MOLECULAR PHYLOGENETICS AND EVOLUTION
Vol. 5, No. 1, February, pp. 2–12, 1996
ARTICLE NO. 0002

Human Coding and Noncoding DNA: Compositional Correlations

OLIVER CLAY, SIMONE CACCIÒ, SERGUEI ZOUBAK,[1] DOMINIQUE MOUCHIROUD,* AND GIORGIO BERNARDI

*Laboratoire de Génétique Moléculaire, Institut Jacques Monod, 75005 Paris, France; and *Laboratoire de Biométrie, Génétique et Biologie des Populations, U.R.A. 243, Université Claude Bernard Lyon I, 69600 Villeurbanne, France*

Received July 27, 1995; revised August 9, 1995

As the correlations between GC levels in third codon positions (GC_3) and intergenic sequence GC levels can be used to assess the distribution of genes in the human genome, they were studied in detail. Previous work from our laboratory has demonstrated the existence of linear correlations between GC levels of exons, introns, third codon positions, 5′ flanking regions of genes, and long genomic DNA sequences (≥10 kb) or DNA molecules (50–100 kb) in which the genes are embedded. The present study confirms and extends the previous results using a larger set of data. Furthermore, an analysis of 4270 human genomic DNA and cDNA sequences has allowed us to confirm a correlation of GC_3 against GC_{1+2}. Recent additions to the sequence database have also allowed separate analyses of the 5′ flanking regions of CpG island and non-CpG island genes as well as analyses of 3′ flanking regions, which suggest that the GC levels of 3′ flanking regions are closer to those of intergenic DNA than are those of other regions of genes.

INTRODUCTION

The human genome is a mosaic of isochores, long (>300 kb) DNA segments which are compositionally homogeneous and belong to a small number of families characterized by different GC levels covering a 30–60% range (GC is the molar fraction of guanine + cytosine). Isochore families L1 + L2 and H1 + H2 + H3 represent the GC-poor 2/3 and the GC-rich 1/3 of the genome, respectively. The isochore organization of the human genome is typical of warm-blooded vertebrates, whereas cold-blooded vertebrates are characterized by a much narrower GC range which never reaches the high levels attained by warm-blooded vertebrates. The different isochore patterns just mentioned are paralleled by different compositional patterns of coding sequences. The different compositional patterns of isochores and coding sequences of warm- and cold-blooded vertebrates represent different genome phenotypes (Bernardi *et al.*, 1985; Bernardi, 1995).

In situ suppression hybridization of human DNA fractions characterized by increasing GC levels on human metaphase chromosomes has clarified the correlations between DNA base composition and chromosomal bands (Saccone *et al.*, 1992, 1993). (i) T(Telomeric)-bands are formed by the GC-richest isochores of the H3 family, and by part of the GC-rich isochores of the H1 and H2 families (with a predominance of the latter); (ii) R′-bands [namely R(Reverse)-bands exclusive of T-bands] are formed, to almost equal extents, by GC-rich isochores of the H1 families (with a minor contribution of the H2 and H3 families) and by GC-poor isochores of the L1 + L2 families; (iii) G(Giemsa)-bands essentially consist of GC-poor isochores from the L1 + L2 families, with a minor contribution of H1 isochores.

The distribution of genes in the human genome is strikingly nonuniform (Bernardi *et al.*, 1985; Bernardi, 1995). Indeed, a low, constant gene concentration is present in the GC-poor isochore families L1 and L2; gene concentration then increases in increasingly GC-richer isochores (isochore families H1 and H2) to attain the highest value (20× higher than in L1 + L2) in the GC-richest isochore family H3, that only represents about 4% of the genome. Because isochore distribution in chromosomes is known (see above), these results also provide information on the distribution of genes in chromosomes. The highest concentration of genes in telomeres is of special interest in view of the association of telomeres with the nuclear membrane and their attachment to the nuclear matrix.

Several findings point to the fact that isochores have not only a structural but also a functional relevance. For example, the GC-richest isochores not only are characterized by the highest gene concentration, but also by the highest concentration of CpG islands (GC-rich, nonmethylated regions typically located upstream of genes), by an open chromatin structure, by the highest transcription and recombination levels, and by extreme codon usage and amino acid encoding.

The reasons for the correlation between gene concentration of isochores and their GC levels have been recently understood (Bernardi, 1995). Indeed, the gene

[1] Permanent address: Institute of Molecular Biology and Genetics, Ukrainian Academy of Sciences, 150, Zabolotnogo Str., 252627, Kiev, Ukraine.

1055-7903/96 $18.00

concentration pattern of the human genome is basically present in all vertebrates. A strong GC increase took place, however, in gene-rich regions at the transition between cold-blooded vertebrates on the one hand and mammals and birds (two separate events in time) on the other.

As the correlations between GC levels in third codon positions (GC_3) and intergenic sequence GC levels can be used to assess the distribution of genes in the human genome (Mouchiroud *et al.,* 1991; Bernardi, 1995), they were studied in detail here.

Previous investigations on human genes (Bernardi *et al.,* 1985; Aota and Ikemura, 1986; Aïssani *et al.,* 1991; Mouchiroud *et al.,* 1991) have shown that consistent compositional correlations exist between (a) the GC levels of third codon positions (GC_3) in the coding regions of genes and the GC levels of DNA molecules or of long (≥10 kb) genomic DNA sequences from the database that contain them; (b) the GC levels of exons and those of the DNA molecules or long sequences that contain them; (c) the GC levels of introns and exons; (d) the GC levels of introns and 5′ flanking sequences of genes; (e) the GC levels of third codon positions (GC_3) and those of first and second codon positions (GC_{1+2}). In addition, it was found that clustered genes had coding sequences that were very similar in composition. Several of the above correlations were obtained from 21 sequenced human genes that had been experimentally localized in DNA fragments (50–100 kb) forming compositional fractions, whose GC level had been determined, and from 32 genes for which long (≥10 kb) nucleotide sequences were available. Compositional correlations between coding and noncoding regions have also been found for the genes of other vertebrates and of plants (Bernardi *et al.,* 1985; Salinas *et al.,* 1988; Montero *et al.,* 1990).

The present report confirms and extends the previous results, using the most recent release of the nucleotide database and incorporating recent experimental data. All of the characteristic GC levels used are defined with respect to the coding regions (CDS), so that bulk retrieval of these values from large numbers of sequences, and screening for sequences in which the GC levels can be reliably assessed, can be almost entirely automated. This latter fact will become important for future refinements and updating of the slopes, intercepts, and correlation coefficients of the compositional correlations, and thus for estimates of the distribution of genes in the genomes of human and other species (see below), as the nucleotide database continues to grow. Since 1982 the total number of bases in GenBank (all species) has doubled approximately every 21 months (GenBank Release Notes, Rel. 88.0; Benson *et al.,* 1994), and if this trend continues, automatability will soon be a prerequisite for comprehensive analyses of the type discussed here. The characteristic GC levels of genes as described here can be retrieved from the database using the general-purpose sequence analysis program systems QUERY (Gouy *et al.,* 1985) and BISANCE (Dessen *et al.,* 1990).

We have already reported that the use of the orthogonal regression, i.e., regression along the principal eigenvector of the variance–covariance matrix, gives a better description of the scatter diagrams than linear regression, since the orthogonal regression line is obtained by minimizing the sum of squares of the orthogonal distances rather than of the distances parallel to the ordinate (see D'Onofrio *et al.,* 1991; Mouchiroud *et al.,* 1991). As the orthogonal regression line typically follows the highest concentration of points, it could also be interpreted as the crest of an adaptive ridge (cf. Wright, 1932) in a landscape defined by the different characteristic GC levels and having the orthogonal regression line as the principal axis of its inertia ellipsoid. Although orthogonal regression is the method of choice here, and the one used to estimate the distribution of genes (see below), we have also included linear regressions for all plots, to allow comparison with previous results which were expressed in terms of linear regression parameters (Aïssani *et al.,* 1991). The mean values of the GC levels in each plot are thus clearly visible as the coordinates of the center of mass of the points, which is at the intersection of the orthogonal and linear regression lines. When comparing different plots, it is often helpful to compare the positions of the center of mass, as these are more easily interpreted than the intercept of a regression line. Also, when regression lines are steep (higher than 2 or 3), it is more meaningful to compare angles than slopes, since the latter have the intrinsic artifact of being very sensitive to small changes in the data in this case, and thus of exaggerating differences between regression lines.

MATERIALS AND METHODS

Analysis of Database Sequences

EMBL Release 42.0 (March 1995) in a GCG environment (Devereux *et al.,* 1984) was systematically searched for all sequences with self-contained CDS fields (features) that had a total length ≥10 kb or that had a long (≥1.5 kb) flank either 5′ of the ATG or 3′ of the stop codon, as well as sufficiently long coding regions (≥300 bp) or total intron length (≥1 kb). Sequences with only "exon" and "intron" specifications but no indication of coding region were not analyzed. ATG and stop codons implied by the features specifications were checked explicitly, and "complete" coding sequences were also checked to ensure that they contained an integral number of codons. Sequences without ATG in the expected position were not included in the analysis to avoid possible contamination of the GC level data. Furthermore, sequences containing neither the ATG nor the stop codon of any gene were omit-

ted from the analysis. Pseudogenes were similarly excluded. To allow automation, only introns interrupting the coding region of a gene were processed, and not those between noncoding exons. In sequences containing two or more duplicated, clustered genes (indicated with an asterisk in the Appendix), a representative (well-flanked) gene was chosen to represent the sequence, and intergenic regions were divided equally between the gene and the one adjacent to it. Finally, three special genes were omitted from the regressions concerning small data sets ($N < 50$): two collagen genes (HSCOL2A1Z) and HSCOL7A1X), as in Aïssani *et al.* (1991), for the reason that the $(Gly\text{-}X\text{-}Y)_n$ repeats of collagen genes, where X is often proline, create an unusual bias toward high GC_{1+2}, since glycine and proline are encoded by GGN and CCN, respectively; and the gene for extracellular superoxide dismutase or SOD3 (HS10116), which was the only example of a gene with a very high GC_3 level (95.6%) due to a CpG island in its 720-bp coding region, yet which is embedded in a 10.1-kb sequence with a GC level of 49.8%, and thus not in an isochore of the GC-richest family as expected. With the 235 nonredundant sequences remaining after this selection, pairwise regressions were performed between characteristic GC levels in which sequences were included in the regressions involving CDS or GC_3 if the total coding region was ≥300 bp, in those involving intron GC levels if the total intron length was ≥1 kb, in those involving (far) 3′ or 5′ flanks if the (far) flank length was ≥1 kb, and in those involving the GC level of the entire sequence if the total length was ≥10 kb. The EMBL mnemonics, lengths, CpG island status, and accession numbers of these sequences are given in the Appendix. Wherever possible, sequences were classified according to whether they contain CpG islands ("CpG island genes") or not ("non-CpG island genes"), according to the latest release of the CpGIsle database (March 1995; see Larsen *et al.*, 1992). Sequences that were not represented in the CpGIsle database were not used in the analyses that addressed this difference. All analyses involving the 5′ flanks of genes were performed separately for the group of sequences with no CpG islands and for those with at least one CpG island. As the results obtained were similar whether the analyses included all CpG island genes or only those genes with CpG islands at the 5′ or "start" regions, the results for this more restricted group are not shown here. Only one sequence (aldolase C gene, HSALDC) contained a 3′ CpG island, and was not an outlier in any of the plots. For the plot of GC_3 against GC_{1+2}, coding cDNA and genomic DNA sequences from the nucleotide database were used after removing incomplete sequences, obvious redundancies, and sequences for the MHC, T cell receptor, and immunoglobulin genes, which are overrepresented in the database and often document different alleles or variants of the same genes. From a 2-dimensional histogram of the remaining 4270 sequences, a contour plot (Fig. 5B) was calculated using the PVWAVE package (Precision Visuals) with the smoothing parameter set to 4.

Localization of Specific Genes

Human DNA was prepared and fractionated as previously described (Zerial *et al.*, 1986), and the DNA fractions were analyzed by analytical ultracentrifugation. Probes were labeled and hybridized to Southern blots by standard techniques (Sambrook *et al.*, 1989). Entries 1–21 are from previous work (Aïssani *et al.*, 1991). Entries 22–26 are from hybridizations performed recently in our laboratory. Entries 27–34 are from an analysis of the chromosomal band Xq28, using a method for determining the GC content of YACs that has been recently developed in our laboratory (DeSario *et al.*, 1995a,b). Only genes that had been previously assigned to the YACs were used in the analysis. Although technically the GC levels of some YACs as well as those of DNA molecules from compositional fractions contributed to the values in Table 1, for simplicity both will be referred to as "fragments" in the text and Figures. In the regressions involving experimentally localized genes, the data for the collagen cluster (entry 13; see above) and for the interferon-α receptor gene (entry 6) were omitted, as in the previous study.

RESULTS

Compositional Correlations of Coding Sequences with Isochores and Long Sequences from the Database

Figure 1A shows a plot of GC levels of CDS from a set of human genes or gene clusters (see Table 1) against the GC levels of DNA molecules from compositional fractions in which the genes or gene clusters were localized by hybridization of appropriate probes. This set includes the 21 genes localized in a previous study (Aïssani *et al.*, 1991). The linear regression yields a slope of 1.10 and a correlation coefficient of 0.82, which are nearly identical to the values observed previously for a smaller data set (1.19 and 0.80; Aïssani *et al.*, 1991). The orthogonal regression line yields a slope of 1.42.

Figure 1B shows a plot of GC levels of coding regions against the GC levels of the corresponding long (≥10 kb) sequences in which they were present. The slope of the linear regression through the points was 0.81 and the correlation coefficient 0.84, values which again are almost identical to those reported in our previous analysis, i.e., 0.81 and 0.85. The orthogonal regression gave a slope of 0.96.

Compositional Correlations of Third Codon Positions with Isochores, Long Sequences from the Database, and Flanking Sequences of Genes

Figure 2A shows a plot of the GC levels of third codon positions (GC_3) from the set of human genes of Fig. 1A against the GC levels of DNA fragments from compositional fractions, again with similar results to those

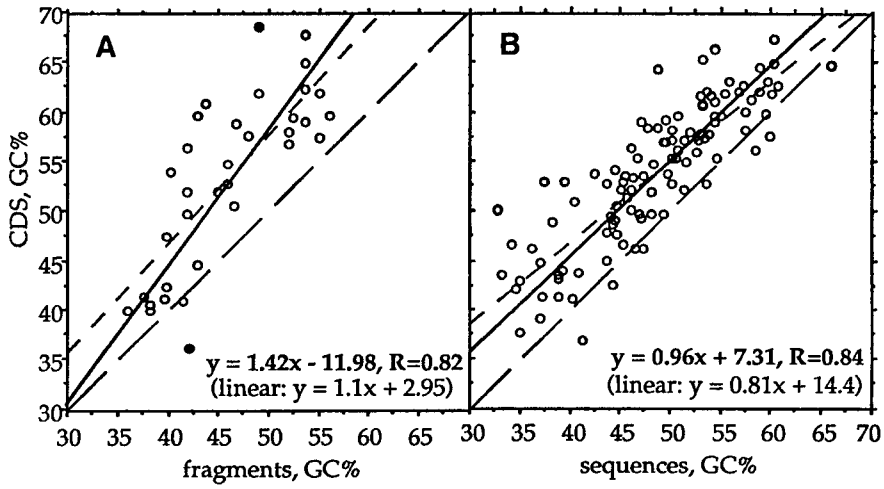

FIG. 1. Plots of GC levels of coding regions (CDS) from human genes against GC levels (A) of the DNA molecules (fragments) from compositional fractions in which the genes were localized (see Table 1) and (B) of long (≥10 kb) genomic DNA sequences containing the genes. The linear regression line (short dashes) and orthogonal regression line (solid) are shown together with equations, correlation coefficient, and slope 1 line (long dashes) for comparison. Two genes that were not used in calculating the regressions (entries 6 and 13 in Table 1) are shown as closed circles (N = 34 (A); 107 (B)).

from the previous study. The slopes were 2.38, i.e., an angle of 67° to the horizontal (previously 2.61, or 69°), and 3.45, i.e., 74°, for the linear and orthogonal regression lines, respectively, and the correlation coefficient was 0.82 (previously 0.82).

Figure 2B shows a plot of GC_3 of genes against the GC levels of database sequences in which they were present, with regression results almost identical to the previous study (Aïssani *et al.,* 1991; see also Mouchiroud *et al.,* 1991). The slopes of the linear and orthogonal regressions were 1.69, or 59° (previously 1.65, or 59°), and 2.31, or 67°, respectively, and the correlation coefficient was 0.83 (previously 0.84).

Figures 2C and 2D show plots of GC_3 against sequences 3′ of the stop codon ("3′ flanks"), and against 3′ flanks after removal of the first 500 bp following the stop codon ("far 3′ flanks"), respectively. The slopes of the linear and orthogonal regressions were 1.29 (52°)/1.20 (50°) and 2.63 (69°)/3.39 (74°), respectively, and the correlation coefficient was 0.66/0.56. These values, with orthogonal regressions closer to those of Fig. 2A than those of Fig. 2B, had no precedents in our previous study, since not enough sequences with long 3′ flanks were yet available. Although the angle and intercept (73.6°, −88.2) for the far 3′ flanks (Fig. 2D) are very similar to those (73.8°, −94.6) of the corresponding plot for the DNA fragments (Fig. 2A), the removal of the sequences and fragments containing CpG islands lowers both slopes to 71°, which is close to the value estimated in a previous study (70°; Mouchiroud *et al.,* 1991). This means that the true slope of GC_3 against the GC level of long stretches of intergenic DNA may not be quite as steep as those in Figs. 2A and 2D. This is also suggested by the regression lines obtained when 5′ flanks of non-CpG island genes are used as the abscissa instead of far 3′ flanks (Figs. 2E–2F; see below).

The choice of a 500-bp "buffer" between a gene and the intergenic DNA flanking it, used here to define the far 3′ flanks (and, by symmetry, the far 5′ flanks; see below), was motivated largely by practical limitations, namely by the scarcity of flanking sequences longer than about 2 kb in the database. Similarly, complete and far flanking regions are defined here relative to the CDS rather than relative to the mRNA or primary transcript, as precise and reliable specifications for the latter are less frequently given in the features (database sequence annotations) than for the CDS. However, 500 nt on either side of the coding sequence should often be sufficient to exclude the parts of the flanking regions in which selection pressures may create local compositional biases. Statistical analyses of 3′ flanking regions of genes (Pesole *et al.,* 1994) suggest that 500 nt will often cover the entire untranslated mRNA and/or primary transcript regions, which could experience selection pressures to maintain secondary mRNA structures or highly conserved regions (HCRs) (Duret *et al.,* 1993). Similarly, the 500 nt immediately 5′ of the ATG will often cover the untranslated untranscribed region as well as the first 100–200 bp upstream of the transcription start site, which experience the strongest pressure to maintain transcription factor binding sites. On the other hand, the local biases introduced by CpG island promoters of genes will typically extend further than 500 nt and can introduce a considerable scatter into correlations involving (far) 5′ flanks of CpG island genes (see below).

To check whether the correlation shown for far 3′ flanks in Fig. 2D is indeed similar to that obtained when only 3′-untranslated regions of the same genes are used, all sequences contributing to Fig. 2D for which the "poly A" or complete primary transcript locations were given in the features (56 of 103 sequences),

TABLE 1

List and Characteristic GC Levels of Human Genes Localized in DNA Fragments

No.	Name	Mnemonic	GC%				
			Exons or CDS	I + II	III	Introns	Fragment
1	β-Globin cluster	HUMHBB	54.0	48.3	65.5	38.5	40.3
2	Coagulation factor IX	HUMFIXG	41.3	44.3	35.3	38.9	39.7
3	HPRT	HUMHPRTB	41.1	41.8	39.7	39.0	41.5
4	Superoxide dismutase	HSSOD1G1	49.7	51.9	45.1	40.1	42.0
5	APP	HSPRA41	52.0	48.8	58.2	36.2	42.0
6	Interferon-α receptor	HSIFNAR	36.8	39.8	30.8	40.7	42.0
7	GART	HSGAGMR	44.8	47.1	40.1		43.0
8	c-mos	HUMCMOS	61.0	55.0	72.9		43.7
9	Vimentine	HUMVIM	56.4	49.0	71.1	41.8	42.0
10	ETS2	HSETS2A	52.8	46.1	66.4		46.0
11	ERG2	HSERG2	54.9	49.5	65.9		46.0
12	MX cluster	HSMXA + B	50.6	44.0	63.7		46.5
13	Collagen cluster	HUMCOLTHA + B	69.4	73.1	60.5		47.0
14	c-myc	HUMMYCC	58.8	50.0	76.4	50.7	46.7
15	Breast cancer induced protein	HSPS2	57.6	52.9	67.1	54.6	48.0
16	Glucose-6 phosphate dehydrogenase	HSG6PDGEN	59.4	46.0	86.2	54.3	52.4
17	c-Ha-ras 1	HUMRASH	59.1	48.1	81.1	69.0	53.7
18	α-Globin cluster	HUMHBA4 + 1	64.9	51.9	90.7	73.3	53.7
19	Propiomelanocortin	HUMPOMC	67.7	56.9	89.2	48.0	53.7
20	c-sis	HSCSIST	62.3	54.5	77.7	59.5	53.7
21	MCF2	HSMCF2PO	41.6	38.0	48.8		37.7
22	Butyrylcholinesterase	HSCHEB	40.0	42.5	34.8		36.1
23	Blym1 transforming gene	HSBLYM1	40.1	40.7	38.9	45.2	38.2
24	Salivary statherin gene	HSSTATH2	40.7	43.6	34.9		38.2
25	Adenylate kinase 1	HSAK1	57.4	48.2	73.2		55.1
26	c-fes/fps	HSFESFPS	61.9	52.7	80.2		55.1
27	Cystathionine β-synthase	HSCBSA	61.9	52.1	81.5		49.0
28	Iduronate 2-sulfatase	HSIDS	52.0	49.2	57.5	47.0	44.9
29	Biglycan	HSBGN2	59.6	46.6	85.6	62.1	42.9
30	L1 cell adhesion molecule	HSML1CAM	59.6	50.7	77.6		56.1
31	Red cone pigment	HSCPR6	56.8	47.3	75.9		52.0
32	Green cone pigment	HSCPGA3	58.1	49.0	76.4		52.0
33	Coagulation factor VIII	HSFVIIIR	42.6	43.2	41.4		39.8
34	Unknown protein	HSRNAC61A	47.5	46.1	50.3		39.8

Note: Entries 1–21 are from Aïssani *et al.* (1991; cf. also Mouchiroud *et al.*, 1991). GC% values of exons or coding region (CDS), first + second positions, introns, and total sequences are given wherever corresponding (genomic DNA or cDNA) sequence information was available.

and in which the 3′ untranslated region was at least 200 bp, were used to calculate the orthogonal regression of GC_3 against GC level of untranslated 3′ regions. The resulting slope and intercept (3.21, or 72.7°; −85.3) were very close to those of the far 3′ flanks, with an almost identical correlation coefficient (0.55).

Figures 2E and 2F show plots for genes that do not contain CpG islands, of GC_3 against sequences 5′ of the ATG codon ("5′ flanks") and against 5′ flanks after removal of the first 500 bp preceding the ATG ("far 5′ flanks"), respectively. Figures 2G and 2H show the corresponding plots for sequences that contain CpG island genes. The large scatter and poor correlation coefficients for the CpG island genes contrast with the much more reliable correlations for the genes containing no CpG islands. In fact, the 5′ flanks were the only regions of the sequences analyzed here that at first did not yield reliable correlations, due to heterogeneities in the height and/or width of CpG islands which are present in approximately half of the long (≥1 kb) sequences available for 5′ flanks. The large scatter and low correlation coefficients visible in Figs. 2G and 2H were also seen in other plots involving 5′ flanks of CpG island genes (e.g., with the GC level of coding regions or of introns as the ordinate) and in those involving all 5′ flanks regardless of CpG island status, but not in those involving only the 5′ flanks of non-CpG island sequences, even though this latter group of genes was smaller than the other groups.

Compositional Correlations of Introns with Coding Sequences and Flanking Sequences of Genes

Figure 3A presents a plot of GC levels of introns against coding regions of the same genes. The linear

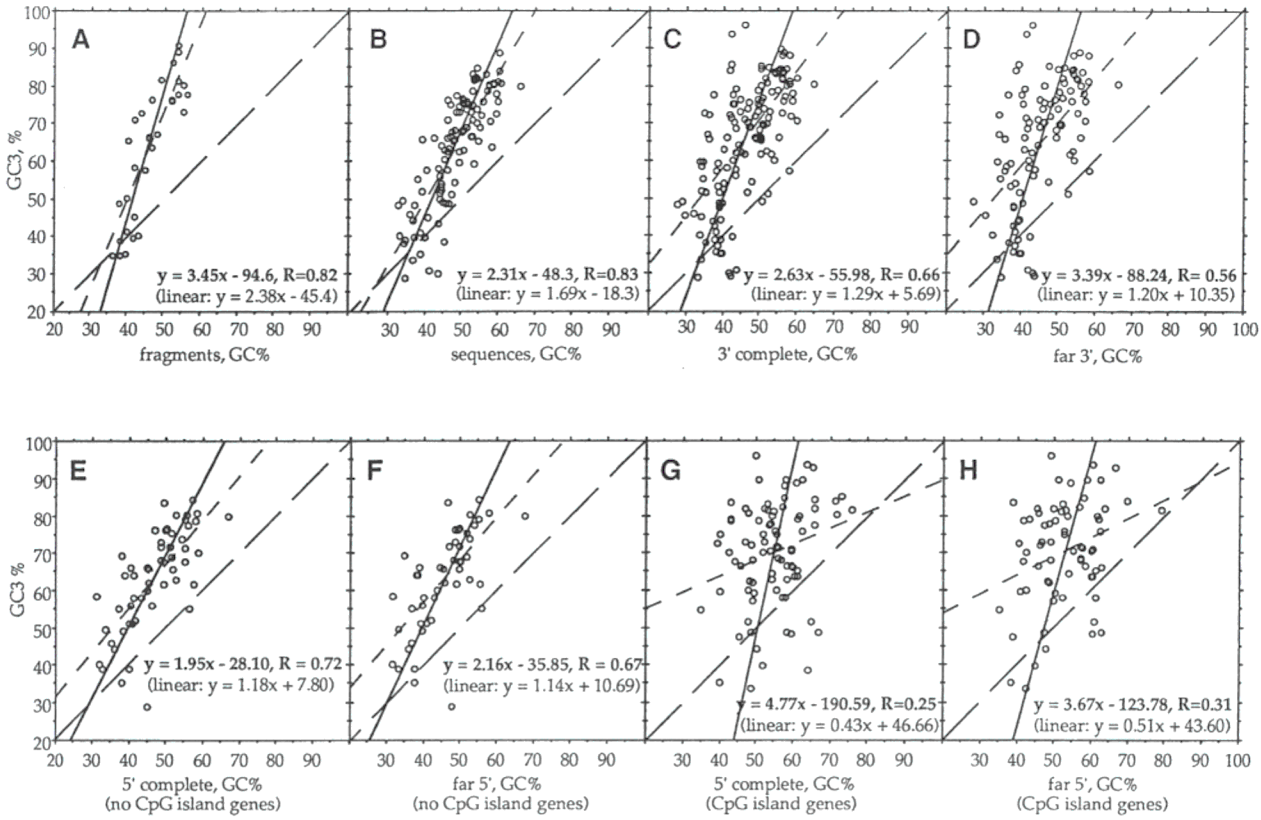

FIG. 2. Plots of GC levels of third codon positions from human genes against the GC levels (A) of DNA molecules (fragments) from compositional fractions in which the genes were localized (see Table 1); (B) of long (≥10 kb) sequences containing the genes; (C) of sequences (≥1 kb) 3′ of the coding region (3′ flanks); (D) of 3′ flanks after the removal of the 500 bp adjacent to the coding region (far 3′ flanks, ≥1kb); (E) of sequences (≥1 kb) 5′ of the coding region (5′ flanks), in sequences containing no CpG islands; (F) of 5′ flanks after the removal of the 500 bp adjacent to the coding region (far 5′ flanks, ≥1 kb), in sequences containing no CpG islands; (G) of 5′flanks (≥1 kb) in sequences containing CpG islands; (H) of far 5′ flanks (≥1 kb) in sequences containing CpG islands. Only sequences classified in the CpGIsle database are shown. Regression lines are shown as in Fig. 1 (N = 32 (A); 105 (B); 139 (C); 103 (D); 54 (E); 49 (F); 85 (G); 71 (H)).

and orthogonal regressions gave slopes of 0.90 (previously 1.0) and 1.20, respectively, with a correlation coefficient of 0.78 (previously 0.77).

Figures 3B and 3C show plots of the GC levels of introns against the GC levels of far 3′ flanks and far 5′ flanks of genes without CpG islands, respectively. These yield slopes of 0.89/0.85 for the linear and 1.18/1.22 for the orthogonal regression, and correlation coefficients of 0.78/0.73. The orthogonal regressions, which have good correlation coefficients, all have slopes slightly higher than 1, whether GC levels of introns are plotted against those of coding regions, far 3′ flanks, or far 5′ flanks of non-CpG island genes. The difference is that for the coding regions plot (Fig. 3A) the center of mass of the points (intersection of orthogonal and linear regression lines) is well below the main diagonal, whereas in the corresponding plots for the far 3′ and far 5′ flanks (Figs. 3B, 3C) it is slightly above it. Thus the orthogonal regression line diverges from the main diagonal with increasing GC in Fig. 3A but approaches it with increasing GC in Figs. 3B and 3C. Also, it should be noted that, whereas in Figs. 1 and 2 almost all the points are above the main diagonal, in Fig. 3A almost all the points are below it. This fact illustrates a general tendency (see Bernardi *et al.*, 1985) of GC levels of coding regions to be consistently higher than those of introns or intergenic sequences.

In summary, the results imply that coding regions typically have higher GC levels than introns, especially in GC-poor regions, and that in turn the introns of a gene tend to have higher GC levels than the surrounding intergenic DNA, especially in GC-rich regions. However, the scatter in Figs. 3B and 3C suggests that this latter tendency is not as strong as the former one.

Compositional Correlations among Flanking Sequences of Genes

Figure 4A (4B) shows plots of (far) 5′ against (far) 3′ GC levels for the few sequences that have both long (far) 3′ and long (far) 5′ flanks and are also listed as non-CpG island sequences in the CpGIsle database.

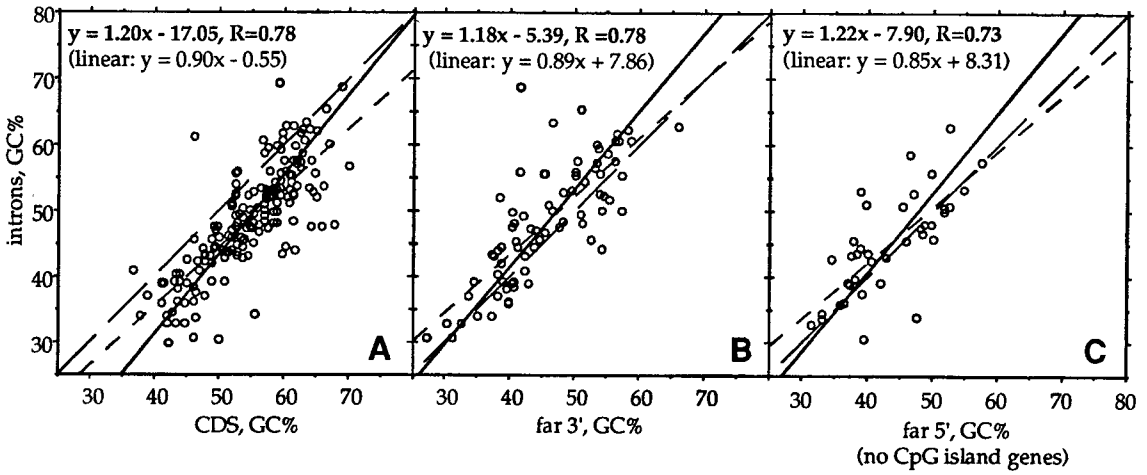

FIG. 3. Plot of GC levels of introns (≥1 kb) against GC levels (A) of coding regions ("CDS", ≥300 nt); (B) of far 3′ flanking regions; (C) of far 5′ flanking regions of genes with no CpG islands. Regression lines are shown as in Fig. 1 (N = 173 (A); 76 (B); 39 (C)).

The slopes of both plots are not far from unity, and the correlation coefficients (0.83, 0.7) are high for these small samples. In contrast, the same plots for a larger sample of genes containing CpG islands (not shown here) showed low correlation coefficients (<0.5). These results are consistent with a general tendency of GC levels to approach similar values on both sides of a gene, namely the GC level of the surrounding isochore.

Compositional Correlations among Codon Positions

Figure 5A shows a plot of the GC levels of third codon positions against the GC levels of first and second codon positions, for a large collection of 4270 cDNA and genomic DNA sequences. Figure 5B shows a contour representation of a 2-dimensional histogram of the same data, in which contours correspond to the relative number of genes in each cell of a 100 × 100 grid after removal of small local fluctuations by smoothing, and in which the highest contours have been omitted for clarity to show the crest of the distribution. This representation clearly shows the shape of the distribution, which is only faintly visible in the corresponding scatterplot due to the high density of points, i.e., indistinguishability of neighboring points even at high resolution.

The plots show a remarkable intrinsic linearity of the distribution, as well as the tendency of the orthogonal regression line (solid line) to follow the crest of the distribution, and the lack of such a tendency for the linear regression line (short dashes). The outliers (bottom right) are largely collagen genes. The orthogonal regression line, with slope 5.64 (79.9°), is very similar to that obtained for a subset of these data that was available 5 years ago (1600 sequences; Mouchiroud *et al.*, 1991), namely 5.07 (78.8°).

DISCUSSION

The GC Levels of GC-Rich Coding Sequences Stand out from the Average Isochore GC Content

While the linear regression of coding sequences GC levels (Fig. 1A) indicates that these values are consistently 5% higher than the corresponding isochore GC (see also Aïssani *et al.*, 1991), the orthogonal regression line shows that the two sets of values practically coincide for GC-poor coding sequences, yet diverge for GC-rich genes, the GC-richer genes being over 10% higher than the corresponding isochore GC. This result is of interest in that it provides an additional difference between GC-poor and GC-rich genes.

This conclusion is not contradicted by the results of Fig. 1B, which essentially show a consistently higher value of coding sequence GC compared to the GC of the entire sequences, and no divergence as in Fig. 1A. The reason for this is that the long sequences from the bank that were used in the plot consist mainly of introns, which are known to be lower in GC than the corresponding coding sequences (Bernardi *et al.*, 1985), with a difference that tends to be similar or smaller in GC-rich genes than for GC-poor genes (see Fig. 3A), thus counteracting the divergence seen in Fig. 1A. On the other hand, the DNA molecules shown in Fig. 1A consist largely of intergenic DNA (cf. also Aïssani *et al.*, 1991).

Other results presented here are also consistent with the idea that introns do not contrast as strongly as intergenic regions with coding regions, especially in GC-rich regions, with respect to their GC levels. The fact that in Fig. 3B the center of mass of the points is above the main diagonal, and that the slope of the orthogonal regression line is greater than unity, indeed suggests that introns tend to have higher GC levels than the in-

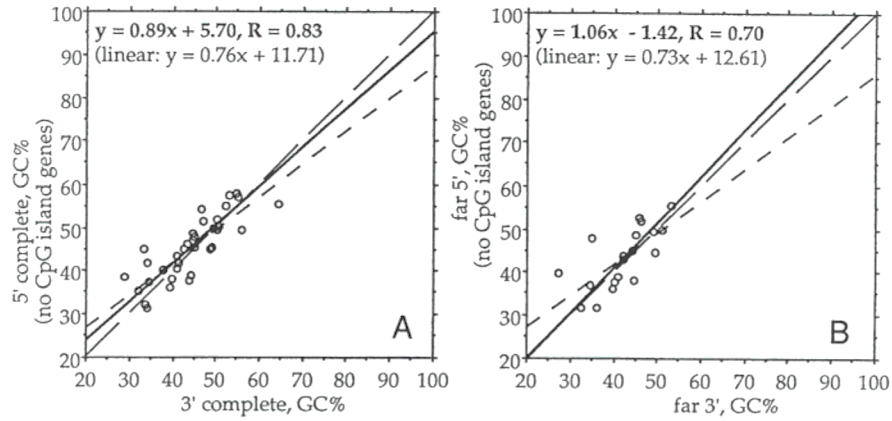

FIG. 4. Plots for sequences containing no CpG islands (A) of GC levels of 5′ flanking regions against GC levels of 3′ flanking regions; (B) of GC levels of far 5′ flanking regions against GC levels of far 3′ flanking regions. Regression lines are shown as in Fig. 1 (N = 35 (A); 19 (B)).

tergenic DNA flanking a gene, especially for GC-rich genes.

Similarly, the fact that the orthogonal regression lines in Figs. 2A, 2C, and 2D have a steeper slope than that in Fig. 2B may be explained, at least partly, by the relative contributions of intergenic DNA to the sequences or molecules analyzed.

The Compositional Correlations of Third Codon Positions with Isochores, Long Sequences from the Database, and Flanking Sequences

The correlation between the GC_3 of genes and the GC levels of the isochores containing them (Fig. 2A) is the base correlation for defining the distribution of genes in the human genome. Its main problem lies in the relatively laborious localization of sequenced genes in compositional DNA fractions, although the advent of PCR has made this task easier. It is, therefore, of interest to find the best equivalent of intergenic sequences in the sequences available in the database. The use of long sequences (Fig. 2B) is obviously not satisfactory for the reason, already mentioned, that noncoding regions in this data set are mainly introns, which have slightly higher GC levels than the intergenic regions. One should, therefore, resort to flanking sequences of genes.

The use of 3′ flanks (Fig. 2C) and far 3′ flanks (Fig. 2D) are thus more satisfactory, and indeed they most closely approach the results of Fig. 2A. The utility of 5′ flanks for predicting intergenic GC levels, or for establishing correlations involving them, is severely limited by the presence of CpG island heterogeneities on one hand, and on the other hand by the fact that non-CpG island genes, which yield good correlations, represent a biased sample in which GC-rich sequences are sys-

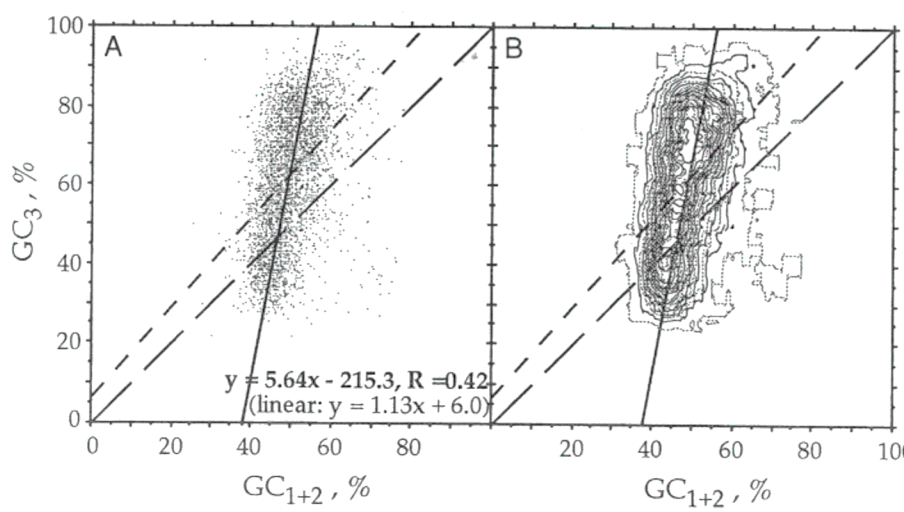

FIG. 5. Plot (A) and smoothed two-dimensional histogram (B) of GC levels of third codon positions against GC levels of first and second codon positions of 4270 human genes. Regression lines are shown as in Fig. 1.

tematically underrepresented (Aïssani and Bernardi, 1991a,b). In particular, these disadvantages limit the deductions and comparisons that are possible using the numerical results shown here of regressions involving 5′ flanks.

Further support for the idea that the GC levels of long far 3′ flanks may be the closest approximations to typical intergenic GC levels currently available from databank sequences was given by individual inspection of moving-window GC plots (using the GCG WINDOW program with a 3-kb window and step size 100 bp; not shown here) of over 40 genomic DNA sequences for genes with long 3′ or 5′ flanks, and for gene clusters with long intergenic regions. In these plots, the immediate 5′ flanks of genes typically had similar GC levels to those of the exons, or higher levels, whereas the 3′ flanks frequently showed a rapid decrease to the GC levels between the genes. This local 5′–3′ asymmetry was also seen around non-CpG island genes, such as those in the β-globin and growth hormone clusters, or in the cytokeratin 20 gene (HSHBB, HSGHCSA, HSCYTOK20). Thus the GC levels of 3′ flanks and far 3′ flanks, rather than those of introns or of the entire sequences (as used in previous studies, when few 3′ sequences were available), seem the most promising estimators of intergenic GC levels. The choice of (far) 3′ flanking regions is suggested also by the fact that almost none of the far or complete 3′ flanks of the database sequences analyzed here had GC levels exceeding 57%, the upper limit for the experimentally analyzed DNA fragments (which consist largely of intergenic DNA), whereas those of all other regions of the same database sequences (exons, third codon positions, introns, 5′ flanks, far 5′ flanks, and entire sequences) often exceed 60%.

It should be kept in mind, however, that although the actual values of GC levels in (far) 3′ flanking regions are probably closer to those of intergenic regions, the accuracy of the compositional correlations involving them, and thus their value as predictors of the GC level of the surrounding isochore, is not as high as for some of the other characteristic GC levels. Among the correlations presented here, the previously established correlation between GC_3 and the GC level of extended sequences is in fact more reliable than that between GC_3 and the GC level of (far) 3′ flanks, as can be seen from its higher correlation coefficient.

The local 5′–3′ asymmetry observed here, as well as the correlations between GC_3 and GC levels of 3′ and 5′ flanking regions, are compatible with previous results on the GC levels of noncoding regions in human mRNAs (Pesole *et al.*, 1994). These results also showed higher GC levels of the untranslated regions of genes on the 5′ side than on the 3′ side, and a linear regression between GC_3 and flanking GC levels with a steeper slope and lower correlation coefficient on the 5′ side than on the 3′ side. In the case of the 3′ mRNA flanks the correlation coefficient (0.67) was almost identical to that obtained here for the complete 3′ flanks (0.66), although the 5′ results differed (0.33 for the mRNA versus 0.72 for the complete DNA flanks), possibly due to heterogeneities from short 5′ flanks in the mRNA sequences that were excluded by the 1-kb size limit for the flanks in our analysis.

Conclusions

The present investigations define in greater detail the "genome equations" which concern coding and noncoding sequences and which represent a "genomic code" (see Bernardi, 1995). The main conclusions are the following. The correlations of Fig. 1 between coding sequences and the isochores embedding them (a) are not trivial, because coding sequences comprise only about 3% of the human genome, whereas noncoding sequences correspond to the remaining 97%; (b) show that while GC poor coding sequences are characterized by GC levels close to those of the isochores embedding them, GC-rich coding sequences tend to exhibit higher levels; and (c) imply that compositional constraints work in the same direction (e.g., toward increasing either GC or AT), although with different amplitudes, on coding sequences and on the isochores surrounding the corresponding genes. The correlation of Fig. 2A between GC_3 and GC levels of the isochores embedding the corresponding genes allows the study of the distribution of genes in the genome (Mouchiroud *et al.*, 1991; Bernardi, 1995). Finally, Fig. 5 shows that a correlation holds among GC levels of different codon positions (GC_3 vs GC_{1+2}), again because of compositional constraints working in the same direction, although not with the same amplitude. This correlation is, in fact, the universal correlation that holds for all eukaryotic and prokaryotic genes (D'Onofrio *et al.*, 1991).

ACKNOWLEDGMENTS

We thank Dr. Giuseppe D'Onofrio for careful reading of the manuscript and for discussions, and Dr. Gabriel Macaya for helpful comments. Simone Cacciò and Serguei Zoubak acknowledge the Federation of European Biochemical Societies for the award of a long-term fellowship.

APPENDIX

EMBL 42.0 mnemonics, CpG island status, lengths, and accession numbers of the 245 gene and gene cluster sequences, of length ≥10 kb and/or with long 3′ or 5′ flank, that were used in calculating one or more of the compositional correlations. Sequences containing more than one gene are shown with an asterisk. Sequences represented in the CpGIsle database are shown with closed circles if the sequence contains one or more CpG island genes, open if no CpG islands were found in the sequence.

APPENDIX

A03736 (3905 ; A03736)
A06939 (5402 ; A06939)
H17BHYD * (21764 ; M84472)
• HS01212 (3718 ; U01212)
HS01882 (13270 ; U01882)
HS04636 (9453 ; U04636)
HS07000 (152141 ; U07000)
HS07977 * (10808 ; U07977)
HS08988 (29243 ; U08988)
• HS10116 (10079 ; U10116)
o HS10685 (3510 ; U10685)
• HS10687 (11495 ; U10687)
HS10689 (4741 ; U10689)
HS15422 * (40573 ; U15422)
HS16812 (6478 ; U16812)
HS17969 (4862 ; U17969)
o HSA1ATP (12222 ; K02212)
o HSA1GLY2 (4944 ; M21540)
• HSACKI10 (6483 ; X14487)
HSACPP1 (10241 ; X74961)
• HSACTH (8658 ; V01510)
• HSADAG (36741 ; M13792)
HSADPRF02 (5167 ; M74493)
• HSADRA2R (3604 ; M23533)
o HSAFPCP (22166 ; M16110)
• HSAGAL (13662 ; M59199)
o HSAGG (4668 ; M11567)
HSAIMBG3 (4627 ; X54818)
• HSAK1 (12229 ; J04809)
• HSALADG (15913 ; X64467)
o HSALBGC (19002 ; M12523)
• HSALDC (6694 ; X05196)
HSALDOA (7530 ; X12447)
• HSALIFA (7614 ; M63420)
• HSALPI (5291 ; J03930)
• HSANT1 (5768 ; J04982)
• HSANT2X (4982 ; M57424)
HSAPOAI1 * (8966 ; J00098)
o HSAPOCII1 (4340 ; M10612)
• HSAPOE4 (5515 ; M10065)
o HSAT3 (14206 ; X68793)
o HSATP1A2 (26668 ; J05096)
• HSATPCP1 (9457 ; X69907)
HSATPCP2 (15016 ; X69908)
HSATPSAS (18004 ; D28126)
• HSATPSYB (10186 ; M27132)
• HSB94 (4180 ; M92357)
o HSBFXIII (33206 ; M64554)
o HSBHSD (9404 ; M38180)
HSBLGR1 * (35962 ; U07561)
• HSBMYH7 (28438 ; M57965)
o HSBNGF (11594 ; V01511)
o HSC1INHIB (18687 ; X54486)
• HSCALCAC (7637 ; X15943)
• HSCAMHCA (31462 ; Z20656)
• HSCD1R3 (6325 ; X14974)
HSCD36A (2561 ; L06849)
HSCD4 (13133 ; M86525)
o HSCDIR2 (10351 ; X14975)
o HSCEL (11502 ; M94579)
• HSCFVII (12850 ; J02933)
o HSCHYMASE (8124 ; M64269)
• HSCKIIBE (5917 ; X57152)
HSCOL7A1X (36631 ; L23982)
• HSCPH70 (6711 ; X52851)
HSCRCANTA (11288 ; Z21818)
o HSCRYGBC * (22775 ; M19364)
o HSCSF1PO (35100 ; X14720)
HSCSN2A (10608 ; L10615)
• HSCST3G (7292 ; X52255)
• HSCYP2D6 (9432 ; M33388)
• HSCYP450 (8028 ; X02612)
HSCYP7A (4959 ; L13460)
• HSCYPIIE (14776 ; J02843)
o HSCYTOK20 (18061 ; X73501)
o HSD1DO (3341 ; X55760)
• HSDAO (9903 ; X78212)
• HSDES01 (11990 ; M63391)
HSDS (22573 ; D26535)
o HSDZA2G (14694 ; D14034)
HSEDN1B (12461 ; J05008)
• HSEF1A (4695 ; J04617)
• HSENO2 (10905 ; X51956)
HSENO3 (7194 ; X56832)
HSENOAL4 (10207 ; X16290)
HSEPOHYDD (24790 ; L29766)
HSFBRG1 (10564 ; M10014)
• HSFESFPS (12263 ; X06292)
o HSFIBRA (7576 ; M64982)
o HSFIXG (38059 ; K02402)
• HSFMR1S (152351 ; L29074)
• HSFOS (6210 ; K00650)
HSFURIN (14977 ; X15723)
o HSG0S19B (4788 ; M24110)
• HSG0S8PP (7345 ; L13391)
HSG17G (5443 ; X80700)
• HSG6PDGEN * (52173 ; X55448)
HSGAD45A (5378 ; L24498)
o HSGAMGLOB * (11393 ; M91037)
o HSGASTA (7739 ; M15958)
o HSGHCSA * (66495 ; J03071)
• HSGLA (12436 ; X14448)
o HSGLUCG2 (10050 ; X03991)
HSGLUT4B (8402 ; M91463)
o HSGPIBAA (6062 ; M22403)
o HSGPIX (3201 ; M80478)
• HSGPV (7452 ; Z23091)
o HSGTRH (7224 ; X15215)
• HSGXA (3957 ; D90150)
• HSHA44G (6826 ; X06956)
HSHAPA (11390 ; D28877)
HSHBA4 * (12847 ; J00153)
HSHBB2 * (73326 ; J00179)
HSHCF1 (17760 ; X79198)
• HSHCF2 (15849 ; M58600)
HSHH1RE (5856 ; D14436)
• HSHKATPC (17201 ; M63962)
• HSHLADPB * (14782 ; X02228)
• HSHLADZA (5691 ; X02882)
• HSHMG14A (8882 ; M21339)
• HSHMG17G (7195 ; X13546)
• HSHMG2A (4341 ; M83665)
• HSHMGIY (10144 ; L17131)
• HSHOX3D (4968 ; X61755)
• HSHOX51 (6305 ; X17360)
o HSHP2HPR * (38542 ; M69197)
• HSHPRT8A (56737 ; M26434)
HSHSD3BA (9127 ; M77144)
• HSHSP90B (8210 ; J04988)
• HSHST (6616 ; J02986)
• HSIBP3 (10884 ; M35878)
HSIDS (36845 ; L35485)
o HSIFD1 * (9937 ; V00531)
• HSIFNAR (32906 ; X60459)
• HSIGF2G (8837 ; X03562)
o HSIL1AG (11970 ; X03833)
o HSIL1B (9721 ; X04500)
o HSIL1RECA (12565 ; X64532)
o HSIL8A (5191 ; M28130)
HSIL8R (13089 ; M99412)
HSIL8RAB (9269 ; L19592)
HSIL9RA (17073 ; L39064)
HSINSTHIG * (12565 ; L15440)
o HSINSU (4992 ; V00565)
• HSINT2 (11608 ; X14445)
HSINTFA4B (3878 ; X55733)
HSKALLIST (9618 ; L28101)
• HSKER101 (6520 ; M24842)
o HSKERTRA (14537 ; M98447)
• HSL7A (5506 ; X52138)
o HSLCATG (6901 ; X04981)
o HSLHDC (32351 ; D16583)
• HSLYL1B (4569 ; M22638)
• HSLYTOXBB (6305 ; L11016)
o HSMGPA (7734 ; M55270)
HSMHA3C7B (8892 ; L29411)
• HSMHCD8A (7319 ; M27161)
o HSMHCP42 (5141 ; M12792)
• HSMHRP1B * (12120 ; L26261)
• HSMK (4638 ; D10604)
• HSMPOG (14638 ; X15377)
o HSMRP8A (4195 ; M21005)
o HSMTS1G (3002 ; Z33457)
HSMYC3L (7011 ; M19720)
• HSMYCL2A (3854 ; J03069)
• HSNCAMX1 (16288 ; Z29373)
o HSNKG5PRO (6746 ; M85276)
HSNM23H1 (8461 ; X75598)
HSNMYC01 (6788 ; M13241)
• HSNUCLEO (10942 ; M60858)
HSODC1A (9373 ; M81740)
• HSOP18A (8058 ; M31303)
o HSOPS (6953 ; K02281)
o HSOSTP (10881 ; D14813)
• HSP3 (4379 ; X12458)
HSP45C17S (8673 ; M63871)
• HSP53G (20303 ; X54156)
HSPACAP (17041 ; X60435)
o HSPAIA (17509 ; J03764)
• HSPBGDA (10024 ; M95623)
HSPBX21 * (10108 ; D28769)
o HSPCI (15571 ; M68516)
• HSPDHAL (17082 ; D90084)
• HSPDHBET (8869 ; D90086)
• HSPKM12 (10368 ; X56494)
• HSPNMTA (4174 ; J03280)
HSPOMC (8658 ; K02406)
o HSPP14B (8076 ; M34046)
• HSPR264SC (4251 ; X75755)
o HSPRB3L (4516 ; X07881)
o HSPRB4S (5813 ; X07882)
• HSPRCA (11725 ; M11228)
HSPREP (4605 ; D43639)
o HSPRF1A (6218 ; M31951)
HSPROFILX (17630 ; M96943)
o HSPROPG (8222 ; X70872)
HSPROSCHY * (13863 ; X71874)
o HSPRPH1 (4946 ; M13057)
• HSRAS1 (6453 ; V00574)
HSRBPA (13646 ; L34219)
• HSRCC1 (34641 ; D00591)
o HSREGA01 (4251 ; J05412)
• HSRETBLAS (180388 ; L11910)
o HSRIGBCHA (11298 ; M89796)
• HSRPS14 (5985 ; M13934)
HSRPS7 (7513 ; Z25749)
HSS100DE * (10952 ; Z18950)
o HSSAA1A (6943 ; X13895)
• HSSAACT (3778 ; M20543)
o HSSEMI (4164 ; M81650)
HSSEMIIB (8224 ; M81651)
• HSSEQX (9416 ; L06237)
• HSSHBGA (6087 ; M31651)
o HSSIAL (6503 ; X52075)
HSSKR2 (11352 ; L31848)
o HSSPBAA (10476 ; M24461)
• HSSPRO (5296 ; X05006)
• HSSRYZ (4737 ; L08063)
HSSTATH2 (4723 ; M32639)
HSSYND1GN (4797 ; Z48199)
HSTA (7666 ; D32046)
o HSTBGA (8769 ; L13470)
• HSTDGF1A (7355 ; M96955)
• HSTEF1 (4443 ; M63896)
• HSTFPB (13865 ; J02846)
• HSTHB (26928 ; M17262)
• HSTKRA (13500 ; M15205)
• HSTM (4593 ; D00210)
o HSTNFAB * (7112 ; M16441)
o HSTPA (36594 ; K03021)
o HSTPALBU (6172 ; L14927)
• HSTPI1G (5000 ; X69723)
• HSTRHYAL (9551 ; L09190)
HSTRINUC (10348 ; L12392)
• HSTS1 (18596 ; D00596)
• HSUBA52G (4555 ; X56997)
• HSUBILP (3583 ; J03589)
• HSUDPCNA (4705 ; M61829)
o HSVCAM1A (5607 ; M73255)
o HSVIPAA (10961 ; M33027)
o HSVITDBP (55136 ; L10641)
HSXBXVIII (12700 ; X71937)
HSYAP65 (5153 ; X80507)
MIHSAPT1 (23065 ; D16561)
S85963 (6152 ; S85963)

REFERENCES

Aïssani, B., and Bernardi, G. (1991a). CpG islands: Features and distribution in the genomes of vertebrates. *Gene* **106:** 173–183.

Aïssani, B., and Bernardi, G. (1991b). CpG islands, genes and isochores in the genome of vertebrates. *Gene* **106:** 185–195.

Aïssani, B., D'Onofrio, G., Mouchiroud, D., Gardiner, K., Gauthier, C., and Bernardi, G. (1991). The Compositional Properties of Human Genes. *J. Mol. Evol.* **32:** 493–503.

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

Benson, D. A., Boguski, M., Lipman, D. J., and Ostell, J. (1994). GenBank. *Nucleic Acids Res.* **22:** 3441–3444.

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

Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M., and Rodier, F. (1985). The mosaic genome of warm-blooded vertebrates. *Science* **228:** 953–958.

DeSario, A., Geigl, E. M., Palmieri, G., D'Urso, M., and Bernardi, G. (1995a). A compositional map of human Xq28. *Proc. Natl. Acad. Sci. USA,* in press.

DeSario, A., Geigl, E. M., and Bernardi, G. (1995b). A rapid procedure for the compositional analysis of yeast artificial chromosomes. *Nucleic Acids Res.* **23:** 4013–4014.

Dessen, P., Fondrat, C., Valencien, C., Mugnier, C. (1990). BISANCE: A French service for access to biomolecular sequence databases. CABIOS **6:** 355–356.

Devereux, J., Haeberli, P., and Smithies, O. (1984). A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res.* **12:** 387–395.

D'Onofrio, G., Mouchiroud, D., Aïssani, B., Gautier, C., and Bernardi, G. (1991). Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. *J. Mol. Evol.* **32:** 504–510.

Duret, L., Dorkeld, F., and Gautier, C. (1993). Strong conservation of non-coding sequences during vertebrate evolution: Potential involvement of post-transcriptional regulation of gene expression. *Nucleic Acid Res.* **21:** 2315–2322.

Gouy, M., Gautier, C., Attimonelli, N., Lanave, C., and Di Paola, G. (1985). ACNUC—Portable retrieval system for nucleic acid sequence database: Logical and physical design and usage. CABIOS **1:**167–172.

Larsen, F., Gundersen, G., Lopez, R., and Prydz, H. (1992). CpG islands as gene markers in the human genome. *Genomics* **13:**1095–1107.

Montero, L. M., Salinas, J., Matassi, C., and Bernardi, G. (1990). Gene distribution and isochore organization in the nuclear genomes of plants. *Nucleic Acids Res.* **18:** 1859–1867.

Mouchiroud, D., D'Onofrio, G., Aïssani, B., Macaya, G., Gautier, C., and Bernardi, G. (1991). The distribution of genes in the human genome. *Gene* **100:** 181–187.

Pesole, G., Fiormarino, C., and Saccone, C. (1994). Sequence analysis and compositional properties of untranslated regions of human mRNAs. *Gene* **140:** 219–225.

Saccone, S., De Sario, A., Della Valle, G., and Bernardi, G. (1992). The highest gene concentrations in the human genome are in T-bands of metaphase chromosomes. *Proc. Natl. Acad. Sci. USA* **89:** 4913–4917.

Saccone, S., De Sario, A., Wiegant, J., Raap, A. K., Della Valle, G., and Bernardi, G. (1993). Correlations between isochores and chromosomal bands in the human genome. *Proc. Natl. Acad. Sci. USA* **90:** 11929–11933.

Salinas, J., Matassi, C., Montero, L. M., and Bernardi, G. (1988). Compositional compartmentalization and compositional patterns in the nuclear genomes of plants. *Nucleic Acids Res.* **16:** 4269–4285.

Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989). "Molecular Cloning: A Laboratory Manual," 2nd ed., Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proc. 6th Int. Congr. Genetics* **1:** 356–366. *Reprinted in: "Evolution: Selected Papers of Sewall Wright"* (W. B. Provine, Ed.), Univ. of Chicago Press, Chicago and London, 1986.

Zerial, M., Salinas, J., Filipski, J., and Bernardi, G. (1986). Gene distribution and nucleotide sequence organization in the human genome. *Eur. J. Biochem.* **160:** 479–485.