Supplementary data

Evolutionary impact of limited splicing fidelity in mammalian genes

Chaolin Zhang[1,2], Adrian R. Krainer[1], Michael Q. Zhang[1]

[1]Cold Spring Harbor Laboratory, NY 11724, USA
[2]Department of Biomedical Engineering, State University of New York, Stony Brook, NY 11794, USA
Corresponding author: Zhang, M.Q. (mzhang@cshl.edu).

Methods

Human and mouse transcript-confirmed exons

Human and mouse transcript (mRNA and/or EST)-confirmed exons were extracted from the Alternative Splicing Database (ASD) (Release 2, April 2005) [1]. The AltSplice database in ASD is a computationally derived collection of alternative splicing (AS) events of human and mouse based on alignment of EST and mRNA sequences to the corresponding genomic sequences with high quality and minimal redundancy. In ASD, all transcript–genome alignments with ambiguities were removed. A confirmed intron is defined by an alignment gap of genomic sequence flanked by two splice sites of known types. A confirmed exon is defined by an alignment match flanked by two confirmed introns; therefore, only internal exons are considered as being confirmed. Confirmed introns and exons that overlap with each other indicate AS events. In human, AltSplice has 16 293 genes, including 9945 (61%) with one or more alternative splicing events. In mouse, AltSplice has 16 352 genes, including 8211 (50%) alternatively spliced ones. The higher percentage of alternatively spliced genes in human is probably due to the higher EST coverage.

In this study, we considered only splicing events involving GT–AG intron boundaries. In total, 133 926 and 121 202 exons, plus 200 nucleotides of flanking intronic sequences, were extracted for human and mouse, respectively. Cassette exons are those included in some transcripts but skipped in others, without affecting the two neighboring exons (denoted as SCE, for simple cassette exons, in ASD). We extracted 10 196 and 5992 cassette exons for human and mouse, respectively. We also compiled a set of 30 892 and 37 313 exons that appear to be constitutively spliced in human and mouse, respectively. These exons were extracted from genes without AS events.

A summary of frame-preserving preference and human-mouse conservation (see below) is given in Table S1 and Figure S1. We also compared other features, such as intron phase bias (data not shown). All these general statistics are similar to and consistent to those reported previously (e.g. [2–4]).

Exon inclusion/skipping level

For each cassette exon, the number of supporting transcripts for the inclusion and the skipping isoforms was also extracted from ASD [1]. The number of supporting transcripts was used as an approximate measure of the abundance of the exon inclusion/skipping isoform, as done previously [5,6]. The ratio of the skipping to inclusion isoform or the ratio of the minor to major isoform (RMM) was used to estimate the relative abundance of the two isoforms.

Previous studies (e.g. [5,7]) have shown that newly evolved splicing isoforms usually have low abundance, whereas original ancestral isoforms remain dominant to minimize the deleterious effects of new isoforms to the organism. During evolution, the new minor isoform becomes more abundant if it has adaptive benefits and is positively selected. Therefore, RMM represents an approximate measure of the evolutionary age and fitness of an AS event.

Frame-preserving preference

An exon is defined as frame-preserving if its length is a multiple of three nucleotides, and as frame-shifting otherwise (e.g. [4]). The inclusion or skipping of a frame-preserving exon will not change the reading frame, thus affecting only the local protein sequence, unless the cassette exon has one or more premature termination codons, which is relatively infrequent. For a set of exons, the frame-preserving preference (FPP) is defined as the fraction of frame-preserving exons out of the total. The standard deviation of the FPP is estimated by a binomial distribution, $\mathrm{std(FPP)} = \mathrm{sqrt}[\mathrm{FPP}\times(1-\mathrm{FPP})/n]$. The statistical significance of the difference in the FPP between two exon groups is tested using a two-way contingency table, (group1 frame-preserving, group1 frame-shifting; group2 frame-preserving, group2 frame-shifting) by Fisher's exact test [4].

To generate the results given in Figure 1b,d, we used cassette exons with ≥10 supporting transcripts and ≥3 transcripts for the minor isoform. The filtering permits a more precise estimate of the relative abundance of the two isoforms. FPPs were calculated for cassette exons with different ranges of relative abundance of the two isoforms. In

particular, we regard an isoform as being rare if the relative abundance is less than 0.1. The thresholds of filtering and intervals were somewhat arbitrary and determined empirically, but the results seem to be robust with different thresholds.

Identification of orthologous exons for human–mouse comparison

Orthologous exon pairs were identified between human and mouse as previously described [4], with minor adaptations. In brief, 19 330 othologous gene pairs were downloaded using the Ensembl BioMart tool (formerly known as EnsMart) [8] (November, 2005). Then, each exon in the human gene was aligned to each exon in the orthologous mouse gene at both the nucleotide and protein levels using CLUSTALW [9]. For the protein-level alignment, nucleotide sequences were translated in all frames. Only those frames without a stop codon were retained for alignment. The reading frame that gave the best amino acid identity in each orthologous comparison was identified. Orthologous exon pairs were defined as those with reciprocal best alignment with nucleotide identity ≥60% and amino acid identity ≥50%. Generally, a real exon pair has a much higher conservation level than the thresholds. We identified mouse orthologous exons for human cassette exons, and vice-versa. We also identified ancestral cassette exons (orthologous cassette exons that can be included and skipped in both species).

To generate the results presented in Figure 2, we used the same filtering criteria as those described above (≥10 supporting transcripts and ≥3 transcripts for the minor isoform). In addition, cassette exons with skip/inc>1 were not analyzed here because rarely included cassette exons are more likely to have been recently exonized and are difficult to match between human and mouse [5]. In other words, rarely included cassette exons that are conserved between human and mouse may represent a very biased sample. Subsets of exons with different relative abundance of the two isoforms were defined similarly as in Figure 1 of the main text. For each subset, synonymous (K_s) and non-synonymous (K_a) mutation rates in the exons, and sequence conservation level in the flanking intronic regions were estimated as described below.

Calculation of synonymous and non-synonymous mutation rates

Following a previously described approach [4,10], the protein alignment generated to identify orthologous exons was used to realign the two nucleotide sequences of each orthologous exon pair. Gaps were removed. Synonymous and non-synonymous substitutions/sites were estimated by the Yang-Nielsen maximum-likelihood method, using the program yn00 in the PAML package [11]. For each subset of exons, the number of substitutions and sites were added up to calculate the overall K_s and K_a mutation rates, respectively, by the ratio of the two sums. The standard deviation of the ratio (K_a or K_s) was estimated by a binomial distribution, as for the estimation of standard deviation of FPP, as described above. The difference in K_s (K_a) for two exon groups was tested using the total number of substitutions/sites and Fisher's exact test, as described above and in previous studies [4].

Intronic sequence conservation

For each orthologous exon pair, we aligned both the upstream and downstream intronic flanking sequences (200 nucleotides in each region) using CLUSTALW. The 50 positions immediately upstream or downstream of the cassette exons were used to estimate the intronic conservation level. For each subset of exons, the average conservation level and standard error were calculated. We used robust estimates, that is, median and scaled MAD (median absolute deviation), which impose no assumption of normality. More precisely, the standard error is estimated by MAD/sqrt(n) , where n is the number of sequences. Note that in the software package R [12], MAD is scaled to be equivalent with the standard deviation for normal distributions.

Comparison of exon length for frame-shifting and frame-preserving exons

To generate the results in Table 1 in the main text, we used all constitutive and cassette exons, as well as ancestral cassette exons. The difference in median of exon size for frame-preserving exons and frame-shifting exons was tested by a Wilcoxon rank sum test. To generate the results presented in Figure S3, we filtered cassette exons by requiring ≥50 supporting transcripts. Exons were then broken down into three subsets, according to the relative abundance of the two isoforms. For each subset, we calculated the average and the standard error by robust estimates, that is, Median and MAD/sqrt(n).

We also examined three additional, readily available datasets of ancestral cassette exons generated by other groups [3,4,13]. Consistent differences between frame-preserving and frame-shifting exons were observed, as with the ASD data (not shown).

Statistical analyses

All statistical analyses and tests were performed in R [12].

Data availability

The original data can be downloaded from ASD. All datasets derived from ASD that were used in this study are freely available upon request.

URLs

ASD: http://www.ebi.ac.uk/asd/
Biomart: http://www.ensembl.org/Multi/martview/

Supplementary references

1. Thanaraj, T.A., *et al.* (2004) ASD: the alternative splicing database. *Nucleic Acids Res.* 32, D64-69
2. Resch, A., *et al.* (2004) Evidence for a subpopulation of conserved alternative splicing events under selection pressure for protein reading frame preservation. *Nucleic Acids Res.* 32, 1261-1269
3. Sorek, R., and Ast, G. (2003) Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res.* 13, 1631-1637
4. Xing, Y., and Lee, C. (2005) Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences. *Proc. Natl. Acad. Sci. USA* 102, 13526-13531
5. Modrek, B., and Lee, C.J. (2003) Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nature Genet.* 34, 177-180
6. Kan, Z., *et al.* (2002) Selecting for functional alternative splices in ESTs. *Genome Res.* 12, 1837-1845
7. Zhang, X.H.F., and Chasin, L.A. (2006) Comparison of multiple vertebrate genomes reveals the birth and evolution of human exons. *Proc. Natl. Acad. Sci. USA* 103, 13427-13432
8. Kasprzyk, A., *et al.* (2004) EnsMart: a generic system for fast and flexible access to biological data. *Genome Res.* 14, 160-169
9. Thompson, J., *et al.* (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673-4680
10. Nekrutenko, A., *et al.* (2003) ETOPE: evolutionary test of predicted exons. *Nucleic Acids Res.* 31, 3564-3567
11. Yang, Z., and Nielsen, R. (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* 17, 32-43
12. Ihaka, R., and Gentleman, R. (1996) R: a language for data analysis and graphics. *J. Comput. Graph. Statist.* 5, 299-314
13. Sugnet, C., *et al.* (2004) Transcriptome and genome conservation of alternative splicing events in humans and mice. *Pac. Symp. Biocomput.*, 66-77
14. Sugnet, C.W., *et al.* (2006) Unusual intron conservation near tissue-regulated exons found by splicing microarrays. *PLoS Computat. Biol.* 2, e4
15. Ule, J., *et al.* (2005) Nova regulates brain-specific splicing to shape the synapse. *Nature Genet.* 37, 844-852

Table S1. Frame-preserving preference and exonic/intronic conservation of orthologous exons in human and mouse

Exon type (hs17 vs mm5)	Exon number	Frame-preserving preference (%)	Exon nucleotide identity(%)	Exon amino acid identity(%)	Upstream identity(%)	Downstream identity(%)
exon vs exon	86063	39.6(0.2)	87.8(0.0)	93.8(0.0)	60.0(0.0)	56.0(0.1)
constitutive vs constitutive	9645	40(0.5)	87(0.1)	93(0.0)	60(0.1)	56(0.1)
exon vs cassette	4700	48(0.7)	88(0.1)	92(0.2)	64(0.2)	58(0.3)
cassette vs exon	2956	49(0.9)	89(0.1)	93(0.2)	66(0.3)	62(0.3)
cassette vs cassette	809	62(1.7)	93.0(0.3)	94.6(0.4)	78.0(0.6)	70.0(0.7)

The median and the standard error are shown. Cassette exons overall are more similar to constitutive exons than ancestral cassette exons.

Table S2. Frame-preserving preference of tissue-specific cassette exons from the literature

	Frame-preserving	All	Fraction(%)	Data source
Brain-specific	106	171	62	[14]
Muscle-specific	17	28	61	[14]
Validated Nova targets[a]	29	35	83	[15]

[a]A few exons not matched in ASD were excluded.

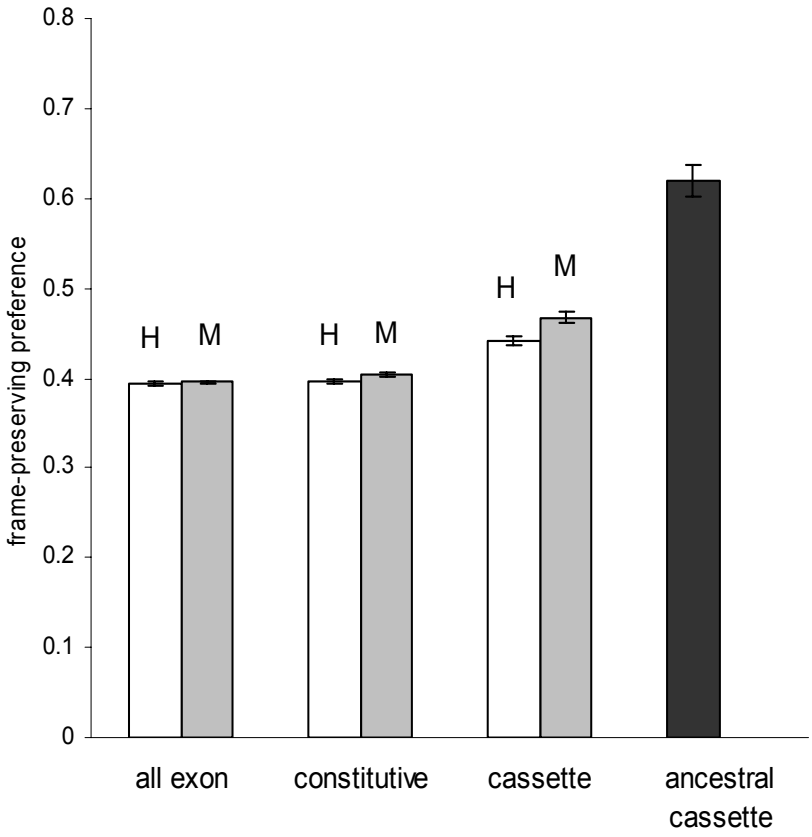

Figure S1. Frame-preserving preference (FPP) of all exons, constitutive exons, cassette exons and ancestral cassette exons in human (H) and mouse (M). The error bars show the standard deviation estimated from a binomial distribution. Cassette exons overall are more similar to constitutive exons than to ancestral cassette exons.

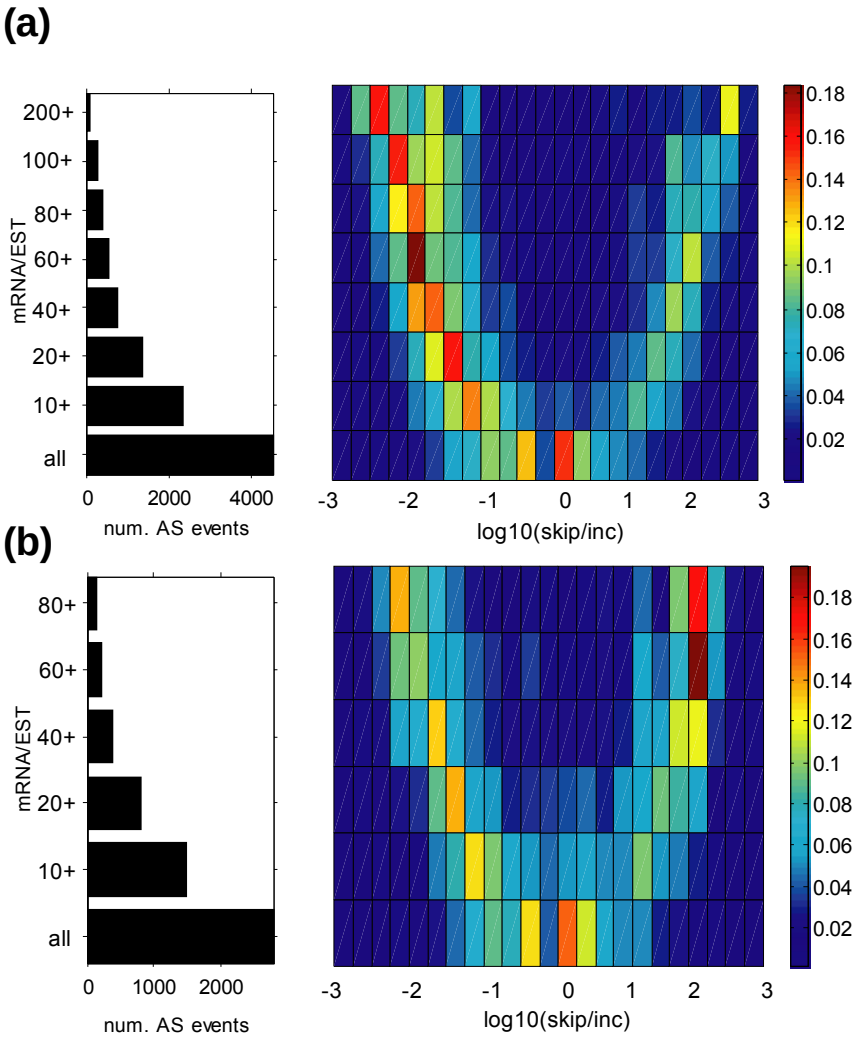

Figure S2. Distribution of frame-preserving cassette exons in terms of relative isoform abundance. **(a)** Human and **(b)** mouse data are shown. Refer to the legend of Figure 1 in the main text for details. The bimodal distribution (when a threshold of supporting transcripts was applied) is very similar to that observed for all cassette exons. This suggests that as the transcriptome is sampled more deeply, it is easier to find low-abundance splicing isoforms. This low abundance is largely independent of NMD. Also note that, in both Figure 1 and Figure S2, the peak of cassette exons with rare-skipping is almost twice that of cassette exons with rare-inclusion, which implies that leaky or aberrant exon skipping is more prevalent than inclusion. As an alternative interpretation, it is easier for random mutations to attenuate splicing signals than to create them in intronic sequences. These observations cannot be explained by NMD either.

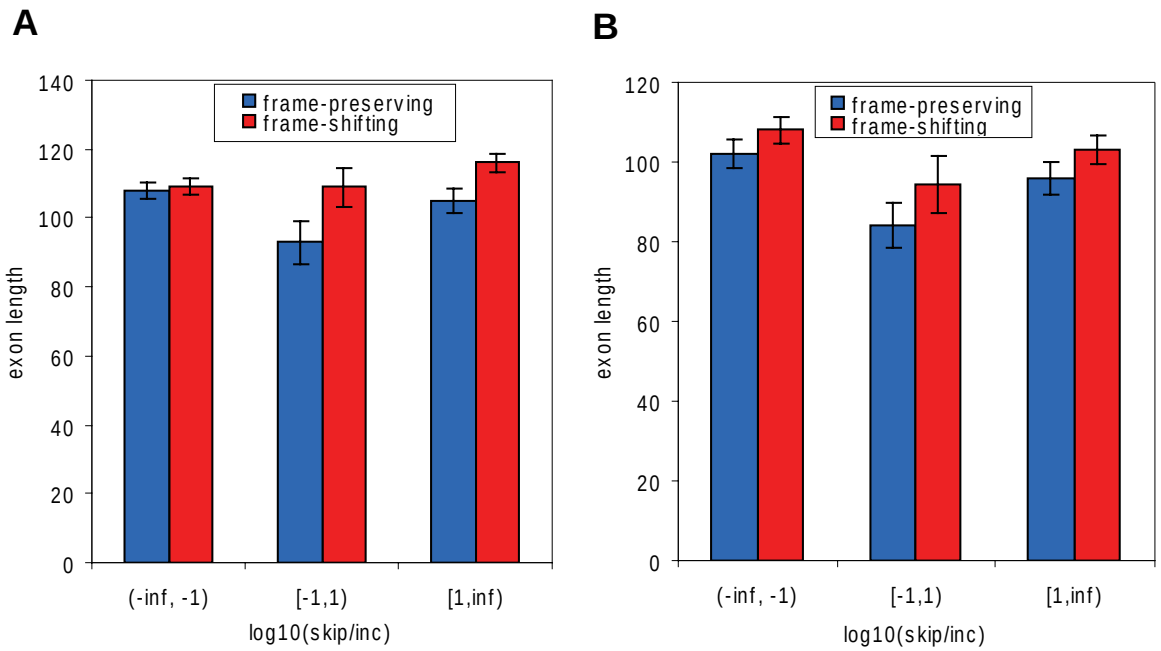

Figure S3. Comparison of exon size for cassette exons with different inclusion levels. Cassette exons with ≥50 supporting transcripts were divided into three bins according to skipping-to-inclusion ratio (<0.1, between 0.1 and 10 and ≥10). Frame-preserving and frame-shifting exons were compared separately (shown in blue and red respectively). The bars show median exon sizes. Error bars show standard errors. **(a)** Human and **(b)** mouse data are shown. Note that exons included at intermediate levels are shorter than those predominantly skipped or included. The difference seems to be larger for frame-preserving exons than frame-shifting ones. If suboptimal exon definition by the spliceosome is the primary reason for the shorter size of AS exons, rarely included exons should be even shorter, which contradicts our observation.