High Intron Sequence Conservation Across Three Mammalian Orders Suggests Functional Constraints

Matthew P. Hare and Stephen R. Palumbi†*[1]

*Biology Department, University of Maryland, College Park; and †Department of Organismic and Evolutionary Biology, Harvard University

Several studies have demonstrated high levels of sequence conservation in noncoding DNA compared between two species (e.g., human and mouse), and interpreted this conservation as evidence for functional constraints. If this interpretation is correct, it suggests the existence of a hidden class of abundant regulatory elements. However, much of the noncoding sequence conserved between two species may result from chance or from small-scale heterogeneity in mutation rates. Stronger inferences are expected from sequence comparisons using more than two taxa, and by testing for spatial patterns of conservation in addition to primary sequence similarity. We used a Bayesian local alignment method to compare approximately 10 kb of intron sequence from nine genes in a pairwise manner between human, whale, and seal to test whether the degree and pattern of conservation is consistent with neutral divergence. Comparison of the three sets of conserved gapless pairwise blocks revealed the following patterns: The proportion of identical intron nucleotides averaged 47% in pairwise comparisons and 28% across the three taxa. Proportions of conserved sequence were similar in unique sequence and general mammalian repetitive elements. We simulated sequence evolution under a neutral model using published estimates of substitution rate heterogeneity for noncoding DNA and found pairwise identity at 33% and three-taxon identity at 16% of nucleotide sites. Spatial patterns of primary sequence conservation were also nonrandomly distributed within introns. Overall, segments of intron sequence closer to flanking exons were significantly more conserved than interior intron sequence. This level of intron sequence conservation is above that expected by chance and strongly suggests that intron sequences are playing a larger functional role in gene regulation than previously realized.

Introduction

Many fewer genes have been found in the human genome than was anticipated. This apparent lack of genomic complexity might be explained in part by a high frequency of alternative splicing (Croft et al. 2000), but it also suggests that other gene regulatory mechanisms may be more diverse or elaborate than currently recognized. The dominant paradigm for eukaryotic gene regulation involves binding of transcription factors at *cis*-enhancer sequences, many of which are near the 5′ end of genes (Lewin 1997). However, our view of regulatory mechanisms may be limited by the low capacity for experimental methods to discover dispersed transcription factor binding sites or sequence domains with novel regulatory functions. The genome sequences now available make noncoding sequence comparisons across taxa a potentially rapid method for discovering functional constraints related to gene regulation.

Twenty-five percent of the human genome consists of introns, transcribed noncoding spacer sequences interleaved between amino acid coding exons within genes (Venter et al. 2001). Transcription qualitatively distinguishes introns from other noncoding sequences in the genome, but among transcripts, introns are also quantitatively abundant. On average in humans, 95% of the primary transcript length consists of intron RNA that is subsequently spliced out (Venter et al. 2001). If transcription is a prerequisite for certain kinds of regulatory function such as *trans*-acting RNA modifiers, then the abundance of introns make them a likely source of these transcripts (Mattick and Gagen 2001). To date, however, the only general patterns of sequence constraint documented for introns occur in small terminal splice junctions and 3′ end branch sites (Mount 1982; Lewin 1997), facultatively coding intron sequence in genes that are alternatively spliced, and the first (furthest 5′) intron of genes (Venter et al. 2001). Early analyses of intron sequence evolution suggested that, on average, introns evolve at comparable rates to pseudogenes and synonymous sites in coding sequences, sites where selective forces are believed to be absent or very low (Hughes and Yeager 1997; Li 1997). Assumptions of intron neutrality based on these generalizations have motivated increased use of intron variation for historical analyses (Palumbi 1996; Friesen 2000; Chen and Li 2001; Hare 2001).

Rapid progress on the genomic sequencing of model taxa and recent improvement of sequence alignment algorithms have made it possible to make large-scale comparisons of homologous intron sequences between distantly related pairs of taxa (Drosophila spp.: Bergman and Kreitman 2001; human/mouse: Wasserman et al. 2000; Jareborg, Birney, and Durbin 1999; Levy, Hannenhalli, and Workman 2001; Dermitzakis et al. 2002; *Caenorhabditis* spp.: Shabalina and Kondrashov 1999). Using a variety of alignment methods, these studies have found 12% to 28% of intron sequence to be highly conserved. Conserved blocks of noncoding sequence are presumed to reflect functional constraints when the comparison involves species so divergent that random nucleotide substitutions and insertion/deletion (indel) events should have erased most sequence similarity under a neutral model of evolution (Jareborg, Birney, and Durbin 1999). Human and mouse, for example, have diverged for 80 Myr (Kumar and Hedges 1998). Given a substitution rate of 0.0037 substitutions per site per Myr (Li 1997), neutral sequence divergence of 60% between human and mouse is expected (40% without correction for multiple

[1] Present address: Hopkins Marine Station, Stanford University.

Key words: phylogenetic footprinting, intron, noncoding sequence conservation, gene regulation, interspersed repeats.

E-mail: matt.hare@umail.umd.edu.

Mol. Biol. Evol. 20(6):969–978. 2003
DOI: 10.1093/molbev/msg111

hits), close to the 75% divergence expected for nonhomologous sequences.

Known transcription factor (TF) binding sites were shown to be overrepresented in conserved blocks of human/mouse intron sequence (Levy, Hannenhalli, and Workman 2001), and this ability of sequence comparisons to reveal conserved binding sites has been called ''phylogenetic footprinting'' (Tagle et al. 1988). However, even in well-studied genes, known TF binding sites represent a small proportion of the noncoding sequence conserved between human and mouse (Wasserman et al. 2000), implying the existence of many additional TF binding sites, the presence of novel regulatory features, or chance stretches of sequence similarity. Exploring the extent and pattern of noncoding sequence conservation and testing it against null models of neutral evolution is a critical step toward understanding the functional role these sequences play, if any, beyond TF binding.

There are several sources of potential ambiguity in the results from noncoding sequence comparisons. First, the observed level and pattern of conservation can be sensitive to the alignment methods used (Bergman and Krietman 2001). Second, any process that generates substitution rate heterogeneity among nucleotide sites, such as mutational ''cold spots,'' can produce blocks of conserved sequence across large divergence times (Clark 2001). Finally, the proportion of sequence blocks conserved by chance will be greater with comparison of less divergent sequences (Duret and Bucher 1997). Comparison of both more divergent sequences and more taxa can strengthen inferences of functional constraint from sequence conservation. In the example above, for example, the binomial probability of human–mouse identity (0.60) across 20 bases is 3.7×10^{-5}. If a third independent lineage is included in the comparison, the probability falls to almost 10^{-9}. Inferences of functional constraint are also strengthened when there is conservation of secondary sequence features such as spacing between conserved blocks of sequence (Ludwig et al. 2000; Kim 2001; Webb et al. 2002) or RNA folding free energy (Stephan and Kirby 1993; Kirby, Muse, and Stephan 1995; Leicht et al. 1995; Clark, Leicht, and Muse 1996).

Unfortunately, lineage-specific indel events in noncoding sequences generally prohibit useful multiple sequence alignments across large divergence times. Previous efforts to test for noncoding sequence constraints among three taxa have been based on multiple pairwise alignments using mismatch and gap penalty parameters (Dubchak et al. 2000; Dermitzakis et al. 2002). Here, we apply a powerful Bayesian alignment algorithm to generate and compare local gapless pairwise alignments of intron sequences from three species representing different mammalian orders. We show striking patterns of sequence conservation and conserved block spacing over and above random expectations, even assuming a moderate level of nucleotide substitution rate heterogeneity. The extent and distribution of this intron sequence conservation strongly suggests that these introns harbor novel functional domains that are likely to play a role in the regulation of gene expression.

Methods

One intron or partial intron from each of nine genes was sequenced from a balaenid whale and a phocid seal. Genes and introns were haphazardly selected based on favorable polymerase chain reaction (PCR) results with conserved primers (Lyons et al. 1997) and favoring introns $>$ 500 base pairs (bp) long. At each locus, 50 to 150 bp of flanking exon was sequenced in addition to the intron to help confirm orthology across taxa. Both strands of PCR products were either sequenced directly or after cloning. At least three clones were sequenced to control for *taq* misincorporation errors. The data from four genes included a complete intron sequence whereas introns from the remaining genes were missing a central or end portion of the intron (fig. 1). Sequences lacking the middle portion of the intron (*HEXB*, *CD40L*, *CAMK IV*) were analyzed as separate 5′ and 3′ loci for a total of 12 loci from 9 genes. Whale or seal sequences were Blasted (Altschul et al. 1997) to retrieve the orthologous human sequence. Whales, seals, and humans represent divergent orders of placental mammals that diverged from each other 80 to 90 million years ago (Kumar and Hedges 1998) and thus represent good candidates for tripartite sequence comparisons.

For each locus, the Web-based Bayesian Block Aligner (BBA, Wasserman et al. 2000, http://bayesweb.wadsworth.org/cgi-bin/bayes_align12.pl) was used with default parameters to compare the three taxa in every pairwise combination (human/whale [H/W], human/seal [H/S], seal/whale [S/W]). Genomically repetitive sequences were not removed or masked before alignment. The BBA employs the sampling algorithm of Sankoff (1972) which, instead of constraining the local alignment optimization with a gap penalty term, seeks gapless alignment blocks with at most $k-1$ interblock intervals. An alignment produced by this algorithm consists of a series of gapless blocks interspersed with interblock intervals that may have dissimilar sequence or may be missing homologous sequence in one or the other taxon. The default BBA implementation of this algorithm uses a PAM1 DNA similarity matrix and an uninformed prior distribution for k ($k \leq 20$), meaning that any number of conserved blocks less than 20 is considered equally likely. A recursive sampling of the many possible alignments in proportion to their exact joint posterior probability is used to calculate the probability that any given base in sequence one is aligned with a particular base in sequence two (Zhu, Liu, and Lawrence 1998). The PAM1 matrix defines an expectation of one nucleotide difference every 100 bp with transitions three times more common than transversions. This matrix imposes a conservative bias toward short blocks with high sequence similarity that is justified to minimize the potential for false positives under the Bayesian procedure (Zhu, Liu, and Lawrence 1998).

One of the benefits of this Bayesian alignment procedure is that the marginal posterior probabilities associated with each base reflect the uncertainty in k, the total number of conserved blocks. However, defining block boundaries based on the posterior probabilities is somewhat arbitrary. The default block definition implemented

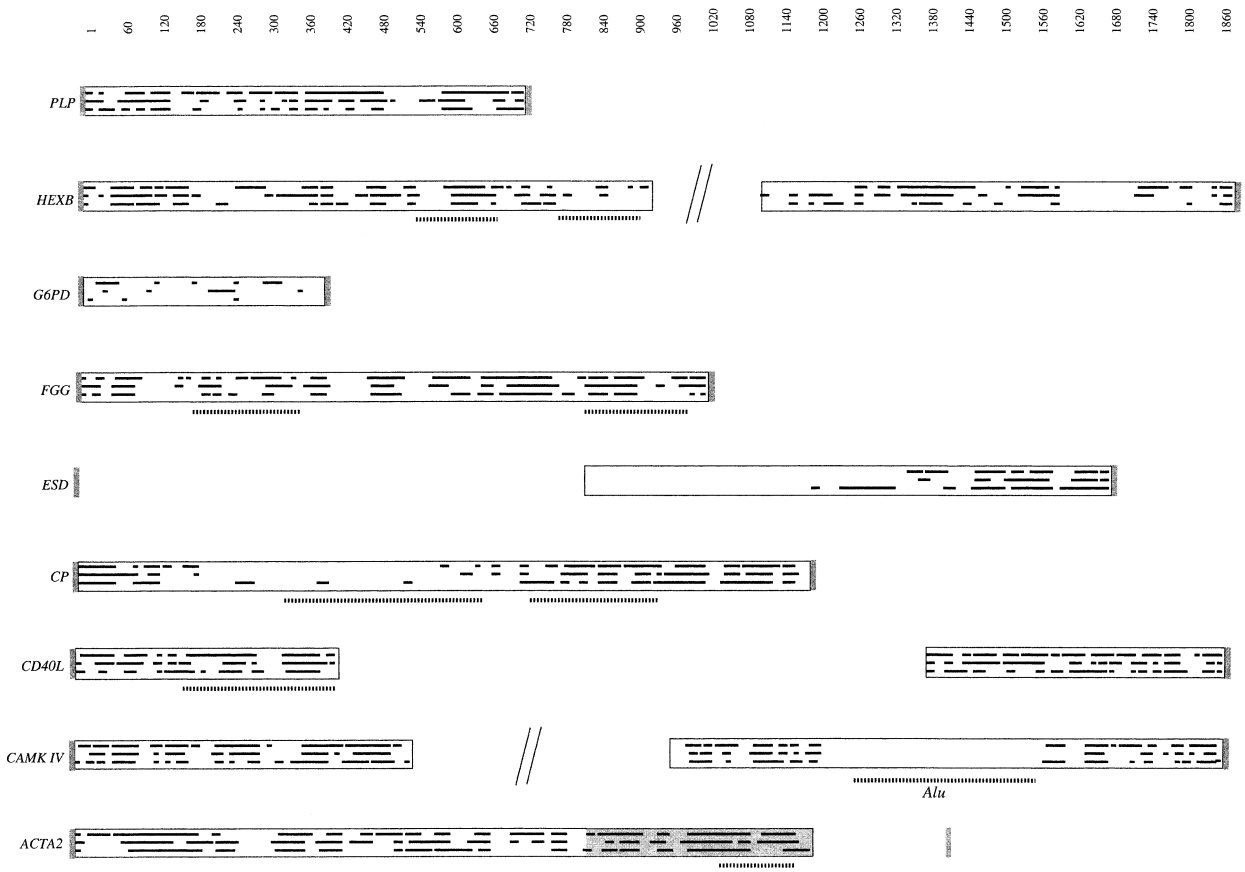

FIG. 1.—Three-taxon comparison of conserved gapless blocks in introns from nine genes. For each gene, the extent of significant whale/seal, whale/human and seal/human (from top to bottom, respectively) pairwise gapless sequence blocks is shown as black bars positioned horizontally according to the human intron sequence numbering shown at the top. Vertical gray bars represent exons, and boxes delimit the intron sequence that was analyzed. *ACTA2* has an alternatively spliced segment that is shaded. Double slashes indicate a large unanalyzed region out of scale with the numbering. Striped bars under boxes represent the position and extent of GMRE and alu repeats at each locus.

in the BBA is to start at the marginal posterior probability maximum (P_{max}) and expand the block to each side until a position is reached whose marginal posterior probability is $< 0.25(P_{max})$. This procedure is repeated for the remaining P_{max} until k blocks have been defined. In practice, this generated some very short blocks (3–8 bp) at some loci. To be conservative, we only analyzed blocks with $P_{max} \geq 0.5$. The shortest such block was 7 bp.

All pairwise sequence alignments were colinear, that is, blocks were sequential in nucleotide position for each taxon and were nonoverlapping. However, using the human sequence as a positional reference for comparing the H/W, H/S, and S/W pairwise results (e.g., fig. 2), nontransitive conflicts occurred when, for example, the same segment of seal sequence aligned with two positionally different segments in whale and human. The noncolinearity produced by these conflicts in the three-taxon comparison was corrected by removing whichever conflicting block had a lower P_{max} or by trimming a block if the conflict involved $\leq$4 bp. Less than 1% of the aligned sequence was removed for this reason.

For pairwise comparisons, the proportion of intron nucleotides within conserved blocks and identity between the two species was calculated for each locus relative to the total intron length analyzed for each species. To measure three-taxon conservation, segments of sequence where all three pairwise comparisons had positionally coincident gapless blocks were analyzed as three-taxon blocks (fig. 2). The number of nucleotides in three-taxon blocks and the number of identical nucleotides in those blocks were used to calculate proportions relative to the total length of analyzed human sequence. These comparisons and calculations were done using Microsoft Excel. The human-specific alu element was subtracted from the CAMK sequence length before calculating the proportion of conserved sequence.

Neutral sequence evolution without insertions or deletions was simulated for 1,000 bp using the HKY substitution model (Hasegawa, Kishino, and Yano 1985) assuming 0.0037 substitutions per site per MY, transition/transversion *ratio* = 2.0, 80 Myr divergence, and 39% GC content (empirical average for the 12 loci examined here) (Seq-Gen version 1.04, Rambaut and Grassly 1997). Gamma-distributed substitution rate heterogeneity among sites was incorporated in some simulations.

Results

Intron sequence lengths collected from whale and seal ranged from 342 to 1,507 bp per gene for a total of

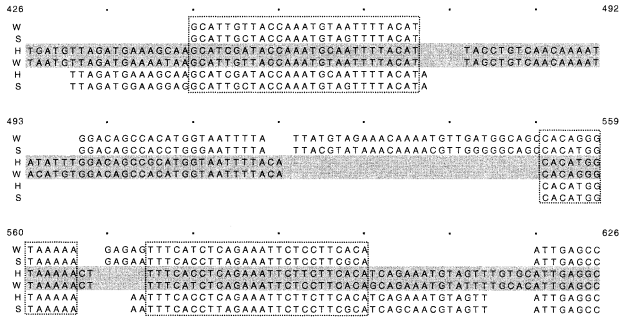

FIG. 2.—Overlap of pairwise gapless sequence blocks from whale/seal (W/S), human/whale (H/W), and human/seal (H/S) comparisons arranged from left to right and wrapping, positioned according to the human sequence numbering in *ACTA2*. The middle pairwise blocks (H/W) are shaded for visual clarity. Three-taxon blocks are boxed.

8,088 bp in whale and 8,650 bp in seal (GenBank accession numbers AY196798 to AY196819). Orthologous human intron sequence analyzed had a total sequence length of 9,120 bp (table 1).

Overall, 163 three-taxon blocks were identified (e.g., fig. 2) containing a total of 2,978 bp. These conserved blocks accounted for 34% of the total human sequence (SD across genes = 15). Slightly less intron sequence was in conserved blocks *and* showed identity across the three species (28 ± 12%). In other words, the BBA identified highly conserved blocks within which only 14% of nucleotides differed among species on average (SD across genes = 5). Using three-taxon block sequences to calculate Kimura 2-parameter distances between each pair of species showed no significant differences in substitution rate. The median length of gapless three-taxon blocks was 17 bp (table 2). Median block lengths were smaller than the mean in the skewed block length distributions (fig. 3).

For comparison to previous studies comparing two species, the proportion of sequence in two-taxon blocks *and* showing identity, averaged across taxa with different amounts of total intron sequence, was 28%–66% per locus (grand mean = 47%, SD = 13; table 1). The median block lengths for the two-taxon comparisons, 22 to 27 bp (table 2), were significantly larger than for three-taxon blocks (nonparametric median test, $P < 0.0001$).

The proportion of nucleotides contained in three-taxon blocks did not differ significantly among taxa but showed significant heterogeneity among the twelve loci (analysis of variance [ANOVA] $F = 20.7$, $P < 0.0001$). The heterogeneity among loci was not correlated with either GC content or total intron size in humans (Spearman, $P > 0.05$; table 1), but was correlated with intron position ($P \leq 0.05$). In addition, the mean GC content of three-taxon blocks across genes was not significantly different than for the total sequence analyzed (paired *t*-test, $P > 0.5$).

Simulations

We simulated sequence evolution along one kb of DNA to test whether the observed extent of intron sequence conservation is consistent with a simple pattern of neutral evolution. The model of evolution was parameterized using empirical averages and did not include insertions or deletions. After applying a homogeneous rate of substitution among sites to simulate the divergence of three sequences from a common ancestor, BBA analysis revealed only one three-taxon block 18 nucleotides long (0.02% of total length), far below the observed level of intron sequence conservation. Because mutational processes can produce substitution rate heterogeneity and increase the number and extent of conserved blocks, we also simulated neutral evolution with a level of substitution rate heterogeneity estimated previously for vertebrate noncoding DNA (gamma shape parameter of 2.0, Lum et al. 2000). A total of 10 three-taxon blocks (16% of total sequence length) resulted from BBA analysis, with a mean block length of 46 bp (range 8 to 109 bp). The three

Table 1
Features of Introns Studied and the Proportion of Total Sequence Contained in Pairwise and Three-Taxon (2-Way and 3-Way) Gapless Blocks *and* Showing Identity

Locus	Position in Gene	Total Human Intron Length	Mean Sequence Length Analyzed[a]	% GC Content		% Sequence	
				Analyzed Intron	3-Way Blocks	2-Way Identity[b]	3-Way Identity[c]
CD40L	1	1860	870	37	38	66	45
ACTA2[d]	2	1422 (556)	869	34	36	50	30
PLP	2	697	709	49	47	55	33
CAMK IV[d]	6	2514 (291)	1036	34	29	60	38
FGG	6	980	939	42	42	53	29
HEXB	6	8186	1598	35	33	42	19
G6PD	7	364	328	60	66	28	12
ESD	7	1680	622	30	29	32	16
CP	8	1497	1200	40	33	40	27
Mean				38.4[e]	36.3[e]	47	28

[a] Sequence length analyzed is averaged across all three species.

[b] Pairwise proportions are means of six estimates per locus; three two-taxon comparisons, each yielding two proportions because taxa have different total sequence lengths.

[c] Three-taxon proportions are calculated relative to the human sequence total length.

[d] The alternatively spliced portion of *ACTA2* and the alu element in human *CAMK IV* (lengths in parentheses) were not included in the sequence length analyzed or in calculations of GC content and percent conservation.

[e] Weighted mean.

Table 2
Block Length Distribution Parameters from Pairwise and Three-Taxon Comparisons

Comparison	Min–Max	Mean	Median	Mode
Human × seal	7–126	28.7	22	13
Human × whale	7–99	31.8	25	19
Whale × seal	8–160	36.3	27	19
H × W × S	2–76	20.5	17	13

NOTE.—Distributions were not significantly different from lognormal (Kolmogorov-Smirnov $P > 0.05$).

pairwise comparisons showed 26% to 33% of simulated sequence in conserved blocks. Thus, the total amount of sequence conservation produced by this simulation was roughly half that observed between human, whale, and seal.

Spatial Patterns of Conservation

We tested for spatial patterns in the degree of sequence conservation by calculating a conservation index at each nucleotide site across the three pairwise alignments. At each site the index ranged from 0 to 6, depending on (1) whether the site was part of a gapless block and (2) whether the site was identical for each of the three two-taxon comparisons. For six of nine genes where we have sufficient data, there was no significant difference in levels of conservation between 330 bp at the 3′ and 5′ intron ends (paired *t*-test, $P > 0.05$). We then combined data from the 3′ and 5′ ends, as well as from the 5′ end of *ACTA2* (the 3′ end is alternatively spliced). In these combined data, the 165-bp intron edge segments were significantly more conserved than the adjacent interior 165-bp segments (paired *t*-test, two-tailed $P = 0.04$). Although there was a great deal of scatter, the conservation index scores decreased with increasing distance from the closest exon (fig. 4). Linearity of the function could not be rejected ($P > 0.25$; Zar 1984, p. 278), and the linear function was significant even though conservation varied considerably among loci ($P = 0.02$, $r^2 = 0.04$, $n = 124$).

Another spatial feature of blocks is their distance from one another—the interblock interval. Interblock interval lengths might vary between taxa because of insertion and deletion events, or they might be constrained by selection if the relative spacing of conserved blocks is important for function. In two-taxon comparisons, after excluding interblocks where one species had contiguous blocks (the other had a unique insertion), there was a total of 429 pairwise interblock intervals. Of these, 103 (24%) were identical in length between the two species. In three-taxon comparisons, 27 out of 147 interblock intervals (18%) had identical length in all three taxa. In every comparison the distribution of interblock intervals was approximately lognormal (Kolmogorov-Smirnov $P > 0.05$) with a mean greater than the median. To try to pinpoint interblock intervals with less variation across taxa than average, we plotted mean interblock interval length among taxa against the standard deviation of those lengths (fig. 5). As expected, the standard deviation generally increased for longer interblock intervals.

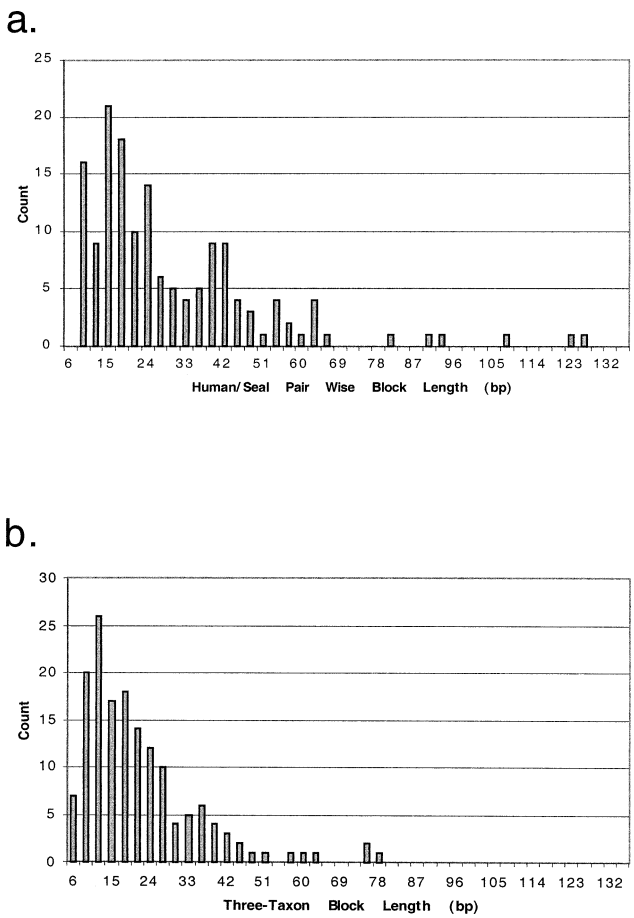

FIG. 3.—Frequency histogram of conserved gapless block lengths for the human-seal pairwise comparisons (*a*) and the three-taxon com parisons (*b*).

Conservation of interblock intervals was most pronounced along the abscissa, where moderately long interblock lengths showed relatively low standard deviations across taxa. While a model of indel evolution is required to generate a priori expectations for the degree of interblock interval conservation expected by chance, the pattern observed in human/seal comparisons (fig. 5*a*, open points) suggests ad hoc criteria defining potentially constrained interblock intervals. Three-taxon comparisons with mean interblock length among taxa > 40 bp and SD < 5 are highlighted as open points in figure 5*b* and described in table 3. Most of the highlighted points in figure 5*a* and *b* correspond to the same sequence segments. Conservation of interblock length is particularly pronounced in two genes, *ACTA2* and *FGG*, with four and five conserved interblock intervals, respectively. Most shorter interblock intervals also were conserved in these genes.

Genomic Uniqueness of Conserved Blocks

It is standard practice to mask repetitive elements prior to sequence comparison so that they do not cause spurious alignments, and because these elements are

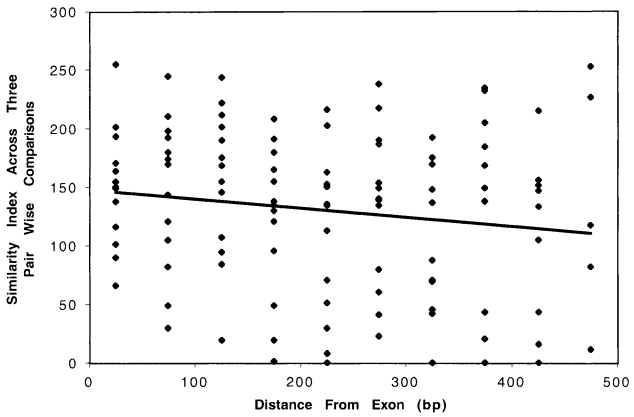

FIG. 4.—Conservation index for the 5′ and 3′ end of each intron summed within 50-bp nonoverlapping windows and plotted relative to the distance from the closest exon. The decrease in conservation with distance from exon is significant ($P = 0.02$) but explains only 4% of the variation.

perceived as ''junk'' pseudogene DNA. Our alignment and comparison methods are not disrupted by the presence of repetitive elements, so we aligned sequences without masking known elements and retrospectively analyzed the pattern and degree of sequence conservation within and outside of general mammalian repetitive elements (GMREs). When GMREs are found at the same locus in multiple mammalian orders (orthology), they most likely stem from an element radiation that predated the mammalian radiation (Smit and Riggs 1995; Gilbert and Labuda 1999).

To identify known repetitive elements in the intron sequences we used RepeatMasker (version 07/16/2000 using the ''other mammal'' subgroup of Repbase version 03/30/2000). In the intron sequences we analyzed, this search identified 30 interspersed repeats (SINEs, LINEs, and MaLR elements), 12 simple or low complexity repeats, and one tRNA. For the purpose of calculating the proportion of conserved sequence, GMREs found in only one species by RepeatMasker were considered present at orthologous positions in other species if a conserved block was found within any part of the element by the BBA. The total length of all interspersed repetitive elements was 1,910 bp, and that of orthologous GMREs was 1,527 bp, 20% and 16% of the analyzed human intron sequence length, respectively (fig. 1). Interspersed repeats found in only one species included one *Alu* element in human (fig. 1), one MAR1- SINE and one tRNA in seal, and one bovine-specific SINE element in whale. Among the five loci that contained GMREs, two results indicate that the level of conservation was similar in GMREs and the surrounding unique sequence. The proportion of three-taxon identities contained within repetitive elements ranged from 10% to 56% per locus, and was correlated with the total proportion of GMRE nucleotide sites at each locus (Spearman $r = 0.90$, $P = 0.037$). Also, the average proportion of three-taxon identities in nonrepetitive intron sequence (total sequence analyzed minus all interspersed elements) was 29% (SD = 12) for three-taxon blocks and 50% (SD = 14) for pairwise blocks, no different than results using the total sequence.

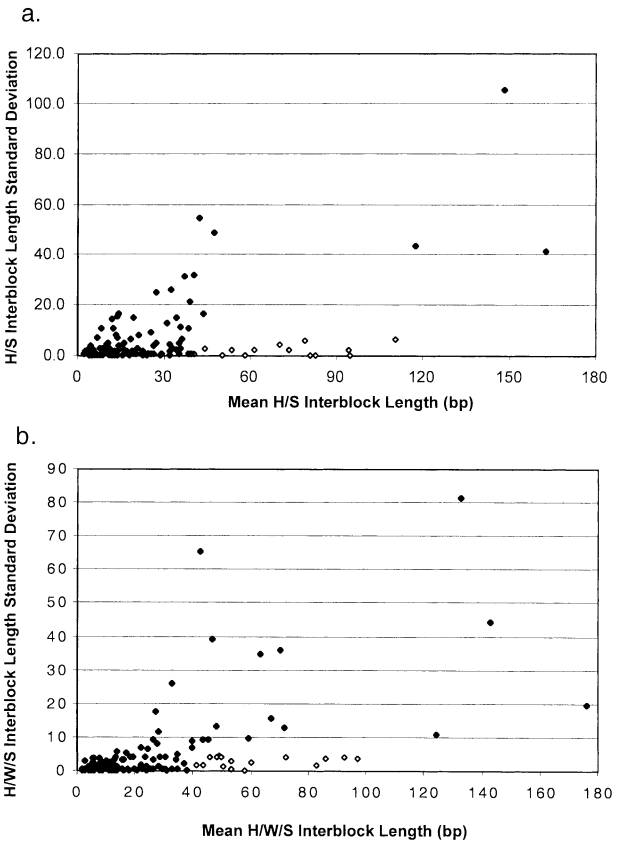

FIG. 5.—Mean length of interblocks relative to standard deviation of interblock lengths among taxa in the H/S pairwise comparisons (*a*) and the three-taxon comparisons (*b*) for all loci. Interblock length is the number of human nucleotides between consecutive blocks. Open symbols highlight the most highly conserved interblock lengths that are further described in table 3.

To test the genomic uniqueness of conserved sequences, human sequence from conserved three-taxon blocks outside of GMREs was used to query human GenBank with MEGABLAST. The few matches found outside homologous genes generally had probabilities ≥ 0.05 of occurring by chance, indicating that nonrepetitive conserved intron sequences are unique in the genome, not represented at moderate frequency as expected if motifs were important for coregulation of several genes.

Discussion

Our analysis has uncovered an unexpectedly high degree of conservation in intron sequence, manifest by high levels of primary sequence identity and in the conserved spacing of gapless sequence blocks. To strengthen these sequence-based inferences, we compared intron sequences across three mammalian orders instead of relying on the standard two-taxon application of phylogenetic footprinting. With the comparison of three taxa in this study, the proportion of conserved sequence found between two taxa averaged 47%, and that among three taxa was 28%. The 40% reduction in conserved sequence among three versus two taxa suggests that pairwise results are inflated because either (1) a large fraction of sequence was conserved by chance or (2) a large proportion of

Table 3
Interblock Interval Lengths Conserved Between Three-Taxon Blocks in Human, Whale, and Seal

Locus	Interblock Position in Locus[a]	Human Sequence Position[b]	Human Interblock Length	Whale Interblock Length	Seal Interblock Length	Repetitive Element Overlap[c]
ACTA2	6	471	82	85	82	
ACTA2	9	643	45	44	42	
ACTA2	10	710	96	88	93	
ACTA2[d]	**16**	**1126**	**50**	**52**	**50**	**LINE/L2**
CAMK IV-L	9	374	43	41	40	
CAMK IV-R	9	822	68	76	73	
CP	5	812	51	44	52	LINE/L1
FGG	3	208	95	101	95	SINE/MAR1
FGG	7	422	58	58	58	
FGG	9	610	90	84	84	
FGG	12	853	56	51	52	SINE/MIR
FGG	14	990	63	58	60	SINE/MIR
HEXB-L	7	466	51	51	44	
HEXB-R	1	381	132	129	112	
HEXB-R	6	621	54	46	48	
PLP	7	307	46	50	42	
PLP	15	706	54	53	53	

[a] Interblock position is based on the 5′ to 3′ numbering of three-taxon interblock intervals.
[b] Human sequence position listed is the 3′ end of the preceding block.
[c] The identity of repetitive elements overlapping conserved interblock intervals.
[d] The boldface *ACTA2* interblock is within an alternatively spliced segment of the intron.

functionally constrained intron sequence evolved in a lineage-biased manner. Our simulations show that moderate rates of substitution rate heterogeneity, expected to result in part from mutational processes, can explain much of the conserved sequence observed in pairwise and three-taxon comparisons under a strictly neutral model of sequence evolution without indels. As a result, blocks of noncoding sequence conserved over long divergence times do not necessarily indicate selective constraints, even when observed across more than two taxa. Nonetheless, our simulations indicate that typical levels of substitution rate heterogeneity in noncoding sequence can explain only about half of the intron conservation observed here. The ''excess'' sequence conservation, taken together with the finding of conserved spacing between some sequence blocks, strongly suggests that intron sequences are playing a larger functional role than previously realized. If 28% of intron sequence is evolutionarily constrained on average, as shown here, and if introns constitute 25% of the entire genome, this suggests that over 1/16 of the genome may be playing a hidden functional role.

Spatial Patterns of Conservation

Two kinds of spatial regularity were found in the pattern of sequence conservation across three mammalian taxa. First, the 5′ and 3′ ends of introns, adjacent to exons, were more highly conserved than the adjacent interior portion of introns. This pattern was observed on a sequence length scale too large to be explained solely by intron splice junctions that extend only 6 to 10 bp into the 5′ and 3′ ends of introns, or by the 3′ branch site that involves 7 bp at a distance of 18–40 bp from the 3′ end (Lewin 1997). Also, stronger conservation of intron edges in this sample was not caused by repetitive elements, because levels of conservation were similar in repetitive and unique sequence (see also Jareborg, Birney, and Durbin 1999). These results corroborate a similar trend reported for human–mouse intron comparisons by Jareborg, Birney, and Durbin (1999).

The second pattern of spatial conservation found in this study involved the spacing between conserved gapless blocks. Experimental studies have demonstrated that transcription factor binding sites are often co-localized within 5′ regulatory regions (Wingender et al. 2001) and position effects occur between regulatory elements and transcriptional units (Willoughby, Vilalta, and Oshima 2000). If indel positions are random, then over time, interblock intervals should diverge in size among taxa, particularly if indel rates are lineage-specific (Petrov et al. 2000). Longer interblock intervals are expected to diverge faster than short intervals for a given indel rate. Chance convergence on similar interblock spacing, perhaps as a result of indels that ''cancel out,'' is much less likely in three-taxon comparisons than in two-taxon comparisons. Thus, the high degree of three-taxon interblock length conservation in *ACTA2* and *FGG* suggests spatially coordinated functions among blocks. Alternative explanations are possible, but the paucity of data on rates and patterns of indel evolution make it difficult to construct meaningful tests (Saitou and Ueda 1994; Ludwig et al. 2000; Kim 2001).

Interspersed Repetitive Elements

Assuming that the interspersed nature of repetitive elements prevents homogenization by concerted evolution, most orthologous GMRE sequences are pseudogenes that should diverge between taxa at a neutral rate (Dermitzakis et al. 2002). Phylogenetic footprinting provides a test of the null hypothesis that the degree of sequence conservation, and perhaps functional constraint, is equivalent in unique and GMRE noncoding DNA. This hypothesis could not be rejected for the introns studied here. Unique

and GMRE sequences in introns evolve similarly; either they are both evolving under selective constraints or they are both affected by neutral mechanisms preserving sequence similarity.

The Link Between Sequence Conservation and Function

Our simulations demonstrated that the amount of pairwise noncoding sequence conservation observed in most studies is expected to result from a random pattern of nucleotide substitution with moderate rate heterogeneity. Mutation rates are known to vary locally and be dependent on sequence context, such as the tendency for C to mutate to T in CpG dinucleotides (Nachman and Crowell 2000; Chen and Li 2001). Spatial variation in selective constraints can also cause substitution rate heterogeneity, however. Distinguishing mutational and selective causes of rate heterogeneity will be much more challenging than measuring the degree and pattern of conservation. Progress will require more informed models of mutationally induced rate heterogeneity for substitutions and indels, perhaps achievable through studies of experimental evolution or pseudogene variation (Petrov 2002). Ultimately, although computational methods can be used to demonstrate that conserved noncoding DNA is enriched for known transcription factor binding sites (Wasserman et al. 2000; Levy, Hannenhalli, and Workman 2001), experimental approaches will be required to confirm the functional significance of particular blocks of conservation.

What functional roles are conserved introns likely to be playing? Recent estimates of the frequency of alternative splicing in the human genome indicate that more than 50% of genes have two or more alternatively spliced transcripts (International Human Genome Sequencing Consortium 2001), although most of this variation is probably due to exon skipping rather than facultative coding of intron sequences in different tissues (Croft et al. 2000). Evidence for alternative splicing exists for only one intron studied here, *ACTA2* (fig. 1). The strictly noncoding portion of the *ACTA2* intron had 37% of sites in three-taxon blocks, whereas the facultatively coding portion was 49% conserved. This contrast suggests that if alternative splicing were responsible for much of the overall intron conservation, then the higher conservation of intron edge sequences would be more pronounced.

Despite the historical allusion to sites of DNA binding made by the term phylogenetic ''footprinting,'' transcription factor binding sites constitute only the most familiar possibility for regulatory function that can be revealed through sequence comparisons. For example, genes encoding nucleolar proteins contain various small nucleolar RNAs transcribed from their introns (Rebane et al. 1998; Rebane and Metspalu 1999). The fact that introns are co-transcribed with their home exons make these noncoding sequences the most likely candidates for coordinated expression of *trans*-acting RNA factors that could modulate transcription or facilitate gene-to-gene interactions (Mattick and Gagen 2001). Several studies have found similar levels of sequence conservation in intergenic versus intron sequence, however, indicating that the link between transcription and noncoding sequence function may be weak (Jareborg, Birney, and Durbin 1999; Shabalina and Kondrashov 1999; Bergman and Krietman 2001).

Comparison with Other Studies

Several factors make it difficult to compare results from this sample of introns with previously reported patterns of noncoding sequence conservation. Most important, the various sequence alignment methods used can affect results (Bergman and Kreitman 2001), and only Wasserman and colleagues (2000) used the powerful but computationally demanding Bayesian Block Aligner that we used here. Second, some previous studies examined a nonrandom sample, focusing on noncoding regions in genes with known tissue-specific regulation or with experimentally determined *cis*-regulatory activity (Wasserman et al. 2000; Bergman and Kreitman 2001). Third, our small sample of introns may not be genomically representative because of limited chromosome sampling or other inadvertent biases. For example, the intron examined in *CD40L* is the first (furthest 5′) intron in that gene, where a higher density of regulatory sequences often reside (Levy, Hannenhalli, and Workman 2001; Venter et al. 2001). After removal of *CD40L* from this data set, the amount of sequence conservation was still very high, with proportions of identity in pairwise and three-taxon blocks averaging 45% (SD = 0.11) and 25% (SD = 0.09) of intron sequence length, respectively. Regardless of whether this small sample of introns proves to be genomically representative, our results indicate an extraordinary level of sequence conservation and presumed functional constraint in otherwise unremarkable and unstudied introns.

Implications and Conclusions

Evidence for widespread but variable evolutionary constraints in noncoding sequence has immediate repercussions for the use of these loci in population genetic and historical inferences made under the assumption of neutral evolution. Introns have been particularly attractive targets for analyses based on neutral theory because their positional homology within genes is phylogenetically conserved (Stoltzfus et al. 1997; Davidson et al. 2000) and flanking exons support the design of PCR primers (Palumbi 1996; Hare 2001). If 30% of intron sequence positions are not free to vary on average, then substitution rate estimates may be altered severalfold (Palumbi 1989). Moreover, substitution rates will vary dramatically from intron to intron, depending on the proportion of constrained sequence. Noncoding sequence constraints imply that some nucleotide substitutions in these regions will be slightly deleterious. Background selection against these variants is expected to reduce intraspecific intron variability below that expected for neutral nuclear sequences, disrupting coalescent predictions of the three-times rule (Palumbi, Cipriano, and Hare 2001).

When seeking neutral variants for historical and evolutionary analyses, the most predictable source may not be

in noncoding DNA, but rather in protein-coding regions. Synonymous sites are relatively neutral and are expected at a maximum frequency of 25% within protein-coding sequences (Li 1997). Although introns are desirable markers because they have a higher density of silent sites on average (this study), the location of those sites is unpredictable without phylogenetic comparisons. Also, general patterns of nucleotide and indel substitution are relatively uncharacterized for noncoding DNA. Nonetheless, it may prove desirable to use phylogenetic footprinting to select noncoding DNA markers with relatively low selective constraints for use in historical analyses. Comparisons made here indicate that, for mammals, the seventh introns in ESD and G6PD are relatively free of selective constraints.

Acknowledgments

Tissue samples or DNA were kindly provided by P. Best, R. Bonde, M. Brown, A. Dizon, S. Krauss, W. McLellan, R. Slade, and M. Wainstein. Thanks to A. M. Reich, N. Sherman, and S. Beck for data collection. Comments made by D. Rand and two anonymous reviewers helped improve this manuscript. This research was supported by Environmental Protection Agency grant R827110 to S.R.P.

Literature Cited

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped Blast and PSI-Blast: a new generation of protein database search programs. Nucleic Acids Res. **25**:3389–3402.

Bergman, C. M., and M. Kreitman. 2001. Analysis of conserved noncoding DNA in Drosophila reveals similar constraints in intergenic and intronic sequences. Genome Res. **11**: 1335–1345.

Chen, F., and W. Li. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am. J. Hum. Genet. **68**:444–456.

Clark, A. G. 2001. The search for meaning in noncoding DNA. Genome Res. **11**:1319–1320.

Clark, A. G., B. G. Leicht, and S. V. Muse. 1996. Length variation and secondary structure of introns in the *Mlc1* gene in six species of Drosophila. Mol. Biol. Evol. **13**:471–482.

Croft, L., S. Schandorff, F. Clark, K. Burrage, P. Arctander, and J. Mattick. 2000. ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. Nat. Genet. **24**:340–341.

Davidson, H., M. S. Taylor, A. Doherty, A. C. Boyd, and D. J. Porteous. 2000. Genomic sequence analysis of *Fugu rubripes* CFTR and flanking genes in a 60 kb region conserving synteny with 800 kb of human chromosome 7. Genome Res. **10**:1194–1203.

Dermitzakis, E. T., A. Reymond, R. Lyle et al. (11 co-authors). 2002. Numerous potentially functional but non-genic conserved sequences on human chromosome 21. Nature **420**: 578–582.

Dubchak, I., M. Brudno, G. Loots, L. Pachter, C. Mayor, E. Rubin, and K. Frazer. 2000. Active conservation of noncoding sequences revealed by three-way species comparisons. Genome Res. **10**:1304–1306.

Duret, L., and P. Bucher. 1997. Searching for regulatory elements in human noncoding sequences. Curr. Opin. Struct. Biol. **7**: 399–406.

Friesen, V. L. 2000. Introns. Pp. 274–294 *in* A. J. Baker, ed. Molecular methods in ecology. Blackwell Science, Oxford.

Gilbert, N., and D. Labuda. 1999. CORE-SINEs: eukaryotic short interspersed retroposing elements with common sequence motifs. Proc. Natl. Acad. Sci. USA **96**:2869–2874.

Hare, M. P. 2001. Prospects for nuclear phylogeography. Trends Ecol. Evol. **16**:700–706.

Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22**:160–174.

Hughes, A. L., and M. Yeager. 1997. Comparative evolutionary rates of introns and exons in murine rodents. J. Mol. Evol. **45**:125–130.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. Nature **409**:860–921.

Jareborg, N., E. Birney, and R. Durbin. 1999. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. Genome Res. **9**:815–824.

Kim, J. 2001. Macro-evolution of the *hairy* enhancer in Drosophila species. Mol. Dev. Evol. **291**:175–185.

Kirby, D. A., S. V. Muse, and W. Stephan. 1995. Maintenance of pre-mRNA secondary structure by epistatic selection. Proc. Natl. Acad. Sci. USA **92**:9047–9051.

Kumar, S., and S. B. Hedges. 1998. A molecular timescale for vertebrate evolution. Nature **392**:917–920.

Leicht, B. G., S. V. Muse, M. Hanczyc, and A. G. Clark. 1995. Constraints on intron evolution in the gene encoding the myosin alkali light chain in Drosophila. Genetics **139**: 299–308.

Levy, S., S. Hannenhalli, and C. Workman. 2001. Enrichment of regulatory signals in conserved non-coding genomic sequence. Bioinformatics **17**:871–877.

Lewin, B. 1997. Genes VI. Oxford University Press, New York.

Li, W.-H. 1997. Molecular evolution. Sinauer Associates, Sunderland, Mass.

Ludwig, M. Z., C. Bergman, N. H. Patel, and M. Kreitman. 2000. Evidence for stabilizing selection in a eukaryotic enhancer element. Nature **403**:564–567.

Lum, J. K., M. Nikaido, M. Shimamura, H. Shiimodaira, A. M. Shedlock, N. Okada, and M. Hasegawa. 2000. Consistency of SINE insertion topology and flanking sequence tree: quantifying relationships among cetartiodactyls. Mol. Biol. Evol. **17**:1417–1424.

Lyons, L. A., T. F. Laughlin, N. G. Copeland, N. A. Jenkins, J. E. Womack, and S. J. O'Brien. 1997. Comparative anchor tagged sequences (CATS) for integrative mapping of mammalian genomes. Nat. Genet. **15**:47–56.

Mattick, J., and M. Gagen. 2001. The evolution of controlled multitasked gene networks: the role of introns and other noncoding RNAs in the development of complex organisms. Mol. Biol. Evol. **18**:1611–1630.

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

Nachman, M. W., and S. L. Crowell. 2000. Estimate of the mutation rate per nucleotide in humans. Genetics **156**: 297–304.

Palumbi, S. R. 1989. Rates of molecular evolution and the fraction of nucleotide positions free to vary. J. Mol. Evol. **29**: 180–187.

———. 1996. Nucleic acids II: the polymerase chain reaction. Pp. 205–247 *in* D. Hillis, C. Moritz, and B. Mable, eds. Molecular systematics. Sinauer Associates, Sunderland, Mass.

Palumbi, S. R., F. Cipriano, and M. P. Hare. 2001. Predicting nuclear gene coalescence from mitochondrial data: The three-times rule. Evolution **55**:859–868.

Petrov, D. A. 2002. DNA loss and evolution of genome size in Drosophila. Genetica **115**:81–91.

Petrov, D. A., T. A. Sangster, J. S. Johnston, D. L. Hartl, and K. L. Shaw. 2000. Evidence for DNA loss as a determinant of genome size. Science **287**:1060–1062.

Rambaut, A., and N. C. Grassly. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. **13**:235–238.

Rebane, A., and A. Metspalu. 1999. U82, a novel snoRNA identified from the fifth intron of human and mouse nucleolin gene. Biochim. Biophys. Acta–Gene Struct. Expr. **1446**: 426–430.

Rebane, A., R. Tamme, M. Laan, I. Pata, and A. Metspalu. 1998. A novel snoRNA (U73) is encoded within the introns of human and mouse ribosomal protein S3a genes. Gene **210**:255–263.

Saitou, N., and S. Ueda. 1994. Evolutionary rates of insertion and deletion in noncoding nucleotide sequences of primates. Mol. Biol. Evol. **11**:504–512.

Sankoff, D. 1972. Matching sequences under deletion/insertion constraints. Proc. Natl. Acad. Sci. USA **69**:4–6.

Shabalina, S., and A. Kondrashov. 1999. Pattern of selective constraint in *C. elegans* and *C. briggsae* genomes. Genet. Res. Camb. **74**:23–30.

Smit, A. F. A., and A. D. Riggs. 1995. MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation. Nucleic Acids Res. **23**:98–102.

Stephan, W., and D. A. Kirby. 1993. RNA folding in Drosophila shows a distance effect for compensatory fitness interactions. Genetics **135**:97–103.

Stoltzfus, A., J. J. M. Longsdon, J. D. Palmer, and W. F. Doolittle. 1997. Intron ''sliding'' and the diversity of intron positions. Proc. Natl. Acad. Sci. USA **94**:10739–10744.

Tagle, D. A., B. F. Koop, M. Goodman, J. L. Slightom, D. L. Hess, and R. T. Jones. 1988. Embryonic epsilon and gamma globin genes of a prosimian primate (*Galago crassicaudatus*): nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. J. Mol. Biol. **203**:439–455.

Venter, J. C., M. D. Adams, E. W. Myers et al. (271 co-authors). 2001. The sequence of the human genome. Science **291**: 1304–1351.

Wasserman, W. W., M. Palumbo, W. Thompson, J. W. Fickett, and C. E. Lawrence. 2000. Human-mouse genome comparisons to locate regulatory sites. Nat. Genet. **26**:225–228.

Webb, C. T., S. A. Shabalina, A. Y. Ogurtsov, and A. S. Kondrashov. 2002. Analysis of similarity within 142 pairs of orthologous intergenic regions of *Caenorhabditis elegans* and *Caenorhabditis briggsae*. Nucleic Acids Res. **30**:1233–1239.

Willoughby, D. A., A. Vilalta, and R. G. Oshima. 2000. An Alu element from the K18 gene confers position-independent expression in transgenic mice. J. Biol. Chem. **275**: 759–768.

Wingender, E., X. Chen, E. Fricke et al. 2001. The TRANSFAC system on gene expression regulation. Nucleic Acids Res. **29**:281–283.

Zar, J. H. 1984. Biostatistical analysis. Prentice-Hall, Englewood Cliffs, N.J.

Zhu, J., J. S. Liu, and C. E. Lawrence. 1998. Bayesian adaptive sequence alignment algorithms. Bioinformatics **14**:25–39.

David Rand, Associate Editor

Accepted February 12, 2003