4 The vertebrate genome: isochores and chromosomal bands

G. BERNARDI
Institut Jacques Monod, France

1 Introduction

Vertebrate genomes are mosaics of *isochores*, namely of long, compositionally homogeneous DNA segments which can be subdivided into a small number of families characterized by different GC levels. In the human genome, which is representative of a number of mammalian genomes and, more broadly, of the genomes of warm-blooded vertebrates, the compositional spectrum of isochores ranges between 30% and 60% GC, and five families of isochores have been identified, two GC-poor families, L1 and L2, representing together 62% of the genome, and three GC-rich families, H1, H2 and H3, representing 22%, 9% and 3%, respectively. The remaining 4% of the genome consists of satellite and ribosomal DNAs, which can also be visualized as isochores, because of their homogeneous base composition (Bernardi, 1989).

2 Compositional patterns and compositional correlations

The *compositional distribution* of large (*ca.* 100 Kb) DNA fragments (such as those forming current DNA preparations) represents a *compositional pattern* that reflects the *isochore pattern*. Other compositional patterns are represented by the compositional distributions of exons (and of their codon positions) and of introns.
These compositional patterns characterize *genome phenotypes* (Bernardi and Bernardi, 1986), which are very different in cold- and warm-blooded vertebrates. The main differences are that the former are much less heterogeneous in composition, are generally GC-poorer, and never attain very high GC levels compared to the latter (Bernardi and Bernardi, 1990a,b; 1991). Smaller compositional differences exist among the genomes of either cold-blooded or warm-blooded vertebrates.

Chromosomes Today Volume 11. Edited by A.T. Sumner and A.C. Chandley. Published in 1993 by Chapman & Hall, London. ISBN 0 412 47670 3

Compositional correlations hold between exons (and their codon positions) and the isochores in which they are embedded, as well as between exons and the corresponding introns (Bernardi et al., 1985; 1988; Bernardi and Bernardi, 1985; 1986; Bernardi, 1989; Aïssani et al., 1991; D'Onofrio et al., 1991; Mouchiroud et al., 1991). These compositional correlations link, in a linear fashion, the coding sequences and the non-coding sequences which surround them, or are contained in them.

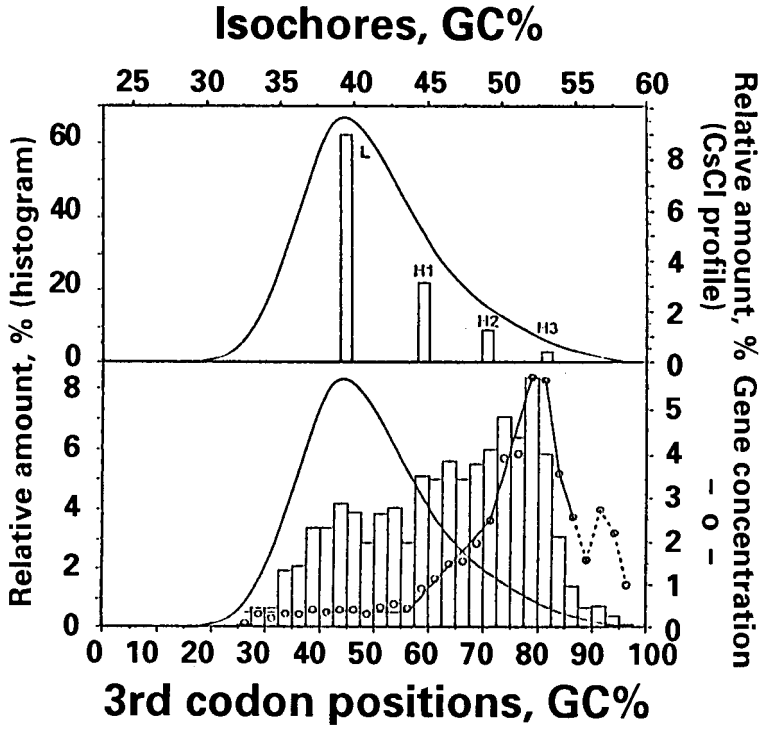

Fig. 1. Top. A histogram of relative amounts of isochore families (or "major DNA components") L (L1 + L2) H1, H2, H3 from the human genome. The upper scale concerns the GC levels of the isochores, as well as those of a CsCl profile of human DNA.

Bottom. Histogram of relative amounts of human genes divided in classes according to GC levels of third codon positions. The profile of gene concentration in the human genome is also shown. The profile was obtained by dividing the relative amounts of genes in each 2.5% GC interval of the histogram by the corresponding relative amounts of DNA, as deduced from the CsCl profile. (Modified from Mouchiroud et al., 1991).

3 The gene distribution in the human genome

The compositional correlation which links GC levels of third codon positions of human genes with the GC levels of the extended sequences in which the genes are located can be used in order to assess gene distribution in the different isochore families and to quantify the finding (Bernardi et al., 1985) that gene distribution in the human genome is strikingly non-uniform. This approach (Fig. 1) has shown that, while 34% of all genes currently present in gene banks are contained in isochore families L1 and L2, 38% are contained in H1 and H2, and 28% in H3 (Mouchiroud et al., 1991). If the gene sample used is representative of all human genes, and if account is taken of the different relative amounts of isochore families (see Introduction), gene concentration in H3 would be 16 times higher than in L1+L2 and 8 times higher than in H1+H2. These ratios are, however, probably underestimated, because housekeeping genes, which are likely to be more abundant in H3 than in other isochore families (see next section), are currently very much under-represented in gene banks (Mouchiroud et al., 1991).

The results just described indicate that increasing gene concentrations are accompanied by increasing GC levels in the genome of warm-blooded vertebrates; the evolutionary process underlying this phenomenon will be discussed later. Very interestingly, the gradient of gene concentration is paralleled by a series of changes in a number of properties which have functional significance. This will be illustrated by describing the extreme situation found in the GC-richest isochore family H3.

4 The human genome core

The isochore family H3 corresponds to a genome compartment endowed with very remarkable properties for which the name *genome core* is proposed (see also next section). This family has not only the highest GC level and the highest gene concentrations, but also the highest concentrations of CpG doublets (Bernardi, 1985), the only potential sites of methylation in vertebrates, and the highest concentrations of CpG islands (Aïssani and Bernardi, 1991a,b), which are very GC-rich sequences characterized by abundant, unmethylated CpG doublets (Bird, 1986). Since CpG islands, which are located in the 5'flanking sequences of genes, are preferentially associated with housekeeping genes (Gardiner-Garden and Frommer, 1987), the latter should be more abundant in H3 than in the other isochore families.

The coding sequences of the H3 isochores are much higher in GC level than their genomic environment, compared with those from other isochore families, especially from GC-poor isochores (Aïssani et al., 1991). Moreover, these genes and

their associated CpG islands are characterized by a particular chromatin structure, with nucleosome-free regions, absence or scarcity of histone H1, and acetylation of histones H3 and H4 (Tazi and Bird, 1990; see also Aïssani and Bernardi, 1991a,b). These properties make these chromosomal regions more "open", as also indicated by their sensitivity to nuclease attack (Kerem et al., 1984).

The H3 isochore family presumably has the highest level of transcription because of its very high concentration of genes, and especially of housekeeping genes. It also has the highest recombination rate, possibly because of its "open" chromatin structure and to the abundance of repetitive sequences, like Alu sequences and minisatellites. The very high recombination rate of H3 isochores may also be largely responsible for the much higher rate of karyotypic rearrangements (and speciation) shown by warm-blooded vertebrates compared to cold-blooded vertebrates (Bernardi, 1992). Indications exist that the H3 isochores may be the main integration regions for the majority of (GC-rich) retroviral sequences (see Rynditch et al., 1991; Zoubak et al., 1992).

The H3 isochore family has an extremely biased codon usage, a number of codons being absent or very scarce because of the very high GC levels in third codon positions, and an extreme amino acid utilization, which favors aminoacids corresponding to codons having only G and/or C in the first two codon positions (D'Onofrio et al., 1991), namely arginine (quartet codons), alanine, glycine and proline, rather than those corresponding to codons with only A and/or T in those positions, like lysine, or those corresponding to codons having both G/C and A/T in first and second codon positions, like serine. Finally, the location of the H3 family of isochores in metaphase chromosomes will be discussed in section 6.

5 The evolutionary origin of the genome core

In terms of base composition, the genome of warm-blooded vertebrates appears to comprise a *paleogenome*, characterized by GC-poor isochores which have not changed in composition relative to the corresponding isochores of cold-blooded vertebrates, and a *neogenome*, characterized by isochores which have become GC-rich (Bernardi, 1989; see Fig. 2). Very interestingly, the GC increase in the genome of warm-blooded vertebrates concerns only a minority of the genomes, about one third of it, which contains, however, the majority of the genes, at least two thirds of them. Indeed, as already mentioned, GC increases parallel gene concentration (Fig. 1).

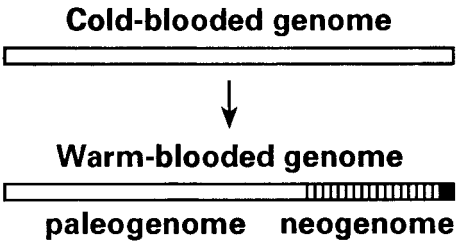

Fig. 2. Scheme of the compositional genome transition accompanying the emergence of warm-blooded from cold-blooded vertebrates. The compositionally homogeneous, GC-poor genomes of cold-blooded vertebrates are changed into the compositionally heterogeneous genomes of warm-blooded vertebrates. The latter comprise a paleogenome (corresponding to about two thirds of the genome) which did not undergo any large compositional change and a neogenome (corresponding to the remaining one third of the genome, with the GC-richest part only representing 3% of the genome). In the scheme, the mosaic structure of the warm-blooded vertebrate genome is neglected; GC-poor isochores (open bar), GC-rich isochores (hatched bar) and GC-richest isochores (black bar) are represented as three contiguous regions. Gene concentration increases from GC-poor to GC-rich to GC-richest isochores (From Bernardi, 1992).

The main reason for proposing the name of *genome core* for the GC-richest isochore family of the human genome is that the strikingly non-uniform gene distribution observed for the human genome and, in particular, the existence of isochores with very high gene concentrations appear to be shared by all warm-blooded vertebrates and, very probably, by all vertebrates.

Indeed, several data point to the fact that the formation of GC-rich and GC-richest isochores is due to events (consisting in the fixation of biased mutations) superimposed on a gene concentration pattern already present in cold-blooded vertebrates. First of all, in spite of significant, yet relatively minor differences, the genomes of mammals are very similar in their compositional organization. This was shown by the analysis of DNA (Sabeur et al., paper in preparation) which indicated fundamentally similar isochore patterns, and by the analysis of homologous coding sequences (Mouchiroud et al., paper in preparation) which showed very similar GC levels in all codon positions and particularly in third codon positions. If one recalls that compositional correlations exist between coding sequences (and their codon positions) and the large DNA fragments in which they are located, the two facts mentioned above indicate that the compositional patterns of mammals,

and, therefore, the corresponding gene concentration patterns, are at least largely conserved. The comparison of mammalian (human) and avian (chicken) genomes and of their homologous coding sequences also indicate a large similarity although, expectedly, a lesser one. This similarity suggests, in turn, that the reptiles from which both classes of warm-blooded vertebrates derived, also showed comparable gene concentration patterns. The same conclusion could also be reached on another ground, namely that, if such was not the case, a tremendous reshuffling of gene distribution should have occurred at the transition between reptiles and warm-blooded vertebrates (Bernardi and Bernardi, 1990b).

Second, in some genomes from cold-blooded vertebrates a certain compositional heterogeneity exists. In these cases, a plot of GC levels of third codon positions of genes against GC levels of the corresponding genomes, as derived from modal buoyant densities, shows a deviation towards lower values (Bernardi and Bernardi, 1991). This is consistent with the fact that these genes are mainly present in the GC-richest compartments of the corresponding genomes. In other words, this finding suggests that the gene concentration pattern is already similar to that of warm-blooded vertebrates, even if the actual GC levels are quite different, compositional differences being modest in cold-blooded vertebrates.

Third, preliminary experiments (paper in preparation) have shown that the H3 isochore family from man cross-hybridizes with the GC-richest fractions not only from other mammals, but also from cold-blooded vertebrates.

An implication of the evolutionary conservation of the gene concentration pattern and especially of the genome core, is that the increase in genome size from fishes to mammals (polyploidy being neglected here) should largely involve an expansion of intergenic sequences located in gene-poor regions, but not of those located in the genome core. Indeed, if such expansion involved the genome core, it would strongly decrease its gene concentration.

6 Isochores and chromosomal bands

A number of findings indicate that GC-poor isochores are located in G(iemsa) bands, whereas GC-rich isochores are located in R(everse) bands of human metaphase chromosomes (see Bernardi, 1989, and Bickmore and Sumner, 1989, for reviews; see also Medrano et al. 1988, and Schmid and Guttenbach, 1988). In the latter case, at least, the correspondence cannot be a direct one, for the simple reason that GC-rich and GC-poor isochores are in a 1:2 ratio, whereas R- and G-bands are in a 1:1 ratio (Gardiner et al., 1990). This may mean (i) that standard R-bands comprise more "thin" G-bands (only detected at high resolution) than

standard G-bands comprise "thin" R-bands (as suggested by Ikemura and Aota, 1988); and/or (ii) that DNA compaction is higher in G-bands than in R-bands. Moreover, while GC-poor isochores, which are present in G-bands, differ very little from each other in composition and represent 62% of the genome, GC-rich isochores, which are located in R-bands, encompass a wide GC range and represent only 22%, 9% and 3%, respectively , of the genome (Bernardi et al., 1985; Bernardi, 1989). This situation should lead by itself to inter and/or intra-band compositional heterogeneity in R-bands.

An approach developed in order to solve the problem of the correlations between isochores and chromosomal bands is *compositional mapping* (Bernardi, 1989). This consists in hybridizing probes corresponding to landmarks that are localized on a physical map, or on a chromosome band map, to compositional DNA fractions. This approach allows to assess the GC level of about 200 Kb around the landmarks

Compositional mapping, as applied to the long arm of human chromosome 21 (Gardiner et al., 1990), showed that practically all probes for loci present in G-bands hybridized to GC-poor isochores, whereas probes located in R-bands hybridized to either GC-poor or GC-rich isochores. In other words, G-bands are GC-poor and at least largely homogeneous in base composition, whereas R-bands are compositionally heterogeneous. The GC-richest region of the long arm of the human chromosome 21, hybridizing on isochore family H3, was shown to correspond to the telomeric band, which is a T-band (Dutrillaux, 1973), namely one of the 20 or so most heat-denaturation-resistant R-bands (Fig. 3).

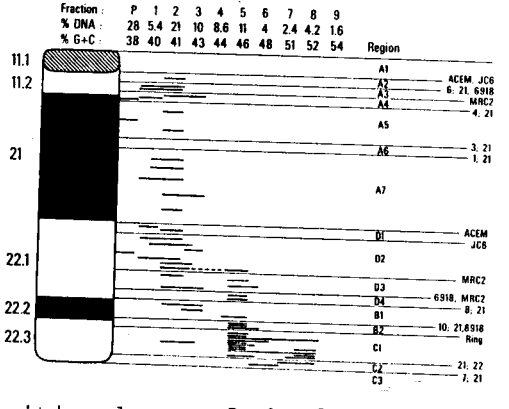

Fig. 3. Compositional map of the long arm of human chromosome 21. Long horizontal lines indicate positions of the breakpoints associated with the rearranged chromosomes listed at the right of the figure. Short horizontal lines indicate the compositional fractions to which probes located in different chromosomal regions hybridized. (From Gardiner et al., 1990).

The meaning of T-bands was not clear for twenty years. Indeed, the high resistance to temperature denaturation of these bands could be due to high GC levels in the corresponding DNA, or to particular chromatin structures or to other reasons. The work of Ambros and Sumner (1987) showed that T-bands are characterized by DNA having high GC levels. While this result pointed to the involvement of DNA, it still did not say, however, whether these GC-rich regions corresponded to repeated or single-copy DNA.

Recent investigations suggested a preferential location of very GC-rich genes in T-bands (Bernardi, 1989; Gardiner et al., 1990; Ikemura and Wada, 1991; De Sario et al., 1991). Moreover, while the telomeric repeats common to all chromosomes hybridized on isochore families H1, H2 and H3, telomere probes corresponding to T-bands hybridized on the GC-richest isochore family, H3, whereas telomeric probes corresponding to R-bands hybridized on GC-rich families H1 and H2 (De Sario et al., 1991). A clear solution of the problem was obtained, however, when *in situ* suppression hybridization on human metaphase chromosomes of the single-copy sequences from the H3 isochore family showed (Fig. 4) that these sequences were precisely located at T-bands (Saccone et al., 1992).

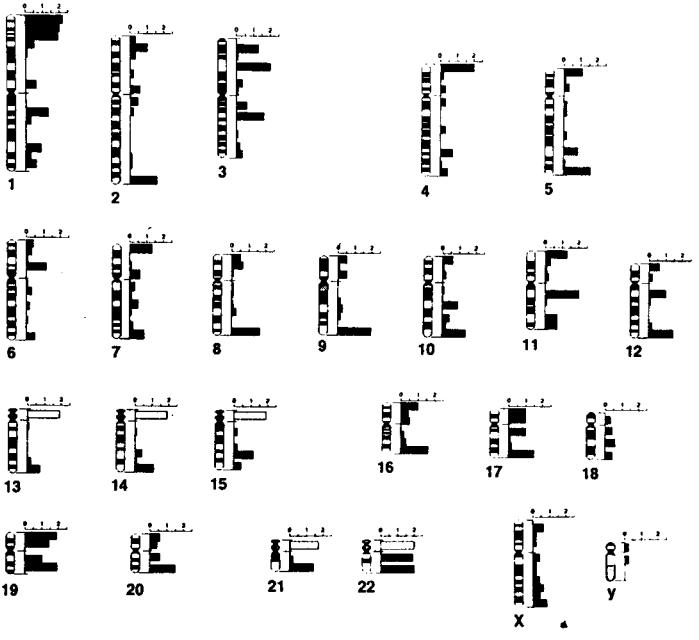

Fig. 4. Histogram (solid bars) showing the distribution of sequences hybridizing with the H3 DNA fraction on human G-banded chromosomes. Open bars correspond to rDNA. Scales are percentage of total number of signals. (From Saccone et al., 1992).

It should be noted that high gene concentrations in some specific T-bands had been occasionally reported. For instance, most of the small nuclear DNA genes, a family of 30-125 genes per haploid genome, were localized at 1p36.3 (Naylor et al., 1984) which is a T-band. Along the same line, one could quote the telomeric localization of ribosomal genes on the short arms of chromosomes 13,14,15,21 and 22. These bands could, in fact, be considered as a sub-class of T-bands. It should also be noted that evidence exists at least for some of the intercalary T-bands to be the result of telomere fusion. For instance, the intercalary T-band 11q13 is the result of a pericentric inversion juxtaposing the telomeric region to the centromere (Dutrillaux, 1979).

Very recent compositional mapping results on the Xq26-Xqter region (Pilia et al., 1992, submitted) confirmed the GC-poorness and compositional homogeneity of G-bands, as well as the heterogeneity of R-bands, and revealed isochores of the H3 family even in chromosomal regions, like the telomeric Xq28 band, which did not correspond to T-bands. This might be due to the fact that the GC-richest region of Xq is narrower than usual T-bands. It is, therefore, possible that a number of T-bands are still escaping cytogenetic detection.

If several T-bands have not yet been detected, the high concentration of genes in telomeres indicated by *in situ* hybridization (Saccone et al., 1992) becomes an even more conspicuous property of the vertebrate genome, whose implications certainly deserve further investigations. Indeed, human telomeres are tightly associated with the nuclear matrix, via their TTAGGG repeats forming the terminal 2-30 Kb of chromosomes (De Lange, 1992), as well as with the nuclear envelope (Henderson and Larson, 1991).

In some cases, both telomeric and intercalary T-bands make up a large proportion of chromosomes. This is true of chromosome 20,21 and, more so, of chromosomes 19 and 22 (in the case of chromosomes 21 and 22, the T-bands on the short arm are ribosomal bands; see above). This raises two questions. The first one is whether these high concentrations of H3 isochores lead to overall increased GC contents of those chromosomes. This appears to be true since, as shown by Korenberg and Engels (1978), these four chromosomes are the richest in GC. The second question, is whether a high gene density can be detected in chromosomes which have high densities of H3 isochores. This also appears to be the case since gene densities are the highest for chromosomes 11,17,19,22 and X (McKusick, 1991) all of which are very high in H3 (except for X in which the high gene density is, in all likelihood, the result of investigations concentrating on this chromosome).

7 References

Aissani,B. and Bernardi, G. (1991a) CpG islands : features and distribution in the genome of vertebrates. **Gene**, 106, 173-183.

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

Aissani, B. D'Onofrio,G. Mouchiroud, D. Gardiner, K. Gautier, C. and Bernarid, G. (1991) The compositional properties of human genes. **J. Mol. Evol.**, 32,497-503.

Ambros, P.F. and Sumner, A.T. (1987) Correlation of pachytene chromomeres and metaphase bands of human chromosomes, anddistinctive properties of telomeric regions. **Cytogen. Cell Genet.**, 44,223-238.

Bernardi, G. (1985) The organization of the vertebrate genome and the problem of the CpG shortage. in **Chemistry, Biochemistry and Biology of DNA Methylation** (eds G.L.Cantoni and A. Razin), Alan Liss, New York, N.Y., pp. 3-10.

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.

Bernardi, G. and Bernardi, G. (1985) Codon usage and genome composition. **J. Mol. Evol.**, 22,363-365.

Bernardi, G. and Bernardi, G. (1986) Compositional constraints and genome evolution. **J. Mol. Evol.**, 24,1-11.

Bernardi, G. Mouchiroud, D. Gautier, C. and Bernardi, G. (1988) Compositional patterns in vertebrate genomes : conservation and change in evolution. **J. Mol. Evol.**, 28,7-18.

Bernardi, G. (1989) The isochore organization of the human genome. **Ann. Rev. Genet.**, 23,637-661.

Bernardi, G. and Bernardi G. (1990) Compositional patterns in the nuclear genomes of cold-blooded vertebrates. **J. Mol. Evol.**, 31,265-281.

Bernardi, G. and Bernardi, G. (1990b) Compositional transitions in the nuclear genomes of cold-blooded vertebrates. **J. Mol. Evol.**, 31,282-293.

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

Bernardi, G. (1992) Genome organization and species formation in vertebrates. **J. Mol. Evol.**, in press.

Bickmore, W.A. and Sumner A.T.(1989) Mammalian chromosome banding - an expression of genome organization. **TIG**, 5,144-148.

Bird, A. (1986) CpG-rich islands and the function of DNA methylation. **Nature**, 321,209-203.

De Lange, T. (1992) Human telomores are attached to the nuclear matrix. **EMBO J.**, 11,717-724.

De Sario, A. Aissani, B. and Bernardi, G. (1991) Compositional properties of telomeric regions from human chromosomes. **FEBS Letters**, 295,22-26.

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

Dutrillaux, B. (1973) Nouveau système de marquage chromosomique : les bandes T. **Chromosoma**, 41,395-402.

Dutrillaux,B.(1979) Chromosomal evolution in primates : tentative phylogeny from *Microcebus murinus* (Prosimian) to man. **Hum. Genet.**, 48,251-314.

Gardiner,K. Aissani, B. and Bernardi, G. (1990b) A compositional map of human chromosome 21. **EMBO J.**, 9,1853-1858.

Gardiner-Garden, M. and Frommer, M. (1987) CpG islands in vertebtate genomes. **J. Mol. Biol.**, 196,261-282.

Henderson, E.R. and Larson, D.D.(1991) Telomeres - what's new at the end? **Curr. Opin. Genet. & Develop.**, 1,538-543.

Ikemura, T.and 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. **Nucl. Acids Res.**, 13,1915-1922.

Kerem, B.-S. Goiten,R. Diamond,G. Cedar, H. and Marcus, M. (1984) Mapping of DNAase I sensitive regions of mitotic chromosomes. **Cell**, 38,493-499.

Korenberg, J.R. and Engels, W.R.(1978) Base ratio, DNA content, and quinacrine-brightness of human chromosomes. **Proc. Natl. Acad. Sci. USA**, 75,3382-3386.

Medrano,L. Bernardi,G. Couturier,J. Dutrillaux,B. and Bernardi, G. (1988) Chromosome banding and genome compartmentalization in fishes. **Chromosoma**, 96,178-183.

McKusick, V.A. (1991) Genomic mapping and how it has progressed. **FASEB J.**, 26,50-64.

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

Naylor, S.L. Zabel,B.U. Manser,T. Gesteland, R. and Sakaguchi, A.Y. (1984) Localization of human U1 small nuclear RNA genes to band p36.3 of chromosome 1 by *in situ* hybridization. **Somat.Cell & Mol.Genet.**, 10,307-313.

Pilia, G. Little,R.D. Aissani,B. Bernardi, G. and Schlessinger, D.(1992) Isochores and CpG islands in YAC contigs covering the q26.1-qter region of the human X chromosome. Submitted for publication.

Rynditch, A. Kadi,F. Geryk,J. Zoubak,S. Svoboda, J. and Bernardi, G. (1991) The isopycnic, compartmentalized integration of Rous sarcoma virus sequences. **Gene**, 106,165-172.

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

Schmid, M. and Guttenbach, M. (1988) Evolutionary diversity of reverse (R) fluorescent chromosome bands in vertebrates. **Chromosoma**, 97,101-114.

Tazi, J. and Bird, A.P. (1990) Alternative chromatin structure at CpG islands. **Cell**, 60,909-920.

Zoubak, S. Rynditch, A. and Bernardi, G. (1992) Compositional bimodality and evolution of retroviral genomes. **Gene**, in press.