DOI: 10.1534/genetics.104.026351

Selective and Mutational Patterns Associated With Gene Expression in Humans: Influences on Synonymous Composition and Intron Presence

Josep M. Comeron[1]

Department of Biological Sciences, University of Iowa, Iowa City, Iowa 52242

Manuscript received January 8, 2004
Accepted for publication March 23, 2004

ABSTRACT

We report the results of a comprehensive study of the influence of gene expression on synonymous codons, amino acid composition, and intron presence and size in human protein-coding genes. First, in addition to a strong effect of isochores, we have detected the influence of transcription-associated mutational biases (TAMB) on gene composition. Genes expressed in different tissues show diverse degrees of TAMB, with genes expressed in testis showing the greatest influence. Second, the study of tissues with no evidence of TAMB reveals a consistent set of optimal synonymous codons favored in highly expressed genes. This result exposes the consequences of natural selection on synonymous composition to increase efficiency of translation in the human lineage. Third, overall amino acid composition of proteins closely resembles tRNA abundance but there is no difference in amino acid composition in differentially expressed genes. Fourth, there is a negative relationship between expression and CDS length. Significantly, this is observed only among genes with introns, suggesting that the cause for this relationship in humans cannot be associated only with costs of amino acid biosynthesis. Fifth, we show that broadly and highly expressed genes have more, although shorter, introns. The selective advantage for having more introns in highly expressed genes is likely counterbalanced by containment of transcriptional costs and a minimum exon size for proper splicing.

POPULATION genetics theory predicts that traits under weak selection in species with large effective population size (N_e; *e.g.*, *Saccharomyces cerevisiae*, *Drosophila melanogaster*, *Caenorhabditis elegans*, etc.), will exhibit a weaker (if any) signature of selection in humans, with a smaller long-term N_e (Li and Sadler 1991; Wall 2003). Additionally, the ability to detect this signature at a genomic level in the human lineage is complicated by other factors. The most significant one is the heterogeneous structure of the genome, which is observed at the level of nucleotide composition (*i.e.*, isochoric structure; Bernardi 1995; Nekrutenko and Li 2000) and also for features such as expression patterns (D'Onofrio 2002; Lercher *et al.* 2002) and gene structure [*i.e.*, length of the coding sequence (CDS) and introns (Mouchiroud *et al.* 1991; Duret *et al.* 1995; Lander *et al.* 2001)]. Whether these large-scale genomic features have coevolved due to selection is under debate (Duret *et al.* 2002; Lercher *et al.* 2003), but this multifaceted mosaic structure clearly makes the distinction between coincidental and causative associations not trivial. The second confounding factor is the large fraction (>30%) of human genes showing alternative splicing forms (Lander *et al.* 2001), which might obscure the fingerprints of selection on exon composition and intron presence and size. Finally, recent genomic analyses in species with no clear isochore structure such as yeast, Drosophila, and *C. elegans* show that gene composition, length of CDS, and intron features are not independent parameters (Powell and Moriyama 1997; Moriyama and Powell 1998; Comeron *et al.* 1999; Coghlan and Wolfe 2000; Comeron and Kreitman 2000, 2002; Hey and Kliman 2002), and therefore their response to different levels of expression and to selection in general ought to be investigated simultaneously.

The influence of selection on gene composition, especially on the unequal use of synonymous codons, has been the archetypal example of a trait under weak selection in species with large N_e (Sharp and Li 1986; Li 1987; Shields *et al.* 1988; Kliman and Hey 1993; Moriyama and Hartl 1993; Hartl *et al.* 1994; Akashi 1995, 1996, 2003; Powell and Moriyama 1997; Comeron *et al.* 1999; Duret and Mouchiroud 1999; Llopart and Aguade 2000; Begun 2001; McVean and Vieira 2001; Duret 2002; Hey and Kliman 2002; Carlini and Stephan 2003). Indeed, two different features are observed in several model eukaryotes such as *Saccharomyces cerevisiae*, *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Arabidopsis thaliana*. First, differences in synonymous codon usage are associated with differences in expression levels (Ikemura 1985; Sharp and Li 1987; Duret and Mouchiroud 1999; Duret 2000). Second, highly ex-

[1] *Address for correspondence:* Department of Biological Sciences, University of Iowa, 212 Biology Bldg. (BB), Iowa City, IA 52242. E-mail: josep-comeron@uiowa.edu

pressed genes show a set of synonymous codons that correspond mainly to abundant tRNAs (Ikemura 1985; Moriyama and Powell 1997; Duret and Mouchiroud 1999; Coghlan and Wolfe 2000; Duret 2000). The combination of both observations strongly supports the action of selection at the level of translational efficiency (*i.e.*, translational selection), increasing either accuracy or speed of translation (Ikemura 1985; Bulmer 1991; Kurland 1992).

Nevertheless, the evidence supporting translational selection in humans has been—at best—arguable (Eyre-Walker 1999; Iida and Akashi 2000; Urrutia and Hurst 2001; Duret 2002; Galtier 2003) and the isochoric structure of the genome is, with certainty, the most influential factor shaping synonymous composition. Other factors, such as multiple-splicing forms, methodological biases introduced by the use of serial analysis of gene expression (SAGE) or expressed sequence tag (EST) studies (Margulies *et al.* 2001), pooling of data from different tissues, and the overlaying effect of selection at certain synonymous sites influencing pre-mRNA structures (Shen *et al.* 1999; Duan *et al.* 2003), might all have played a part in the inability to detect reliable patterns of translational selection in the human lineage.

On the other hand, it is well known that intron size and presence vary considerably between homologous genes. Several studies have applied population-genetic techniques to provide a primary insight on modes of selection that could explain the proliferation and maintenance of spliceosomal introns as well as their variation in size (Stephan *et al.* 1994; Leicht *et al.* 1995; Carvalho and Clark 1999; Comeron and Kreitman 2000; Llopart *et al.* 2002; Lynch 2002; Schaeffer 2002; Parsch 2003). Moreover, recent studies in humans (Castillo-Davis *et al.* 2002; Urrutia and Hurst 2003) suggest the action of selection favoring short introns in highly expressed genes, possibly due to the beneficial effects of reducing transcriptional costs (time and energy; Castillo-Davis *et al.* 2002). Nonetheless, the evolution of introns cannot be understood independently of the known impact of intronic sequences on downstream mRNA metabolism (Sun and Maquat 2000; Zhou *et al.* 2000; Le Hir *et al.* 2001; Yu *et al.* 2002), splicing efficiency (Klinz and Gallwitz 1985; Sterner *et al.* 1996), and overall gene regulation.

Here, we report a comprehensive study of the influence of gene expression on both composition and intron presence and size in human protein-coding genes with no evidence of multiple-splicing forms. The application of several approaches to take into account background effects and the study of expression (*i.e.*, transcription) in different tissues based on microarray data allow detection of the signature of natural selection on both traits and transcription-associated mutational biases.

MATERIALS AND METHODS

Sequence data: Genomic sequences were obtained from GenBank, Build 31 (January 3, 2003) available at ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/. Information on alternative splicing was extracted from the National Center for Biotechnology Information (NCBI) Reference Sequence database (http://www.ncbi.nlm.nih.gov/RefSeq/ index.html; January, 2003) using nonredundant human mRNA sequences (hs.gbff) for known genes (*i.e.*, predicted genes were not used). Only nuclear genes with complete information on protein-coding sequence, a CDS >300 bp, and no evidence of multiple-splicing forms were included in this study. Altogether, we obtained information for a total of 11,441 genes (list available upon request).

Microarray data: Expression data for human tissues were obtained from a high-throughput gene expression study of the normal mammalian transcriptome (Su *et al.* 2002; http://expression.gnf.org; April 2003) based on hybridization to high-density arrays (GLP91 platform; Affymetrix U95A). A total of 5280 genes (4876 with introns) overlapped with those described above and were used in this study. Presence/absence for each probe/transcript (the Absolute Call) was determined for each sample from the reference series GSE96 using Affymetrix MAS4 algorithm, as reported in NCBI's Gene Expression Omnibus (Edgar *et al.* 2002). Levels of expression were investigated using positive AD values only in genes with validated presence. Nineteen tissues were investigated using a total of 45 samples: adrenal gland, brain (fetal and adult), liver (fetal and adult), heart, kidney, lung, ovary pool, pancreas, pituitary gland, placenta, prostate, salivary gland, spinal cord, spleen, testis, thymus, thyroid, trachea, and uterus.

Two different measures of expression were used in this study: breadth of expression (Expression$_{\text{breadth}}$; Duret and Mouchiroud 2000; Urrutia and Hurst 2001), which is the number of tissues in which transcription is detected (ranging up to 19 tissues), and the level of transcription within each of these 19 tissues (Expression$_{\text{level}}$). Genes are defined as ubiquitously or narrowly expressed if they are expressed in >14 (1497 genes) and <3 (1732 genes) tissues, respectively. We have avoided using measures of expression based on pooled or mean levels of transcription from different tissues because these latter measures are more dependent on breadth of expression (Urrutia and Hurst 2003) and they are highly sensitive to the particular set of tissues chosen for the analysis and to possible differences in overall expression levels. We also avoided using transcription information based on SAGE because of its known GC content bias (Margulies *et al.* 2001) that could not only influence the analysis of expression and composition but also generate spurious clustering of expression across the genome in association with different isochores.

BLAST searches and CpG islands: Local alignments between human and mouse (*Mus musculus*) orthologous intron sequences were used to estimate the number of conserved sites in introns by applying BLASTn searches (Altschul *et al.* 1997). BLASTn searches are highly sensitive to the set of parameters used, most conspicuously when fairly divergent sequences are compared, but there is no reason to presume a systematic bias when comparing broadly and narrowly expressed genes. We used a word size set to 11 and masked off segments of the human introns that have low compositional complexity and human repeats (http://www.ncbi.nlm.nih.gov/blast). The presence of CpG islands was predicted with the program NewCpGreport (EMBOSS v.2.3.1. package) using default parameters. As expected, both approaches reveal higher percentages of CpG islands and conserved sites in first introns than in other introns, which is a positive control for these methods.

Statistical analyses: All correlation coefficients reported in this study were obtained using all genes independently, avoiding the approach of subdividing genes into groups to later investigate relationships among groups. Note that this latter approach is equally valuable to detect statistically significant associations but it cannot be used to assess the actual strength of association. Statistical analyses were carried out using Statistica for Windows v.6 (StatSoft, Tulsa, OK).

RESULTS

Gene expression and nucleotide composition

Gene expression and amino acid composition: Total amino acid composition and tRNA abundance show a positive correlation in prokaryotes and in several eukaryotes, supporting the concept that selection has maximized translation efficiency at the amino acid level (Ikemura 1985; Sharp and Li 1986; Bulmer 1991; Kurland 1992; Duret 2000; Akashi and Gojobori 2002; Akashi 2003). We also observe this trend in human proteins, with an overall amino acid composition that strongly resembles the relative number of corresponding tRNA genes: nonparametric Spearman's rank correlation, $R = +0.527$; $P = 0.017$. [We use the number of isoaccepting tRNA genes as a proxy for cellular abundance of each tRNA (Percudani *et al.* 1997; Duret 2000; Lander *et al.* 2001; Akashi 2003)]. But, contrary to results using SAGE (Urrutia and Hurst 2003) and EST data, when microarray data are used there is no detectable influence of expression on amino acid usage. No amino acid increases its frequency in a gene, in any of the 19 tissues under study, in association with $\text{Expression}_{\text{level}}$ (see materials and methods) after sequential Bonferroni correction for multiple tests (Rice 1989). Congruently with this lack of influence of expression on amino acid usage, there is no evidence for a better fit to tRNA abundance in highly expressed than in poorly expressed genes. For instance, amino acid composition of highly and poorly expressed genes in brain shows similar association with tRNA abundance ($R = +0.548$, $P = 0.012$ and $R = +0.568$, $P = 0.009$, respectively).

Gene expression and synonymous base composition: Many studies have shown that GC content at the third positions of codons (GC3) is greater than GC content at introns (GCi) or at the first and second position of codons (GC12) (our data set shows an average GC content of 58.3, 49.4, and 45.7%, for GC3, GC12, and GCi, respectively). This overall difference has been used, at times, as an argument in favor of selection on synonymous base composition (although see discussion).

Previous studies based on EST data (Duret 2002) revealed a negative relationship between GC3 and $\text{Expression}_{\text{breadth}}$ (see materials and methods), and this is also observed using microarray data ($R = -0.120$, $P < 1 \times 10^{-12}$). On the other hand, the study of $\text{Expression}_{\text{level}}$, which is the measure of expression that is expected to be associated with translational efficiency, shows the opposite tendency, with GC3 increasing with $\text{Expression}_{\text{level}}$ in all 19 tissues with R ranging between $+0.078$ ($P = 6 \times 10^{-6}$) and $+0.393$ ($P < 1 \times 10^{-12}$). The same trend is observed for GC12 and GCi, increasing significantly with $\text{Expression}_{\text{level}}$ also in all 19 tissues, with R between $+0.055$ ($P = 0.0015$) and $+0.234$ ($P < 1 \times 10^{-12}$) for GC12 and between $+0.142$ and $+0.433$ ($P < 1 \times 10^{-12}$) for GCi. The covariation of GC3 and GCi with $\text{Expression}_{\text{level}}$ exposes a strong nonselective component not specific to synonymous composition, with two obvious possible causes for this observation: isochores and transcription-associated mutational biases (TAMB).

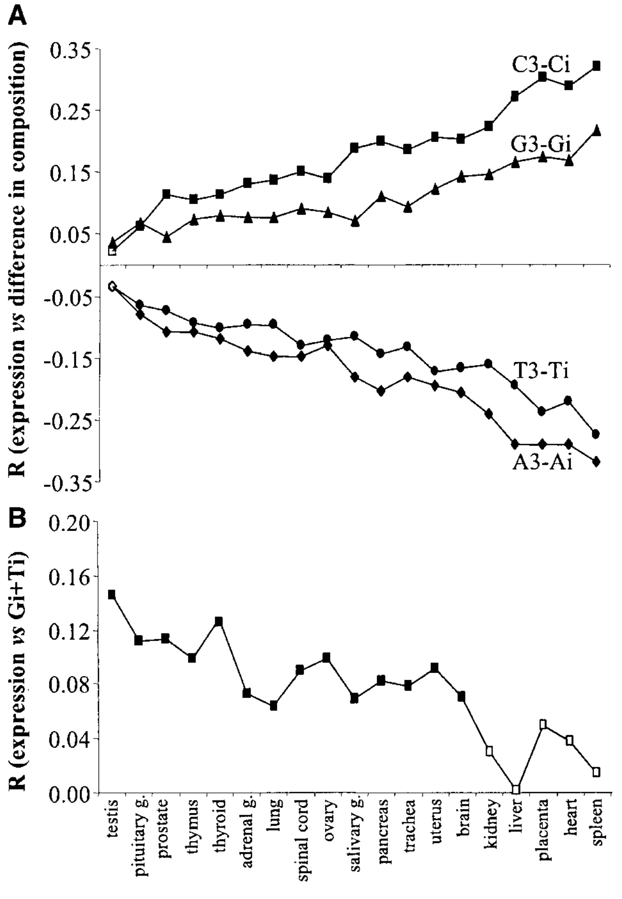

Figure 1.—(A) Association (Spearman's rank correlation R) between levels of expression within each tissue and the difference in composition between third position of codons and introns. Tissues are ordered according to the overall influence of $\text{Expression}_{\text{level}}$. (B) Association ($R$) between expression within each tissue and the G + T content in the coding strand of introns (Gi + Ti) as a measure of strand asymmetry. Open symbols indicate nonsignificant ($P > 0.05$) association after sequential Bonferroni correction.

We investigated the influence of expression on synonymous base composition relative to that in introns for each nucleotide separately. Figure 1A shows the correlation coefficient, R, between $\text{Expression}_{\text{level}}$ and the difference in base composition between the third position of codons and introns. In testis and, to a lesser degree, in prostate and pituitary gland, the influence of expression on synonymous composition can be explained by an equivalent influence of expression on intron composition. On the other hand, tissues like spleen, heart, placenta, liver, or kidney, show that C and, to a lesser degree, G content at synonymous sites increases with expression beyond mutational tendencies operating on whole transcripts, a first indication of translational selection in these tissues.

Transcription-associated mutational biases: Transcription-associated repair is expected to cause strand asymmetries, increasing G relative to C and T relative to A content of the coding strand (Sullivan 1995; Green *et al.* 2003). Then, we investigated consequences of TAMB by measuring G + T content in the coding strand of introns (Gi + Ti; Green *et al.* 2003). As shown in Figure 1B, genes expressed in testis show the greatest influence of expression on Gi + Ti ($R = +0.146$, $P < 1 \times 10^{-12}$), evidencing TAMB, followed by genes expressed in thyroid, prostate, and pituitary gland. Conversely, genes expressed in liver, spleen, kidney, heart, and placenta show no evidence of TAMB. Note that tissues for which a change in synonymous composition cannot be explained by a similar change in intron composition are the same tissues showing weakest TAMB.

Set of optimal synonymous codons in highly expressed genes: To investigate consequences of translational selection, we have looked into the set of synonymous codons that increase in frequency with expression (*i.e.*, optimal codons). A caveat, however, should be mentioned since translational selection depend on aspects of protein translation while we used levels of transcription due to the scarcity of information on protein amounts. Here, we assume that transcript levels are strongly correlated with protein levels. Table 1 shows the difference in the relative synonymous codon usage (RSCU; Sharp and Li 1987; Duret and Mouchiroud 1999; Duret 2000) between highly and poorly expressed genes (ΔRSCU). Highly and poorly expressed genes were defined as the 25% with highest and lowest levels, respectively, of detectable transcription within each tissue, to allow for differences in overall expression levels among tissues. We avoided using the effective number of codons (Wright 1990), which is a measure of overall codon bias that is not influenced by the number of codons under study (Wright 1990; Comeron and Aguade 1998), because it does not directly correct for differences in background composition. On the other hand, a measure of codon bias that corrects for background composition such as MCB (Urrutia and Hurst 2001) is strongly influenced by the length of the CDS in a nonlinear manner (Urrutia and Hurst 2001) and it exposes heterogeneity of synonymous base composition among amino acids rather than bias in synonymous codon usage that might, or might not, be consistent among amino acids. Moreover, neither of these two methods reveal the codons that increase in frequency with expression.

There is no clear prediction on which tissues are most likely to exhibit the strongest link between expression and translational selection. However, we can predict on the basis of the previous results that tissues showing no or minimal TAMB should be those exhibiting clearer, if any, patterns. Congruently, tissues least influenced by TAMB (*e.g.*, liver, spleen, heart, placenta, and kidney) reveal a strong effect of expression on the usage of synonymous codons, with several codons showing a positive ΔRSCU (Table 1). On the other hand, genes expressed in testis (with strongest TAMB) expose only two synonymous codons increasing significantly with expression (the same two codons showing the strongest effect of expression in tissues with no TAMB). These results, however, could be explained without invoking selection if highly expressed genes cluster in GC-rich isochores (see Introduction) and therefore we have compared highly and poorly expressed genes in GC-rich and GC-poor isochores separately (we defined three isochore categories with equivalent gene numbers based on GCi). A conservative definition of optimal codons then refers to codons showing strong positive ΔRSCU in both GC-rich and GC-poor isochores in tissues with no evidence of TAMB (*e.g.*, liver and spleen). A total of 17 optimal codons are consistently observed in tissues with no detectable TAMB, 12 C- and 5 G-ending codons (see Table 1).

In total, the frequency of optimal codons in a gene (Fop) increases with $\text{Expression}_{\text{level}}$ in all tissues, with R ranging from +0.100 (testis) to +0.406 (spleen; $P < 1 \times 10^{-12}$ in all cases). Another measure that explores the overall adaptation of codon usage taking into account background mutational biases (isochoric and/or TAMB) is the ratio of GC-ending optimal to GC-ending nonoptimal codons ($\text{GC}_{\text{optimal}}/\text{GC}_{\text{nonoptimal}}$) in a gene (a ratio that should be computed using only amino acids with both optimal and nonoptimal GC-ending codons, *i.e.*, four- and sixfold degenerate amino acids). As expected if the set of optimal codons properly describes consequences of translational selection, all tissues show a significant increase of $\text{GC}_{\text{optimal}}/\text{GC}_{\text{nonoptimal}}$ with $\text{Expression}_{\text{level}}$, with R ranging from +0.139 to +0.307 ($P < 1 \times 10^{-12}$ in all cases). Note that this latter analysis reveals that the influence of expression on synonymous codon usage is also detectable, although to a much lesser degree, in tissues such as testis where a nonselective component in association with expression (*i.e.*, TAMB) is the main influence on synonymous composition. Figure 2 shows the relationship between $\text{GC}_{\text{optimal}}/\text{GC}_{\text{nonoptimal}}$ and $\text{Expression}_{\text{level}}$ for genes expressed in liver, the tissue with least evidence of TAMB.

Frequency of optimal codons and CDS length: In yeast, Drosophila, and nematodes, measures of codon usage bias correlate negatively with CDS length (Moriyama and Powell 1998; Comeron *et al.* 1999; Duret and Mouchiroud 1999). We show that this is also the

TABLE 1

Difference in codon usage (ΔRSCU) between highly and poorly expressed genes in three different tissues

		Spleen			Liver			Testis		
		All	GC rich	GC poor	All	GC rich	GC poor	All	GC rich	GC poor
Asn	*AAC*[a]	0.37++	0.28+	0.36++	0.32++	0.27+	0.38++	0.09	0.03	0.16
	AAT	−0.37	−0.28	−0.36	−0.32	−0.27	−0.38	−0.09	−0.03	−0.16
Asp	*GAC*	0.36++	0.35++	0.31++	0.30++	0.27+	0.28+	0.04	−0.02	0.09
	GAT	−0.36	−0.35	−0.31	−0.30	−0.27	−0.28	−0.04	0.02	−0.09
Cys	*TGC*	0.35++	0.29+	0.30++	0.27+	0.20+	0.27+	0.06	0.00	0.04
	TGT	−0.35	−0.29	−0.30	−0.27	−0.18	−0.27	−0.06	0.00	−0.04
Gln	CAA	−0.33	−0.28	−0.33	−0.28	−0.28	−0.26	−0.13	−0.10	−0.11
	CAG	0.33++	0.28+	0.33++	0.28+	0.28+	0.26+	0.13	0.10	0.11
Glu	GAA	−0.49	−0.44	−0.46	−0.40	−0.38	−0.41	−0.14	−0.05	−0.19
	GAG	0.49++	0.44++	0.46++	0.40++	0.38++	0.41++	0.14	0.05	0.19
His	*CAC*	0.41++	0.41++	0.36++	0.32++	0.32++	0.34++	0.09	0.07	0.07
	CAT	−0.41	−0.41	−0.36	−0.32	−0.32	−0.34	−0.09	−0.07	−0.07
Lys	AAA	−0.45	−0.41	−0.43	−0.36	−0.33	−0.43	−0.14	−0.09	−0.15
	AAG	0.45++	0.41++	0.43++	0.36++	0.33++	0.43++	0.14	0.09	0.15
Phe	*TTC*	0.43++	0.39++	0.41++	0.37++	0.34++	0.40++	0.12	0.06	0.13
	TTT	−0.43	−0.39	−0.41	−0.37	−0.34	−0.40	−0.12	−0.06	−0.13
Tyr	*TAC*	0.38++	0.31++	0.35++	0.31++	0.26+	0.33++	0.10	0.05	0.17
	TAT	−0.38	−0.31	−0.35	−0.31	−0.26	−0.33	−0.10	−0.05	−0.17
Ile	ATA	−0.35	−0.20	−0.20	−0.32	−0.19	−0.24	−0.11	−0.04	−0.06
	ATC	0.79++	0.47++	0.49++	0.67++	0.40++	0.48++	0.18	0.02	0.16
	ATT	−0.44	−0.26	−0.29	−0.35	−0.21	−0.24	−0.07	0.01	−0.10
Ala	GCA	−0.49	−0.21	−0.24	−0.37	−0.16	−0.21	−0.16	−0.04	−0.09
	GCC	0.61++	0.23+	0.29+	0.51++	0.21+	0.27+	0.14	0.02	0.09
	GCG	0.14	0.08	0.07	0.10	0.03	0.05	0.02	−0.01	0.04
	GCT	−0.26	−0.10	−0.12	−0.24	−0.08	−0.12	0.00	0.03	−0.04
Gly	GGA	−0.59	−0.25	−0.29	−0.46	−0.20	−0.26	−0.21	−0.12	−0.11
	GGC	0.51++	0.23+	0.22+	0.45++	0.22+	0.22+	0.14	0.04	0.12
	GGG	0.26+	0.11	0.14	0.16	0.06	0.08	0.01	0.00	−0.01
	GGT	−0.18	−0.08	−0.07	−0.15	−0.08	−0.05	0.07	0.08	0.00
Pro	CCA	−0.39	−0.16	−0.18	−0.35	−0.19	−0.14	−0.11	−0.07	−0.08
	CCC	0.56++	0.23+	0.25+	0.48++	0.22+	0.26+	0.12	0.05	0.09
	CCG	0.17	0.08	0.08	0.14	0.07	0.05	0.03	0.00	0.05
	CCT	−0.34	−0.15	−0.15	−0.27	−0.11	−0.16	−0.05	0.02	−0.06
Thr	ACA	−0.42	−0.18	−0.20	−0.40	−0.17	−0.22	−0.12	−0.04	−0.08
	ACC	0.59++	0.33++	0.29+	0.53++	0.28+	0.33++	0.19	0.06	0.10
	ACG	0.19	0.07	0.06	0.18	0.08	0.05	0.03	0.01	0.05
	ACT	−0.36	−0.23	−0.15	−0.31	−0.18	−0.16	−0.10	−0.03	−0.08
Val	GTA	−0.37	−0.17	−0.18	−0.32	−0.15	−0.18	−0.08	−0.02	−0.04
	GTC	0.24+	0.12	0.11	0.18	0.09	0.09	0.03	0.00	0.02
	GTG	0.62++	0.26+	0.30++	0.51++	0.24+	0.26+	0.15	0.01	0.09
	GTT	−0.50	−0.21	−0.23	−0.38	−0.17	−0.17	−0.10	0.01	−0.06
Arg	AGA	−1.00	−0.29	−0.31	−0.84	−0.26	−0.28	−0.33	−0.07	−0.11
	AGG	−0.22	−0.07	−0.07	−0.18	−0.05	−0.10	−0.16	0.01	−0.11
	CGA	−0.17	−0.06	−0.05	−0.13	−0.05	−0.02	−0.04	−0.02	−0.02
	CGC	0.75++	0.21+	0.24+	0.66++	0.23+	0.25+	0.27+	0.05	0.13
	CGG	0.67++	0.21+	0.20+	0.47++	0.18	0.13	0.18	0.01	0.10
	CGT	−0.03	0.00	−0.01	0.03	0.02	0.03	0.08	0.02	0.00
Leu	TTA	−0.51	−0.14	−0.16	−0.44	−0.14	−0.15	−0.16	−0.05	−0.04
	TTG	−0.33	−0.11	−0.11	−0.25	−0.08	−0.11	−0.03	0.01	−0.03
	CTA	−0.19	−0.06	−0.05	−0.20	−0.07	−0.05	−0.08	−0.02	−0.03
	CTC	0.34++	0.09	0.09	0.28+	0.08	0.09	0.06	−0.01	0.01
	CTG	1.16++	0.34++	0.36++	0.96++	0.31++	0.35++	0.31++	0.04	0.13
	CTT	−0.46	−0.12	−0.13	−0.35	−0.10	−0.12	−0.09	0.02	−0.05
Ser	AGC	0.50++	0.15	0.14	0.43++	0.13	0.15	0.04	−0.01	0.05
	AGT	−0.37	−0.13	−0.11	−0.34	−0.11	−0.11	−0.12	0.00	−0.06
	TCA	−0.38	−0.08	−0.13	−0.36	−0.09	−0.14	−0.10	−0.03	−0.04
	TCC	0.44++	0.14	0.14	0.41++	0.13	0.15	0.17	0.03	0.08
	TCG	0.11	0.03	0.03	0.11	0.03	0.04	0.06	0.01	0.01
	TCT	−0.30	−0.10	−0.08	−0.25	−0.08	−0.10	−0.05	0.00	−0.04

[a] Optimal codons (in underlined italic) defined by a positive ΔRSCU (>0.2) in both GC-rich and GC-poor isochores in tissues with no detectable TAMB (liver and spleen). ++, ΔRSCU > 0.3; +, ΔRSCU > 0.2.

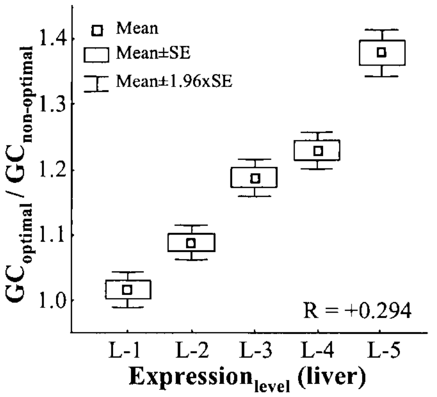

FIGURE 2.—Relationship between the level of expression in the liver and the ratio of GC-ending optimal to GC-ending nonoptimal codons in a gene. Genes are divided according to their level of transcription into five classes with equivalent sample size, from lowest (L-1) to highest (L-5) level. Only four- and sixfold degenerate amino acids were used (see text). Spearman's rank correlation (*R*) is measured using all genes independently.

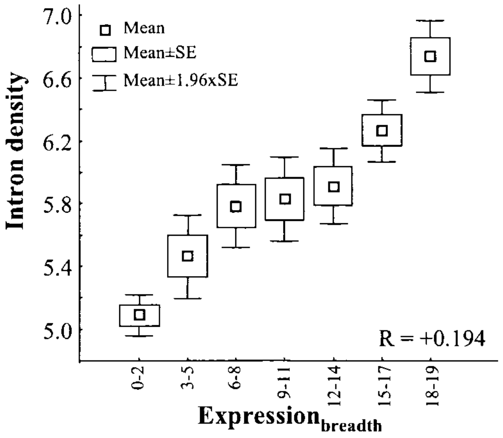

FIGURE 3.—Relationship between gene expression ($Expression_{breadth}$) and intron density (number of introns per kilobase of CDS) in human genes. *R* is measured using all genes independently.

case in humans, with a negative relationship between Fop and CDS length ($R = -0.076$, $P < 1 \times 10^{-12}$), although this is not unexpected because both parameters exhibit a significant association with $Expression_{level}$, positive and negative (see below), respectively. More informative are multivariate analyses showing that Fop decreases with CDS length after taking into account $Expression_{level}$ in the different tissues: *B* ranges from -0.067 ($P = 6 \times 10^{-4}$) to -0.135 ($P < 1 \times 10^{-12}$).

Gene expression, CDS length, and intron presence

Gene expression and CDS length: There is a negative relationship between expression and the length of the CDS. This effect is detected using $Expression_{breadth}$ ($R = -0.088$, $P = 1.4 \times 10^{-10}$) and $Expression_{level}$ in any of the 19 tissues investigated (*R* ranges between -0.118 and -0.204, $P < 1 \times 10^{-12}$ in all cases). Equivalent results based on pooled SAGE data have been recently reported (URRUTIA and HURST 2003), and this study broadens the validity of this relationship in humans.

Significantly, the negative correlation between expression and CDS length is observed only among genes with introns ($R = -0.111$ for $Expression_{breadth}$ and *R* between -0.124 and -0.211 for $Expression_{level}$, $P < 1 \times 10^{-12}$ in all cases). In contrast, there is no detectable correlation among genes without introns using either $Expression_{breadth}$ or $Expression_{level}$ (all associations are statistically nonsignificant after sequential Bonferroni correction). Nevertheless, genes without introns are usually shorter (average 984 bp) than genes with introns (average 1505 bp) and one could argue that the relationship between expression and CDS length is detected only among genes with intermediate/long CDS. We have then analyzed genes with introns and short CDS, a subset with an average CDS length of 983 bp, and observed again a negative relationship between expression and CDS length, for both $Expression_{breadth}$ and $Expression_{level}$ in all tissues ($P < 3 \times 10^{-10}$). This latter result indicates that the observed distinct behavior of genes with and without introns is not attributable to differences in CDS length.

Gene expression and intron density: Several cellular processes associated with intron presence and size might influence the final amount of mRNA correctly transcribed, spliced, and exported to the cytoplasm. In this regard, various selective models (see DISCUSSION) forecast an association of levels of gene expression with differences in intron presence and size among genes.

Predictably, intron number increases with the length of CDS ($R = +0.665$, $P < 1 \times 10^{-12}$). Therefore, to investigate the influence of expression on intron presence and because of the aforementioned association between expression and CDS length, we have studied measures of intron presence relative to the size of the CDS (*i.e.*, intron density; number of introns per kilobase of CDS). Intron density increases with any measure of expression: $Expression_{breadth}$ ($R = +0.194$, $P < 1 \times 10^{-12}$) and $Expression_{level}$ in all tissues (*R* ranging between $+0.114$ and $+0.204$; $P < 1 \times 10^{-12}$ in all cases; Figure 3). The same results are obtained when multiple regression analyses that account for variation in CDS length are performed: $B = +0.185$ ($P < 1 \times 10^{-12}$) for $Expression_{breadth}$ and *B* ranges from $+0.034$ ($P = 0.01$) to $+0.107$ ($P < 1 \times 10^{-12}$) for $Expression_{level}$. These results are not caused simply by intron-less genes being narrowly/lowly expressed because the same trend is obtained when only genes with introns are analyzed: $R = +0.177$ and $R > +0.102$ ($P < 1 \times 10^{-12}$ in all cases), for $Expression_{breadth}$ and $Expression_{level}$, respectively.

The tendency of broadly/highly expressed genes to

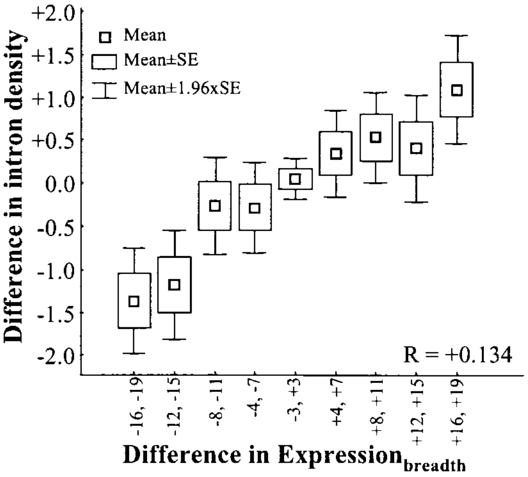

Figure 4.—Comparison of gene expression (Expression$_{breadth}$) and intron density (number of introns per kilobase of CDS) between physically adjacent genes. When comparing two adjacent genes, difference indicates Expression$_{breadth}$ (or intron density) of the second gene minus that of the first gene in the order present in the genomic sequences. *R* is measured using all independent comparisons.

have higher intron density is somewhat unexpected under the hypothesis that highly expressed genes tend to reduce transcriptional costs (time and energy; Castillo-Davis *et al.* 2002). This hypothesis was proposed on the basis of a negative relationship between expression levels and intron size pooling EST data and among genes expressed in brain (Castillo-Davis *et al.* 2002) or using SAGE data (Urrutia and Hurst 2003). Our analyses also show this same trend, with a negative association between expression and average intron size using Expression$_{breadth}$ ($R = -0.075$, $P = 1 \times 10^{-7}$) and Expression$_{level}$ in all 19 tissues, with *R* ranging from -0.038 ($P = 0.008$) to -0.248 ($P < 1 \times 10^{-12}$). Altogether, these results suggest that the reduction in intron size may well be the result of selection to reduce transcriptional costs but also that other factors might operate favoring an increase in intron density in highly/broadly expressed genes even though this will increase transcript length.

Possible influence of isochores: We considered the possible coincidental basis to our previous results because both transcription patterns and gene structures differ among isochores (Mouchiroud *et al.* 1991; Duret *et al.* 1995; Zoubak *et al.* 1996; Lander *et al.* 2001; D'Onofrio 2002). We compared patterns of expression and intron density between physically adjacent genes, both with expression data, hence removing background tendencies even under a restrictive definition of isochore (Nekrutenko and Li 2000). This analysis is possible only by using Expression$_{breadth}$ because too few adjacent genes show expression data in the same tissue. As shown in Figure 4, when two adjacent genes differ in breadth of expression, the gene expressed in a greater number of tissues shows a higher density of introns (2377 gene pairs, $R = +0.134$, $P < 1 \times 10^{-12}$).

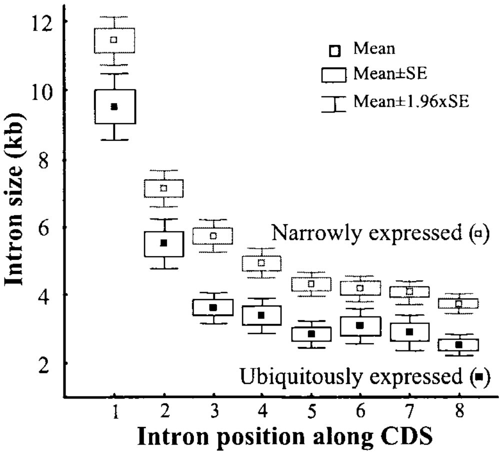

Figure 5.—Intron size along (from proximal to terminal position) human genes. Comparison between ubiquitously (Expression$_{breadth}$ >14 tissues; 1497 genes) and narrowly (Expression$_{breadth}$ <3 tissues; 1732 genes) expressed genes.

Possible influence of intron position along genes: In humans, as in most eukaryotes, first introns (*i.e.*, introns closer to the start codon) tend to be longer than other introns. Indeed, there is a reduction in intron size with respect to their position along a gene (from 5′ to 3′; $R = -0.161$, $P < 1 \times 10^{-12}$). Note that there is a gradual decay in intron length from proximal to terminal positions along a CDS and this is observed in both ubiquitously and narrowly expressed genes ($R = -0.146$ and $R = -0.172$, respectively; $P < 1 \times 10^{-12}$ in both cases). Therefore, intron size and intron number are not independent parameters and we have investigated the influence of expression on intron size taking into account intron position. We confirmed the general trend of ubiquitously expressed genes having shorter introns than narrowly expressed genes and report that this tendency is observed at any given intron position (Figure 5). Further, the relative difference in intron size is least obvious for the first intron (a reduction of ~17%) compared to any other intron where the reduction is close to 30%, hence supporting the perception that first introns contain a larger number of regulatory elements for transcription control (Majewski and Ott 2002).

DISCUSSION

We have used two measures of gene expression based on microarray data: the number of tissues in which a gene is transcribed (breadth of expression) and the level of transcription within tissues. This design has allowed us to investigate the different behavior of both measures and, more importantly, to query for differences among tissues, a required feature if we wanted to assess possible consequences of TAMB since they are expected to be tissue specific. The results shown here indicate that the mutational/selective tendencies associated with both measures of expression (Expression$_{\text{breadth}}$ and Expression$_{\text{level}}$) should not be presumed to be alike; we show comparable outcomes when studying CDS size and intron presence and size and show opposite tendencies for gene composition.

This study also illustrates that results based on SAGE or EST data are equivalent to those based on gene chips when investigating intron features but very different, even conflicting, when investigating base composition. The incongruence between methods is likely caused by the known bias in the quantitative aspect (*i.e.*, Expression$_{\text{level}}$) of the SAGE and EST methods relative to GC content (Margulies *et al.* 2001; Duret 2002). On the other hand, qualitative studies (*i.e.*, Expression$_{\text{breadth}}$ since it is based on presence/absence) are expected to be comparable between SAGE/EST and microarray data, as observed. Another noteworthy difference between this and previous studies is that we have used only genes that have not yet shown evidence of multiple-splicing forms. The use of genes with multiple-splicing forms would introduce a certain degree of ambiguity when composition is investigated because constitutive and facultative exons differ in synonymous GC content (Iida and Akashi 2000).

Certainly, the major determinants of various gene features in mammals are the isochore in which they are located and the functional properties of the encoded proteins. In accordance, in this study we show associations that, although with great statistical significance, explain individually only a small percentage of the overall variance in gene composition or intron features (2–16% of the overall variance).

Gene expression and amino acid composition: Selection at the level of amino acid composition might favor reducing energetic costs of amino acid biosynthesis (Akashi and Gojobori 2002) or act in association with the abundance of tRNAs for each amino acid, increasing translation accuracy or reducing translation costs of proofreading (time and energy). There is no *a priori* reason to expect the fitness consequence of an amino acid misincorporation to be proportional to the degree of expression of a protein (Bulmer 1991). Conversely, the fitness costs associated with amino acid biosynthesis and proofreading will increase with expression. In the prokaryotes *Escherichia coli* and *Bacillus subtilis*, the usage of less energetically costly amino acids increases in abundant proteins (Akashi and Gojobori 2002). In *C. elegans* (Duret 2000) and yeast (Akashi 2003), highly transcribed protein-coding genes show a stronger correlation between amino acid composition and tRNA abundance than do poorly transcribed genes, supporting the proposal that selection minimizes translational costs. In humans, the overall amino acid composition of proteins also matches tRNA abundance but there is no support for different amino acid composition in differentially expressed genes. Therefore, the data suggest that the coevolution of amino acid composition and tRNA abundance in the human lineage is driven by selection to minimize amino acid misincorporation during translation and not to reduce translational costs.

Selection and mutational biases on synonymous composition: The observed difference in GC content between synonymous sites and introns (see results) can be explained under a nonselective scenario by arguing that transposable elements (TEs), which have a reduced GC content (Duret *et al.* 1995), represent a frequent component of many introns. The difference between synonymous sites and introns then might just reflect the recent insertion of TEs in introns, especially in genes with long introns (Duret and Hurst 2001). In partial agreement with this proposal, there is a strong tendency for long introns to have reduced GC content ($R = -0.507$, $P < 1 \times 10^{-12}$). Nevertheless, our analyses indicate that highly transcribed genes show the strongest compositional difference between synonymous sites and introns while, at the same time, have shorter introns and a reduced number of TEs in introns (Castillo-Davis *et al.* 2002). This would argue against the possibility that the positive association between expression and compositional difference between synonymous sites and introns is a consequence of TE presence.

In addition to a strong effect of isochores, we have also detected the influence of transcription-associated mutational biases evidenced by compositional strand bias in introns. Although TAMB is expected to be apparent only in genes expressed in germline cells (Hanawalt 1994; Svejstrup 2002), recent analyses suggest that some level of germline transcription may involve a large fraction of human genes (Green *et al.* 2003; Majewski 2003). Here, we report that the influence of expression on strand bias varies widely among tissues, with genes expressed in testis showing the greatest influence while genes expressed in tissues such as liver, spleen, heart, placenta, and kidney show no evidence of TAMB. Genes expressed in tissues with significant TAMB will be subject to conflicting mutational and selective pressures on synonymous composition beyond the isochore effects. As a result, tissues showing TAMB also reveal the least obvious influence of selection on synonymous codon usage. Conversely, tissues showing little or no evidence of TAMB are those in which selection on synonymous composition is better observed.

Translational selection in humans: Overall, the results shown here are evidence that selection on synonymous codons is operating at a detectable level in the human lineage. As predicted by population genetics theory, however, the signature of translational selection is less conspicuous in humans than in species with much larger N_e. Indeed, genomic patterns of selection on synonymous codons are distinguished only after taking into account the strong influence of background composition (isochores) and tissue-specific features such as TAMB.

We propose a set of 17 synonymous optimal codons selectively favored in highly expressed genes. All optimal codons are GC ending and they resemble the set proposed for *D. melanogaster* more closely than that for *C. elegans* or *A. thaliana* (Duret and Mouchiroud 1999). The comparison between optimal codons and gene copy numbers of isoaccepting tRNAs (expected to reflect cellular tRNA abundance) shows a good, although not perfect, association, with 14 of the 17 optimal codons being decoded by the most frequent isoaccepting tRNA according to classical rules of codon-anticodon interactions (Ikemura 1985). In agreement with the proposal of translational selection, two amino acids (glycine and proline) show a corresponding change in codon preference and tRNA abundance when *C. elegans* and humans are compared, generating in both species precise, although different, matches between optimal codons and the most frequent isoaccepting tRNA. For instance, in the case of glycine, the optimal codon and the anticodon of the most frequent tRNA in *C. elegans* are GGA and UCC, respectively (Duret 2000); in humans, the optimal codon and the anticodon of the most frequent tRNA are GGC and GCC, respectively. Certainly, the use of optimal codons will increase our capability to explore further consequences of translational selection at both intra- and interspecific levels. Further, the exposure of translational selection in the human lineage is a factor that should be introduced into evolutionary analyses that often assume neutrality of all synonymous mutations.

Gene expression and CDS size: The negative relationship between expression and protein size reported in *S. cerevisiae* (Akashi 2003) and *C. elegans* (Jansen and Gerstein 2000) has been explained by the selective advantage of reducing energetic costs of amino acid biosynthesis in highly expressed genes (Akashi and Gojobori 2002; Akashi 2003). On the other hand, the overall excess of deletions over insertions described in many eukaryotes, including mammals (Ogata *et al.* 1996; Ophir and Graur 1997), and the possibility of transcription-associated deletions could also generate a nonselective association between transcription rates in germinal cells and a reduction in protein size. Thus, in multicellular organisms a negative relationship between expression and protein size does not require a selective explanation unless such a relationship is observed among genes not transcribed in germinal cells.

We have shown a negative association between protein size and Expression$_{\text{breadth}}$. Because broadly expressed genes are also more likely to be expressed in germinal cells (Duret and Mouchiroud 2000), this observation alone would not rule out a mutational (transcription-associated) cause. However, the same trend is also observed using Expression$_{\text{level}}$, including tissues with no detectable mutational trends expected in germinal cells, hence supporting a selective explanation for the association between expression and CDS size in the human lineage. Interestingly, this trend is specific to genes with introns, suggesting that protein size is not the sole factor playing a role in this relationship. Thus, the results indicate that the association between expression and CDS size should be investigated not only by selective models based on total protein size (*e.g.*, on costs of amino acid biosynthesis) but also in conjunction with models based on the evolutionary/metabolic consequences of exon size and intron presence (see below).

Gene expression and intron presence and size: Previous reports showed that short introns are favored in highly expressed genes and this study confirms this trend in a wide range of different tissues. Altogether, these results support the hypothesis of a measurable selective advantage for having small transcripts to reduce transcriptional costs (time and energy; Castillo-Davis *et al.* 2002). Significantly, we show a counterbalancing trend that is not caused by background tendencies, instigating broadly/highly expressed genes to have higher intron density in the human lineage. One could argue that genes expressed in many tissues are more likely to have more introns because they are more likely to be alternatively spliced, but these multiple-splicing forms have not yet been detected. Nevertheless, the same trend is observed using Expression$_{\text{level}}$ in specific tissues, ruling out the possibility of a spurious relationship between intron density/number and, at least, Expression$_{\text{level}}$.

Selective causes favoring intron presence: A heterogeneous group of selective causes might associate intron presence in protein-coding genes with levels of correct gene products. At this level, the advantage for having higher intron density would be counterbalanced by a minimum exon size required for proper splicing (Upholt and Sandell 1986; Dominski and Kole 1991) and restrictions on transcription costs that are likely to be species specific, hence explaining differences among species.

A first possibility is that genes with a broader and/or higher expression require an increased number of regulatory signals in different introns. We have applied two indirect approaches to investigate the presence of regulatory regions (see materials and methods for details). Our survey of CpG islands reveals an equivalent presence in narrowly and ubiquitously expressed genes, with an average of 1.67 and 1.64 islands per gene, respectively, comparing genes with >10 introns. In the second approach, we applied BLASTn searches to identify conserved segments of noncoding DNA as a proxy for functionally important sequences (Hardison *et al.* 1997; Jareborg *et al.* 1999; Wasserman *et al.* 2000; Shabalina *et al.* 2001). The comparison of human and mouse orthologous sequences reveals that the total number of conserved sites in introns does not increase with breadth of expression. On the contrary, narrowly expressed genes have an average of 473 conserved sites in introns compared to 372 in ubiquitously expressed genes (Mann-Whitney *U*-test, $P = 0.020$), with percentages of conserved sites of 2.9 and 2.5%, respectively (*U*-

test, $P = 0.029$). In all, these indirect analyses suggest that differences in intron number are not likely a consequence of an increased number of regulatory signals distributed in different introns.

A second explanation for the observed association between gene expression and intron density might be related to the influence of introns on mRNA metabolism (Sun and Maquat 2000; Zhou *et al.* 2000; Le Hir *et al.* 2001; Yu *et al.* 2002) and splicing efficiency (Klinz and Gallwitz 1985; Sterner *et al.* 1996). The so-called exon-exon junction complexes (EJC) are deposited upstream of intron positions after splicing (Kataoka *et al.* 2000; Le Hir *et al.* 2000, 2001) and there is evidence that EJC enhance export efficiency of spliced mRNAs to the cytoplasm (Zhou *et al.* 2000; Le Hir *et al.* 2001). Also, splicing factors might promote transcriptional elongation (Fong and Zhou 2001). Therefore, selection could be acting at the level of intron density to increase mRNA transport and/or transcriptional elongation, especially in highly expressed genes.

Another selective model for intron presence is associated with the deleterious consequences of linkage between sites under selection, a phenomenon termed the Hill-Robertson effect (Hill and Robertson 1966; Felsenstein 1974; see also Li 1987; Kliman and Hey 1993; Comeron *et al.* 1999; McVean and Charlesworth 2000; Tachida 2000; Betancourt and Presgraves 2002; Comeron and Kreitman 2002; Hey and Kliman 2002). Specifically, Comeron and Kreitman (2000, 2002) have proposed that the Hill-Robertson effect might be detectable at an intragenic level in many eukaryotes due to the prevalence of mutations under weak selection in coding regions. Under this model, introns (generally with a reduced frequency of sites under selection compared to exons) will reduce the Hill-Robertson effect at the intragenic level, *i.e.*, intron-containing genes would exhibit increased effectiveness of selection. Then, all else being equal, highly expressed genes would benefit from high intron density to maximize the consequences of selection on amino acid and synonymous composition. A fraction of replacement (amino acid changing) mutations in many species are likely under weak selection and our report of selection on synonymous mutations increases the likelihood of detectable Hill-Robertson effect within genes in the human lineage, particularly in highly expressed genes. Upcoming large-scale population genetics analyses based on polymorphism and divergence data will allow testing of these possibilities.

I thank A. Llopart for fruitful discussions and many comments on the manuscript. C. Castillo-Davis, J. Hey, and S. Schaeffer provided useful suggestions to improve this manuscript. This research was partly funded by The Old Gold Summer Fellowship Program (University of Iowa).

LITERATURE CITED

Akashi, H., 1995 Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. Genetics **139:** 1067–1076.

Akashi, H., 1996 Molecular evolution between *Drosophila melanogaster* and *D. simulans*: reduced codon bias, faster rates of amino acid substitution, and larger proteins in *D. melanogaster.* Genetics **144:** 1297–1307.

Akashi, H., 2003 Translational selection and yeast proteome evolution. Genetics **164:** 1291–1303.

Akashi, H., and T. Gojobori, 2002 Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis.* Proc. Natl. Acad. Sci. USA **99:** 3695–3700.

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

Begun, D. J., 2001 The frequency distribution of nucleotide variation in Drosophila simulans. Mol. Biol. Evol. **18:** 1343–1352.

Bernardi, G., 1995 The human genome: organization and evolutionary history. Annu. Rev. Genet. **29:** 445–476.

Betancourt, A. J., and D. C. Presgraves, 2002 Linkage limits the power of natural selection in Drosophila. Proc. Natl. Acad. Sci. USA **99:** 13616–13620.

Bulmer, M., 1991 The selection-mutation-drift theory of synonymous codon usage. Genetics **129:** 897–907.

Carlini, D. B., and W. Stephan, 2003 In vivo introduction of unpreferred synonymous codons into the Drosophila Adh gene results in reduced levels of ADH protein. Genetics **163:** 239–243.

Carvalho, A. B., and A. G. Clark, 1999 Intron size and natural selection. Nature **401:** 344.

Castillo-Davis, C. I., S. L. Mekhedov, D. L. Hartl, E. V. Koonin and F. A. Kondrashov, 2002 Selection for short introns in highly expressed genes. Nat. Genet. **31:** 415–418.

Coghlan, A., and K. H. Wolfe, 2000 Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae.* Yeast **16:** 1131–1145.

Comeron, J. M., and M. Aguade, 1998 An evaluation of measures of synonymous codon usage bias. J. Mol. Evol. **47:** 268–274.

Comeron, J. M., and M. Kreitman, 2000 The correlation between intron length and recombination in Drosophila: dynamic equilibrium between mutational and selective forces. Genetics **156:** 1175–1190.

Comeron, J. M., and M. Kreitman, 2002 Population, evolutionary and genomic consequences of interference selection. Genetics **161:** 389–410.

Comeron, J. M., M. Kreitman and M. Aguade, 1999 Natural selection on synonymous sites is correlated with gene length and recombination in Drosophila. Genetics **151:** 239–249.

Dominski, Z., and R. Kole, 1991 Selection of splice sites in pre-mRNAs with short internal exons. Mol. Cell. Biol. **11:** 6075–6083.

D'Onofrio, G., 2002 Expression patterns and gene distribution in the human genome. Gene **300:** 155–160.

Duan, J., M. S. Wainwright, J. M. Comeron, N. Saitou, A. R. Sanders *et al.*, 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.

Duret, L., 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, and Arabidopsis. Proc. Natl. Acad. Sci. USA **96:** 4482–4487.

Duret, L., and D. Mouchiroud, 2000 Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol. Biol. Evol. **17:** 68–74.

Duret, L., D. Mouchiroud and C. Gautier, 1995 Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. J. Mol. Evol. **40:** 308–317.

Duret, L., M. Semon, G. Piganeau, D. Mouchiroud and N. Galtier, 2002 Vanishing GC-rich isochores in mammalian genomes. Genetics **162:** 1837–1847.

Edgar, R., M. Domrachev and A. E. Lash, 2002 Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. **30:** 207–210.

Eyre-Walker, A., 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.

Felsenstein, J., 1974 The evolutionary advantage of recombination. Genetics **78:** 737–756.

Fong, Y. W., and Q. Zhou, 2001 Stimulatory effect of splicing factors on transcriptional elongation. Nature **414:** 929–933.

Galtier, N., 2003 Gene conversion drives GC content evolution in mammalian histones. Trends Genet. **19:** 65–68.

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

Hanawalt, P. C., 1994 Transcription-coupled repair and human disease. Science **266:** 1957–1958.

Hardison, R. C., J. Oeltjen and W. Miller, 1997 Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. Genome Res. **7:** 959–966.

Hartl, D. L., E. N. Moriyama and S. A. Sawyer, 1994 Selection intensity for codon bias. Genetics **138:** 227–234.

Hey, J., and R. M. Kliman, 2002 Interactions between natural selection, recombination and gene density in the genes of Drosophila. Genetics **160:** 595–608.

Hill, W. G., and A. Robertson, 1966 The effect of linkage on limits to artificial selection. Genet. Res. **8:** 269–294.

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.

Jansen, R., and M. Gerstein, 2000 Analysis of the yeast transcriptome with structural and functional categories: characterizing highly expressed proteins. Nucleic Acids Res. **28:** 1481–1488.

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.

Kataoka, N., J. Yong, V. N. Kim, F. Velazquez, R. A. Perkinson *et al.*, 2000 Pre-mRNA splicing imprints mRNA in the nucleus with a novel RNA-binding protein that persists in the cytoplasm. Mol. Cell **6:** 673–682.

Kliman, R. M., and J. Hey, 1993 Reduced natural selection associated with low recombination in *Drosophila melanogaster*. Mol. Biol. Evol. **10:** 1239–1258.

Klinz, F. J., and D. Gallwitz, 1985 Size and position of intervening sequences are critical for the splicing efficiency of pre-mRNA in the yeast Saccharomyces cerevisiae. Nucleic Acids Res. **13:** 3791–3804.

Kurland, C., 1992 Translational accuracy and the fitness of bacteria. Annu. Rev. Genet. **26:** 29–50.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody *et al.*, 2001 Initial sequencing and analysis of the human genome. Nature **409:** 860–921.

Le Hir, H., M. J. Moore and L. E. Maquat, 2000 Pre-mRNA splicing alters mRNP composition: evidence for stable association of proteins at exon-exon junctions. Genes Dev. **14:** 1098–1108.

Le Hir, H., D. Gatfield, E. Izaurralde and M. J. Moore, 2001 The exon-exon junction complex provides a binding platform for factors involved in mRNA export and nonsense-mediated mRNA decay. EMBO J. **20:** 4987–4997.

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.

Lercher, M. J., A. O. Urrutia and L. D. Hurst, 2002 Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nat. Genet. **31:** 180–183.

Lercher, M. J., A. O. Urrutia, A. Pavlicek and L. D. Hurst, 2003 A unification of mosaic structures in the human genome. Hum. Mol. Genet. **12:** 2411–2415.

Li, W. H., 1987 Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. J. Mol. Evol. **24:** 337–345.

Li, W. H., and L. A. Sadler, 1991 Low nucleotide diversity in man. Genetics **129:** 513–523.

Llopart, A., and M. Aguade, 2000 Nucleotide polymorphism at the RpII215 gene in *Drosophila subobscura.* Weak selection on synonymous mutations. Genetics **155:** 1245–1252.

Llopart, A., J. M. Comeron, F. G. Brunet, D. Lachaise and M. Long, 2002 Intron presence-absence polymorphism in Drosophila driven by positive Darwinian selection. Proc. Natl. Acad. Sci. USA **99:** 8121–8126.

Lynch, M., 2002 Intron evolution as a population-genetic process. Proc. Natl. Acad. Sci. USA **99:** 6118–6123.

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.

Margulies, E. H., S. L. Kardia and J. W. Innis, 2001 Identification and prevention of a GC content bias in SAGE libraries. Nucleic Acids Res. **29:** E60.

McVean, G. A., and B. Charlesworth, 2000 The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. Genetics **155:** 929–944.

McVean, G. A., and J. Vieira, 2001 Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in Drosophila. Genetics **157:** 245–257.

Moriyama, E. N., and D. L. Hartl, 1993 Codon usage bias and base composition of nuclear genes in Drosophila. Genetics **134:** 847–858.

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

Moriyama, E. N., and J. R. Powell, 1998 Gene length and codon usage bias in *Drosophila melanogaster, Saccharomyces cerevisiae* and *Escherichia coli.* Nucleic Acids Res. **26:** 3188–3193.

Mouchiroud, D., G. D'Onofrio, B. Aissani, G. Macaya, C. Gautier *et al.*, 1991 The distribution of genes in the human genome. Gene **100:** 181–187.

Nekrutenko, A., and W. H. Li, 2000 Assessment of compositional heterogeneity within and between eukaryotic genomes. Genome Res. **10:** 1986–1995.

Ogata, H., W. Fujibuchi and M. Kanehisa, 1996 The size differences among mammalian introns are due to the accumulation of small deletions. FEBS Lett. **390:** 99–103.

Ophir, R., and D. Graur, 1997 Patterns and rates of indel evolution in processed pseudogenes from humans and murids. Gene **205:** 191–202.

Parsch, J., 2003 Selective constraints on intron evolution in Drosophila. Genetics **165:** 1843–1851.

Percudani, R., A. Pavesi and S. Ottonello, 1997 Transfer RNA gene redundancy and translational selection in *Saccharomyces cerevisiae.* J. Mol. Biol. **268:** 322–330.

Powell, J. R., and E. N. Moriyama, 1997 Evolution of codon usage bias in Drosophila. Proc. Natl. Acad. Sci. USA **94:** 7784–7790.

Rice, W. R., 1989 Analyzing tables of statistical tests. Evolution **43:** 223–225.

Schaeffer, S. W., 2002 Molecular population genetics of sequence length diversity in the Adh region of Drosophila pseudoobscura. Genet. Res. **80:** 163–175.

Shabalina, S. A., A. Y. Ogurtsov, V. A. Kondrashov and A. S. Kondrashov, 2001 Selective constraint in intergenic regions of human and mouse genomes. Trends Genet. **17:** 373–376.

Sharp, P. M., and W. H. Li, 1986 An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol. **24:** 28–38.

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

Shen, L. X., J. P. Basilion and V. P. Stanton, Jr., 1999 Single-nucleotide polymorphisms can cause different structural folds of mRNA. Proc. Natl. Acad. Sci. USA **96:** 7871–7876.

Shields, D., P. Sharp, D. 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.

Stephan, W., V. S. Rodriguez, B. Zhou and J. Parsch, 1994 Molecular evolution of the metallothionein gene Mtn in the *melanogaster* species group: results from *Drosophila ananassae.* Genetics **138:** 135–143.

Sterner, D. A., T. Carlo and S. M. Berget, 1996 Architectural limits on split genes. Proc. Natl. Acad. Sci. USA **93:** 15081–15085.

Su, A. I., M. P. Cooke, K. A. Ching, Y. Hakak, J. R. Walker *et al.*, 2002 Large-scale analysis of the human and mouse transcriptomes. Proc. Natl. Acad. Sci. USA **99:** 4465–4470.

Sullivan, D. T., 1995 DNA excision repair and transcription: implications for genome evolution. Curr. Opin. Genet. Dev. **5:** 786–791.

Sun, X., and L. E. Maquat, 2000 mRNA surveillance in mammalian cells: the relationship between introns and translation termination. RNA **6:** 1–8.

Svejstrup, J. Q., 2002 Mechanisms of transcription-coupled DNA repair. Nat. Rev. Mol. Cell. Biol. **3:** 21–29.

Tachida, H., 2000 Molecular evolution in a multisite nearly neutral mutation model. J. Mol. Evol. **50:** 69–81.

Upholt, W. B., and L. J. Sandell, 1986 Exon/intron organization of the chicken type II procollagen gene: intron size distribution suggests a minimal intron size. Proc. Natl. Acad. Sci. USA **83:** 2325–2329.

Urrutia, A. O., and L. D. Hurst, 2001 Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. Genetics **159:** 1191–1199.

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

Wall, J. D., 2003 Estimating ancestral population sizes and divergence times. Genetics **163:** 395–404.

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.

Wright, F., 1990 The ‘effective number of codons’ used in a gene. Gene **87:** 23–29.

Yu, J., Z. Yang, M. Kibukawa, M. Paddock, D. A. Passey *et al.*, 2002 Minimal introns are not “junk.” Genome Res. **12:** 1185–1189.

Zhou, Z., M. J. Luo, K. Straesser, J. Katahira, E. Hurt *et al.*, 2000 The protein Aly links pre-messenger-RNA splicing to nuclear export in metazoans. Nature **407:** 401–405.

Zoubak, S., O. Clay and G. Bernardi, 1996 The gene distribution of the human genome. Gene **174:** 95–102.

Communicating editor: S. W. Schaeffer