Please cite this article in press as: Ramensky et al., Positive Selection in Alternatively Spliced Exons of Human Genes, The American Journal of Human Genetics (2008), doi:10.1016/j.ajhg.2008.05.017

Positive Selection in Alternatively Spliced Exons of Human Genes

Vasily E. Ramensky,[1,*] Ramil N. Nurtdinov,[2] Alexei D. Neverov,[3,4] Andrei A. Mironov,[2,3] and Mikhail S. Gelfand[2,5]

Alternative splicing is a well-recognized mechanism of accelerated genome evolution. We have studied single-nucleotide polymorphisms and human-chimpanzee divergence in the exons of 6672 alternatively spliced human genes, with the aim of understanding the forces driving the evolution of alternatively spliced sequences. Here, we show that alternatively spliced exons and exon fragments (alternative exons) from minor isoforms experience lower selective pressure at the amino acid level, accompanied by selection against synonymous sequence variation. The results of the McDonald-Kreitman test suggest that alternatively spliced exons, unlike exons constitutively included in the mRNA, are also subject to positive selection, with up to 27% of amino acids fixed by positive selection.

Alternative splicing (AS) of genes is the processing of an RNA transcript into various mRNA molecules (and hence proteins) by the inclusion of some exons and the exclusion of others. It is one of the major sources of eukaryotic proteome complexity,[1,2] with the estimates of the fraction of alternatively spliced genes in the human genome gradually increasing over time, from 35%[3] up to 70%–80%.[4,5] Alternative splicing also seems to be an important mechanism of accelerated genome evolution, associated with relaxation of the selection pressure against certain types of evolutionary events on both the level of the exon-intron structure and the level of individual nucleotides.[2,6,7] In particular, AS significantly relaxes the selection pressure against amino acid substitutions,[8] with stronger manifestation in the C-terminal alternatives.[9] Simultaneous increase in the conservation of synonymous sites is also observed.[8] Species-specific AS is also associated with lower amino acid conservation in constitutive exons,[10] with evolutionarily young exons evolving at a higher rate.[11] However, none of the listed studies attempted to distinguish between positive selection and relaxed negative selection.

A common measure of the degree of evolutionary constraint on a sequence is the ratio *K*a/*K*s of nonsynonymous substitutions per nonsynonymous site (*K*a) to synonymous substitutions per synonymous site (*K*s), with higher values of the ratio indicating weaker negative selection acting on a sequence.[9] In the past years, several approaches have been applied to the detection of positive selection in the human genome,[12,13] but not all of them can be used in the case of closely intermixed genomic regions, characteristic of genomic AS. However, the availability of the SNP data makes it possible to apply a more-advanced analysis via the McDonald-Kreitman test,[14] which enables the detection of positive Darwinian selection in the presence of negative selection.[15,16] Here, we investigate the differences between the rates of synonymous and nonsynonymous polymorphism and of human-chimpanzee variation in alternative and constitutive regions of human genes. The novelty of the study is in the simultaneous analysis of polymorphism and divergence data in the coding regions of two types, revealing not only the relaxation of negative selection pressure associated with AS[8,9,11] but also the presence of a significant difference in positive-selection strength between the alternative and the constitutive regions.

The splicing alternatives were derived from EDAS,[17] a database containing genomic, protein, mRNA, and EST sequences, and then processed by IsoformCounter.[18] The latter is a conservative algorithm for the construction of a representative and nonredundant set of protein isoforms compatible with AS data for a gene (see ref. [18] for isoform-quality-validation details). A generated isoform is accompanied by the start and stop codons, the reading frame, and the effective number of sequences representing each coding region, thus enabling markup of a gene into alternative and constitutive fragments with known reading frames, as shown in Figure S1. Exons not completely aligned to the chimpanzee genome, genes without polymorphism data, and sequence variations in the first and last codons of a fragment were not considered. Exons were defined as conserved if they had donor and acceptor sites conserved in at least one of the orthologous mouse and dog genes and as nonconserved if either the exon or the entire gene was not conserved in both dog and mouse. To avoid contamination by misspliced or otherwise nonfunctional isoforms, we considered only conserved exons or nonconserved but still functional exons from genes for which at least 60 EST sequences were known. This procedure generated 52,151 constitutive and 14,196 alternative gene exons in 6672 human genes. Similar to Xing

[1]Bioinformatics and Systems Biology Lab; Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, 119991, Russia; [2]Department of Bioengineering and Bioinformatics, Moscow State University, Moscow, 119992, Russia; [3]Bioinformatics Lab; State Scientific Center GosNIIGenetika, Moscow, 117545, Russia; [4]Center of Molecular Diagnostics; Central Research Institute of Epidemiology, Moscow, 111123, Russia; [5]Research and Training Center "Bioinformatics", Institute for Information Transmission Problems (Kharkevich Institute), Russian Academy of Sciences, Moscow, 127994, Russia
*Correspondence: ramensky@imb.ac.ru

DOI 10.1016/j.ajhg.2008.05.017. ©2008 by The American Society of Human Genetics. All rights reserved.

and Lee,[8] we define the exon-inclusion level A, $0 < A \leq 1$, as the total number of all protein, mRNA, and EST sequences that cover the corresponding exon and include it completely divided by the total number of sequences covering the exon. All regions were then divided into three major classes: minor ($0 < A < 2/3$; N = 4397 fragments), major ($2/3 \leq A < 1$; N = 9799), and constitutive ($A = 1$; N = 52,151).

Furthermore, 6465 SNPs from build 125 of the dbSNP database[19] and 50,649 human-chimpanzee substitutions derived from the whole-genome alignments[20] were mapped to the genes. In order to overcome the biased nature of the polymorphism data submitted to dbSNP, we considered only SNPs marked as "validated" and obtained by the following methods:[21] EGP_SNPS (NIEHS/Seattle data), CSHL-HAPMAP, TSC-CSHL, SC, SC_JCM, SC_SNP, BCM_SSAHASNP, SSAHASNP, WI_SSAHASNP, or WUGSC_SSAHASNP. SNP frequencies were derived from genotype data in dbSNP build 128. This forms the basic set of SNPs. We also considered separately a subset of 4977 SNPs from the basic set observed in either James Watson's Personal Genome Sequence[22] or Celera's Individual A (dbSNP method WGSA-200403). Also, 6499 SNPs from the Applera data[16] mapped to the AS genes were used as a separate test set.

The *K*a/*K*s ratios for SNPs and divergence (Figure 1) are very close for constitutive and major alternative regions but differ more than two-fold from those for functional minor alternative regions, confirming that lower negative selection against amino acid substitutions is especially characteristic of these fragments. These observations hold when only neutral SNPs are considered or when conserved and nonconserved exons are considered separately (the complete data from Figure 1 and Table 1 are given in Table S1). Table 1 also contains the results of the McDonald-Kreitman test, in which the numbers of synonymous polymorphisms (*P*s) and substitutions (*D*s) and the numbers of nonsynonymous polymorphisms (*P*a) and substitutions (*D*a) are compared in a contingency table. A statistically significant excess of nonsynonymous substitutions relative to polymorphisms ($Da/Ds > Pa/Ps$) implies positive selection that provides fixation of advantageous mutations. The fraction α of fixed amino acid substitutions driven by positive selection is then estimated as $\alpha = 1-(Pa/Ps)/(Da/Ds)$.[23] This value equals 0.27 for the minor alternatives, suggesting that they experience positive selection. For major alternatives and constitutive exons, $\alpha < 0$ indicates purifying selection (Table 1A). The Fisher test value *F* given in the last column of Table 1 and Table S1 reflects the probability that the difference between *D*a/*D*s and *P*a/*P*s is random. In the basic set, it equals 0.001 for minor alternatives. Taking into consideration that synonymous and nonsynonymous SNPs are not independent in individual genes, we performed 10,000 permutation tests. In each test, types of 52,151 constant and 4397 minor alternatively spliced exons from the basic set were randomly permuted between the two categories of exons, and the corresponding values were calculated for each category. The observed value $\alpha = 0.27$ falls within the 0.0015 quantile of thus generated random distribution of α. The result of the permutation test does not provide direct evidence for positive selection, but it gives an independent support to the statistical significance of the *D*a/*D*s excess over *P*a/*P*s in the minor alternative exons.

We have repeated the analysis on SNPs from the basic set observed in two individual genomes (Table S1F). The values of α for constitutive, major, and minor exons were –0.05, –0.02, and 0.23, respectively, confirming the basic result albeit with a lower statistical significance due to the smaller number of SNPs in the subset (Fisher test value 0.02 for the minor alternatives). The results of the McDonald-Kreitman test on SNPs from the Applera dataset mapped to the genes with AS are given in Table S1G. Formally, α equals 0.24 for minor isoforms and is in good agreement with previous results. However, the statistical significance of this observation is low (the Fisher test value is 0.13), due to the small number of SNPs in the minor alternative exons. The primer-selection procedure described in ref. [16] yielded a low fraction of SNPs from this dataset mapping to gene introns (< 40%). This suggests that the coverage provided by this set is skewed toward constitutive exons.

Positive selection in the minor alternatives is still observed when the exons are split into conserved ($\alpha = 0.35$, Table 1B), and nonconserved ($\alpha = 0.23$, Table 1C) categories. On the other hand, the α values in the constitutive and major alternative exons differ between these two groups, with negative selection ($\alpha < 0$) characteristic for the conserved exons and positive selection ($\alpha > 0$) dominating in the nonconserved ones. These data also confirm the similar behavior of the constitutive and major alternative exons.[24] Thus, all types of nonconserved exons seem to experience positive selection.

The above analysis could be compromised by the fact that the McDonald-Kreitman test relies on the assumption that the polymorphism accounted for is neutral. To eliminate potentially mildly deleterious SNPs that might have accumulated in the human population, e.g., as a result of rapid population expansion, we have repeated the test with SNPs for which the frequency of the derived (new) allele is known and not less than 5% (Table 1D and Table S1D) or 10% (Table S1E). With these presumably neutral SNPs, for minor alternatives the α fractions of substitutions fixed by positive selection are then estimated as 0.27 and 0.29, respectively, with Fisher's test value ~0.01. This moderate increase conforms with the expectation that weakly relaxed selection on mildly deleterious minor-alternative SNPs would inflate the Pa number and therefore decrease the α value, thus making our basic test with all SNPs rather a conservative estimate. Positive selection also appears in the major regions (α = 0.19 and 0.22 for the 5% and 10% SNPs, respectively), suggesting that in the total sample weaker positive selection in the major exons is masked by the presence of nonneutral SNPs.

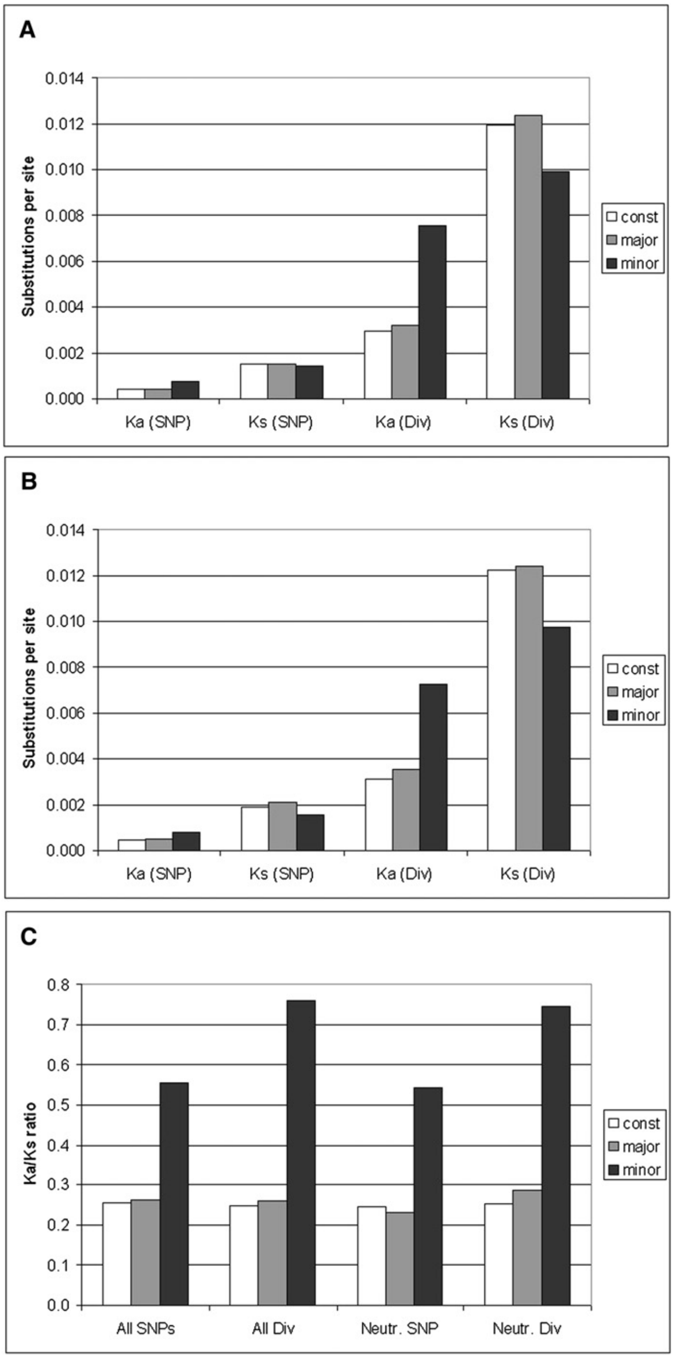

Figure 1. Substitution Rates in Constitutive, Major and Minor Alternatively Spliced Exons of Human Genes

(A) Nonsynonymous (*K*a) and synonymous (*K*s) substitution rates for SNPs (SNP) and human-chimpanzee divergence substitutions (Div). The *K*a values are lower and the *K*s values are higher in the minor exons.

(B) Same as (A), but only neutral SNPs (with frequency > 5%) are considered.

(C) The *K*a/*K*s ratios in the minor exons are consistently higher for SNPs and divergence in the total sample and when only neutral SNPs are considered.

The synonymous divergence and polymorphism rates are slightly lower in minor AS exons (Table S1A). This observation can be explained by RNA-level selection acting on synonymous sites of alternatively spliced exons in order to maintain the splicing-regulation motifs, ESEs (exonic splicing enhancers) in particular.[8,25–28] In the case of presumably neutral polymorphisms (Table S1D), the reduction of *K*s values between constitutive and minor exons is close for polymorphism and divergence: 22% and 25%, respectively. The *P*a/*P*s and *D*a/*D*s ratios used in the McDonald-Kreitman test are therefore almost equally affected by this fact, and the observed positive selection is indeed driven by nonsynonymous variation. However, when all SNPs are considered (Table S1A), the *K*s value for polymorphisms in minor exons is only 7% lower than that for constitutive exons, suggesting that a fraction of synonymous SNPs might be mildly deleterious for the former.

The excess of nonsynonymous SNPs and fixed substitutions in the minor alternatively spliced fragments of human genes suggests that they are under lower purifying-selection pressure. As shown by the results of the McDonald-Kreitman test, they also experience positive selection, observed both in conserved and nonconserved exons. With some sacrifice of statistical significance, this result is also confirmed by two more-conservative subsets with high-frequency SNPs, providing the strict case of presumably neutral variation. These tests, along with those restricted to SNPs from two personal genomes providing uniform genome coverage, seem to be a reasonable compromise in the present-day absence of an "ideal" perfectly ascertained set of SNPs that is sufficiently large to represent minor protein isoforms. One should be aware, however, that the nonsystematic nature of the data may be a potential source of artifactual effects in polymoprhism-related studies.

Positive selection could not be observed in earlier studies relying on comparisons of *K*a/*K*s values in the alternative and constitutive regions of pairwise-aligned mammalian genes,[9] because the latter technique does not allow one to distinguish between positive selection and relaxed negative selection. The advantage of the McDonald-Kreitman test is that it detects the fraction of between-species amino acid diversity that cannot be explained by neutral variation.[23] In our case, the increase of nonsynonymous diversity in minor isoforms is greater than that expected due to relaxation of negative selection estimated from the polymorphism levels and can naturally be interpreted as a trace of positive selection.

As more mammalian genomes are being sequenced, an opportunity to use the methods for estimating positive selection based on representative multiple alignments will appear as a sufficient number of reliable genomic sequences become available.[29]

Unexpectedly, the fraction of amino acid substitutions fixed by positive selection in the conserved minor alternative exons is higher than that in nonconserved exons

Table 1. SNP and Divergence Density Values and the McDonald-Kreitman Test

Exon Type	Number of Exons	*Ka*/*Ks* (SNP)	*Ka*/*Ks* (Div)	α	Fisher Test Significance
All Exons[a]					
Constitutive	52151	0.26	0.25	−0.04	0.118
Major	9799	0.26	0.26	−0.02	0.442
Minor	4397	0.55	0.76	0.27	0.001
Conserved Exons[b]					
Constitutive	42590	0.25	0.23	−0.09	0.008
Major	7218	0.26	0.23	−0.16	0.069
Minor	2023	0.40	0.62	0.35	0.009
Nonconserved Exons[c]					
Constitutive	9561	0.29	0.32	0.09	0.056
Major	2581	0.27	0.34	0.21	0.052
Minor	2374	0.66	0.85	0.23	0.026
All Exons, SNPs with Derived Allele Frequency ≥ 5%[d]					
Constitutive	30060	0.25	0.25	0.04	0.149
Major	4801	0.23	0.29	0.19	0.015
Minor	2026	0.54	0.74	0.27	0.012

The table contains the ratio values *Ka*/*Ks* of nonsynonymous substitutions per nonsynonymous site (*Ka*) to synonymous substitutions per synonymous site (*Ks*) for three types of gene regions. *Ka* and *Ks* denote both polymorphism densities and divergence, with discriminating labels "SNP" and "Div," respectively. The last column shows the significance computed by the Fisher test applied to the four numbers organized in a 2 × 2 contingency table.

[a] All exons.

[b] Conserved exons.

[c] Nonconserved exons from genes covered by ≥ 60 ESTs.

[d] Genes with at least one SNP with known frequency, only neutral SNPs (both validated and nonvalidated) with the frequency of the derived (new) allele ≥ 5%.

($\alpha = 0.35$ versus $\alpha = 0.23$). This fact could be explained by contamination of the minor-isoform sample by aberrantly spliced, nonfunctional exons, blurring the evidence for selection. Indeed, nonconserved minor alternative exons are always suspicious,[30] and thus this result should be re-examined as more data become available. It should be noted, however, that the fact that positive selection is observed in both conserved and nonconserved minor exons provides independent support for the main finding of this study. This is consistent with a recent theory, according to which new exons emerge by fixation of aberrant splicing events, with subsequent upregulation of functionally useful variants. Indeed, because minor alternative exons, unlike constitutive and major alternative ones, seem to be relatively young,[11,24,29] they are a natural substrate for positive selection.

Supplemental Data

One figure and one table are available with this paper online at http://www.ajhg.org/.

Acknowledgments

We are grateful to Fyodor Kondrashov, Shamil Sunyaev, and Ekaterina Ermakova for valuable discussions and to Alexander Favorov for a subroutine for calculation of the Fisher test. This study was partially supported by grants from the Howard Hughes Medical Institute (55001056) and the Russian Fund of Basic Research (07-04-00343), as well as by the Russian Academy of Sciences programs "Cellular and Molecular Biology" (M.G. and V.R.) and "Human molecular polymorphism" (V.R.). No conflict of interest is associated with the manuscript.

Received: January 15, 2008
Revised: April 8, 2008
Accepted: May 30, 2008
Published online: June 19, 2008

Web Resources

The SNP dataset from James Watson's Personal Genome Sequence was downloaded from ftp://jimwatsonsequence.cshl.edu

References

1. Graveley, B.R. (2001). Alternative splicing: increasing diversity in the proteomic world. Trends Genet. *17*, 100–107.
2. Lareau, L.F., Green, R.E., Bhatnagar, R.S., and Brenner, S.E. (2004). The evolving roles of alternative splicing. Curr. Opin. Struct. Biol. *14*, 273–282.
3. Mironov, A.A., Fickett, J.W., and Gelfand, M.S. (1999). Frequent alternative splicing of human genes. Genome Res. *9*, 1288–1293.
4. Johnson, J.M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P.M., Armour, C.D., Santos, R., Schadt, E.E., Stoughton, R., and Shoemaker, D.D. (2003). Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science *302*, 2141–2144.
5. Kampa, D., Cheng, J., Kapranov, P., Yamanaka, M., Brubaker, S., Cawley, S., Drenkow, J., Piccolboni, A., Bekiranov, S., Helt, G., et al. (2004). Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. Genome Res. *14*, 331–342.
6. 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.
7. Xing, Y., and Lee, C. (2006). Alternative splicing and RNA selection pressure - evolutionary consequences for eukaryotic genomes. Nat. Rev. Genet. *7*, 499–509.
8. Xing, Y., and Lee, C. (2005). Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences. Proc. Natl. Acad. Sci. USA *102*, 13526–13531.
9. Ermakova, E.O., Nurtdinov, R.N., and Gelfand, M.S. (2006). Fast rate of evolution in alternatively spliced coding regions of mammalian genes. BMC Genomics *7*, 84.
10. Cusack, B.P., and Wolfe, K.H. (2005). Changes in alternative splicing of human and mouse genes are accompanied by faster evolution of constitutive exons. Mol. Biol. Evol. *22*, 2198–2208.

11. Zhang, X.H., and Chasin, L.A. (2006). Comparison of multiple vertebrate genomes reveals the birth and evolution of human exons. Proc. Natl. Acad. Sci. USA *103*, 13427–13432.
12. Sabeti, P.C., Schaffner, S.F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T.S., Altshuler, D., and Lander, E.S. (2006). Positive natural selection in the human lineage. Science *312*, 1614–1620.
13. Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E.H., McCarroll, S.A., Gaudet, R., et al. (2007). Genome-wide detection and characterization of positive selection in human populations. Nature *449*, 913–918.
14. McDonald, J.H., and Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. Nature *351*, 652–654.
15. Fay, J.C., Wyckoff, G.J., and Wu, C.I. (2002). Testing the neutral theory of molecular evolution with genomic data from Drosophila. Nature *415*, 1024–1026.
16. Bustamante, C.D., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M.T., Glanowski, S., Tanenbaum, D.M., White, T.J., Sninsky, J.J., Hernandez, R.D., et al. (2005). Natural selection on protein-coding genes in the human genome. Nature *437*, 1153–1157.
17. Nurtdinov, R.N., Neverov, A.D., Mal’ko, D.B., Kosmodem’ianskii, I.A., Ermakova, E.O., Ramenskii, V.E., Mironov, A.A., and Gel’fand, M.S. (2006). Biofizika *51*, 589–592.
18. Neverov, A.D., Artamonova, I.I., Nurtdinov, R.N., Frishman, D., Gelfand, M.S., and Mironov, A.A. (2005). Alternative splicing and protein function. BMC Bioinformatics *7*, 266.
19. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. *29*, 308–311.
20. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res. *12*, 996–1006.
21. Kondrashov, F.A., Ogurtsov, A.Y., and Kondrashov, A.S. (2006). Selection in favor of nucleotides G and C diversifies evolution rates and levels of polymorphism at mammalian synonymous sites. J. Theor. Biol. *240*, 616–626.
22. Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.J., Makhijani, V., Roth, G.T., et al. (2008). The complete genome of an individual by massively parallel DNA sequencing. Nature *452*, 872–876.
23. Smith, N.G., and Eyre-Walker, A. (2002). Adaptive protein evolution in Drosophila. Nature *415*, 1022–1024.
24. Lev-Maor, G., Goren, A., Sela, N., Kim, E., Keren, H., Doron-Faigenboim, A., Leibman-Barak, S., Pupko, T., and Ast, G. (2007). The “alternative” choice of constitutive exons throughout evolution. PLoS Genet. *3*, e203.
25. 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.
26. 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.
27. Orban, T.I., and Olah, E. (2001). Purifying selection on silent sites–a constraint from splicing regulation? Trends Genet. *5*, 252–253.
28. 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.
29. Alekseyenko, A.V., Kim, N., and Lee, C.J. (2007). Global analysis of exon creation versus loss and the role of alternative splicing in 17 vertebrate genomes. RNA *13*, 661–670.
30. Sorek, R., Shamir, R., and Ast, G. (2004). How prevalent is functional alternative splicing in the human genome? Trends Genet. *20*, 68–71.