# Positive selection acting on splicing motifs reflects compensatory evolution

Shengdong Ke,[1] Xiang H.-F. Zhang,[1,2] and Lawrence A. Chasin[3]

*Department of Biological Sciences Columbia University New York, New York 10027, USA*

**We have used comparative genomics to characterize the evolutionary behavior of predicted splicing regulatory motifs. Using base substitution rates in intronic regions as a calibrator for neutral change, we found a strong avoidance of synonymous substitutions that disrupt predicted exonic splicing enhancers or create predicted exonic splicing silencers. These results attest to the functionality of the hexameric motif set used and suggest that they are subject to purifying selection. We also found that synonymous substitutions in constitutive exons tend to create exonic splicing enhancers and to disrupt exonic splicing silencers, implying positive selection for these splicing promoting events. We present evidence that this positive selection is the result of splicing-positive events compensating for splicing-negative events as well as for mutations that weaken splice-site sequences. Such compensatory events include nonsynonymous mutations, synonymous mutations, and mutations at splice sites. Compensation was also seen from the fact that orthologous exons tend to maintain the same number of predicted splicing motifs. Our data fit a splicing compensation model of exon evolution, in which selection for splicing-positive mutations takes place to counter the effect of an ongoing splicing-negative mutational process, with the exon as a whole being conserved as a unit of splicing. In the course of this analysis, we observed that synonymous positions in general are conserved relative to intronic sequences, suggesting that messenger RNA molecules are rich in sequence information for functions beyond protein coding and splicing.**

**[Supplemental material is available online at www.genome.org.]**

In addition to sequences that specify their cognate polypeptides, the open reading frames of eukaryotic messenger RNAs (mRNAs) are likely to contain additional information governing the functioning of these molecules; these functions include exon splicing and mRNA transport, localization, stability, and translation. If so, the evolution of mRNA sequence would be constrained by purifying selection against the loss of this information. This selection should be evident by examining the rate and quality of base substitutions that do not alter protein coding (synonymous substitutions) but may alter one of the abovementioned functions.

Mutation rates in open reading frames have been extensively characterized by two parameters: $K_a$, the rate of mutations that result in amino acid substitutions; and $K_s$, the rate at synonymous sites. The latter has most often been assumed to reflect neutral change and has been used to normalize the mutation rates in a given gene, with the $K_a/K_s$ ratio serving as an inverse measure of protein sequence conservation (Hurst 2002). More recently, the neutrality of $K_s$ has been challenged, with accumulating evidence for selection acting at synonymous sites (for reviews, see Chamary et al. 2006; Xing and Lee 2006). Some of this evidence has been based on comparisons between mutation rates in predicted exonic splicing enhancers (ESEs) versus control sequences, concluding that the former tend to be conserved. Thus, mutations that disrupt ESEs have been selected against among human single nucleotide polymorphisms (SNPs), especially those near exon extremities (Fairbrother et al. 2004; Carlini and Genut 2006), and the $K_s$ within human ESEs defined by human–mouse comparisons was seen to be lower than that in non-ESE sequences (Parmley et al. 2006). Another approach has been to compare $K_s$ in alternatively versus constitutively spliced exons. $K_s$ has been found to be lower in the former (Xing and Lee 2005; Chen et al. 2006), which is in keeping with the idea that alternatively spliced mammalian exons have a higher density of splicing regulatory signals.

A problem in interpreting $K_s$ values as experimental variables is deciding what to use as a background, i.e., sequences that more closely reflect neutral change. This problem is exacerbated when examining sets of specific motifs, since mutation rates can vary widely depending on the exact sequences neighboring the sites in question (Arndt et al. 2003; Zhang and Gerstein 2003; Hwang and Green 2004; Siepel and Haussler 2004), making a comparison to a control context all the more important. We have approached this problem by (1) measuring changes in deep intronic sequences ($K_i$) to estimate the neutral mutation rate; (2) comparing the rates of ESE and exonic splicing suppressor (ESS) creation and disruption to rates in control motif sets with similar sequence characteristics; and (3) testing the idea that when mutations predicted to compromise splicing occur, they tend to be accompanied by the accumulation of compensatory mutations that are predicted to encourage splicing. Comparing the human genome to chimpanzee and macaque, we observed the following: (1) $K_s$ is significantly slower than $K_i$, supporting the idea of RNA function beyond protein coding. (2) ESEs have been disrupted at a lower rate than control sequences and have been created at a higher rate. In the first such analysis of predicted ESS motifs, we found the converse to be true, in agreement with their proposed role in silencing splicing. (3) When splicing efficiency in one species is predicted to be decreased due to the loss of an ESE, the gain of an ESS, or the weakening of a splice-site consensus sequence, there is a trend toward compensation in the form of ESE gains, ESS losses, or strengthening of a splice site. We conclude that the exon as a whole is conserved as a unit of splicing and

**[1] These authors contributed equally to this work.**
**[2] Present address: Memorial Sloan–Kettering Cancer Center, 1275 York Ave., New York, NY 10065, USA.**
**[3] Corresponding author.**
**E-mail lac2@columbia.edu; fax (212) 531-0425.**
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.070268.107.

that other units of function may also be contributing to exon homeostasis.

## Results and Discussion

### Overall $K_s$ vs. $K_i$

To ask whether synonymous sites are subject to selection, we measured overall mutation rates by comparing the sequences of human open reading frames to those of chimpanzee and macaque. We made the conservative choice of using only constitutively expressed human exons because alternatively spliced exons exhibit a lower $K_s$ than constitutive exons (Chen et al. 2006) and are thought to contain more splicing regulatory sequences (Itoh et al. 2004). The human sequences were collected from mRNA and EST databases (see Methods) and aligned to the other primate genomes. As an estimate of neutral mutation rates, we enumerated changes in intron sequences. We ignored sequences located within 100 nt of exons, as these regions are known to harbor splicing regulatory signals (Louie et al. 2003; Zhang et al. 2003, 2005c) and tend to be conserved (Sorek and Ast 2003; Sugnet et al. 2004; Xing and Lee 2005; Yeo et al. 2005). We also excluded highly repeated sequences, as they may be subject to distinctive evolutionary pressures, and we purposely ignored changes at CpG sites to avoid domination of the data by that highly mutable dinucleotide.

As can be seen in the first two lines of Table 1, $K_a$ amounts to 20%–30% of $K_s$, reflecting the greater selective pressure to maintain protein sequence compared to RNA sequence. The more interesting comparison here is between $K_s$ and $K_i$; the former is ~75% of the latter in both primate comparisons. This substantial difference is the expected result if selection is operating at synonymous sites, implying a function for these nucleotides in processes such as splicing, stability, transport, or translation efficiency. Although the data are consistent with this idea, it is also possible that the constraint on synonymous site changes is being caused by their close linkage to amino acid coding nucleotides. In the absence of recombination among these closely linked (intragenic) nucleotides, selection for or against a change in an amino acid coding nucleotide could reduce the pool of variants at synonymous sites (Charlesworth et al. 1993; Kim 2004; Comeron and Guthrie 2005). This mechanism may underlie the strong positive correlation seen between $K_s$ and $K_a/K_s$ comparing humans and rodents (Wyckoff et al. 2005). On the other hand, Birky and Walsh (1988) have argued that there are no net effects on neutral mutations from this phenomenon, and no correlation between $K_s$ and $K_a/K_s$ was seen when comparing less divergent species (human–chimp; Wyckoff et al. 2005) such as those used here. We conclude that the low $K_s/K_i$ values seen here probably reflect purifying selection at synonymous sites.

### $K_s$/$K_i$ in splicing motifs

Since the results of an overall $K_s/K_i$ ratio of <1 could originate from diverse selective pressures, we turned to a more pointed hypothesis to examine the evolutionary behavior of splicing motifs: If predicted exonic splicing enhancer (ESE) sequences are generally functional, they would tend to be conserved and thus influence the value of $K_s$ within these motifs. Analogously, the creation of exonic splicing silencer motifs would tend to be avoided. To test this idea we compiled a list of such motifs by combining our previously described (Zhang and Chasin 2004) set of predicted putative ESEs (PESEs) with RESCUE-ESEs predicted by Fairbrother et al. (2002). Similarly, we combined our previously described (Zhang and Chasin 2004) set of predicted putative ESSs (PESSs) with the experimentally determined FAS-hex3 set described by Wang et al. (2004). To effect this merger, we converted our PESE and PESS octamer motifs to their most commonly embedded hexamers (see Methods). The two sources of each motif set were represented approximately equally. We term the classified changes limited to these motifs as $K_{a\ \mathrm{disrupt}}$ for mutations that alter an amino acid and disrupt a motif, $K_{s\ \mathrm{create}}$ for mutations that do not alter an amino acid and create a motif, and so on. We compared human and macaque sequences, using dog as an outgroup to distinguish motif disruption from motif creation. The results apply to all changes in both primates. It was important to use merged lists in order to avoid misclassification of mutational events. For example, a change that disrupts one of our original PESEs but simultaneously creates a RESCUE-ESE would be misclassified as an ESE disruption if the RESCUE-ESE motifs were not taken into account.

The rate of change at the synonymous sites within predicted ESE and ESS motifs will depend on both selection and mutability. The mutability of these motif sets will depend critically on their sequence, since the mutation rate at any particular nucleotide position is highly influenced by surrounding nucleotides. We diminished this effect by measuring the rate of change of each set of motifs when located in intronic regions (as described above), and we used these values for normalization. While we think that normalization of changes to a given motif at a control location is essential, it does raise the possibility that our results will be exaggerated if selective forces acting in exons are mirrored by forces acting in the opposite direction in introns (e.g., in favor of ESE disruption to prevent splicing of pseudo-exons). Results for disruption of ESEs and ESSs in constitutive exons compared in human and macaque are shown in the top part of the first data column of Supplemental Table S1. The $K_{a\ \mathrm{disrupt}}/K_{s\ \mathrm{disrupt}}$ odds ratio for ESEs is 0.192 (95% confidence interval 0.188–0.195) in the human–macaque comparison, the low value reflecting selection for amino acid sequences. However, this ratio is 6% lower than the overall $K_a/K_s$ odds ratio of 0.204 (95% confidence interval 0.202–0.205) seen in Table 1, suggesting selection against the loss of splicing information embedded within protein coding information.

The more incisive comparison is between $K_s$ and $K_i$. $K_{s\ \mathrm{disrupt}}$ for ESEs is significantly lower than $K_{i\ \mathrm{disrupt}}$, yielding a $K_{s\ \mathrm{disrupt}}/K_{i\ \mathrm{disrupt}}$ odds ratio of 0.702. This value is considerably less than 1, as expected if predicted ESE motifs are indeed subject to purifying selection, and is ~10% lower than the overall $K_s/K_i$. To ask

**Table 1.** Comparing mutation rates at synonymous, nonsynonymous, and intronic sites

| | $K_s$ | $K_a$ | $K_i$ | $K_a$:$K_s$ OR[a] (95% CI) | $K_a$:$K_i$ OR (95% CI) | $K_s$:$K_i$ OR (95% CI) |
|---|---|---|---|---|---|---|
| Human–chimp, constitutive exons | 0.0073 | 0.0023 | 0.0099 | 0.313 (0.308–0.319) | 0.231 (0.227–0.235) | 0.729 (0.720–0.738) |
| Human–macaque, constitutive exons | 0.0418 | 0.0088 | 0.0531 | 0.204 (0.202–0.205) | 0.158 (0.157–0.160) | 0.777 (0.773–0.782) |

[a] Odds ratio.

whether ESEs are more conserved than other motifs at synonymous sites, we created a control set of an equal number of non-ESE motifs for comparison. Devising such a control set is nontrivial, because the ESE set is composed of families of highly related overlapping sequences. For example, if one ESE differs from seven other ESEs by only a single nucleotide, then the probability of a random mutation disrupting this ESE is (18 − 7)/18. On the contrary, if it is similar to no others, then this probability is 18/18. Thus, a compilation of random non-ESE hexamers (mostly unrelated) could not be compared on the same basis. We decided to use the reverse complements of ESEs as a non-ESE control set. This set of motifs maintains the similarities among the hexamers, forming a control set that has the same "coherence" as the experimental set, and has the same CG dinucleotide content as well. There are some possible dangers in such a control set: The purine to pyrimidines ratio is systematically reversed; and the reverse complements of ESEs and ESSs may be selected against and for, respectively, if secondary structures play a role in the availability of these motifs. While such an effect would tend to exaggerate the differences between the experimental and control set, it would not distort the result, as the exaggeration itself rests on the functionality of the ESE and ESS sets.

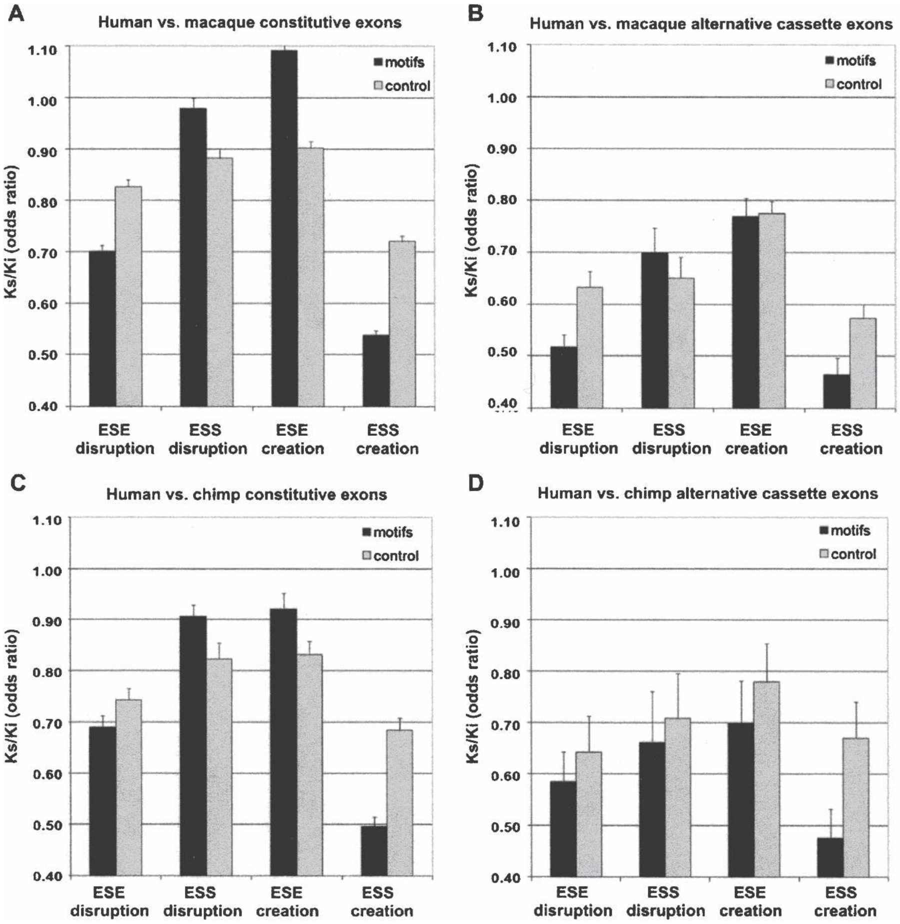

**Figure 1.** Rates of ESE and ESS creation and disruption. Rates have been normalized for change at intronic sites, expressed as $K_s/K_i$. Black bars represent ESE and ESS motifs (but purged of reverse complements); the gray bars represent two separate control motif sets, being largely the reverse complements of the ESE and ESS sets (see text). Error bars show 95% confidence intervals. (*A*,*B*) Human–macaque comparisons; (*C*,*D*) human–chimpanzee comparisons. (*A*,*C*) Constitutive exons; (*B*,*D*) alternatively spliced exons.

The mutation rates for ESEs, ESSs, and their control sets in constitutively and alternatively spliced exons are shown for the human–macaque comparison in Supplemental Table S1; the key $K_s/K_i$ values (presented as odds ratios) are summarized graphically in Figure 1, A and B. Two principal conclusions can be drawn from Figure 1A, the comparison of human and macaque constitutive exons. The first is that both ESE disruptions and ESS creations occur at much lower rates at synonymous sites in exons than in introns (black bars at left and right, ratios <1), an indication of purifying selection acting on both types of these splicing-negative events. The accompanying gray bars show that the control motif sets are also under purifying selection by this criterion ($K_s/K_i < 1$), in accordance with the overall $K_s/K_i$ results shown in Table 1. However, the $K_s/K_i$ values for the splicing motifs are much lower than those of the controls, suggesting that selection for splicing function is an important evolutionary pressure.

Purifying selection against ESE disruption was first shown by Fairbrother et al. (2004) using human SNPs as a measure of change. More recently, conservation of ESEs at synonymous sites has been demonstrated by mammalian comparative genomics (Yeo et al. 2004; Xing and Lee 2005; Parmley et al. 2006; Stadler et al. 2006). Our data confirm these observations using a new method for taking motif-specific variation in mutation rate into account ($K_s/K_i$) and extend them to suggest that the non-creation of ESSs should also be considered a purifying selection.

A second conclusion drawn from Figure 1A is that splicing motifs appear to be subject to positive selection: Creation of ESEs and disruption of ESSs occur at much higher rates than the analogous changes in the control motif sets (the two middle bar sets in Figure 1A; see also Supplemental Table S1, row 4, columns 3–6). Resch et al. (2007) found evidence for positive selection at synonymous sites in numerous human genes. The results presented here suggest that much of this positive selection can be ascribed to splicing motifs. All four differences between splicing motifs and the controls shown in Figure 1A were also seen when we analyzed SNPs (with chimpanzee as the out-group), but because the number of changes were much smaller, the statistical significances were marginal (data not shown). An adaptation of the McDonald–Kreitman test (McDonald and Kreitman 1991) was applied to this data but failed to indicate positive selection here (data not shown). This result could have been predicted, since this test is based on the assumption that positively selected changes show a higher frequency between species (fixed mutations) than within a species (polymorphisms). It is clear that ESE selection can in fact be seen in human SNP data (Fairbrother et al.

2004; Carlini and Genut 2006). We discuss the significance of this positive selection further below.

We also examined exons that are alternatively spliced in humans (cassette-type, the most frequent form of alternative splicing), measuring changes in both species and making the conservative assumption that most would be alternatively spliced in both primates. By definition, this analysis is limited to conserved alternatively spliced exons. These exons exhibit a somewhat different behavior than their constitutive counterparts. Evidence for purifying selection is seen here also, but to a considerably greater extent than in constitutive exons (Fig. 1B, left and right black bars); the mutation rate for ESE disruption and ESS creation at synonymous sites being reduced to only about half the intronic rate. Thus, maintenance of splicing motifs appears even more critical for alternative exons than for constitutive exons. Chen et al. (2006), using human–rat and rat–mouse comparisons, also concluded that non–protein-coding information is more highly conserved in alternative compared to constitutive exons, based on their observation of lower $K_s/K_a$ values in the former. Our results provide evidence that splicing motifs are contributing to the greater conservation of alternative vs. constitutive exon sequence. Interestingly, the $K_s/K_i$ values for non-ESE and non-ESS control motifs are also lower compared to constitutive exons, implying that functions other than splicing are also more critical for alternative exons. Remarkably, the alternative exons exhibit less positive selection than their constitutive counterparts for changes that would increase splicing (ESE creation and ESS disruption, middle black bars); indeed, they do not differ from the control sets in this respect. These results are in keeping with the need for alternatively spliced exons to maintain a balance among splicing elements so as to allow a combination of exon skipping and exon inclusion and with the observation that conserved alternative exons tend to conserve their quantitative regulation (Modrek and Lee 2003; Kan et al. 2005; Rukov et al. 2007). Thus, existing motifs are strongly conserved, but increases in splicing efficiency are not tolerated. An analogous inference was reached by Garg and Green (2007) for splice-site sequences: Weak splice-site scores were conserved in alternative exons so as to maintain inefficient splicing. The evidence for purifying selection seen for ESS motifs in alternative exons (a $K_s/K_i$ value of 0.70, equal to that showing ESE conservation in constitutive exons) suggests that silencing (as well as weak enhancement) plays an important role in the mechanism of alternative splicing.

We repeated this analysis comparing human to chimpanzee, with the macaque as an out-group to distinguish motif disruption from motif creation. These comparisons led to the same conclusions for constitutively spliced exons (Fig. 1C), but few conclusions can be drawn for alternatively spliced exons here due to the paucity of data (Fig. 1D).

The inference that ESEs and ESSs are subject to both negative and positive selection is based on significant differences between these splicing motifs and the control sets. If the control sets themselves contain many motifs subject to selection, then the differences exhibited by the ESEs and ESSs would be underestimating the selective pressure. To see whether motifs other than splicing elements are conserved, we compared $K_s/K_i$ values for all hexamers. Hexamers were divided into ESEs, ESSs, and ~3500 non-ESE/ESSs; a histogram of the distribution of $K_s/K_i$ values for these three sets is shown in Figure 2. Indeed, the non-ESE/ESS hexamers comprise a distribution not unlike the ESEs, with many members exhibiting even higher conservation (lower $K_s/K_i$ values) than the ESEs. This result should not be surprising, as these motifs may comprise sequence elements that function in the transport of mRNA from nucleus, mRNA localization in the cytoplasm, mRNA stability, and mRNA translation efficiency, as well as yet-unrecognized splicing signals. It might be interesting to identify families of hexamers with low $K_s/K_i$ values and test consensus sequences for function in the processes listed above. A more quantitative indication of ESE and ESS function than comparison to arbitrary control motif sets is probably the direct comparison to rates of change in intronic regions, i.e., the $K_s/K_i$ value itself. These ratios are strikingly low: ≤0.70 for ESE disruptions and ≤0.54 for ESS creations among constitutive exons, with even lower values in alternative exons. The $K_i$ for splicing motifs may not be entirely neutral if splicing motifs are evolving in introns in a manner opposite that in exons so as to keep pseudo-exons from being spliced. If this were the case, the $K_s/K_i$ can be viewed as the sum of the selective forces acting to enhance splicing in exons and silence it in introns. Even if intronic selection is making a contribution here, it seems safe to conclude that splicing signals play a major role in shaping the sequence of exons.

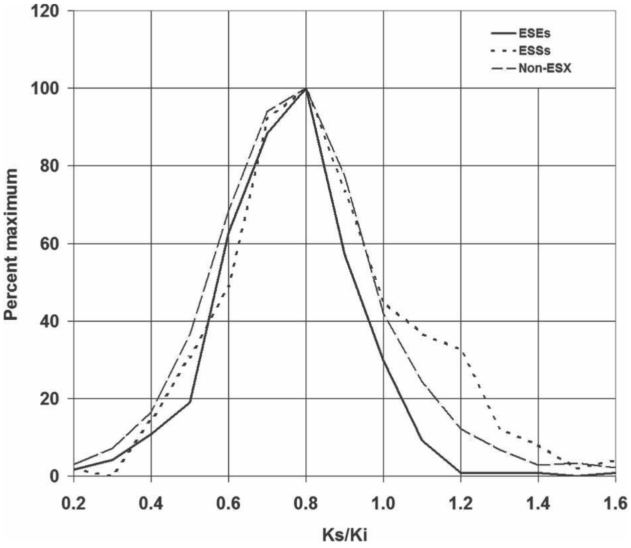

**Figure 2.** Distribution of $K_s/K_i$ ratios for hexamers. Histograms of $K_s/K_i$ values for ESEs, ESSs, and remaining hexamers are shown. Data include the full 469 ESE and 246 ESS sets (not purged of reverse complements) and 3448 non-ESXs and exclude changes at CpG sites and 33 mostly CpG-rich hexamers that yielded no data.

The evidence for positive selection found here for splicing motifs in constitutive exons is at first glance surprising. Splicing is a very old process, which one would think would be well honed in contemporary organisms. Nevertheless, the data indicate that, even in the 6 million years since the human–chimpanzee divergence, selection is continuing for ESE creation and ESS disruption (Fig. 1C). We can think of several explanations for this result: (1) There are a significant number of orthologous exon pairs in which the human exon is constitutively spliced but the macaque or chimpanzee exon is alternatively spliced, and constitutive splicing is favored. Although the human constitutive exons were chosen as being constitutive on the basis of their lack of EST isoforms, we have no such assurance for the two other primates, for which sufficient EST data are lacking. Thus, ESE creation and ESS disruption in human compared to macaque could sometimes reflect a switch from alternative splicing to constitutive splicing. Indeed, many exons do show species specific alternative splicing in human–mouse comparisons (Pan et al. 2005). However, among conserved exons, the majority show conservation of alternative splicing patterns (Modrek and Lee 2003; Kan et al. 2005; Pan et al. 2005; Rukov et al. 2007). Furthermore, about half of the data in Figure 1 is based on splicing-positive changes in the nonhuman primate, where this explanation could not hold as the human exons are all constitutive. (2) The exigencies of protein evolution result in ESE disruptions and ESS creations in some exons due to changes at nonsynonymous sites. New ESEs must be created or ESSs disrupted in order to restore efficient splicing to such exons, and most of these new mutations survive at synonymous sites. (3) A more general extension of (2) is that, despite pressure to the contrary, ongoing mutation in general does result in ESE disruptions and ESS creations that must be counterbalanced by ongoing ESE creations and ESS disruptions. We next set out to see if there was evidence for the latter two explanations in the form of splicing-negative changes being compensated by splicing-positive changes.

## Evolutionary compensation of exonic splicing-negative motif changes by exonic splicing-positive motif changes

Regardless of the evolutionary pressure for conservation of ESEs and against conservation of ESSs, the fact remains that differences that are predicted to decrease splicing in one species relative to the other are abundant in human and macaque exons. Macaques and humans have evolved from some common ancestor, and both species have maintained constitutive exon splicing despite the presence of these potentially deleterious mutations. We reasoned that many of these negative changes should be matched by compensatory positive changes. The compensatory mutations would include nonsynonymous as well as synonymous mutations, mutations at splice sites, and mutations in flanking intronic sequences. In this model, each exon together with its flanks is conserved as a splicing unit that must maintain a functional balance between positive and negative elements; or, for constitutive exons, at least a dominance of positive elements over negative elements.

We set out to test this idea by simply making two-way comparisons, considering human exons as functional mutant versions of their macaque orthologs and vice versa. By abandoning the use of an outgroup we lose the ability to assign a direction to any given change (i.e., assignment as a creation or disruption) as well as the ability to deduce the order of the mutational events (i.e., whether a compensatory positive change preceded or succeeded a negative change). However, the use of an outgroup here would hide some compensatory changes by discarding positions where all three species differ and would decrease the number of informative exons (data not shown). Additionally, differences at CpG dinucleotides and nonsynonymous changes were now taken into account, as their exclusion could conceal compensatory changes.

We defined an exonic splicing-negative mutation (spl−) as a single base substitution that results in the disappearance of an ESE or the appearance of an ESS and a splicing-positive mutation (spl+) as resulting in the appearance of an ESE or the disappearance of an ESS (see Methods). We describe most of the results in terms of how the human exons differ from macaque, but for all experiments we repeated the analysis for how macaque exons differ from human, as that analysis provides a comparable amount of information and demonstrates the reproducibility of the data.

We used a set of ~50,000 human–macaque constitutive exon pairs for which the two exons of each pair were of equal length and contained no gaps in the alignment. Each human exon was examined for exonic splicing motif differences relative to its macaque ortholog, including both synonymous and nonsynonymous differences. To normalize for the effects of exon length, we calculated the frequency of differences as spl+ or spl− mutations per nucleotide for each exon and classified the exons according to their spl− frequencies. For each class we then calculated the frequency of spl+ mutations that had also occurred. As can be seen in Figure 3A (black bars), the frequency of spl+ mutations increases as the frequency of spl− mutations increases. Although this result is consistent with the splicing compensation model, we were concerned that other factors could contribute to the correlation. For instance, the exon population exhibits a range of overall mutation frequencies; thus, exons that have suffered many total mutations would be expected to have a high frequency of both spl− and spl+ mutations. To take such correlations into account, we designed a simulated set of mutant exons derived from macaque exons. For each macaque–human exon pair, we replaced each of the synonymous mutations in the human exon with mutations of the same exact base difference but at random synonymous positions. The restriction to synonymous differences was imposed to avoid mutations that could affect protein function and therefore would be less likely to have been conserved. It should be noted that these restrictions made our simulation quite conservative, in that the randomly placed differences were sometimes placed coincidently at the same positions as the naturally occurring mutations. The frequency of spl+ mutations in this simulated set did in fact increase with the frequency of spl− mutations (Fig. 3A, white bars) but to a significantly lesser degree than in the real set. Subtracting the frequency of mutations produced in the simulation from the real exon data isolates the splicing compensation effect (Fig. 3B).

Next, we repeated the analysis but restricted the differences to those where directionality could be determined by reference to dog as an outgroup. Thus, in this experiment spl− changes represent ESE disruptions and ESS creations and spl+ changes represent ESE creations and ESS disruptions in either the human or the macaque (combining the data for the two species). The results (Fig. 3C) are very similar to those found for the two-way differences although slightly damped, as expected.

An additional control experiment was carried out in which mutation in randomly chosen sets of motifs drawn from non-ESE and non-ESS hexamers were enumerated. In contrast to the result with the splicing motifs, no difference between the real exons and the simulated exons was evident (Fig. 3D).

We repeated these analyses viewing macaque exons as functional mutant derivatives of human exons and obtained similar results (Supplemental Fig. S2). We conclude that there is a tendency for exonic splicing-negative motif changes to be compensated by exonic splicing-positive motif changes.

## Compensation between splice-site differences and exonic splicing motif differences

Mutations that disrupt ESEs or create ESSs might also be compensated by an increase in the strength of one or both of the splice sites of the exon. By strength here we mean agreement to the splice-site consensus sequence, as scored by a position-specific scoring matrix. By the same token, mutations that decrease the strength of a 3′ and/or 5′ splice site of an exon could

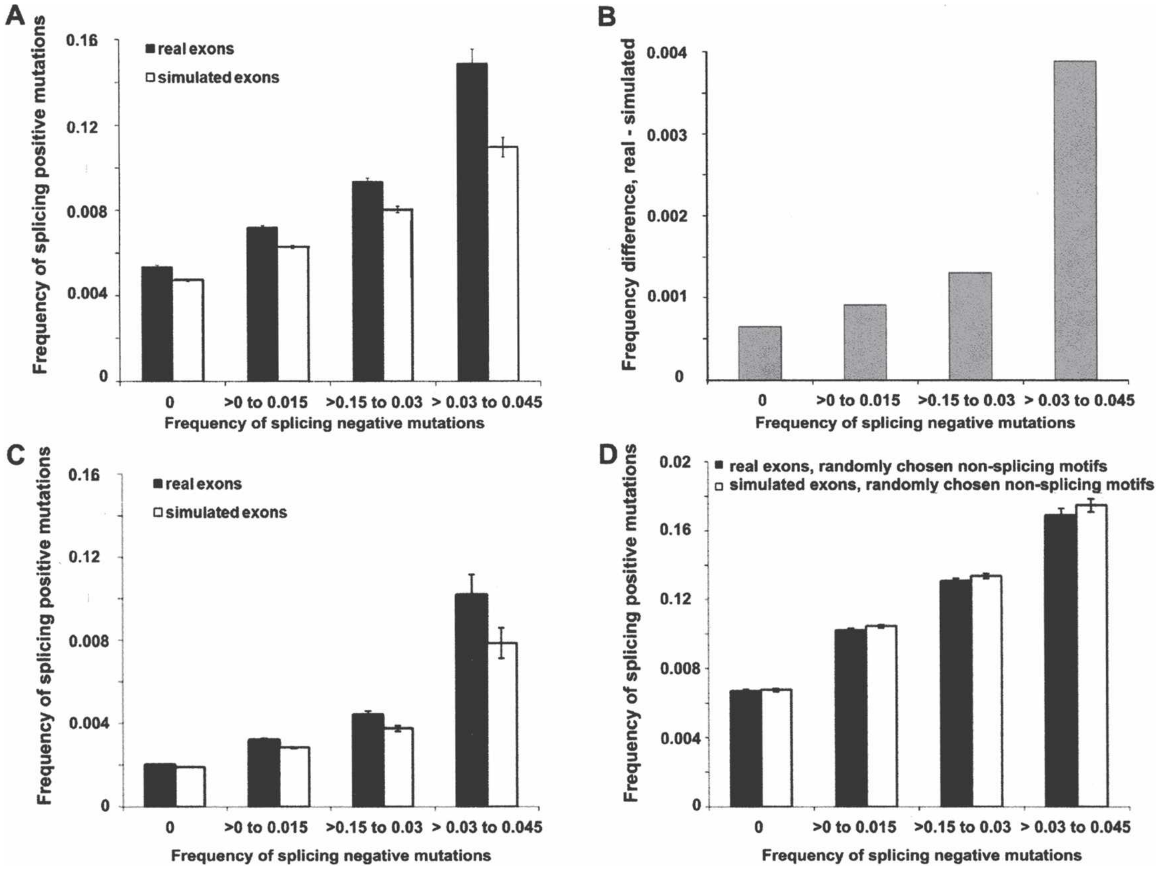

**Figure 3.** Evolutionary compensation of exonic splicing-negative motif changes by exonic splicing-positive motif changes. (*A*) Comparing human to macaque: exons were separated into four groups according to their frequency of exonic splicing-negative events (events per nt) as indicated. Exons with frequencies of 0–0.015, 0.015–0.030, and 0.030–0.045 generally contain 1 or 2, 3 or 4, and 5 or 6 splicing-negative differences, respectively. Exon pairs with frequencies higher than 0.045 have been omitted as many of them are poorly aligned. Black bars: real exons; white bars, simulated exons. (*B*) Net difference between the real and simulated sets is shown for each splicing-negative mutation category. (*C*) Same analysis as in *A*, except restricted to splicing changes in which the directionality of the change was determined by the use of dog as an out-group. Thus, in this panel the mutations in human exons can be considered ESE and ESS creations and disruptions rather than simply differences from macaque. (*D*) Same analysis as in *A*, but measuring differences using two randomly chosen non-overlapping sets of equal numbers of non-ESEs and of non-ESSs as controls. Similar results comparing macaque exons to human exons are presented in Supplemental Fig. S2.

be compensated by the creation of ESEs or the disruption of ESSs. To investigate this sort of compensation, we combined the positive and negative motif differences undergone by an exon, defining the exonic splicing motif difference (ESMD) as the frequency of splicing-positive differences (ESE appearances and ESS disappearances) in human relative to macaque minus the frequency of splicing-negative differences (ESS appearances and ESE disappearances). A positive ESMD is predicted to promote splicing, while a negative ESMD would discourage splicing. We asked whether a decrease in ESMD was correlated with a strengthening of one or both splice-site sequences and whether a weakening of a splice site was correlated with an increase in ESMD for an exon. Splice-site strength was measured by calculating a consensus value (CV) using a position-specific scoring matrix (Senapathy et al. 1990; Zhang et al. 2005b). To decrease noise, a score difference threshold of 5 on this scale was applied. As can be seen in Figure 4A, human exons with CVs that are identical to their macaque counterparts show little net change in ESMD. In contrast, exons with a weaker splice-site sequence exhibit an overall increase in ESMD, and those that have suffered a decrease in ESMD are associated with a stronger splice-site sequence. Both of these differences between unchanged and changed splice-site strengths are significant, with *P*-values ≤ 0.01, and control non-ESE/non-ESS motifs do not show these differences. A compensatory effect can also be seen by enumerating the number of exons that have undergone such changes. Among exons that have acquired a weaker splice site, those with an increase in EMSD outnumber those with a decrease (Fig. 4B, left, $P < 0.006$). The opposite is true for exons with stronger splice sites (Fig. 4B, right, $P < 0.0006$). In contrast, EMSD increases and decreases are equally represented among exons with no change in CV (Fig. 4B, middle; $P = 0.60$). Again, we repeated this analysis from the macaque point of view and obtained similar results (Supplemental Fig. S3). We conclude that changes in splicing regulatory motifs can compensate for changes in splice-site sequence strength and vice versa.

Despite the overall trend toward compensation seen in Figure 4, it is evident there are large numbers of exons that do not follow a compensatory route. This discrepancy may be more apparent than real and is not so surprising given our limited knowledge of the elements that contribute to splicing. CV scores were used as an index of splice-site strength, yet we know they are

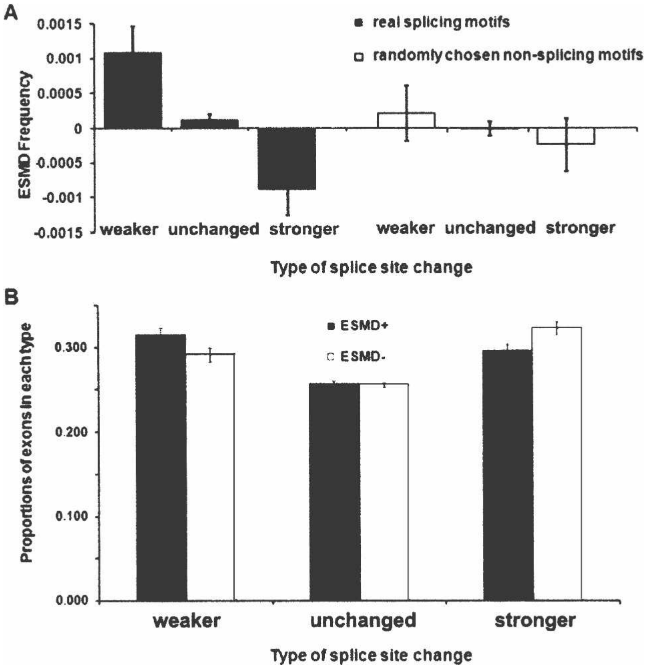

**Figure 4.** Compensation between splice-site changes and exonic splicing motif changes (comparing human to macaque). (*A*) Splicing-positive splice-site changes correlate with splicing-negative motif changes and vice versa. The exonic splicing motif difference (ESMD) is defined as the frequency of splicing-positive changes (ESE creations and ESS disruptions) minus the frequency of splicing-negative changes (ESS creations and ESE disruptions) in human relative to macaque. A positive ESMD is predicted to promote splicing while a negative ESMD would discourage splicing. Exons in the "weaker" set have one splice site at which the CV score (consensus values) has decreased by at least 5 on a CV scale of 0–100 and in which the other splice site has not increased by >5. Exons in the "stronger" set are defined in the opposite way. Exons in the "unchanged" set show no change at all in CV score for both 3′ and 5′ splice sites. The results of a control using non-ESE/ESS motifs, as described in the legend to Fig. 3D, are shown on the *right*. (*B*) Proportion of exons that have undergone changes in splice-site sequence and splicing motifs reflects compensation. Standard errors are indicated in both panels. The total number of exons in the weaker, unchanged, and stronger sets are 2897, 20,855, and 3394, respectively. Similar results comparing macaque exons to human are presented in Supplemental Fig. S3.

poor predictors of splice sites (Lear et al. 1990; Sun and Chasin 2000); a predicted "strengthening" can even have the opposite effect (Carothers et al. 1993). Similarly, we do not yet understand the possible combinatorial rules for ESE and ESS interaction. An ESE can act as an ESS and vice versa depending on position (Kanopka et al. 1996; Goren et al. 2006). Finally, compensatory changes may be acting via changes in intronic motifs flanking these exons (Zhang et al. 2003, 2005c; Yeo et al. 2005); we did not search for intronic changes here in the absence of validated global sets of such motifs. The predicted compensatory trends were evident in our experiments despite these limitations.

## Specific examples of predicted compensation or noncompensation

Some specific examples of the types of changes represented in these statistics are shown in Figure 5. In Figure 5A we show the loss of a single predicted ESE in human (top sequence) compared to macaque (bottom sequence), together with the loss of a single ESS, a potential case of motif compensation. In Figure 5B, a 12-point weakening of a 5′ splice site is shown, with a potential compensation in the form of a gain in four ESEs and the loss of one ESS. In Figure 5C we show a counter example, in which no compensation is seen for the loss of three ESEs and the gain of two ESSs. Admittedly, we cannot know the order of these events, i.e., whether they represent compensation for or tolerance of a deleterious mutation. These examples suggest that it may be possible to learn more about the effect of context on splicing by the analysis of the two different combinations of elements that lead to (presumably) productive splicing of orthologous exons, as well as by the experimental manipulation and testing of such exon pairs. Moreover, negative examples (see Fig. 5C) may serve as useful datasets for discovering additional elements involved in splicing regulation.

## Global maintenance of exonic splicing motif numbers

The compensation model also predicts that to the extent that human and macaque exons diverged from those of their common ancestor their gains and losses of splicing motifs should be about the same overall. As can be seen in Figure 6, the average difference in motif numbers for each exon pair is close to zero for

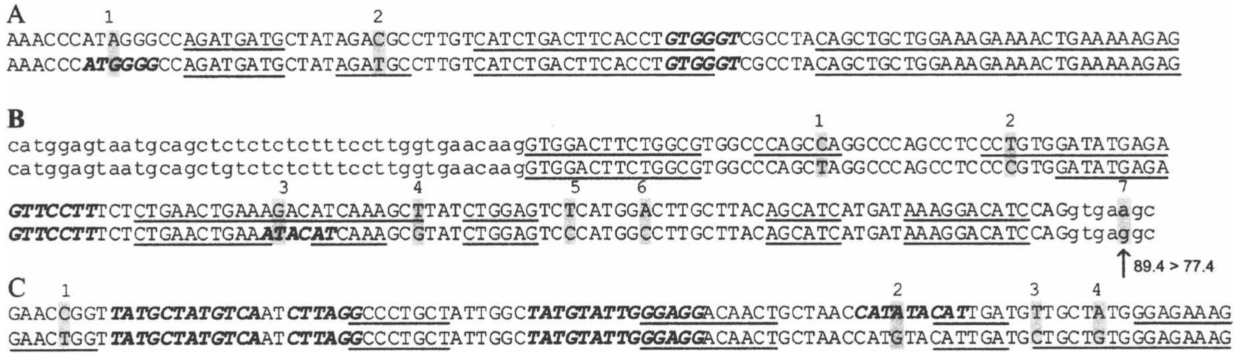
A 1 2
AAACCCATAGGGCCAGATGATGCTATAGACGCCTTGTCATCTGACTTCACCTGTGGGTCGCCTACAGCTGCTGGAAAGAAACTGAAAAAGAG
AAACCCATGGGGCCAGATGATGCTATAGATGCCTTGTCATCTGACTTCACCTGTGGGTCGCCTACAGCTGCTGGAAAGAAACTGAAAAAGAG
B 1 2
catggagtaatgcagctctctctctttccttggtgaacaagGTGGACTTCTGGCGTGGCCCAGCCAGGCCCAGCCTCCCTGTGGATATGAGA
catggagtaatgcagctgtctctctttccttggtgaacaagGTGGACTTCTGGCGTGGCCCAGCTAGGCCCAGCCTCCCCGTGGATATGAGA
3 4 5 6 7
GTTCCTTTCTCTGAACTGAAAGACATCAAAGCTTATCTGGAGTCTCATGGACTTGCTTACAGCATCATGATAAAGGACATCCAGgtgaagc
GTTCCTTTCTCTGAACTGAAATACATCAAAGCGTATCTGGAGTCCCATGGCCTTGCTTACAGCATCATGATAAAGGACATCCAGgtgaggc
89.4 > 77.4
C 1 2 3 4
GAACCGGTTATGCTATGTCAATCTTAGGCCCTGCTATTGGCTATGTATTGGGAGGACAACTGCTAACCATATACATTGATGTTGCTATGGAGAAAG
GAACTGGTTATGCTATGTCAATCTTAGGCCCTGCTATTGGCTATGTATTGGGAGGACAACTGCTAACCATGTACATTGATGCTGCTGTGGAGAAAG

**Figure 5.** Examples of compensatory and noncompensatory changes. *Top* sequences are human, *bottom* sequences macaque. Mutations are shaded, ESEs are underlined, ESSs are in bold italics, exons are in *upper* case, and introns are in *lower* case. (*A*) Decrease of an ESE in human compensated by the decrease of an ESS. Mutation 2 results in the absence of an ESE, potentially compensated by the presence of an ESS (mutation 1). (*B*) Weakened splice site compensated by an increase of an ESE or decrease of an ESS or both. Mutation 7 weakens the 5′ splice-site CV score from 89.4 to 77.4 (arrow), while mutations 1, 2, 3, and 4 result in ESE appearances and mutation 3 disrupts an ESS as well. Mutations 5 and 6 are predicted to be neutral. (*C*) Decrease in ESEs and increase in ESSs with no compensation. Mutations 1, 3, and 4 disrupt ESEs and mutation 2 creates two overlapping ESSs, with no further changes in the exon.

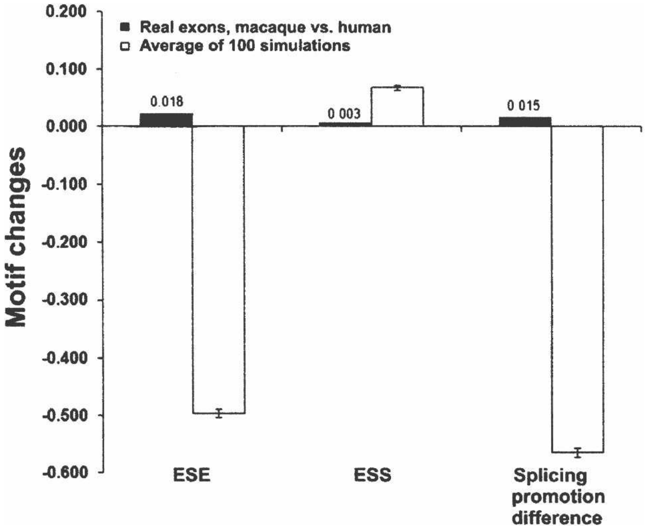

**Figure 6.** Global maintenance of exonic splicing motifs (comparing human to macaque). Motif change is defined as the number of ESE or ESS hexamers in a human exon minus the number in its macaque ortholog. Splicing promotion for an exon is defined as the ESE number minus the ESS number for that exon. In the simulated sets, the same number and type of differences seen in human exons were placed randomly among synonymous sites of macaque exons, as described in the text. Similar results simulating macaque differences in human exons are shown in Supplemental Fig. S4.

both ESEs and ESSs and is even smaller if considered as splicing promotion (number of ESEs minus number of ESSs for each exon). As a control for this measurement, we once again made use of the simulated changes described above, reproducing the same exact base differences from macaque to human but placing them at randomly chosen synonymous positions. The differences in the simulated sets were one or two orders of magnitude greater than those seen for real exons (Fig. 6). Without any evolutionary constraint to maintain splicing efficiency, the randomly placed mutations tend to disrupt ESEs and create ESSs. Furthermore, the variance of differences of predicted splicing motifs for each exon pair is significantly smaller in the real exon data compared to the simulated set, demonstrating a tighter distribution of differences (Table 2). For instance, the variance for splicing promotion differences is 9.09 ± 0.006 for simulated mutations compared to 8.24 for the real exons ($P < 10^{-22}$). Thus, the real differences between human and macaque exons are constrained, with exons being conserved as splicing units capable of efficient inclusion in each species. Both human and macaque exons reach an optimal composition of motifs to allow efficient splicing, albeit sometimes by different routes. Similar results were obtained from the analysis of differences in macaque relative to human (where the simulation is distinct; Supplemental Fig. S4; Supplemental Table S2) and when the directionality of the changes was determined using the dog as an out-group (data not shown).

**Table 2.** Variance of the distribution of differences in number of motifs between human and macaque orthologous exons

| | Variance | | |
|---|---|---|---|
| Motif difference | Real set | Simulated set[a] | *P*-value |
| ESE | 5.96 | 6.30 | $<10^{-11}$ |
| ESS | 1.72 | 1.95 | $<10^{-15}$ |
| Splicing promotion | 8.24 | 9.09 | $<10^{-22}$ |

[a] Random positioning of human mutations relative to macaque.

## Summary and implications

### *Predicted ESE and ESS motifs exhibit evolutionary behavior consistent with their proposed function*

In our previous work (Zhang and Chasin 2004; Zhang et al. 2005a), we defined putative splicing motifs as 8-mers; here they were converted to 6-mers to enable their amalgamation with analogous motifs described by the Burge laboratory (Fairbrother et al. 2002; Wang et al. 2004). The latter were defined using different strategies, and the sets from the two laboratories were given approximately equal weight in the combined sets used here. We found that these predicted exonic splicing enhancer and silencer motifs exhibited evolutionary behavior consistent with roles in promoting and inhibiting splicing, respectively. This distinctive evolutionary behavior supports the overall quality of these sets as functional splicing motifs. In particular, the ESEs and ESSs exhibited opposite evolutionary trends, consistent with their interpretation as elements that act with a considerable degree of autonomy. This contrary behavior would not be expected if ESEs and ESSs are considered to be facultative motifs entirely dependent on context ("ESRs," as proposed by Goren et al. 2006). This is not to say that this autonomy is complete, as there is experimental evidence that context can indeed determine the action of a splicing motif (e.g., Kanopka et al. 1996; Goren et al. 2006). Rather, we conclude that ESEs and ESSs usually and overall act according to their designation.

### *Synonymous sites are subject to strong purifying selection*

Over the last few years, different strategies have been used to show that bases at synonymous sites overall are conserved, rather than being neutral in evolution. Our findings support this idea with the application of a little-used but straightforward normalization method, that of comparing the rate of synonymous change to that of intronic change. Although there have been some differences in the magnitude and direction of this ratio (for review, see Chamary et al. 2006), our results are close to those of Hellmann et al. (2003) who found about a 40% lower $K_s$ than $K_i$ comparing human to chimpanzee. We found that conservation at synonymous sites is not limited to splicing motifs, implying that many other non–protein-coding sequence elements are functioning in mRNA. The pervasive presence of such elements could be a major determinant of codon usage bias.

### *ESEs and ESSs are subject to purifying selection*

The results reported here extend the findings of ESE conservation reported in two recent studies (Parmley et al. 2006; Stadler et al. 2006). Parmley et al. (2006) reported low $K_s$ values for ESEs compared to non-ESEs and showed that the conservation was not due to differential mutability caused by base composition or CpG content. Our use of $K_i$ extends this important control by normalizing for the mutability of each individual hexameric sequence. Stadler et al. (2006) did take such mutability into account by normalizing to the rate of change of each motif at 1000 randomly chosen genomic locations. However, nonsynonymous as well as synonymous changes were scored, allowing for some of the conservation seen to be attributable to selection for protein function. The $K_s/K_i$ measurement used here eliminates protein-coding effects. It also allowed us to see a strong conservation of non-ESE motifs in exons compared to introns, with the implication that an mRNA molecule contains many additional functional elements in addition to those for protein coding and splicing. ESSs have been less studied, although their conservation in regions between alternatively used 5′ or 3′ splice sites has been demonstrated (Stadler et al. 2006; Wang et al. 2006). Here we have presented evidence that mutations resulting in the creation of ESSs are also selected against, which would constitute a purifying selection. Finally, our use of an out-group has allowed us to track events as creations and disruptions rather than just as changes.

#### *ESEs and ESSs are subject to positive selection*

Using $K_i$ as a normalizing measure, Resch et al. recently presented evidence for positive selection at synonymous sites in several hundred human genes (Resch et al. 2007). The genomewide results reported here implicate positive selection for ESE creation and for ESS destruction as contributors to this effect. This positive selection is operating in constitutive exons but not alternative exons, as might be expected if efficient splicing is selected against in alternative exons.

#### *Compensatory changes conserve exons as splicing units*

Our previous experimental tests indicated that most ESEs in human exons are required for efficient splicing; i.e., extensive redundancy was not present (Zhang et al. 2005a). That result, coupled with evidence for positive selection reported here, prompted us to search for evidence of compensation for changes that discourage splicing by changes that encourage it. We interpret this compensation in the following way. Despite a purifying selection against changes that would tend to compromise efficient splicing, such mutations nonetheless continue to appear. Compensatory changes, the creation of new ESEs, the disruption of old ESSs, and the strengthening of splice-site sequences, act to ameliorate the effects of the deleterious mutations and are positively selected. The result is that net changes in splicing efficiency are minimized, as was evidenced also by the global maintenance of exonic splice motif numbers. Thus, the compensatory model paints a picture of exon evolution as a dynamic interplay between helpful and harmful mutations, continuously at work.

## Methods

### Multiple alignments of orthologous exons and introns

Human ESTs and mRNA sequences were downloaded from the UniGene database (ftp://ftp.ncbi.nih.gov/repository/UniGene/Homo_sapiens/Hs.seq.all.gz) and were aligned to the assembled genomic sequences obtained from (ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/) using Sim4. Only ESTs that span at least two exon-exon joints were considered. A perl script was written to retrieve the intron flanks from the alignment. Genes that exhibited no intron-exon joints were excluded. The coordinates of the exons and introns on the assembled genome were recorded. Based on these coordinates, orthologous sequences from chimpanzee, macaque and dog genomes were extracted from a 17-genome multi-alignment available at http://hgdownload.cse.ucsc.edu/goldenPath/hg17/multiz17way/. For introns, we ignored the highly repeated sequences, which have been premasked in the alignment files. We also ignored intronic sequences within 100 nt of exons, since these sequences are likely to harbor intronic splicing signals (Zhang et al. 2003, 2005c) and so be subject to selection. We surveyed 16.7 million nt of introns. In the human–macaque comparisons we surveyed 59,221 constitutive exons comprising 7.1 million nt. The corresponding numbers for the human–chimpanzee comparison were 56,949 and 6.8 million. Alternatively spliced cassette exons were chosen solely on the basis of the human phenotype. We surveyed 3696 alternative cassette exons comprising 0.40 million nt in the human–macaque comparison; the corresponding numbers for the human–chimpanzee comparison were 3675 and 0.40 million.

### Predicted ESEs, ESSs, and control motif sets

We combined predicted exonic splicing signals from our previous studies, putative ESEs (PESEs), and putative ESSs (PESSs) with RESCUE-ESE signals and FAS-hex3 ESSs, respectively, from Burge and colleagues (Fairbrother et al. 2002; Wang et al. 2004; Zhang and Chasin 2004; Zhang et al. 2005a). We merged the sequences from both groups to make a more complete list of splicing signals. For calculation of $K_s/K_i$, we counted hexamer frequencies in PESE octamers, collecting hexamers that occurred six or more times. We then took the union of RESCUE-ESEs and PESE hexamers. PESS hexamers and FAS-hex3 ESSs (Wang et al. 2004) were merged in the same way. These operations resulted in an ESE list and an ESS list comprising 469 and 246 hexamers, respectively. The ESE and ESS lists are each made up of families of sequences; sequences within each family vary slightly from one another, strongly influencing their propensity to be disrupted or created. To prepare control motif sets with this same degree of coherence, we generated the reverse complements of the hexamers in the ESE list and in the ESS list. The complementary sets will also exhibit the same G+C content and will be subject to identical position effects (if any) from their situation on the opposite strand. These control lists contained sequences that were also classified as ESEs or ESSs. We purged the hexamers that were common to the original lists and control lists, ending up with 403 ESE hexamers, 199 ESS hexamers and the same numbers of control hexamers.

Slightly different sets of ESE and ESS hexamers were used to study compensation. Hexamers that occurred seven or more times in the PESE octamer set (238) were merged with 238 RESCUE-ESE hexamers (Fairbrother et al. 2002) resulting in an ESE hexamer set of 400 unique sequences (omitting one that was also in the ESS set). Similarly, 120 PESS-derived hexamers were merged with 103 FAShex3 hexamers (Wang et al. 2004) to yield 217 unique ESS hexamers. Note that these hexamers sets were not purged of complementary sequences as above, as these sets were designed to be as complete as possible to maximize detection of compensatory changes. These hexamers (hexESEs and hexESSs) are listed in Supplemental Table S3. We also created larger hexamer sets (836 ESE hexamers and 477 ESS hexamers) using more relaxed criteria for their occurrence within the longer motifs sets (three or more for PESEs and PESSs and FAShex2 hexamers [Stadler et al. 2006]) and obtained essentially the same results.

PESE and PESS octamers were taken from the updated online lists at http://cubweb.biology.columbia.edu/pesx.

### Calculation of $K_s$, $K_a$, and $K_i$ and the odds ratio of $K_s$ and $K_i$

$K_s$ and $K_a$ were calculated according to Li (1993). To compute $K_i$, we first computed the proportion of nonidentical sites in the aligned (no gap) segments of introns and then used a Poisson distribution to estimate the average substitution rates.

To discriminate the direction of mutations, we used the outgroups as ancestral references and considered synonymous/

intronic substitutions that disrupt or create ESEs or ESSs separately. A substitution is considered to disrupt an ESE if it converts an ancestral ESE to a non-ESE in either in-group species. Conversely, a substitution is said to create an ESE if it converts a non-ESE in the out-group genome to an ESE in either in-group genome. Changes involving CpG dinucleotides present in any species were ignored. The frequencies of synonymous substitutions that disrupt or create ESEs or ESSs are designated as $K_{\mathrm{s\ disrupt}}$ and $K_{\mathrm{s\ create}}$, respectively. Similarly, $K_{\mathrm{i\ disrupt}}$ and $K_{\mathrm{i\ create}}$ represent the frequencies of intronic substitutions that change ESE and ESS sequences. The odds ratio (OR) was used to compare the ratios of mutation rates; log(OR) approximately follows a normal distribution with simple standard deviations (Bland and Altman 2000). For $K_{\mathrm{s}}$ and $K_{\mathrm{i}}$, OR was approximated by

$$\mathrm{OR} = \frac{N_{ss}/(N_s - N_{ss})}{N_{is}/(N_i - N_{is})} \quad \text{or} \quad \frac{K_s/(1 - K_s)}{K_i/(1 - K_i)},$$

where $N_{\mathrm{ss}}$ is the number of synonymous substitutions, $N_{\mathrm{s}}$ is the number of synonymous sites examined, $N_{\mathrm{is}}$ is the number of intronic substitutions, and $N_{\mathrm{i}}$ is the total number of alignable intronic nucleotides examined. The confidence interval of ORs can be estimated using the logarithm of ORs, which approximately follows a normal distribution with a standard variance of

$$\sigma[\ln(\mathrm{OR})] = \sqrt{[1/(N_s - N_{ss})] + (1/N_{ss}) + [1/(N_i - N_{is})] + (1/N_{is})}.$$

See Bland and Altman (2000) for details.

### Splicing motif changes

The creation of an ESE was defined as an event in which at least one ESE was created by a single base substitution at a position at which there was no ESE among the six hexamers overlapping the base that was changed. The disruption of an ESE was defined as an event in which a single base substitution removed any and all ESE hexamers overlapping the base that was changed. ESS creation and disruption were similarly defined. A splicing-negative event (spl−) is defined as a mutation that either creates an ESS or disrupts an ESE; a splicing-positive event (spl+) is defined as either disrupting an ESS or creating an ESE. Compensatory changes were measured at all exonic sites, including CpG dinucleotides and nonsynonymous sites.

### Simulated mutations

We simulated human exon mutants by retaining all exonic nonsynonymous mutations while randomly replacing synonymous mutations with the same number and type (e.g., T to A) of exonic synonymous mutations compared to their macaque exon ortholog. This strict exon simulation maximally mimicked the synonymous mutational differences between human and macaque and sought to create a set of changes in which only selection for splicing effects had been minimized. In the same way, we simulated macaque exon mutants with respect to human orthologs. In the simulation of human exon mutants using dog as an outgroup, the restrictions were even stricter as only those mutations that were unique in human (macaque and dog being the same) were randomly replaced. The use of simulated exons as a control circumvents the problem posed by exons that have suffered an unusually large number of mutations, as these would show a correlation between mutations of any sort, and will be evident in the control.

### Grouping exons with splicing-negative changes

Exons were separated into four groups according to their frequency of exonic splicing-negative events relative to their ortholog. Exons in the first group have zero exonic splicing-negative events; exons in the second group have a frequency up to 0.015 spl− mutations per nucleotide, and so on, up to 0.045. There were generally 1–2, 3–4, and 5–6 spl− mutations in the 0–0.15, 0.015–0.03, and 0.03–0.045 frequency sets, respectively. Exons with frequencies higher than 0.045 were omitted as most of them exhibited poor overall alignments.

### Compensation between splice-site differences and exonic splicing motif differences

The exonic splicing motif difference (ESMD) is defined as the number normalized by exon length of splicing-positive differences (ESE creations and ESS disruptions) in human relative to macaque minus the number of splicing-negative differences (ESS creations and ESE disruptions). A positive ESMD is predicted to enhance splicing while a negative ESMD would weaken splicing. Splice-site strengths were estimated using consensus values (CV) as described previously (Zhang et al. 2005b).

## Acknowledgments

We thank Dennis Weiss for insightful comments in the course of this work and for a critical reading of the manuscript, and Mauricio Arias, Shulian Shang, David Berman, Ron Gejman, John Cacciatore, and Harmen Bussemaker for useful discussions. We thank an anonymous reviewer for suggesting the McDonald–Kreitman test. This work was supported by a grant from the NIH (GM072740) to L.A.C.

## References

Arndt, P.F., Burge, C.B., and Hwa, T. 2003. DNA sequence evolution with neighbor-dependent mutation. *J. Comput. Biol.* **10:** 313–322.

Birky Jr., C.W. and Walsh, J.B. 1988. Effects of linkage on rates of molecular evolution. *Proc. Natl. Acad. Sci.* **85:** 6414–6418.

Bland, J.M. and Altman, D.G. 2000. Statistics notes. The odds ratio. *BMJ* **320:** 1468.

Carlini, D.B. and Genut, J.E. 2006. Synonymous SNPs provide evidence for selective constraint on human exonic splicing enhancers. *J. Mol. Evol.* **62:** 89–98.

Carothers, A.M., Urlaub, G., Grunberger, D., and Chasin, L.A. 1993. Splicing mutants and their second-site suppressors at the dihydrofolate reductase locus in Chinese hamster ovary cells. *Mol. Cell. Biol.* **13:** 5085–5098.

Chamary, J.V., Parmley, J.L., and Hurst, L.D. 2006. Hearing silence: Non-neutral evolution at synonymous sites in mammals. *Nat. Rev. Genet.* **7:** 98–108.

Charlesworth, B., Morgan, M.T., and Charlesworth, D. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134:** 1289–1303.

Chen, F.C., Wang, S.S., Chen, C.J., Li, W.H., and Chuang, T.J. 2006. Alternatively and constitutively spliced exons are subject to different evolutionary forces. *Mol. Biol. Evol.* **23:** 675–682.

Comeron, J.M. and Guthrie, T.B. 2005. Intragenic Hill–Robertson interference influences selection intensity on synonymous mutations in *Drosophila*. *Mol. Biol. Evol.* **22:** 2519–2530.

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

Fairbrother, W.G., Holste, D., Burge, C.B., and Sharp, P.A. 2004. Single nucleotide polymorphism–based validation of exonic splicing enhancers. *PLoS Biol.* **2:** e268. doi: 10.1371/journal.pbio.0020268.

Garg, K. and Green, P. 2007. Differing patterns of selection in alternative and constitutive splice sites. *Genome Res.* **17:** 1015–1022.

Goren, A., Ram, O., Amit, M., Keren, H., Lev-Maor, G., Vig, I., Pupko, T., and Ast, G. 2006. Comparative analysis identifies exonic splicing regulatory sequences—The complex definition of enhancers and silencers. *Mol. Cell* **22:** 769–781.

Hellmann, I., Zollner, S., Enard, W., Ebersberger, I., Nickel, B., and Paabo, S. 2003. Selection on human genes as revealed by comparisons to chimpanzee cDNA. *Genome Res.* **13:** 831–837.

Hurst, L.D. 2002. The $K_{\mathrm{a}}/K_{\mathrm{s}}$ ratio: Diagnosing the form of sequence evolution. *Trends Genet.* **18:** 486.

Hwang, D.G. and Green, P. 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc. Natl. Acad. Sci.* **101:** 13994–14001.

Itoh, H., Washio, T., and Tomita, M. 2004. Computational comparative analyses of alternative splicing regulation using full-length cDNA of various eukaryotes. *RNA* **10:** 1005–1018.

Kan, Z., Garrett-Engele, P.W., Johnson, J.M., and Castle, J.C. 2005. Evolutionarily conserved and diverged alternative splicing events show different expression and functional profiles. *Nucleic Acids Res.* **33:** 5659–5666. doi: 10.1093/nar/gki834.

Kanopka, A., Muhlemann, O., and Akusjarvi, G. 1996. Inhibition by SR proteins of splicing of a regulated adenovirus pre-mRNA. *Nature* **381:** 535–538.

Kim, Y. 2004. Effect of strong directional selection on weakly selected mutations at linked sites: Implication for synonymous codon usage. *Mol. Biol. Evol.* **21:** 286–294.

Lear, A.L., Eperon, L.P., Wheatley, I.M., and Eperon, I.C. 1990. Hierarchy for 5′ splice site preference determined in vivo. *J. Mol. Biol.* **211:** 103–115.

Li, W.H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol.* **36:** 96–99.

Louie, E., Ott, J., and Majewski, J. 2003. Nucleotide frequency variation across human genes. *Genome Res.* **13:** 2594–2601.

McDonald, J.H. and Kreitman, M. 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* **351:** 652–654.

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. *Nat. Genet.* **34:** 177–180.

Pan, Q., Bakowski, M.A., Morris, Q., Zhang, W., Frey, B.J., Hughes, T.R., and Blencowe, B.J. 2005. Alternative splicing of conserved exons is frequently species-specific in human and mouse. *Trends Genet.* **21:** 73–77.

Parmley, J.L., Chamary, J.V., and Hurst, L.D. 2006. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Mol. Biol. Evol.* **23:** 301–309.

Resch, A.M., Carmel, L., Marino-Ramirez, L., Ogurtsov, A.Y., Shabalina, S.A., Rogozin, I.B., and Koonin, E.V. 2007. Widespread positive selection in synonymous sites of mammalian genes. *Mol. Biol. Evol.* **24:** 1821–1831. doi: 10.1093/molbev/msm100.

Rukov, J.L., Irimia, M., Mork, S., Lund, V.K., Vinther, J., and Arctander, P. 2007. High qualitative and quantitative conservation of alternative splicing in *Caenorhabditis elegans* and *Caenorhabditis briggsae*. *Mol. Biol. Evol.* **24:** 909–917.

Senapathy, P., Shapiro, M.B., and Harris, N.L. 1990. Splice junctions, branch point sites, and exons: Sequence statistics, identification, and applications to genome project. *Methods Enzymol.* **183:** 252–278.

Siepel, A. and Haussler, D. 2004. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol. Biol. Evol.* **21:** 468–488.

Sorek, R. and Ast, G. 2003. Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res.* **13:** 1631–1637.

Stadler, M.B., Shomron, N., Yeo, G.W., Schneider, A., Xiao, X., and Burge, C.B. 2006. Inference of splicing regulatory activities by sequence neighborhood analysis. *PLoS Genet.* **2:** e191. doi: 10.1371/journal.pgen.0020191.

Sugnet, C.W., Kent, W.J., Ares Jr., M., and Haussler, D. 2004. Transcriptome and genome conservation of alternative splicing events in humans and mice. *Pac. Symp. Biocomput.* **9:** 66–77.

Sun, H. and Chasin, L.A. 2000. Multiple splicing defects in an intronic false exon. *Mol. Cell. Biol.* **20:** 6414–6425.

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

Wang, Z., Xiao, X., Van Nostrand, E., and Burge, C.B. 2006. General and specific functions of exonic splicing silencers in splicing control. *Mol. Cell* **23:** 61–70.

Wyckoff, G.J., Malcom, C.M., Vallender, E.J., and Lahn, B.T. 2005. A highly unexpected strong correlation between fixation probability of nonsynonymous mutations and mutation rate. *Trends Genet.* **21:** 381–385.

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.* **102:** 13526–13531.

Xing, Y. and Lee, C. 2006. Can RNA selection pressure distort the measurement of $K_a/K_s$? *Gene* **370:** 1–5.

Yeo, G., Hoon, S., Venkatesh, B., and Burge, C.B. 2004. Variation in sequence and organization of splicing regulatory elements in vertebrate genes. *Proc. Natl. Acad. Sci.* **101:** 15700–15705.

Yeo, G.W., Van Nostrand, E., Holste, D., Poggio, T., and Burge, C.B. 2005. Identification and analysis of alternative splicing events conserved in human and mouse. *Proc. Natl. Acad. Sci.* **102:** 2850–2855.

Zhang, X.H. and Chasin, L.A. 2004. Computational definition of sequence motifs governing constitutive exon splicing. *Genes & Dev.* **18:** 1241–1250.

Zhang, Z. and Gerstein, M. 2003. Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucleic Acids Res.* **31:** 5338–5348.

Zhang, X.H., Heller, K.A., Hefter, I., Leslie, C.S., and Chasin, L.A. 2003. Sequence information for the splicing of human pre-mRNA identified by support vector machine classification. *Genome Res.* **13:** 2637–2650.

Zhang, X.H., Kangsamaksin, T., Chao, M.S., Banerjee, J.K., and Chasin, L.A. 2005a. Exon inclusion is dependent on predictable exonic splicing enhancers. *Mol. Cell. Biol.* **25:** 7323–7332.

Zhang, X.H., Leslie, C.S., and Chasin, L.A. 2005b. Computational searches for splicing signals. *Methods* **37:** 292–305.

Zhang, X.H., Leslie, C.S., and Chasin, L.A. 2005c. Dichotomous splicing signals in exon flanks. *Genome Res.* **15:** 768–779.

*Received August 14, 2007; accepted in revised form January 16, 2008.*