Evidence for Purifying Selection Against Synonymous Mutations in Mammalian Exonic Splicing Enhancers

Joanna L. Parmley, J. V. Chamary, and Laurence D. Hurst

Department of Biology and Biochemistry, University of Bath, Bath, United Kingdom

Silent sites in mammals have classically been assumed to be free from selective pressures. Consequently, the synonymous substitution rate (K_s) is often used as a proxy for the mutation rate. Although accumulating evidence demonstrates that the assumption is not valid, the mechanism by which selection acts remain unclear. Recent work has revealed that the presence of exonic splicing enhancers (ESEs) in coding sequence might influence synonymous evolution. ESEs are predominantly located near intron-exon junctions, which may explain the reduced single-nucleotide polymorphism (SNP) density in these regions. Here we show that synonymous sites in putative ESEs evolve more slowly than the remaining exonic sequence. Differential mutabilities of ESEs do not appear to explain this difference. We observe that substitution frequency at four-fold synonymous sites decreases as one approaches the ends of exons, consistent with the existing SNP data. This gradient is at least in part explained by ESEs being more abundant near junctions. Between-gene variation in K_s is hence partly explained by the proportion of the gene that acts as an ESE. Given the relative abundance of ESEs and the reduced rates of synonymous divergence within them, we estimate that constraints on synonymous evolution within ESEs causes the true mutation rate to be underestimated by not more than ~8%. We also find that K_s outside of ESEs is much lower in alternatively spliced exons than in constitutive exons, implying that other causes of selection on synonymous mutations exist. Additionally, selection on ESEs appears to affect nonsynonymous sites and may explain why amino acid usage near intron-exon junctions is nonrandom.

Introduction

At least in mammals, synonymous (silent) sites have long been assumed to be free from the pressures of natural selection (Eyre-Walker 1991; Sharp et al. 1995). If synonymous mutations are neutral (King and Jukes 1969; Kimura 1977) then the rate of synonymous substitution can be employed to measure the point mutation rate (e.g., Eyre-Walker and Keightley 1999; Keightley and Eyre-Walker 2000). Recently, however, there has been mounting evidence against this line of thought (Iida and Akashi 2000; Bustamante, Nielsen, and Hartl 2002; Hellmann et al. 2003; Keightley and Gaffney 2003; Urrutia and Hurst 2003; Chamary and Hurst 2004; Comeron 2004; Chamary and Hurst 2005*b*; Lavner and Kotlar 2005; Lu and Wu 2005). For example, constitutively and alternatively spliced exons differ in GC content at third (largely synonymous) sites (Iida and Akashi 2000).

What might be the mechanism for selection at so-called silent sites in exons? The classical model, that selection favors efficient translation (e.g., Ikemura 1985; Bulmer, Wolfe, and Sharp 1991; Akashi and Eyre-Walker 1998; Duret 2002), may not apply in mammals (Duret 2002; dos Reis, Savva, and Wernisch 2004) (but see Urrutia and Hurst 2003; Comeron 2004; Lavner and Kotlar 2005). Some evidence suggests that synonymous sites might be of importance in mRNA secondary structure and stability (Duan and Antezana 2003; Duan et al. 2003; Capon et al. 2004; Chamary and Hurst 2005*b*). Here we consider the possibility that purifying selection acts at synonymous sites to ensure efficient pre-mRNA splicing (Willie and Majewski 2004; Chamary and Hurst 2005*a*).

Exons are classically thought to be defined by sequence located within introns: the 5′ splice site, branch point, and 3′ splice site (Robberson, Cote, and Berget 1990). However, this tripartite signal (Fairbrother and Chasin 2000) is often necessary but not sufficient for intron excision. In human introns, these signals contain only half the required information for accurate splicing (Lim and Burge 2001). The polypyrimidine tract is important for regulating alternative splicing (Spellman et al. 2005). Exonic splicing enhancers (ESEs) are oligonucleotide sequences that are abundant in both constitutively and alternatively spliced exons (Tian and Kole 1995; Coulter, Landree, and Cooper 1997; Liu, Zhang, and Krainer 1998; Schaal and Maniatis 1999; Fairbrother et al. 2002). Most ESEs are thought to function though the binding of serine/arginine-rich proteins, which help instigate spliceosome assembly and localization (Wang et al. 2004). The Burge/Sharp group recently developed a computational method (Fairbrother et al. 2002; Fairbrother et al. 2004*b*) that identifies candidate hexameric sequences with ESE activity (for a brief summary of how these are defined, see *Materials and Methods*). The density of these ESE hexamers increases as one approaches intron-exon junctions (Supplementary Fig. 1, Supplementary Material online; Fairbrother et al. 2004*a*). ESE activity is optimal within ~70 nucleotides of splice sites, although the effect is dependent on the strength of the enhancer, with potent enhancers exerting an influence at double this distance (Graveley, Hertel, and Maniatis 1998).

Prior evidence suggests that codon choice is biased owing to the presence of ESEs and biased against intronic splicing enhancers (Willie and Majewski 2004; Chamary and Hurst 2005*a*), e.g., the codon GAA is common in ESEs and is increasingly preferred over its synonym GAG near intron-exon boundaries. It is unclear, however, whether this explains all the trends in codon bias as a function of distance from exonic ends (S. T. Eskesen, F. N. Eskesen, and Ruvinsky 2004; Chamary and Hurst 2005*a*). Consistent with a preference for ESEs at particular exonic locations, at least two genes exhibit a marked reduction in the synonymous rate of evolution in regions containing an

Key words: codon usage bias, mutation rate, purifying selection, splicing, synonymous sites.

E-mail: l.d.hurst@bath.ac.uk.

Mol. Biol. Evol. 23(2):301–309. 2006
doi:10.1093/molbev/msj035
Advance Access publication October 12, 2005

ESE (BRCA1: Hurst and Pal 2001; Liu et al. 2001; Orban and Olah 2001; CFTR: Pagani, Raponi, and Baralle 2005). More generally, it has been reported that single-nucleotide polymorphism (SNP) density decreases as one approaches the ends of exons (Majewski and Ott 2002) and that this can be explained by increasing ESE density (Fairbrother et al. 2004*a*; see also Carlini and Genut 2005). Although some ESEs appear to be conserved over the course of evolution (Yeo et al. 2004), it has not previously been demonstrated that the fixation of certain mutations have been opposed by natural selection because they occur within ESEs. Consequently, here we ask whether putative ESEs are associated with a lower rate of synonymous evolution and, if they are, what impact this might have had on estimates of the mutation rate (μ) derived from the rate of synonymous nucleotide substitution (K_s).

Materials and Methods

Alignments of Orthologous Mammalian Genes

We downloaded the 7,645 human-chimpanzee-mouse orthologues used by Clark et al. (2003) from http://www.sciencemag.org/cgi/content/full/302/5652/1960/DC1, using only those alignments where each of the three sequences contained a start codon and a terminal stop codon. Alignment trios containing sequences with lengths that were not multiples of three or contained internal stop codons were discarded. Sequences from the remaining trios were translated and aligned at the amino acid level using MUSCLE, http://www.drive5.com/muscle, after which the peptide sequences were used to reconstruct the nucleotide alignment.

Determining the Location of Intron-Exon Junctions

The GeneID (LocusLink) numbers in the annotation file were used to derive the human RefSeq identifiers at http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene. We then compared the human sequences in the alignments to those in the RefSeq files, retaining only those which were of same length and >99% identical. The RefSeq identifier was then used to identify genomic sequence (hence exon structure of the human coding sequence [CDS]) at Ensembl, http://www.ensembl.org/Homo_sapiens/exportview. We justify the use of the exon structure from human genes to define intron-exon junctions in other mammals because such structures are highly conserved (Roy, Fedorov, and Gilbert 2003). We ignored Ensembl genomic files where the CDS of the associated RefSeq was not the same length as that derived from the genomic annotation. For the 972 genes remaining, the intron-exon junctions in the alignments were reconstructed from the genomic sequence.

Obtaining Exonic Splicing Enhancers and Silencers

Candidate ESEs and exonic splicing silencer (ESS) sequences were identified by assaying whether oligonucleotide motifs exhibit splicing activity in vivo. The 238 human (Fairbrother et al. 2002) and 380 mouse (Yeo et al. 2004) ESE hexamers were determined using Relative Enhancer and Silencer Classification by Unanimous Enrichment (RESCUE), a computational approach followed by experimental validation. Briefly, the method identifies motifs that are: (1) significantly enriched in exons relative to introns and (2) significantly more frequent in exons with weak nonconsensus splice sites than in exons with strong consensus splice sites (Fairbrother et al. 2004*b*). Motifs that match these criteria are then grouped into clusters, after which representatives from each cluster are tested for ESE activity in vivo using a splicing reporter system. ESS motifs were identified by screening a library of random decamers for splicing activity in an in vivo reporter system (Wang et al. 2004). Human and mouse ESEs were downloaded from the RESCUE-ESE Web Server, http://genes.mit.edu/burgelab/rescue-ese, while human ESSs came from the supplementary data of Wang et al. (2004), http://www.download.cell.com/supplementarydata/cell/119/6/831/DC1index.htm.

Identification of ESEs and ESSs Within CDS

Defining sequence as ESE or ESS is nontrivial, so we took several different approaches. In principle, a putative ESE within an alignment could be defined as sequence present in one, either, or both species. Although one might imagine that the latter is the best definition because it is the most restrictive, human and mouse ESEs are very similar (e.g., 175/238 human hexamers are also found in mouse) and so this protocol may well end up isolating slow evolving sequence, rather than ESE. Consider the following hypothetical human-mouse alignment:

Human GAAGAATTT
Mouse CCCGAAGAA

If the hexamer GAAGAA is only identified in one species (by "human masking" or "mouse masking"), six of the nine sites are considered to be associated with the ESE (underlined) and 3 nucleotide substitutions have occurred. Under our most stringent definition of an ESE ("human + mouse masking"), only the three sites (GAA) that are within hexamers in both species are considered. Note that it is not the alignments but the sequences themselves that are scanned for the presence of putative ESEs/ESSs (gaps are collapsed and then later reinserted). Non-ESE regions were defined as the remaining unmasked sequence.

Evolutionary Rate Estimation

Nonsynonymous (K_a) and synonymous (K_s) substitution rates were estimated with the Li method (Li 1993) using the Kimura 2-parameter model. Whenever possible, to control for heterogeneity in mutation/substitution rates between genes (e.g., Lercher, Chamary, and Hurst 2004), differences in rates between putative ESE and non-ESE were performed by paired analyses using *t*-tests or one-sample Wilcoxon signed-rank tests. To minimize the effect of noise when sampling short sequence, we only considered pairs of sequences (ESE vs. non-ESE) where neither rate estimate was unusually high for the comparison (human-chimpanzee $K_a < 0.01$ and $K_s < 0.03$; human-mouse $K_a < 0.2$ and $K_s < 0.75$).

Frequency of Substitutions as a Function of Distance from Intron-Exon Junctions

Each exon was divided in two, with the first half being considered the 5′ end and the second the 3′ end. Under this

protocol no given site can be counted more than once. Running toward the interior of an exon, the distance from the intron-exon junction is the number of nucleotides (including gaps) from the junction pertinent to the half-exon. If a given site was fourfold degenerate in both species, we incremented the count of the number of sites at that distance and the number of substitutions where appropriate.

We also obtained ESE hexamers predicted to be predominantly active at the 5′ and 3′ ends of exons. The human ESE clusters were kindly provided by Will Fairbrother and the mouse 5′ and 3′ ESEs by Gene Yeo. Masking 5′ ends using ESEs with 5′ activity and 3′ ends with 3′ ESEs does not qualitatively affect our results (data not shown).

Comparison of Alternative and Constitutive Exons

We obtained the "training" set of exons (Yeo et al. 2005) from ACEScan, http://genes.mit.edu/acescan, where we have high confidence that exons have been conserved as being alternative or constitutive between human and mouse. The mouse and human exons were aligned at the nucleotide level using ClustalX. Exons in which the number of single-base indels in the alignment was not a multiple of three were eliminated (16 of the alternative exons and 24 of the constitutive ones). For the remainder we calculated the Tamura-Nei distance (Tamura and Nei 1993). For each of the three possible reading frames, we followed the method of Xing and Lee (2005) to ascribe the correct frame. After translating all exons in each of the three frames and eliminating those containing a stop codon, for each exon we calculated K_a for each of the remaining frames and employed the frame with the lowest K_a as the reading frame.

Results

Synonymous Evolution Is Slower in ESEs

If selection acts to preserve splicing activity (Yeo et al. 2004), the rate of synonymous substitution (K_s) should be lower in putative ESEs when compared with non-ESE sequence. To investigate this we scanned a data set of chimpanzee-human-mouse orthologues (Clark et al. 2003) for the presence of 238 putative human (Fairbrother et al. 2002) and 380 mouse (Yeo et al. 2004) ESE hexamers. As ESEs have yet to be identified in chimpanzees, here we report data for the human-mouse comparison, although the use of human hexamers as a "chimpanzee" set yields qualitatively the same results (Supplementary Table 1, Supplementary Material online; additional data available upon request). Similarly, as many ESEs are conserved (Yeo et al. 2004), one can also identify "mammalian" enhancers. This too gives similar results (Supplementary Table 2, Supplementary Material online).

As it is unclear on a priori grounds whether we should consider putative ESEs as being present in one or both species, we employ various masking protocols to identify sites that might be associated with putative ESEs. The first method identifies ESE sites as those that occur within human hexamers in human sequence (human masking). The second considers ESE sites to be those that are within mouse hexamers (mouse masking). Using more stringent definitions, we can also define ESE sites to be those present within hexamers in both sequences (human + mouse masking). This involves masking human hexamers in human sequence and mouse hexamers in mouse sequence, realigning the masked sequences (based on the original unmasked alignment), and then identifying those sites in the alignment where both sequences are putatively ESE.

In all masking permutations, we find that the synonymous substitution rate in putative ESEs is lower than that in non-ESEs (table 1; Supplementary Table 1, Supplementary Material online). The magnitude of the reduction in K_s is dependent on the masking protocol. The difference in K_s is relatively modest when masking hexamers in single species (~5%) but quite large in the more stringent double masking (~35%).

Reduced K_s Within ESEs Is Not Due to a Skewed CpG Distribution

Sites within CpG dinucleotides are known to be hypermutable (Bird 1980; Cooper and Krawczak 1989; Sved and Bird 1990), and ESEs are typically purine rich (Blencowe 2000) (in combined human/mouse hexamers A = 42.5%, G = 25.7%, C = 17.9%, and T = 13.9%). Consequently, it is possible that the reduction in K_s is an artefact owing to non-ESE sequence having a higher concentration of CpGs. However, after repeating the above analysis, this time omitting CG/GC pairs in either sequence, we again find that putative ESEs evolve more slowly than non-ESEs

Table 1
Differences in the Rate of Synonymous Evolution Between Putative ESE and Non-ESE Sequence in Human-Mouse Alignments

Masking Protocol[a]	Non-ESE[b]	ESE[b]	N[c]	P[d]
Human	0.4484 ± 0.0042	0.4117 ± 0.0054	812	8×10^{-11}
Human non-CpG	0.3378 ± 0.0041	0.3006 ± 0.0053	848	1×10^{-12}
Mouse	0.4440 ± 0.0040	0.4377 ± 0.0048	854	0.0538
Mouse non-CpG	0.3343 ± 0.0041	0.3184 ± 0.0048	889	8×10^{-5}
Human + mouse	0.4701 ± 0.0042	0.2896 ± 0.0053	815	3×10^{-103}
Human + mouse non-CpG	0.3488 ± 0.0041	0.2157 ± 0.0048	797	3×10^{-77}

[a] The sequences in which putative ESE motifs are masked. For human + mouse, these are the sites that are identified as being associated with ESEs in both species.

[b] The mean synonymous substitution rate (±SEM).

[c] The number of genes analyzed in pairwise comparisons.

[d] The significance of the difference between ESE and non-ESE (P values from paired t-tests).

(table 1). In fact, the previously marginally nonsignificant difference in the mouse masking now becomes significant. We conclude that the decreased K_s in ESEs cannot be explained by differential abundances of hypermutable CpGs.

Reduced K_s Within ESEs Is Not Due to a Skewed Nucleotide Distribution

The above test considers a class of well-known hypermutable sites. However, different nucleotides may themselves have different mutabilities (see e.g., Chamary and Hurst 2004). More generally, we can ask whether, controlling for skewed nucleotide contents, ESEs still have unusually low synonymous rates of evolution. Moreover, it is also possible that the reduction in K_s is a result of searching for relatively little sequence (particularly in human + mouse masking) which will artificially isolate slowly evolving sequences.

To examine these possibilities we performed a simulation. In each of 1,000 randomizations, we generated a set of simulated hexamers of the same average nucleotide composition as the real ESE hexamers. These simulated sets are then used to carry out human, mouse, and the human + mouse (stringent) maskings. For each gene, the difference between the real and the simulants was expressed as a *Z*-score, the number of standard deviations the observed K_s (from real ESEs) is away from the mean K_s of the simulated ESEs. Under a null hypothesis that the reduced K_s in ESE is due to the masking protocol and/or skewed nucleotide content in ESEs, the *Z*-score distribution should have an average that is not significantly different from zero. Alternatively, if putative ESEs evolve slowly, then their K_s should be significantly lower than the average of the simulants, i.e., a negative *Z*-score. Under the three protocols studied, we found that this was indeed the case (human masking median $Z = -0.293$, $P < 0.0001$; mouse median $Z = -0.214$, $P < 0.0001$; human + mouse median $Z = -0.17$, $P = 0.015$). We conclude that the low K_s in putative ESEs is not owing to skewed nucleotide content or any bias introduced by the masking process.

Substitution Frequency at Fourfold Degenerate Sites Declines Near Intron-Exon Junctions, Which Is Partially Explained by the Presence of ESEs

While the above results are consistent with a model in which ESE sequence is under selection to retain their function, there exists a further possibility. ESE density is known to be highest near intron-exon junctions. If, for some other reason, sequence in the near vicinity of such junctions are under stronger selection (or experience low mutation rates), then ESEs would have lower rates of evolution than either non-ESE sequence or our simulated ESEs, both of which may be relatively more common in exonic interiors. For example, exon-exon junctions tend to occur at or around the position of nucleosome formation (Kogan and Trifonov 2005). If nucleosomal or perinucleosomal sequence is more conserved than the average, then we may expect ESEs to be slow evolving, but only because they tend to be near nucleosomes. Note too that there may well be patterns of nucleotide usage across exons that are not explained by ESE presence/absence (S. T. Eskesen, F. N. Eskesen, and

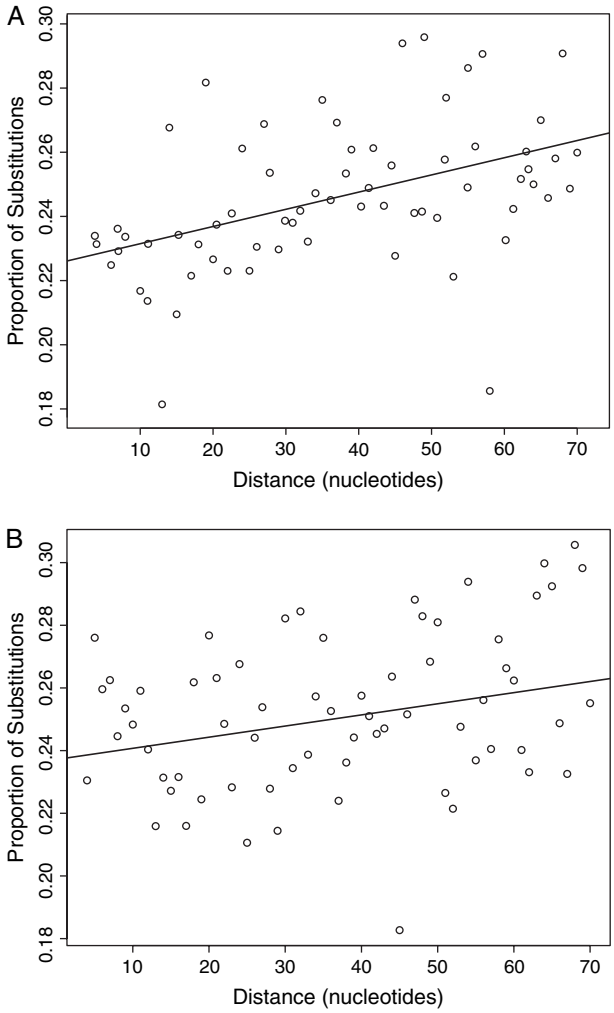

FIG. 1.—Frequency of substitutions at fourfold degenerate sites in human-mouse alignments as a function of distance from intron-exon junctions, at (*A*) the 5′ end of exons (slope = 0.2260; $R^2 = 0.1995$; $P = 9 \times 10^{-05}$) and (*B*) the 3′ end (slope = 0.2372; $R^2 = 0.0660$; $P = 0.0203$). The lines of best fit are derived by linear regression and weighted by the number of sites.

Ruvinsky 2004; Chamary and Hurst 2005*a*). We can therefore ask whether, given their location in proximity to the junctions, ESEs evolve slower than non-ESEs and whether this alone is adequate to explain the reduced SNP density near intron-exon junctions (Fairbrother et al. 2004*a*).

The frequency of substitutions at fourfold degenerate sites was assessed as a function of distance from both the 5′ and 3′ ends of exons, without masking ESE/non-ESE but ignoring CpGs. This analysis strongly suggests that synonymous mutations are increasingly opposed as one approaches the end of an exon (fig. 1). Studies looking at SNP density have suggested that such selection only extends about 30 nt into exons (Majewski and Ott 2002; Fairbrother et al. 2004*a*), but we observe an effect that is closer to the biased codon choice data (~100 nt, Willie and Majewski 2004; Chamary and Hurst 2005*a*).

Given the possible discrepancy in the scale of the effect, we then asked whether it is likely owing to a reduced rate of evolution in ESEs coupled with their greater

Table 2
ANCOVA Between Putative ESE and Non-ESE Sequences for the Substitution Frequency at Fourfold Synonymous Sites as a Function of Distance from Intron-Exon Junctions in Human-Mouse Alignments

		5′ End of Exons		3′ End of Exons	
Masking Protocol[a]	Parameter	Estimate[b]	P[c]	Estimate[b]	P[c]
Human non-CpG	Distance	0.0005 ± 0.0001	7×10^{-5}	0.0003 ± 0.0001	0.0137
	Level	0.0254 ± 0.0054	7×10^{-6}	0.0214 ± 0.0060	0.0005
Mouse non-CpG	Distance	0.0005 ± 0.0001	0.0002	0.0003 ± 0.0001	0.0123
	Level	0.0231 ± 0.0053	3×10^{-5}	0.0376 ± 0.0051	2×10^{-11}
Human + mouse non-CpG	Distance	0.0005 ± 0.0001	0.0001	0.0003 ± 0.0001	0.0244
	Level	0.0886 ± 0.0061	$<2 \times 10^{-16}$	0.1036 ± 0.0067	$<2 \times 10^{-16}$

[a] The sequences in which putative ESEs are masked.

[b] The ''Estimate'' for ''Distance'' is the slope of the regression line (±SEM) for the substitution frequency at fourfold sites in ESEs plotted against the distance from the intron-exon junction. There is no difference between the slopes derived from ESE and non-ESE sequences ($P > 0.05$). The estimate for ''Level'' is the difference between the slopes (±SEM) for ESE and non-ESE.

[c] For Distance, the P value indicates whether the common slope (ESE was used) is significant. For Level, the P value indicates whether there is a difference between ESEs and non-ESEs while controlling for the distance from the junction, i.e., to determine whether, at a given distance from the junction, the proportion of substitutions at fourfold sites differs between ESE and non-ESE.

proximity to intron-exon junctions or to some more general underlying cause. Under the first model, we expect both ESE rates of evolution and non-ESE rates of evolution to show no trend as a function of the distance from the junction, but with the ESE synonymous rates lower than those of the non-ESEs. In the second case, we might expect ESE and non-ESE to show the same trend of increasing synonymous divergence as a function of distance from the junction and no difference in the rates of evolution controlling for distance from junction.

These hypotheses were tested by analysis of covariance (ANCOVA) in which the distance from the junction was the covariate, and ESE and non-ESE sequence were the two factors/groups (NB there is no significant interaction term, so the assumptions of ANCOVA are upheld, $P > 0.05$). The difference in rates between the groups was always significant controlling for the distance from the junction ("Level" in table 2). This strongly suggests that ESEs are slow evolving even controlling for their differential abundance near junctions (table 2 and fig. 2). In all cases, there remains an effect whereby all sequences evolve marginally slower if closer to the junction ("Distance" in table 2). This suggests the presence of some weak force affecting substitution rates as a function of the distance from the junction independent of ESE presence or absence. As the effect is weak, however, we cannot rule out the possibility that it arises as a consequence of missing true ESEs in our classification.

The Effect of ESEs on Evolution at Nonsynonymous Sites

Here we have concentrated on how conservation of ESEs can influence synonymous mutations and codon usage. In principle, however, ESEs could also affect nonsynonymous mutations. This may well be the case as K_a is lower in putative ESEs (table 3). Moreover, as ESEs are generally purine rich (Blencowe 2000), it is interesting to ask whether amino acids specified by purine-rich codons are also more abundant near junctions. If so, we should expect the effect to be most strikingly seen for usage of lysine (AAA and AAG), A being the most common nucleotide in ESEs followed by G. This is indeed observed (fig. 3). However, while AG-rich codons tend to be employed near boundaries, at least for the 3′ end, the effect is more striking for AT-rich codons (Supplementary Fig. 2, Supplementary Material online). This suggests a pressure toward A and T rather than A and G and might hint at some other force (e.g., Chamary and Hurst 2005*a*). This is unlikely to be nucleosome associated as in mouse and human these are associated with G and C (Kogan and Trifonov 2005).

Discussion

Our analyses demonstrate that ESEs are under purifying selection. As the enhancer regions do not discriminate between synonymous and nonsynonymous sites, it is perhaps unsurprising that both classes of site are under constraint due to the presence of ESEs, most profoundly at the end of exons. This finding tempts several questions. First, assuming selection on splicing enhancers is the only mode of selection on synonymous mutations, to what extent might one underestimate the mutation rate when extrapolating from synonymous divergence? Second, is it likely that this is the only mechanism of selection on synonymous mutations? To address the latter issue, we examine alternative exons, these being known to have lower synonymous substitution rates than constitutive ones from the same gene (Iida and Akashi 2000; Xing and Lee 2005). Finally, we ask about implications of the finding that codon usage and rates of evolution are unusual in the vicinity of intron-exon junctions.

Selection on ESEs Has a Modest Effect on Underestimation of the Mutation Rate

Under the supposition that synonymous sites evolve neutrally, their rate of evolution has been used as a measure of the mutation rate (see e.g., Eyre-Walker and Keightley 1999; Keightley and Eyre-Walker 2000). Assuming selection on ESEs to be the only form of selection at synonymous sites, how much might this method underestimate the real mutation rate? To address this issue we need to know what proportion of the sequence is functional splicing

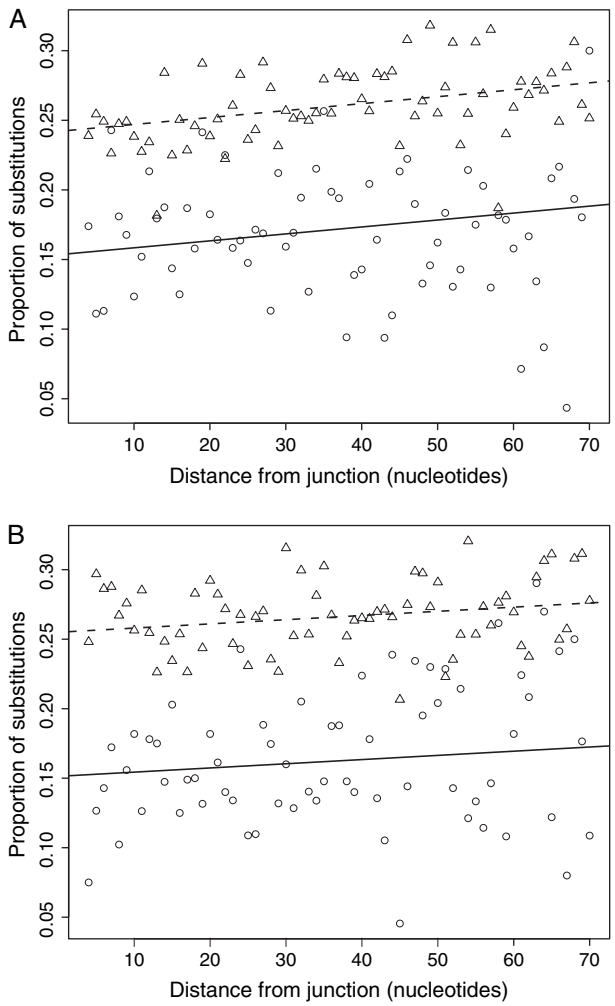

FIG. 2.—Frequency of substitutions at fourfold degenerate sites in human-mouse alignments as a function of distance from intron-exon junctions in ESE (circles and solid lines) and non-ESE (triangles and dashed lines) sequences, at (*A*) the 5′ end of exons and (*B*) the 3′ end. The weak trends are shown for sites within ESEs at the 5′ (*A*, slope = 0.1658; $R^2 = 0$; $P = 0.6831$) and 3′ end (*B*, slope = 0.1369; $R^2 = 0.0773$; $P = 0.0130$), and non-ESE sequence at the 5′ (*A*, slope = 0.2396; $R^2 = 0.1625$; $P = 0.0004$) and 3′ end (*B*, slope = 0.2575; $R^2 = 0.0143$; $P = 0.1664$). The lines of best fit are derived by linear regression and weighted by the number of nucleotide sites. The ESE masking is by the human + mouse protocol.

enhancer and what, on the average, is the reduction in the rate of evolution within ESEs.

We have employed three different methods to define putative ESEs. Enhancers identified within a single species (mouse or human) show a modest 1%–11% reduction in their rate of evolution (depending on whether we ignore CpGs, table 4). Sequence defined as ESE in both mouse and human have a more striking ~38% reduction in their rate compared with non-ESE regions (table 4). However, the more stringent definition defines less of the sequence as being in enhancer. When we factor in the proportion of sequence that is putatively ESE, the three methods all suggest that the net reduction in K_s, owing to the presence of ESEs, is modest. It may be as low as 2% and unlikely to be much more than 8% (table 4). This suggests that correction for the presence of ESEs will not have a major effect on estimates of the mutation rate, not least because the margin of error associated with estimates of the number of generations between any two mammalian taxa is vastly more error prone and alterations here will have a much more profound effect.

Selection on ESEs Is Only One Form of Selection on Synonymous Mutations

Conservation of ESEs is unlikely to be the only form of selection at synonymous sites. In terms of splicing, biased codon usage may also reflect an avoidance of certain sequences that might be associated with cryptic splice sites (S. T. Eskesen, F. N. Eskesen, and Ruvinsky 2004; but see Chamary and Hurst 2005*a*). Additionally, we have not considered the contribution of ESS sequence, although we find that masking the 133 decamers that have been systematically identified in humans (Wang et al. 2004) does not alter our conclusions (Supplementary Table 3, Supplementary Material online). Importantly, the strongest signal for selection that has been seen so far is a high stability of cytosine at third sites (Chamary and Hurst 2004). This is not obviously explained by a role in the splicing process (Chamary and Hurst 2005*a*) because ESEs are AG rich and C poor. The cause of the C preference remains unclear, but a role in mRNA stability is supported by some data (Chamary and Hurst 2005*b*). There may also be other factors that constrain synonymous evolution, such as the need to bind antisense transcripts. Therefore, we cannot conclude that selection on silent sites has not lead to a significant underestimate of the mutation rate.

Selection on ESEs Does Not, for the Most Part, Explain Low Synonymous Rates in Alternative Transcripts

Another way to address whether other forms of selection act at synonymous mutations is to ask whether it is a greater abundance of and/or stronger selection on ESEs that might explain why alternatively spliced exons have unusually low rates of synonymous evolution (Iida and Akashi 2000; Xing and Lee 2005). To address this, we examined a carefully curated set of conserved alternative and constitutive exons (Yeo et al. 2005). We see that mean substitution rates in alternative exons (Tamura-Nei distance = 0.069 ± 0.004; $N = 225$) is lower ($P < 0.0001$ by Mann-Whitney *U*-test) than that in constitutive exons (0.123 ± 0.001, $N = 5{,}045$). This is owing to a much lower rate of evolution at both synonymous sites and, in contrast to prior analyses (Iida and Akashi 2000; Xing and Lee 2005), nonsynonymous sites, although the effect is more dramatic for the former. Examining exons with a minimum of 30 codons, for example, we find that the mean K_s is lower in alternative exons (0.115 ± 0.02; $N = 51$) compared to constitutive exons (0.311 ± 0.009; $P < 0.0001$ by Mann-Whitney *U*-test) while K_a in alternative exons (0.058 ± 0.008) is lower than that in constitutives (0.103 ± 0.002; $P = 0.0003$ by Mann-Whitney *U*-test). The reduced K_s is not due to alternative exons possessing more ESEs, as we find that there is no consistent difference in the proportion of putative enhancer sequence between the two classes

Table 3
Differences in the Rate of Amino Acid Evolution Between Putative ESE and Non-ESE Sequence in Human-Mouse Alignments

Masking Protocol[a]	Non-ESE[b]	ESE[b]	N[c]	P[d]
Human	0.0526 ± 0.0015	0.0473 ± 0.0015	862	5×10^{-9}
Human non-CpG	0.0394 ± 0.0013	0.0404 ± 0.0015	874	0.5685
Mouse	0.0524 ± 0.0015	0.0503 ± 0.0015	890	0.0147
Mouse non-CpG	0.0396 ± 0.0013	0.0402 ± 0.0014	908	0.4211
Human + mouse	0.0545 ± 0.0016	0.0343 ± 0.0013	838	2×10^{-68}
Human + mouse non-CpG	0.0418 ± 0.0015	0.0298 ± 0.0013	815	1×10^{-34}

[a] The sequences in which putative ESEs are masked.
[b] The mean nonsynonymous substitution rate (±SEM).
[c] The number of genes analyzed in pairwise comparisons.
[d] The significance of the difference between ESE and non-ESE (P values from paired t-tests).

of exons (Supplementary Table 4, Supplementary Material online). Is then the reduced rate of evolution especially noticeable in ESEs, and is it seen in non-ESE parts of alternative transcripts?

As regards the second issue, the rate of synonymous evolution in non-ESE sequence of alternative exons is over 50% lower than that for non-ESE parts of constitutive exons (Supplementary Table 5, Supplementary Material online). This strongly suggests that selection on ESEs cannot fully explain why alternative exons are slow evolving. Although the data are noisy, our best evidence suggests that ESEs in alternative transcripts have K_s values that are slightly lower than that of non-ESE in the same alternative exon (Supplementary Table 6, Supplementary Material online). The causes of the unusually low rates of evolution in conserved alternative exons deserve further scrutiny.

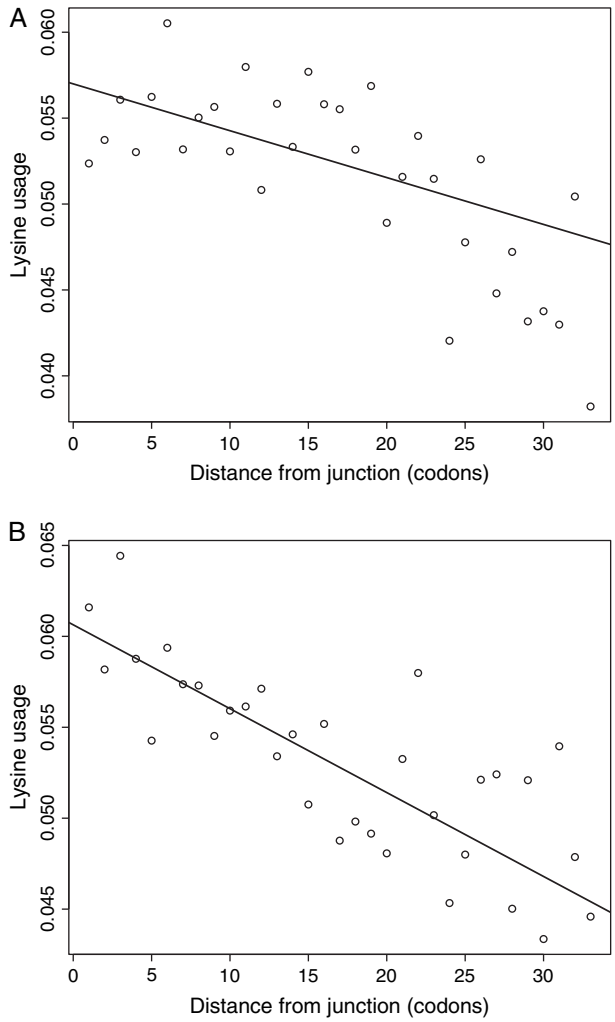

FIG. 3.—Lysine residue usage as a function of distance from intron-exon junctions, at (*A*) the 5′ end of exons ($R^2 = 0.2936$; $P = 0.0007$) and (*B*) the 3′ end ($R^2 = 0.6364$; $P = 2 \times 10^{-8}$). The lines of best fit are derived by linear regression and weighted by the number of codons.

Implications of Stronger Selection Near Intron-Exon Junctions

One consequence of all the evidence for skewed nucleotide composition (Louie, Ott, and Majewski 2003; S. T. Eskesen, F. N. Eskesen, and Ruvinsky 2004) and biased codon usage (Willie and Majewski 2004; Chamary and Hurst 2005*a*) near intron-exon boundaries is that it adds layers of complexity to the interpretation of prior results. First, the conventional application of $K_a/K_s > 1$ as an indication of positive selection should be treated with caution as this may be owing to reduced K_s rather than elevated K_a (Pond and Muse 2005), as previously described in at least two genes (BRCA1 [Hurst and Pal 2001; Liu et al. 2001; Orban and Olah 2001] and CFTR [Pagani, Raponi, and Baralle 2005]). Further, several recent reports find evidence for systematic codon bias that is not explained by background nucleotide content (Urrutia and Hurst 2003; Comeron 2004; Lavner and Kotlar 2005). For example, highly expressed genes exhibit the greatest bias (Urrutia

Table 4
The Contribution of Purifying Selection at Synonymous Sites in Putative ESEs to Underestimates of the Mutation Rate (μ) in Mammals

Masking Protocol[a]	K_s Reduction[b] (%)	ESE Coverage[c] (%)	μ Underestimation (%)
Human	8.19	30.42	2.49
Human non-CpG	11.03	30.42	3.36
Mouse	1.41	40.30	0.57
Mouse non-CpG	4.74	40.30	1.91
Human + mouse	38.39	21.77	8.36
Human + mouse non-CpG	38.15	21.77	8.31

[a] The sequences in which putative ESEs are masked.
[b] The difference in the synonymous substitution rate between ESE and non-ESE.
[c] The proportion of sequence covered by ESE sites.

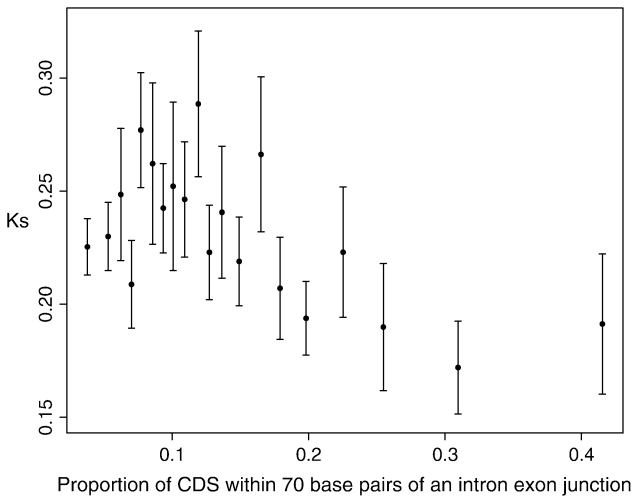

FIG. 4.—The synonymous substitution rate (K_s) as a function of the proportion of CDS within 70 bp of an intron-exon junction. The data is split into 20 bins with equal numbers of genes ($N = 48$) in each bin.

and Hurst 2003). As intron density also varies with expression parameters (Comeron 2004), these results may be artefacts of biased codon usage in the proximity of intron-exon junctions. Indeed, when we consider the relationship between K_s and the proportion of the CDS within 70 nt of the junction, we observe a significant negative correlation (fig. 4; Spearman rank correlation $\rho = -0.15$, $P < 0.0001$). To factor out any such effects, we recommend that one should exclude those regions of exons within about 70 nt on either side of junctions.

The potential impact of ESE presence on nonsynonymous substitution rates has numerous corollaries. First, this makes it difficult to ask whether a certain protein domain is under purifying selection. A low K_a may be evidence for this, but it could also be explained by selection on an ESE rather than the protein. To examine in detail such claims, one should also ask whether the DNA specifying the domain is near an intron-exon junction and matches known ESEs. The skewed amino acid usage near intron-exon boundaries has two possible interpretations. First, that at the time of insertion, a viable intron can only be tolerated if there are already ESEs present in the near vicinity. Second, that after insertion, the process of splicing is subject to selection, with choice of amino acids around junctions being determined in part by the efficiency of splicing of flanking introns. These are not mutually incompatible. To establish whether the first is true, one would need to identify new introns within the mammal lineage. These are remarkably rare (Roy, Fedorov, and Gilbert 2003) (see also Sry in marsupials, O'Neill et al. 1998). Conversely, if loss of an intron is not followed by adjustment of amino acid content, this would suggest that amino acid content was dictated by the protein level considerations rather than splicing regulation.

Supplementary Material

Supplementary Figs. 1 and 2 and Supplementary Tables 1–6 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank the anonymous referees for suggestions. J.L.P. and J.V.C. are funded by the United Kingdom Biotechnology and Biological Sciences Research Council.

Literature Cited

Akashi, H., and A. Eyre-Walker. 1998. Translational selection and molecular evolution. Curr. Opin. Genet. Dev. **8**:688–693.

Bird, A. P. 1980. DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res. **8**:1499–1504.

Blencowe, B. J. 2000. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. Trends Biochem. Sci. **25**:106–110.

Bulmer, M., K. H. Wolfe, and P. M. Sharp. 1991. Synonymous nucleotide substitution rates in mammalian genes: implications for the molecular clock and the relationship of mammalian orders. Proc. Natl. Acad. Sci. USA **88**:5974–5978.

Bustamante, C. D., R. Nielsen, and D. L. Hartl. 2002. A maximum likelihood method for analyzing pseudogene evolution: implications for silent site evolution in humans and rodents. Mol. Biol. Evol. **19**:110–117.

Capon, F., M. H. Allen, M. Ameen, A. D. Burden, D. Tillman, J. N. Barker, and R. C. Trembath. 2004. A synonymous SNP of the corneodesmosin gene leads to increased mRNA stability and demonstrates association with psoriasis across diverse ethnic groups. Hum. Mol. Genet. **13**:2361–2368.

Carlini, D. B., and J. E. Genut. 2005. Synonymous SNPs provide evidence for selective constraint on human exonic splicing enhancers. J. Mol. Evol. (in press).

Chamary, J. V., and L. D. Hurst. 2004. Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents: evidence for selectively driven codon usage. Mol. Biol. Evol. **21**:1014–1023.

———. 2005*a*. Biased codon usage near intron-exon junctions: selection on splicing enhancers, splice-site recognition or something else? Trends Genet. **21**:256–259.

———. 2005*b*. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. Genome Biol. **6**:R75.

Clark, A. G., S. Glanowski, R. Nielsen et al. (17 co-authors). 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science **302**:1960–1963.

Comeron, J. M. 2004. Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence. Genetics **167**:1293–1304.

Cooper, D. N., and M. Krawczak. 1989. Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. Hum. Genet. **83**:181–188.

Coulter, L. R., M. A. Landree, and T. A. Cooper. 1997. Identification of a new class of exonic splicing enhancers by in vivo selection. Mol. Cell. Biol. **17**:2143–2150.

dos Reis, M., R. Savva, and L. Wernisch. 2004. Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res. **32**:5036–5044.

Duan, J., and M. A. Antezana. 2003. Mammalian mutation pressure, synonymous codon choice, and mRNA degradation. J. Mol. Evol. **57**:694–701.

Duan, J., M. S. Wainwright, J. M. Comeron, N. Saitou, A. R. Sanders, J. Gelernter, and P. V. Gejman. 2003. Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. Hum. Mol. Genet. **12**:205–216.

Duret, L. 2002. Evolution of synonymous codon usage in metazoans. Curr. Opin. Genet. Dev. **12**:640–649.

Eskesen, S. T., F. N. Eskesen, and A. Ruvinsky. 2004. Natural selection affects frequencies of AG and GT dinucleotides at the 5′ and 3′ ends of exons. Genetics **167**:543–550.
Eyre-Walker, A. 1991. An analysis of codon usage in mammals: selection or mutation bias? J. Mol. Evol. **33**:442–449.
Eyre-Walker, A., and P. D. Keightley. 1999. High genomic deleterious mutation rates in hominids. Nature **397**:344–347.
Fairbrother, W. G., and L. A. Chasin. 2000. Human genomic sequences that inhibit splicing. Mol. Cell. Biol. **20**:6816–6825.
Fairbrother, W. G., D. Holste, C. B. Burge, and P. A. Sharp. 2004*a*. Single nucleotide polymorphism-based validation of exonic splicing enhancers. PLoS Biol. **2**:e268.
Fairbrother, W. G., R. F. Yeh, P. A. Sharp, and C. B. Burge. 2002. Predictive identification of exonic splicing enhancers in human genes. Science **297**:1007–1013.
Fairbrother, W. G., G. W. Yeo, R. Yeh, P. Goldstein, M. Mawson, P. A. Sharp, and C. B. Burge. 2004*b*. RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. Nucleic Acids Res. **32**:W187–W190.
Graveley, B. R., K. J. Hertel, and T. Maniatis. 1998. A systematic analysis of the factors that determine the strength of pre-mRNA splicing enhancers. EMBO J. **17**:6747–6756.
Hellmann, I., S. Zollner, W. Enard, I. Ebersberger, B. Nickel, and S. Paabo. 2003. Selection on human genes as revealed by comparisons to chimpanzee cDNA. Genome Res. **13**: 831–837.
Hurst, L. D., and C. Pal. 2001. Evidence for purifying selection acting on silent sites in BRCA1. Trends Genet. **17**:62–65.
Iida, K., and H. Akashi. 2000. A test of translational selection at 'silent' sites in the human genome: base composition comparisons in alternatively spliced genes. Gene **261**:93–105.
Ikemura, T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. **2**:13–34.
Keightley, P. D., and A. Eyre-Walker. 2000. Deleterious mutations and the evolution of sex. Science **290**:331–333.
Keightley, P. D., and D. J. Gaffney. 2003. Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents. Proc. Natl. Acad. Sci. USA **100**:13402–13406.
Kimura, M. 1977. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature **267**:275–276.
King, J. L., and T. H. Jukes. 1969. Non-Darwinian evolution. Science **164**:788–798.
Kogan, S., and E. N. Trifonov. 2005. Gene splice sites correlate with nucleosome positions. Gene **352**:57–62.
Lavner, Y., and D. Kotlar. 2005. Codon bias as a factor in regulating expression via translation rate in the human genome. Gene **345**:127–138.
Lercher, M. J., J. V. Chamary, and L. D. Hurst. 2004. Genomic regionality in rates of evolution is not explained by clustering of genes of comparable expression profile. Genome Res. **14**:1002–1013.
Li, W. H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. **36**:96–99.
Lim, L. P., and C. B. Burge. 2001. A computational analysis of sequence features involved in recognition of short introns. Proc. Natl. Acad. Sci. USA **98**:11193–11198.
Liu, H. X., L. Cartegni, M. Q. Zhang, and A. R. Krainer. 2001. A mechanism for exon skipping caused by nonsense or missense mutations in BRCA1 and other genes. Nat. Genet. **27**:55–58.
Liu, H. X., M. Zhang, and A. R. Krainer. 1998. Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. Genes Dev. **12**:1998–2012.
Louie, E., J. Ott, and J. Majewski. 2003. Nucleotide frequency variation across human genes. Genome Res. **13**:2594–2601.
Lu, J., and C. I. Wu. 2005. Weak selection revealed by the whole-genome comparison of the X chromosome and autosomes of human and chimpanzee. Proc. Natl. Acad. Sci. USA **102**: 4063–4067.
Majewski, J., and J. Ott. 2002. Distribution and characterization of regulatory elements in the human genome. Genome Res. **12**: 1827–1836.
O'Neill, R. J., F. E. Brennan, M. L. Delbridge, R. H. Crozier, and J. A. Graves. 1998. De novo insertion of an intron into the mammalian sex determining gene, SRY. Proc. Natl. Acad. Sci. USA **95**:1653–1657.
Orban, T. I., and E. Olah. 2001. Purifying selection on silent sites—a constraint from splicing regulation? Trends Genet. **17**:252–253.
Pagani, F., M. Raponi, and F. E. Baralle. 2005. Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. Proc. Natl. Acad. Sci. USA **102**:6368–6372.
Pond, S. K., and S. V. Muse. 2005. Site-to-site variation of synonymous substitution rates. Mol. Biol. Evol. (in press).
Robberson, B. L., G. J. Cote, and S. M. Berget. 1990. Exon definition may facilitate splice site selection in RNAs with multiple exons. Mol. Cell. Biol. **10**:84–94.
Roy, S. W., A. Fedorov, and W. Gilbert. 2003. Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. Proc. Natl. Acad. Sci. USA **100**:7158–7162.
Schaal, T. D., and T. Maniatis. 1999. Selection and characterization of pre-mRNA splicing enhancers: identification of novel SR protein-specific enhancer sequences. Mol. Cell. Biol. **19**:1705–1719.
Sharp, P. M., M. Averof, A. T. Lloyd, G. Matassi, and J. F. Peden. 1995. DNA-sequence evolution: the sounds of silence. Phil. Trans. R. Soc. Lond. B **349**:241–247.
Spellman, R., A. Rideau, A. Matlin, C. Gooding, F. Robinson, N. McGlincy, S. N. Grellscheid, J. Southby, M. Wollerton, and C. W. Smith. 2005. Regulation of alternative splicing by PTB and associated factors. Biochem. Soc. Trans. **33**:457–460.
Sved, J., and A. Bird. 1990. The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. Proc. Natl. Acad. Sci. USA **87**:4692–4696.
Tamura, K., and M. Nei. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial-DNA in humans and chimpanzees. Mol. Biol. Evol. **10**:512–526.
Tian, H., and R. Kole. 1995. Selection of novel exon recognition elements from a pool of random sequences. Mol. Cell. Biol. **15**:6291–6298.
Urrutia, A. O., and L. D. Hurst. 2003. The signature of selection mediated by expression on human genes. Genome Res. **13**:2260–2264.
Wang, Z., M. E. Rolish, G. Yeo, V. Tung, M. Mawson, and C. B. Burge. 2004. Systematic identification and analysis of exonic splicing silencers. Cell **119**:831–845.
Willie, E., and J. Majewski. 2004. Evidence for codon bias selection at the pre-mRNA level in eukaryotes. Trends Genet. **20**:534–538.
Xing, Y., and C. Lee. 2005. Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences. Proc. Natl. Acad. Sci. USA. **102:** 13526–13531.
Yeo, G., S. Hoon, B. Venkatesh, and C. B. Burge. 2004. Variation in sequence and organization of splicing regulatory elements in vertebrate genes. Proc. Natl. Acad. Sci. USA **101**:15700–15705.
Yeo, G. W., E. Van Nostrand, D. Holste, T. Poggio, and C. B. Burge. 2005. Identification and analysis of alternative splicing events conserved in human and mouse. Proc. Natl. Acad. Sci. USA **102**:2850–2855.

Kenneth Wolfe, Associate Editor

Accepted October 10, 2005