Chem. Listy 90, 549 – 553 (1996)

P02 THE ORGANIZATION OF THE HUMAN GENOME

GIORGIO BERNARDI

Laboratoire de Genetique Moleculaire, Institut Jacques Monod, 2 Place Jussieu, 75005 Paris, France

KEY WORDS: Genome, Isochore, Compositional pattern, Gene distribution, Human genome core

The term *genome* was coined three quarters of a century ago by a German botanist Winkler[1] to designate the haploid chromosome set. While current textbooks of Molecular Biology do not yet go beyond the purely operational definition of the eukaryotic genome as the sum total of genes and of intergenic sequences, a number of molecular biologists have been thinking for some time that the genome is more than the sum of its parts. This implies the existence of structural and functional interactions between the minority of coding sequences and the majority of non-coding sequences. This general and rather vague concept has been changed into a precise one by the discovery, in our laboratory, of specific genome properties. These properties, which have mainly been defined by investigations on the nuclear genome of vertebrates[2-5] will be briefly outlined here. They comprise the isochore organization, the compositional patterns of DNA fragments (i.e. DNA molecules) and of coding sequences, the compositional correlations between coding and non-coding sequences and, above all, the gene distribution and its associated functional properties.

The mammalian genomes are mosaics of **isochores** (see Fig. 1a), namely of long (> 300 Kb) DNA segments that are homogeneous in base composition and range from 30 to 60% GC (ref.[6,7]). This is an extremely wide range, almost as wide as that covered by all bacterial DNAs (25-72% GC). 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 (Fig. 1b).

The **compositional distributions** of large (> 100 Kb) genome fragments, such as those forming routine DNA preparations, of exons (and particularly of their third codon positions) and of introns represent **compositional patterns**[2]. These correspond to genome phenotypes[8], in that they 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 Fig. 2).

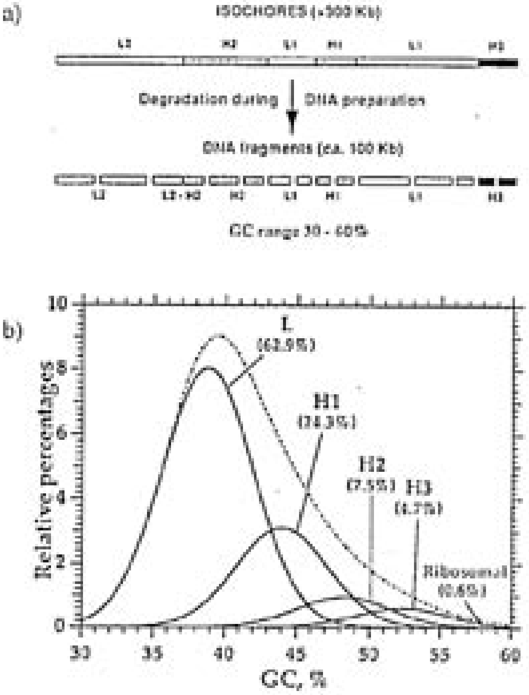

Fig. 1. (a) **Scheme of the isochore organization of the human genome.** This genome, which is a typical mammalian genome, is a mosaic of large (> 300 Kb) DNA segments, the isochores. These are compositionally homogeneous (above a size of 3 Kb) and can be partitioned into a small number of families, 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% (from ref.[4]).
(b) **The isochore families from the human genome.** The relative amounts of major DNA components derived from isochore families L (i.e., L1 + L2), H1, H2, H3 (see ref.[10]) are superimposed on the CsCl profile of human DNA (from ref.[11]).

Compositional correlations[2] exist (Fig. 3) between exons (and their codon positions) and isochores, as well as between exons and introns[9,10]. These correlations concern, therefore, coding and non-coding sequences and are not trivial since coding sequences only make up about 3% of the genome, whereas non-coding sequences correspond to 97% of the genome. The compositional correlations repre-

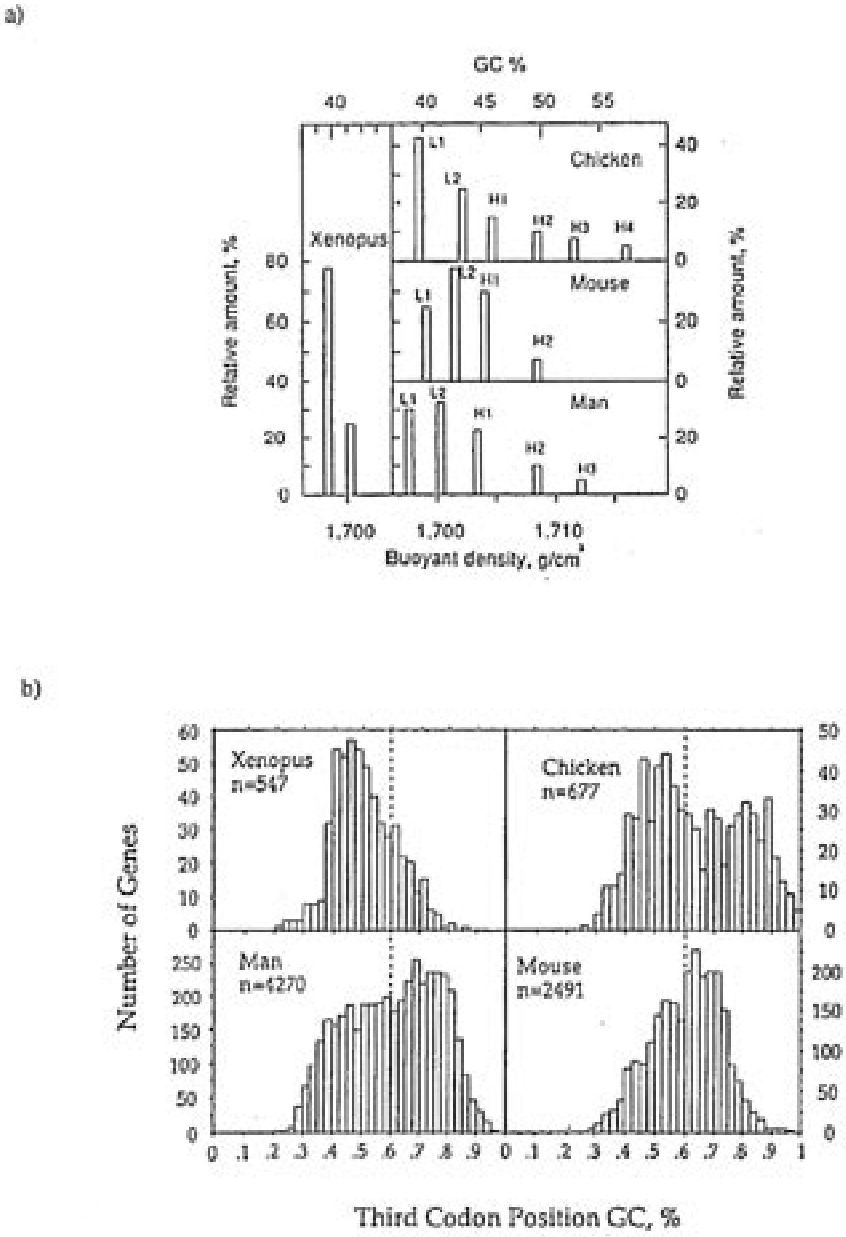

Fig. 2. (a) Compositional patterns of vertebrate genomes. Histograms showing the relative amounts, modal buoyant densities and GC levels of the major DNA components from *Xenopus*, chicken, mouse and man, as estimated after fractionation of DNA by preparative density gradient in the presence of a sequence-specific DNA ligand (Ag+ or BAMD; BAMD is bis (acetato mercuri methyl) dioxane). The major DNA components are the families of large DNA fragments (see Fig. 1) derived from different isochore families. Satellite and minor DNA components (such as rDNA) are not shown in these histograms (from ref.[4]).

(b) Compositional distribution of third codon positions from vertebrate genes. The number of genes taken into account is indicated. A 2.5% GC window was used. The broken line at 60% GC is shown to provide a reference (from ref.[4]).

sent a **genomic code**[4]. It should be noted that a universal correlation holds among GC levels of codon positions (third positions against first and/or second positions). Both the genomic code and the **universal correlation** are apparently due to compositional constraints working in the same direction (towards GC or AT), although to different extents on coding and non-coding sequences, as well as on different codon positions.

The compositional correlations between GC3 (the GC level of third codon positions) and isochore GC have a practical interest in that they allow to position the coding sequence histogram of Fig. 2 relative to the CsCl profile of Fig. 1 and to assess the **gene distribution** in the human genome[5,11,12]. In fact, if one divides the relative number of genes per histogram bar by the corresponding relative amount of DNA, one can see that gene concentration is low in GC-poor isochores, increases with increasing GC in isochore familes H1 and H2, and reaches a maximum in isochore family H3, which exhibits at least a 20-fold higher gene concentration compared to GC-poor isochores (Fig. 4).

The H3 isochore family has been called the **human genome core**[4], 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, as well as by the scarcity of histone H1, the acetylation of histones H3 and H4 and wider nucleosome spacing[13], with the highest transcription and recombination levels and with the earliest replication timing. The genes of the genome core have the highest GC3 levels relative to their flanking sequences, have the shortest exons and introns[14], exhibit an extreme codon usage and encode prote-

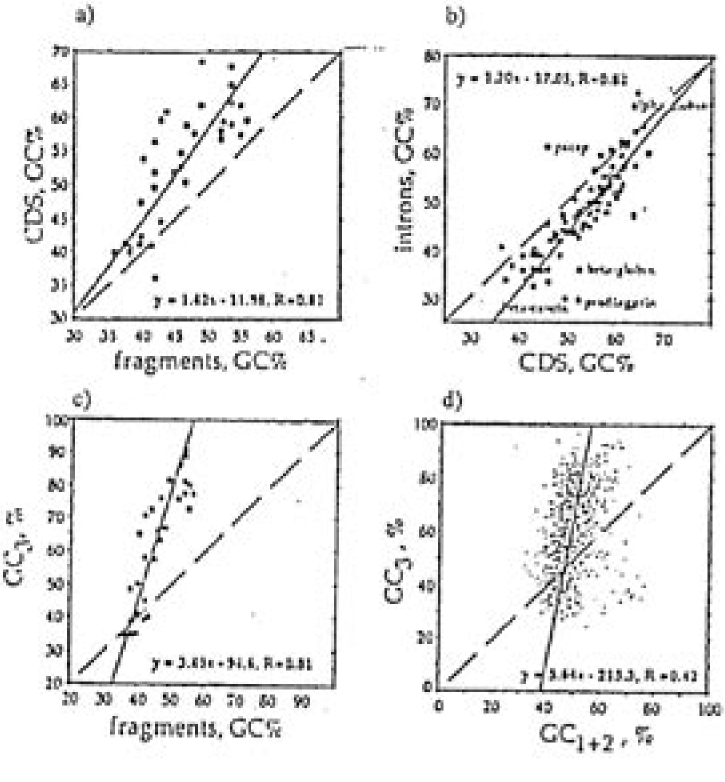

Fig. 3. (a) GC levels of coding sequences (CDS) are plotted against GC levels of the isochores in which they were experimentally localized; (b) intron GC is plotted against the GC levels of the corresponding coding sequences; (c) GC3 is plotted against the GC levels of the isochores containing the corresponding genes; (d) GC3 is plotted against GC1+2. In all plots, orthogonal (solid line) relationships are shown along with the diagonal (slope = 1), the equations, and the correlation coefficient (from ref.[10]).

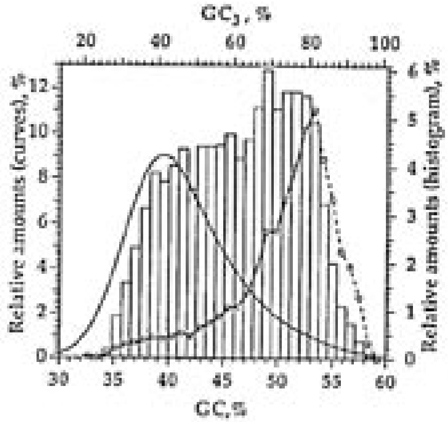

Fig. 4. Profile of gene concentration in the human genome as obtained by dividing the relative amounts of genes in each 2.5% GC interval of the histogram by the corresponding relative amounts of DNA deduced from the CsCl profile. The apparent decrease in gene concentration for very high GC values (broken 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 (from ref.[21]).

ins characterized by amino acid frequencies differing from those of proteins encoded by GC-poor isochores[15].

The human genome core is located in about 30 T(elomeric)-bands[16], which are essentially formed by GC-rich isochores (mainly of the H2 and H3 families). In contrast, R'-bands, namely the R(everse) bands exclusive of T-bands, comprise both GC-rich isochores (of the H1 family) and GC-poor isochores. R' bands, in turn, can be divided (at a 400 band resolution) into about 30 T' bands which contain gene H3 isochores and about 140 R" bands which do not[17]. Finally, G(iemsa) bands are formed almost exclusively by GC-poor isochores[18] (Fig. 5). The difference in GC level between G-bands and T-bands is about 15%. About 20% of genes are present in G-bands and about 80% in R-bands (60% of them in T-bands).

It should be stressed that the gene distribution reported for the human genome seems to have been conserved in evolution, genes showing their highest concentration in the GC-richest isochores of all vertebrates[5].

As already mentioned, 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

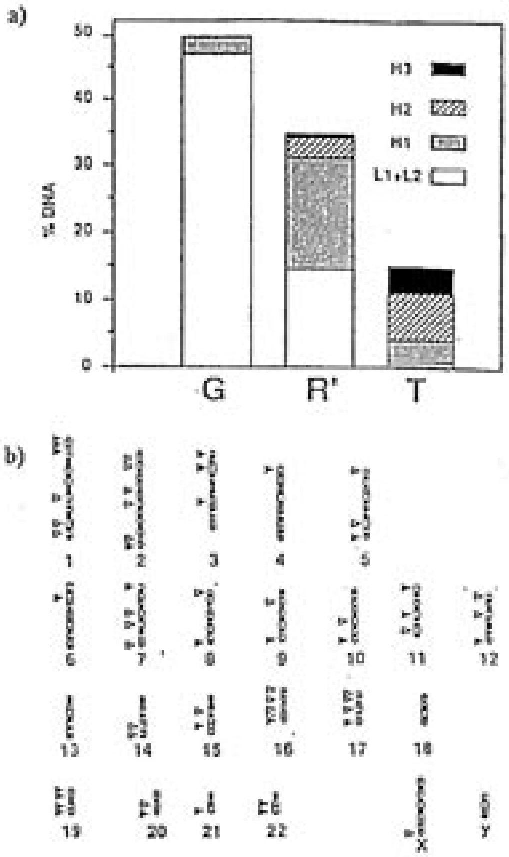

Fig. 5. (a) A scheme of the relative amounts of isochore families L1 + L2, H1, H2 and H3 in G-bands, R'-bands and T-bands; R'-bands are R-bands exclusive of T-bands (from ref.[18]).
(b) Distribution of sequences hybridizing DNA from H3 isochores on human chromsomes. T bands (solid arrow) correspond to strong signals, T' bands (open arrows) to medium signals. The remaining R bands correspond to undetectable signals (from ref.[17]).

different from the compositional patterns of cold-blooded vertebrates which exhibit a much lower degree of heterogeneity and are characterized by metaphase chromosomes which do exhibit a R-banding. These different genome phenotypes of warm- *versus* cold-blooded vertebrates are due to compositional changes. While the gene-poor, GC-poor isochores have undergone little or no compositional change in vertebrates genomes, the gene-rich, GC-rich isochores are those which underwent compositional changes in evolution.

In the case of homologous mammalian genes, it has been possible to show that third codon position synonymous substitutions exhibit frequencies and compositions which strongly suggest natural selection[19,20]. Under these circumstances, the compositional changes in non-coding sequences, which are correlated wiht those occurring in third codon positions, suggest that non-coding sequences are not junk DNA, but must fulfill some functional role.

REFERENCES AND NOTES

1. Winkler H.: *Verbreitung und Ursache der Parthenogenesis im Pflanzen-und Tierreich*. Fischer, Jena 1920.
2. Bernardi G., Olofsson B., Filipski J., Zerial M., Salinas J., Cuny G., Meunier-Rotival M., Rodier F.: Science *228*, 953 (1985).
3. Bernardi G.: Ann. Rev. Genet., *23*, 637 (1989).
4. Bernardi G.: Gene *135*, 57 (1993).
5. Bernardi G.: Annu. Rev. Genet. *29*, 445 (1995).
6. Thiery J.P., Macaya G., Bernardi G.: J. Mol. Biol. *108*, 219 (1976).
7. Macaya G., Thiery J.P., Bernardi G.: J. Mol. Biol. *108*, 237 (1976).
8. Bernardi G., Bernardi G.: J. Mol. Evol. 24, 1 (1986).
9. Aïssani B., D'Onofrio G., Mouchiroud D., Gardiner K., Gautier C., Bernardi G.: J. Mol. Evol. *32*, 497 (1991).
10. Clay O., Cacciò S., Zoubak S., Mouchiroud D., Bernardi G.: Mol. Phylogenet. Evol. *5*, 2 (1996).
11. Mouchiroud D., D'Onofrio G., Aïssani B., Macaya G., Gautier C., Bernardi G.: Gene *100*, 181 (1991).
12. Zoubak S., Clay O., Bernardi G.: In press.
13. Tazi J., Bird A.: Cell *60*, 909 (1991).
14. Duret L., Mouchiroud D., Gautier C.: J. Mol. Evol. *40*, 308 (1995).
15. D'Onofrio G., Mouchiroud D., Aïssani B., Gautier C., Bernardi G.: J. Mol. Evol. *32*, 504 (1991).
16. Saccone S., De Sario A., Della Valle G., Bernardi G.: Proc. Natl. Acad. Sci. USA *89*, 4913 (1992).
17. Sacone, S., Cacciò S., Kusuda J., Andreozzi L., Bernardi, G.: In press.
18. Saccone S., De Sario A., Wiegant J., Rap A.K., Della Valle G., Bernardi G.: Proc. Natl. Acad. Sci USA *90*, 11929 (1993).
19. Cacciò S., Zoubak S., D'Onofrio G., Bernardi G.: J. Mol. Evol. *40*, 280 (1995).
20. Zoubak S., D'Onofrio G., Cacciò S., Bernardi G., Bernardi, G.: J. Mol. Evol. *40*, 293 (1995).