Association of Intron Phases with Conservation at Splice Site Sequences and Evolution of Spliceosomal Introns

Manyuan Long and Michael Deutsch

Department of Ecology and Evolution, University of Chicago

How exon-intron structures of eukaryotic genes evolved under various evolutionary forces remains unknown. The phases of spliceosomal introns (the placement of introns with respect to reading frame) provide an opportunity to approach this question. When a large number of nuclear introns in protein-coding genes were analyzed, it was found that most introns were of phase 0, which keeps codons intact. We found that the phase distribution of spliceosomal introns is strongly correlated with the sequence conservation of splice signals in exons; the relatively underrepresented phase 2 introns are associated with the lowest conservation, the relatively overrepresented phase 0 introns display the highest conservation, and phase 1 introns are intermediate. Given the detrimental effect of mutations in exon sequences near splice sites as found in molecular experiments, the underrepresentation of phase 2 introns may be the result of deleterious-mutation-driven intron loss, suggesting a possible genetic mechanism for the evolution of intron-exon structures.

Introduction

Most eukaryotic nuclear genes share the common structural feature that nuclear introns disrupt protein-coding regions. Studies of such gene structures have revealed that intron-exon structures were subjected to various evolutionary changes (Long, DeSouza, and Gilbert 1995; Gilbert, DeSouza, and Long 1997) which include intron loss, intron gain, and intron drift as major forms. An ample literature documents clear cases of intron loss in many gene families (e.g., Fink 1987; Charlesworth, Liu, and Zhang 1998), but concrete evidence for the other two forms of intron evolution is still being sought, and a few cases have been reported (Giroux et al. 1994; Stoltzfus et al. 1997; Logsdon, Stoltzfus, and Doolittle 1998; O'Neill et al. 1998). The evolutionary forces driving such evolutionary changes in intron-exon structures remain unexplained, despite extensive studies (Palmer and Logsdon 1991; Long, DeSouza, and Gilbert 1995; DeSouza, Long, and Gilbert 1996; DeSouza et al. 1996; Gilbert, DeSouza, and Long 1997; Long and DeSouza 1998). One challenge is that the distribution across species of introns in a given gene often provides little information. Recently arising genomic analysis, or database analysis (Long, DeSouza, and Gilbert 1995; Long, Rosenberg, and Gilbert 1995), may provide an efficient approach to characterization of the variation of eukaryotic gene structures.

Intron phases, the positions of introns between or within codons (Sharp 1981), are a conservative character of eukaryotic gene structures, because any phase change requires either compensatory double mutations or more complex molecular mechanisms. There are three intron phases: phase 0 introns are located between codons, phase 1 introns are located between the first and second nucleotides of a codon, and phase 2 introns are located between the second and third nucleotides. General functions or phenotypic effects of spliceosomal introns have not been reported except in exceptional cases where introns harbor regulatory elements or code for snoRNAs (Tycowski, Shu, and Steitz 1996; Long and DeSouza 1998). Remarkably, when a large number of introns in independent genes in the databases were analyzed, it was found that a dominant proportion of introns are of phase 0, showing that introns tend to keep codons intact, and the proportions of both phase 1 and phase 2 introns are significantly lower than random expectation (Fedorov et al. 1992; Long, Rosenberg, and Gilbert 1995; Tomita, Shimizu, and Brutlag 1996).

For overrepresented phase 0 introns, two alternative explanations were offered based on alternative hypothetical origins of introns. The direct explanation is that because primitive genes might code for short peptides, such as independent functional or structural units, the modern genes evolved greater complexity by assembling primitive genes in an intron-exon structure, and all introns were of phase 0. According to this, the exon theory of genes (Gilbert 1987), phase 1 and 2 introns were features derived by intron movement. Alternatively, overrepresented phase 0 introns could also be explained by insertion to preferred sites of genes, ''proto-splice sites'' (Dibb and Newman 1989), although so far no such sites that fit the distribution of intron phases have been found (Long et al. 1998). Although both approaches seek historical reasons for phase 0 overrepresentation, several interesting problems remain. Consider the hypothesis that phase 1 and 2 introns result from movement of phase 0 introns as a consequence of recombination of primitive genes. In this scenario, a one-nucleotide shift toward the 5′ codon would create a phase 2 intron, while a one-nucleotide shift toward the 3′ codon would yield a phase 1 intron. If shifts in either direction are equally likely, then the proportions of phase 1 and 2 introns should be similar. This prediction differs from the observed unequally probable phase distribution.

Since the introns have to be spliced before mature mRNAs are made from pre-mRNAs, the mechanism involved in the splicing process may be related to the distribution of intron phases. One important factor is

Key words: intron phase, splice site conservation, intron loss, evolution of intron-exon structures, natural selection.

Address for correspondence and reprints: Manyuan Long, Department of Ecology and Evolution, University of Chicago, 1101 East 57th Street, Chicago, Illinois 60637. E-mail: mlong@midway.uchicago.edu.

Mol. Biol. Evol. 16(11):1528–1534. 1999

how signals for recognizing splice sites are distributed and changed, a mutation parameter that may lead to the loss of introns. In the early 1980s, some conserved sequences were found adjacent to splice sites in both exons and introns (Mount 1982). Stephens and Schneider (1992) further characterized the distribution of conserved nucleotides in the splice sites of 1,800 human introns using logo analysis, a quantitative graphical display for information contents in the intron and exon nucleotide sites near the splice boundaries. A similar analysis was extended to other model species later (Long et al. 1998). These statistical analyses have located important nucleotide sites and elucidated their relative importance within and among the sites at splice boundaries and branching points.

A general picture of the molecular interactions between the nucleotides at splice sites and the components of spliceosomes, the machinery to recognize and cut introns accurately, has been described (Horowitz and Krainer 1994; Reed 1996). Using cross- linking in a mammalian system, the human HeLa cell line, Wyatt, Sontheimer, and Steitz (1992) and Sontheimer and Steitz (1993) identified interactions between the two nucleotides immediately upstream of intron 5′ donor sites and U5 snRNA. By suppressing point mutations in *Saccharomyces cerevisiae*, Newman and Norman (1991, 1992) demonstrated that Watson-Crick pairing between U5 snRNA and exon nucleotides near both 5′ and 3′ splice sites plays an essential role in splicing reactions, which was further confirmed to be general in a genomic analysis of splice sites in yeast (Long, DeSouza, and Gilbert 1997). Many point mutations immediately upstream of 5′ donor sites have been identified and shown to abort splicing (Rymond and Rosbash 1992; Aebi et al. 1987). All of these observations and experiments show interactions between spliceosomes and nucleotide sequences in the exon regions near splice sites; perturbation of these interactions would have deleterious effects on an organism.

Are the exon regions close to splice sites (donor and acceptor) related to the distribution of intron phases? We analyzed the phase distribution of spliceosomal introns and the degree of conservation of the exon sequences near splice sites in a large exon database that contains all recently published introns and exons in eukaryotic genes. (For the sake of convenience, we will simply use the word ‘‘intron’’ to refer to the spliceosomal intron). We report here a significant correlation between the distribution of intron phases and conservation of the splicing signals in the exon nucleotides near splice sites and explore the evolutionary implications of such a correlation for the evolution of eukaryotic genes.

Materials and Methods

Splice Site Sequence Databases and Exon Databases

Two types of databases were constructed from GenBank 106: (1) exon databases and (2) splice site sequence databases. The procedure for constructing exon databases was similar to that of Long, Rosenberg, and Gilbert (1995). Briefly, using an earlier computer program (Long, Rosenberg, and Gilbert 1995), we selected all the genes that contain known intron-exon structures as defined in the feature tables to form a raw database. Pseudogenes, genes with questionable protein lengths as defined by feature tables, and putative *Caenorhabditis elegans* genes identified by computer programs were discarded to avoid possible artifacts. Then we wrote a program to calculate various parameters such as the positions and phases of introns. In total, there are 16,989 genes in the raw database, each with defined intron positions and phases. To avoid possible bias brought about by uneven representation of different gene families, we used the GBPURGE program to purge all redundant sequences at a criterion of 20% amino acid identity: If two protein sequences shared an identity defined by fasta3 program (Pearson 1996) as greater than 20%, we kept the sequence with more exons and deleted the other sequence. We deleted all splice sequences for introns lacking consensus GT . . . AG, because such sequences may not be correct splice sequences (the small number of AT . . . AC class introns [Hall and Padgett 1994; Tarn and Steitz 1997] that were also excluded in the computation does not change the result of statistical analyses).

Conservation Measurement of Splice Sites

To measure the conservation of the exon sequences that flank the introns, we aligned all splice sequences in the splice sequence databases and calculated the percentages of four nucleotides, A, T, G, and C, in each of 10 sites upstream or downstream of the splice boundaries of donor and acceptor sites. We then used two approaches to measure conservation. The first one, a now standard approach for analyzing sequence conservation around splice sites, is an information measurement method proposed by Schneider et al (1986) and Schneider (1991) which has the advantage over conventional goodness-of-fit tests, such as the chi-square test, that it is additive across sites and free from the influence of sample size. The second approach is simply to calculate the percentage of the most conservative consensus sites.

Information Analysis

The information for each base is calculated according to the equation

$$R_s(l) = 2 - H(l) - e(n),$$

where $H(l) = -\Sigma\, f_i(l)\log_2 f_i(l)$, $f_i(l)$ being the frequency of nucleotide i (A, T, G, or C) at base l, and $e(n) = 3/(2n \log_e 2)$ is a term for correcting sample size n (the total number of splice sequences). In the case of the splice sequence database, $e(n)$ is negligible, since n is large. $R_s(l) = 0$ indicates no conservation, while $R_s(l)$ reaches the maximum of 2, when only one nucleotide is present, indicating the highest conservation.

Logo analysis (Schneider and Stephens 1990), a graphic description of sequence conservation, was used. Based on the additive property of $R_s(l)$, we calculated the total information by summing over all five con-

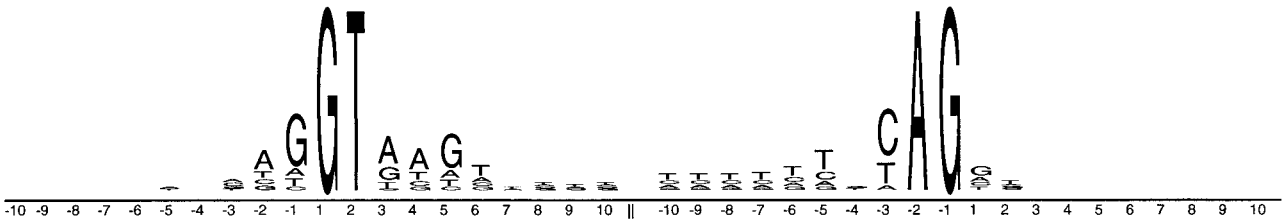

FIG. 1.—The logo of distribution of splice signals in independent eukaryotic genes. Only the exon and introns regions near splice sites are shown. The total height at each position is the ''information'' in that position; the information of GT at positions 1 and 2 at 5′ splice sites and that of AG at positions −1 and −2 at 3′ splice sites reach the highest information (2.0). The height of each letter represents its percentage in total bases counted in that position; the order of the letters in each position reflects the relative magnitudes of the proportions of these bases, from the highest to the lowest percentage.

served exon sites (positions −3, −2, and −1 at the 5′ splice site and positions 1 and 2 at the 3′ splice site):

$$R = \sum_{-1 \text{ to } -3} R_s(l_{donor}) + \sum_{1 \text{ to } 2} R_s(l_{acceptor}),$$

with the variance

$$V = 5\left(\sum_{\text{all nb}} P_{nb} H_{nb}^2 - (EH_{nb})^2 \right),$$

where P_{nb} is the probability $(n!/(na!nc!ng!nt!)\ P_a^{na} P_{mc}^{nc} P_g^{ng} P_t^{nt})$ for the particular combination n = na + nc + ng + nt (see Schneider et al. 1986). The maximum value of R is 10, and the minimum is 0. The statistical test for the significance of R values of three phases followed the method developed by Schneider et al. (1986) and Schneider (http://www-lecb.ncifcrf.gov/~toms/delila/ttest.html), which used a t-test $(t = (R_1 - R_2)/(S_1^2/N_1 + S_2^2/N_2)^{1/2})$ for comparison of two R values with small sample sizes, where S^2 and N are variance and sample size, respectively, and subscripts 1 and 2 indicate samples 1 and 2, respectively. In this investigation, the sample sizes for three intron phases are very large (larger than 3,000), and the standard normal distribution is used in the test.

Percentage of the Most Conservative Splice Sites

For conserved sequences at exon sites, we simply calculate the percentage of the G at the last nucleotide in the exon at the 5′ splice sites and the linked G at the first nucleotide position in the exon at the 3′ splice sites, because the G's at the two sites are generally the most conserved (Sontheimer and Steitz 1993; Reed 1996). This is designated the G|G site, where ''|'' represents an intron. The statistical significance of the differences in G|G sites between the three intron phases were tested using an approximate normal distribution (z-test; see Blank 1980), $z = (P_i - P_j)/S$, where $S = [pq(1/N_1 + 1/N_2)]^{1/2}$ and $p = (P_1 + P_2)/(N_1 + N_2)$, and P_1 and P_2 are counts of G|G in the samples with N_1 and N_2 individuals, separately. This test was used because of the large sample sizes (N_1 and N_2 are larger than 3,000). We also used a test of independence for a 2×2 table to test the significance of the percentages of G|G sites between different phases (Sokal and Rohlf 1995, section 17.6); results were similar to those of the z-test.

Results

The final exon database included 2,895 genes that were independent or quasi-independent and contained 17,044 introns. We calculated the proportions of the three intron phases in this database. The results were similar to those reported previously (Long, Rosenberg, and Gilbert 1995; Tomita, Shimuzu, and Brutlag 1996): 51% phase 0 introns, 27% phase 1 introns, and 22% phase 2 introns.

For the splice sequence database, we wrote a program to collect sequences of 10 nucleotides on both the intron and exon sides of each splice site (both donor and acceptor sites) from all genes for which defined intron-exon structures with intron sequences were available. Then, we used the splice sequences in the genes of the purged exon databases to form a raw splice sequence database for further analysis. After we deleted redundant sequences, we generated a final splice sequence database that contained 16,295 independent splice sequences. Intron phase proportions from this database are approximately equal to those calculated from the exon database (the differences are smaller than 1%).

The logo in figure 1 describes the distribution of conservation of the nucleotides near donor and acceptor splice sites for eukaryotic intron-containing genes in general. It shows that the highest conservation is within introns and that there is also a limited conservation in the adjacent exon sites, a similar situation to that described for human introns (Stephens and Schneider 1992) and those of other model organisms (Long et al. 1998). The logos for the total splice sequence database (fig. 1) and the three subdatabases (unpublished data) show a steady decline of information from position −1 to position −3 in the exon regions near donor sites (5′ splice sites) (after position −4, the information becomes indistinguishable from random noise). In the exon regions near acceptor sites (3′ splice sites), there is limited conservation in positions 1 and 2.

We conducted similar conservation analyses on the subdatabases of splice sequences for the three intron phases separately. Since the signals within intron regions are similarly high among three intron phases, we focus on the exon regions near the splice sites.

Table 1 shows the results of these calculations. Surprisingly, there is a significant correlation between the proportions and degrees of conservation among the three

Table 1
Intron Phase Proportion and Conservation

	Phase		
	0	1	2
Proportion (%)	51	27	22
Conservation			
Information	2.03 (0.58×10^{-3})	1.38 (1.18×10^{-3})	1.08 (1.34×10^{-3})
G|G percentage	41 (5.41×10^{-3})	35 (7.19×10^{-3})	28 (7.47×10^{-3})

NOTE.—The individual R_s(l) values in five exon positions (−3, −2, −1, 1, and 2) were as follows: for phase 0, 0.16, 0.55, 0.89, 0.30, and 0.13; for phase 1, 0.10, 0.18, 0.89, 0.14, and 0.07; and for phase 2, 0.07, 0.37, 0.47, 0.12 and 0.05. The standard errors for information contents are listed in parentheses, and the standard deviations for G|G percentages in parentheses were calculated as square root of $p(1 - p)N$; p and N are the proportion of G|G type sequences and the sample size, respectively. The differences in the G|G percentages between the three intron phases are significant at $P \gg 10^{-9}$. The sample sizes for calculating information contents of phase 0, 1, and 2 were 8,280, 4,395, and 3,609, respectively. The database contained 17,044 introns.

intron phases. The information amount for phase 0 introns is the highest, followed by the information amount for phase 1; the information amount for phase 2 is the lowest. The differences between the three phases are statistically very significant (all three z values are $\gg z_{P=10^{-9}} = 6.11$: $z_{p_0\text{-}p_1} = 529.1$; $z_{p_0\text{-}p_2} = 622.5$; $z_{p_1\text{-}p_2} = 186.8$).

Previous experiments indicated some direct interaction between consensus sequences and some component small nuclear ribonucleoproteins of spliceosomes (Steitz 1992). Similarity to the consensus sequence may have an important role in splice site recognition (Sontheimer and Steitz 1993; Reed 1996). We generated an exon consensus sequence for position −1 near the 5′ donor site and position 1 near the 3′ acceptor site as G|G, since these positions have high information amounts. We then calculated the percentage of actual sequences that are identical to this consensus. The results are listed in table 1, which shows, again, a correlation with the distribution of intron phases: phase 0 introns have the highest percentage of consensus-like sequences, phase 1 introns have an intermediate percentage, and phase 2 introns have the lowest percentage. The differences in percentages between the three phases are statistically very significant (all three z values are $\gg z_{P=10^{-9}} = 6.11$: $z_{p_0\text{-}p_1} = 6.59$; $z_{p_0\text{-}p_2} = 7.00$; $z_{p_1\text{-}p_2} = 13.4$).

Thus, the two different methods that measure the conservation of the splice sites lead to the same conclusion, that there is strong correlation between intron phase proportions and the degrees of conservation in exon splice sites.

To look for possible exceptions in different intron-containing organisms, we repeated the analysis of intron phase distribution and splice site conservation in those model organisms for which large numbers of introns and exons were sequenced. Using the same technique to build and purge overall database sequences at a criterion of 20% identity and using the same GenBank 106, we separated the databases of exons and splice sequence databases from *Homo sapiens, Mus musculus, Gallus gallus, Rattus norvegicus, Arabidopsis thaliana, Schizosacchromyces pombe*, and *Drosophila* (since the numbers of sequenced independent introns in individual *Drosophila* species are small, we analyzed the pooled data of all *Drosophila* species). A correlation similar to that of the overall databases was found between intron phase proportions and splice conservation in each of these organisms (table 2). Hence, the correlation shown in the overall databases is not an artifact of pooling different patterns, supporting our general analysis.

Table 2
Intron Phase and Conservation at Splice Sites in Six Model Organisms

	Phase		
	0	1	2
Human (4,620 introns)			
Percentage	46	33	21
Information (bits)	2.42 (0.002)	1.79 (0.003)	1.62 (0.005)
Mouse (1,565 introns)			
Percentage	45	36	19
Information (bits)	2.31 (0.007)	1.74 (0.009)	1.47 (0.016)
Rat (513 introns)			
Percentage	47	36	17
Information (bits)	2.54 (0.020)	1.95 (0.027)	1.14 (0.054)
Chicken (329 introns)			
Percentage	50	34	16
Information (bits)	2.45 (0.030)	2.23 (0.044)	1.61 (0.088)
Arabidopsis thaliana (8,214 introns)			
Percentage	57	22	21
Information (bits)	2.18 (0.001)	1.69 (0.003)	1.34 (0.003)
Schizosacchromyces pombe (1,387 introns)			
Percentage	44	29	27
Information (bits)	1.08 (0.007)	1.03 (0.012)	0.58 (0.011)
Drosophila (1,173 introns)			
Percentage	45	29	26
Information (bits)	1.43 (0.009)	0.900 (0.014)	0.73 (0.016)

NOTE.—"Percentage" refers to the proportions of three intron phases in various organisms. The standard errors for information contents are listed in parentheses.

Discussion

We observed that the proportion of an intron phase is correlated with the extent of conservation in the adjacent exon sequences. The sequences flanking phase 2 introns are the most variable. The two observations that lead to the same conclusion suggest that the observed correlations are not statistical artifacts but may implicate biological connections between the two parameters we

observed, the proportions of intron phases and the changes in splice sequences in exon regions.

The first interpretation is that the three classes of introns differ with regard to the recognition signaling role of the exon sequences immediately adjacent to the splice site. The exon sequences near the phase 2 introns are the least important to recognition by spliceosomes, so less splicing functional constraint leads to more variable sequences to meet the coding requirement in various genes, a situation similar to the neutral evolution of homologous genes (Kimura 1983). Functionally, the exon sequences flanking phase 0 introns, which have developed the most conservative use of nucleotides, are most important to splicing. This scenario is unlikely, because there is no indication that a spliceosome is ''aware'' of the reading frame (E. J. Sontheimer, personal communication).

The second interpretation is based on the empirical fact that mutations in the exon sequences flanking introns are often detrimental to splicing. The direct change in exon nucleotides at 5′ splice sites of phase 0 introns was found to yield reduced or abolished splicing in both yeast and rabbit genes (Felber, Orkin, and Hamer 1982; Seraphin and Rashbash 1990). Spontaneous mutations involving the GT dinucleotide at 5′ splice sites were found to inactivate normal splicing but activate cryptic splicing sites G|GT in human thalassemic genes (Treisman et al. 1982). The phases of the mutated introns in these observations included phases 0 and 2, suggesting that the detrimental effects of the mutations may not depend on the phases of introns. Thus, there could be stronger negative selection on the more variable exon boundaries, since they would more frequently generate deleterious changes. Therefore, the low frequency of phase 2 introns could be the consequence of deleterious-mutation-driven intron loss, since the exon regions flanking these introns are the most variable and create the highest deleterious mutation rates. This hypothesis, which differs from the first interpretation, offers an explanation for the correlations between intron phase distributions in independent eukaryotic genes and exon conservation.

While the consequence of the correlation between conservation of exon sequences flanking introns and intron phase proportions is clear, the molecular mechanism responsible for the variation of the conservation among the three intron phases remains unknown. A most attractive model, proposed by Fichant (1992), holds that the nucleotide composition in exon regions might depend on the balance between the conflicting demands of coding constraint and spliceosomal functional constraint. It is plausible that the variation in conservation reflects such a balance. However, one simple prediction from this theory, that nucleotides in wobble positions should meet the requirement for constraint better than first and second codon positions, is not supported by the information distribution at the five exon sites. For example, the amounts of information at position −1 for phases 0 and 1 are equal, although the −1 site in phase 0 introns is a wobble position, while the −1 site in phase 1 is the first codon position (table 1). Evidently, the balance between spliceosome and amino acid choice is more complex than this simple prediction. One possibility is that the exon sequences across phase 1 and 2 introns may be subject to more protein-coding constraint than the exon sequences near phase 0 introns, whose −1 nucleotide at the 5′ side is in wobble position. Thus, different genes would choose different amino acids in the boundary of phase 1 and 2 introns. This is not an unreasonable assumption, given the discovery that there are excess phase 0 introns, instead of phase 1 and 2 introns, in the boundary regions between protein modules, where the coding constraints may be lower than those in the regions inside the modules (DeSouza et al. 1998).

The uneven distribution of intron phases is likely a mixed consequence of historical events and natural selection. The ancient origin of introns may yield overrepresentation of phase 0 introns; subsequent intron drift could produce equal portions of phase 1 and 2 introns. Then, as a result of negative selection, phase 2 introns have been lost more quickly than phase 0 or 1 introns and thus become the least represented.

Compared with the previous search for historical reasons, the correlation found in this report sheds additional light on the evolution of the eukaryotic gene: the impact of evolutionary forces. Thus, the evolution of eukaryotic gene structures may not be entirely a process of neutral evolution. A more complete picture requires both historical and mechanistic approaches. We propose that deleterious-mutation-driven intron loss may have played a significant role in the evolution of eukaryotic gene structures, because variation in exon regions may, along with other factors that also affect splicing reactions (Sontheimer and Steitz 1993; Reed 1996), significantly affect recognition of introns by spliceosomes. The study reported in this paper may represent a new approach to dealing with the mechanisms responsible for the evolution of gene structures.

Finally, the roles that various evolutionary forces play in molecular evolution have mainly been concerned with interpreting variation in DNAs or proteins (Kimura 1983; Gillespie 1991; Li 1997). In particular, natural selection and genetic drift have become increasingly recognized as factors that shape the patterns of sequence variation (Kreitman and Akashi 1995). This study suggests that a new evolutionary force, deleterious-mutation-driven intron loss, may also control the evolution of gene structures and provides a useful conceptual extension of molecular evolution to the understanding of the origin and evolution of new gene structures.

Acknowledgments

We thank E. J. Sontheimer, C.-I. Wu, and J. Spofford for critical reading of the manuscript; W. Gilbert, M. Kreitman, P. A. Sharp, T. Nagylaki, C. Bergman, and the members of our laboratory for stimulating discussions; T. Schneider and R. Block for help in compiling the programs for the information analysis downloaded from Schneider's web sites (http://www-lecb.ncifcrf.gov/~toms); the David and Lucile Packard Foundation

for a Packard Fellowship in Science and Engineering; and the National Science Foundation for support.

LITERATURE CITED

AEBI, M., H. HORNIG, R. A. PADGETT, J. REISER, and C. WEISSMAN. 1987. Sequence requirements for splicing of higher eukaryotic nuclear pre-mRNA. Cell **47**:555–565.

BLANK, L. 1980. Statistical procedures for engineering, management, and science. McGraw-Hill, New York.

CHARLESWORTH, D., F. L. LIU, and L. ZHANG. 1998. The evolution of the alcohol dehydrogenase gene family by loss of introns in plants of the genus Leavenworthia (Brassicaceae). Mol. Biol. Evol. **15**:552–559.

DESOUZA, S. J., M. LONG, and W. GILBERT. 1996. Introns and gene evolution. Genes Cells **1**:493–505.

DESOUZA, S. J., M. LONG, R. J. KLELN, S. ROY, and W. GILBERT. 1998. Toward 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**: 5094–5099.

DESOUZA, S. J., A. STOLTZFUS, J. LOGSDON, M. LONG, W. FISHER, and W. MARTIN. 1996. Origin and evolution of introns. HMS Beagle: a BioMedNet publication. Issue 1 (Feb. 1) at http://hmsbeagle.com.

DIBB, N. J., and A. J. NEWMAN. 1989. Evidence that introns arose at proto-splice site. EMBO J. **8**:2015–2022.

FEDOROV, A., G. SUBOCH, M. BUJAKOV, and L. FEDOROVA. 1992. Analysis of nonuniformity in intron phase distribution. Nucleic Acids Res. **20**:2553–2557.

FELBER, B. K., S. H. ORKIN, and D. H. HAMER. 1982. Abnormal RNA splicing causes one form of alpha thalassemia. Cell **29**:895–902.

FICHANT, G. A. 1992. Constraints acting on the exon positions of the splice site sequence and local amino acid composition of the protein. Hum. Mol. Genet. **1**:259–267.

FINK, G. R. 1987. Pseudogenes in yeast? Cell **49**:5–6.

GILBERT, W. 1987. The exon theory of genes. Cold Spring Harbor Symp. Quant. Biol. **52**:901–905.

GILBERT, W., S. J. DESOUZA, and M. LONG. 1997. Origin of genes. Proc. Natl. Acad. Sci. USA **94**:7698–7703.

GILLESPIE, J. H. 1991. The causes of molecular evolution. Oxford University Press, New York.

GIROUX, M. J., M. CLANCY, J. BAIER, L. INGHAM, D. MCCARTY, and L. C. HANNAH. 1994. De novo synthesis of an intron by the maize transposable element. Proc. Natl. Acad. Sci. USA **91:**8507–8511.

HALL, L., and R. A. PADGETT. 1994. Conserved sequences in a class of rare eukaryotic nuclear introns with non-consensus splice sites. J. Mol. Biol. **239**:357–365.

HOROWITZ, D. S., and A. R. KRAINER. 1994. Mechanisms for selecting 5′ splice sites in mammalian pre-mRNA splicing. Trends Genet. **10**:100–106.

KIMURA, M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge, England.

KREITMAN, M., and H. AKASHI. 1995. Molecular evidence for natural selection. Annu. Rev. Ecol. Syst. **26**:403–422.

LI, W.-H. 1997. Molecular evolution. Sinauer, Sunderland, Mass.

LOGSDON, J. M. 1998. The recent origins of spliceosomal introns revisited. Curr. Opin. Genet. Dev. **8**:637–648.

LOGSDON, J. M. JR., A. STOLTZFUS, and W. F. DOOLITTLE. 1998. Molecular evolution: recent cases of spliceosomal intron gain? Curr. Biol. **8**:R560–R563.

LONG, M., and S. J. DESOUZA. 1998. Intron-exon structures: from molecular to population biology. Pp. 143–178 *in* R. S. VERMA, ed. Advances in genome biology Vol. 5A. JIA Press, Greenwich, Conn., and London.

LONG, M., S. J. DESOUZA, and W. GILBERT. 1995. Evolution of the intron-exon structure of eukaryotic genes. Curr. Opin. Genet. Dev. **5**:774–778.

———. 1997. The yeast splice site revisited: new exon consensus from genomic analysis. Cell **91**:739–740.

LONG, M., S. J. DESOUZA, C. ROSENBERG, and W. GILBERT. 1998. Relationship between ''proto-splice sites'' and intron phases: evidence from dicodon analysis. Proc. Natl. Acad. Sci. USA **94**:219–313.

LONG, M., C. ROSENBERG, and W. GILBERT. 1995. Intron phase correlations and the evolution of the intron/exon structure of genes. Proc. Natl. Acad. Sci. USA **92**:12495–12499.

MOUNT, S. M. 1982. A catalogue of splice junction sequences. Nucleic Acids Res. **10**:459–472.

NEWMAN, A. J., and C. NORMAN. 1991. Mutations in yeast U5 snRNA alter the specificity of 5′ splice-site cleavage. Cell **65**:115–123.

———. 1992. U5 snRNA interacts with exon sequences at 5′ and 3′ splice sites. Cell **68**:743–754.

O'NEILL, R. J. W., F. E. BRENNAN, M. L. DELBRIDGE, R. H. CROZIER, and J. A. M. GRAVES. 1998. De novo insertion of an intron into the mammalian sex determining gene, SRY. Proc. Natl. Acad. Sci. USA **95**:1653–1657.

PALMER, J. D., and J. M. LOGSDON JR. 1991. The recent origins of introns. Curr. Opin. Genet. Dev. **1**:470–477.

PEARSON, W. R. 1996. Effective protein sequence comparison. Methods Enzymol. **226**:227–258.

REED, R. 1996. Initial splice-site recognition and pairing during pre-mRNA splicing. Curr. Opin. Genet. Dev. **6**:215–220.

RYMOND, B. C., and M. ROSBASH. 1992. Yeast pre-mRNA splicing. Pp. 143–192 *in* J. R. BROACH, J. R. PRINGLE, and E. W. JONES, eds. The molecular and cellular biology of the yeast saccharomyces: gene expression. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

SCHNEIDER, T. D. 1991. Theory of molecular machines. II. Energy dissipation from molecular machines. J. Theor. Biol. **148**:125–137.

SCHNEIDER, T. D., and R. M. STEPHENS. 1990. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. **18**:6097–6100.

SCHNEIDER, T. D., G. D. STORMO, L. GOLD, and A. EHRENFEUCHT. 1986. Information content of binding sites on nucleotide sequences. J. Mol. Biol. **188**:415–431.

SERAPHIN, B., and M. RASHBASH. 1990. Exon mutations uncouple 5′ splice site selection from U1 snRNA pairing. Cell **63**:619–629.

SHARP, P. A. 1981. Speculations on RNA splicing. Cell **23**: 643–646.

SOKAL, R. R., and F. J. ROHLF. 1995. Biometry. 3rd edition. W. H. Freeman, New York.

SONTHEIMER, E. J., and J. A. STEITZ. 1993. The U5 and U6 small nuclear RNAs as active site components of the spliceosome. Science **262**:1989–1996.

STEITZ, J. A. 1992. Splicing takes a holiday. Science **257**:888–889.

STEPHENS, R. M., and T. D. SCHNEIDER. 1992. Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites. J. Mol. Biol. **228**: 1124–1136.

STOLTZFUS, A., J. M. LOGSDON, J. D. PALMER, and W. F. DOOLITTLE. 1997. Intron ''sliding'' and the diversity of intron positions. Proc. Natl. Acad. Sci. USA **94**:10739–10743.

TARN, W.-Y., and J. A. STEITZ. 1997. Pre-mRNA splicing: the discovery of a new spliceosome doubles the challenge. Trends Biochem. Sci. **22**:132–137.

TOMITA, M., N. SHIMIZU, and S. BRUTLAG. 1996. Introns and reading frames: correlation between splicing sites and their codon positions. Mol. Biol. Evol. **13**:1219–1223.

TREISMAN, R., N. J. PROODFOOT, M. SHANDER, and T. MANIATIS. 1982. A single-base change at a splice site in a beta 0-thalassemic gene causes abnormal RNA splicing. Cell **29**: 903–911.

TYCOWSKI, K. T., M. D. SHU, and J. A. STEITZ. 1996. A mammalian gene with introns instead of exons generating stable RNA products. Nature **379**:464–466.

WYATT, J. R., E. J. SONTHEIMER, and J. A. STEITZ. 1992. Site-specific cross-linking of mammalian U5 snRNP to the 5′ splice site before the first step of pre-mRNA splicing. Genes Dev. **6**:2542–2553.

ANTONY DEAN, reviewing editor

Accepted August 3, 1999