J Mol Evol (1993) 36:201–213

Journal of Molecular Evolution



# Relationship Between G + C in Silent Sites of Codons and Amino Acid Composition of Human Proteins

David W. Collins and Thomas H. Jukes

Space Sciences Laboratory, University of California, Berkeley, 6701 San Pablo Avenue, Oakland, CA 94608, USA

**Summary.** We have investigated the relationship between the G + C content of silent (synonymous) sites in codons and the amino acid composition of encoded proteins for approximately 1,600 human genes. There are positive correlations between silent site G + C and the proportions of codons for Arg, Pro, Ala, Trp, His, Gln, and Leu and negative ones for Tyr, Phe, Asn, Ile, Lys, Asp, Thr, and Glu. The median proteins coded by groups of genes that differ in silent-site G + C content also differ in amino acid composition, as do some proteins coded by homologous genes. The pattern of compositional change can be largely explained by directional mutation pressure, the genetic code, and differences in the frequencies of accepted amino acid substitutions; the shifts in protein composition are likely to be selectively neutral.

**Key words:** G + C content — Silent sites — GC pressure — Directional mutation pressure — Human genome — Codon usage — Amino acid composition — Neutral theory

The G + C content of DNA varies widely among organisms, especially microorganisms. Among bacterial species the mean G + C content ranges from approximately 25% to 75%, but the amount of intragenomic heterogeneity is small (Rolfe and Meselson 1959; Sueoka et al. 1959). Vertebrate genomes, however, are characterized by a relatively large intragenomic variability in base composition (Sueoka 1959) and this variability is reflected in sequence data (e.g., Ikemura and Aota 1988; Ikemura et al. 1990). It is likely that the human and other vertebrate genomes are mosaics of fairly large regions of similar base composition ("isochores") belonging to a few distinct classes (Bernardi et al. 1985). These may be associated with chromosomal banding. GC-rich genes are most often associated with chromosomal R-(Giemsa pale, quinacrine dull) bands and AT-rich genes with G-(Giemsa dark, quinacrine bright) bands (Ikemura and Wada 1991). Human genes with high G + C in third codon (mostly silent) positions are typically surrounded by GC-rich sequences, whereas those with low third position G + C are embedded in relatively AT-rich sequences (Aota and Ikemura 1986).

Sueoka (1962, 1988, 1992) has proposed that variations in G + C content may be due to directional mutational pressure. According to this theory, the G + C content of DNA is determined by the effective base conversion rates $u$ (G·C to A·T) and $v$ (A·T to G·C). If these rates are unequal, a directional mutation pressure will result. For example, if $v$ is greater than $u$, GC pressure will result; the G + C content at equilibrium is $v/(u + v)$. These pressures could result from biases in the process of DNA replication and repair (reviewed in Filipski 1990). GC pressure was demonstrated experimentally by Cox and Yanofsky (1967), who found that the genomic G + C content of *E. coli* increased

*Offprint requests to:* D.W. Collins

following mutations in the Treffers mutator gene (*mutT*). A mutation in *mutT* results in a high frequency of A·T to C·G transversions. The normal *mutT* protein degrades 8-oxo-7,8-dihydro-2′-dGTP, which may otherwise be incorporated opposite A during DNA replication (Maki and Sekiguchi 1992). The human genome may be subject to a range of mutational biases which could be responsible for its mosaic structure (Filipski 1990; Sueoka 1992).

Changes in genomic G + C content are accompanied by small but significant shifts in the amino acid composition of proteins. Sueoka (1961) examined bacterial species ranging in G + C content from 35% to 72% and showed that mean DNA base composition was correlated with the amino acid composition of total bacterial protein. The protein from more GC-rich organisms was found to have more Pro, Ala, Gly, Arg, and less Lys, Asn, Tyr, Phe, Ile. These effects have also been observed in comparisons of homologous gene sequences from bacteria and mitochondria (Jukes and Bhushan 1986; Muto and Osawa 1987; Filipski 1990) and viruses (Karlin et al. 1990). Typically, large differences in silent-site G + C content are accompanied by smaller differences in the G + C content of replacement sites (mostly codon first and second positions) (e.g., Jukes and Bhushan 1986). For samples of human genes, positive correlations have been found between the G + C content of codon third positions and the G + C content of codon first and second positions (Sueoka 1988, 1992; Aissani et al. 1991; D'Onofrio et al. 1991). In the present work, we attempt to characterize the influence of this effect on the composition of human proteins.

## Methods and Results

*Sample of Human Genes.* We used codon usage tables to calculate the silent-site G + C content and amino acid composition coded by human genes. Wada et al. (1991) have compiled codon usage from the GenBank genetic sequence data bank (Bilofsky and Burks 1988). We obtained a magnetic tape containing codon usage for 2,681 human gene sequences extracted from release 69.0 of GenBank (Wada et al. 1991; T. Ikemura personal communication).

The same gene may be represented by more than one entry and LOCUS name in GenBank. As an example, an identical 963-bp mRNA sequence is present in six different GenBank files, variously defined as endonexin II (GenBank LOCUS name HUMENN), lipocortin-V (HUMLC5), blood coagulation inhibitor (HUMBCI), placental anticoagulant protein PAP (HUMATC), placental anticoagulant PP4 (HUMPAP4), and vascular anticoagulant (HUMVAC) (ACCESSION numbers m18366, d00172, j03745, m21731, m19384, x12454).

The following procedure was used to filter out data from redundant and nearly identical sequence submissions. A database of codon tables for 2,681 GenBank entries was sorted according to the following criteria: (1) total number of codons, (2) the Euclidean distance of the counts of the four nucleotides from the origin, (3) the number of occurences of the nucleotide A, and (4) the number of occurrences of the nucleotide C. Identical and nearly identical sequences necessarily have similar codon usage; the effect of the sorting procedure is to bring these together. The amino acid counts specified by each database entry were then compared to the those of the preceding entry in the sorted list, using a simple distance measure (Collins et al. 1992):

$$\text{Distance} = \left[ (L_1 - L_2)^2 + \sum_{i=1}^{20} (N_{i1} - N_{i2})^2 \right]^{1/2}$$

where $N_{i1}$ and $N_{i2}$ represent the counts of synonymous codons for each of the 20 amino acids and $L_1$ and $L_2$ represent the total number of codons. Additionally, the terms in the sum corresponding to Cys and Trp were multiplied by five and those corresponding to Gly, Pro, Phe, Leu, and Tyr were multiplied by two. If the calculated distance was less than 10, the corresponding record was deleted from the sample. The choice of 10 was arbitrary; 937 records were deleted on this basis. This procedure also removed closely related sequences that do not represent redundant data-bank submissions. For example, HUMCNPG, green cone photoreceptor pigment, was retained but HUMCNPR, the closely related red cone photoreceptor having 96% amino acid sequence identity (Nathans et al. 1986), was excluded. Entries having fewer than 100 codons were also deleted; there were 137 of these. The filtered sample represents 1,607 human gene sequences containing 762,710 codons. These are listed in order of increasing silent-site G + C content and identified by GenBank LOCUS name in the Appendix. A floppy disk containing the codon usage data is available upon request.

### *Silent Sites, Replacement Sites, GC- and AT-Coded Amino Acids*

We define codon silent sites as those nucleotide positions that may potentially undergo substitution without changing the meaning of the codon. In the "universal" genetic code, these are the first-position nucleotides of Arg (CGR, AGR) and Leu (CTR, TTR) codons[1] and the third-position nucleotides of all codons except Met (ATG) and Trp (TGG). The other codon positions are replacement sites. These distinctions help to differentiate between changes in base composition which influence amino acid composition and those which only produce changes in synonymous codon usage. For example, a correlation between the G + C content of codon first and third positions is ambiguous: it may indicate either changes in amino acid coding or changes in synonymous codon usage, since both first and third positions are a mix of silent and replacement sites. There are 61 codons for amino acids. Of these, most (53/61) codon first positions are replacement sites, and most (59/61) third positions are silent. All second positions are replacement

[1] R = A or G, Y = C or T(U), N = any base. Silent sites are underlined.

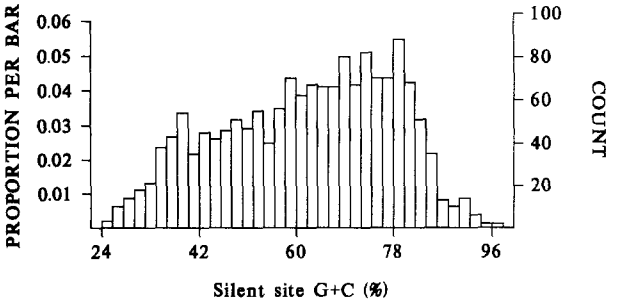



**Fig. 1.** G + C content of codon silent sites, 1,607 human genes. Silent sites are defined in the text. The sample was culled from a larger collection of 2,681 coding sequences from release 69.0 of the GenBank database.

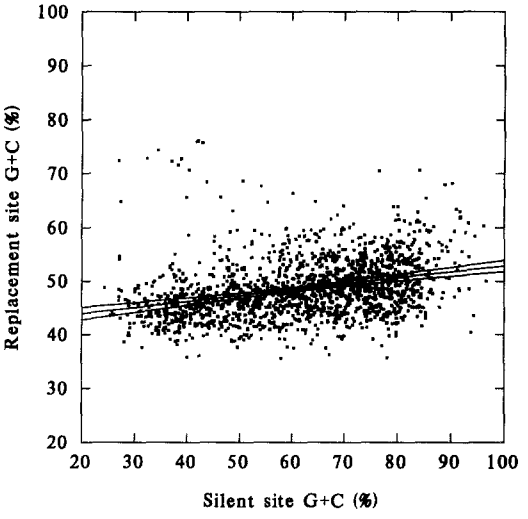



**Fig. 2.** G + C content of replacement vs silent sites in codons, 1,607 human genes. A regression line is plotted with 99% confidence intervals ($r = 0.31$, $P < 0.001$). No outlying points have been excluded.

sites. Because the genetic code provides an alternative codon containing either a G·C or A·T base pair for all amino acids having two or more codons, the G + C content of silent sites is free to vary from 0 to 100%, independently of amino acid composition.

GC-coded and AT-coded amino acids are defined as those with replacement sites occupied exclusively by G/C and A/T, respectively. The GC-coded amino acids are Ala (GCN), Arg (CGN, AGR), Gly (GGN), and Pro (CCN) and the AT-coded amino acids are Asn (AAY), Ile (ATY, ATA), Lys (AAR), Phe (TTY), and Tyr (TAY). Note that Met (ATG) is not AT-coded because the third position is a replacement site occupied by "G." Similarly, Arg (CGN, AGR) is GC-coded because the first positions of AGR codons are silent sites.

### *Correlation Between Silent-Site G + C Content and Amino Acid Composition*

For each gene in the sample we used the corresponding codon table to calculate the G + C content of silent and replacement sites and the amino acid composition of the protein. This was done with the aid of microcomputers and spreadsheet software.

Figure 1 is a histogram representing the silent G + C contents of the sample of genes. The shape of this broad, possibly multimodal distribution is familiar from previous studies of third-position (mostly silent) G + C content in human genes (Karlin et al. 1990; Bernardi and Bernardi 1991). The silent-site G + C content of the sample ranges from 24.3% (GenBank LOCUS HUMACADM, acyl-CoA dehydrogenase) to 96.6% (HUMMIFA, migration inhibitory factor MIF). This wide variation in synonymous codon usage is typical of vertebrate genomes (Ikemura and Aota 1988; Ikemura et al. 1990; Wada et al. 1991).

The G + C content of silent vs. replacement sites for the sample of genes is plotted in Fig. 2. Although the points are widely scattered, there is a significant positive association ($r = 0.31$, $P < 0.001$). Figures 3 and 4 show the relationships between silent G + C and the fractions of GC-coded (Ala + Arg + Gly + Pro) and AT-coded amino acids (Asn + Ile + Lys + Phe + Tyr) for the sample of genes. The correlations are significantly positive and negative, respectively.

Table 1 summarizes the correlations between silent G + C and the fraction of codons for each of the 20 amino acids. The levels of the individual AT-coded amino acids are all negatively correlated with silent-site G + C. The levels of the individual GC-coded amino acids are positively and significantly correlated with silent G + C with the exception of Gly. The sample contains several genes (e.g., collagens, proline-rich proteins) that are extreme outliers with respect to content of glycine and proline codons. These are mostly low in silent G + C and exert a large influence on the correlations. When the 20 most proline-rich proteins are removed from the sample, the correlation between Gly and silent G + C is significantly positive ($P < 0.001$).

Among amino acids with "mixed" coding (i.e., replacement sites occupied by both G·C and A·T), Trp, His, Gln, and Leu are positively correlated with silent G + C; Asp, Thr, and Glu are negatively correlated. The correlations of Met, Cys, Val, and Ser are not significantly different from zero.

### *Average Protein Composition*

The preceding results suggest that the average composition of human proteins may vary, depending on the pattern of synonymous codon usage of the

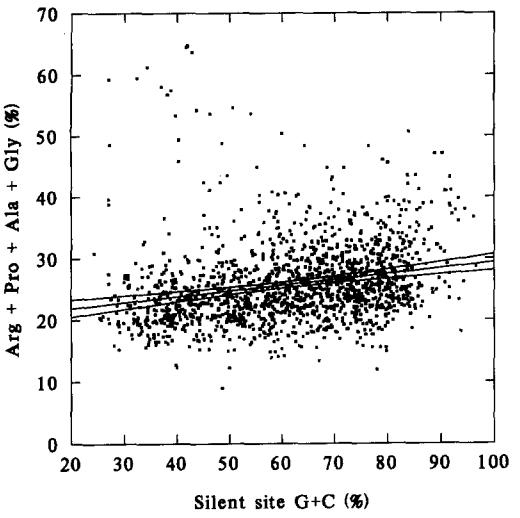



**Fig. 3.** Percentage of codons for GC-coded amino acids (Arg, Pro, Ala, Gly) vs silent-site G + C content, 1,607 human genes. The regression line is shown with a 99% confidence interval ($R = 0.22$, $P < 0.001$).

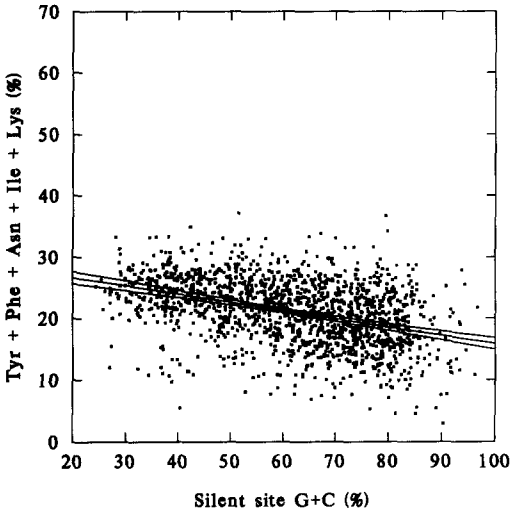



**Fig. 4.** Percentage of codons for AT-coded amino acids (Tyr, Phe, Asn, Ile, Lys) vs silent-site G + C content, 1,607 human genes ($R = 0.41$, $P < 0.001$).

genes. In order to quantify these differences, we divided the sample of genes into three groups (low, medium, high) on the basis of silent site G + C content. The low group consisted of 473 genes having from 24% to 52% silent G + C; the medium group consisted of 578 genes having from 52% to 71% silent G + C; and the high group consisted of 556 genes having 71–97% silent G + C. The divisions were chosen to place, as nearly as possible, equal numbers of codons into three groups; the decision to use three groups was arbitrary. We then compared the average proteins coded by these groups.

The composite codon usage of the three groups is presented in Table 1. Increases in the proportion of the GC-coded amino acids Ala and Arg (but not Gly and Pro) and decreases in AT-coded amino acids are present across the three groups (from low to high). The largest changes are for Lys and Arg; the ratio of Lys to Arg codons is 1.47:1 for the low group and decreases to 0.81:1 for the high group. There is a stepwise increase in the ratio of GC- to AT-coded amino acids across the three groups. For amino acids with mixed coding, the proportions are relatively constant except for Leu, which shows a stepwise increase (Table 1).

The distributions of amino acid content of samples of proteins are often skewed to the right; consequently the median rather than mean composition may provide a better representation. The median protein compositions coded by the three groups of genes are shown in Figs. 5 and 6. Regular increases in the fractions of GC-coded amino acids (Arg, Pro, Ala, Gly) and decreases in AT-coded amino acids (Tyr, Phe, Asn, Ile, Lys) are apparent (Fig. 5). Among GC-coded amino acids, the increases in Arg and Ala are greater than for Pro and Gly. For AT-coded amino acids the decreases in Lys, Ile, and Asn are larger than for Tyr and Phe. For amino acids with mixed coding, there are small but stepwise decreases in Thr and Glu, a small increase in Gln, and a larger increase in Leu.

### *Homologous Genes—Steroid Hormone Receptors, Ras Proto-Oncogenes*

There are relatively few opportunities to compare homologous human genes differing greatly in base conposition, but some examples are presented in Table 2. The amino acid composition and silent-site G + C content coded by three *ras* genes are compared, as are those of three members of the steroid hormone receptor superfamily. Silent G + C is again more subject to increase than is G + C in replacement sites. Even in these small samples, the trends in amino acid composition resemble those in Table 1. The ratio of GC- to AT-coded amino acids increases with silent site G + C.

In the case of the three *ras* proteins, which are very similar in sequence, most of the change in amino acid composition is concentrated in a 25-residue C-terminal variable region which is "not required for any of the known biochemical functions of the protein" (de Vos et al. 1988). In K-*ras* (30% silent G + C) this region contains 13 AT-coded amino acids (12 Lys + 1 Ile) and one GC-coded amino acid (Gly). In H-*ras* (82% silent G + C) there are only 4 AT-coded amino acids (3 Lys, 1 Asn) and 6 GC-coded amino acids (3 Pro, 2 Gly, 1 Arg). Thus

**Table 1.** Average composition of protein coded by 1,607 human genes[a]

| Group | No. of codons | | | 1,607 human genes protein composition | | |
|---|---|---|---|---|---|---|
| | Low | Medium | High | | | |
| Silent G + C | 24–52% | 52–71% | 71–97% | Avg. | SD | (*r*) vs |
| No. genes | 473 | 578 | 556 | (%) | (%) | silent G + C |
| *(a) AT-coded amino acids* | | | | | | |
| Tyr (TAY) | 7,999 | 7,619 | 7,123 | 3.0 | 1.3 | −0.11*** |
| Phe (TTY) | 9,784 | 9,625 | 9,356 | 3.8 | 1.6 | −0.06* |
| Asn (AAY) | 11,272 | 9,904 | 8,501 | 3.8 | 1.6 | −0.33*** |
| Ile (ATY/A) | 13,121 | 11,300 | 10,178 | 4.5 | 1.9 | −0.29*** |
| Lys (AAR) | 16,986 | 14,436 | 12,092 | 6.0 | 2.9 | −0.29*** |
| *(b) Amino acids with mixed coding* | | | | | | |
| Trp (TGG) | 3,000 | 3,447 | 3,674 | 1.3 | 1.0 | 0.11*** |
| Met (ATG) | 5,161 | 4,929 | 5,045 | 2.0 | 1.1 | −.03 |
| Cys (TGY) | 5,978 | 5,757 | 6,212 | 2.4 | 2.1 | .04 |
| His (CAY) | 5,838 | 6,235 | 6,181 | 2.4 | 1.2 | .06* |
| Gln (CAR) | 10,879 | 11,252 | 11,798 | 4.4 | 1.9 | .06* |
| Asp (GAY) | 12,977 | 12,862 | 12,718 | 5.0 | 1.9 | −0.09*** |
| Thr (ACN) | 14,496 | 14,242 | 13,490 | 5.5 | 1.8 | −0.09*** |
| Val (GTN) | 15,868 | 15,760 | 16,307 | 6.3 | 1.9 | .02 |
| Glu (GAR) | 17,506 | 17,555 | 17,186 | 6.8 | 3.1 | −0.07** |
| Ser (TCN/AGY) | 18,672 | 19,152 | 18,219 | 7.3 | 2.5 | .00 |
| Leu (CTN/TTR) | 21,714 | 23,962 | 26,202 | 9.7 | 2.8 | 0.29*** |
| *(c) GC-coded amino acids* | | | | | | |
| Arg (CGN/AGR) | 11,534 | 13,640 | 14,963 | 5.5 | 2.4 | 0.23*** |
| Pro (CCN) | 15,156 | 15,150 | 15,513 | 5.8 | 3.1 | 0.06* |
| Ala (GCN) | 16,288 | 17,651 | 19,805 | 7.3 | 2.7 | 0.25*** |
| Gly (GGN) | 18,883 | 18,855 | 18,488 | 7.3 | 3.4 | 0.02 |
| Total | 253,112 | 253,333 | 253,051 | 100.0 | | |
| *(c)/(a)* | 1.05 | 1.23 | 1.46 | 1.23 | | |

[a] The first three columns list the total number of codons for each amino acid coded by the "low," "medium," and "high" groups of genes. The low, medium, and high groups consist of genes with silent G + C contents ranging from 24 to 52% (473 genes), 52 to 71% (578 genes), and 71 to 97% (556 genes), respectively. The sequence files are identified by GenBank LOCUS name in the Appendix. Initiation codons are not included in "Met" and stop codons have been omitted. The fourth column lists the mean protein composition coded by the sample of 1,607 genes. The last column lists the correlation coefficient (*r*) for silent-site G + C content (%) and amino acid abundance (%) in the sample of genes (*$P < 0.05$, **$P < .01$, ***$P < 0.001$)

the nonfunctional region is more subject to change, as would be expected from the neutral theory.

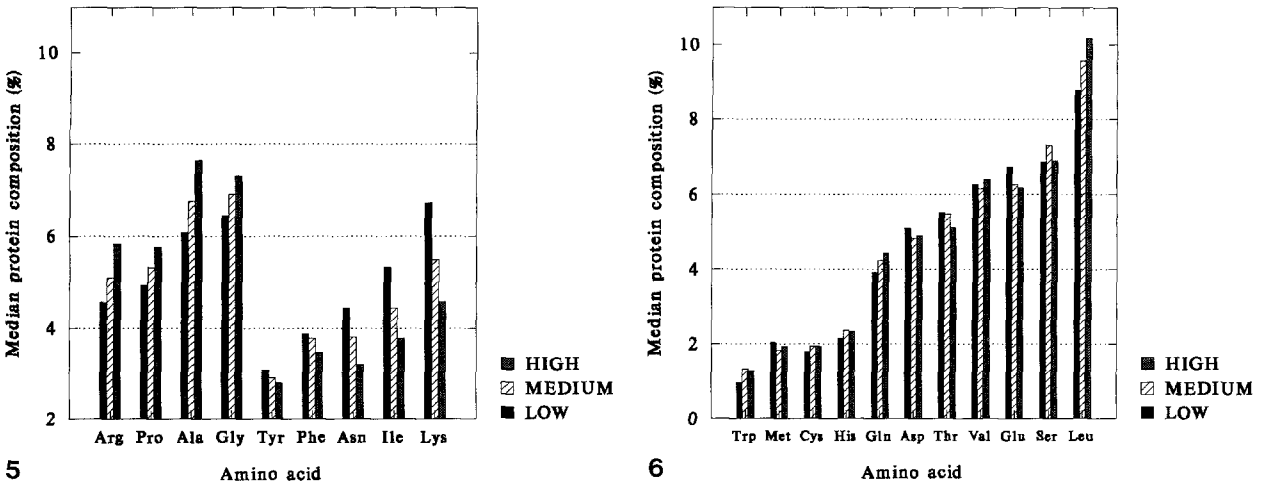
The effect of GC pressure on amino acid composition should be related to the fraction of sites in the protein which is free to accept amino acid substitutions. If an amino acid sequence is perfectly conserved, only the pattern of synonymous codon usage may "respond" to GC pressure. For example, the human genome contains two genes for calmodulin having divergent codon usage (26% and 68% G + C in silent sites), but in spite of this disparity, these sequences code identical proteins (Fischer et al. 1988).

## Discussion

### *Amino Acid Mutability*

In general, we have found that GC-coded amino acids increase with silent G + C, and AT-coded amino acids decrease. However, the "response" of the individual amino acids differs substantially in magnitude (Table 1, Figs. 5, 6). Among GC-coded amino acids, the increases in Pro and Gly with silent G + C (Table 1, Fig. 5) are smaller than those of Ala and Arg. For AT-coded amino acids, the decreases in Tyr and Phe are small relative to Asn, Ile, Lys (Table 1, Fig. 6). Similar results were obtained by Karlin et al. (1990), who compared the amino acid usage of homologous proteins in related viruses differing greatly in genomic G + C content. Several amino acids with "mixed" coding are also correlated with silent-site G + C content.

The pattern of compositional change may be explained by considering differences in rates of amino acid substitution in addition to coding. For example, amino acid exchanges involving Pro (56) and Gly (49) residues are fixed less frequently during evolution than those involving Arg (65) and Ala




**Figs. 5, 6.** Median protein composition coded by genes with low, medium, and high silent G + C content. Content of GC-coded (Arg, Pro, Ala, Gly) and AT-coded (Tyr, Phe, Asn, Ile, Lys) amino acids (Fig. 5). Content of amino acids with "mixed" coding (Trp, Met, Cys, His, Gln, Asp, Thr, Val, Glu, Ser, Leu) (Fig. 6).

**Table 2.** Amino acid composition and silent G + C content of two groups of dispersed homologous human genes[a]

| | *ras* oncogenes | | | Steroid hormone receptors | | |
|---|---|---|---|---|---|---|
| Gene | K-*ras* | N-*ras* | H-*ras* | HAPRA | VDR | EAR3 |
| Silent G + C (%) | 29.6 | 44.1 | 81.5 | 50.5 | 76.1 | 85.4 |
| Replace G + C (%) | 42.2 | 43.9 | 45.9 | 45.2 | 45.4 | 53.5 |
| Chromosome | 12p12.1 | 1p22 | 11p15.5 | 3 | ? | 5 |
| Total codons | 188 | 188 | 188 | 447 | 426 | 424 |
| | (No. codons) | | | (No. codons) | | |
| *AT-coded* | | | | | | |
| Tyr (TAY) | 9 | 9 | 9 | 10 | 8 | 12 |
| Phe (TTY) | 6 | 6 | 5 | 16 | 18 | 14 |
| Asn (AAY) | 4 | 6 | 5 | 13 | 13 | 16 |
| Ile (ATY/A) | 15 | 11 | 11 | 25 | 24 | 26 |
| Lys (AAR) | 16 | 12 | 11 | 32 | 24 | 14 |
| *(a) Total* | 50 | 44 | 41 | 96 | 87 | 82 |
| *GC-coded* | | | | | | |
| Arg (CGN,AGR) | 12 | 10 | 12 | 23 | 29 | 26 |
| Pro (CCN) | 5 | 5 | 6 | 28 | 22 | 31 |
| Ala (GCN) | 9 | 10 | 11 | 21 | 20 | 36 |
| Gly (GGN) | 11 | 14 | 13 | 20 | 20 | 37 |
| *(b) Total* | 37 | 39 | 42 | 92 | 91 | 130 |
| *(b)/(a)* | 0.74 | 0.89 | 1.02 | 0.96 | 1.05 | 1.59 |

[a] K-*ras* = c-Ki-*ras* oncogene (GenBank accession number K03209, Hirai et al. 1985), N-*ras* = N-*ras* proto-oncogene (X00642, Brown et al. 1984), H-*ras* = c-Ha-*ras*1 proto-oncogene (J00277, Sekiya et al. 1984), HAPRA = hepatocellular carcinoma hormone receptor (Y00291, de The et al. 1987), VDR = vitamin D receptor (J03258, Baker et al. 1988), EAR3 = v-erbA-related hormone receptor (X12795, Miyajima et al. 1988). Sequences were obtained from the GenBank database (Bilofsky and Burks 1988)

(100). (The numbers represent the relative mutabilities of the amino acids given in Dayhoff 1978.) Accordingly, the fractions of Pro and Gly may be expected to evolve at a lower rate in response to mutational biases relative to other GC-coded amino acids. Similarly, mutations involving Phe (41) and Tyr (41) are accepted less often than are those involving the other AT-coded amino acids: Asn (134), Ile (96), and Lys (56) (Dayhoff 1978). The large increase and decrease in Arg and Lys, respectively (Table 1, Fig. 5), may be linked; Lys and Arg are readily exchanged through single point mutations from Arg (AGR) to Lys (AAR).

### *The Case of Leucine*

A similar argument may explain the increase in Leu (CTN, TTR) (Table 1, Fig. 6) with silent G + C

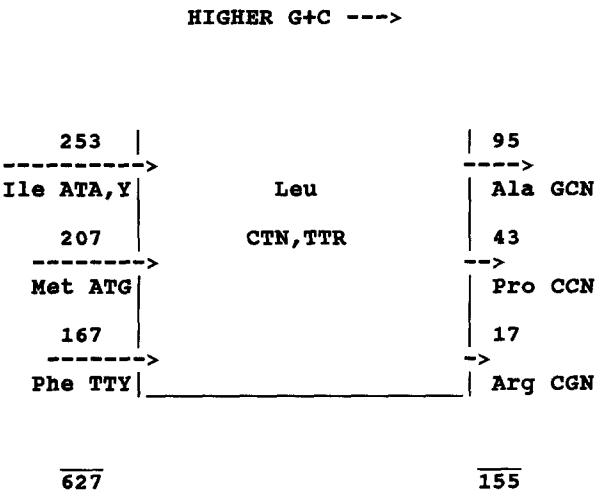



Fig. 7. Schematic representation of a possible mechanism for increase in the fraction of leucine codons under GC pressure. A mutational bias favoring G + C (GC pressure) would preferentially suggest mutations in the directions of the *arrows*. New leucine codons would be generated from codons with lower G + C (Ile, Met, Phe), but leucine codons would also be lost through mutation to GC-coded amino acids (Ala, Pro, Arg). However, the latter mutations are fixed less often than the former, causing a net increase in the number of Leu codons. The numbers above each *arrow* indicate the relative frequency of accepted point mutations between leucine and the indicated amino acid (from Dayhoff 1978).

despite relatively AT-rich mixed coding. If the underlying pattern of nucleotide substitution is biased in the direction of G + C (GC pressure), one might expect new leucine codons to be generated through point mutations of Ile (ATA, ATY), Phe (TTY), and Met (ATG), but simultaneously eliminated via mutation of other Leu codons to Pro (CCN), Arg (CGN), and Ala (GCN, two-base change). These pathways are not selectively equivalent, however, and therefore may not balance. Exchanges between Leu and Ile/Phe/Met (lower GC) are fixed about four times as frequently as those involving Ala/Pro/Arg (higher GC) (Dayhoff 1978). Consequently, the fraction of leucine codons may be predicted to increase under GC pressure, rather than remain constant. This scheme is diagrammed in Fig. 7.

### *Compositional Change and Neutral Evolution*

There are two broad viewpoints on heterogeneity in the G + C content of DNA. These differ fundamentally with regard to the role of natural selection. The first viewpoint holds that compositional variation is maintained by positive Darwinian selection for G + C content per se or for particular amino acids in proteins (Bernardi and Bernardi 1986a,b). In contrast, the second viewpoint attributes compositional variation to directional mutation biases, rather than to selective advantage (Sueoka 1988); the compositional shifts are viewed as selectively neutral phenomena. The neutral theory of molecular evolution postulates that "nucleotide substitutions inherently take place in DNA as a result of point mutations followed by random genetic drift" (Jukes and Kimura 1984). The substitution rate is highest in the absence of selective constraints—for example, in pseudogenes. Lower rates indicate the presence of constraints imposed by negative selection, which rejects deleterious mutations.

Bernardi and Bernardi (1986a,b) have proposed that GC-rich regions of chromosomes in warm-blooded vertebrates have evolved to accommodate higher body temperatures because G·C pairs have a higher melting point than A·T pairs. They offer selectionist explanations for the observed regional variations, attributing them to "compositional constraints." Higher GC levels in mRNAs are proposed to "increase their base-pairing and stability." Additionally, increases in G + C content of coding sequences "lead to thermodynamically more stable proteins" (Bernardi and Bernardi, 1986b). Specifically, alanine and arginine "are most frequently acquired in thermophiles: and produce an increase in heat stability, while serine and lysine, which diminish stability, are correspondingly lost."

Sueoka (1992) has criticized the proposal that large-scale variations in base composition are maintained by natural selection because point mutations away from the hypothetical optimal G + C content would have to be considered deleterious on the basis of an infinitesimal effect on the heat stability of DNA. Furthermore, some thermophilic organisms are high in A + T (Filipski 1990).

Alanine and arginine may accumulate under GC pressure via neutral mutations because they are GC-coded, rather than via positive selection operating at the protein level. Filipski compared sequences from (1) *Thermus thermophilus* (high GC, thermophilic) and *Saccharomyces cerevisiae* (high AT, mesophilic) and (2) from *Streptomyces limosus, Streptomyces hygroscopicus* (high GC, mesophilic), and *Dictyoglomus thermophilium* (high AT, thermophilic). Alanine content was found to correlate with genomic G + C rather than with the thermophilicity of the organisms. We have found that serine, listed with lysine as decreasing thermal stability, is not negatively correlated with G + C in silent sites (Table 1, Fig. 6). The decrease in lysine that accompanies increasing GC may proceed by selectively neutral mutations from Lys (AAR) to Arg (AGR) suggested by GC pressure, rather than from selective advantage.

According to the neutral theory of molecular evolution, nucleotide changes take place and are fixed by genetic drift. If mutations in the direction AT → GC predominate, then the GC content of

DNA will increase, although this trend may be opposed by negative selection to some extent. The neutral theory predicts that this increase will take place most readily in silent positions of codons and this is the case. There is a much wider variation in G + C content of codon silent sites than replacement sites (Fig. 2). Replacement substitutions are fixed less frequently than silent substitutions (Li et al. 1985), and there are varying degrees of constraint acting against amino acid exchanges. However, many replacements are neutral, when they do not interfere appreciably with protein function. This occurs in many proteins—for example, in globins (Perutz 1983). In the case of three members of the *ras* family of proto-oncogenes, the variation in amino acid composition is concentrated in a nonessential region.

## Conclusion

We have determined the actual changes in coded amino acid content that accompany increases in the silent G + C content of human genes and have observed correlations between silent-site G + C and the levels of the majority of the 20 amino acids. Amino acids with codons high in G + C were higher in genes with high content of G + C in silent sites, and the opposite was true for amino acids with codons high in A + T. The largest changes are for amino acids that may undergo conservative replacements, e.g., Arg/Lys and Ile/Leu. Less mutable amino acids such as Phe show smaller increases. Codons with mixed G + C content in replacement sites were in some cases positively and in other cases negatively correlated with G + C content of silent sites. Positive correlation is conspicuous in the case of leucine, and we suggest that this results from the lower selective cost of mutations to leucine from isoleucine, methionine, and phenylalanine relative to mutations of leucine to amino acids with codons high in G + C.

The strength of these associations, both positive and negative, can be accounted for by the genetic code, amino acid mutability, and the neutral theory.

*Acknowledgments*. We thank Dr. Toshimichi Ikemura for kindly providing us with a magnetic tape containing codon usage from GenBank and Dr. Noboru Sueoka and an anonymous referee for comments. This work was supported by NIH grant R01 HG00312-03.

## References

Aïssani B et al (1991) The compositional properties of human genes. J Mol Evol 32:493–503

Aota S, Ikemura T (1986) Diversity of G + C content at the third position of codons in vertebrate genes and its cause. Nucleic Acids Res 14:6345–6355

Baker AR et al (1988) Cloning and expression of full-length cDNA encoding human vitamin D receptor. Proc Natl Acad Sci USA 85:3294–3298

Bernardi G et al (1985) The mosaic genome of vertebrates. Science 228:953–958

Bernardi G, Bernardi G (1986a) Compositional constraints and genome evolution. J Mol Evol 24:1–11

Bernardi G, Bernardi G (1986b) The human genome and its evolutionary context. Cold Spring Harbor Symp Quant Biol 51:479–487

Bernardi G, Bernardi G (1991) Compositional properties of nuclear genes from cold-blooded vertebrates. J Mol Evol 33:57–67

Bilofsky HS, Burks C (1988) The GenBank genetic sequence data bank. Nucleic Acids Res 16:1861–1864

Brown R et al (1984) Mechanism of activation of an N-ras gene in the human fibrosarcoma cell line HT1080. EMBO J 3:1321–1326

Collins DW et al (1992) Numerical classification of coding sequences. Nucleic Acids Res 20(6):1405–1410

Cox EC, Yanofsky C (1967) Altered base ratios in the DNA of an *Escherichia coli* mutator strain. Proc Natl Acad Sci USA 58:1895–1902

Dayhoff MO (1978) Atlas of protein sequence and structure, vol 5, suppl 3, National Biomedical Research Foundation, Silver Spring, MD

de The H et al (1987) A novel steroid thyroid hormone receptor-related gene inappropriately expressed in human hepatocellular carcinoma. Nature 330:667–670

de Vos AM et al (1988) Three-dimensional structure of an oncogene protein: catalytic domain of human c-H-ras p21. Science 239:888–893

D'Onofrio G et al (1991) Correlation between the compositional properties of human genes, codon usage, and amino acid composition of proteins. J Mol Evol 32:504–510

Filipski J (1990) Evolution of DNA sequence. Contribution of mutation bias and selection to the origin of chromosomal compartments. Adv Mutagenesis Res 2:1–54

Fischer R et al (1988) Multiple divergent mRNAs code for a single human calmodulin. J Biol Chem 263:17055–17062

Hirai H et al (1985) Activation of the c-K-ras oncogene in a human pancreas carcinoma. Biochem Biophy Res Commun 127:168–174

Ikemura T, Aota S (1988) Global variation in G + C content along vertebrate genome DNA. Possible correlation with chromosome band structures. J Mol Biol 203:1–13

Ikemura T et al (1990) Giant G + C% mosaic structures of the human genome found by arrangement of GenBank human DNA sequences according to genetic positions. Genomics 8:207–216

Ikemura T, Wada K (1991) Evident diversity of codon usage patterns of human genes with respect to chromosome banding patterns and chromosome numbers; relation between nucleotide sequence data and cytogenetic data. Nucleic Acids Res 19:4333–4339

Jukes TH, Kimura M (1984) Evolutionary constraints and the neutral theory. J Mol Evol 21:90–92

Jukes TH, Bhushan V (1986) Silent nucleotide substitutions and G + C content of some mitochondrial and bacterial genes. J Mol Evol 24:39–44

Karlin S et al (1990) Contrasts in codon usage of latent versus productive genes of Epstein-Barr virus: data and hypotheses. J Virol 64(9):4264–4273

Li W-H et al (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol 2(2):150–174

Maki H, Sekiguchi M (1992) MutT protein specifically hydrolyses a potent mutagenic substrate for DNA synthesis. Nature 355:273–275

Miyajima N et al (1988) Identification of two novel members of erbA superfamily by molecular cloning: the gene products of the two are highly related to each other. Nucleic Acids Res 16:11057–11074

Muto A, Osawa S (1987) The guanine and cytosine content of genomic DNA and bacterial evolution. Proc Natl Acad Sci USA 84:166–169

Nathans J et al (1986) Molecular genetics of human color vision: the genes encoding blue, green and red pigments. Nature 232:193–202

Perutz M (1983) Species adaptation in a protein molecule. Mol Biol Evol 1:1–28

Rolfe R, Meselson M (1959) The relative homogeneity of microbial DNA. Proc Natl Acad Sci USA 45:1039–1043

Sekiya T et al (1984) Molecular cloning and the total nucleotide sequence of the human c-Ha-ras-1 gene activated in a melanoma from a Japanese patient. Proc Natl Acad Sci USA 81: 5384–5388

Sueoka N et al (1959) Heterogeneity in deoxyribonucleic acids. II. Dependency of the density of deoxyribonucleic acids on guanine-cytosine content. Nature 183:1429–1431

Sueoka N (1961) Correlation between base composition of deoxyribonucleic acid and amino acid composition of protein. Proc Natl Acad Sci USA 47:1141–1149

Sueoka N (1962) On the genetic basis of variation and heterogeneity of DNA base composition. Proc Natl Acad Sci USA 48:582–592

Sueoka N (1988) Directional mutation pressure and neutral molecular evolution. Proc Natl Acad Sci 85:2633–2657

Sueoka N (1992) Directional mutation pressure, selective constraints and genetic equilibria. J Mol Evol 34:95–114

Wada K et al (1991) Codon usage tabulated from the GenBank genetic sequence data. Nucleic Acids Res 19 (Suppl):1981–1986

Received February 27, 1992/Revised June 3, 1992

**Appendix 1.** Coding sequence data

24 ACADM
25 BFXIII
26 CAM
26 HELB#3
27 COL3AI
27 EVI2B3P
27 CYES1
27 DLDH
27 RNPB1A
27 RNPA2A
27 RASRPB
27 CA1XIA
28 FAPAPC
28 HPLK
28 IREBPA
29 ANTARNP
29 6ORO
29 HAAB
29 MCH
29 SBLA
29 GLYCA
29 PMPCA
29 P68A
29 IFNRA
29 KRASM
30 RASK25*
30 MALENAD
30 PMMPP1
30 MUT13*
30 KAD
31 AMY12*
31 PROS30
31 MCP
31 CHEB
31 AMYA110*
31 SEM
31 IL6GP
31 CD46Q
31 PTYPH
31 CHEBG4*
31 EVI22
31 IL76*
32 CDC2
32 GAPB
32 NPM
32 TPPIIA
32 MITF1
32 MACT
32 SRTR2A
32 C1A2
32 TR29
33 RBS
33 CNRA1
33 PROSA
33 UCHL3A
33 SNONR
33 SNOAR
33 BCAT
33 33DPTP
33 TR211
33 MRL3R
33 PROP2AA
33 RAB2
33 CNRAB
34 MSCA
34 SNRNPD
34 LFA3R
34 SPROTR
34 TOPII
34 CIX
34 PRPC
34 FIX
34 LDHX
34 DBLPRO
34 CALCBP
34 LACI09*
34 PYHBASB
35 FUMA
35 TKFER
35 ANTLF3
35 EIF2A
35 ASP
35 LCA
35 CBF
35 IFNRG
35 DGIGLY
35 CCC5
35 ETFA
35 RP19
35 GLYCQ
35 ANTCD36
36 NKG2B
36 ZFY
36 NKG2A
36 PLA2A
36 TFIIS
36 UG2BA
36 GAPA
36 C9M
36 LCAR
36 DCKATPB
36 PHSR2
36 HSF2
36 FISP
36 VIPHM6*
36 GST
36 RELH2
36 FNRB
36 ARAA5*
36 NKG2C
36 SCF
37 HELB#1
37 CS1PA
37 CREB
37 SCFA2
37 HMGCOA
37 KGF
37 HGFHL
37 PRPD
37 PHSR1
37 CREBA#2
37 BFDNA
37 FBRG#1
37 DCREB
37 FBRG#2
37 PTH2
37 TPX1A
37 TRA1
37 PRPC4B
37 HPRTA1A
38 POLYAB
38 ROSA
38 APOH
38 GAGMR
38 SNEXIN
38 PCCAR
38 IL1RA
38 AFP
38 VLA2A
38 CBP
38 P70S6KB
38 CNR
38 ALBGC#2
38 ALBAF1
38 CERP
38 NUCLEO
38 COR2M
38 BRAF
38 POLB
38 G25KA
38 FABP
38 UDPGTA
38 PRPE
38 NKG2D
38 HFSP
38 PC1Q1
38 PP2AB
38 DUG
38 DBLTP
38 HNRNPA
38 PKCBA
38 MRNAEN
39 CALLA
39 ELF2
39 PLAST
39 LSZA
39 NF1AA
39 ARF2A
39 C7A
39 PRPF
39 HELA#2
39 PAM12
39 UDPGT
39 FVIIIC
39 EBUR13*
39 FASANT
39 VTNR
39 CAM1V
39 P70S6KA
39 PPKKA
39 FBRA
39 ALPHLA
39 ADH5CH1
39 ADH4C1
39 B1LYM
39 FOL5*
39 EB2CR2
39 CFTRM
40 IDE#2
40 LAMP1B
40 OATC
40 ZFX
40 DAFB
40 CLGNA
40 MRA
40 SNRAA
40 OMGPA
40 DAFA#2
40 ANTN
40 TS11
40 ASNS
40 GHR
40 IL2S3*
40 EGFRER
40 U1C
40 GCRB
40 PRPH2
40 PAP4
40 GCRA
40 QBPCA
41 RALA
41 CSAE
41 ANX3
41 NMTDC
41 POLDNAA
41 ADH5C3
41 TGFBB
41 COOTAA
41 ELI#1
41 TFRR
41 LDHBR
41 VCAM1
41 MHCC6A
41 ALDC13*
42 INTA6R
42 AMD
42 RAL
42 BCKDHA
42 LKHA4
42 ELASF
42 SIALO
42 STEAA
42 EL20#1*
42 EL20#2*
42 KUP
42 TSH2*
42 MDR1
42 PDHBET
42 GABAAA1
42 ECPG
42 SGII
42 LGTPA
42 IFNG
42 HRGA
42 STROM2
42 CYCLINA
43 KIN10#1*
43 ADH6
43 ELIA
43 LIPCR
43 HMG1
43 RIREM1
43 STROMR
43 HF10
43 II56KD
43 EL20#3*
43 CD2A
43 EDNMRN
43 IL5R
43 METPOA
43 HPBA
43 OTC10*
43 FVA
43 PUMP1
43 AACTA1
43 ELF1B
43 CAPPTA
43 EMP41
43 CN2
43 ATPSY
44 HEXB13*
44 ANTCD2
44 DMDR
44 CAIR
44 BPPTK
44 COL1A42*
44 ALBP
44 RPS12
44 CARMC

44 TCRGA
44 XYPFLA
44 RPS6
44 TRGC64
44 TGLIP
44 7AH
44 RCYP3
44 CMYBLA
44 RASAB
44 MDR3
44 UMPS
44 LDHA7*
44 TCRGAD
44 THD
44 ZFX1
44 GC
44 HPROT
44 KUPMR
44 CYPHLP
44 RPS6A
44 ELI#2
44 ARG8*
45 HGFR
45 CYPCN
45 SRICPA
45 TCRGC
45 ODCA2
45 CCG1
45 GLUCG2
45 CR1
45 LAMBP
45 SOD1
45 IL7AA
45 ACKII
45 H33G4*
45 TCRDR
45 GAP43A
45 MCR
45 HSC70
46 ADH1CB
46 KITCR
46 RODSA
46 BPGM3*
46 PGK2G
46 OSTRO
46 LDL100
46 HMG14
46 MAC2
46 BCGF
46 COL8A1
46 TCRGR
46 RPS24A#2
46 FXI
47 TCOBI
47 ASF#1
47 TDTA
47 IGIIDO
47 DONT11*
47 CYPAX
47 INTB6A
47 MONAP
47 NFKB34
47 MGPA
47 EF1AR
47 EP2AA
47 ISG2*
47 LYAM9*
47 AGALAR
47 PLNHR
47 GABAAS
47 TYRM
48 MAP4
48 HAP2A
48 IGB7
48 CDC25HS
48 RPL17
48 PRG
48 IREBP
48 RNPC2A
48 HIS2AZ
48 LHHCGR
48 TYRA
48 ATPFIB
48 C1RS
48 PKCAMD
48 PAI2A
48 C1S
48 ATPSYB
48 ALIPOA
48 CTAP3
48 7B2
48 HNRNA
48 CATHL
48 CD44E
48 U2AR
48 SF2P33
48 CYPB
48 TGFB1B
48 TOP19*
49 PLG24*
49 CCBL
49 TEF1
49 CDR34
49 PPEPB
49 TFIIDA
49 LBP
49 RETAA
49 TCR3G6*
49 TFIID
49 TBG
49 BRANK2
49 CYL
49 MDMCSF
49 PRLR
49 ARNTA
49 GMP140
49 CYP19
49 LYN
49 REPA
49 C8BS
49 CCDR
49 ELAM9*
49 FCERI
49 A2TPI
49 LAMB
50 PORAC
50 YB1A
50 LAMB33*
50 CALBR
50 GPPSBC
50 FMO1
50 RASAD
50 GRP78
50 BADPTA
50 AMPD1
50 PNU4*
50 PSG3A
50 ZNF7
50 PZPHEP
50 PAPB
50 MAOAA
50 B2M2*
50 ET
50 P18
50 PSGB1A
50 SPDMAT
50 SPP
50 C8AS
50 RASAA
50 ADH229*
50 TUMP
50 ARGCAA
50 THYMA
51 HAPRA
51 HK2A
51 AIXIII
51 H1T
51 5NUASE
51 RPS25
51 TCRT3E
51 PROAF
51 ARP450
51 IL1P
51 PPARP1
51 THYMAA
51 PDHA12*
51 CD1C6*
51 CATR
51 CSPG1A
51 P65
51 RPS4Y
51 HK1A
51 RASAC
51 VINC
51 CYCAA
52 SYTA
52 PIP
52 RAFR
52 CD1C
52 PSBGA14*
52 CD1B6*
52 FGF53*
52 EAP
52 FCRHA
52 RPL31
52 MAOB
52 PHH
52 RHOB
52 CAP
52 PLASTA
52 FGF5A#2
52 C4BAA
52 IL2RA
52 ITI2
52 OCT1A
52 TCAR
52 TFPB
52 SCAR
52 GSTC
52 MIC2A
52 GLYCA4*
53 IGFBPS
53 CAMPR2
53 RPA70KD
53 XCGD
53 SAPR
53 CLMF35
53 CFCGRI6*
53 MEA
53 ACYLHYD
53 PDGFARA
53 GPPSBA
53 IL2RB
53 RPS4X
53 SPTA01
53 PSBGAA
53 CEASG5
53 BMP3A
53 SLK
53 LPLR
53 MLCAB
53 NAKATP2*
53 CA2
53 CNTFG
53 GPBPS
53 ATCT2*
53 PSG6A
54 PSG12
54 LAMBB
54 CDW44A
54 A2M
54 GPPSBD
54 A1CKII
54 PSGA
54 TGFB2A
54 RAG1
54 CD38
54 P47
54 NLF1
54 TCAYE
54 HPA2B
54 UPCP12*
54 CHROMB
54 M6PR
54 ETSR
54 HP1G5*
54 BDNFC
54 ATPAR
54 C4A2
54 IFNAH
54 NKSFP35
54 AGPRO
54 FDX
54 ACHRM2
54 SYB1A5*
55 FSHRE
55 ALAS1R
55 CD59A
55 BDNF
55 PPROA
55 IFNAIP
55 MLC3F
55 TROPCR
55 SPTCS
55 SAA1A
55 F13A15*
55 SLIPG
55 NMOR
55 BAT2B4*
55 TM2CEA
55 FERC
55 CHYMASE
55 KER673*
56 ACKI10
56 APOD
56 CYCR
56 HPRG5*
56 PSPB
56 IFNA01
56 LSP1Q1
56 HISAB
56 IFNAN
56 GTLPA
56 THRAA
56 CX43
56 IDSX
56 NCAX
56 CD1A
56 TRHA
56 MYOHP
56 CD1A6*
56 CKMT
56 GDHL
56 TF
56 GFB#3
56 GBP1
57 CYB5
57 TROPSR
57 GABAAB1
57 RPS14
57 AFPA4
57 RPZH21
57 FSH3*
57 IFNF
57 CERBA
57 AMYB19*
57 PAP4A
57 HUGBR2
57 BMP2A
57 BGALRP
57 TK14
57 DOCKP
57 RASFAB
57 A20
57 HSP90B
57 MNSOD
57 APA4R
57 BGPAB#1
57 ALFUC
58 TRO
58 BGPI
58 HER3A
58 MHDRARN
58 SAA
58 PALC
58 PALFAP
58 PROP2AB#3
58 HCPB
58 BLAST1
58 EMBPA
58 HLRPR
58 PTPAAA
58 IFNAB
58 BCTHA
58 FCGRB
58 GIF
58 PPPB1A
58 PTKJAK1
58 IFNB1
58 AGG
58 QRE
58 CRPGA
58 AHSG
58 BGPAA
58 RIBIIR
58 CGM7
58 MHHSPHO
58 FIBAA
58 LIC
58 PFKM23
58 HXMA7*
58 HUGBR1
58 RNP7008*
58 5AR
58 PHIDYIN
58 CRYBA6*
58 BGAL
58 UROD
58 IFNAII
58 INTAZ
58 YAVREB
59 MHCAG1
59 HLDOBR
59 LYB2
59 ARSBX
59 IFNATD
59 P40CYT
59 MHDRA
59 CMYCQ
59 BAT3A
59 PCC
59 CRIPTO
59 CLMF40
59 ASAM
59 TSHX
59 CARAA
59 ASPAT
59 ORF
59 USFMR
59 CGM1A
59 CD284#2*
59 LCTHA
59 ANTNC
59 SSARO
59 GLYPL
59 IGFIIR
59 FCGRA
59 ALDOB9*
59 LACTA
59 IFNB2R
59 GLI
59 PKCL
59 ANT2X
59 ACROS
59 MBPA
59 GLUSYN
59 HXBP1
59 STSB
59 PTHL3*
60 IGGK
60 BFR
60 OSF1
60 KSAMAA
60 PPARP2
60 PLP6*
60 MHRD5*
60 PP15
60 ETS1A

60 MRCOX4*
60 CGM1B
60 ERCC3A
60 CEAF
60 CA1V
60 1433
60 GFIAB4*
60 FCREC
60 RNPAB
60 BGPAB#2
60 BHA14*
60 GP34M
60 TS1
60 FCRII
61 MHDQADR
61 THRR
61 VPF
61 MBPZ
61 IGGFCRA
61 GRFCIG
61 HLADQA
61 MBP17K
61 PGP95
61 HBP
61 G6PA
61 DBI
61 FLAP5*
61 HSDI
61 GALT
61 IIP
61 UNG
61 LCT17#2*
61 PEMP
61 IL1RAA
61 ATP
61 THYP
61 IL1C
61 CSFM
61 IMP
61 CAIII7*
61 KRT10A
61 CPT
61 TCAXB
61 FGF4H
61 MUCAB
61 IRP
61 PLPDM
61 AT3X6*
61 COXCA
61 ALDCG
61 SMCK
62 NOXF
62 MXB
62 PGRR
62 HRSR
62 P53C
62 T519#2
62 MLN
62 GRP5E
62 AR
62 TNSCN
62 EGFAA
62 HBB222*
62 TM30R
62 AFH
62 WRSAA
62 ALR
62 PLAX
62 CD43#2
62 DHPR
62 APOC2G
62 OAS08*
62 HDC
62 CDW40
62 ETMAGA
62 GAST
62 PKCB2A
63 ERG2
63 CGPRA
63 FERG2*
63 PPARP0
63 OAS07*
63 CRFBP
63 FMLP
63 MBPC
63 PRPS2
63 PBGDR2
63 TUBAK
63 GABAR
63 NT3A
63 GYPCAA
63 KBLOOD
63 HODB3*
63 RPS17
63 HELAGT#2
63 ILRA
63 PBGDR
63 ANPCR
63 OGCB
63 PROZII
63 MBPB
63 3B5H5E
63 HBEGF
63 THROMR
63 BN51
63 FAB
63 EGFRS
63 FIGRE
63 HEM1
63 FIGRD
64 RISDAD
64 MXA
64 TGLH
64 GHVA#2
64 FOSB#1
64 MAS
64 FOLLI2*
64 HBGF3*
64 SHBGA
64 ERG11
64 ASPX
64 PROZI
64 HELAGT#1
64 ARX
64 MHCSE
64 IGLAM2
64 HBB#4
64 CRYABA
64 GSTMUA
64 MAX#2
64 SGLT1
64 RPOLAA
64 GCB1
64 NK4
64 P42LA
64 ARB
64 D1DO
64 PINCAM
64 ANDREC
64 CSK2B
64 MYCTR
64 IRF1
64 UKPM
64 ET3
64 NGFBA2
64 MLC2
65 CTHG
65 MGDMT
65 ALASR
65 EGFRN
65 NKG5
65 RHPAA
65 ALAS2R
65 ENOA
65 GF1A
65 ETS2A
65 RETSA#2
65 HCF2
65 RPL32
65 RIBIR
65 GHG
65 GLI3A
65 GCB
65 CTSE
65 P42SA
65 PTHL
65 EGR2A
65 GALTA
65 GFIAB5*
65 AP2AA
65 NID
65 TPI
65 PROT2
65 SECPA
65 CSDF1
65 MAX#1
65 ANFA
65 PTHL4*
66 PRL7*
66 CYPIIE
66 A1AR2*
66 PRPOA
66 PRP
66 ACTCA4*
66 PC2A
66 DNFA
66 GFIBP
66 HBB#2
66 HBB#3
66 FAH
66 ENOG
66 XRCC1
66 CA6
66 FAPS
66 PTHL2*
66 MYLA1
66 HEXKIN
66 FOSB#2
66 APOAII
66 DCDK
66 NFM
66 GPIBAA
66 CYP2BB
66 GBR#2
66 QM
66 MHBA123
66 KER654*
66 MH3C2
66 PKM2L
66 MYOD1R
66 TAUI
67 CD53
67 MYF5
67 BMP2B
67 RCC1B
67 UBIQAA
67 GPIIIAA
67 MYHC
67 ZP3
67 HPS12
67 A1ATB
67 TAUA
67 TRL
67 CAM3X1*
67 ASM
67 A1ACM
67 C1INHB
67 CRPR
67 MYCL2A
67 TCII
67 ANT1
67 AUTAN64
67 IGFIB
67 FABPLA
67 3OCTR
67 ACTSG7*
67 I6REC
67 MYLCA#1
67 A1GP06*
67 CALCR4*
67 HEMOB
67 PAR
67 FLA1A
67 PPTRH03*
67 ALD
67 MHSXA
67 RFPA
67 CBPE
68 HOX329
68 LAMP1A
68 REN10*
68 CYP2BA
68 ALAD
68 INHBA
68 RENT3*
68 IGFBP1A
68 CALCR5*
68 GFII
68 HCR
68 UBI13
68 BPIAA
68 TRBP
68 A1ATZ
68 ACHRA
68 IL4R
68 LMGP
68 ET2A
68 PP11A
68 PLC
68 HIS3PRM
68 FGL2
68 PCAR
68 IL4
68 KALX
68 CYP178*
68 SOMI
68 ATCT4
68 REGB
69 ANTCD
69 TKR
69 CYPBA
69 GIPX6*
69 MLC3NM
69 FNRA
69 TMPKMR
69 NM23H2S
69 RNP70K
69 HLADRBA
69 CRCMUT
69 UBA52P
69 IBSUB
69 INTB5A
69 TRKR
69 CSF1M3*
69 GLUTRA
69 TBBM40
69 CD19W07*
69 IGLV
69 PSAG
69 HMGIA
69 PTCAA
69 INTB7A
69 MHDRBA
69 ADRBR
69 B61
69 FBPB
69 TCBYZ
69 SCYLP
69 BETGLA
69 CTSB
69 TCBXA
69 ROSSAA
69 A1MICR
69 RNAGLA
69 IL10
69 HLASBA#1
69 UBA52C
69 RNP7011#2
69 IGLBV
69 SRAA
70 MYOL1
70 CALRTR
70 KTRAN
70 P50NFKB
70 ICOA
70 SYN
70 RNP7011#1
70 MHDRBU
70 LAPA
70 MAC1A
70 ERMCF
70 VTNSP
70 PHK
70 IGFBP4
70 SPRO
70 ABL
70 INHA
70 VIL2
70 KERC15
70 ANTCD9
70 CNPB5*
70 FBP
70 MYC3L
70 C4AA2*
70 ENKB4*
70 HOXB
70 CBG
70 TGFB3A
70 FOS
70 COX4AA
70 TPO15*
70 G3PDC
70 A1GLY2
71 TPOB
71 LOX15A
71 GFRIL
71 CYP45C
71 SYB2A5*
71 PLAT
71 CRYGX2*
71 TGFA
71 GLCB
71 INSR
71 PDGA7*
71 FGF2H
71 P45SC9*
71 HIS4
71 PEC12L
71 FGF3H
71 ANTLA
71 CRYGA2*
71 ANK
71 AGP1A
71 ERP
71 TIMPR
71 PROPERD
71 CP45IV
71 VIM
71 INSRA
71 ACHRB
71 MLC1SA
71 TNFR
71 TCBYY
71 KER2A
71 MUC18A
71 AP2
71 CRE
71 LT
71 NAGA
71 TNFRII
72 INTLEU8
72 TPAR
72 INCP3*
72 TCF1A
72 CAD
72 NFH4*
72 ALDHIR
72 VACB
72 PDGFA6*
72 HA44G
72 SYN1E13#1
72 ASGPR2
72 PPE
72 HISAC
72 CLG4Q13*
72 GOAQ10#2*
72 VWFR1
72 FGFAA

72 POVRA
72 ELA308*
72 HBGFA
72 ALRM
72 FCER
72 TCF1B
72 INT07*
72 EPP10*
72 SAPD1
72 CMOS
72 JGEBFR
72 KEREP9*
72 ICAMA1M
72 ADAM2
72 ASL
73 VILLR
73 GNBPB3
73 SA#1
73 CYP27
73 GLBA
73 PDEAA
73 THB
73 CD8B
73 PF4A
73 HAAG
73 MYCRT
73 C1R
73 ERCC1
73 CD14R
73 TCF1C
73 TFLS
73 ALDH13*
73 IRGT
73 H4
73 PGEX11*
73 PROF
73 GADD45
73 ACP5
73 TIR
73 HOXA
73 AK1
73 CGBS08
73 TBP1
73 KERUHS
73 HISAA
73 HER2A
73 IL2RBC
73 RETPO
73 PKLR
73 AFL2*
73 CYC1
73 HLAATE
73 PKA
73 SB2BR
73 LAP11*
73 ANGG5*
73 LIGAA
74 APOC3B
74 MHDRDQ
74 POVRB
74 LSP1A
74 GPIIBA
74 PLCA
74 GNAS6#2*
74 MHDRO3*
74 CINHP
74 FIBUB
74 HOM4
74 ERBT1
74 BMYB
74 GCSFRD
74 ASA
74 STROL3
74 OCT2A#1
74 TGFBC
74 GMCSFRB
74 GNAS6#3*
74 GSA2R
74 FMSCPO
74 CHRAA
74 ELA3A
74 ATP1A2
74 RALBPC
74 C3
74 TROPIA
74 ARAF1R
74 RAP2
74 CNTFR
74 YUBG1
74 TCPTK
74 XEH
74 GSA1R
74 PLA2A2*
74 NAGB
74 CRBP2*
74 LDLR18*
74 THBP
74 PAIR
74 CS3
74 GRFP5#2*
74 PLA
74 ALDA
75 CNPG6*
75 GRFP5#1*
75 ABPA
75 HCKA
75 FIBUA
75 ARF5A
75 PDGFRA
75 ICAM4*
75 IGLR141
75 KER18
75 RARG
75 MHDPL
75 CETP7*
75 MYCM
75 GLUT5
75 ADXR#2
75 927A
75 ADREDB#1*
75 CSFGM
75 PLAPL
75 ADREDB#2*
75 HBLOD
75 HLA1EA
75 BCTHB
75 GCSFR3
75 U1RNPA
75 PPR
75 GROG5
75 NLK
75 UMOD
75 PDGFA7*
75 RHOC9
75 CCK3*
75 FIBUC
75 GH#1
75 MRP14
75 IGHAF
76 PROTP
76 BNPA
76 TPMYOC
76 GATA
76 GCSFR4
76 COLIP
76 DNASEI
76 NFLG
76 THRA1A
76 LCKAA
76 GHCSA#2
76 HLDQWB
76 LOX5A
76 GAPJR
76 PGSR
76 CYPB3*
76 TNTS
76 PLPSPC#1
76 GLYSA
76 G19P1A
76 VDR
76 GHV
76 SERDHY
76 HPSNA
76 HISH4
76 H2B2H2#2
76 LCTHB
76 PLPSPC#2
76 TRK2H
76 ELAP2B
76 PANMU
77 PYGM20*
77 FGR
77 NORTR
77 KERIA8*
77 SPARC
77 GH#2
77 SPTB#2
77 APOJ
77 PEPC9*
77 PEPCA9*
77 RAP1GAP
77 TNTSA
77 GROB5
77 MH
77 SPARC10*
77 ETR103
77 OCS3*
77 ASFA
77 KERI8*
77 SPTB#1
77 KDEL
77 NEKAR5*
77 CLI
77 ANTP
77 ERYTH
77 APHOL
77 MHTRP
77 PIM1
77 BMP1A
77 HLADZA
77 LDLRRL
77 ARF6A
77 TROPA
77 IMPH
77 CSP40
77 IGHAE2*
77 C45AII
77 MHB27D
77 ELS2
78 RBP
78 C8G
78 CYPIIF
78 GFIBPA
78 CYP2DG
78 LEC14K
78 PGP
78 CPIIA3A
78 MHCCB2
78 COF
78 PSPBA
78 LOX5
78 RDS
78 CENPB
78 ASLA
78 LAMAR
78 SCL
78 CYIIA4A
78 IL2AB
78 TNFAA
78 SRF
78 SISPDG
78 PEPD
78 4COLA
78 MAL
78 NKB
78 PP14A
78 MHCW1C
78 TK
78 CNPDEA
78 RYR
78 PKBR
78 MGI3*
78 FKBP13
78 EDG
78 ACHE
79 CYP345
79 IRBPG
79 IRBPG4*
79 MH6
79 MHA3
79 MHCGE2*
79 LCAT
79 PRF1A
79 CYP4A7*
79 SCAD
79 GCSFG
79 PRPHOS1
79 P2A
79 GP
79 FGFR4
79 INV2
79 TRNB
79 PEP9*
79 NMYCA#1
79 CEAPX
79 TACEB
79 FX8*
79 ARP1
79 HISH2B
79 HLA11E
79 MHCA1A
79 TRKPOA
79 CMP8*
79 ENDOA2#1
79 GTUB#1
79 CERA
79 CP21OHC
79 LYL1B
79 OXYGR
79 ALPI1
79 CHRM
79 FFI2A
79 ACTAR
79 ALPP
79 MHDNDRW
79 CRYGQ2*
79 HIS10G
80 P3A
80 ENDOA2#5
80 FFI2B
80 AHCY2
80 FVII#2
80 GFAP
80 CRF
80 KEREP
80 LIPBSA
80 AMIPEP
80 TNCS
80 ZNFBPAA
80 RPS11
80 MLC2A
80 IFP
80 THY1A
80 A2PIG6*
80 MYP
80 GLUTRN
80 RNAPII
80 CFXII3*
80 KER19
80 ALPHA
80 ATPGG
80 GIAA
80 GIP2A8*
80 SCL7*
80 PDECGF
80 AICEB
80 HKATPC
80 CKMA
80 TACEA
80 GGTX
80 MDP4
80 MHGM
80 ALPPB
80 LAMC
80 BCL3AA
80 IFNIN3
80 ISK
80 MHCACA#1
80 E12A
80 CD37
80 CD7
80 TFAA
80 ACHRG8*
80 C5AAR
81 GCSF
81 FVII#1
81 EDHB17
81 MGPHB
81 DRD2A
81 ATPK14*
81 FESFPS
81 GCSFR
81 TIMP2
81 P58GTA
81 HNF1
81 P971
81 KER7E9*
81 ATPA23*
81 PNMT
81 PCD
81 UBILP
81 BMYH7
81 TGASE
81 CYPDB1
81 CYP2D6
81 TUBBM
81 TROPI
81 GAA
81 IL11
81 ASGPR1
81 ARF1BA
81 ACHRM4
81 GGT
81 INSPR
81 NADPHO
82 H2B2H2#1
82 BCR
82 CANPR
82 ADRA2RA
82 INTBE4
82 DBHRB
82 GATA3M
82 LHB
82 DES
82 GATA3R
82 PEROXP
82 TNC2*
82 ECK
82 KER8
82 PRC7*
82 BCL2B
82 18D
82 BCL2A
82 GAA01*
82 PRCM
82 CTF1
82 MZF1
82 TCXAAA
83 LAP
83 ALIFA
83 INT2
83 APRTA
83 INTB4R
83 DKERB
83 RAR
83 RASR2*
83 SHIIIC
83 GSTPI
83 CYS3A3*
83 CGB
83 RETREC
83 UDPG
83 18U
83 UDPCNA#2
83 GKNASE
83 A2MGRAP
83 NGFR
83 JUNCAA

83 CYTOK
83 PFKLA
83 TGFB
83 FKMKA
83 TRY
83 HD5DR
84 ETR101
84 PROSYN#1
84 PCA
84 COMTA
84 MYELA
84 OPS
84 SRC11*
84 HOX14
84 ELFTL
84 SPYRAT
84 PKCGA
84 APOA4B
84 G6PDA
84 TRPY1B
84 IGFBP5A
84 GAAA
84 PLAKO
84 THX
84 LORAA
84 INT1G
84 APOAIB
84 EF2AB
84 ARB3A
84 THM
84 G6PDG13*
84 RACPC
84 R2IMP
84 PTBMR
84 THR
84 RACB
85 PGAMM2*
85 TAPA1
85 JUNA
85 GA16
85 PGPIX#1
85 BGN3*
85 MAG
85 GXA
85 HPGI
85 MIS
85 EAR3
85 HIP3K
85 HPBS
86 ELFT
86 HTH1R
86 ACTGA
86 MYCPOB
86 LEUELA
86 CNP
86 CTRP
86 TGFBD
86 GDF1#1
86 BIGFII
86 CATD5*
86 IHRP
87 SPERSYN
87 MYOD
87 PCHSUCA
87 INHBB2*
87 FGFR3
88 GLYPIC
88 IGFBP5
88 BCL1
88 NCBLCA
88 POMC9*
88 SAACT
88 TBB5
89 GDF1#2
89 HSP27
89 ISGA2*
89 SKIR
90 HBA4#2
90 APOE4
90 GPIB
91 ADRA2C
91 GA733A
91 HSP70D
91 MHHSP
91 CPGISL
91 G0S2A
92 CKB
92 GLTH1
92 RHOB6
92 JUNDR
92 OTNPI
92 OTCB
92 D4DOP
92 ADRA2R
93 EAR2
93 HST
93 VPNP
94 HBA1
94 GTPBRPA
94 IDB
95 ADRB1
96 SODEC
97 MIFA

[a] 1,607 coding sequences are identified by GenBank LOCUS name and are listed in order of increasing silent site G + C ("%G + C"). The first three characters of each LOCUS name (HUM) have been omitted.