# Similar Rates but Different Modes of Sequence Evolution in Introns and at Exonic Silent Sites in Rodents: Evidence for Selectively Driven Codon Usage

*Jean-Vincent Chamary and Laurence D. Hurst*

Department of Biology and Biochemistry, University of Bath, Bath, United Kingdom

In mammals divergence at fourfold degenerate sites in codons ($K_4$) and intronic sequence ($K_i$) are both used to estimate the mutation rate, under the supposition that both evolve neutrally. Does it matter which of these we use? Using either class of sequence can be defended because (1) $K_4$ is the same as $K_i$ (at least in rodents) and (2) there is no selectively driven codon usage (hence no systematic selection on third sites). Here we re-examine these findings using 560 introns (for 136 genes) in the mouse-rat comparison, aligned by eye and using a new maximum likelihood protocol. We find that the rate of evolution at fourfold sites and at intronic sites is similar in magnitude, but only after eliminating putatively constrained sites from introns (first introns and sites flanking intron-exon junctions). Any approximate congruence between the two rates is not, however, owing to an underlying similarity in the mode of sequence evolution. Some dinucleotides are hypermutable and differently abundant in exons and introns (e.g., CpGs). More importantly, after controlling for relative abundance, all dinucleotides starting with A or T are more prevalent in mismatches in exons than in introns, whereas C-starting dinucleotides (except CG) are more common in introns. Although C content at intronic sites is lower than at flanking fourfold sites, G content is similar, demonstrating that there exists a strong strand-specific preference for C nucleotides that is unique to exons. Transcription-coupled mutational processes and biased gene conversion cannot explain this, as they should affect introns and flanking exons equally. Therefore, by elimination, we propose this to be strong evidence for selectively driven codon usage in mammals.

## Introduction

In mammals, divergences of two classes of sequence are regularly used to estimate the mutation rate: fourfold degenerate sites in exons (Eyre-Walker and Keightley 1999; Keightley and Eyre-Walker 2000) and intronic DNA (Chang et al. 1994; Chang and Li 1995; Chang, Hewett-Emmett, and Li 1996; Huang et al. 1997). That the rates of point substitution at both classes of site ($K_4$ and $K_i$, respectively) are valid measures is supported by two important findings. First, there is the finding that in rodents $K_4$ (or $K_s$) and $K_i$ are approximately equal (Hughes and Yeager 1997), suggesting that the mode of evolution (putatively neutral) is the same in the two classes of sequence. Second, unlike most taxa (e.g., bacteria, yeast, fly, and nematode), there is no selectively driven codon usage in mammalian genes (Eyre-Walker 1991; Smith and Hurst 1999*a*; Kanaya et al. 2001; Duret 2002). This is evidenced by (among other things) the lack of correspondence between the usage of a codon and iso-acceptor tRNA abundance (Duret 2002).

Both of these findings require re-analysis, particularly because accumulating evidence suggests that neither fourfold degenerate sites nor introns are entirely free of constraint. Hughes and Yeager (1997) used complete intron sequences and all introns from a gene (except those too difficult to align). While noting that the splicing control regions that flank intron-exon junctions are subject to selective constraint, they reasonably argued that the number of such sites is too small to matter. However, recent evidence suggests that sequence conservation associated with splice sites may extend relatively far away from intron-exon boundaries (Majewski and Ott 2002; Waterston et al. 2002; Hare and Palumbi 2003; Sorek and Ast 2003). Majewski and Ott (2003), for example, showed that human SNP density and SINE insertion frequency is lower in the first and last 20 bp of introns and constraint may extend up to 200 bp into intronic sequence. The extent of conservation may well differ between the 5′ and 3′ ends (Majewski and Ott 2002; Sorek and Ast 2003). Given uncertainty over the size of the conserved region, we start by estimating its average size. We then purge our intronic alignments of these regions.

Mammalian introns also contain other motifs that could be under purifying selection, such as transcription factor (TF) binding sites (e.g., Rossi and de Crombrugghe 1987; Katai et al. 1992; Kawada et al. 1999; Suen and Goss 2001). The presence of such control elements may explain why transgene expression can be 10–100 times more efficient when introns are added to cDNA clones (Brinster et al. 1988). An estimated 10% of mouse introns contain regulatory elements, of which a fraction overlaps with CpG islands (Waterston et al. 2002).

Intron-associated regulatory elements are believed to be nonrandomly distributed within a gene. For example, they tend to be located in close proximity to the start codon, and thus in the first intron within the coding sequence (Sakurai et al. 2002). This in turn may explain (Sakurai et al. 2002) why the intron in single-intron genes tends to be located 5′ end (see also Fink 1987; Mourier and Jeffares 2003). Numerous reports (e.g., Oshima, Abrams, and Kulesh 1990; Rohrer and Conley 1998; Chan et al. 1999) describe control elements in first introns (see also all the above references describing introns with TF binding sites). Although only a few studies have compared all introns derived from the same gene (e.g., Palmiter et al. 1991; Jonsson et al. 1992), these report that the first intron has the greatest impact in modulating expression. In contrast, a systematic in silico analysis on a large data set failed to identify a higher frequency of TF binding sites in first introns (Levy,

Key words: codon usage bias, point substitution rate, purifying selection, introns, fourfold degenerate sites, dinucleotides.

E-mail: l.d.hurst@bath.ac.uk.

*Mol. Biol. Evol.* 21(6):1014–1023. 2004
DOI:10.1093/molbev/msh087
Advance Access publication March 10, 2004

 

Hannenhalli, and Workman 2001). Although this may reflect our poor abilities to detect control elements computationally, some control elements have been described in non-first introns (e.g., Lothian and Lendahl 1997; Hural et al. 2000). The 5′ end of first introns may be of particular importance in transcriptional control (Majewski and Ott 2002).

If it is a general property of first introns to harbour more control elements, we would expect them to evolve slower than the other "non-first" introns, all things being equal. However, as first introns tend to be larger (Hawkins 1988; Smith 1988), a higher number of constrained sites need not imply a higher density of constrained sites. Indeed, Levy, Hannenhalli, and Workman (2001) report that, if anything, first introns evolve faster. However, this analysis came from the mouse-human comparison in which alignment of freely evolving sites is unreliable (Jareborg, Birney, and Durbin 1999). In contrast, we find that first introns evolve slower and that the reduced point substitution rate is in part owing to a greater abundance of CpG islands. Given this evidence for the likely action of purifying selection on intronic sites, we then ask whether $K_4$ becomes significantly lower than $K_i$ after removing constrained sites. Such a finding could undermine the use of $K_4$ as an estimator of the mutation rate.

The second issue we investigate is whether selection on synonymous sites exists and, more specifically, whether we can detect selectively driven codon usage. Recent direct evidence shows that synonymous mutations can be highly deleterious (Duan et al. 2003). Further, codon usage bias has recently been reported (Urrutia and Hurst 2003) to be greater in highly expressed genes (see also Debry and Marzluff 1994). Comparably, constitutively expressed exons have a higher GC content than those that are alternatively expressed (Iida and Akashi 2000). It is possible that as many as 40% of fourfold degenerate sites are under selection (Hellmann et al. 2003).

To address this issue, here we ask whether the substitution processes in introns and at synonymous sites are comparable. Given that certain dinucleotides can be differently abundant in exons and introns (e.g., "differential CpG content," Subramanian and Kumar 2003; see also Hellmann et al. 2003), we analyze all possible dinucleotides and ask whether there is a discrepancy in their relative abundances between fourfold degenerate sites and introns. We then determine the "mutability" of each dinucleotide, i.e., how often a given dinucleotide is associated with a mismatch in introns and at fourfold sites. Additionally, a dinucleotide could be equally abundant in the two classes of sequence, but have different forces operating on it. To examine this we investigate the "stability" of each dinucleotide, i.e., the rate of involvement of a given dinucleotide in a mismatch, after controlling for the abundance of the dinucleotide.

## Materials and Methods

### Data Set of Orthologous Mouse-Rat Genes

We expanded upon a data set of over 40 mouse-rat orthologs (Hughes and Yeager 1997; Smith and Hurst 1998), where orthology was determined through HOVERGEN (the Homologous Vertebrate Gene Database, Release 42, available at www.hgmp.mrc.ac.uk; Duret, Mouchiroud, and Gouy 1994). Each gene pair is considered orthologous only if, within the gene family tree, there is no non-rodent lineage between the mouse and rat branches and if at least one non-rodent sequence is present as an outgroup. Orthology was further validated through syntenic comparisons using LocusLink (www.ncbi.nlm.nih.gov/LocusLink/) and the Rat Genome Database (RGD) Virtual Comparative Map tool (http://rgd.mcw.edu/VCMAP/).

A list of 5,339 rat genes was downloaded from the RGD (http://rgd.mcw.edu/pub/data_release/GENES). The corresponding sequence entries were extracted from GenBank (www.ncbi.nih.gov/Genbank). Each GenBank file was scrutinized for the presence of annotations describing the location of every exon. This returned 231 matches for complete genes possessing at least one intron. Excluding the rat genes for which a confirmed mouse ortholog had already been identified, each of the remaining 189 rat sequences was Blasted (Altschul et al. 1990) against the complete mouse genome at Ensembl (version 14.30.1, www.ensembl.org/Mus_musculus/blastview/), returning 126 hits.

We examined each rat gene in HOVERGEN to identify the mouse ortholog. If the mouse ortholog had the introns described, the HOVERGEN-derived GenBank files replaced Blast ones (which occurred in 44 cases). If the HOVERGEN-described mouse ortholog was the same as the Blast return, but lacking in introns (i.e., only mRNA described), then we retained the Blast entry. In 23 cases, HOVERGEN described an unambiguous ortholog, different from the Blast match. These were eliminated from our data set. For the remaining 58 genes for which HOVERGEN did not specify a mouse ortholog (usually because no rat sequence was available for the gene family), the orthology of the Blast sequence to the rat sequence was confirmed by ensuring that the genes were syntenic with the orthologous region in rat, that intron pairs were well-aligned overall, and that the estimated rate of protein evolution was within normal bounds for the mouse-rat comparison ($K_a < 0.2$).

Of the remaining data set of 142 genes, a further 2 were excluded due to poor sequence annotation and another 4 because they were located on the X-chromosome (these having lower substitution rates than autosomal genes [Hurst and Ellegren 1998]). Syntenic comparisons and the RGD list of accession numbers for each gene were used to ensure that there was no redundancy within the data set. For statistical analyses, we report sample size as "$N_g$" if evaluated on a gene-by-gene basis and as "$N_i$" on an intron-by-intron basis.

### Sequence Alignments

Coding sequence was extracted from GenBank files using GBPARSE (http://sunflower.bio.indiana.edu/~wfischer/Perl_Scripts/). Alignment of the translated sequences was carried out using PILEUP. Nucleotide alignments were reconstructed from the amino acid

sequence alignments using AA2NUC (available from L.D.H.).

Only internal introns located between coding exons were analyzed. Two methods were used for aligning introns: (1) manually (by-eye) and (2) using MCALIGN, a stochastic maximum likelihood (ML)–based program incorporating a Monte Carlo algorithm (P. D. Keightley and T. Johnson, unpublished data, http://homepages.ed.ac.uk/eang33/mcinstructions.html). MCALIGN is based on a model of noncoding sequence evolution that is built upon the frequency of indel events relative to nucleotide substitutions. We executed the program using the rodent intron parameters provided. Seven massive introns (>7 kb, two of which were first introns) proved too difficult to align. Our final data set consisted of 136 orthologous genes possessing 560 introns.

## Estimating Rates of Evolution

Both $K_4$, the number of substitutions per fourfold degenerate site within exons, and $K_i$, the intronic substitution rate, were estimated using the algorithmic method of Tamura and Nei (1993). To obtain genic $K_i$ values, intronic substitution rates were weighted according to the number of bases compared per individual intron alignment (Smith and Hurst 1998). The indel rate, $K_{indel}$, was calculated as the total number of indels per base pair of the alignment.

## Accounting for Alignment Artifact

After estimating the $K_i$ per intron by the two alignment methods, we defined a conservative set and a liberal set. The former contains the lower estimate of $K_i$, and the latter contains the higher value. For the intron alignments in the conservative alignment set, 516 introns were drawn from the by-eye set and the remainder from the ML set. Both alignment methods agree exactly on both $K_{indel}$ and $K_i$ for 84 of our orthologous introns. We define this as our "tight" set ($N_g = 50$). This subset consists mainly of short introns (mean length 122 bp, as opposed to 604 bp).

Alignment-induced noise is commonly minimized by rejecting difficult-to-align introns from analysis (e.g., Hughes and Yeager 1997; Smith and Hurst 1998). Instead, we filtered out ambiguous regions of alignments to produce a fifth alignment set containing all introns ($N_i = 560$). We did this by applying the Gblocks program (www1.imim.es/~castresa/Gblocks/Gblocks.html; Castresana 2000) to the alignments in the conservative set under the default parameters.

## Sequence Conservation at Intron-Exon Junctions

For a given distance away from the intron-exon boundary (running $5' \rightarrow 3'$ for the $5'$ end, $3' \rightarrow 5'$ for the $3'$ end), we calculated the frequency among all introns of mismatches at the site in question. Mismatches were defined in two ways: (1) a nucleotide aligned against a different nucleotide (i.e., those that contribute to estimates of $K_i$) and (2) a nucleotide matched with a gap. Ambiguous nucleotides (N) were ignored.

## Detecting Control Elements and Transcription Factor Binding Sites

To ask whether first introns contain more regulatory elements, and whether the presence of regulatory elements predicts the rate of evolution, we assayed introns for the presence of CpG islands and transcription factor (TF) binding sites.

CpG islands tend to be located in regions with higher than expected numbers of CpG dinucleotides in the coding strand (Gardiner-Garden and Frommer 1987). We used the recommended settings for detecting CpG islands at the $5'$ of genes implemented by CPGREPORT (available as part of EMBOSS, www.ebi.ac.uk/emboss/cpgplot/), which carries out a running-sum (not window) analysis to identify CpGs associated with putative islands. CpGs in putative islands are not necessarily present as a continuous array, so we also define islands by the extent to which the CpGs in such islands are clustered within a CpG-rich region. We therefore define island density by whether we require a minimum of 0, 1, 5, 10, 15, or 20 clustered island-associated CpGs to define an island. The density of putative CpG islands within a given intron was taken as the mean of the densities in mouse and rat.

To identify TF binding sites, we scanned our sequences for exact matches to well-characterized vertebrate TF binding sites as described in TRANSFAC (Wingender et al. 2000). We did this by employing TFSCAN within EMBOSS (www.hgmp.mrc.ac.uk/Software/EMBOSS/Apps/tfscan.html). We then masked the corresponding input sequences at the positions where hits were found. We then determined, for each intron, the density of putative TF binding sites. We also examined mouse TF binding sites alone. While this qualitatively affects the observed densities, it does not affect any of the patterns that we observe. We report only the data from the vertebrate collection.

## Dinucleotide Content and Mismatch Rates

Any given mismatch in an alignment is associated with four dinucleotides (e.g., ATT/ACT is associated with AT, AC, TT, and CT). At fourfold sites, however, there can be no dinucleotides starting with A that have the mismatch at the second site, as there are no fourfold degenerate codons with A at their second site. Therefore, to give a fairer impression of the forces acting in exons and introns, in exons we only consider dinucleotides in which the first base is the one at the fourfold site. To ensure comparability, in introns we only count dinucleotides as they occur at the first site of the pair. For each dinucleotide we calculate, for fourfold degenerate sites and introns: (1) the relative abundance (frequency of occurrence) of the dinucleotide; (2) "mutability," the probability with which the dinucleotide is associated with a mismatch; and (3) "stability," the probability with

**Table 1**
**Estimates for Rates of Evolution from Mouse-Rat Orthologous Introns and Exons**

| Rate[a] | Alignment Set[b] | | | | | |
|---|---|---|---|---|---|---|
| | By-Eye | Max. Likelihood | Conservative | Liberal | Tight | Gblocks |
| $K_i$ | 0.1478 ± 0.0282 | 0.1701 ± 0.0356 | 0.1468 ± 0.0274 | 0.1710 ± 0.0358 | 0.1454 ± 0.0495 | 0.1443 ± 0.0266 |
| $K_i$ first | 0.1442 ± 0.0399 | 0.1616 ± 0.0513 | 0.1431 ± 0.0394 | 0.1627 ± 0.0513 | 0.1375 ± 0.0448 | 0.1405 ± 0.0391 |
| $K_i$ non-first | 0.1533 ± 0.0293 | 0.1791 ± 0.0406 | 0.1526 ± 0.0286 | 0.1798 ± 0.0408 | 0.1466 ± 0.0518 | 0.1504 ± 0.0271 |
| $K_a$ | 0.0421 ± 0.0391 | | | | | |
| $K_s$ | 0.1718 ± 0.0513 | | | | | |
| $K_4$ | 0.1824 ± 0.0723 | | | | | |

[a] Mean point substitution rate (± SD). For a given gene, the intronic substitution rate ($K_i$) is the mean across all introns weighted by the size (number of aligned sites) of each intron. Exonic substitution rates ($K_a$, $K_s$, and $K_4$) are shown for estimates from all 136 genes.

[b] The number of genes (number of introns), $N_g$ ($N_i$), used to calculate $K_i$ (excluding the tight alignment set) = including all introns 136 (560), first 134 (134), non-first 118 (426). For the tight set, $N_g$ ($N_i$) = all introns 50 (84), first introns 16 (16), non-first introns 42 (68).

which the dinucleotide is associated with a mismatch, per occurrence of the dinucleotide.

## Results

### Comparing Estimates of Evolutionary Rates Generated by Alternative Alignment Methods

How confident can we be that our estimates of intronic point substitution rates are accurate and relatively unaffected by alignment artifacts? We attempted to resolve this problem by aligning each orthologous intron pair using two different methods. Although there was a strong correlation between the values obtained per intron ($R^2 = 0.678$, $P < 0.001$, $N_i = 560$), the maximum likelihood (ML) program tends to produce slightly higher estimates for $K_i$ than the by-eye method (table 1). For the 84 introns in the tight alignment set, $K_i$ is very similar to that in the conservative alignment set. Thus, these are likely to be minimum estimates, whereas the liberal set reflects realistic upper estimates.

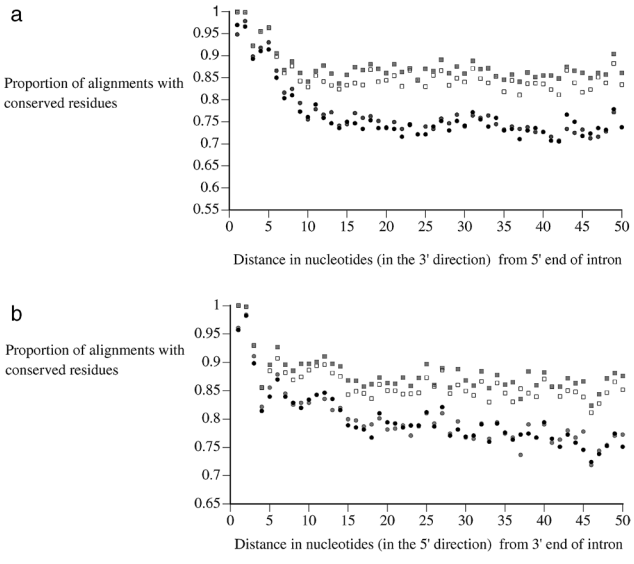


FIG. 1.—Sequence conservation at intronic sites flanking intron-exon junctions as a function of distance from the junction at the (*A*) 5′ end and (*B*) 3′ end. Conservation is defined with (circles) and without (squares) counting gaps as informative sites. The alignment methods are by-eye (grey) and using a maximum likelihood protocol (black/white).

### Constraint on Intronic Sites: How Much Sequence Flanking Intron-Exon Junctions Is Conserved?

Considering figure 1, we suggest that, to be conservative, intron sequence within the first and last 20 bp should be excluded from analysis. As expected (Reed and Maniatis 1985), there is selection against substitutions at the first and last two intronic sites adjacent to splice junctions (fig. 1). In contrast, at the 3′ end, we do not find evidence that the 7-nucleotide branch site, commonly located 18–40 nucleotides upstream (Reed and Maniatis 1985), is located at any single well-conserved region. However, though the branch site is the preferred site for lariat formation, it is not essential (Zhuang, Goldstein, and Weiner 1989). On the other hand, we observe minor peaks in conservation approximately 12 bp and 19 bp upstream of the 3′ end, which may represent two alternative locations for the branch site (fig. 1*B*). As there is not enough evidence to justify exclusion of further nucleotides (e.g., up to 200 bp), we quote substitution rates for all analyses given below, after excluding only the 20 bp at each end.

### Constraint on Intronic Sites: First Introns Contain More CpG Islands and Evolve Slower

Previous reports have shown that first introns can enhance gene expression to a greater degree than other introns from the same gene. Does this mean that first introns evolve slower? As we report in table 2, we find this to be the case (see also table 1). The obvious explanation is that first introns possess more control elements. The possession of CpG islands could have a twofold effect in slowing first intron evolution, not only by imposing functional constraints, but also by demethylation of CpG dinucleotides that would otherwise be hypermutable. We observe that first introns are richer in CpGs belonging to putative CpG islands (table 3). In contrast, we find no evidence that first introns have a higher density of transcription factor binding sites (see table *A* in the Supplementary Material online). The latter is a weak test, particularly because the short length of TF binding sites results in high rates of degeneration and spontaneous emergence ("turnover") (Dermitzakis and Clark 2002). Indeed, 30%–40% of binding sites known to be present in humans cannot be detected in rodents (Dermitzakis and

**Table 2**
**Differences in Rates of Evolution and GC Content Between First and Non-first Introns from the Same Gene (One-Sample Wilcoxon Signed-Rank Tests, $N_g = 116$)**

| | Alignment Set | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Conservative | | | Liberal | | | Gblocks | | |
| Rate/GC Content | *P*-value | Mean first | Mean Non-first | *P*-value | Mean First | Mean Non-first | *P*-value | Mean First | Mean Non-first |
| $K_i$ | 0.031 | 0.1452 ± 0.0397 | 0.1515 ± 0.0283 | 0.002 | 0.1648 ± 0.0514 | 0.1783 ± 0.0402 | 0.029 | 0.1432 ± 0.0398 | 0.1494 ± 0.0268 |
| $K_i$ non-CpG | 0.026 | 0.1274 ± 0.0403 | 0.1351 ± 0.0284 | 0.001 | 0.1457 ± 0.0521 | 0.1598 ± 0.045 | <0.001 | 0.1176 ± 0.0395 | 0.1398 ± 0.0510 |
| GCi | 0.765 | 50.692 ± 7.936 | 50.483 ± 8.055 | | | | | | |

Clark 2002). In mouse, the proportion of G + C nucleotides within regulatory elements is generally higher than overall genomic GC content (Waterston et al. 2002). However, we do not find a significant difference between mean GCi of first introns and non-first introns (table 2).

We expect that introns with control elements should evolve slower than those without. At least for putative CpG islands, this is so (table *B* in the Supplementary Material online). Could this, along with the greater abundance of CpG islands in first introns, entirely explain why first introns evolve slowly? Do first introns without CpG islands then have the same rate of evolution as non-first introns also lacking such islands? First introns still evolve more slowly (online supplementary table *C*), suggesting that CpG island presence only partly explains the difference between first and non-first introns. We find no correlation between putative TF binding site density and $K_i$ (online supplementary table *D*).

We presumed that the excess of CpGs found in first introns reflects unmethylated CpG islands. The higher density of CpG-rich regions in first introns might instead represent methylated hypermutable CpGs, rather than conservation of control elements (CpG islands). To examine this we asked whether, after masking CpG dinucleotides in all introns, first introns have a lower rate of evolution than non-first. We find that the significance of the difference in $K_i$ increases after masking (table 2), indicating that the CpGs present are constrained putative islands, not hypermutable sites.

Given that in silico methods have a high false positive rate (Fickett and Hatzigeorgiou 1997; Wasserman et al. 2000), can we be confident that removal of first introns also eliminates most control elements? We address this issue by asking whether, before and after removal of first introns, there is heterogeneity between introns in their rate of evolution. We classified introns according to their relative position within a given gene, i.e., 1 = first intron, 2 = second, etc. As expected, if all introns are analyzed, we observe highly significant heterogeneity in $K_i$ between introns at different positions (e.g., liberal set: $P = 0.0019$, Kruskal-Wallis, df = 16). Importantly, after removal of first introns this heterogeneity disappears ($P = 0.1671$, df = 15). Similarly, there exists a highly significant correlation between intron position and $K_i$ when first introns are included ($\rho^2 = 0.0278$, $P < 0.0001$, Spearman rank, $N_i = 560$) but not after their removal ($\rho^2 = 0.0037$, $P = 0.212$, $N_i = 426$). These findings are consistent with the notion that most functionally constrained elements are located in first introns.

## Exclusion of Constrained Sites Leads to a Similar $K_4$ and $K_i$

If we include first introns in our estimate of $K_i$, we find that fourfold sites evolve faster than intronic ones (independent of the alignment set used, $P < 0.05$, one-sample Wilcoxon signed-rank tests, $N_g = 136$, tight set $N_g = 50$). In this regard, we fail to replicate the prior results (Hughes and Yeager 1997; Smith and Hurst 1998). However, if we exclude first introns, we find that $K_4 = K_i$ in the ML and liberal sets ($P > 0.3$, one-sample Wilcoxon signed-rank tests, $N_g = 118$) but not in the other sets ($P < 0.001$, one-sample Wilcoxon signed-rank tests, $N_g = 118$). We can conclude that fourfold sites do not evolve slower than intronic sites, and that after putatively constrained sites are removed from introns, $K_4$ and $K_i$ become more similar in magnitude.

## Association of Dinucleotides with Mismatches

Given that the rate of sequence evolution in introns (excluding sites under purifying selection) and at fourfold sites are about the same, should we conclude that the

**Table 3**
**Differences in CpG Island Density Between First and Non-first Introns from the Same Gene (Paired *t*-tests, $N_g = 116$)**

| | Minimum Number of Clustered Island-Associated CpGs Required to Define a Putative CpG Island[a] | | | | | |
|---|---|---|---|---|---|---|
| Density | 0 | 1 | 5 | 10 | 15 | 20 |
| *P*-value | 0.005 | 0.006 | 0.004 | 0.004 | 0.039 | 0.077 |
| Ratio | 1.3876 | 1.6459 | 3.3275 | 6.3027 | 9.4976 | 14.3903 |
| Mean first | 0.0381 ± 0.0407 | 0.0284 ± 0.0426 | 0.0167 ± 0.0424 | 0.0114 ± 0.0366 | 0.0063 ± 0.0261 | 0.0047 ± 0.0231 |
| Mean non-first | 0.0275 ± 0.0260 | 0.0173 ± 0.0263 | 0.0050 ± 0.0214 | 0.0018 ± 0.0125 | 0.0007 ± 0.0067 | 0.0003 ± 0.0038 |

[a] Putative CpG islands become increasingly more difficult to detect as the threshold for the number of island-associated CpGs required to define an island increases, so when 20 clustered CpGs are required, the majority of introns have no detectable CpG cluster and the significance level is dependent on the number ($N_i = 15$) in which the islands can be detected.

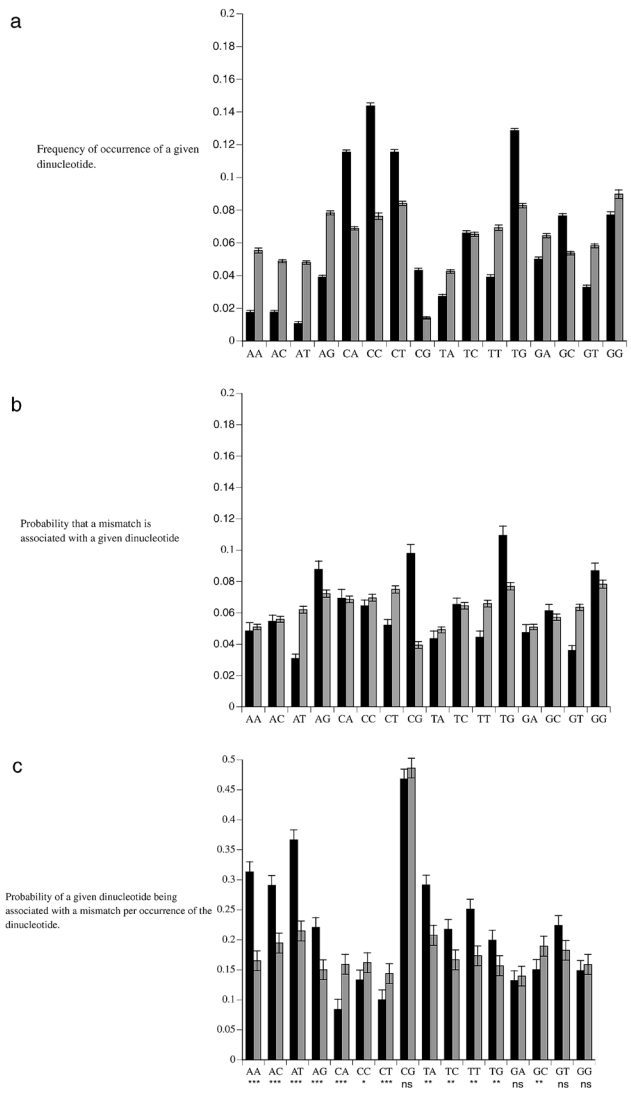


FIG. 2.—At fourfold sites it is the first base of the dinucleotide that occurs at the fourfold site. (*A*) Relative abundance of dinucleotides at fourfold degenerate sites in exons (black bars) and in introns (grey bars). (*B*) Mutability of dinucleotides at fourfold degenerate sites in exons (black bars) and in introns (grey bars). Mutability is defined as the frequency of occurrence of a given dinucleotide in a mismatch. (*C*) Stability of dinucleotides at fourfold degenerate sites in exons (black bars) and in introns (grey bars). Stability is defined as the frequency of occurrence of a given dinucleotide in a mismatch, per incidence of the dinucleotide (i.e., controlling for differences in the relative abundance between fourfold and intronic sites). The significance of differences is indicated by: ns = $P > 0.05$, *$P < 0.05$, **$P < 0.01$, and ***$P < 0.001$.

process of evolution in the two classes are equivalent? Analysis of dinucleotides strongly suggests that we should not. As previously reported in primates (Hellmann et al. 2003; Subramanian and Kumar 2003), CpGs in rodent sequences are more abundant in exons than in introns (fig. 2*A*; CpG content is 7% in exons and 3% in introns). The same is true for all dinucleotides starting with a C. Given the hypermutability of CpG dinucleotides (Bird 1980; McClelland and Ivarie 1982; Cooper and Krawczak 1989; Sved and Bird 1990), our observed excess of TG pairs in exons compared with introns is also expected. The other striking feature is the dearth of the A-starting dinucleotides at fourfold sites in exons.

If all things were equal, we should expect that the probability that a dinucleotide is associated with a mismatch should simply be proportional to the frequency of occurrence of the dinucleotide. However, the C-starting dinucleotides (excluding CpG) are no more likely to be found at a mismatch in exons than in introns (fig. 2*B*). Likewise, mismatches at A-starting dinucleotides (excluding AT) occur at either comparable frequencies in exons and introns or are more abundant in exons, counter to the occurrence of the dinucleotide itself. In other words, the abundance of mismatch-associated dinucleotides per occurrence of the dinucleotide is not the same in exons and introns (fig. 2*C*). Although CpGs are equally likely to be associated with a mismatch in exons and introns, these are the exception. Other C-starting dinucleotides are more "stable" in exons. In contrast, A- and T-starting dinucleotides are more unstable in exons than in introns.

Theoretically, these results could reflect an artifact. Consider a sequence alignment with substitutions randomly located. All things being equal, the number of occurrences of a dinucleotide that is associated with a mismatch could be lower than that of any dinucleotides that happen not to be associated with mismatches. This is simply because association with a mismatch reduces the count of the dinucleotide, as it is present in only one of the two sequences at the site of the mismatch. To exclude this potential bias we reperformed the analysis, but this time we added one to the count of each dinucleotide when it was associated with a mismatch. None of the significant results shown in figure 2*C* are rendered nonsignificant and vice versa.

The apparent instability of A and T at fourfold sites and the opposite apparent stability of C might tempt us to suppose that single nucleotide effects (rather than dinucleotide effects) are the sole interest. However, in exons, the stability of dinucleotides with A or T at the first position is dependent on the nucleotide at the second position (ANOVA on stability of dinucleotides in exons: A-starting dinucleotides, $F = 5.34$, df = 3, $P = 0.001$; T-starting dinucleotides in exons, $F = 4.09$, df = 3, $P = 0.007$; N.B. this result is also found when we adjust for the putative artifact).

## Discussion

We have shown that removing putatively constrained sites from introns renders their rate of evolution similar to that of silent sites in the flanking exons. This is consistent with the notion that $K_4$ and $K_i$ both may be measuring the background mutation rate. However, we have also shown that this similarity in rates of evolution is more likely to be a happy accident, rather than an indication of equivalence in the mode of sequence evolution.

Notably, there are discrepancies in exons and introns in the frequency of occurrence of given dinucleotides associated with mismatches, after controlling for their abundances. Consider, for example, AA and AC. The rate of evolution (mismatch rate) is approximately the same in exons and introns (fig. 2*B*) but only because these

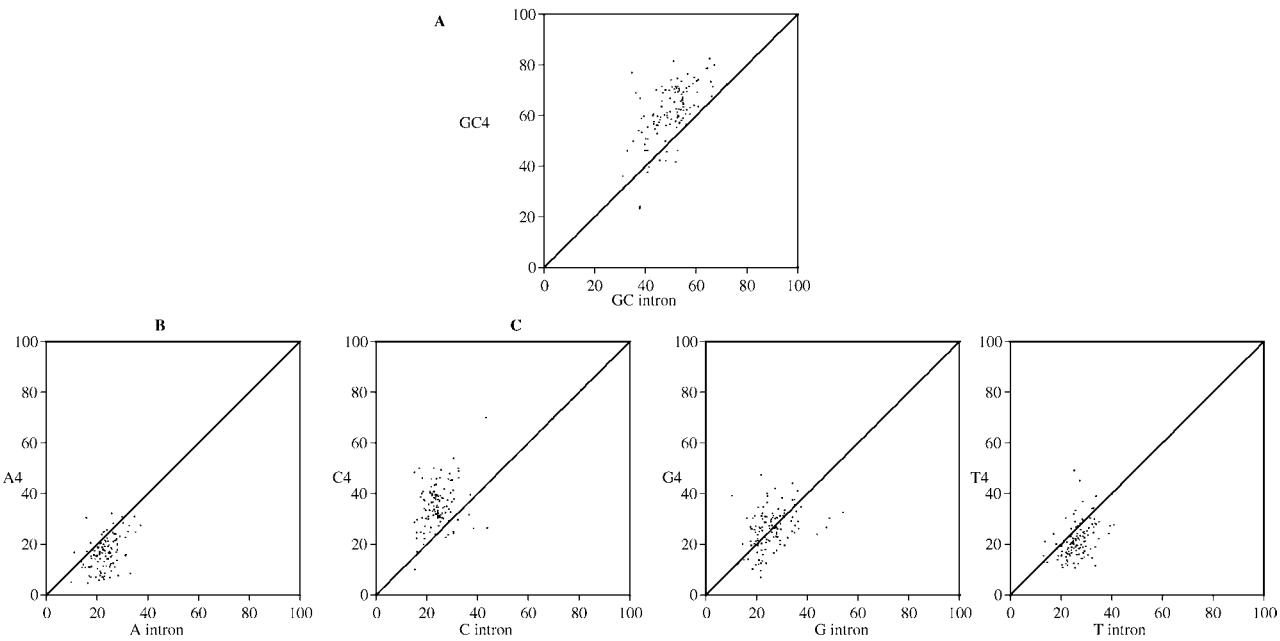


FIG. 3.—Relationship between nucleotide content in introns and at fourfold degenerate sites in flanking exons. (*A*) GC content, (*B*) A content, (*C*) C content, (*D*) G content, and (*E*) T content. The line indicates equality of content.

dinucleotides are both especially rare (fig. 2*A*) and especially unstable in exons (fig. 2*C*). In contrast, the evolution of CpG dinucleotides is more as classically supposed (Cooper and Krawczak 1989; Sved and Bird 1990); the mutability of CpGs appears to be independent of context (intron vs. exon; fig. 2*C*) and the different effects on exons and introns relates solely to different abundances (fig. 2*B*). To account for the latter observation, we need only account for the different CG dinucleotide contents and need not evoke a difference in the mode of evolution of exons and introns. As noted, however, CG is the exception, and only three other dinucleotides (GA, GT, and GG) show the same stability in exons as in introns (fig. 2*C*).

In addition, there is a strand-specific, as well as an exon-specific, enrichment of C nucleotides (fig. 3). Were the effect not strand-specific, we should expect both G and C content in exons to be higher than that in introns. However, C content at fourfold sites is much higher than that in the flanking introns (fig. 3*C*; $P < 0.0001$, one-sample Wilcoxon signed-rank test, $N_g = 136$), while G content is not significantly different between the two (fig. 3*D*; $P = 0.69$, one-sample Wilcoxon signed-rank test, $N_g = 136$). We are not aware of any previous reports of this unexpected difference. These discrepancies cannot be the result of transcription-coupled processes (Green et al. 2003; Majewski 2003), nor to biased gene conversion (Galtier et al. 2001), both of which should affect introns and flanking exons equally. By elimination, we conclude that this is consistent with selectively driven codon usage.

This suggestion is, however, unorthodox. The classical model for selectively driven codon usage bias suggests that its function is to increase the efficiency (rate or accuracy; Duret 2002) of mRNA translation. Evidence for this comes from observations of highly expressed genes having greater bias and that the skews in codon usage reflect iso-acceptor tRNA abundance. Co-adaptation between codon usage, expression rate, and/or tRNA abundance have been described in the worm (Stenico, Lloyd, and Sharp 1994; Duret and Mouchiroud 1999; Duret 2000; Castillo-Davis and Hartl 2002), in the fruitfly (Shields et al. 1988; Moriyama and Powell 1997; Duret and Mouchiroud 1999), and in yeast and bacteria (for reviews, see Ikemura 1985; Sharp et al. 1995), but not in humans (Duret 2002). More generally, in mammals it is usually presumed that the effective population size is too small to allow selection on synonymous codon usage (Sharp et al. 1995) and thus that codon usage bias reflects background isochore GC content (Eyre-Walker 1991; Sharp et al. 1995).

How can we then suggest that there exists selectively driven codon usage, while the abundance of iso-acceptor tRNAs is not skewed (Duret 2002)? Despite this absence of skew, modified patterns of codon usage can affect expression levels in mammals (e.g., Kim, Oh, and Lee 1997). Rather than the product of translational selection, codon usage bias may be the result of selection on mRNA secondary structure (Carlini, Chen, and Stephan 2001), stability, and half-life. Importantly, a recent in vitro study in humans (Duan et al. 2003) has shown that synonymous mutations can be deleterious because of their effect on mRNA secondary structure, reducing stability (see also Gottlieb et al. 1999). Similarly, in bacteria, exchanging synonymous "major" (frequently-used) codons for "minor" ones can result in up to 10-fold reductions in the half-life of mRNA in vivo (Deana and Reiss 1993; Deana, Ehrlich, and Reiss 1996; Deana, Ehrlich, and Reiss 1998). The reduced synonymous substitution rate at the 5′ end of both bacterial (Eyre-Walker and Bulmer 1993) and rodent (Smith and Hurst 1999*b*) coding sequences may also

reflect selection on mRNA secondary structure. The same process may explain, in part, our observation that first introns evolve slowly, even after allowing for the presence of CpG islands (table *B* in the Supplementary Material online).

There are at least two explanations for the effects of silent mutations. On the one hand, the mutations may alter the folding properties, and therefore the stability, of the mRNA. Importantly, several in silico studies report that natural mRNAs (including vertebrate sequences) are more stable than artificial variants that are identical in all regards, other than their synonymous codon usage (Seffens and Digby 1999 [but see Workman and Krogh 1999]; Cohen and Skiena 2003; Katz and Burge 2003). Such arguments fit within the broader context, advocated by Vinogradov (2001*a*, 2001*b*, 2003), that sequence composition reflects selection on physical properties of nucleic acids, such as bendability.

Alternatively, codon usage might alter the ability of mRNA to bind RNA-metabolizing proteins, such as those that bind AU-rich motifs (usually present in 3′ UTRs [Caput et al. 1986; Shaw and Kamen 1986]) and induce rapid mRNA degradation (e.g., Bohjanen et al. 1991). AT-avoidance in exons might therefore minimize the probability of this occurring. The rarest dinucleotide at fourfold degenerate sites in exons is AT (fig. 2*A*), and, given its abundance, it is unusually unstable (fig. 2*C*).

If the preference for C and avoidance of A nucleotides in exons (figs. 2*A* and 3) is the result of selection on mRNA half-life, then we predict that modified versions of mRNAs in mammals, decreasing the abundance of the stable dinucleotides and increasing that of the unstable ones, should, on average, increase mRNA decay rates. For those post-transcriptionally-regulated, this need not be the case (Seffens and Digby 1999). In principle, our prediction could be tested in vitro.

The skew in C but not G content has a bearing on the interpretation of the finding that GC4 is usually greater than GCi (for references, see Duret and Hurst 2001), as also seen in our data set (fig. 3*A*). This is potentially consistent with selection favoring a higher GC content in exons (Hughes and Yeager 1997; Eyre-Walker 1999). This possible interpretation was criticized, as introns also have more GC-poor transposable element (TE) insertions (Duret and Hurst 2001). Vinogradov (2001*b*) counter-argued that TEs cannot account for all of the distortion seen. In addition, that we observe a strand- and exon-dependent enrichment of C, but not of G, is not obviously consistent with the TE model. The model likewise fails to account for the rarity with which C nucleotides feature as a mismatch in exons, given their frequency of occurrence.

## Acknowledgments

We thank Martin Lercher, Csaba Pal, and Araxi Urrutia Odabachian for discussion, and we thank Laurent Duret and Jacek Majewski. We are grateful to two anonymous reviewers for insightful comments on an earlier version of the manuscript. J.V.C. and L.D.H. are funded by the United Kingdom Biotechnology and Biological Sciences Research Council.

## Literature Cited

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J. Mol. Biol. **215**:403–410.

Bird, A. P. 1980. DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res. **8**:1499–1504.

Bohjanen, P. R., B. Petryniak, C. H. June, C. B. Thompson, and T. Lindsten. 1991. An inducible cytoplasmic factor (AU-B) binds selectively to AUUUA multimers in the 3′ untranslated region of lymphokine mRNA. Mol. Cell Biol. **11**:3288–3295.

Brinster, R. L., J. M. Allen, R. R. Behringer, R. E. Gelinas, and R. D. Palmiter. 1988. Introns increase transcriptional efficiency in transgenic mice. Proc. Natl. Acad. Sci. USA **85**:836–840.

Caput, D., B. Beutler, K. Hartog, R. Thayer, S. Brown-Shimer, and A. Cerami. 1986. Identification of a common nucleotide sequence in the 3′-untranslated region of mRNA molecules specifying inflammatory mediators. Proc. Natl. Acad. Sci. USA **83**:1670–1674.

Carlini, D. B., Y. Chen, and W. Stephan. 2001. The relationship between third-codon position nucleotide content, codon bias, mRNA secondary structure and gene expression in the drosophilid alcohol dehydrogenase genes Adh and Adhr. Genetics **159**:623–633.

Castillo-Davis, C. I., and D. L. Hartl. 2002. Genome evolution and developmental constraint in Caenorhabditis elegans. Mol. Biol. Evol. **19**:728–735.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. **17**:540–552.

Chan, R. Y., C. Boudreau-Lariviere, L. M. Angus, F. A. Mankal, and B. J. Jasmin. 1999. An intronic enhancer containing an N-box motif is required for synapse- and tissue-specific expression of the acetylcholinesterase gene in skeletal muscle fibers. Proc. Natl. Acad. Sci. USA **96**:4627–4632.

Chang, B. H. J., D. Hewett-Emmett, and W.-H. Li. 1996. Male-to-female ratios of mutation-rate in higher primates estimated from intron sequences. Zool. Studies **35**:36–48.

Chang, B. H. J., and W.-H. Li. 1995. Estimating the intensity of male-driven evolution in rodents by using X-linked and Y-linked *Ube-1* genes and pseudogenes. J. Mol. Evol. **40**:70–77.

Chang, B. H. J., L. C. Shimmin, S. K. Shyue, D. Hewett-Emmett, and W.-H. Li. 1994. Weak male-driven molecular evolution in rodents. Proc. Natl. Acad. Sci. U.S.A. **91**:827–831.

Cohen, B., and S. Skiena. 2003. Natural selection and algorithmic design of mRNA. J. Comput. Biol. **10**:419–432.

Cooper, D. N., and M. Krawczak. 1989. Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. Hum. Genet. **83**:181–188.

Deana, A., R. Ehrlich, and C. Reiss. 1996. Synonymous codon selection controls in vivo turnover and amount of mRNA in Escherichia coli bla and ompA genes. J. Bacteriol. **178**:2718–2720.

———. Silent mutations in the *Escherichia coli* ompA leader peptide region strongly affect transcription and translation in vivo. Nucleic Acids Res. **26**:4778–4782.

Deana, A., and C. Reiss. 1993. Stability of messenger RNA of *Escherichia coli* ompA is affected by the use of synonymous codon. C R Acad. of Sci. III **316**:628–632.

Debry, R. W., and W. F. Marzluff. 1994. Selection on silent sites in the rodent H3 histone gene family. Genetics **138**:191–202.

Dermitzakis, E. T., and A. G. Clark. 2002. Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. Mol. Biol. Evol. **19**:1114–1121.

Duan, J., M. S. Wainwright, J. M. Comeron, N. Saitou, A. R. Sanders, J. Gelernter, and P. V. Gejman. 2003. Synonymous

mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. Hum. Mol. Genet. **12**:205–216.

Duret, L. 2000. tRNA gene number and codon usage in the C-elegans genome are co-adapted for optimal translation of highly expressed genes. Trends Genet. **16**:287–289.

———. 2002. Evolution of synonymous codon usage in metazoans. Curr. Opin. Genet. Dev. **12**:640–649.

Duret, L., and L. D. Hurst. 2001. The elevated GC content at exonic third sites is not evidence against neutralist models of isochore evolution. Mol. Biol. Evol. **18**:757–762.

Duret, L., and D. Mouchiroud. 1999. Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, Arabidopsis. Proc. Natl. Acad. Sci. USA **96**:4482–4487.

Duret, L., D. Mouchiroud, and M. Gouy. 1994. HOVERGEN—a database of homologous vertebrate genes. Nucleic Acids Res. **22**:2360–2365.

Eyre-Walker, A. 1991. An analysis of codon usage in mammals: selection or mutation bias? J. Mol. Evol. **33**:442–449.

———. 1999. Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. Genetics **152**:675–683.

Eyre-Walker, A., and M. Bulmer. 1993. Reduced synonymous substitution rate at the start of enterobacterial genes. Nucleic Acids Res. **21**:4599–4603.

Eyre-Walker, A., and P. D. Keightley. 1999. High genomic deleterious mutation rates in hominids. Nature **397**:344–347.

Fickett, J. W., and A. C. Hatzigeorgiou. 1997. Eukaryotic promoter recognition. Genome Res. **7**:861–878.

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

Galtier, N., G. Piganeau, D. Mouchiroud, and L. Duret. 2001. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics **159**:907–911.

Gardiner-Garden, M., and M. Frommer. 1987. CpG islands in vertebrate genomes. J. Mol. Biol. **196**:261–282.

Gottlieb, B., D. M. Vasiliou, R. Lumbroso, L. K. Beitel, L. Pinsky, and M. A. Trifiro. 1999. Analysis of exon 1 mutations in the androgen receptor gene. Hum. Mut. **14**:527–539.

Green, P., B. Ewing, W. Miller, P. J. Thomas, NISC Comparative Sequencing Program, and E. D. Green. 2003. Transcription-associated mutational asymmetry in mammalian evolution. Nat. Genet. **33**:514–517.

Hare, M. P., and S. R. Palumbi. 2003. High intron sequence conservation across three Mammalian orders suggests functional constraints. Mol. Biol. Evol. **20**:969–978.

Hawkins, J. D. 1988. A survey on intron and exon lengths. Nucleic Acids Res **16**:9893–9908.

Hellmann, I., S. Zollner, W. Enard, I. Ebersberger, B. Nickel, and S. Paabo. 2003. Selection on human genes as revealed by comparisons to chimpanzee cDNA. Genome Res. **13**:831–837.

Huang, W., B. H. J. Chang, X. Gu, D. Hewett-Emmett, and W. H. Li. 1997. Sex differences in mutation rate in higher primates estimated from AMG intron sequences. J. Mol. Evol. **44**:463–465.

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

Hural, J. A., M. Kwan, G. Henkel, M. B. Hock, and M. A. Brown. 2000. An intron transcriptional enhancer element regulates IL-4 gene locus accessibility in mast cells. J. Immunol. **165**:3239–3249.

Hurst, L. D., and H. Ellegren. 1998. Sex biases in the mutation rate. Trends Genet. **14**:446–452.

Iida, K., and H. Akashi. 2000. A test of translational selection at 'silent' sites in the human genome: base composition comparisons in alternatively spliced genes. Gene **261**:93–105.

Ikemura, T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. **2**:13–34.

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.

Jonsson, J. J., M. D. Foresman, N. Wilson, and R. S. McIvor. 1992. Intron requirement for expression of the human purine nucleoside phosphorylase gene. Nucleic Acids Res. **20**:3191–3198.

Kanaya, S., Y. Yamada, M. Kinouchi, Y. Kudo, and T. Ikemura. 2001. Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. J. Mol. Evol. **53**:290–298.

Katai, H., J. D. Stephenson, C. P. Simkevich, J. P. Thompson, and R. Raghow. 1992. An AP-1-like motif in the first intron of human Pro alpha 1(I) collagen gene is a critical determinant of its transcriptional activity. Mol. Cell. Biochem. **118**:119–129.

Katz, L., and C. B. Burge. 2003. Widespread selection for local RNA secondary structure in coding regions of bacterial genes. Genome Res. **13**:2042–2051.

Kawada, N., T. Moriyama, A. Ando, T. Koyama, M. Hori, T. Miwa, and E. Imai. 1999. Role of intron 1 in smooth muscle alpha-actin transcriptional regulation in activated mesangial cells in vivo. Kidney International **55**:2338–2348.

Keightley, P. D., and A. Eyre-Walker. 2000. Deleterious mutations and the evolution of sex. Science **290**:331–333.

Kim, C. H., Y. Oh, and T. H. Lee. 1997. Codon optimization for high-level expression of human erythropoietin (EPO) in mammalian cells. Gene **199**:293–301.

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

Lothian, C., and U. Lendahl. 1997. An evolutionarily conserved region in the second intron of the human nestin gene directs gene expression to CNS progenitor cells and to early neural crest cells. Eur. J. Neurosci. **9**:452–462.

Majewski, J. 2003. Dependence of mutational asymmetry on gene-expression levels in the human genome. Am. J. Hum. Genet. **73**:688–692.

Majewski, J., and J. Ott. 2002. Distribution and characterization of regulatory elements in the human genome. Genome Res. **12**:1827–1836.

———. 2003. Amino acid substitutions in the human genome: evolutionary implications of single nucleotide polymorphisms. Gene **305**:167–173.

McClelland, M., and R. Ivarie. 1982. Asymmetrical distribution of CpG in an 'average' mammalian gene. Nucleic Acids Res. **10**:7865–7877.

Moriyama, E. N., and J. R. Powell. 1997. Codon usage bias and tRNA abundance in Drosophila. J. Mol. Evol. **45**:514–523.

Mourier, T., and D. C. Jeffares. 2003. Eukaryotic intron loss. Science **300**:1393–1393.

Oshima, R. G., L. Abrams, and D. Kulesh. 1990. Activation of an intron enhancer within the keratin 18 gene by expression of c-fos and c-jun in undifferentiated F9 embryonal carcinoma cells. Genes Dev. **4**:835–848.

Palmiter, R. D., E. P. Sandgren, M. R. Avarbock, D. D. Allen, and R. L. Brinster. 1991. Heterologous introns can enhance expression of transgenes in mice. Proc. Natl. Acad. Sci. USA **88**:478–482.

Reed, R., and T. Maniatis. 1985. Intron sequences involved in lariat formation during pre-messenger RNA splicing. Cell **41**:95–105.

Rohrer, J., and M. E. Conley. 1998. Transcriptional regulatory elements within the first intron of Bruton's tyrosine kinase. Blood **91**:214–221.

Rossi, P., and B. de Crombrugghe. 1987. Identification of a cell-specific transcriptional enhancer in the first intron of the mouse alpha 2 (type I) collagen gene. Proc. Nat. Acad. Sci. USA **84**:5590–5594.

Sakurai, A., S. Fujimori, H. Kochiwa, S. Kitamura-Abe, T. Washio, R. Saito, P. Carninci, Y. Hayashizaki, and M. Tomita. 2002. On biased distribution of introns in various eukaryotes. Gene **300**:89–95.

Seffens, W., and D. Digby. 1999. mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. Nucleic Acids Res. **27**:1578–1584.

Sharp, P. M., M. Averof, A. T. Lloyd, G. Matassi, and J. F. Peden. 1995. DNA-sequence evolution: the sounds of silence. Phil. Trans. R. Soc. Lond. B **349**:241–247.

Shaw, G., and R. Kamen. 1986. A conserved AU sequence from the 3′ untranslated region of GM-CSF mRNA mediates selective mRNA degradation. Cell **46**:659–667.

Shields, D. C., P. M. Sharp, D. G. Higgins, and F. Wright. 1988. Silent sites in Drosophila genes are not neutral—evidence of selection among synonymous codons. Mol. Biol. Evol. **5**:704–716.

Smith, M. W. 1988. Structure of vertebrate genes: a statistical analysis implicating selection. J. Mol. Evol. **27**:45–55.

Smith, N. G. C., and L. D. Hurst. 1998. Sensitivity of patterns of molecular evolution to alterations in methodology: a critique of Hughes and Yeager. J. Mol. Evol. **47**:493–500.

———. 1999*a*. The causes of synonymous rate variation in the rodent genome: can substitution rates be used to estimate the sex bias in mutation rate? Genetics **152**:661–673.

———. 1999*b*. The effect of tandem substitutions on the correlation between synonymous and nonsynonymous rates in rodents. Genetics **153**:1395–1402.

Sorek, R., and G. Ast. 2003. Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. Genome Res. **13**:1631–1637.

Stenico, M., A. T. Lloyd, and P. M. Sharp. 1994. Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases. Nucleic Acids Res. **22**:2437–2446.

Subramanian, S., and S. Kumar. 2003. Neutral substitutions occur at a faster rate in exons than in noncoding DNA in primate genomes. Genome Res. **13**:838–844.

Suen, T. C., and P. E. Goss. 2001. Identification of a novel transcriptional repressor element located in the first intron of the human BRCA1 gene. Oncogene **20**:440–450.

Sved, J., and A. Bird. 1990. The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. Proc. Nat. Acad. Sci. USA **87**:4692–4696.

Tamura, K., and M. Nei. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial-DNA in humans and chimpanzees. Mol. Biol. Evol. **10**:512–526.

Urrutia, A. O., and L. D. Hurst. 2003. The signature of selection mediated by expression on human genes. Genome Res. **13**:2260–2264.

Vinogradov, A. E. 2001*a*. Bendable genes of warm-blooded vertebrates. Mol. Biol. Evol. **18**:2195–2200.

———. 2001*b*. Within-intron correlation with base composition of adjacent exons in different genomes. Gene **276**:143–151.

———. DNA helix: the importance of being GC-rich. Nucleic Acids Res. **31**:1838–1844.

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.

Waterston, R. H., K. Lindblad-Toh, E. Birney et al. (222 co-authors). 2002. Initial sequencing and comparative analysis of the mouse genome. Nature **420**:520–562.

Wingender, E., X. Chen, R. Hehl, H. Karas, I. Liebich, V. Matys, T. Meinhardt, M. Pruss, I. Reuter, and F. Schacherer. 2000. TRANSFAC: an integrated system for gene expression regulation. Nucleic Acids Res. **28**:316–319.

Workman, C., and A. Krogh. 1999. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. Nucleic Acids Res. **27**:4816–4822.

Zhuang, Y. A., A. M. Goldstein, and A. M. Weiner. 1989. UACUAAC is the preferred branch site for mammalian mRNA splicing. Proceedings National Academy of Sciences U.S.A. **86**:2752–2756.

Pekka Pamilo, Associate Editor

Accepted December 30, 2003