing also could often be aligned to cow or chicken (Fig. 2D and data not shown). Using our RNA-seq data derived from three different mouse strains, we observed that nearly a quarter of mouse-specific exons were detected in just two of the three strains, while for ancient exons nearly 99% were detected in all three mouse strains. This observation could result from incomplete detection of (often lowly-expressed) species-specific exons, or from polymorphic splicing of such exons; distinguishing between these possibilities would require substantial additional sequencing.

We next sought to identify features associated with new exon creation. We observed that more than 60% of new internal exons in mouse are derived from unique intronic sequence. In most cases, these exons aligned to sequences in the orthologous intron in rat (Fig. 3A). Applying a similar approach to identify novel exons in human – using criteria designed to allow mapping to repetitive elements (Methods) – yielded a similarly high proportion of unique mapping (~54%) (Fig. S4A). Alu elements, a class of primate-specific SINE repeats, have previously been implicated as a major source of new exons in primates(*16, 18, 19*). Here, we found that ~19% of exons we classified as human-specific overlap with Alus (Fig. S4), and a similar proportion of mouse-specific exons (~18%) overlap with rodent SINEs, which are also thought to derive from 7SL RNA(*20*) (Fig. 3B). Thus, our analysis indicated that rodent SINEs have contributed to new exon creation in rodents to a similar extent as Alus have in primates.

The observed proportions of SINE-derived exons exceeded the genomic background frequencies of SINE elements (Fig. 3C; Fig. S4B), but were 2- to 3-fold lower than the proportions that derived from unique genomic sequence in both organisms. This observation

contrasts with previous suggestions that Alus are a predominant source of new exons in primates, instead demonstrating the primacy of unique sequences. The differences in conclusions from previous studies likely result from differences in data sources (RNA-seq versus EST) and analytical procedures that are expected to make our analysis less biased and more sensitive to detection of low-abundance isoforms(*18*). Other types of repetitive elements (LINEs, LTRs and others) have together contributed slightly more species-specific exons than have SINEs in both human and mouse (Fig. 3B, Fig. S4A).

Altered splicing motifs and shortened upstream introns are associated with exon creation

Mutations that create or disrupt splice site motifs frequently cause changes in splicing patterns over evolutionary time periods(*10, 16*). While the vast majority of mouse-specific exons contained minimal splice site dinucleotides (GT or GC at the 5' splice site, AG at the 3' splice site), ~47% of homologous "proto-exon" sequences in rat lacked these minimal splicing motifs (Fig. 3D). This observation suggests that mutations that create splice site dinucleotides may contribute to up to about half of exon creation events. However, other types of changes must explain the remaining cases, where minimal splice site motifs were present in rat but no evidence of splicing was detected in rat tissues. For about 43% of these remaining cases, a substantive increase in the strength of one or both extended splice site motifs was observed in mouse (Fig. S5), suggesting that strengthening of existing minimal splice site motifs also contributes to creation and/or maintenance of new exons, but that other types of changes must also play a major role.

Motifs present in the body of an exon or in the adjacent introns can enhance or suppress exon inclusion(*21*). We found that mouse-specific exons contain a higher density of exonic

splicing enhancer (ESE) motifs and a lower density of exonic splicing silencer (ESS) motifs than their associated rat proto-exons (Fig. 3E). Thus, both gain of enhancing motifs and loss of silencing motifs are likely to contribute to creation and/or maintenance of novel exons. No significant difference in density of intronic splicing silencer (ISS) motifs was seen, but we did observe a higher density of intronic splicing enhancer (ISE) motifs adjacent to mouse-specific exons relative to homologous rat sequences, suggesting that changes in flanking sequences may also contribute to exon creation (Fig. 3D). Together, the high frequency of changes to splice site and regulatory motifs associated with new exons suggests that most exon creation results from changes in *cis* rather than changes to the levels of activities of *trans*-acting factors.

Intron length is associated with a number of splicing properties, and longer flanking introns tend to be associated with lower inclusion levels of alternative exons(*22*). We therefore asked whether changes in intron length might be associated with species-specific exons. Notably, we found that the distance between the exons flanking `M----` exons was shorter on average than the distance between the homologous exons in rat (rat distance exceeded mouse by 1.3-fold on average; interquartile range: 0.9-fold to 1.7-fold; Fig. 3F). The distance between the exons flanking `-R---` exons was even shorter, with the distance between homologous mouse exons exceeding that of rat-specific exons by 1.7-fold on average (Fig. S6). These observations suggest that substantial changes in intron length often accompany exon creation. Comparing the lengths of the introns flanking each species-specific exon to each other, we observed that the intron downstream of `M----` exons was 1.2-fold longer on average than the upstream intron, compared to no difference between the homologous regions in rat (Fig. 3G), with a somewhat smaller effect observed for `-R---` exons (Fig. S6). When examining rodent-specific exons, we observed a similar bias towards presence of a longer downstream intron in both mouse and rat.

Comparison to an outgroup (rhesus macaque) indicated that the differences in flanking intron length in the rodent lineages that acquired new exons most often reflect upstream deletions rather than downstream insertions (Fig. S7). Older groups of exons showed no such bias, suggesting that exons may acquire tolerance for expansion of the upstream intron over time, as other splicing determinants strengthen. Together, these data suggest that deletions upstream of proto-exons favor creation and/or maintenance of novel exons. Previously, shortening of an upstream intron was associated with enhancement of exon inclusion in a minigene reporter context(*23*), but the generality of this effect and its evolutionary impact have not been explored.

Having identified a number of genomic changes associated with species-specific exons, we asked about the relative contributions of each. To compare the magnitudes of these different types of changes using a standard scale, we converted them all to z-scores, using the standard deviation of each type of change observed between mouse and rat in ancient (MRQCG) exons as a scale. We observed relatively small z-scores (< 0.4) for changes to each *cis*-motif type. However, upstream intronic deletions had an average z-score of ~ 0.75, comparable to the sum of the z-scores of all cis-motifs analyzed (Fig. 3H). This observation suggests that upstream intronic deletions may contribute to creation of novel exons to an extent comparable to that of changes in known classes of splicing regulatory elements.

Upstream indels are associated with nucleosome occupancy and RNA Pol II pausing over novel exons

Intron length can impact splicing in multiple ways. Shortening of introns can promote exon inclusion in splicing reporters, with a larger effect observed for the upstream intron(*23*), possibly by promoting intron definition or by enhancing exon juxtaposition following exon definition.

Lengthening of the downstream intron can also impact splicing through effects on the kinetics of transcription relative to splicing(*24*). Conceivably, changes in the lengths of either flanking intron might also promote exon creation by impacting nucleosome positioning. Suggesting functional links between nucleosomes and splicing, nucleosomes are often positioned near the centers of internal exons, exon-associated nucleosomes have higher density of the H3K36me3 histone modification(*25-27*), and histone modifications can impact recruitment of splicing factors(*28*). We used published micrococcal nuclease (MNase) sequencing data from digestion of chromatin to identify nucleosome-protected regions in the vicinity of mouse-specific exons. We observed a stronger enrichment for nucleosome positioning over mouse-specific exons which had shortened upstream introns (relative to rat) compared to ancient exons or to mouse-specific exons without upstream shortening, or to mouse regions orthologous to rat-specific exons (Fig. 4A; $P < 10^{-4}$ for all three comparisons by modified Kolmogorov-Smirnov (KS) test). This association suggested a connection between upstream deletions and changes in nucleosome positions. While indels in either the upstream or downstream intron could potentially impact nucleosome positioning on an exon, upstream deletions may be more likely than other types of indels to promote exon inclusion by the other mechanisms discussed above. Therefore, upstream deletions are likely enriched near new exons both because they promote intron definition or exon juxtaposition, and possibly because they can alter nucleosome positioning. Because the relationship between nucleosome positioning and splicing is less understood, we chose to further explore this potential connection.

It has been proposed that nucleosomes can function as molecular "speed bumps" to slow down RNA polymerases as they transcribe through exons(*24*). This effect may contribute to exon inclusion by increasing the time available for splicing machinery associated with RNA

polymerase II to associate with the exon and commit it to splicing(*29, 30*). Mutations that slow down RNA polymerase elongation are reported to enhance recognition of exons with weak splice sites(*31*). We hypothesized that changes promoting stronger nucleosome positioning over novel exons might slow polymerase elongation and thereby act to promote splicing. To test this hypothesis, we used available global run-on-sequencing (GRO-seq) data, which detects nascent transcription. Using data from ref(*32*), we observed a strong GRO-seq peak over ancient exons, almost twice the background level (Fig. 4B; $P < 10^{-4}$, modified KS test). This observation suggests that polymerases slow down by almost twofold while transcribing through these exons; much smaller effects in this direction have been observed using Pol II ChIP-seq data(*25, 27*). When considering mouse-specific exons with an upstream intronic deletion, we observed a GRO-seq peak ~37% above the nearly flat background of mouse-specific exons without upstream deletions ($P < 10^{-4}$, modified KS test). Thus, although further studies will be required to confirm, the observations above suggest a model in which upstream deletions that enhance nucleosome positioning on exons may slow polymerase elongation and enhance inclusion of new exons.

Recent studies of the genetic basis for gene expression variation have also identified thousands of genetic variants associated with altered levels of splicing between human individuals(*33-35*). If changes in nucleosome positioning often impact splicing, as implicated above, we reasoned that polymorphisms associated with variation in splicing (splicing-quantitative trait loci or sQTLs) that are structural variants located in introns upstream of the associated splicing event might exert their effects through impacts on nucleosome positioning in the vicinity of the affected exons. To explore this idea, we used sQTLs identified in genotyped human lymphoblastoid cell lines studied by the GEUVADIS Consortium(*35*) and MNase-seq

data from a subset of these individuals(*36*). This combination allowed us to evaluate the nucleosome positioning in individuals with different genotypes at the associated sQTL. Considering upstream intron indel sQTLs of at least a minimum length, we observed that the genotype associated with increased exon inclusion also had increased nucleosome density in the vicinity of the associated exon (Fig. 4C; $P < 0.01$, modified KS test). We also found that stronger differences in nucleosome density were associated with indel sQTLs located closer to the affected exon, as expected if these indels directly impact nucleosome placement (Fig. S8). Together, the data in Figure 4 implicate upstream intronic indels in changes in nucleosome positioning and splicing between both species and individuals.

New exon splicing is associated with species-specific increases in expression

We next asked what effects new exons have on the genes in which they arise. Since the majority of species-specific exons we identified were non-coding (Fig. 2B), we examined effects on gene expression. Intron-mediated enhancement is a well-established though incompletely understood phenomenon in which introduction of a (possibly heterologous) intron or exon into a gene or minigene often leads to higher expression of the gene(*37-39*), through effects on mRNA export, cleavage/polyadenylation, stability or other mRNA properties(*40, 41*). Furthermore, studies of mRNA stability have often observed strong correlations between the density of exon junctions in the open reading frame and mRNA half-life (*42, 43*). During evolution, creation of a novel exon in an intron will increment the number of introns and exons in a transcript. Here, we observed significantly higher expression (in mouse) of genes containing mouse-specific exons relative to their rat orthologs in corresponding rat tissues (Fig. 5A). This effect was specific to those mouse tissues where the new exon was included ("spliced in" to the mRNA), consistent with a positive

effect of splicing on steady state expression levels (Fig. 5A). The inclusion of a new exon was associated with an average increase in gene expression of ~10% (Fig. 5A, inset). To confirm our classification of these mouse-specific exons as internal exons (Fig. 1), we counted RNA-seq reads supporting their 3' and 5' splice junctions. We observed similar 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. S9).

An alternative way to measure the effects of the splicing of a new exon on expression is to measure the ratio of the mean expression in tissues where the exon is included to the mean expression in tissues where the exon is excluded, a ratio we call the "exon-associated expression index" (EEI). Dividing the EEI in the species which contains the exon to the ratio of mean expression values in the same sets of tissues in the species lacking the exon yields an exon-associated expression ratio (EER), values of which should be distributed symmetrically around one (on a log scale) under the null hypothesis that the splicing of the new exon does not affect gene expression. This approach controls for certain technical factors that could impact estimation of expression levels between different species. Comparing EER values for genes containing mouse-specific exons or rat-specific exons to shuffled controls (Fig. 5B), we observed significantly elevated ratios (~1.1) in both cases, consistent with the 10% increase in expression observed above (Fig. 5A), and further supporting the alternative hypothesis that "splicing in" of new exons enhances gene expression.

These observations suggest a widespread impact of splicing on gene expression. To further explore this phenomenon, we considered exons whose presence in the transcriptome is ancient, but which undergo exon skipping only in mouse (38). We observed that the species-specific skipping in mouse of these exons was associated with lower gene expression relative to

rat, suggesting that reduced splicing of these genes in mouse contributes to reduced expression (Fig. 5C). Furthermore, we observed a dose-dependent effect, where lower exon inclusion was associated with a stronger decrease in gene expression (Fig. 5D). We also observed a positional effect, with the strongest effects on expression associated with exons located at the 5' ends of genes, and virtually no effect seen for exons located in the 3' end of the gene (Fig. 5E). Most of the species-specific exons analyzed in Figure 5A were located near the 5' ends of genes (Fig. 2C). It should be noted that our ability to detect exon skipping and to detect novel exons both require a certain minimum level of expression, raising the possibility of detection biases. However, we have taken measures to counteract such biases (Methods), and we note that any such bias would tend to reduce the effect on expression observed in Fig. 5C-E, not enhance it.

Analyzing genes that have exhibit species-specific increases in gene expression, we observed substantial enrichment for genes containing species-specific exons (Fig. S10). This observation, together with the results shown in Figure 5, suggest that splicing has a general role in promoting gene expression and that changes in splicing may represent a major contributor to the evolution of gene expression.

Methods

RNA-seq and genome builds

Data from from mouse, rat, rhesus, cow, and chicken were processed as in(*1*) using TopHat v1.1.4(*2*) and Cufflinks v1.0.2(*44*). 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(*1*) were used in this analysis. As done in that study, we only considered single copy genes. We flagged and removed 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(*3*), and pairwise alignments from BLASTZ(*45*). In addition, we further attempted to align exons without a genomic aligned region not expressed in chicken to the genome of each species using BLAT(*5*) to reduce a false negative rate of finding an aligned region, taking the best matching region and requiring a minimum 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 value 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 interpret 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(*6, 46*). We only consider unambiguous age assignments (i.e. if there are multiple equally parsimonious assignments that would yield different ages, then the exon is 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 0 (new), 25, 90, 110 and 300 if there were aligned regions in rat, rat/rhesus, rat/rhesus/cow and rat/rhesus/cow/chicken. Similarly, its splicing age was assigned to similar categories if there were aligned regions expressed in rat, rat/rhesus, rat/rhesus/cow and rat/rhesus/cow/chicken (Fig. 1B).

Note: we only considered exons detected in the previous RNA-seq study(*1*). This was done to mitigate the effects of prior transcript annotation quality on our results since, for instance, mouse and rhesus annotations (by proxy from converting human annotations) would be expected to be much better than cow or rat. This approach will miss annotated exons only included in embryonic tissues, for instance, 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 we don't have comparable data.

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 were categorized as skipped exons (SE) while exons with PSI > 97 in all expressed tissues were defined as constitutive exons (CE) for each individual. We required an exon be evaluated in 3 or more tissues for this classification, since the probability of detecting exon skipping increases with the number of tissues considered. In Fig. 2A, the proportion of exons that are skipped or constitutive were calculated by SE/(SE+CE) and CE/(SE+CE) respectively, where SE and CE represents the number of alternative spliced exons and the number of constitutive exons.

Transcripts' open-reading frames (ORFs) were annotated as in(*1*). 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 Fig. 2B, 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 Fig. 3A, exons were categorized into intronic, 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 which does not overlap any gene, 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 the 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 (poorly sampled categories with low counts). Exons not overlapping any repeats were assigned to the “unique" group in Fig. 3B.

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 Fig. 3C. In Fig. 3D, exonic splicing enhancers (ESEs) from(*47*), exonic splicing silencers (ESSs) from(*48*), intronic splicing enhancers (ISEs) from(*49*), and intronic splicing silencers (ISSs) from(*50*) 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 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 (Fig. 3E). For Fig. 3F, the length downstream mouse intron was divided by the length of the upstream mouse intron. A similar ratio was calculated in rat, where the downstream intron (or the remainder of the intron downstream of the aligned exon region for mouse-specific exons) was divided by the upstream intron (or upstream remainder of the intron).

Z-score conversion for comparisons

For each change considered (changes in ISEs, ISSs, ESEs, ESSs, 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. For each new exon, Each change was then calculated for each new exon and converted to a z-score using the values calculated in the ancient group.

Nucleosome localization and GRO-seq analyses

We downloaded the MNase-seq data from(*36*) from GEO (accession no. GSE36979). We mapped the reads with Bowtie v0.12.7(*51*) 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(*52*) 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(*32*) 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 normalize each region, this should not affect our results. These data were then processed in the same manner as the MNase-seq data.

To investigate the impact of intronic structural variants on nucleosome localization (Fig. 5C), we downloaded the following files:

- the sQTL table EUR373.exon.cis.FDR5.all.rs137.txt.gz from the Geuvadis consortium(*35*),
- Gencode v12 ref(*53*), matching the annotations used in the Geuvadis study from http://www.gencodegenes.org,
- MNase-seq data from individuals included in the Geuvadis study from(*36*),
- Genotype data for these individuals from the 1000 Genomes Project(*54*) 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 (*35*) 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.

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 / rat using gene expression from(*1*). We considered the following cases in Fig. 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 (Fig. 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 (Fig. S10) 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, downsampling to match the shorter length. We then applied DEseq(*55*) 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 (Fig. 5C), 2) for exons alternative spliced in mouse with different inclusion levels but constitutive in rat (Fig. 5D) and 3) for first and last exons alternative spliced in mouse but constitutive in rat (Fig. 5E). For case 1), we used ancient exons included in both species as control whereas for case 2) and 3), we used constitutive exons included in both mouse and rat as control.

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.

Figure Legends

Figure 1. Identification and classification of species- and lineage-specific exons.

(A) A schematic diagramming our bioinformatic pipeline to identify species- and lineage-specific exons (Methods). We considered every exon in the target species (here, mouse) and aligned it to other exons in the same gene to filter out exons arising from exon duplication(*56*). We filtered out initial and terminal exons to focus on analyses of splicing. We used multiple alignments between species studied here to assign an orthologous region to each exon in other species and used parsimony to interpret the pattern of genomic presence or absence as the genomic age. We then looked for an overlapping exon in the orthologous region to determine if the mouse exon was spliced in a given species, and interpreted this pattern of presence or absence of splicing as a splicing age.

(B) 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 2. Evolutionarily young exons differ from older exons in many properties.

(A) The proportions of exons of various ages that are alternatively or constitutively spliced is shown.

(B) The proportion exons of various ages that contain coding sequence (CDS) or are entirely non-coding (NC) is shown.

(C) The proportions of non-coding exons of various ages that are located in non-coding transcripts (nc tx), or in 5' or 3' UTRs of coding transcripts is shown.

(D) The distributions of genomic ages of exons with splicing ages `M----` or `MRQ--` are shown.

(E) The proportion of mouse exons of various ages that were detected in only 2 out of 3 individuals or where the splicing status (alternative or constitutive) in one individual

differed from the other two mice.

(F) Average-linkage hierarchical agglomerative clustering of samples (vertical axis) or exons (horizontal axis) based solely upon PSI values of mouse-specific exons. The tissue of origin of each sample is colored according to the key at left and the PSI value is visualized in the heat map (center).

Figure 3. A variety of genomic changes are associated with novel exon splicing.

(A) Proportion of mouse-specific exons that map to different classes of genomic regions in rat.

(B) Proportion of mouse-specific exons that overlap with various classes of repeats.

(C) Proportion of mouse genome that belong to various repeat categories.

(D) The proportion of new mouse exons with specific splice site dinucleotide sequences in mouse and rat.

(E) The change in splicing regulatory element number in various regions in and around a new exon associated with its creation (mean ± SEM).

(F) The change in length of the entire intron region between rat and mouse. The length in rat is plotted as a percentage of the length in mouse (mean ± SEM).

(G) 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 (rat) (mean ± SEM). The rat bar in the `M----` class is hatched to represent the fact that it is not an exon in rat.

(H) The magnitude of each change associated with splicing of `M----` exons was converted into a z-score based upon the distribution of such changes between mouse and rat in MRQCG exons. Motifs that are expected to promote splicing are colored in green and changes that are expected to inhibit splicing are shown in red.

Figure 4. Upstream intronic deletions are associated with increased exonic nucleosome occupancy and transcription pausing.

(A) Nucleosome positioning (measured by MNase protection) around various sets of exons.

(B) Density of global run-on sequencing (GRO-seq) reads, showing the position of elongating RNA Pol II.

(C) Nucleosome positioning (measured by protection from MNase treatment) around exons with a structural sQTL in the upstream intron binned by sQTL genotype.

Figure 5. Inclusion of new exons is associated with increased species-specific gene expression changes.

Throughout the figure, statistical significance by Mann-Whitney U test is indicated by asterisks (* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$, ***** $P < 0.00001$).

(A) Fold change in gene expression between mouse and rat. Inset: mean ± SEM of displayed distributions.

(B) Mean EEI values, calculated as the ratio of the mean gene expression in tissues where a novel exon is included to the mean expression in tissues where inclusion of the exon is not detected. This ratio is calculated in a related species with matched tissues, and the ratio of these two values is plotted (mean ± SEM).

(C) Fold change in gene expression between mouse and rat in genes where an ancestrally present exon has become skipped in mouse.

(D) Fold change in gene expression between mouse and rat in genes where an old exon has become skipped in mouse, binned by the PSI of the exon in the tissue.

(E) Fold change in gene expression between mouse and rat in genes where an old exon has become skipped in mouse, binned by location of the exon within the gene.

Supplementary Figure Legends

Figure S1. Number of exons of various ages in the human lineage.
(As in Fig. 1B.) 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. Expression clustering of mouse-specific constitutive exons.
Average-linkage hierarchical agglomerative clustering of samples (rows) and exons (columns) based on gene expression values of genes containing mouse-specific constitutively spliced exons (as in Fig. 2F). The tissue of origin of each sample is colored according to the key at left and expression levels are visualized in a heat map.

Figure S3. Examples of new exons that alter protein sequence features.
(A) A portion of the mouse aprataxin (APTX) gene is shown (Ensembl ID ENSMUSG00000028411), together with homologous sequences from rat. For each, chromosome coordinates (mouse chr4, rat chr5) are shown at bottom, with RNA-seq read density from three tissues shown above in green, light blue and red (arcs represent splice junction reads), and transcript structures shown in dark blue (the gray region in rat is the genomic segment homologous to the mouse-specific exon). Predicted protein features are shown at top: the mouse-specific exon is predicted to encode an N-terminal signal peptide based on analysis by SignalP (*57*).
(B) As in part (A), 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 chr8. This region contains a mouse-specific exon that is predicted to encode transmembrane domains by TMHMM (*58*).

Figure S4. Repeat class of human-specific exons.
As in Fig. 3B,C.
(A) Proportion of human-specific exons that match various categories of repeats.
(B) Proportion of human genome belonging to various repeat categories.

Figure S5. Splice site strength changes associated with exons already containing minimal splice sites.
Splice site strength was measured by the MaxEnt method(*59*).

Figure S6. Intron length changes associated with rat-specific exons.
As in Fig. 3F-G.
(A) The change in length of the entire intron region between rat and mouse. The length in mouse is plotted as a percentage of the length in rat (mean ± SEM).
(B) 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). The mouse bar in the `-R---` class is hatched to represent the fact that it is not an exon.

Figure S7. Intron length changes associated with mouse-specific exons compared to macaque (instead of rat as in Fig. 3F,G).
(A) The change in length of the entire intron region between macaque and mouse. The length in macaque is plotted as a percentage of the length in mouse (mean ± SEM).
(B) 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). The macaque bars in the `M----` and `MR---` classes are hatched to represent the fact that they are not an exon.

Figure S8. Nucleosome positioning near human exons with upstream structural sQTLs.
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.

Figure S9. Comparison of 3' splice site and 5' splice site junction read density.
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.

Figure S10. New exons in genes with species-specific changes in expression.
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.

References

1. J. Merkin, C. Russell, P. Chen, C. B. Burge. *Science (New York, NY)* **338**, 1593 (Dec 21, 2012).

2. C. Trapnell, L. Pachter, S. L. Salzberg. *Bioinformatics* **25**, 1105 (May 1, 2009).

3. B. Paten, J. Herrero, K. Beal, S. Fitzgerald, E. Birney. *Genome research* **18**, 1814 (Nov, 2008).

4. M. Blanchette *et al*. *Genome research* **14**, 708 (Apr, 2004).

5. W. J. Kent. *Genome research* **12**, 656 (Apr, 2002).

6. A. V. Alekseyenko, N. Kim, C. J. Lee. *RNA* **13**, 661 (May, 2007).

7. B. Modrek, C. J. Lee. *Nature genetics* **34**, 177 (Jun, 2003).

8. X. H. Zhang, L. A. Chasin. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 13427 (Sep 5, 2006).

9. C. B. Nielsen, B. Friedman, B. Birren, C. B. Burge, J. E. Galagan. *PLoS biology* **2**, e422 (Dec, 2004).

10. R. K. Bradley, J. Merkin, N. J. Lambert, C. B. Burge. *PLoS biology* **10**, e1001229 (Jan, 2012).

11. A. C. Marques, I. Dupanloup, N. Vinckenbosch, A. Reymond, H. Kaessmann. *PLoS biology* **3**, e357 (Nov, 2005).

12. M. T. Levine, C. D. Jones, A. D. Kern, H. A. Lindfors, D. J. Begun. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 9935 (Jun 27, 2006).

13. E. V. Kriventseva, M. S. Gelfand. *Journal of biomolecular structure & dynamics* **17**, 281 (Oct, 1999).

14. C. Giorgi *et al*. *Cell* **130**, 179 (Jul 13, 2007).

15. X. Gao, M. Lynch. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 20818 (Dec 8, 2009).

16. G. Lev-Maor, R. Sorek, N. Shomron, G. Ast. *Science* **300**, 1288 (May 23, 2003).

17. J. Brosius, S. J. Gould. *Proceedings of the National Academy of Sciences of the United States of America* **89**, 10706 (Nov 15, 1992).

18. Z. Wang, M. Gerstein, M. Snyder. *Nature reviews. Genetics* **10**, 57 (Jan, 2009).

19. R. Sorek, G. Ast, D. Graur. *Genome research* **12**, 1060 (Jul, 2002).

20. J. O. Kriegs, G. Churakov, J. Jurka, J. Brosius, J. Schmitz. *Trends in genetics : TIG* **23**, 158 (Apr, 2007).

21. A. J. Matlin, F. Clark, C. W. Smith. *Nat Rev Mol Cell Biol* **6**, 386 (May, 2005).

22. G. W. Yeo, E. Van Nostrand, D. Holste, T. Poggio, C. B. Burge. *Proc Natl Acad Sci U S A* **102**, 2850 (Feb 22, 2005).

23. K. L. Fox-Walsh *et al*. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 16176 (Nov 8, 2005).

24. D. L. Bentley. *Nature reviews. Genetics* **15**, 163 (Mar, 2014).

25. N. Spies, C. B. Nielsen, R. A. Padgett, C. B. Burge. *Molecular cell* **36**, 245 (Oct 23, 2009).

26. H. Tilgner *et al*. *Nature structural & molecular biology* **16**, 996 (Sep, 2009).

27. S. Schwartz, E. Meshorer, G. Ast. *Nature structural & molecular biology* **16**, 990 (Sep, 2009).

28. R. F. Luco *et al*. *Science* **327**, 996 (Feb 19, 2010).

29. F. Q. Gunderson, E. C. Merkhofer, T. L. Johnson. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 2004 (Feb 1, 2011).

30. H. L. Zhou *et al*. *Proceedings of the National Academy of Sciences of the United States of America* **108**, E627 (Sep 6, 2011).

31. J. Y. Ip *et al*. *Genome research* **21**, 390 (Mar, 2011).

32. M. U. Kaikkonen *et al*. *Molecular cell* **51**, 310 (Aug 8, 2013).

33. E. Lalonde *et al*. *Genome research* **21**, 545 (Apr, 2011).

34. J. K. Pickrell *et al*. *Nature* **464**, 768 (Apr 1, 2010).

35. T. Lappalainen *et al*. *Nature* **501**, 506 (Sep 26, 2013).

36. D. J. Gaffney *et al*. *PLoS genetics* **8**, e1003036 (2012).

37. J. J. Jonsson, M. D. Foresman, N. Wilson, R. S. McIvor. *Nucleic acids research* **20**, 3191 (Jun 25, 1992).

38. D. Mascarenhas, I. J. Mettler, D. A. Pierce, H. W. Lowe. *Plant molecular biology* **15**, 913 (Dec, 1990).

39. R. D. Palmiter, E. P. Sandgren, M. R. Avarbock, D. D. Allen, R. L. Brinster. *Proceedings of the National Academy of Sciences of the United States of America* **88**, 478 (Jan 15, 1991).

40. A. Nott, S. H. Meislin, M. J. Moore. *RNA* **9**, 607 (May, 2003).

41. S. Lu, B. R. Cullen. *RNA* **9**, 618 (May, 2003).

42. N. Spies, C. B. Burge, D. P. Bartel. *Genome research* **23**, 2078 (Dec, 2013).

43. L. V. Sharova *et al*. *DNA research : an international journal for rapid publication of reports on genes and genomes* **16**, 45 (Feb, 2009).

44. C. Trapnell *et al*. *Nature protocols* **7**, 562 (Mar, 2012).

45. S. Schwartz *et al*. *Genome research* **13**, 103 (Jan, 2003).

46. M. Roy, N. Kim, Y. Xing, C. Lee. *RNA* **14**, 2261 (Nov, 2008).

47. W. G. Fairbrother, R. F. Yeh, P. A. Sharp, C. B. Burge. *Science* **297**, 1007 (Aug 9, 2002).

48. Z. Wang *et al*. *Cell* **119**, 831 (Dec 17, 2004).

49. Y. Wang, M. Ma, X. Xiao, Z. Wang. *Nature structural & molecular biology* **19**, 1044 (Oct, 2012).

50. Y. Wang, Z. Wang. *Methods* **65**, 350 (Feb, 2014).

51. B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg. *Genome Biol* **10**, R25 (2009).

52. H. Li *et al*. *Bioinformatics* **25**, 2078 (Aug 15, 2009).

53. T. Derrien *et al*. *Genome research* **22**, 1775 (Sep, 2012).

54. G. R. Abecasis *et al*. *Nature* **491**, 56 (Nov 1, 2012).

55. S. Anders, W. Huber. *Genome biology* **11**, R106 (2010).

56. F. A. Kondrashov, E. V. Koonin. *Human molecular genetics* **10**, 2661 (Nov 1, 2001).

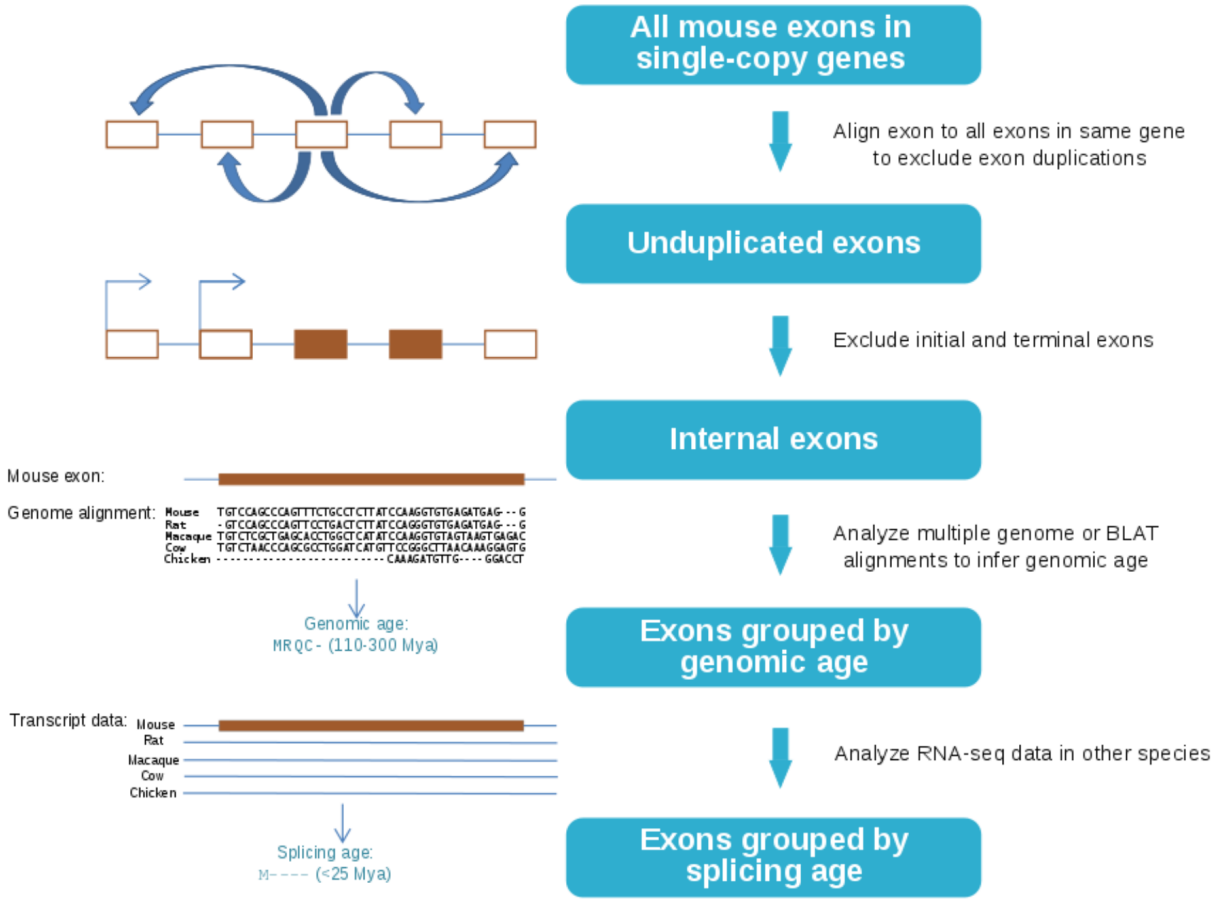
57. J. D. Bendtsen, H. Nielsen, G. von Heijne, S. Brunak. *Journal of molecular biology* **340**, 783 (Jul 16, 2004).

58. L. Kall, A. Krogh, E. L. Sonnhammer. *Journal of molecular biology* **338**, 1027 (May 14, 2004).

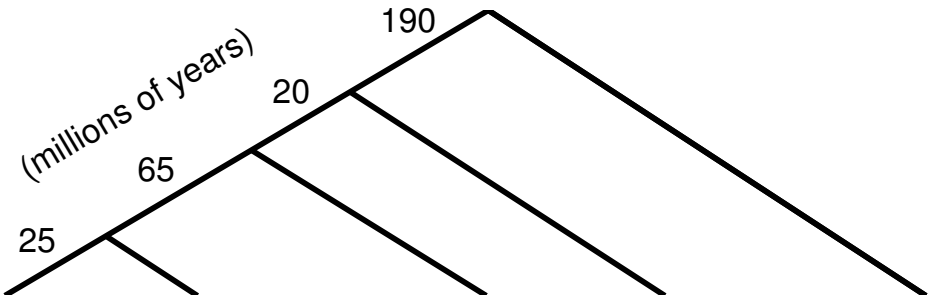
59. G. Yeo, C. B. Burge. *J Comput Biol* **11**, 377 (2004).

Figure 1

A



B



splicing age	abbr	mouse	rat	macaque	cow	chicken	no. of exons
	Present	M	R	Q	C	G	
Mya	absent	–	–	–	–	–	
0-25	M----						1089
25-90	MR---						361
90-110	MRQ--						792
110-300	MRQC-						4868
300+	MRQCG						47750

Figure 2

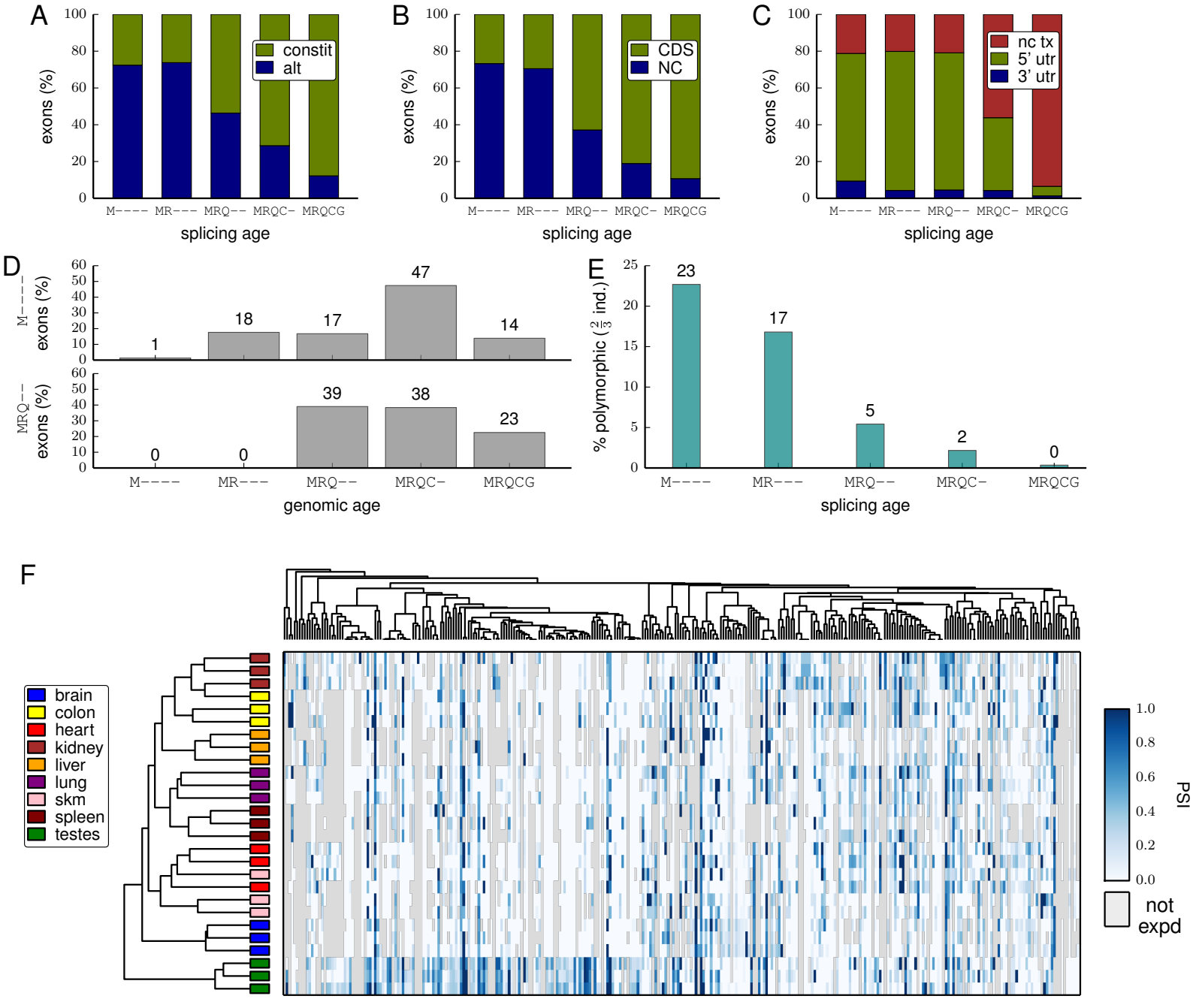


Figure 3

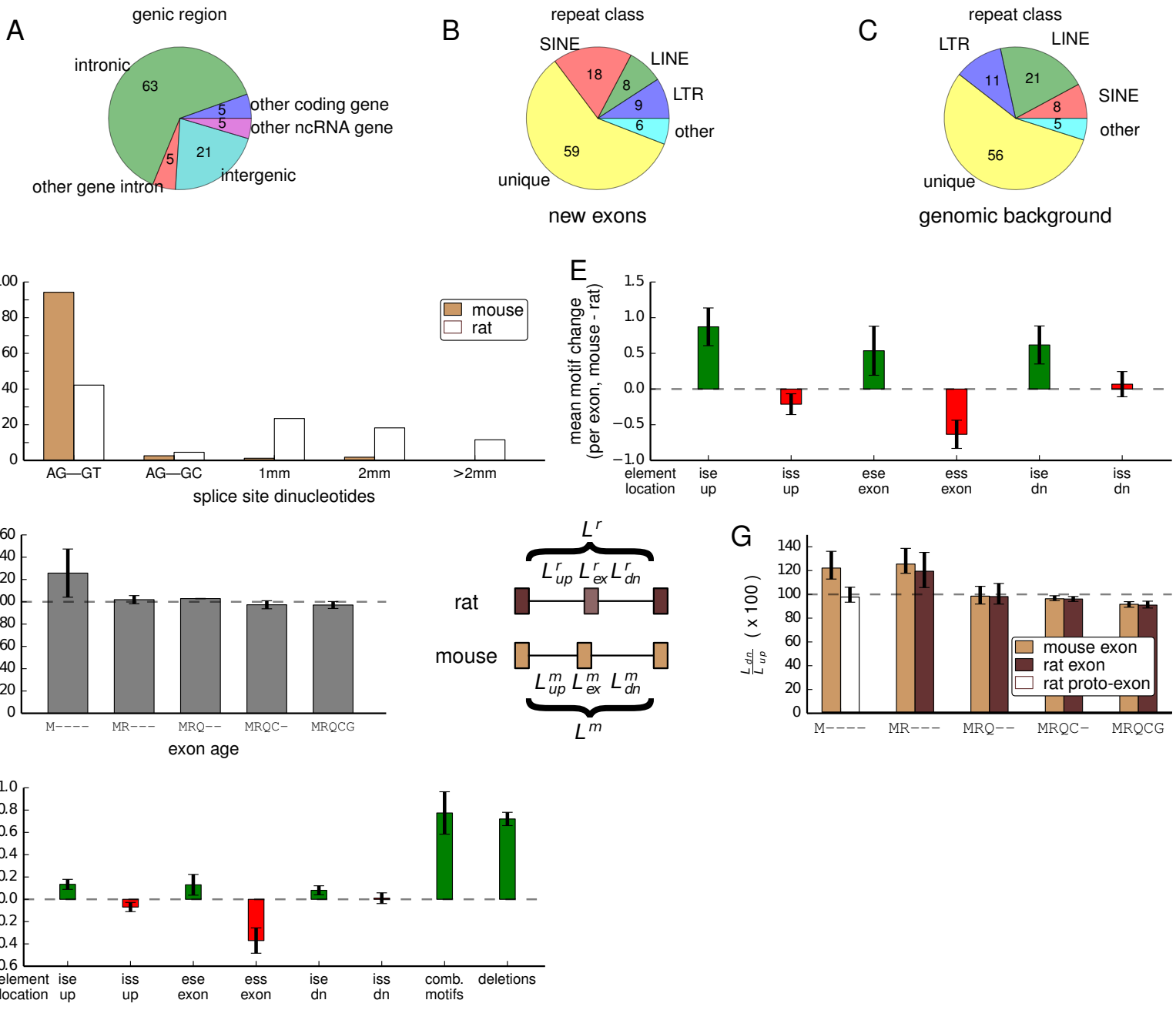


Figure 4

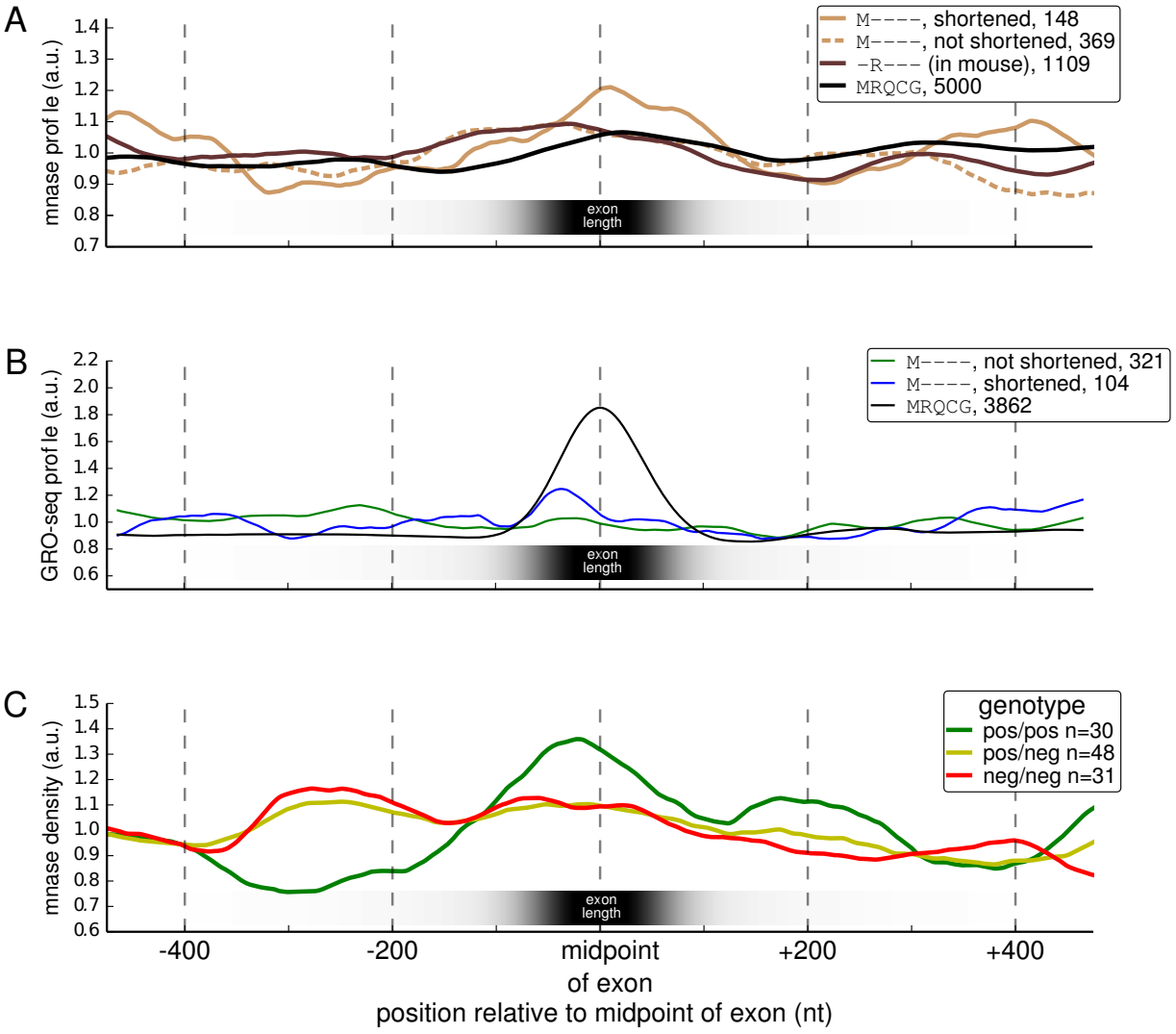


Figure 5