J Mol Evol
DOI 10.1007/s00239-013-9555-2

ORIGINAL ARTICLE

Testing for Natural Selection in Human Exonic Splicing Regulators Associated with Evolutionary Rate Shifts

Rodrigo F. Ramalho · Sahar Gelfman · Jorge E. de Souza · Gil Ast · Sandro J. de Souza · Diogo Meyer

Received: 13 August 2012 / Accepted: 9 March 2013

Abstract Despite evidence that at the interspecific scale, exonic splicing silencers (ESSs) are under negative selection in constitutive exons, little is known about the effects of slightly deleterious polymorphisms on these splicing regulators. Through the application of a modified version of the McDonald–Kreitman test, we compared the normalized proportions of human polymorphisms and human/rhesus substitutions affecting exonic splicing regulators (ESRs) on sequences of constitutive and alternative exons. Our results show a depletion of substitutions and an enrichment of SNPs associated with ESS gain in constitutive exons. Moreover, we show that this evolutionary pattern is also present in a set of ESRs previously involved in the transition from constitutive to skipped exons in the mammalian lineage. The similarity between these two sets of ESRs suggests that the transition from constitutive to skipped exons in mammals is more frequently associated with the inhibition than with the promotion of splicing signals. This is in accordance with the hypothesis of a constitutive origin of exon skipping and corroborates previous findings about the antagonistic role of certain exonic splicing enhancers.

Keywords Alternative splicing · Human polymorphism · Exonic splicing regulators · MK test

Introduction

Splicing is the process by which introns are removed from an mRNA precursor and exons are ligated to form a mature mRNA. During this process, several *cis* and *trans* factors are involved. Besides the canonical *cis* factors (e.g., splicing sites, branch point and polypyrimidine tract), splicing regulators—short sequences located in exons and introns—have an important role in assisting the spliceosome to correctly recognize exon/intron boundaries. Exonic splicing regulators (ESRs) can be divided in two groups, according to their function in splicing: exonic splicing enhancers (ESEs) and exonic splicing silencers (ESSs) which promote and inhibit the inclusion of exons in mRNA, respectively.

Through alternative splicing, great transcript diversity is generated, and currently more than 80 % of human genes are known to present splicing variants (Wang et al. 2008). There are three main hypotheses for the origin of alternative exons: (a) "exonization" of Alu elements (Sorek et al. 2002), (b) gene/exon duplication (Letunic et al. 2002), and (c) weakening of purifying selection in the splicing *cis*-elements of constitutive exons (Lev-Maor et al. 2007). Among these, only the last one is able to explain the high

Electronic supplementary material The online version of this article (doi:10.1007/s00239-013-9555-2) contains supplementary material, which is available to authorized users.

R. F. Ramalho (✉) · D. Meyer
Departamento de Genética e Biologia Evolutiva, Instituto de Biociências, Universidade de São Paulo, São Paulo, SP 05508-900, Brazil
e-mail: rfrusp@gmail.com

D. Meyer
e-mail: diogo@ib.usp.br

R. F. Ramalho · J. E. de Souza · S. J. de Souza
Instituto de Bioinformática e Biotecnologia, Ribeirão Preto, SP, Brazil

S. Gelfman · G. Ast
Department of Human Molecular Genetics & Biochemistry, Sackler Faculty of Medicine, Tel-Aviv University, 69978 Ramat Aviv, Israel

S. J. de Souza
Instituto do Cérebro, Universidade Federal do Rio Grande do Norte (UFRN), Natal, RN, Brazil

Published online: 26 March 2013

frequency of alternative exons observed in the genomes of various eukaryotes, especially in mammals (Kim et al. 2007), while the other two can only explain a minority of alternative exons in the human genome because less than 5 % of these contain Alus (Sorek et al. 2002) and less than 20 % are associated with duplicated exons (Letunic et al. 2002).

The evolutionary mechanism underlying the hypothesis of a constitutive origin of alternative exons was inferred combining two sources of information: differences in the splicing pattern of orthologous genes in mammals and non-mammals, and the phylogenetic position of modification in *cis*-splicing signals (Lev-Maor et al. 2007). These authors observed that the transition from constitutive splicing in non-mammals to exon skipping in mammals is based on the weakening of splice site (ss) strength, associated with the fixation of certain ESR motifs. Based on this result, a model was proposed, predicting that the weakening of the 5′ ss is the main driving force in the transition from constitutive to alternative splicing and that following this weakening, there is a gain of ESR functionality to properly regulate the exon inclusion level (Lev-Maor et al. 2007). Although this hypothesis attributes to the fixation of ESRs an important functional role in the switching of splicing pattern, it does not clearly define the major function of these ESRs. Do they mainly act as splicing silencers, inhibiting the exon inclusion, or as splicing enhancers, promoting exon inclusion?

Despite this uncertainty, there is a strong pattern that gives a clue to the answer: Several studies suggest that depletion in the density of ESSs is characteristic of constitutive exons (Ke et al. 2008; Wang et al. 2004; Xiao et al. 2007; Zhang et al. 2009). Wang et al. (2004) showed that ESSs are significantly depleted in constitutive exons relative to introns and also depleted in constitutive exons with a weaker ss relative to constitutive exons with a strong ss. Similar conclusions were reported by Xiao et al. (2007).

A second line of evidence, supporting the importance of ESS density in defining whether an exon will be constitutive or alternative, was provided by the analysis of duplicated genes (Zhang et al. 2009). Using pairs of paralogous exons that exhibit a constitutive and an alternative copy, these authors showed that the density of ESSs is lower in the constitutive copy than in the alternative one.

Further evidence of purifying selection on ESSs in constitutive exons was described by Ke et al. (2008). Using synonymous divergence between a human and macaque, the authors showed that ESS creation occurs at a much lower rate in constitutive exons than in introns. Additionally, the rate of ESS creation in constitutive exons was lower than the rate of creation of non-ESS motifs for this same set of exons. Overall, these lines of evidence suggest that a gain in ESSs could explain the transition from constitutive to alternative splicing.

Assuming that the presence of ESRs in constitutive exons is essential for promoting exon skipping, the above evidence suggests that the ESRs fixed in skipped exons after the split between mammals and non-mammals should act as silencers, down-regulating exon inclusion level.

In the present study, we investigated the evolution of hexamers (a sequence of 6 nt in length) that are putatively involved with the constitutive to alternative transition in vertebrate evolution (Gelfman et al. 2012; Lev-Maor et al. 2007). Our approach differs from other studies because we directly contrasted the evolutionary rate of these splicing regulators at the populational and interspecies levels. The comparison of SNP data with interspecies divergence provides an opportunity to investigate how evolution occurs in a relatively recent time scale (within human population).

The nearly neutral model of molecular evolution (Ohta 1973) predicts that slightly deleterious mutations are more prone to accumulate within populations, while over long spans of time, they are removed by natural selection with high probability. Therefore, the comparison of DNA polymorphisms with substitutions allows us to test the hypothesis that skipped exons are accumulating slightly deleterious mutations and thus weakening the *cis*-splicing signals of constitutive exons. Specifically, we test if the hexamers identified in human-skipped exons and fixed in mammals present a significant excess of polymorphism relative to substitutions when located in human constitutive exons. Assuming that the mammalian-fixed hexamers in skipped exons act as silencers, we expect to detect a stronger signal of purifying selection on mutations that create these ESRs in constitutive exons. Such a pattern would indicate that genetic changes create favorable conditions for the transition from constitutive to alternative splicing.

By using a modified version of the McDonald–Kreitman test (MK test, 1991), we find that the hexamers associated with the origin of exon skipping in mammals (called rate-shifted herein) present a depletion of substitutions and an excess of polymorphisms when located in constitutive exons. This evolutionary pattern, typical of loci under purifying selection, reaches a higher statistical significance in constitutive exons with a weak ss.

Based on a similar pattern found here compared with that observed for previously known ESSs, we suggest that the rate-shifted hexamers identified in skipped exons might have inhibitory activity on exon inclusion, corroborating previous results about the context-dependent function of ESRs (Goren et al. 2006; Jumaa and Nielsen 1997; Kanopka et al. 1996; Solis et al. 2008; Ule et al. 2006).

Materials and Methods

Primary Exon Database

For all the data mining described below, we started the procedures using an in-house exon database consisting of 60,383 internal alternative exons and 70,801 internal constitutive exons (de Souza et al. 2011; Galante et al. 2004, 2007). From all alternative exons, 34,669 skipped exons were used for further analyses. Briefly, the definition of exonic and intronic regions was based on the genomic coordinates of cDNA sequences classified as "mRNA" in GenBank. All the analyzed genes have at least one RefSeq identification code. To increase the reliability of the splicing events identified, we chose for further analysis only those events that were supported by at least two ESTs from two distinct libraries using information from the Evoc database (http://www.evocontology.org/) (see Supplemental material).

Polymorphism Dataset

We obtained the SNP data from the low coverage pilot phase of the 1,000 genomes project (Consortium 2010). We chose the African sample since it has more SNPs than other samples, increasing the power of our statistical tests. The SNP data were downloaded from ftp://ftptrace.ncbi.nih.gov/1000genomes/ftp/pilot_data/release/2010_07/low_coverage/snps/. We used ANNOVAR (Wang et al. 2010) to annotate all SNPs and filter the synonymous ones for further analysis. In total, 37,080 synonymous SNPs were identified. Also, we identified 2,277 internal skipped exons and 10,232 internal constitutive exons with at least one synonymous SNP from the 1,000 genomes project data.

Substitution Dataset

We downloaded the genomic alignments of human and rhesus (axtNet format) from the UCSC genome browser (http://hgdownload.cse.ucsc.edu/). These files can be obtained at http://hgdownload.cse.ucsc.edu/goldenPath/hg18/vsRheMac2/.

Using a local Perl script, we parsed this file to find the chromosome, the genomic coordinate, and the nucleotide for each divergent site between the two species. Approximately 180 million substitutions were found for the whole genomes and approximately 650,000 for the human exome.

We used the ANNOVAR software to annotate the substitutions, entering the divergent nucleotides found at each site as alleles from a biallelic SNP.

We identified 1,581 orthologous internal skipped exons and 7,575 orthologous internal constitutive exons containing at least one synonymous substitution (between human and rhesus) and one synonymous SNP.

Identification of Rate-Shifted ESRs

We applied the strategy described in Gelfman et al. (2012) upon 28,970 skipped exons to identify functionally important hexamers for splicing regulation. The same strategy was applied to 67086 constitutive exons, providing a comparsion set. This strategy is based on the two steps described below.

Detection of Site-Specific Deceleration of Evolutionary Rate

For the detection of evolutionary rate shifts in the exonic sequences, we used a phylogenetic approach implemented in the Covarion program (Pupko and Galtier 2002). The Covarion program is based on maximum-likelihood site-specific evolutionary rate estimates at every site of a given multiple sequence alignment. Given a specific branch point, the program estimates for each site the rate for one sub-tree and the rate for the other sub-tree, both extracted from a phylogenetic tree consisting of twelve vertebrate species assuming the Jukes and Cantor substitution model. For a given site, statistically different rates for the two sub-trees reveal a probable change in functional constraints. The significance of the difference between the two groups was assessed using a likelihood-ratio test. This approach allowed the detection of sites that had highly different rates in two sub-groups, consistent with a change in the selection pressure of the site.

We only identified positions that harbor rate shifts to conservation. These are the positions where the rate of the sub-tree containing human was significantly lower than the rate of the sub-tree of the non-mammalian species and correspond to the cases where the position became conserved in mammals. Next, each sequence was given a rate-shift score normalized to sequence length. The score was calculated as follows:

$$\text{score} = \text{RS} \times (100/L),$$

where RS is the sum of rate shifts in the sequence (rate shifts to conservation) and L is the sequence length. The final score was calculated as the average rate-shift abundance in exons (normalized per length) and was repeated for the entire alternative or constitutive exons data (exons with no rate shifts received a value of 0).

Applying the Covarion Method to Predict Splicing *Cis*-Regulatory Elements

The prediction of regulatory sequences was done for all possible 4,096 hexamers for a given group of exonic sequences (skipped or constitutive). The number of sites that were shifted to conservation was counted for each hexamer for both kinds of exons. We then compared this number to the expected number assuming rate-shifted sites are randomly distributed. p values were calculated using Fisher's exact test and corrected for multiple testing (FDR, $\alpha < 0.01$). Hexamers with p values lower than 0.05 were considered functionally important. All the hexamers identified by this phylogenetic approach (from constitutive and skipped exons) will be called rate-shifted motifs herein and are available as an Online Resource. Note that a motif identified as a rate-shifted ESR in one exon category does not necessarily means its complete absence in sequences of the other exon category, only that it has not attained the statistical status of being a rate-shifted motif in this latter exon category.

The resulting set of hexamers identified from skipped exon sequences (rate-shifted$_{\text{skipped}}$) was tested for signals of natural selection. Hexamers identified from constitutive exon sequences were analyzed in the same way, providing a control.

ESR Datasets

We used a Perl script to search for exact matches between the rate-shifted hexamers, six ESE sets (SF2_IgM, SRP40, SRP55, SC35, RESCUE, and PESE), and one ESS set (FAS-hex2). The SF2_IgM, SRP40, SRP55, and SC35 ESE sets were discovered in vitro by the SELEX procedure (Liu et al. 1998, 2000) and each contains binding sites for four distinct SRPs. Regarding the SELEX-ESEs, only those oligomers with a score equal to or higher than the threshold scores defined by the original study were considered as ESEs. The other two ESE sets (RESCUE and PESE) were predicted in silico. Motifs that make up the RESCUE set were computationally identified based on their enrichment in a set of constitutive exons with a weak ss relative to constitutive exons with a strong ss and also to introns (Fairbrother et al. 2002). The PESE set contains motifs enriched in constitutive exons relative to pseudoexons and 5′ UTR of intronless genes (Zhang and Chasin 2004). For ESSs, we used the list of 176 FAS-hex2 hexamers with a silencer function (Wang et al. 2004), discovered with assays based on splicing in cultivated cells with a fluorescent system that reports the silencer role of random motifs.

Ss Scoring

We used the MaxEntScan web server—http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq_acc.html—as described by Yeo and Burge (2004) to estimate the 5′- and 3′-ss strengths for all constitutive exons containing substitutions and polymorphisms. We then used the median of the ss score distribution (ss score median = 8.5) as a threshold to define the categories of exons with weak and strong ss.

The 3′ ss was defined as the region including the 20 intronic nucleotides upstream of the exon and the first three exonic nucleotides, whereas the 5′ ss was defined as the region including the three terminal exonic nucleotides and the first six downstream intronic nucleotides.

Mapping SNPs and Divergent Sites onto ESR Motifs

For each of the polymorphic or divergent sites, we created sequence tags by extracting the site plus ten flanking nucleotides at either side (human genome assembly hg18). The tags were extracted with respect to the same strand orientation of the RefSeq gene which contains the SNP. Next, each tag was tested for its status as containing ESRs by aligning it with all known ESRs from each set.

Defining the Ancestral and Derived Alleles of SNPs

In order to distinguish the SNPs that create or disrupt rate-shifted hexamers, we compared the SNP alleles to the orthologous sites of the chimpanzee (*Pan troglodytes*) and rhesus macaque (*M. mulatta*) genomes. To perform this, we used a Perl script that reads the genomic alignment between human and rhesus to retrieve the orthologous nucleotides in the rhesus genome for each SNP coordinate. The SNP data which we use (provided by the 1,000 genome project) already contain the annotation of ancestral allele, relative to the chimpanzee genome. Only SNPs with ancestral alleles identical to orthologous positions of rhesus and chimpanzee genomes were further analyzed. In this way, ancestral and derived tags (around the SNPs) were defined, and the polarity of the change (i.e., creation or disruption of rate-shifted hexamers) was determined.

Modified MK Test

The usual implementation of the MK test uses data from SNPs and substitutions to compare, through a 2 × 2 contingency table, the proportion of variants from two distinct functional categories (for example, synonymous and non-synonymous changes) located in a given gene. One of these categories is assumed to be evolving neutrally, allowing a formal test of whether the ratio between functional/neutral variant sites differs between the intraspecific (polymorphism) and interspecific levels (substitutions). In our modified version of this test, we compare mutations that create ESRs (our supposedly functional class) with those that do not alter the ESR status (which are putatively neutral). We avoid the confounding effects of selection at the amino acid level by only using synonymous variants.

The null hypothesis of homogeneity for contingency tables was tested using the χ^2 distribution, implemented in the function "chisq.test" of R statistical software (http://www.r-project.org/). For each cell of a contingency table, we calculated the normalized deviation to the expected proportions as (observed − expected)/(expected). The expected values are calculated by assuming independence among categories. The normalized deviation for the cells that contain the counts of possibly deleterious changes (i.e., ESR changing mutations) was used to graphically summarize the results of the MK test. For these graphs, the *Y*-axis was called "slightly deleterious deviation from neutral expectation."

Figure 1 presents a schema of the SNPs and substitutions' data flow until their final application on modified MK tests.

Comparing Different Estimates of the Neutrality Index (NI)

The result of the MK test is commonly summarized by the NI which is simply the odds ratio of the 2 × 2 table and is calculated as NI $= (P_c/P_m)/(D_c/D_m)$, where *P* and *D* refer to the number of sites which are polymorphic and divergent, respectively, and the subscripts c and m refer to variations that change or maintain the ESR status. Values greater than 1 suggest that ESR-altering mutations are slightly deleterious and under negative selection, while NI less than 1 suggests positive selection for ESR change. Our modified version of the MK test was applied using two distinct approaches. Initially, we simply counted the total number of functional or neutral synonymous SNPs and substitutions found within a given set of human exons, without distinguishing if they are present in the same gene or not. Next, in order to check if this approach could lead to a bias, we applied a gene-based approach, counting the neutral or functional SNPs and substitutions for each gene. By using the gene-based approach, we could estimate the odds ratio (NI) for a contingency table created with the sum of SNPs and substitutions from all genes (called $\mathrm{NI}_{\mathrm{pool}}$) and also apply two statistical methods (Tarone–Greenland and Mantel–Haenszel) implemented by the software DOFE (http://www.lifesci.sussex.ac.uk/home/Adam_Eyre-Walker/Website/Software.html) to infer the NI and the confidence interval (CI) for the set of analyzed genes.

Table 1 Density of rate-shifted ESR sets on human exonic sequences

	Rate-shifted$^{a}_{\mathrm{constitutive}}$ (198)	Rate-shifted$_{\mathrm{skipped}}$ (145)
Constitutive (10232 exons)	0.05063	0.04286
Skipping (2277 exons)	0.05077	0.04348
Exon pool (12509 exons)	0.05067	0.04294

[a] Corrected by size differences among ESR sets. Median of 1,000 sets with 145 motifs each

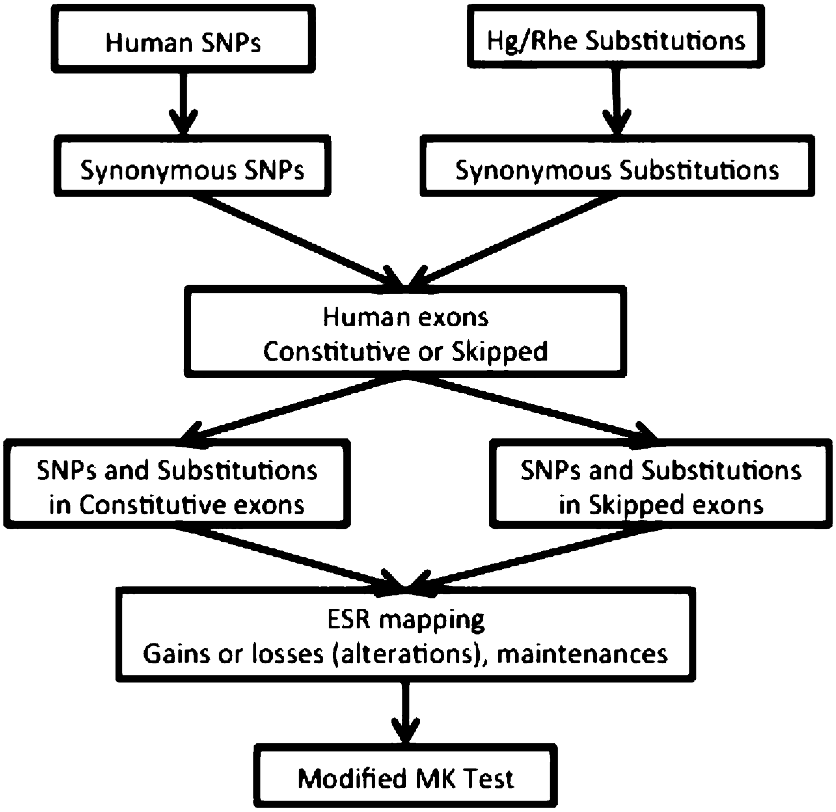

Fig. 1 Schema of the main steps used to filter variant sites further tested with modified version of MK test

Table 2 Comparison between the density of rate-shifted$_{skipped}$ ESRs and several public ESR sets on human exonic sequences

	RESCUE[a] (238)	PESE[a] (2060)	SC35[a] (2585)	SF2_IgM[a] (924)	SRP40[a] (671)	ESS[a] (176)
Exon pool (12509 exons)	0.05738	0.00313	0	0.00793	0.00840	0.02137

[a] Corrected by size differences among ESR sets. Median of 1,000 sets with 145 motifs each

Gene Expression Rank

To create the gene expression rank, we downloaded RNA-SEQ data from 64 experiments derived from tissues samples and cell lines and available at SRA (http://www.ncbi.nlm.nih.gov/sra). In total, 15 distinct human tissues were represented. All these data were aligned against the human genome using the software Bowtie. We obtain the RPKM using a Perl script which is based on the following formula RPKM = $((10^9) * \text{nreads})/(\text{treads} * \text{gene\_len})$, where nreads represents the number of reads per gene, treads represents the total number of reads in the cDNA library, and gene_len represents the gene length.

For each gene, we calculated the mean RPKM and then we obtained the quartiles from this distribution to create four categories of gene expression level.

Table 3 Comparison between the density of rate-shifted$_{skipped}$ ESRs and SRP55 ESR set on human exonic sequences

	Rate-shifted$_{skipped}$[a]	SRP55 (133)
Exon pool (12509 exons)	0.03947	0.0253

[a] Corrected by size differences among ESR sets. Median of 1,000 sets with 133 motifs each

Results and Discussion

The rate-shift analysis applied to sequences of human-skipped exons revealed 145 hexamers (called rate-shifted$_{skipped}$,

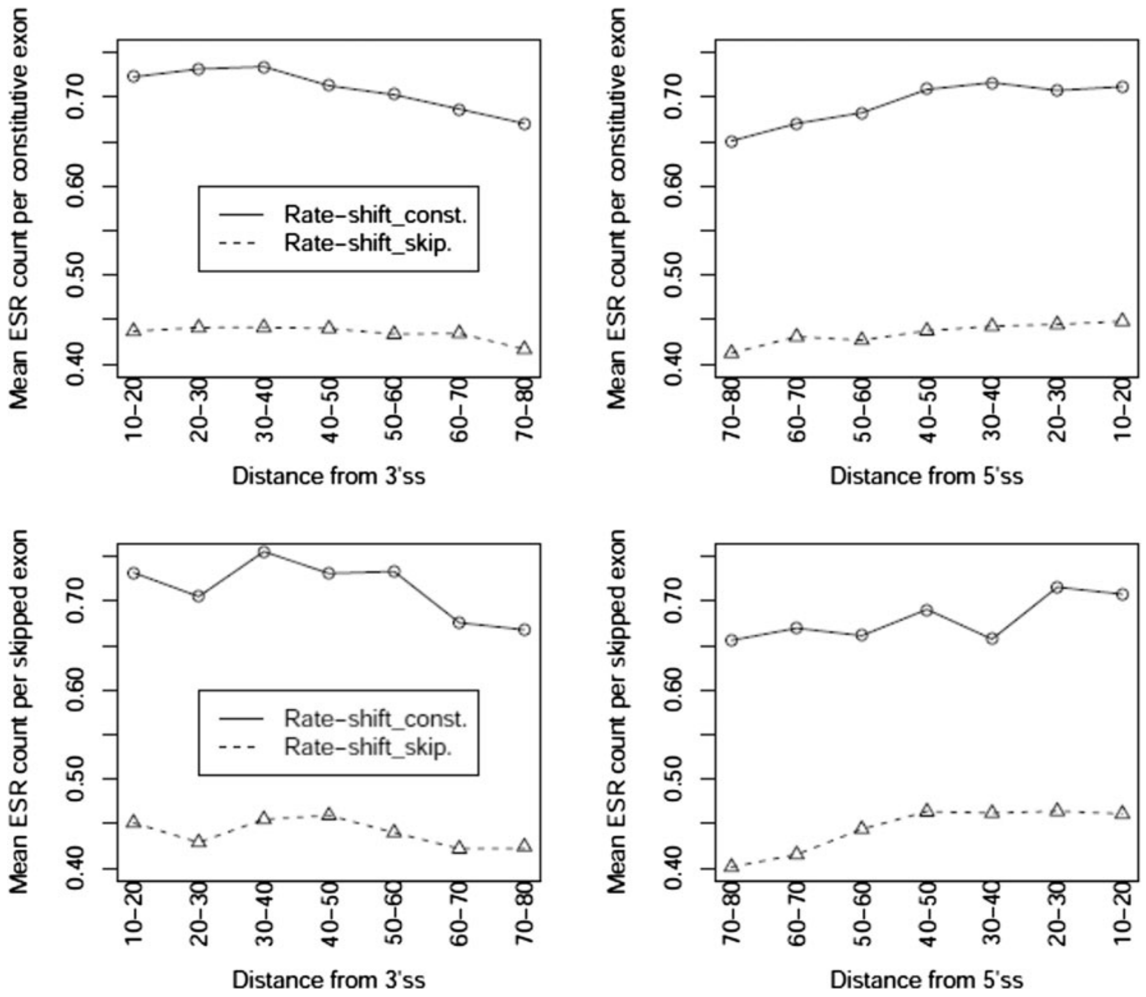

Fig. 2 Occupation profiles of rate-shifted ESRs in human constitutive (*upper panel*) and skipped (*bottom*) exons

herein). Although our method is designed to identify motifs under long-term purifying selection, several different classes of functional elements could contribute to the identified motifs. Therefore, our first analysis was the alignment of these rate-shifted hexamers with known ESRs (Supplementary Table S1). A great majority (116 hexamers; 80 %) were found in putative ESR groups and only nine (6 %) were identical to known ESSs. This result suggests that the rate-shifted$_{\text{skipped}}$ motifs represent potential ESRs. A caveat of this approach, however, is the high number of public motifs described as ESRs.

To further explore this hypothesis, we compared the density of rate-shifted hexamers in human exons with the major public ESR sets (Tables 1, 2, 3) and analyzed the occupation profile of the rate-shifted ESRs relative to ss (Fig. 2).

We observed that rate-shifted motifs show similar density relative to other public ESR hexamers (RESCUE and SRP55, Tables 2, 3 respectively). Only few differences in density were observed for rate-shifted ESRs either in constitutive or skipped exons, even though rate-shifted$_{\text{constitutive}}$ ones were slightly denser in human exons than rate-shifted$_{\text{skipped}}$ motifs (Table 1). The density of the ESR sets composed by hexamers (RESCUE, SRP55 and rate-shifted) is strikingly higher in human exon sequences than other ESR set composed by heptamers and octamers (including SELEX-ESEs and PESEs). This could be explained by the higher number of all possible heptamers ($4^7 = 16{,}384$) and octamers ($4^8 = 65{,}536$) relative to hexamers ($4^6 = 4{,}096$). Next, we analyzed the occupation profile of the rate-shifted motifs (Fig. 2) and found that their distribution is biased toward exon borders. This pattern is in accordance with the results of Gelfman et al. (2012). Moreover, the occupation profile of rate-shifted ESRs resembles the pattern of RESCUE–ESEs, an ESR set with several reports of negative selection in the literature (Carlini and Genut 2006; Fairbrother et al. 2004).

When applied to constitutive exons, the rate-shift analysis identified 198 significant motifs (called rate-shifted$_{\text{constitutive}}$). Most of these hexamers (173 hexamers; 87 %) were fully aligned with known ESEs and four (2 %) were identical to ESSs. Interestingly, only 14 rate-shifted hexamers were identical between both sets of rate-shifted hexamers (constitutive vs. skipped exons).

We observed that the rate-shifted$_{\text{skipped}}$ hexamers have a higher proportion of unknown motifs (those absent in all sets considered herein) than the rate-shifted$_{\text{constitutive}}$ hexamers (29 (∼20 %) and 25 (∼12 %) unknown hexamers, respectively). Moreover, the proportion of known ESSs among the rate-shifted$_{\text{skipped}}$ hexamers was higher (∼6 %) than for rate-shifted$_{\text{constitutive}}$ (∼2 %) (χ^2, p value ∼ 0.04 for both tests; 20 vs. 12 % and 6 vs. 2 %; one-tailed test). These differences suggest that for skipped exons, the evolutionary rate shifts affect more often ESRs with inhibitory rather than enhancing activity.

Table 4 Modified MK test applied for ESSs located within (a) constitutive exons and (b) skipped exons

	Create ESS	Maintain ESS
(a) Constitutive exons		
Substitution	2421 (−1.2 %)	925 (+3.2 %)
Polymorphism	959 (+3.1 %)	311 (−8.4 %)
NI = 1.18		
χ^2 p value = 0.03		

Genes	NI$_{\text{pool}}$	p value-Pool	NI-Mantel–Haenszel (CI)	NI-Tarone–Greenland (CI)
2051	1.13	0.1	1.14 (0.92–1.42)	1.21 (0.96–1.54)

	Create ESS	Maintain ESS
(b) Skipped exons		
Substitution	437 (−0.8 %)	185 (+1.9 %)
Polymorphism	213 (+1.6 %)	82 (−4 %)
NI = 1.09		
χ^2 p value = 0.60		

Genes	Pool	p value-Pool	NI-Mantel–Haenszel (CI)	NI-Tarone–Greenland (CI)
496	1.10	0.6	0.94 (0.58–1.59)	0.84 (0.49–1.45)

Only synonymous mutations that create or maintain ESSs were used. The respective values for the NI and its CIs are presented below each contingency table

To further explore the tendency of the silencer functionality for the rate-shifted hexamers found in skipped exons, we applied two approaches. First, we checked if our modified version of the MK test could detect signals of natural selection acting on mutations that create or disrupt well-known ESSs. Our expectation was to find purifying selection acting against mutations that create ESSs on constitutive exons. Second, we applied the same test to rate-shifted$_{\text{skipped}}$ motifs and compared the results of these two tests.

Using the modified MK test, we compared a total of 3,501 synonymous substitutions and 1,425 synonymous SNPs that create (probably deleterious) or maintain (probably neutral) ESS motifs in constitutive exons (Table 4a). We found a significant excess of SNPs that create ESSs, relative to ESS creating substitutions (p value <0.05, Fig. 3). Using the gene-based approach, the deviation tendency remained the same as for the pooled contingency table, but the χ^2 test result was only marginally significant (p value = 0.1). This result confirmed previous evidence that ESSs are under purifying selection in constitutive exons (Ke et al. 2008; Wang et al. 2004; Xiao et al. 2007; Zhang et al. 2009).

In contrast, when we applied the MK test to 652 substitutions and 325 SNPs that create ESSs in skipped exons, the proportions of probably deleterious SNPs and substitutions were not significantly different (p value = 0.48, Table 4b). This result is in agreement with a scenario in which ESSs from skipped exons are evolving according to the neutral model.

Consistent with the above-mentioned result—that the gain of silencers is weakly deleterious in constitutive exonic sequences—we observed that the contingency table for ESS disruptions presented an opposite deviation relative to ESS creations, i.e., substitutions are enriched relative to polymorphisms (Supplementary Table S2a). This result suggests that changes that maintain ESSs in constitutive exons are more deleterious than those that disrupt them.

Next, we applied the modified version of the MK test to look for signatures of natural selection on the 145 rate-shifted$_{\text{skipped}}$ hexamers. Note that these putative ESRs are expected to be involved in the regulation of alternative splicing in mammals (Lev-Maor et al. 2007). For comparison purposes, we analyzed these ESRs in two distinct contexts: (a) when located in skipped exons and (b) when located in constitutive exons.

Similar to what was observed for previously described ESSs, the MK test applied to variant sites that cause creation or maintenance of these putative ESRs showed a significant excess of polymorphisms with respect to substitutions in constitutive exons (p value = 0.01, Table 5a; Fig. 4). On the other hand, for skipped exons, the significance is non-significant (p value = 0.56, Table 5b). These results are consistent with the hypothesis that ESRs play a role in the regulation of exon-skipping events. In light of our hypothesis of a constitutive origin for the skipped exons, we believe that these ESRs act mainly as silencers, inhibiting the *cis*-splicing signals. The observation of a signal of natural selection against gains of rate-shifted$_{\text{skipped}}$ ESRs on sequences of constitutive exons corroborates the inhibitory function of these ESRs on exon inclusion.

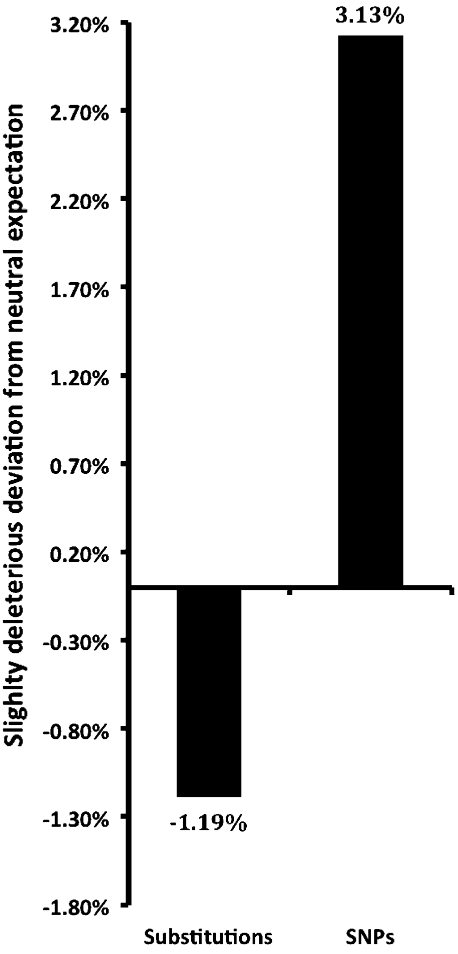

Fig. 3 Normalized difference between the observed and expected counts of substitutions and polymorphisms which create ESS within constitutive exons. *Bars* represent the enrichment (*positive* values) or depletion (*negative* values) of slightly deleterious related to the expectation based on supposedly neutral mutations. $\chi^2 p$ value <0.05

Notably, the deviation from neutral expectation observed for rate-shifted$_{\text{skipped}}$ hexamers located in constitutive exons is very small (Fig. 4). In fact, this observation is not surprising given that we analyzed only synonymous single nucleotide variations (SNPs and substitutions), which evolve in a predominantly neutral manner.

We next investigated if the above signature of purifying selection for constitutive exons reflects differences in nucleotide composition between constitutive and skipped exons and is not necessarily associated with splicing regulation. We checked this by analyzing the set of 198 rate-shifted$_{\text{constitutive}}$ hexamers. Assuming the existence of a nucleotide composition bias, we expect that this set of motifs (identified from constitutive exons) should present a signal of purifying selection when located in skipped exons. However, this was not the case, since the MK tests applied to synonymous variant sites located in this set did not show significant signals of natural selection either in constitutive or skipped exons (Supplementary Table S3). This result suggests that for constitutive exons, the events of evolutionary rate shifts captured by our method are probably caused by selection at the protein level. This could explain the lack of evidence of natural selection using synonymous sites. Therefore, we conclude that our method of motif identification could not, by itself, explain the observed signal of purifying selection.

Next, we examined how gene expression could be interfering with our MK test results regarding selective signatures. To this end, we divided our gene expression rank (see "Materials and Methods" section) in quartiles and then counted for each MK table presented in the text the number of genes inside each quartile of gene expression distribution. The results are presented in Supplementary Fig. 1. This figure shows the frequency of analyzed genes inside each one of the four categories of gene expression (1st, 2nd, 3rd, and 4th quartiles of RPKM distribution). The results show that the frequency of the gene in each quartile is very similar between all the MK tables presented in the manuscript. Therefore, our conclusion is that the observed differences between the results of the MK tests could not be attributed to differences in the expression levels of genes used in each test, given that low, medium, and high expressed genes are present in similar proportions in all presented MK tests.

We next examined if the fact of a significant p value for constitutive exons sequences (and not for skipped) was a consequence of a higher absolute number of SNPs and substitutions within this category, leading to increased power in the statistical test. If this were true, a reduction in the absolute number of single nucleotide variants from this exon category would diminish the statistical significance.

To this end, we sub-divided the set of constitutive exons according to an important biologic feature, which governs the splicing process—the strength of the ss. Our intent was to investigate if the accumulation of slightly deleterious polymorphisms observed for the whole set of constitutive exons was in any way associated with ss strength.

Generally, constitutive exons have a stronger ss than alternative exons (Clark and Thanaraj 2002; Shepard et al. 2011). This implies that ESRs are probably less functional in constitutive than in alternative exons. However, among the constitutive exons, the ss score is variable. Assuming that a stronger ss allows better recognition and therefore less dependency on ESRs, it is predictable that constitutive exons with a weaker ss would be more sensitive to the inhibitory effects of ESS on exon inclusion. We tested this

Table 5 Modified MK test applied for rate-shifted ESRs identified from sequences of human-skipped exons (rate-shifted$_{\text{skipped}}$ ESRs) located within (a) constitutive exons and (b) skipped exons

	Create rate-shifted$_{\text{skipped}}$	Maintain rate-shifted$_{\text{skipped}}$
(a) Constitutive exons		
Substitutions	4615 (−0.8 %)	1276 (+3.2 %)
Polymorphism	1595 (+2.5 %)	373 (−9.5 %)
NI = 1.18		
χ^2 p value = 0.01		

Genes	Pool	p value-Pool	NI-Mantel–Haenszel (CI)	NI-Tarone–Greenland (CI)
2560	1.12	0.08	1.07 (0.91–1.27)	1.17 (0.95–1.41)

	Create rate-shifted$_{\text{skipped}}$	Maintain rate-shifted$_{\text{skipped}}$
(b) Skipped exons		
Substitutions	899 (−0.5 %)	238 (+1.8 %)
Polymorphism	352 (+1.2 %)	85 (−4.7 %)
NI = 1.09		
$\chi^2 p$ value = 0.56		

Genes	Pool	p value-Pool	NI-Mantel–Haenszel (CI)	NI-Tarone–Greenland (CI)
648	1.06	0.73	0.99 (0.64–1.58)	1.09 (0.64–1.95)

Only synonymous mutations that create or maintain rate-shifted$_{\text{skipped}}$ ESRs were used. The respective values for the NI and its CIs are presented below each contingency table

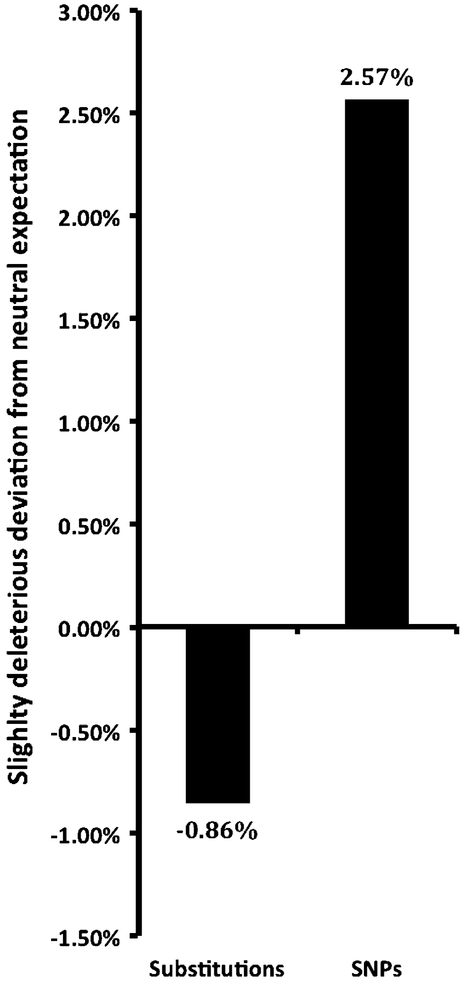

Fig. 4 Normalized difference between the observed and expected counts of substitutions and polymorphisms which create rate-shifted$_{\text{skipped}}$ ESRs within human constitutive exons. $\chi^2 p$ value <0.05

hypothesis using the modified MK approach to ESS, creating changes within constitutive exons with a strong or weak ss. The results show an increase in the enrichment of polymorphisms (with respect to substitutions), which accompanies a weakening of ss strength (Fig. 5; Supplementary Tables S4a–d). This result suggests that purifying selection is stronger on changes that occur in exons with a weaker ss. It also refutes the above-mentioned hypothesis that the significant signal of purifying selection on constitutive exons was only due to a higher absolute number of variant sites analyzed.

Similar results were observed by Xiao et al. (2007), who analyzed ESS motifs and found that the purifying selection against ESS is stronger in constitutive exons with a weaker ss. Moreover, it is important to note that among the exons with a weak ss, those with a weak 5′ ss show stronger signal of natural selection against the rate-shifted hexamers than those with a weak 3′ ss (Fig. 5). This result is in accordance with the major role of 5′ ss in stabilizing the spliceosome complex (Robberson et al. 1990), especially in mammals—where exon definition predominates (Sterner et al. 1996; Xiao et al. 2007). These results are again concordant with the findings of Xiao et al. (2007), who showed that changes in 5′ ss strength better predict the evolution of ESS than changes in the 3′ ss strength.

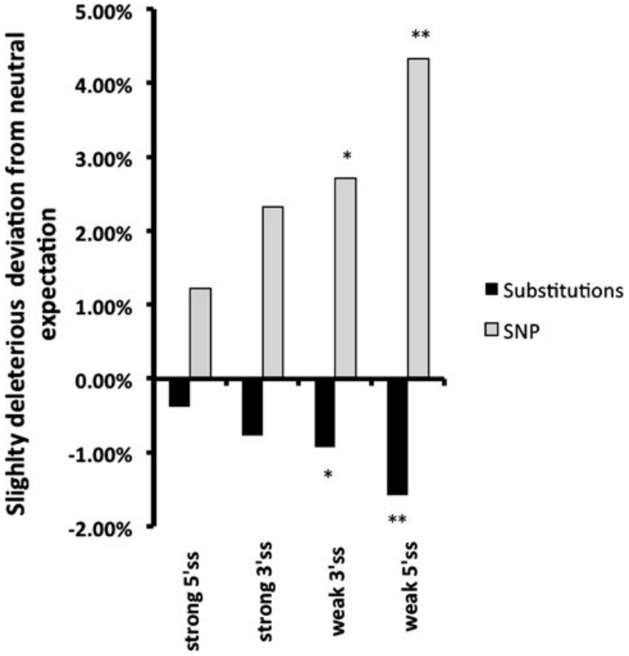

Fig. 5 Normalized difference between the observed and expected counts of substitutions and polymorphisms which create rate-shifted$_{\text{skipped}}$ ESRs within human constitutive exons according to ss strength. Constitutive exon sequences were divided according to the ss score and tested separately. Only constitutive exons with weak ss showed significant results for the modified version of the MK test. **χ^2 p value <0.01. *Marginally significant (χ^2 p value = 0.05)

Concluding Remarks

Certain evolutionary rate-shift events may be caused by selective pressure acting at the amino acid level (Pupko and Galtier 2002). However, this is unlikely to explain our results, showing that synonymous mutations located in rate-shifted ESRs display significant deviations from neutral expectations. Moreover, the fact that the purifying selection signal varies depending on the kind of ESR alteration (creation or loss) suggests that selection is acting at the RNA level.

A majority of the rate-shifted hexamers that we identified in this study are identical (or completely aligned) to known ESEs, including several binding sites for SRPs. However, we believe that those identified from skipped exons act as silencers. This hypothesis is supported by our finding of a higher proportion of known ESS and lower proportion of ESEs among them (relative to ESRs identified in constitutive exons). Moreover, these hexamers present an evolutionary pattern similar to known ESSs, i.e., both are under purifying selection in constitutive exons. These findings are in accordance with previous reports about the context-dependent function of ESRs (Goren et al. 2006; Jumaa and Nielsen 1997; Kanopka et al. 1996; Solis et al. 2008; Ule et al. 2006). Further experimental analysis will be necessary to prove the silencer function of these rate-shifted hexamers.

Finally, this study provides the first attempt to demonstrate consistent evidence that, at the populational level, slightly deleterious mutations that occur in constitutive exons reduce their potential to be recognized during the splicing process. Given that the MK test revealed an excess of SNPs creating ESRs with putative silencer function (relative to substitutions) within constitutive exons, we conclude that these SNPs should reduce both the strength of splicing signals and exon recognition itself. This result strengthens the hypothesis of a constitutive origin for skipped exons and predicts that many human exons that are defined solely as constitutive should present some splicing variants at low level. The new technology of deep RNA sequencing applied to large populational samples will allow better detection of those cases.

Acknowledgments We thank Daniel Ohara for technical support, Dr. Pedro Galante for providing the exon database, and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) for the PhD Fellowship (2007/59721-8) to RFR and for the grant (2009/09127-8) to DM. SG is a Fellow of the Edmond J. Safra Bioinformatics Program at Tel-Aviv University.

Conflict of interest The authors declare that they have no conflict of interest.

References

Carlini DB, Genut JE (2006) Synonymous SNPs provide evidence for selective constraint on human exonic splicing enhancers. J Mol Evol 62:89

Clark F, Thanaraj TA (2002) Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. Hum Mol Genet 11:451

Consortium GP (2010) A map of human genome variation from population-scale sequencing. Nature 467:1061

de Souza JE, Ramalho RF, Galante PA, Meyer D, de Souza SJ (2011) Alternative splicing and genetic diversity: silencers are more frequently modified by SNVs associated with alternative exon/intron borders. Nucleic Acids Res 39:4942

Fairbrother WG, Yeh RF, Sharp PA, Burge CB (2002) Predictive identification of exonic splicing enhancers in human genes. Science 297:1007

Fairbrother WG, Holste D, Burge CB, Sharp PA (2004) Single nucleotide polymorphism-based validation of exonic splicing enhancers. PLoS Biol 2:E268

Galante PA, Sakabe NJ, Kirschbaum-Slager N, de Souza SJ (2004) Detection and evaluation of intron retention events in the human transcriptome. RNA 10:757

Galante PA, Vidal DO, de Souza JE, Camargo AA, de Souza SJ (2007) Sense-antisense pairs in mammals: functional and evolutionary considerations. Genome Biol 8:R40

Gelfman S, Burstein D, Penn O, Savchenko A, Amit M, Schwartz S, Pupko T, Ast G (2012) Changes in exon–intron structure during vertebrate evolution affect the splicing pattern of exons. Genome Res 22:35

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

Jumaa H, Nielsen PJ (1997) The splicing factor SRp20 modifies splicing of its own mRNA and ASF/SF2 antagonizes this regulation. EMBO J 16:5077

Kanopka AM, Mühlemann O, Akusjärvi G (1996) Inhibition by SR proteins of splicing of a regulated adenovirus pre-mRNA. Nature 381:535

Ke S, Zhang XH, Chasin LA (2008) Positive selection acting on splicing motifs reflects compensatory evolution. Genome Res 18:533

Kim E, Magen A, Ast G (2007) Different levels of alternative splicing among eukaryotes. Nucleic Acids Res 35:125

Letunic I, Copley RR, Bork P (2002) Common exon duplication in animals and its role in alternative splicing. Hum Mol Genet 11:1561

Lev-Maor G, Goren A, Sela N, Kim E, Keren H, Doron-Faigenboim A, Leibman-Barak S, Pupko T, Ast G (2007) The “alternative” choice of constitutive exons throughout evolution. PLoS Genet 3:e203

Liu HX, Zhang M, Krainer AR (1998) Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. Genes Dev 12:1998

Liu HX, Chew SL, Cartegni L, Zhang MQ, Krainer AR (2000) Exonic splicing enhancer motif recognized by human SC35 under splicing conditions. Mol Cell Biol 20:1063

McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in *Drosophila*. Nature 351:652

Ohta T (1973) Slightly deleterious mutant substitutions in evolution. Nature 246:96

Pupko T, Galtier N (2002) A Covarion-based method for detecting molecular adaptation: application to the evolution of primate mitochondrial genomes. Proc Biol Sci 269:1313

Robberson BL, Cote GJ, Berget SM (1990) Exon definition may facilitate splice site selection in RNAs with multiple exons. Mol Cell Biol 10:84

Shepard PJ, Choi EA, Busch A, Hertel KJ (2011) Efficient internal exon recognition depends on near equal contributions from the 3′ and 5′ splice sites. Nucleic Acids Res 39:8928

Solis AS, Peng R, Crawford JB, Phillips JA, Patton JG (2008) Growth hormone deficiency and splicing fidelity: two serine/arginine-rich proteins, ASF/SF2 and SC35, act antagonistically. J Biol Chem 283:23619

Sorek R, Ast G, Graur D (2002) Alu-containing exons are alternatively spliced. Genome Res 12:1060

Sterner DA, Carlo T, Berget SM (1996) Architectural limits on split genes. Proc Natl Acad Sci USA 93:15081

Ule J, Stefani G, Mele A, Ruggiu M, Wang X, Taneri B, Gaasterland T, Blencowe BJ, Darnell RB (2006) An RNA map predicting Nova-dependent splicing regulation. Nature 444:580

Wang Z, Rolish ME, Yeo G, Tung V, Mawson M, Burge CB (2004) Systematic identification and analysis of exonic splicing silencers. Cell 119:831

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB (2008) Alternative isoform regulation in human tissue transcriptomes. Nature 456:470

Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38:e164

Xiao X, Wang Z, Jang M, Burge CB (2007) Coevolutionary networks of splicing *cis*-regulatory elements. Proc Natl Acad Sci USA 104:18583

Yeo G, Burge CB (2004) Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol 11:377

Zhang XH, Chasin LA (2004) Computational definition of sequence motifs governing constitutive exon splicing. Genes Dev 18:1241

Zhang Z, Zhou L, Wang P, Liu Y, Chen X, Hu L, Kong X (2009) Divergence of exonic splicing elements after gene duplication and the impact on gene structures. Genome Biol 10:R120