Published online February 24, 2004

Nucleic Acids Research, 2004, Vol. 32, No. 4 ***1261–1269***
DOI: 10.1093/nar/gkh284

Evidence for a subpopulation of conserved alternative splicing events under selection pressure for protein reading frame preservation

Alissa Resch, Yi Xing, Alexander Alekseyenko, Barmak Modrek and Christopher Lee*

Molecular Biology Institute, Institute for Genomics and Proteomics, and Department of Chemistry and Biochemistry, University of California, Los Angeles, Los Angeles, CA 90095-1570, USA

Received October 15, 2003; Revised January 9, 2004; Accepted January 26, 2004

ABSTRACT

Recently there has been much interest in assessing the role of alternative splicing in evolution. We have sought to measure functional selection pressure on alternatively spliced single-exon skips, by calculating the fraction that are an exact multiple of 3 nt in length and therefore preserve protein reading-frame in both the exon-inclusion and exon-skip splice forms. The frame-preservation ratio (defined as the number of exons that are an exact multiple of three in length, divided by the number of exons that are not) was slightly above random for both constitutive exons and alternatively spliced exons as a whole in human and mouse. However, orthologous exons that were observed to be alternatively spliced in the expressed sequence tag data from two or more organisms showed a substantially increased bias to be frame-preserving. This effect held true only for exons within the protein coding region, and not the untranslated region. In five animal genomes (human, mouse, rat, zebrafish, *Drosophila*), we observed an association between these conserved alternative splicing events and increased selection pressure for frame-preservation. Surprisingly, this effect became stronger as a function of decreasing exon inclusion level: for alternatively spliced exons that were included in a majority of the gene's transcripts, the frame-preservation bias was no higher than that of constitutive exons, whereas for alternatively spliced exons that were included in only a minority of the gene's transcripts, the frame-preservation bias increased nearly 20-fold. These data indicate that a subpopulation of modern alternative splicing events was present in the common ancestors of these genomes, and was under functional selection pressure to preserve the protein reading frame.

INTRODUCTION

Over the last several years alternative splicing has emerged as an increasingly important contributor to genomic complexity and gene function (1–8). In the human genome more than 30 000 alternative splice relationships have been identified from expressed sequence tag (EST) and mRNA sequences mapped on the genomic sequence (9), approximately doubling the number of transcript forms expected from the estimated 32 000 genes (5). Confronted with so many new splice forms from high-throughput EST sequencing, it is natural to ask whether these splice forms are functional and if so how they contribute to regulation of gene function (10–15).

We can categorize several possible impacts of alternative splicing based on its effect on the transcript and protein products. First, it can affect mRNA processing and stability, for example by inducing nonsense mediated decay (NMD) (13). Secondly, it can replace, extend or truncate the protein's N- or C-termini. Thirdly, it can add or remove an internal segment of the protein sequence, without altering the protein sequence before or after this point. For this third case, the exonic region that is alternatively spliced must be an exact multiple of 3 nt in length; we will refer to such an alternative splicing event as 'frame-preserving' (Fig. 1). Alternative splicing of an exonic region that is not an exact multiple of 3 nt will change the reading frame of subsequent exons; we will refer to this as 'frame-switching'.

Intuitively, one might expect strong selection pressure for alternatively spliced exons to be an exact multiple of 3 nt in length, both because this enables functional units within a protein to be switched on or off in a modular fashion, and also because alternative splices that shift the protein reading frame are more likely to induce nonsense-mediated decay (13). Indeed, there is some evidence that alternatively spliced exons in the human genome are biased to be a multiple of three in length (16). There is also evidence from studies of intron phase that constitutive exons show selection for frame-preservation (17). Intron phase zero is defined to be an intron that sits between two codons; phase one when the intron is between the first and second nucleotide of a codon; and phase two when the intron is between the second and third nucleotides of a codon. Gilbert and co-workers have defined exons that are flanked by introns with equal phase values as

*To whom correspondence should be addressed. Tel: +1 310 825 7374; Fax: +1 310 267 0248; Email: leec@mbi.ucla.edu

'symmetrical'; this is equivalent to our definition that an exon is 'frame-preserving'. Among constitutive exons, it has been reported that 'symmetrical' exons are observed more frequently than expected by random chance, consistent with the suggestion that there is selection pressure for exons to be frame-preserving (17). Taking intron phase bias in protein coding regions into account, the expected frequency of 'symmetrical' (i.e. frame-preserving) exons under a random model is 39.8% (17), but the actual frequency is substantially higher (45% for constitutive exons in protein coding regions) (17). Of course, there are likely to be many specific cases where there is not necessarily selection pressure for an exon to be frame-preserving.

For any group of alternatively spliced exons, we define the frame-preservation ratio to be the ratio of the number of exons that are an exact multiple of three in length, divided by the number that are not. This metric provides a simple but potentially revealing measure of functional selection pressure on alternatively spliced exons. Recently, there has been great interest in evaluating the potential contribution of alternative splicing to evolution of mammalian genomes (12,16,18 –21). In this paper, we describe evidence of significant functional selection pressure during the evolution of individual alternative splicing events. We have analyzed orthologous exons from five genomes ranging from *Drosophila* to human, to see which factors in evolution have resulted in selection pressure for frame-preserving exons. Our analysis is based on previously described methods for elucidating gene structure and alternative splicing from mapping of mRNA and EST sequences onto the genomic sequence (6,9,19). In the first part of this paper we focus specifically on exon skipping events (where a single exon is observed to be included in one transcript form, but skipped in another transcript form, and is flanked by constitutive exons). In the latter part of the paper we look at alternative 5′ and alternative 3′ splicing events.

MATERIALS AND METHODS

Alternative splicing analysis

We detected alternative splice forms for human, mouse, rat, zebrafish and *Drosophila* by mapping mRNA and EST sequences onto genomic sequence as previously described (6) using the following data: (i) UniGene EST data (22) from January 2002 and July 2003 (human and mouse), December 2002 (rat), July 2002 (zebrafish and *Drosophila*); (ii) genomic sequence data, July 2003 (human and mouse), July 2002 (rat, zebrafish and *Drosophila*). All exon boundaries were checked for consensus splice site sequences.

Alternative splicing and ortholog data (below) were loaded into a relational database (MySQL) and analyzed via SQL queries. We used EST counts to calculate the exon inclusion level for alternatively spliced exons as previously described (19,23). *P*-values were calculated using two methods: Fisher's exact test (24), and direct numerical integration as previously described (25).

We calculated the probability of obtaining a given number of exons that are alternatively spliced in one genome, from orthologous exons in another genome. Given n candidate exons, out of N total exons in a given genome, we calculated the probability of finding by random chance at least m alternatively spliced exons (out of the M total alternatively spliced exons identified in that genome), according to the hypergeometric distribution (24):

$$p(i \geq m \mid N, M, n, random) = \sum_{i=m}^{n} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

For example, out of the N = 66 441 total number of exons in our mouse data set, M = 3072 were observed to be alternatively spliced single-exon skips. Starting from human alternatively spliced single-exon skips, we were able to map n = 1514 to an orthologous mouse exon. Of these, m = 181 were found to be alternatively spliced single-exon skips in the mouse EST data. Thus, the *P*-value for obtaining this result by random chance calculated from the above expression is $10^{-30.6}$.

Orthologous exon detection

We have employed a conservative method for identifying orthologous exons, as previously described (19). Briefly, to identify orthologous genes between human and mouse, we used pairs of genes reported as reciprocal best matches in the HomoloGene database (26), using HomoloGene data from July 2003. To identify orthologous genes for human versus rat, and human versus zebrafish, we used HomoloGene data from July 2002. We used all orthologous pairs of genes that we mapped successfully onto the genomic sequence in our splicing calculation. However, it should be emphasized that this approach is limited by the HomoloGene data, and identifies only a fraction of all possible orthologous genes in these genomes. We matched exons within orthologous gene pairs as previously described (19). For each exon sequence, we used BLAST to search the genomic sequence of the orthologous gene in the other organism, using a 10^{-5} expectation cutoff. We used RepeatMasker to screen out repetitive sequences for each BLAST step. To align orthologous exons, we used dynamic programming global alignment (27) implemented in the program POA (28) as described in Modrek and Lee (19). To identify orthologous exons from *Drosophila*, we searched a database of human or mouse exon sequences for matches to a given *Drosophila* exon sequence, using TBLASTX with a 10^{-10} expectation cutoff. We required several criteria for assigning an ortholog for a *Drosophila* exon: (i) the TBLASTX hit must cover the full-length of the exon allowing at most 10 nt of mismatch total at the ends; (ii) the length of the *Drosophila* exon must match that of the homologous exon to within 25%; in most cases the exons had the same length or differed by only a few nucleotides; (iii) the inferred amino acid sequences must have at least 50% identity; (iv) for previously characterized genes, the candidate orthology was checked against published literature.

Alternative 5′ and alternative 3′ splicing analysis

We also identified alternative 5′ and 3′ splicing in human and mouse, by looking for alternative splices that selected different splice sites in a single exon, altering the length of this exon that was included in transcripts (see Fig. 6). For

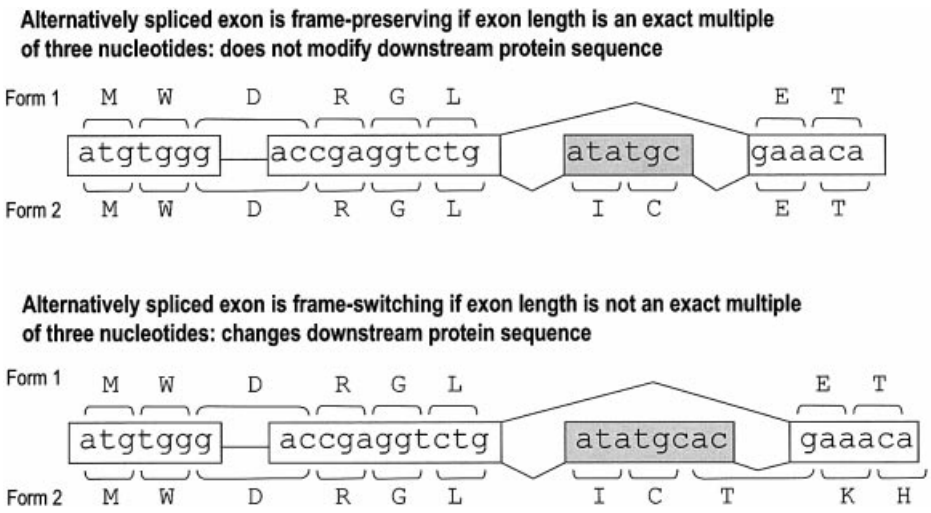

Figure 1. Exon length determines whether an alternatively spliced single-exon skip is frame-preserving or frame-switching. We define an alternatively spliced exon as frame-preserving if its length is an exact multiple of 3 nt, as its alternative splicing will not alter the protein reading frame of subsequence exons (top).

alternative 5′ splicing, we detected 1164 cases in human and 776 in mouse; for alternative 3′ splicing, we detected 2008 in human and 1322 in mouse. The criteria for matching orthologous exons were modified from the description above, simply by adding a requirement of 80% sequence identity within the constitutively included region of the exon, and 70% sequence identity for the alternatively included region of the exon.

URLs

Our results will be made available upon publication at http://www.bioinformatics.ucla.edu/ASAP. We used the UniGene (ftp://ftp.ncbi.nih.gov/repository/UniGene/) and (ftp://ftp.ncbi.nih.gov/pub/HomoloGene/) databases for our analysis. We downloaded genomic sequence from ftp://ftp.ncbi.nih.gov/genomes/H_sapiens (human sequence 2002), ftp://ftp.ensembl.org/pub/current_human/ (human sequence, 2003), ftp://ftp.ensembl.org/pub/current_mouse/ (mouse sequence), ftp://rat-ftp.hgsc.bcm.tcm.edu/pub/analysis/rat/chromosome (rat sequence), ftp://ftp.ensembl.org/pub/assembly/zebrafish/Zv2release/ (zebrafish sequence) and http://www.fruitfly.org/sequence/release3download.shtml (*Drosophila* sequence).

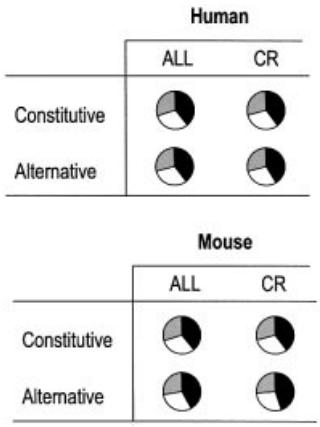

Figure 2. Genome-wide analysis of exon frame-preservation in human and mouse. Data for human (top) and mouse (bottom) are subdivided into constitutive exons (Constitutive) versus alternatively spliced single-exon skips (Alternative). CR indicates the subset of exons within each group that are within the protein coding region. Each pie chart shows the fraction of exons whose lengths are frame-preserving (exon length an exact multiple of 3 nt, black), versus frame-switching (white, gray).

RESULTS

Alternative splicing in the human and mouse genomes does not appear to preserve frame

We have performed separate genome-wide analyses of the exon frame-preservation ratio in human, and also in mouse (Fig. 2 and Table 1), comparing constitutive exons (defined as those that are always included in the transcript) and alternatively spliced single-exon skips (individual exons that are included in some transcripts but skipped in other transcripts for that gene, and flanked by constitutive exons). In our data set of 44 312 constitutive exons and 7112 alternatively spliced exons from the 2003 human data set, and 58 326 constitutive exons and 3072 alternatively spliced exons from the 2002 mouse data set, constitutive exons showed little bias to be frame-preserving (38.8% were frame-preserving in human, and 39.3% in mouse). Alternatively spliced exons were similar: 39.7% were frame-preserving in human, and 42.0% in mouse. Out of all possible exon lengths, only one-third (33%) are exact multiples of 3 nt. Therefore, the expected probability by pure random chance that a given exon is frame-preserving (an exact multiple of 3 nt in length) is 33%. Thus, both constitutive and alternatively spliced exons are 6–9% more likely to be frame-preserving than expected by this simple random model, but alternatively spliced exons are only 1–3% more likely to be frame-preserving than constitutive exons.

For exons within the protein coding region, alternatively spliced exons were slightly more likely to be frame-preserving (41.6% in human, 44.7% in mouse) than constitutive exons (39.7% in human, 39.5% in mouse). Taking intron phase bias in protein coding regions into account, the expected frequency

Table 1. Exon frame-preservation in human and mouse[a]

	All			CR			UTR		
	Frame 0	Frame 1	Frame 2	Frame 0	Frame 1	Frame 2	Frame 0	Frame 1	Frame 2
Human									
Constitutive exons	17 205	13 611	13 496	13 242	10 151	9996	3399	2888	2731
Alternatively spliced	2820	2167	2125	2092	1463	1477	818	756	711
Mouse									
Constitutive exons	22 897	17 824	17 605	17 983	13 893	13 647	3907	3229	3166
Alternatively spliced	1289	924	859	804	516	478	349	318	300

Throughout this table 'frame *N*' means the remainder after dividing the exon length by three; thus 'frame 0' is frame-preserving (an exact multiple of 3 nt), while frame 1 and frame 2 are frame-switching (not an exact multiple of three). CR, protein coding region; UTR, untranslated region.
[a]See Figure 2.

of frame-preserving exons under a random model is 39.8% (17), and for constitutive exons has been observed to range as high as 45% (17). Thus, alternatively spliced exons do not appear to show a marked increase in frame-preservation relative to constitutive exons.

In evaluating alternatively spliced exons identified from EST data, several studies have checked whether these exons are conserved in the orthologous gene sequences of related organisms (12,16,18–21). Significant sequence conservation of an exon in two or more related genomes implies selection pressure for conservation of that exon, suggesting that it is more likely to be functional. To assess whether genomic conservation of an exon between human and mouse might correlate with increased selection pressure for frame-preserving exons, we examined a subset of alternatively spliced exons from each genome that were conserved in the genomic sequence of the orthologous gene in the other genome (Fig. 3). Orthologous exon pairs between human and mouse showed a high level of homology (87% sequence identity, on average). However, these conserved orthologous exons showed no increase in frame-preservation (41.2% in human, 34.7% in mouse), even when restricted to just those within the protein coding region (40.9% in human, 38.0% in mouse). It should be emphasized that these data represent exons that were conserved in the genomic sequence of both genomes, but which were not necessarily observed to be alternatively spliced in ESTs from both species.

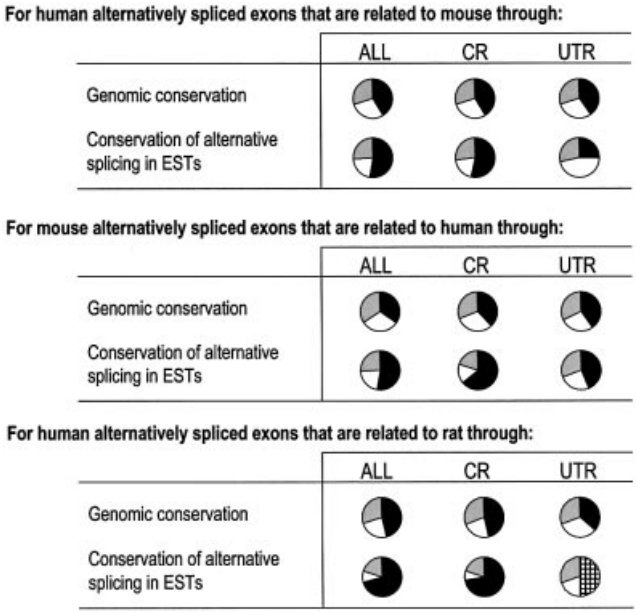

Figure 3. Frame-preservation selection pressure for orthologous alternatively spliced single-exon skips. Alternatively spliced exons from one organism were matched to orthologous exons in a second genome (see Materials and Methods), and divided into two categories. Genomic conservation, conserved in the genomic sequence (but not observed to be alternatively spliced in the second organism); conservation of alternative splicing in ESTs, also observed to be alternatively spliced in ESTs from the second organism. CR indicates the subset of exons within each group that are within the protein coding region; UTR indicates the subset of exons within each group that were within the untranslated region. Each pie chart shows the fraction of exons whose lengths are frame-preserving (exon length an exact multiple of 3 nt, black), versus frame-switching (white, gray). The human versus rat UTR data had insufficient counts for statistical significance.

Conserved alternative splicing events show increased selection pressure to preserve frame

It has previously been reported that human alternatively spliced exons have a high probability to be frame-preserving, if the orthologous exon in mouse is also observed to be alternatively spliced (16). We have examined this question via comparison of alternative splicing data from three genomes (human, mouse and rat; Fig. 3 and Table 2). In all three cases we observed a significant increase in frame-preservation when the exon is also observed to be alternatively spliced in ESTs from a second species (51.8% for human verus mouse; 51.9% for mouse versus human; 70% for human versus rat). All of these results were statistically significant, with *P*-values of 0.0064–0.000021 (Table 2). As a confirmation that this reflects functional pressure for preserving protein coding, we found that this increase was observed only in exons within the protein coding region (53.0% in human versus mouse; 63.6% in mouse versus human; 72% in human versus rat), and not within the untranslated region (UTR) (26.2% in human versus mouse; 44% in mouse versus human; insufficient counts for human versus rat).

Conservation of a specific characteristic between two related organisms is commonly interpreted as evidence that this characteristic was inherited from (and present in) their common ancestor. Thus, just as conservation of an exon sequence between the human and mouse genomes suggests that this exon was present in their common ancestor, observation of the identical alternative splicing pattern for a given exon in both human and mouse EST data suggests that this alternative splicing event was also present in their common ancestor. The fact that these conserved alternative

Table 2. Comparison of alternatively spliced exons in orthologous genes[a]

	All			CR			UTR		
	Frame 0	Frame 1	Frame 2	Frame 0	Frame 1	Frame 2	Frame 0	Frame 1	Frame 2
Human versus mouse									
Genomic conservation, but not alternatively spliced in mouse	381	264	279	328	228	245	110	78	83
Alternatively spliced in both	86	36	44	71	26	37	11	19	12
P-value	0.0058			0.0047					
Mouse versus human									
Genomic conservation, but not alternatively spliced in human	138	122	138	70	56	58	21	14	17
Alternatively spliced in both	109	47	54	35	9	11	7	4	5
P-value	0.000021			0.00042					
Human versus rat									
Genomic conservation, but not alternatively spliced in both	158	82	104	150	75	101	20	18	17
Alternatively spliced in both	21	3	6	18	2	5	5	2	3
P-value	0.0059			0.0064					

CR, the protein coding region; UTR, untranslated region.
[a]See Figure 3.

splicing events display a markedly different frame-preservation ratio compared with other alternatively spliced exons supports this hypothesis, since it shows that they represent a distinct subpopulation and not just a random sampling of alternatively spliced exons as a whole. We will refer to this subpopulation of alternative splicing events that are observed in ESTs from both organisms as 'conserved alternative splicing events' (12,16), to distinguish them from conserved exons (i.e. conservation of the exon sequence in two or more genomes).

To further assess the hypothesis that conserved alternative splicing events are associated with increased selection pressure for frame-preservation, we analyzed alternative splicing of orthologous exons in five different genomes: human (Hs), mouse (Mm), rat (Rn), zebrafish (Dr) and *Drosophila* (Dm). Since there is much less EST data available for the latter organisms, we identified far fewer alternative splicing events in these organisms (Fig. 4A and Table 3). It should also be noted that the orthologous exon counts in this analysis are limited by the smaller fraction of genes that can reliably be mapped as orthologs in these genomes (see Materials and Methods). For organisms in which the available EST data revealed a relatively small number of alternatively spliced exons (52 in zebrafish, and ~300 each in rat and *Drosophila*), the fraction of these alternative splicing events that were also observed in another organism was high (Fig. 4B). For example, among alternative splice events observed in rat, 38% were also observed in mouse ESTs for the orthologous gene. In contrast, among alternative splice events observed in human, only 12% were also observed in mouse ESTs for the orthologous gene.

Does the conservation of these alternative splice events indicate that they were inherited from a common ancestor, or might they simply have arisen by chance during the separate evolution of these genomes? In all cases, the observed level of conservation of alternative splicing events was significantly higher than expected by random chance, suggesting that they reflect ancestral alternative splicing events that have been inherited from the common ancestors of these organisms. For example, even the lowest rate of conservation of alternative splicing events (observed in human: 12% of human alternatively spliced single-exon skips were also observed to be alternatively spliced as single-exon skips in mouse ESTs) was much greater than expected by random chance. Only 4.6% of mouse exons were observed to be alternatively spliced single-exon skips. Thus, for a randomly selected human exon, there is a 4.6% chance that its orthologous exon in mouse would be observed to be alternatively spliced as a single-exon skip in mouse ESTs. Overall, the probability of obtaining our observed result (181 out of 1514 counts = 12%) by random chance is $<10^{-30}$ (see Materials and Methods). Similarly, whereas only 8.5% of the human exons in our data set were alternatively spliced (single-exon skips), 27.9% (181 out of 649) of mouse alternatively spliced single-exon skips mapped to orthologous human exons that were observed to be alternatively spliced single-exon skips in human ESTs. The probability of this occurring by random chance is $<10^{-47}$. This indicates that these events reflect ancestral alternative splicing events that have been inherited from the common ancestors of these organisms.

The elevated level of conserved alternative splicing events observed in rat, zebrafish and *Drosophila* (Fig. 4B) was associated with a 2-fold increase in the frame-preservation ratio (Fig. 4C). Whereas the total set of alternatively spliced exons detected in human ESTs had almost the same frame-preservation ratio as constitutive exons in human, alternatively spliced exons detected in rat, zebrafish and *Drosophila* had a frame-preservation ratio 2-fold higher than that of the constitutive exons in each species. Thus, throughout these five genomes there appears to be an association between conserved alternative splicing events and increased selection pressure for exons to preserve frame.

Exon inclusion level is negatively correlated with the frame-preservation ratio

We have previously reported that the fractional level of inclusion of an alternatively spliced exon in ESTs (defined as the number of ESTs which include the exon, divided by the total number of ESTs that either include or skip the exon) appears to be an important determinant in exon evolution (19). To examine whether exon inclusion level plays any role in selection pressure for frame-preservation, we calculated the frame-preservation ratio for alternative splicing events that were conserved between human and mouse, grouping them

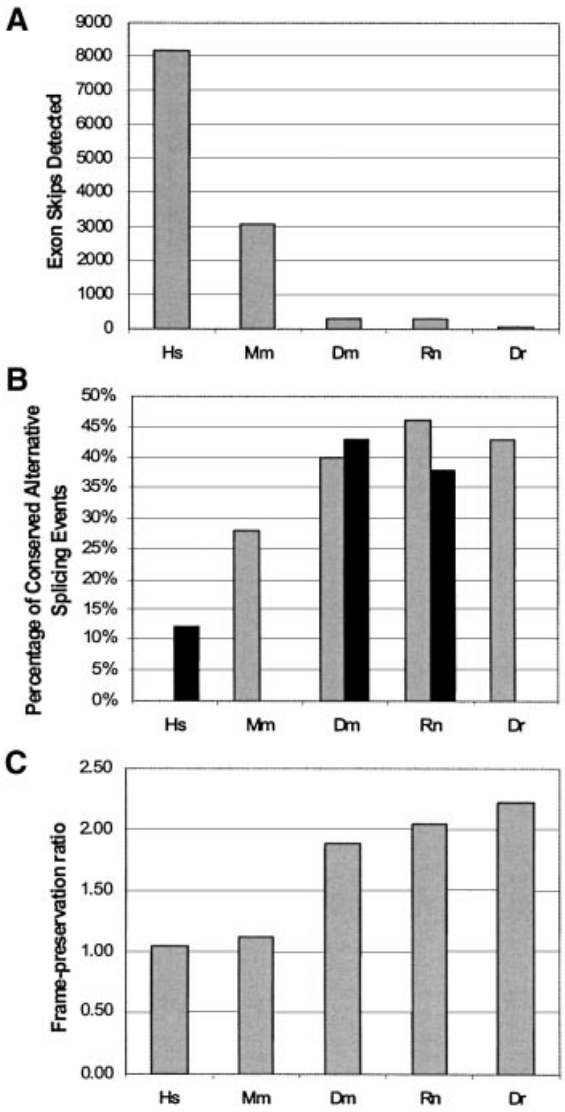

Figure 4. Conservation of alternative splicing single-exon skip events and frame-preservation selection pressure in five animal genomes. Data are shown for human (Hs), mouse (Mm), Drosophila (Dm), rat (Rn) and zebrafish (Dr). (**A**) The number of alternatively spliced single-exon skips detected for each species. (**B**) The fraction of orthologous alternatively spliced single-exon skips in each species that were also observed to be alternatively spliced in a second, reference species (black bars: mouse was used as the reference species; gray bars: human was used as the reference). (**C**) The relative frame-preservation ratio of alternatively spliced single-exon skips for each species. The relative frame-preservation ratio was calculated by taking the number of alternatively spliced exons that were frame-preserving, divided by the number that were not, divided by the same ratio for constitutively spliced exons.

into three sets by exon inclusion level (0–33%, 33–66%, 66–100%). These data show a negative correlation between the exon inclusion level and frame-preservation ratio, both in human and in mouse (Fig. 5 and Table 4). Whereas alternative exons with a high inclusion level had a frame-preservation ratio almost identical to that of constitutive exons, alternative exons with a low inclusion level had a frame-preservation ratio five to seven times higher. For alternative exons within the protein coding region only, this shift is even stronger (18- to 19-fold; data not shown). These results are statistically significant (P-values of $10^{-6.9}$ to $10^{-5.7}$; see Table 4).

Alternative 5′ splicing and alternative 3′ splicing data sets confirm that conserved alternative splicing is associated with selection pressure for frame-preservation

All of the results described so far were restricted to alternative splices that caused exon skipping (i.e. an alternative splice form that excludes an entire exon). To check whether our results are a general pattern, we have repeated our analysis on two different forms of alternative splicing: alternative 5′ splicing (in which alternative usage of two different 5′ splice sites changes the amount of an exon that is retained in the transcript; see Fig. 6), and alternative 3′ splicing (alternative usage of two different 3′ splicing sites). In the total set of alternative 5′ or alternative 3′ splices, the length of the alternative spliced region showed little selection pressure for frame-preservation (Fig. 6 and Table 5). In contrast, the subset of alternative 5′ splices in human that were conserved in mouse again showed significantly increased frame-preservation (Fig. 6). Similar results were observed for mouse alternative 5′ splices, and for alternative 3′ splices in human, and in mouse. Overall, these data are statistically significant (P-value 0.013).

DISCUSSION

EST data require careful interpretation for many reasons: potential experimental artifacts, clustering errors, incomplete coverage issues and sampling bias (7,8,12,29). For example, most alternative splicing researchers would agree that a single EST observation of a novel splice form does not constitute strong evidence of an authentic biological phenomenon. In contrast, sequence conservation in the genome sequences of related organisms is often viewed as highly significant, and has spawned an entire research field (comparative genomics).

From this point of view, our frame-preservation data are surprising. On the one hand, conservation of the exon in the genomic sequence of another species is associated with no significant increase in functional selection pressure (i.e. frame-preservation; compare Figs 2 and 3). On the other hand, observation of an EST from another species displaying the same alternative splice is associated with a significant increase in frame-preservation (see Fig. 3). For example, 40.9% of human alternatively spliced exons in the protein coding region are frame-preserving (even if they are conserved in the mouse genomic sequence), similar to constitutive exons (39.7%). But if such an exon is also observed to be alternatively spliced in rat ESTs, the probability that the exon is frame-preserving rises to 72%. In many cases this determination depended on the observation of a single EST from rat showing an alternative splice. It seems remarkable that observation of a single rat EST can tell us whether the exon's length in the human genome will likely be frame-preserving or not. Although one might be tempted to dismiss such an odd result as a fluke, multiple lines of evidence support it. This was obtained as a statistically significant result in independent comparisons of human exons versus rat ESTs, human exons versus mouse ESTs, and mouse exons versus human ESTs, and supported by further data from rat, zebrafish and *Drosophila*. This is evidently a general pattern, as it was

Table 3. Conservation of alternative splicing in five genomes[a]

	Mapped ESTs	Exon skips	Versus human 2002 Mapped to ortholog	Alternatively spliced in both	Versus mouse 2002 Mapped to ortholog	Alternatively spliced in both
Human	1 992 958	8177			1514	181
Mouse	1 560 419	3072	649	181		
Drosophila	186 815	306	15	6	7	3
Rat	91 590	295	76	35	98	37
Zebrafish	73 754	52	7	3		

[a]See Figure 4.

observed both for exon-skip alternative splicing (Fig. 2), and alternative 5′ and alternative 3′ splicing (Fig. 6). In contrast, no such increase was observed for alternatively spliced exons in the UTR, indicating that this result does reflect selection pressure for maintaining the protein reading frame. Moreover, our data indicate that EST observations are meaningful for frame-preservation not only qualitatively, but quantitatively. The exon inclusion level measured from EST counts has a strong negative correlation with the frame-preservation ratio, which is nearly 20-fold higher for alternatively spliced single-exon skips with a low inclusion level than for those with a high inclusion level.

We hypothesize that there exists a distinct set of alternative splicing events that were present in the common ancestors of the organisms included in this study. We have found that conservation of alternative splicing in ESTs from different organisms was observed much more frequently than expected by random chance (e.g. *P*-value 10^{-30} for the human versus mouse comparison). This suggests that the observed conservation of alternative splicing patterns is not a random coincidence, but instead reflects inheritance of alternative splicing events from a common ancestor. This is supported by the fact that these putative ancestral alternative splicing events display markedly increased frame-preservation, relative to other alternatively spliced exons. For all five of the genomes included in this study, conserved alternative splicing appears to represent a population of exons specialized for modular alteration of protein architecture.

It is interesting to note that these conserved alternative splicing events appear to be enriched among the first, most easily detected alternative splices for each organism. That is, in organisms with less EST data and correspondingly few alternative splices detected, we found a much higher fraction of these events to be observed also in ESTs from another species (see Fig. 4A and B). Further work is needed to assess whether this is a generally valid pattern, and if so, to investigate why. For example, it may be that alternative splicing in highly expressed genes tends to be more ancient in origin than alternative splicing as a whole in any given species, or that alternative splicing in highly expressed genes is more conserved within these five organisms than alternative splicing in other genes is.

The observation that exon inclusion level is negatively correlated with selection pressure for frame-preservation raises additional questions about the role that alternative splicing may have played during the evolution of the human and mouse genomes. At first glance, this result may seem paradoxical. We have previously defined 'major-form' exons as those which are included in the majority of transcripts (exon inclusion level >50%), and 'minor-form' exons as those which are included in only a minority of transcripts (exon inclusion level <50%) (19). The minor-form exons in our data set have less EST evidence than the major-form exons, and good questions have been raised about whether minor-form exons detected from ESTs are real biological forms, and specifically whether they generate functional proteins (12,13,19). In contrast, our data in Figure 5 indicate that minor-form exons show the strongest evidence of functional selection pressure (a

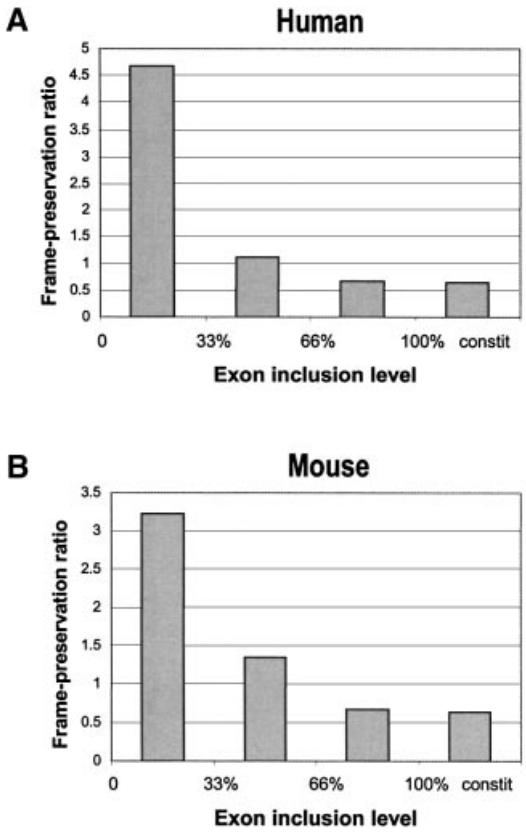

Figure 5. The frame-preservation ratio of conserved alternatively spliced single-exon skips, as a function of exon inclusion level. We divided alternatively spliced single-exon skips into three groups by exon inclusion level (0–33%, 33–66% and 66–100%), and calculated the frame-preservation ratio for each group. As a control, we also calculated the frame-preservation ratio for constitutively spliced exons (constit). (**A**) Human alternatively spliced exons. (**B**) Mouse alternatively spliced exons. Both data sets were restricted to the set of exons that were also observed to be alternatively spliced in the other organism.

nearly 20-fold increase in the frame-preservation ratio), while major-form exons show almost no evidence of functional selection pressure by the same criterion.

How can there be such a striking difference between minor-form versus major-form exons? One possible interpretation is that the alternative splicing of an exon may arise by different mechanisms in these two groups. We can hypothesize that a novel splice form created during evolution is more likely to initially be expressed at a lower level than the original form of the transcript, instead of a higher level. If an exon were originally constitutive (included in 100% of transcripts of a gene), and a new splice that skips the exon were introduced, we would expect the new splice form to be produced as <50% of the gene's transcripts, resulting in the transformation of a constitutive exon to a 'major-form' alternatively spliced exon. Our data show that for constitutive exons there is apparently little selective advantage for frame-preservation, in agreement with previous reports (16). Major-form exons are almost exactly like constitutive exons in their lack of selection pressure for frame-preservation (see Fig. 5), as one might expect under this model. In contrast, if a new exon were introduced into an existing gene by alternative splicing, this model would predict it to arise as a minor-form exon (included in <50% of the gene's transcripts). Thus, the model suggests major-form exons might arise from constitutive exons (by introduction of an alternative splice that skips the exon), whereas minor-form exons might arise from the addition of a new exon in the transcript sequence. One prediction of this model is that minor-form exons would be predicted to be more recent in evolutionary origin than major-form exons. We recently presented independent evidence from comparison of three genomes that this may indeed be the case (19).

Such different mechanisms could give rise to different levels of selection pressure for frame-preservation, such as the following speculative model. In the major-form case, the novel splice form removes an existing segment of the functioning protein. For example, this might remove the regulatory domain from a protein, leaving its catalytic domain constitutively active in tissues expressing the novel splice form. Since the primary functional consequence of such a removal is disruption of an existing functional segment, it does not seem critically important whether this removal is confined strictly to a single exon (by preserving the reading frame of subsequent exons) or affects multiple exons (due to a frameshift). The latter case could still produce a useful functional impact (for example, induction of NMD). In contrast, in the minor-form case, the novel splice form inserts a new sequence into the protein. Unlike the major-form case (removal), this introduces a new sequence segment into the protein that can be positively selected if it produces some functional benefit. For this new minor form to become fixed in the population and retained through the subsequent evolution of multiple descendant genomes as we have observed, it would

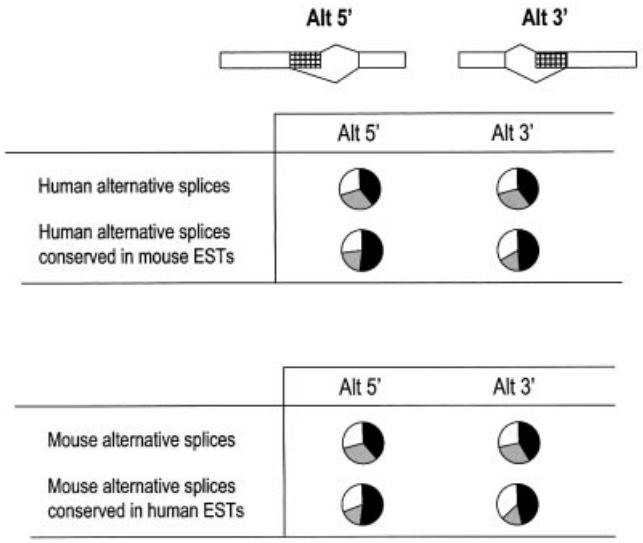

Figure 6. Frame-preservation selection pressure for alternative 5′ and alternative 3′ splicing. Alternatively spliced exons from one organism were matched to orthologous exons in a second genome. Each pie chart shows the fraction of exons whose lengths are frame-preserving (exon length an exact multiple of 3 nt, black), versus frame-switching (white, gray).

Table 4. Frame-preservation ratio as a function of exon inclusion level[a]

	Frame 0	Frame 1	Frame 2	Ratio
Human versus mouse				
Exon inclusion level				
0–33%	28	2	4	4.67
33–66%	20	8	10	1.11
66–100%	38	26	30	0.68
Constitutive exon	17 205	13 611	13 496	0.63
P-value	0.00000012			
Mouse versus human				
Exon inclusion level				
0–33%	29	6	3	3.22
33–66%	35	10	16	1.35
66–100%	45	31	35	0.68
Constitutive exon	20 444	15 970	15 618	0.65
P-value	0.0000019			

[a]See Figure 5.

Table 5. Frame-preservation for alternative 5′ and 3′ splices[a]

	Alternative 5′			Alternative 3′		
	Frame 0	Frame 1	Frame 2	Frame 0	Frame 1	Frame 2
Human						
Alternatively spliced in human ESTs	453	358	353	794	616	598
Alternatively spliced in both human and mouse ESTs	15	6	8	17	6	12
Mouse						
Alternatively spliced in mouse ESTs	298	245	233	545	396	381
Alternatively spliced in both mouse and human ESTs	15	5	9	16	6	13

[a]See Figure 6.

likely be under strong pressure to preserve the reading frame of the rest of the existing protein.

ACKNOWLEDGEMENTS

We wish to thank C. Grasso and M. Quist for their helpful discussions and comments on this work. This work was supported by NIMH/NINDS Grant MH65166, and DOE grant DEFG0387ER60615. A.A. and B.M. are predoctoral trainees supported by NSF IGERT Award DGE-9987641.

REFERENCES

1. Mironov,A.A., Fickett,J.W. and Gelfand,M.S. (1999) Frequent alternative splicing of human genes. *Genome Res.*, **9**, 1288–1293.
2. Brett,D., Hanke,J., Lehmann,G., Haase,S., Delbruck,S., Krueger,S., Reich,J. and Bork,P. (2000) EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett.*, **474**, 83–86.
3. Croft,L., Schandorff,S., Clark,F., Burrage,K., Arctander,P. and Mattick,J.S. (2000) ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. *Nature Genet.*, **24**, 340–341.
4. Kan,Z., Rouchka,E.C., Gish,W.R. and States,D.J. (2001) Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. *Genome Res.*, **11**, 889–900.
5. I.H.G.S Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
6. Modrek,B., Resch,A., Grasso,C. and Lee,C. (2001) Genome-wide analysis of alternative splicing using human expressed sequence data. *Nucleic Acids Res.*, **29**, 2850–2859.
7. Modrek,B. and Lee,C. (2002) A genomic view of alternative splicing. *Nature Genet.*, **30**, 13–19.
8. Brett,D., Pospisil,H., Valcarcel,J., Reich,J. and Bork,P. (2002) Alternative splicing and genome complexity. *Nature Genet.*, **30**, 29–30.
9. Lee,C., Atanelov,L., Modrek,B. and Xing,Y. (2003) ASAP: The Alternative Splicing Annotation Project. *Nucleic Acids Res.*, **31**, 101–105.
10. Boue,S., Vingron,M., Kriventseva,E. and Koch,I. (2002) Theoretical analysis of alternative splice forms using computational methods *Bioinformatics*, **18** (Suppl. 2), S65–S73.
11. Heber,S., Alekseyev,M., Sze,S.H., Tang,H. and Pevzner,P.A. (2002) Splicing graphs and EST assembly problem. *Bioinformatics*, **18** (Suppl. 1), S181–S188.
12. Kan,Z., States,D. and Gish,W. (2002) Selecting for functional alternative splices in ESTs. *Genome Res.*, **12**, 1837–1845.
13. Lewis,B.P., Green,R.E. and Brenner,S.E. (2003) Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc. Natl Acad. Sci. USA*, **100**, 189–192.
14. Black,D.L. (2000) Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell*, **103**, 367–370.
15. Kriventseva,E.V., Koch,I., Apweiler,R., Vingron,M., Bork,P., Gelfand,M.S. and Sunyaev,S. (2003) Increase of functional diversity by alternative splicing. *Trends Genet.*, **19**, 124–128.
16. Thanaraj,T.A., Clark,F. and Muilu,J. (2003) Conservation of human alternative splice events in mouse. *Nucleic Acids Res.*, **31**, 2544–2552.
17. de Souza,S.J., Long,M., Klein,R.J., Roy,S., Lin,S. and Gilbert,W. (1998) Towards a resolution of the introns early/late debate: only phase zero introns are correlated with the structure of ancient proteins. *Proc. Natl Acad. Sci. USA*, **95**, 5094–5098.
18. Kondrashov,F.A. and Koonin,E.V. (2003) Evolution of alternative splicing: deletions, insertions and origin of functional parts of proteins from intron sequences. *Trends Genet.*, **19**, 115–119.
19. Modrek,B. and Lee,C. (2003) Alternative splicing in the human, mouse and rat genomes is associated with an increased rate of exon creation/loss. *Nature Genet.*, **34**, 177–180.
20. Nurtdinov,R.N., Artamonova,I.I., Mironov,A.A. and Gelfand,M.S. (2003) Low conservation of alternative splicing patterns in the human and mouse genomes. *Hum. Mol. Genet.*, **12**, 1313–1320.
21. Sorek,R. and Ast,G. (2003) Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res.*, **13**, 1631–1637.
22. Schuler,G. (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med.*, **75**, 694–698.
23. Xu,Q., Modrek,B. and Lee,C. (2002) Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Res.*, **30**, 3754–3766.
24. Bartoszynski,R. and Niewiadomska-Bugaj,M. (1996) *Probability and Statistical Inference*. John Wiley & Sons, New York.
25. Xu,Q. and Lee,C. (2003) Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences. *Nucleic Acids Res.*, **31**, 5635–5643.
26. Wheeler,D.L., Church,D.M., Lash,A.E., Leipe,D.D., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Tatusova,T.A., Wagner,L. *et al.* (2002) Database resources of the National Center for Biotechnology Information: 2002 update. *Nucleic Acids Res.*, **30**, 13–16.
27. Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.
28. Lee,C., Grasso,C. and Sharlow,M. (2002) Multiple sequence alignment using partial order graphs. *Bioinformatics*, **18**, 452–464.
29. Ewing,B. and Green,P. (2000) Analysis of expressed sequence tags indicates 35,000 human genes. *Nature Genet.*, **25**, 232–234.