Complex selection on 5′ splice sites in intron-rich organisms

Manuel Irimia,[1,5] Scott William Roy,[2,5,6] Daniel E. Neafsey,[3] Josep F. Abril,[1,4] Jordi Garcia-Fernandez,[1] and Eugene V. Koonin[2]

[1]*Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, 08028 Barcelona, Spain;* [2]*National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA;* [3]*Microbial Analysis Group, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA;* [4]*Institute of Biomedicine (IBUB), University of Barcelona, 08028 Barcelona, Spain*

In contrast to the typically streamlined genomes of prokaryotes, many eukaryotic genomes are riddled with long intergenic regions, spliceosomal introns, and repetitive elements. What explains the persistence of these and other seemingly suboptimal structures? There are three general hypotheses: (1) the structures in question are not actually suboptimal but optimal, being favored by selection, for unknown reasons; (2) the structures are not suboptimal, but of (essentially) equal fitness to "optimal" ones; or (3) the structures are truly suboptimal, but selection is too weak to systematically eliminate them. The 5′ splice sites of introns offer a rare opportunity to directly test these hypotheses. Intron-poor species show a clear consensus splice site; most introns begin with the same six nucleotide sequence (typically GTAAGT or GTATGT), indicating efficient selection for this consensus sequence. In contrast, intron-rich species have much less pronounced boundary consensus sequences, and only small minorities of introns in intron-rich species share the same boundary sequence. We studied rates of evolutionary change of 5′ splice sites in three groups of closely related intron-rich species—three primates, five *Drosophila* species, and four *Cryptococcus* fungi. Surprisingly, the results indicate that changes from consensus-to-variant nucleotides are generally disfavored by selection, but that changes from variant to consensus are neither favored nor disfavored. This evolutionary pattern is consistent with selective differences across introns, for instance, due to compensatory changes at other sites within the gene, which compensate for the otherwise suboptimal consensus-to-variant changes in splice boundaries.

[Supplemental material is available online at http://www.genome.org.]

Genomic sequencing of a broad variety of eukaryotes has revealed the ubiquity of seemingly costly genetic traits. A typical genome of a multicellular organism, as well as some genomes of unicellular eukaryotes, contains numerous transposable elements, spliceosomal introns, and gene duplicates, as well as long intergenic and untranslated regions (Britten and Davidson 1971; Lynch 2006). These features impose costs associated with DNA replication (Doolittle 1978), production, processing, translation, and degradation of RNA transcripts (Sapienza and Doolittle 1981; Lewis et al. 2003; Tress et al. 2007; Jaillon et al. 2008; McGuire et al. 2008; Roy and Irimia 2008), and elevated rates of deleterious mutations (Charlesworth et al. 1994; Conne et al. 2000; Lynch and Conery 2003; Lynch 2006).

Formally, there are three possible explanations for the persistence of these seemingly costly elements. First, despite any associated costs, an element could confer a net benefit due to some function (Davidson and Britten 1979; Lewis et al. 2003; Crombach and Hogeweg 2007; Häsler et al. 2007; Roy et al. 2007; Lev-Maor et al. 2008; Urrutia et al. 2008). Second, these elements could indeed be deleterious but persist due to ongoing creation by mutation (e.g., Hickey 1982; Charlesworth et al. 1994), or in the case of fixed elements, due to a lack of mutations removing these elements (e.g., Roy and Hartl 2006). Third, the elements could be neutral or the associated costs could be too small to be efficiently acted upon by selection, with neutral genetic drift dominating the elements' evolutionary fates (Lynch 2006).

We studied the case of variant 5′ splice sites of spliceosomal introns. In nearly all eukaryotes, the four intronic base pairs following the nearly universal GT/C at the splice junction exhibit a clear consensus sequence (Irimia et al. 2007a) (3′ consensus sequences are often limited to a terminal C/TAG, and are not studied here). The 5′ consensus reflects the importance of the 5′ splice site in intron recognition by complementary base pairing to the spliceosomal U1 snRNA (Zhuang and Weiner 1986; Séraphin et al. 1988; Siliciano and Guthrie 1988). However, adherence to the consensus varies significantly across species. In species with very few introns ("intron-poor" species), the majority of introns have the full 6-nucleotide (nt) consensus motif (e.g., 84% in the apicomplexan parasite *Cryptococcus parvum*). In contrast, in all characterized intron-rich genomes, the consensus sequence typically is far from perfect, with rather small minorities of introns possessing the full consensus (e.g., 14% in humans) (Irimia et al. 2007a; Schwartz et al. 2008).

Strict adherence to consensus in intron-poor species presumably indicates general efficient selection against variant boundaries. What evolutionary forces explain the frequent variant boundaries in intron-rich species? First, in some cases variant (nonconsensus) boundaries could, in fact, be optimal. For instance, the consensus motif is not always the most efficiently spliced (e.g., Mayeda and Ohshima 1988), and in some species, extended complementarity to the U1 RNA can inhibit splicing, suggesting that full complementarity of the splice boundary could be selected against in some cases (Staley and Guthrie 1999). For the case

[5]**These authors contributed equally to this work.**
[6]**Corresponding author.**
E-mail royscott@ncbi.nlm.nih.gov; fax (617) 496-5854.

Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.089276.108. Freely available online through the *Genome Research* Open Access option.

of alternatively spliced boundaries, splicing motifs may be under selection to remain weak (Garg and Green 2007; Ke et al. 2008). Second, in some cases variant and consensus boundaries could be equally efficient; for instance, when other exonic or intronic splicing signals fully compensate for variant boundaries (Puig et al. 1999; Zhang and Rosbash 1999; Förch et al. 2002). Finally, variant boundaries could be truly suboptimal, but the difference in fitness could be too small for selection to efficiently eliminate mutations to variant boundaries (Ohta and Kimura 1971).

These hypotheses make qualitatively different predictions about the rates of evolutionary change between consensus and variant boundary nucleotides. We studied rates of nucleotide change at intron boundaries in three groups of closely related intron-rich species. We find evidence for qualitatively different modes of selection acting across sites: Mutations from observed consensus nucleotides to variant nucleotides are efficiently opposed by selection, but mutations from observed variant nucleotides to consensus nucleotides occur at roughly the neutral rate. Thus, observed variant splice-site boundaries in the studied species are neither favored nor opposed by selection, but have fitnesses very near equal to those of an otherwise identical allele with consensus boundaries. We discuss possible explanations of this pattern and implications for the differential evolution of intron-rich and intron-poor species.

Results

Alignments of intron sequences across groups of closely related primate, *Drosophila,* and *Cryptococcus* species

We obtained genome-wide alignments of orthologous intronic sequences for three groups of closely related species: five species of *Drosophila*, four species of *Cryptococcus* basidiomycetous fungi, and three species of primates (Fig. 1). Alignments were filtered to exclude dubious intron predictions (see Methods). The total degree of intronic nucleotide divergence within each group was <0.4, and terminal branches typically showed <0.1 divergence at synonymous sites, with a minimum of 0.006 in the case of human–chimp (Fig. 1), allowing for confident reconstruction of ancestral states and sequence evolution in these groups.

Splice-site consensus sequences

Sequence logos were constructed for the 5′ splice sites of each species studied (Fig. 2A). As expected, we found a consensus sequence that was essentially limited to GTAAGT (or sometimes GTRAGT) at the 5′ splice site for all studied species, with very similar strengths and patterns of consensus sequences between closely related species (data not shown). Consistent with previous studies (Irimia et al. 2007a), the consensus motif was restricted to the first six or seven intronic positions, suggesting that positions beyond position +7 are not subject to appreciable selection on splice sites (Fig. 2A).

Rates of sequence change at splice sites

We next determined the rates of sequence change for different intronic sites (Fig. 2B). Consistent with the apparent lack of selection on splice sites beyond the observed 5′ consensus, overall rates of nucleotide change were roughly constant from position +8 onward (Fig. 2B) (throughout, for clarity of presentation, we present results for these three species, each of which shows typical results for its respective group). We therefore used positions +8 to +16 to obtain expectations for the degree of change at intronic positions in the absence of selection for splice-site motifs.

The expected degrees of change were calculated as follows: for each external branch, we identified sites from positions +8 to +16 with the same nucleotide in all closely related species (e.g., to study evolution along the human branch, we analyzed all sites with identical nucleotides in chimpanzee and macaque; Fig. 2C), which is likely to have been the ancestral nucleotide sequence. We then calculated the fraction of these sites that have undergone a change in the studied species (e.g., human). The degree of change was calculated separately for each of the four possible ancestral nucleotides in order to allow for mutational and/or selection biases in substitution.

Purifying selection on consensus nucleotides

We examined different kinds of change at each of the four splice-site positions (+3 to +6) for each lineage (positions +1 and +2 in this study were limited to the canonical GT dinucleotide). The clearest results were found in the case of change from consensus-to-variant nucleotides (Fig. 3A). Rates of change from consensus-to-variant nucleotides are quite different from those expected under the null model, with the observed rates ranging from two- to 12-fold smaller than expected, with *P*-values <<0.01 for all 44 comparisons (11 species × four positions). The observed/expected ratios (hereafter "Obs/Exp") differed across sites and across lineages (Supplemental

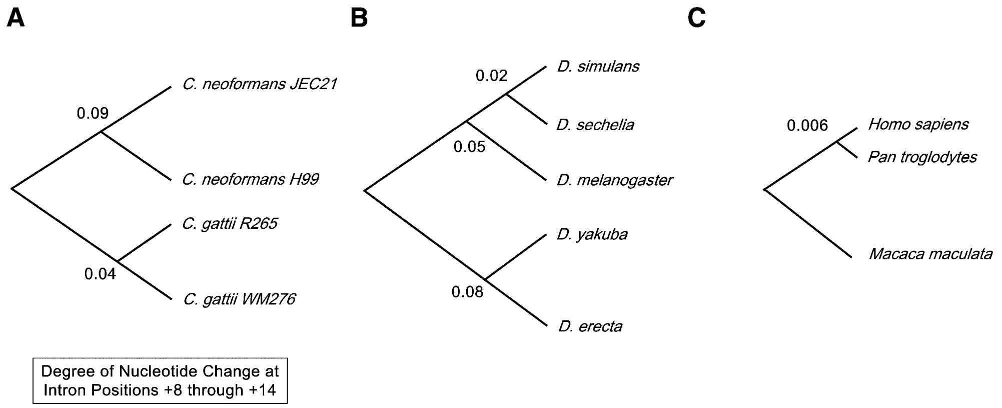

Figure 1. Relationships between studied species. The estimated degree of nucleotide change at intron positions +8 through +14 is given for each external branch.

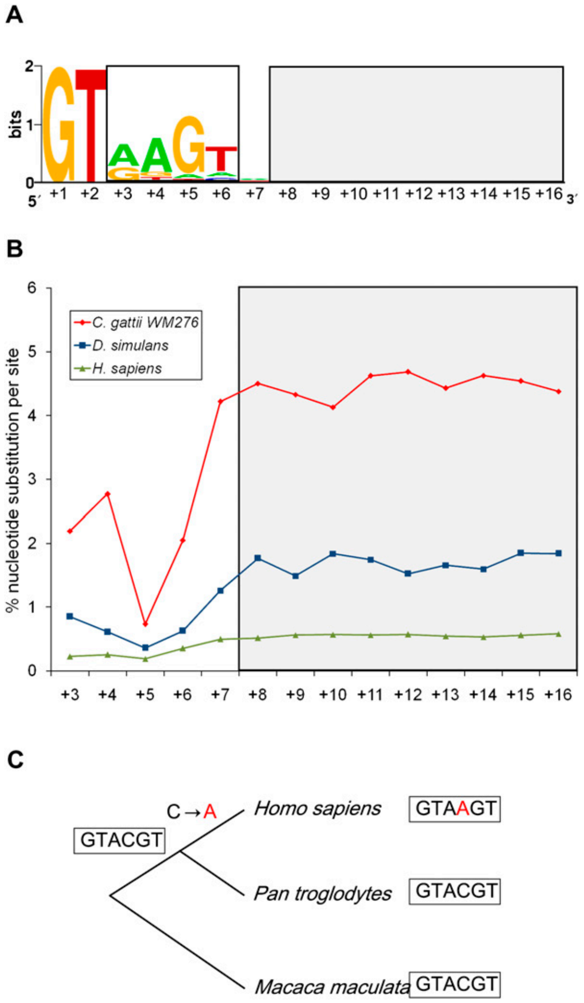

Figure 2. 5′ Boundary structure and methodology. (*A*) Sequence logo for *Cryptococcus gattii* species WM276, showing the typical restriction of 5′ consensus sequences to positions +1 through +6. (*B*) Degree of change for sites +3 through +16 for external branches leading to three representative species. Rates of change at positions +8 through +16 are much higher than at positions +3 through +6 and constant across sites, consistent with lack of selection for specific consensus motifs. (*C*) Illustration of method. For each nucleotide position +3 through +6, we studied boundaries at which the other three positions are conserved with sister and outgroup species (in this case, we study position +4, and thus require conservation at +3, +5, and +6).

Table 1). Figure 3B gives ranges of Obs/Exp ratios for each site for all species within each group of species (four *Cryptococcus*, five *Drosophila*, and two primates [*Homo*/*Pan*]). The Obs/Exp ratios were similar across all species, although *Drosophila* species showed slightly lower values, and *Cryptococcus* slightly higher values (Fig. 3B). These ratios also showed some systematic differences across sites, with site +5 showing the lowest Obs/Exp ratio in 10/11 lineages and sites +3 and +6 showing significantly higher values. However, there were also differences in the apparent relative importance of sites across lineages. For each *Drosophila* and *Cryptococcus* species, sites +3, +4, and +6 showed similar Obs/Exp ratios, whereas in both primates, rates of change at site +4 were comparable to +5, with +3 showing a much higher ratio, and +6 higher still. Also, the difference between +5 and the other sites was larger for *Cryptococcus* (fivefold difference between the Obs/Exp ratios for +5 and other sites) than for *Drosophila* (approximately twofold difference). These differences are consistent with the relative strength of the consensus at each site in genomes from the three

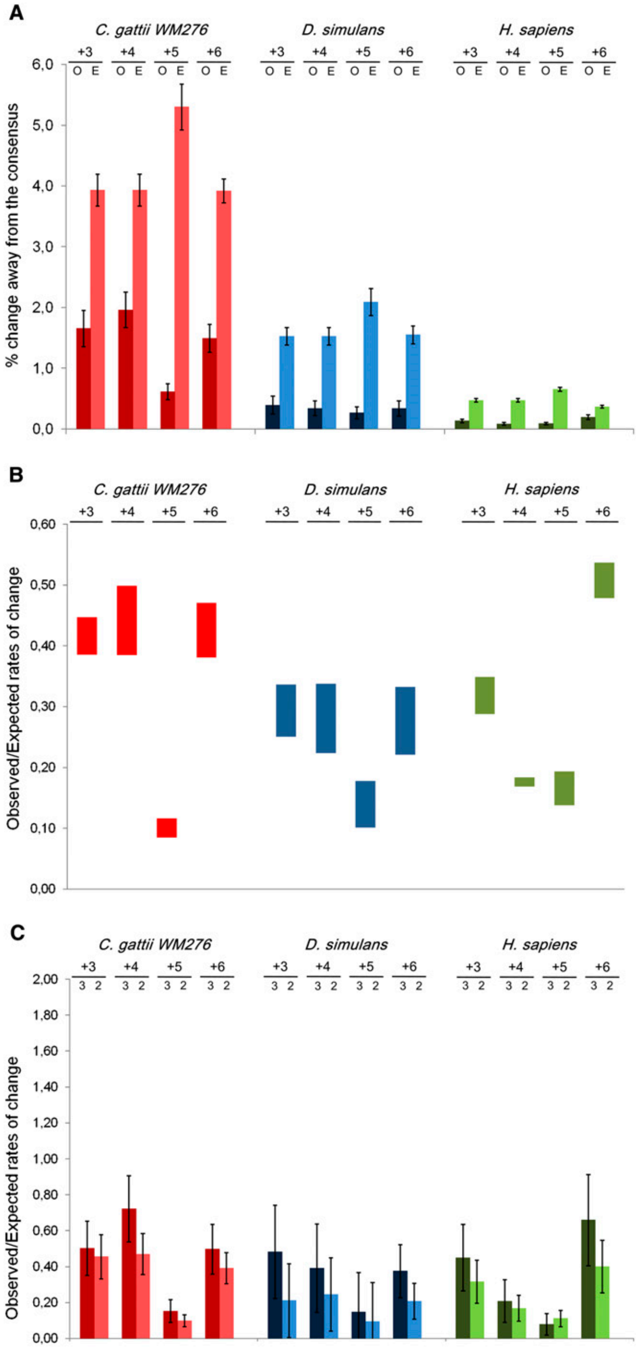

Figure 3. Consensus-to-variant changes for intron positions +3 through +6. (*A*) Observed (dark bars) and expected (light bars) degrees of change along external nucleotide branches are shown for three different species. (*B*) Range of sequence conservation across species within phylogenetic groups. The range of estimated Obs/Exp ratios for species within each group are given for each position (thus the *top*/*bottom* of the bar gives the estimate for the species with the highest/lowest corresponding estimated value within each species group). (*C*) Consensus-to-variant changes based on ancestral boundary strength. For each nucleotide site +3 through +6, Obs/Exp ratios for degree of change are given for boundaries with 4/4 (4) and 3/4 (3) ancestral consensus nucleotides. Error bars correspond to 95% confidence intervals for nucleotide change.

groups: +5 is by far the most highly conserved nucleotide in *Cryptococcus* and less so in *Drosophila*, whereas primates +4 and +5 show similar levels of conservation. These differences likely indicate subtle differences in splice-site recognition among lineages.

Importantly, preferential conservation of consensus nucleotides is found for all boundary subsets with different numbers of matches to consensus sequences (Fig. 3C): Even boundaries with all four consensus nucleotides (m4 boundaries) show significantly diminished rates of consensus-to-variant nucleotide. Thus, selection does not simply maintain some minimum level of adherence to consensus (e.g., opposing further mutations away from consensus at already variant boundaries). Rather, observed consensus nucleotides are generally preserved by purifying selection, regardless of the greater 5′ splice-site boundary context.

Neither purifying nor positive selection on variant-to-consensus changes

The finding that selection opposes changes from consensus-to-variant emphasizes the question of why so many variant nucleotides are observed in intron-rich organisms. There are basically three hypotheses. First, variant nucleotides could be generally suboptimal, opposed by weak selection regardless of context. In this case, rates of variant-to-consensus nucleotide change would be higher than the neutral expectation. Alternatively, the nucleotides at splice boundaries could be generally optimized: Where consensus nucleotides are observed, it is largely because they are favored over variant nucleotides; conversely, where variant nucleotides are observed, they too could typically be favored over consensus or other variant nucleotides. In this case, rates of variant-to-consensus changes would be lower than the neutral expectation. Finally, the presence of numerous variant nucleotides could be due to a lack of selection between consensus and variant nucleotides at those sites, in which case variant-to-consensus changes should occur at the neutral rate.

We found that the observed rates of change for variant-to-consensus in positions +3 to +6 along external branches were not significantly different from those predicted by the null/neutral model (Fig. 4; Supplemental Table 1). These results were consistent for all species and across splice-site positions with the exception of position +3, and suggest that overall selection does not distinguish between consensus and variant nucleotides in these cases. Excluding position +3, nominal statistical significance was reached for only 1/33 cases, and for zero cases, when multiple tests were corrected for by a Bonferroni correction. The exception at position +3 coincides with the clear preference for G above other nonconsensus nucleotides (C and T).

We also studied all changes from variant nucleotides to other nucleotides (e.g., changes from C to A, G or T at position +4; Supplemental Fig. 1). We found that overall rates of change are slightly reduced relative to the expectation. The pattern is clearly different from the pronounced deficit of consensus-to-variant changes (Obs/Exp <0.5 for only 1/44 tests, compared with 43/44 for conserved-to-variant changes); however, there is a noticeable trend for an overall reduction in rate, reaching statistical significance in 22/44 tests (including 9/11 tests for position +3). This reduction is likely due to decreased rates of change from the observed variant nucleotides to alternative variant nucleotides (given that, as shown above, rates of variant-to-consensus change are indistinguishable from the expectation). This decreased rate of change between variant nucleotides is likely to be a combination of: (1) general differences in splicing efficiency between different variant nucleotides at a given position (the clearest case being the preference for G over C or T at position +3); (2) intron-specific differences in splicing efficiency between variant nucleotides, for a given intron; and (3) site-specific mutational and compositional differences such

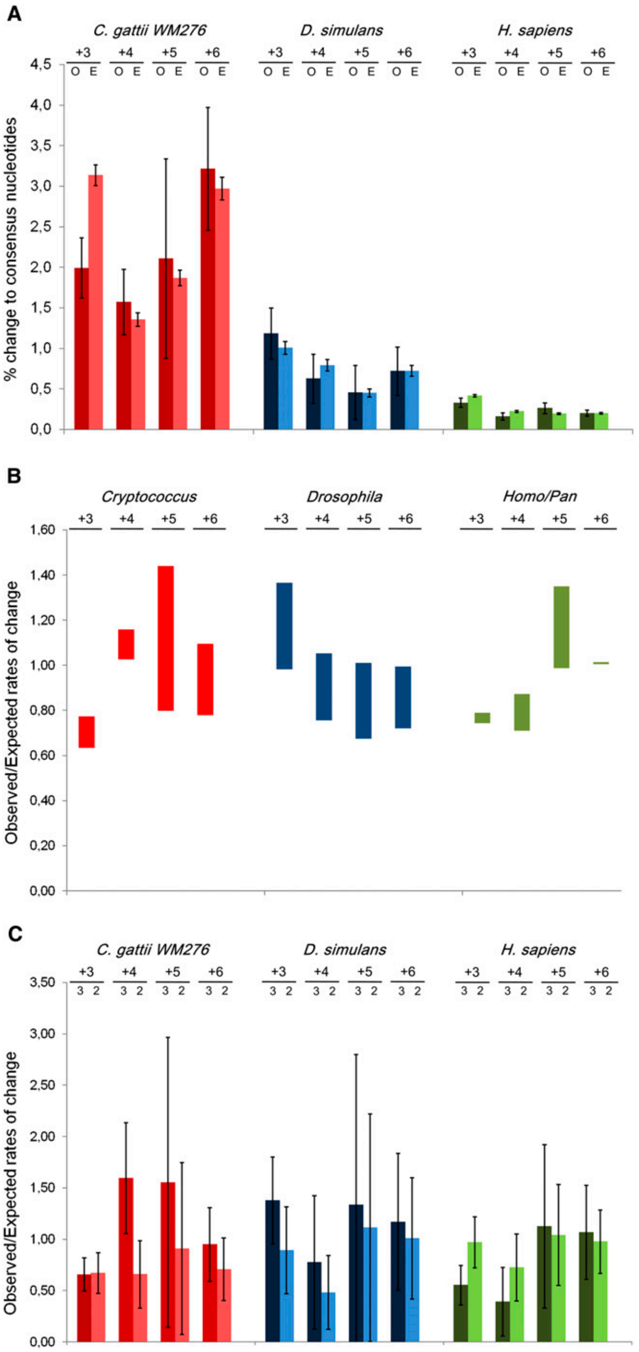

Figure 4. Variant-to-consensus changes for intron positions +3 through +6. (*A*) Observed (dark bars) and expected (light bars) degrees of change along external nucleotide branches are shown for three different species. (*B*) Range of sequence conservation across species within phylogenetic groups. The range of Obs/Exp ratios for species within each group are given for each position. (*C*) Variant-to-consensus changes based on ancestral boundary strength. For each nucleotide site +3 through +6, Obs/Exp ratios for degree of change are given for boundaries with 3/4 (3) and 2/4 (2) ancestral consensus nucleotides. Error bars correspond to 95% confidence intervals for nucleotide change.

that, for instance, variant nucleotide presence might correspond to local compositional biases, leading to decreased rates of change. At present, the small numbers of instances of individual variant nucleotides thwarted our attempts to distinguish between these possibilities (data not shown). In any case, the selective and/or mutational forces acting on these changes are considerably weaker than those working on consensus-to-variant changes.

Overall, we find that observed variant nucleotides are neither generally favored nor disfavored by selection relative to consensus nucleotides at the same site. Instead, observed variant nucleotides appear to change to consensus at neutral rates. These results thus indicate substantial heterogeneity of selection across sites. At sites where consensus nucleotides are observed, they are favored: Variant nucleotides would be suboptimal at these sites. In contrast, at sites where variant nucleotides are observed, these nucleotides are roughly equal in fitness to an allele with a consensus nucleotide, which is otherwise identical. Thus, the answer to the question "why are seemingly suboptimal boundaries observed" appears to be that weak boundaries are not suboptimal when they are observed, but would often be suboptimal where they are not observed.

Boundary context and rates of change of consensus nucleotides

We found that rates of consensus-to-variant changes depend on the overall level of adherence to consensus. More consensus-like boundaries undergo more change to variant nucleotides (Fig. 3C). In general, for nearly every species and every position (39 of 44 tests overall), the rate of changes for full consensus boundaries (four match-to-three match changes, or "4m-to-3m" changes) is higher than that for the 3m-to-2m changes (4m bars in Fig. 3C lower than corresponding 3m bars), and also higher than the overall average across boundaries. Thus, whereas selection opposes 4m-to-3m changes, this selection is weaker than the selection opposing changes away from consensus for already nonconsensus boundaries: The selective difference between 4m and 3m boundaries is smaller than that between 3m and 2m boundaries.

The case is very different for changes from variant to consensus nucleotides. We found no evidence that rates of change are associated with the overall consensus character of the boundary (the examples in Fig. 4C are somewhat misleading, as they suggest greater rates of changes at m3. However, when all 11 species are considered [Supplemental Table 2], no overall trend is apparent). Among all species at all three sites (excluding position +3), we found only two cases in which there was a nominally statistically significant difference in rates between 2m-to-3m changes and 3m-to-4m match changes, and none when multiple tests were taken into account. Nor was there a consistent trend in the direction of the (nonsignificant) differences: in 22/44 of cases 2m-to-3m changes had greater rates; in 22/44, 3m-to-4m changes had greater rates. This furthers the argument that, at least the majority of the observed variant nucleotides are not suboptimal: Even in the case of boundaries with multiple variant nucleotides, there is no evidence of general accelerated change to consensus.

Discussion

Simple and complex genomes: 5′ Intron splice sites

Genomic sequencing of a broad sampling of natural diversity has indicated dramatic diversity in genome architecture and complexity. Some species display striking order and simplicity of genomic structure, with contiguous genes separated by short intergenic distances and few or no repetitive elements. Other species show tremendous complexity, with genes dispersed over vast expanses of selfish, repetitive, and regulatory noncoding sequences, and transcript-encoding (exonic) portions of genes scattered within sometimes much longer transcribed regions. These differences also have parallels in regularity of sequence motif usage, with species showing different levels of preference in codon usage (Sharp and Li 1987), consistency of regulatory sequences (Ohta 2002), and diversity of genic start and stop signals (e.g., Lynch et al. 2005). The evolutionary forces responsible for, and phenotypic implications of these dramatic differences have become a central concern of molecular evolution (Charlesworth et al. 1994; Lynch and Conery 2000; Rogozin et al. 2003; Lynch 2006).

Here, we studied the case of 5′ intron splice sites. Although nearly all studied species show a clear consensus sequence across introns, the level of adherence to consensus varies dramatically, with vast majorities of introns having the perfect consensus in intron-poor species, but only small minorities in intron-rich species. What explains the preponderance of variant boundaries in intron-rich species? Extended complementarity to spliceosomal RNAs can sometimes inhibit splicing (e.g., Staley and Guthrie 1999), and alternatively spliced exons in rodents may be under selection to have boundaries with only weak pairing to spliceosomal RNAs (Garg and Green 2007), suggesting that partially complementary (variant) boundaries could sometimes be more efficient. Alternatively, low efficiency or strength of selection could mean that consensus boundaries are indeed favored, but by too small a factor to ensure strict adherence to consensus (Ohta and Kimura 1971). Finally, the additional splicing signals and splicing factors could compensate for lower intron-snRNA complementarity, rendering otherwise suboptimal variant boundaries equally efficient (e.g., Förch et al. 2002; Ke et al. 2008).

Complex selection on 5′ intron splice-site boundaries

In this work, we compared orthologous intron sequences across sets of closely related species to characterize the effect of selection on the splice boundary sequences. We find qualitative differences in the action of selection on changes from consensus to variant nucleotides, and the reverse, variant to consensus changes. Whereas the rate of change from consensus-to-variant nucleotides is typically several fold lower than expected in the absence of selection, changes from variant-to-consensus nucleotides occur at rates not statistically different from the null/neutral expectation. Our results were very similar across three distantly related sets of intron-rich species, suggesting that the general finding of variant boundaries across intron-rich species (Irimia et al. 2007a) reflects similar evolutionary forces acting across eukaryotes (true confidence in the generality of these results will require analysis of additional clusters of closely related species as appropriate genomic sequences become available).

Differences in selection across sites

This finding is exactly as expected if selection varies across sites, with variant nucleotides having either lesser or equal fitness to consensus nucleotides (selected and nonselected sites, respectively). In this case, selection will ensure mostly consensus nucleotides at selected sites, whereas mutational pressure will lead to mostly variant nucleotides at nonselected sites. Thus, the sets of sites with ancestral consensus and variant nucleotides will be dominated by selected and nonselected sites, respectively, leading to qualitative differences in selection between the two sets of sites.

What determines differences in strength of selection across different introns? In humans, introns are recognized by a host of RNA and protein factors that bind to both the core-splicing elements—5′ intron boundary, branch point site, and 3′ polyT tail—and to more distal signals largely located within exons (e.g., exonic splicing enhancers and silencers, ESEs, and ESSs, respectively). Recent work suggests that exonic splicing signals are common across intron-rich species, and are thus likely to be ancestral to eukaryotes (Warnecke et al. 2008). The presence of ESEs (and other exonic signals) has previously been shown to inversely correlate with adherence to intron boundary consensus (Carmel et al. 2004; Sorek et al. 2004; Roca et al. 2005; Zheng et al. 2005; Lev-Maor et al. 2007), and evolutionary changes in ESEs and ESSs have been shown to associate with changes at splicing boundaries (Ke et al. 2008). These results suggest a complex picture in which differences across introns in the presence of flanking exonic signals lead to qualitative differences in selection on splice sites and in which changes in these exonic signals could lead to differences in the strength of selection on individual splice sites through time.

Differential evolution of intron-poor and intron-rich species

Recently we uncovered an enigmatic trend among eukaryotic genomes: Introns from genomes with high intron density showed only weak adherence to boundary consensus sequences, whereas intron-poor genomes showed generally much stronger adherence to consensus (Irimia et al. 2007a; Irimia and Roy 2008; Schwartz et al. 2008). Both the phylogenetic pattern of intron boundary consensus strength and reconstructions of ancestral intron densities suggest that these intron-poor, strict-boundary lineages arose independently several times throughout eukaryotic evolution from more intron-rich, weaker consensus lineages (Irimia et al. 2007a). These findings suggested a qualitative difference in selection on intron boundary sequences in intron-poor and intron-rich species.

The current results help to clarify this pattern. We find that full-consensus boundaries are efficiently conserved by selection in three independent intron-rich lineages, as is expectedly the case for the vast majority of boundaries in intron-poor species. In this context, the difference between intron-poor and intron-rich species might reflect not a lack of boundaries subject to strong selection in intron-rich species, but rather a lack of neutrally evolving boundary sites in intron-poor species. Given the evidence that, in general, intron-poor species are descended from intron-rich ancestors (Rogozin et al. 2003; Roy and Gilbert 2005; Csurös et al. 2008), this indicates that loss of introns was associated with the extinction of permissive intron boundaries in intron-poor species.

This difference could reflect changes in splicing mechanisms between intron-rich and intron-poor species. Unlike humans and other intron-rich species, *Saccharomyces cerevisiae* and other intron-poor species are not known to extensively utilize such exonic signals (Plass et al. 2008). In this case, compensation for variant splicing boundaries would be far less likely, leading to strong selection for consensus boundaries for most or all introns. This could explain the general strong adherence to consensus motifs in intron-poor boundaries if loss of ESEs and other *cis*-splicing regulators is a general trend across intron-poor species.

One implication of the current work is worth noting: We find no evidence that nonconsensus intron boundaries are disfavored relative to consensus variants. This suggests that introns with nonconsensus boundaries are not generally more costly than are introns with consensus boundaries. If so, this would suggest against the hypothesis that introns in intron-poor species have consensus boundaries due to selection for preferential loss of introns with nonconsensus boundaries (see Irimia and Roy 2008).

Methods

We limited our study to very closely related species, since over greater evolutionary distances there is a significant possibility of multiple changes occurring, particularly in outgroup species, rendering parsimony inaccurate. Genome assemblies of *Cryptococcus neoformans* strains JEC21 and H99, and *Cryptococcus gattii* strains WM276 and R265 were acquired and aligned as previously described (Neafsey and Galagan 2007). Introns and intergenic regions in the alignment were identified according to the annotation produced by the Institute for Genome Research for strain JEC21, which was based on a library of 23,000 full-length cDNA-paired sequencing reads (Loftus et al. 2005). Only introns exhibiting canonical splice sites (GT/AG) in all four strains were analyzed, resulting in the elimination of 3511 coding-region introns and 349 UTR introns.

Intronic regions for the *Homo sapiens* genome (March 2006 assembly) were extracted using the UCSC Genome table browser and uploaded to Galaxy platform (Giardine et al. 2005). Orthologous regions for *Pan troglodytes* (March 2006 assembly) and *Macaca maculata* (January 2006 assembly) genomes were obtained using Lift-Over tool and genomic sequences retrieved for each species. Introns were filtered for redundancy. The first 100 intronic nucleotide positions for each intron were aligned using automated ClustalW.

Drosophila intron alignments were extracted from a 12-species whole-genome alignment created by Anat Caspi using Mercator (Dewey 2007) and MAVID (Bray and Pachter 2003) software (available at http://www.biostat.wisc.edu/~cdewey/fly_CAF1/). Intronic regions were identified using annotation coordinates corresponding to the *D. melanogaster* r4.3 genome release.

Alignments were then filtered to exclude dubious alignments. First, only intron-containing canonical GT/C…AG intron boundaries in all aligned species were retained. Second, all introns with a gap or "N," or otherwise ambiguous residue within the first 16 alignment positions were excluded for the analysis of the overall nucleotide change, and within the first six alignment positions for the comparison of nucleotide changes between stronger and weaker boundaries (the results were similar when filtered for the first 16 alignment positions). A total of 23,670 intron alignments were used for *Cryptococcus*, 25,124 for *Drosophila,* and 112,381 for primates. We did not study alternatively spliced introns separately due to small numbers of overall changes occurring between such closely related species.

A fraction of nucleotide change was calculated for each position under study. The corresponding 95% confidence interval (CI) was calculated for each fraction using the standard formula: $a \pm 1.96\sqrt{\frac{a(1-a)}{N}}$, where N is the total number of studied introns, and a, the fraction of introns with observed changes (Irimia et al. 2007b). Significance values for rates of consensus-to-variant, variant-to-consensus, and variant-to-any other nucleotide were calculated based on the number of standard deviations from unity of the Obs/Exp value, assuming a normal distribution. *P*-values for comparisons between rates of change for boundaries of different levels of adherence to consensus (three vs. four matches for consensus-to-variant changes; two vs. three matches for variant-to-consensus and variant-to-any other nucleotide) were calculated by a 2×2 χ^2 contingency table.

All filtering and indicated analyses were performed by novel Perl scripts.

Due to concerns about annotation quality in the *Cryptococcus* species, the least thoroughly annotated group, we also ran all analyses on the subset of introns for which splicing was confirmed by BLASTN searches against available cDNA sequences. The results were qualitatively and quantitatively very similar (data not shown).

Acknowledgments

M.I., J.G.F., and J.F.A. were funded by the Spanish Ministerio of Educación y Ciencia, through the BFU2005-00252 and BFU2005-00422 grants, respectively, S.W.R. and E.V.K. were funded by the Intramural Research Program of the National Library of Medicine at National Institutes of Health/DHHS, and D.E.N. by funds from the National Institute of Allergy and Infectious Disease and a grant from the National Science Foundation (0450812).

References

Bray N, Pachter L. 2003. MAVID multiple alignment server. *Nucleic Acids Res* **31:** 3525–3526.

Britten RJ, Davidson EH. 1971. Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. *Q Rev Biol* **46:** 111–138.

Carmel I, Tal S, Vig I, Ast G. 2004. Comparative analysis detects dependencies among the 5′ splice-site positions. *RNA* **10:** 828–840.

Charlesworth B, Sniegowski P, Stephan W. 1994. The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature* **371:** 215–220.

Conne B, Stutz A, Vassalli JD. 2000. The 3′ untranslated region of messenger RNA: A molecular "hotspot" for pathology? *Nat Med* **6:** 637–641.

Crombach A, Hogeweg P. 2007. Chromosome rearrangements and the evolution of genome structuring and adaptability. *Mol Biol Evol* **24:** 1130–1139.

Csurös M, Rogozin IB, Koonin EV. 2008. Extremely intron-rich genes in the alveolate ancestors inferred with a flexible maximum-likelihood approach. *Mol Biol Evol* **25:** 903–911.

Davidson EH, Britten RJ. 1979. Regulation of gene expression: Possible role of repetitive sequences. *Science* **204:** 1052–1059.

Dewey CN. 2007. Aligning multiple whole genomes with Mercator and MAVID. *Methods Mol Biol* **395:** 221–236.

Doolittle WF. 1978. Genes in pieces: Were they ever together? *Nature* **272:** 581–582.

Förch P, Puig O, Martínez C, Séraphin B, Valcárcel J. 2002. The splicing regulator TIA-1 interacts with U1-C to promote U1 snRNP recruitment to 5′ splice sites. *EMBO J* **21:** 6882–6892.

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

Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, et al. 2005. Galaxy: A platform for interactive large-scale genome analysis. *Genome Res* **15:** 1451–1455.

Häsler J, Samuelsson T, Strub K. 2007. Useful "junk": *Alu* RNAs in the human transcriptome. *Cell Mol Life Sci* **64:** 1793–1800.

Hickey DA. 1982. Selfish DNA: A sexually transmitted nuclear parasite. *Genetics* **101:** 519–531.

Irimia M, Roy SW. 2008. Evolutionary convergence on highly conserved 3′ intron structures in intron-poor eukaryotes and insights into the ancestral eukaryotic genome. *PLoS Genet* **4:** e1000148. doi: 10.1371/journal.pgen.1000148.

Irimia M, Penny D, Roy SW. 2007a. Coevolution of genomic intron number and splice sites. *Trends Genet* **23:** 321–325.

Irimia M, Rukov JL, Penny D, Roy SW. 2007b. Functional and evolutionary analysis of alternatively spliced genes is consistent with an early eukaryotic origin of alternative splicing. *BMC Evol Biol* **7:** 188. doi: 10.1186/1471-2148-7-188.

Jaillon O, Bouhouche K, Gout J-F, Aury J-M, Noel B, Nowacki M, Serrano V, Porcel BM, Ségurens B, Mouël AL, et al. 2008. Translational control of intron splicing in eukaryotes. *Nature* **451:** 359–362.

Ke S, Zhang XHF, Chasin LA. 2008. Positive selection acting on splicing motifs reflects compensatory evolution. *Genome Res* **18:** 533–543.

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. doi: 10.1371/journal.pgen.0030203.

Lev-Maor G, Ram O, Kim E, Sela N, Goren A, Levanon EY, Ast G. 2008. Intronic *Alus* influence alternative splicing. *PLoS Genet* **4:** e1000204. doi: 10.1371/journal.pgen.1000204.

Lewis BP, Green RE, Brenner SE. 2003. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc Natl Acad Sci* **100:** 189–192.

Loftus BJ, Fung E, Roncaglia P, Rowley D, Amedeo P, Bruno D, Vamathevan J, Miranda M, Anderson IJ, Fraser JA, et al. 2005. The genome of the *Basidiomycetous* yeast and human pathogen *Cryptococcus* neoformans. *Science* **307:** 1321–1324.

Lynch M. 2006. The origins of eukaryotic gene structure. *Mol Biol Evol* **23:** 450–468.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290:** 1151–1155.

Lynch M, Conery JS. 2003. The origins of genome complexity. *Science* **302:** 1401–1404.

Lynch M, Scofield DG, Hong X. 2005. The evolution of transcription-initiation sites. *Mol Biol Evol* **22:** 1137–1146.

Mayeda A, Ohshima Y. 1988. Short donor site sequences inserted within the intron of beta-globin pre-mRNA serve for splicing in vitro. *Mol Cell Biol* **8:** 4484–4491.

McGuire A, Pearson M, Neafsey D, Galagan J. 2008. Cross-kingdom patterns of alternative splicing and splice recognition. *Genome Biol* **9:** R50. doi: 10.1186/gb-2008-9-3-r50.

Neafsey D, Galagan J. 2007. Positive selection for unpreferred codon usage in eukaryotic genomes. *BMC Evol Biol* **7:** 119. doi: 10.1186/1471-2148-7-119.

Ohta T. 2002. Near-neutrality in evolution of genes and gene regulation. *Proc Natl Acad Sci* **99:** 16134–16137.

Ohta T, Kimura M. 1971. Functional organization of genetic material as a product of molecular evolution. *Nature* **233:** 118–119.

Plass M, Agirre E, Reyes D, Camara F, Eyras E. 2008. Co-evolution of the branch site and SR proteins in eukaryotes. *Trends Genet* **24:** 590–594.

Puig O, Gottschalk A, Fabrizio P, Séraphin B. 1999. Interaction of the U1 snRNP with nonconserved intronic sequences affects 5′ splice site selection. *Genes & Dev* **13:** 569–580.

Roca X, Sachidanandam R, Krainer AR. 2005. Determinants of the inherent strength of human 5′ splice sites. *RNA* **11:** 683–698.

Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV. 2003. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr Biol* **13:** 1512–1517.

Roy SW, Gilbert W. 2005. Complex early genes. *Proc Natl Acad Sci* **102:** 1986–1991.

Roy SW, Hartl DL. 2006. Very little intron loss/gain in Plasmodium: Intron loss/gain mutation rates and intron number. *Genome Res* **16:** 750–756.

Roy SW, Irimia M. 2008. Intron mis-splicing: No alternative? *Genome Biol* **9:** 208. doi: 10.1186/gb-2008-9-2-208.

Roy SW, Penny D, Neafsey DE. 2007. Evolutionary conservation of UTR intron boundaries in *Cryptococcus*. *Mol Biol Evol* **24:** 1140–1148.

Sapienza C, Doolittle WF. 1981. Genes are things you have whether you want them or not. *Cold Spring Harb Symp Quant Biol* **45:** 177–182.

Schwartz S, Silva J, Burstein D, Pupko T, Eyras E, Ast G. 2008. Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. *Genome Res* **18:** 88–103.

Séraphin B, Kretzner L, Rosbash M. 1988. A U1 snRNA:pre-mRNA base pairing interaction is required early in yeast spliceosome assembly but does not uniquely define the 5′ cleavage site. *EMBO J* **7:** 2533–2538.

Sharp PM, Li WH. 1987. The codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* **15:** 1281–1295.

Siliciano PG, Guthrie C. 1988. 5′ Splice site selection in yeast: Genetic alterations in base-pairing with U1 reveal additional requirements. *Genes & Dev* **2:** 1258–1267.

Sorek R, Shemesh R, Cohen Y, Basechess O, Ast G, Shamir R. 2004. A non-EST-based method for exon-skipping prediction. *Genome Res* **14:** 1617–1623.

Staley JP, Guthrie C. 1999. An RNA switch at the 5′ splice site requires ATP and the DEAD box protein Prp28p. *Mol Cell* **3:** 55–64.

Tress ML, Martelli PL, Frankish A, Reeves GA, Wesselink JJ, Yeats C, Olason PI, Albrecht M, Hegyi H, Giorgetti A, et al. 2007. The implications of alternative splicing in the ENCODE protein complement. *Proc Natl Acad Sci* **104:** 5495–5500.

Urrutia AO, Ocaña LB, Hurst LD. 2008. Do Alu repeats drive the evolution of the primate transcriptome? *Genome Biol* **9:** R25. doi: 10.1186/gb-2008-9-2-r25.

Warnecke T, Parmley JL, Hurst LD. 2008. Finding exonic islands in a sea of non-coding sequence: Splicing related constraints on protein composition and evolution are common in intron-rich genomes. *Genome Biol* **9:** R29. doi: 10.1186/gb-2008-9-2-r29.

Zhang D, Rosbash M. 1999. Identification of eight proteins that cross-link to pre-mRNA in the yeast commitment complex. *Genes & Dev* **13:** 581–592.

Zheng C, Fu X-D, Gribskov M. 2005. Characteristics and regulatory elements defining constitutive splicing and different modes of alternative splicing in human and mouse. *RNA* **11:** 1777–1787.

Zhuang Y, Weiner AM. 1986. A compensatory base change in U1 snRNA suppresses a 5′- splice site mutation. *Cell* **46:** 827–835.

Received November 16, 2008; accepted in revised form July 9, 2009.

Complex selection on 5′ splice sites in intron-rich organisms

Manuel Irimia, Scott William Roy, Daniel E. Neafsey, et al.

Genome Res. 2009 19: 2021-2027 originally published online September 10, 2009
Access the most recent version at doi:10.1101/gr.089276.108