COMPOSITIONAL CORRELATIONS AND GENE DISTRIBUTION OF THE HUMAN GENOME

OLIVER CLAY, GIUSEPPE D'ONOFRIO,
KAMEL JABBARI, SERGUEI ZOUBAK,
SALVATORE SACCONE, AND GIORGIO BERNARDI

ABSTRACT This review briefly describes the compositional approach to the animals of vertebrate genomes. This approach involves the study of distributions of, and correlations among, the base compositions (GC levels) of different parts of these genomes, such as exons, introns, third codon positions, flanking of genes, and long genomic sequences or fragments spanning genic and intergenic DNA. Properties of the human genome that were inferred using the compositional approach include its organization into isochores, the presence of much higher gene densities in GC rich than in GC poor regions, and the non-uniform concentration of genes in the chromosomal bands.

The term genome was coined over three quarters of a century ago by a German botanist, Hans Winkler (1920), to designate the haploid chromosome set. Although textbooks of molecular biology still do not venture beyond this purely operational definition of the eukaryotic genome (or its more modern variant, the sum total of genes and of intergenic sequences), there is growing evidence that a genome is more than the sum of its parts and, in particular, that there are important structural and functional interactions between the coding regions of the human genome and the vast majority of this genome that is noncoding. Over the past 30 years, the study of the properties of DNA that can be described in terms of base composition, i.e., in terms of GC (the percentage of guanine and cytosine nucleotides), has yielded valuable tools for the analysis and comparison of genomes. The most important compositional properties found in eukaryotic genomes include isochore organization, compositional distributions of DNA fragments and of coding sequences, and compositional correlations between coding and non-coding sequences. These properties, which often have functional and structural correlates, and which have been extensively investigated in the nuclear genomes of vertebrates in our laboratory (see Bernardi et al., 1985; Bernardi 1989, 1993, 1995), will be briefly outlined here.

S. P. Wasser (ed.), Evolutionary Theory and Processes: Modern Perspectives, Papers in Honour of Eviatar Nevo
O. Clay et al. Compositional Correlations and Gene Distribution of the Human Genome, 115-123.
© 1999 *Kluwer Academic Publishers. Printed in the Netherlands.*

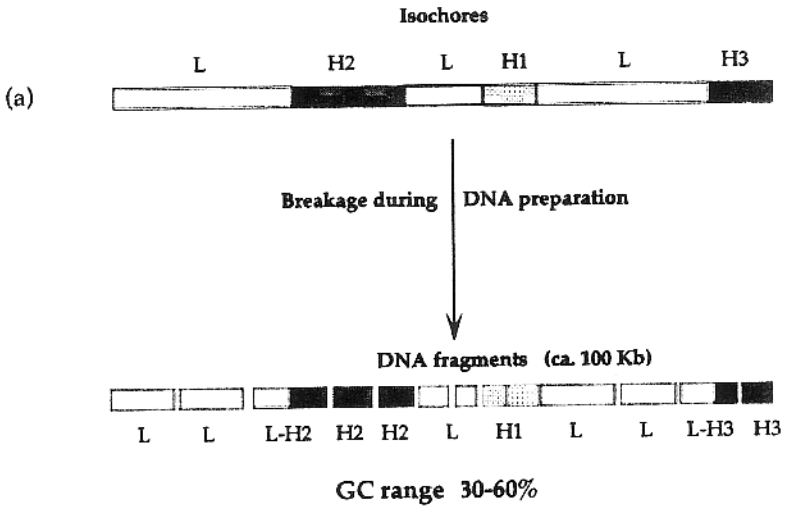
The mammalian genomes are mosaics of isochores (see Figure 1a), namely of long (>300 kb), compositionally homogeneous DNA segments, with base compositions ranging from 30 to 60% GC in the case of the human genome (Thiery et al. 1976; Macaya et al. 1976). This is an extremely wide range, since it is almost as wide as that (about 25-75% GC) covered by all bacterial DNAs. In the human genome, isochores can be assigned to two GC-poor families (L1 and L2) representing 2/3 of the genome, and to three GC-rich families (H1, H2 and H3) forming the remaining 1/3 (Figure 1b).

The compositional distributions of large (>100 kb) genomic fragments, such as those obtained during routine DNA preparations, of exons (and particularly of their third codon positions), and of introns, can be viewed as compositional patterns (Bernardi et al. 1985) that represent genome phenotypes (Bernardi and Bernardi, 1986): these distributions differ characteristically not only between cold- and warm-blooded vertebrates, but also between mammals and birds, and even between murids and most other mammals (see Figure 2).

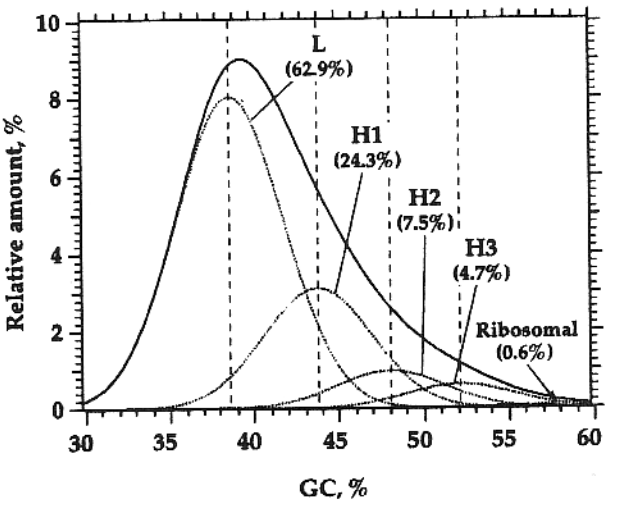
Compositional correlations (Bernardi et al. 1985) exist (Figure 3) between exons (and their codon positions) and the isochores in which they are located, as well as between the exons (and their codon positions) and the introns or flanking regions of genes (Aissani et al. 1991; Clay et al. 1996). The average GC level in each of the functionally or structurally characteristic parts of a sequenced gene (5' flank/promoter, coding exons, the three codon positions, introns, 3' flank, and embedding isochore), where this can be determined, gives a data point. From the data available for a given species, pairwise orthogonal regressions between these characteristic GC levels can then be computed. Each of the resulting pairwise scatterplots (or bivariate probability densities modelling them) can be characterized by an orthogonal regression line (slope and intercept or center of mass) and a correlation coefficient. The resulting compositional correlations have non-trivial implications, since they quantitatively relate coding sequences, which account for only 3% of the human genome, for example, to non-coding sequences, which account for 97% of this genome. For any species, the set of correlations between the characteristic GC levels of its genes or isochores can be viewed as defining a genomic code (Bernardi 1993) of the species. Many of these correlations can be order- or even species-specific. The intragenomic correlation between GC levels of the third codon position and that of the other codon positions for human genes (Figure 3d) is however described by essentially the same orthogonal regression line (major axis) as the corresponding intergenomic correlation representing all eukaryotic and prokaryotic species for which sequence data are available, and is thus a universal correlation (D'Onofrio and Bernardi 1992). Both the genomic code and the universal correlation are apparently due to compositional constraints working in the same direction (towards GC or AT) on coding and non-coding sequences, as well as on different codon positions, yet with different strengths.

The compositional correlations between GC_3 (the GC level of third codon positions) and isochore GC have a practical interest in that they allow one to position the coding sequence histogram of Figure 2b relative to the CsCl profile of Figure 1b, i.e., to the GC distribution of 50-100 kb fragments of human DNA, and thus to assess

Figure 1. (a) Scheme of the isochore organization of the human genome
(b) The isochore families of the human genome



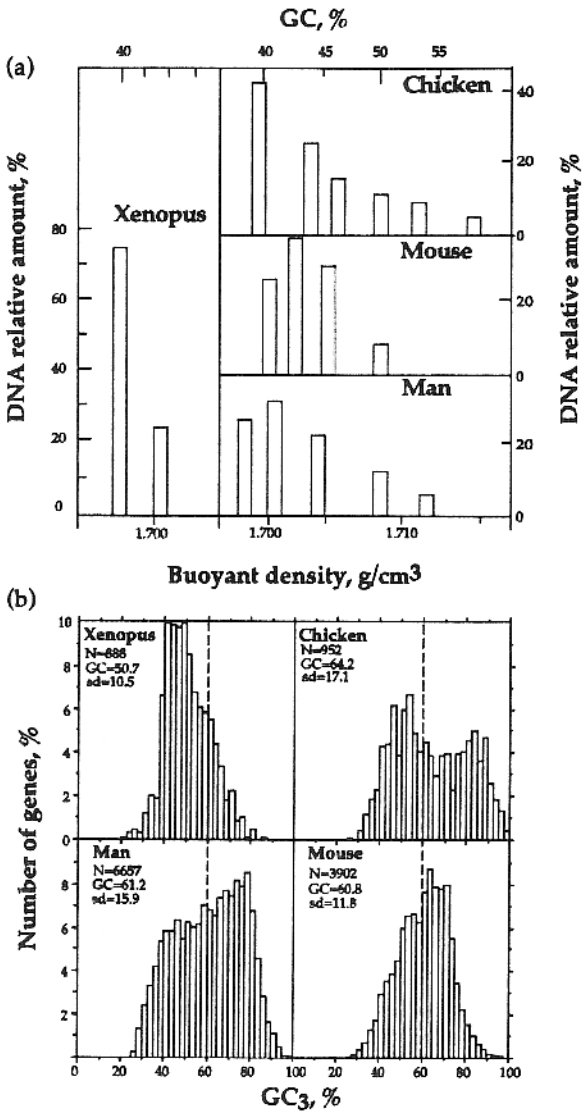
(b)



a) This genome, a typical mammalian genome (Sabeur et al. 1993), is a mosaic of large (>300 kb) DNA segments, the isochores. These are compositionally homogeneous (above a size of 3 kb, and apart from local fluctuations within genes) and can be partitioned into a small number of families, GC-poor (L1 and L2), GC-rich (H1 and H2), and very GC-rich (H3). The GC range of the isochores from the human genome is 30-60% (modified from Bernardi 1993).

b) The contributions of DNA fragments derived from isochore families L (*i.e.*, L1 + L2), H1, H2, H3 are superimposed on the CsCl profile of human DNA (modified from Zoubak et al. 1996).

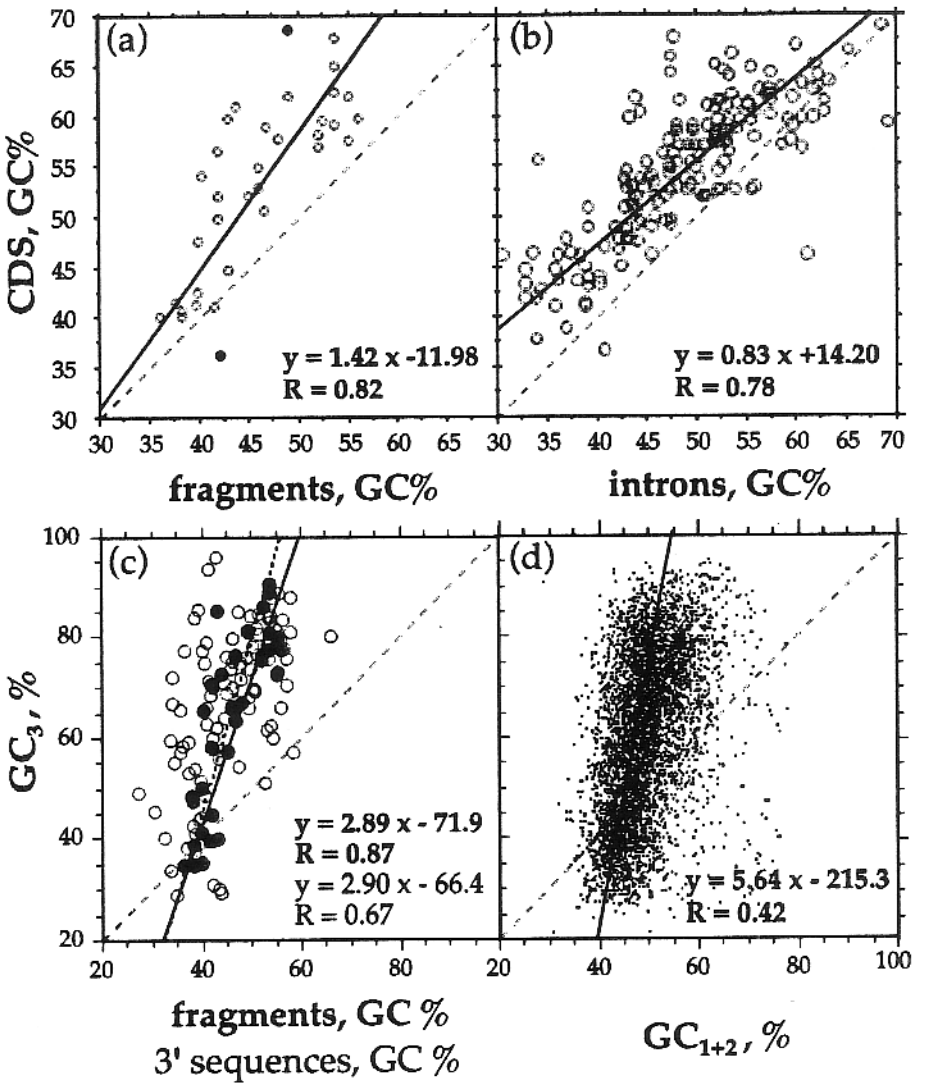
Figure 2. (a) Compositional patterns of vertebrate genomes
(b) Compositional distribution of third codon positions from vertebrate genes



a) Histograms showing the relative amounts, modal buoyant densities and GC levels of the major DNA components from *Xenopus*, chicken, mouse and man, as estimated after fractionation of DNA by preparative density gradient in the presence of a sequence-specific DNA ligand (Ag^+ or BAMD; BAMD is bis (acetato-mercuri-methyl) dioxane). The major DNA components are the families of large DNA fragments derived from different isochore families (see Figure 1). Satellite and minor DNA components (such as rDNA) are not shown in these histograms (modified from Bernardi 1993).

b) The number of genes taken into account is indicated. A 2.5% GC_3 window was used. (Updated from Bernardi 1993).

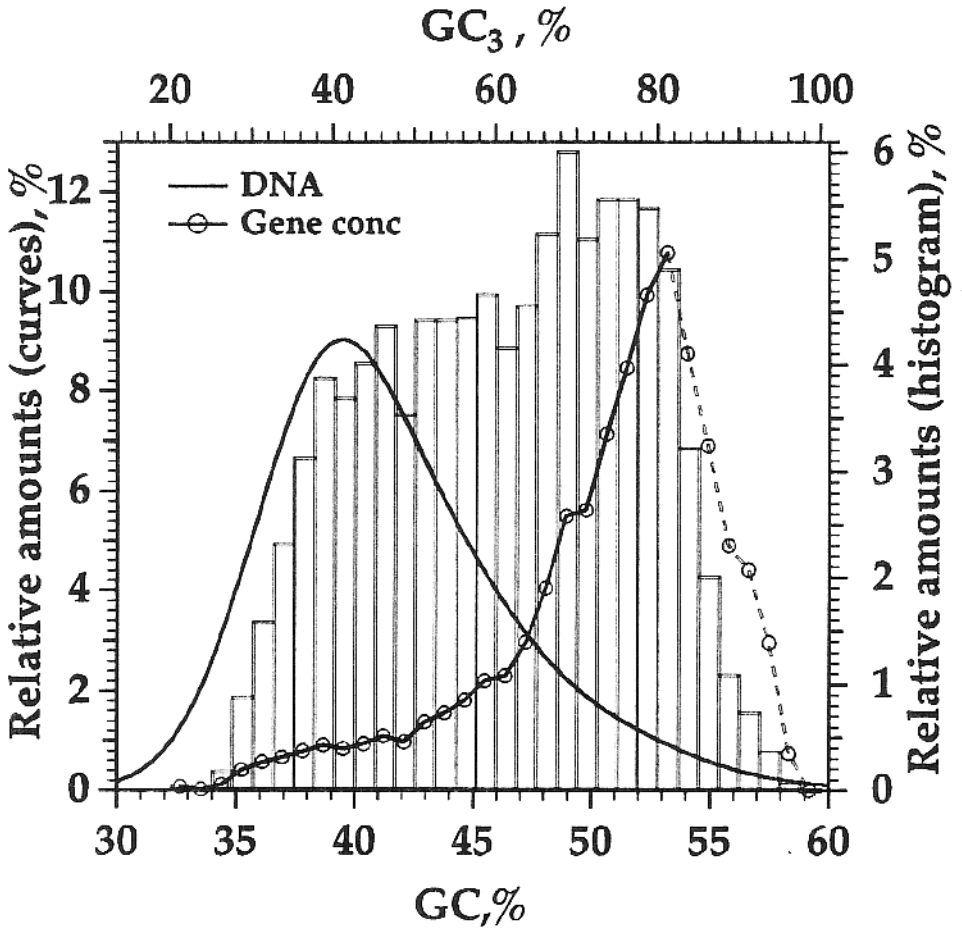
Figure 3. GC levels of human coding sequences (CDS) are plotted (a) against GC levels of the DNA fragments in which they were experimentally localized; (b) against the GC levels of the corresponding introns. GC_3 of human coding sequences is plotted (c) against the GC levels of the DNA fragments containing the corresponding genes (filled circles) or the 3' flanking sequences of the latter (closed circles); (d) against GC_{1+2}



In all plots, orthogonal relationships are shown along with the diagonal (slope = 1), the equations, and the correlation coefficients (modified from Clay et al. 1996). The solid line shown in (c), which was used to align GC_3 and GC scales when calculating the gene concentration profile (GC_3 = 2.92 GC - 74.3; see text, and Figure 4), was determined independently by comparing the decomposition of Fig. 1b with the analogous decomposition of the GC_3 distribution of human genes.

the gene distribution in the human genome (Mouchiroud et al. 1991; Bernardi 1995; Zoubak et al. 1996). In fact, if one divides the relative number of genes per histogram bar by the corresponding relative amount of DNA, one can see that the ratio, namely the (relative) gene concentration, is low in GC-poor isochores, increases with increasing GC in isochore families H1 and H2, and reaches a maximum in isochore family H3, which exhibits at least a 17-fold higher gene concentration compared to GC-poor isochores (Figure 4).

Figure 4. Profile of gene concentration in the human genome as obtained by dividing the relative amounts of genes in each 2.5% GC_3 interval of the GC_3 histogram (N=4270) by the corresponding relative amounts of DNA deduced from the CsCl profile



The apparent decrease in gene concentration for very high GC values (broken line) is due to the presence of rDNA in that region. The last concentration values are uncertain because they correspond to very low amounts of DNA (modified from Zoubak et al. 1996).

The H3 isochore family has also been named the human genome core (Bernardi 1993), because it corresponds to the functionally most significant part of the human genome. Indeed, the H3 isochore family is not only endowed with the highest gene (and CpG island) concentration, but also with an open chromatin structure, as witnessed by the accessibility to DNases, by the scarcity of histone H1, the acetylation of histones H3 and H4 and a wider nucleosome spacing (Tazi and Bird 1991), as well as with the highest transcription and recombination levels and with the earliest replication timing (Federico et al. 1998). The genes of the genome core have the highest GC_3 levels relative to their flanking sequences (see Figure 2a), have the shortest genes and the least frequent and shortest introns (Duret et al. 1995), exhibit an extreme codon usage, and encode proteins characterized by amino acid frequencies differing from those of proteins encoded by GC-poor isochores (D'Onofrio et al. 1991).

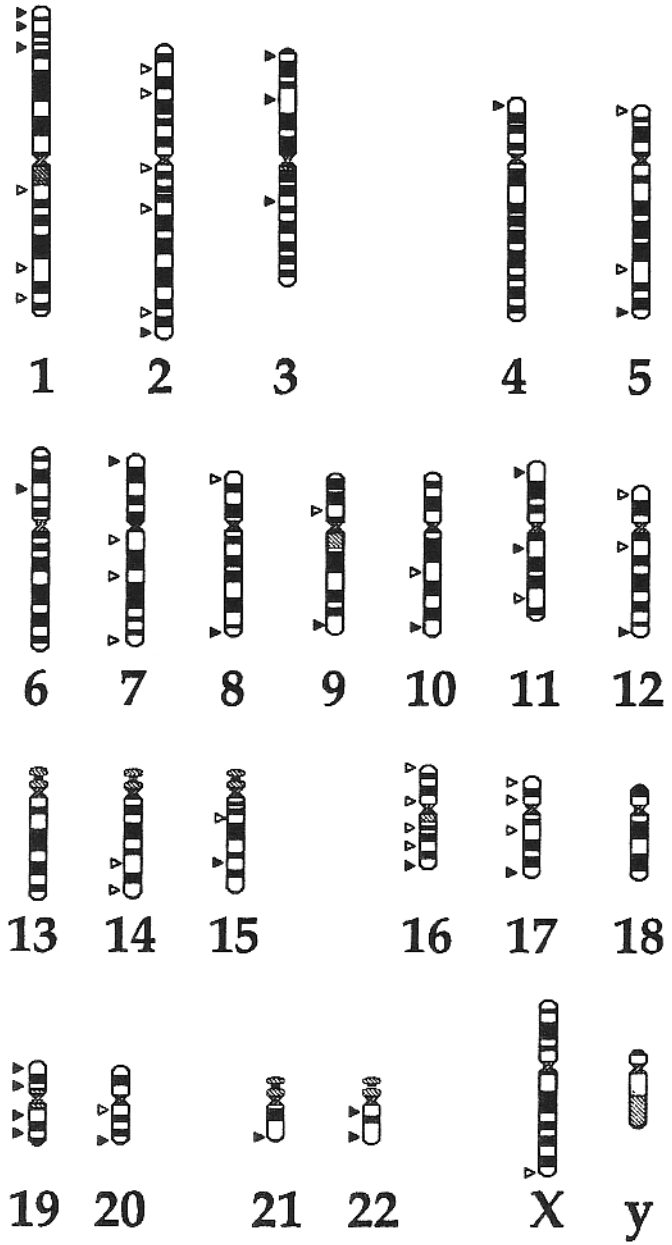
The isochores of the human genome core are located in about 30 $H3^+$ or T(elomeric)-bands (Saccone et al. 1992; 1993; 1996), which are largely formed by GC-rich isochores of the H3 family, and in about 30 $H3^*$ or T' bands, which contain a much smaller proportion of isochores from H3. The remaining 140 $H3^-$ R(everse) bands (at a 400 band resolution) comprise both GC-rich isochores (of the H1 family) and GC-poor isochores, but do not contain H3 isochores. G(iemsa) bands consist almost exclusively of GC-poor isochores (Saccone et al. 1993; see Figure 5).

It should be stressed that the tendency exemplified by the gene distribution of the human genome seems to have been conserved in evolution: the concentration of genes is highest in the GC-richest isochores in all vertebrates (Bernardi 1995).

As mentioned above, the compositional pattern of the human genome, which is typical of the genomes of most mammals and similar to the genomes of birds, is strikingly different from the compositional patterns of cold-blooded vertebrates, which exhibit a much lower degree of heterogeneity and are characterized by metaphase chromosomes that do not exhibit R banding (except to a low extent in some reptiles). These different genome phenotypes of warm- *versus* cold-blooded vertebrates are due to compositional changes. While the gene-poor, GC-poor isochores of cold-blooded vertebrates underwent little or no compositional change during the transition to warm-blooded vertebrates, the gene-rich, GC-rich isochores underwent dramatic compositional changes at that transition.

In the case of homologous mammalian genes, we have been able to show that synonymous substitutions in third codon positions exhibit frequencies, and are associated with compositions, that are highly non-random and suggest the influence of natural selection (Caccio et al. 1995; Zoubak et al. 1995; Alvarez-Vallin et al. 1998). Under these circumstances, the compositional changes in non-coding sequences, which are correlated with those occurring in third codon positions, suggest that non-coding sequences are not junk DNA, but must fulfill some functional role.

Figure 5. Distribution of sequences hybridizing DNA from H3 isochores on human chromosomes. $H3^{+}$ or T bands (solid arrows) correspond to strong signals, $H3^{*}$ or T' bands (open arrows) to medium signals. The remaining $H3^{-}$ R bands correspond to undetectable signals (modified from Saccone et al. 1996).



References

Aissani, B., G. D'Onofrio, D. Mouchiroud, K. Gardiner, C. Gautier, and G. Bernardi. 1991 The compositional properties of human genes. Journal of Molecular Evolution 32: 497-503.

Alvarez-Vallin, F., K. Jabbari, and G. Bernardi. 1998. Synonymous and nonsynonymous substitutions in mammalian genes: intragenic correlations. Journal of Molecular Evolution 46: 37-44.

Bernardi G. 1989. The isochore organization of the human genome. Annual Review of Genetics 23: 637-661

Bernardi, G. 1993. The human genome organization and its evolutionary history: a review. Gene 135: 57-66.

Bernardi, G. 1995. The human genome: organization and evolutionary history. Annual Review Genetics 29: 445-476.

Bernardi, G. and G. Bernardi. 1986. Compositional constraints and genome evolution. Journal of Molecular Evolution 24: 1-11.

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

Caccio, S., S. Zoubak, G. D'Onofrio, and G. Bernardi. 1995. Nonrandom frequency patterns of synonymous substitutions in homologous mammalian genes. Journal of Molecular Evolution 40: 280-292.

Clay, O., S. Caccio, S. Zoubak, D. Mouchiroud, and G. Bernardi. 1996. Human coding and non-coding DNA: compositional correlations. Molecular Phylogenetics and Evolution 5: 2-12.

D'Onofrio, G. and G. Bernardi. 1992. A universal compositional correlation among codon positions. Gene 110: 81-88.

D'Onofrio, G., D. Mouchiroud, B. Aissani, C. Gautier, and G. Bernardi. 1991. Correlations between the compositional properties of human genes, codon usage and aminoacid composition of proteins. Journal of Molecular Evolution 32: 504-510.

Duret, L., D. Mouchiroud, and C. Gautier. 1995. Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. Journal of Molecular Evolution 40: 308-317.

Federico, C., S. Saccone, and G. Bernardi. 1998. The gene-richest bands of human chromosomes replicate at the onset of the S-phase. Cytogenetic and Cell Genetic 80: 83-88.

Macaya, G., J. P. Thiery, and G. Bernardi. 1976. An approach to the organization of eukaryotic genomes at a macromolecular level. Journal of Molecular Biology 108: 237-254.

Mouchiroud D., G. D'Onofrio, B. Aissani, G. Macaya, C. Gautier, and G. Bernardi. 1991. The distribution of genes in the human genome. Gene 100: 181-187.

Sabeur, G., G. Macaya, F. Kadi, and G. Bernardi. 1993. The isochore patterns of mammalian genomes and their phylogenetic implications. Journal of Molecular Evolution 37: 93-108.

Saccone, S., A. De Sario, G. Della Valle, and G. Bernardi. 1992. The highest gene concentrations in the human genome are in T-bands of metaphase chromosomes. Proceedings of the National Academy of Sciences of the U.S.A. 89: 4913-4917.

Saccone, S., A. De Sario, J. Wiegant, A. K. Rap, G. Della Valle, and G. Bernardi. 1993. Correlations between isochores and chromosomal bands in the human genome. Proceedings of the National Academy of Sciences of the U.S.A. 90: 11929-11933.

Saccone, S., S. Caccio, J. Kusuda, L. Andreozzi, and G. Bernardi. 1996. Identification of the gene-richest bands in human chromosomes. Gene 174: 85-94.

Tazi, J., and A. Bird. 1991. Alternative chromatin structure at CpG islands. Cell 60: 909-920.

Thiery, J.P., G. Macaya, and G. Bernardi. 1976. An analysis of eukaryotic genomes by density gradient centrifugation. Journal of Molecular Biology 108: 219-235.

Winkler, H. 1920. *Verbreitung und Ursache der Parthenogenesis im Pflanzen- und Tierreich*, Fischer, Jena.

Zoubak, S., G. D'Onofrio, S. Caccio, G. Bernardi, and G. Bernardi. 1995. Specific compositional patterns of synonymous positions in homologous mammalian genes. Journal of Molecular Evolution 40: 293-307.

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