Annu. Rev. Genetics. 1995. 29:445–76

THE HUMAN GENOME: Organization and Evolutionary History

Giorgio Bernardi
Laboratoire de Genétique Moléculaire, Institut Jacques Monod, 2 Place Jussieu, 75005 Paris, France

KEY WORDS: chromosomes, evolution, genome, isochores, vertebrates

ABSTRACT

This review outlines briefly the compositional properties of the vertebrate genome, namely its isochore organization, the compositional patterns of DNA molecules and of coding sequences, the compositional correlations between coding and noncoding sequences, and the relationships between isochores and chromosomal bands. It then deals with the fundamental properties of the vertebrate genome, namely the distribution of genes and its associated functional features. Finally, it considers how the structural and functional organization of the human genome (and of the genomes of warm-blooded vertebrates in general) arose in evolution.

CONTENTS

INTRODUCTION

The term genome was coined three quarters of a century ago by a German botanist, Hans Winkler (113), to designate the haploid chromosome set of

eukaryotes. Surprisingly, in current textbooks of molecular biology, this term, now become so popular because of the human genome project, is either not even defined or defined (often only tacitly) in a purely operational way, namely as the sum total of genes and intergenic sequences of a haploid cell. The genome is, however, more than the sum of its parts, essentially because structural, functional, and evolutionary interactions occur among different regions of the genome and, more specifically, between coding and noncoding sequences.

This review deals with the advances made on the intertwined issues of genome organization, function, and evolution in vertebrates since a previous article in this series (9). Because of space limitations, the emphasis will not be on details but rather on the coherent and comprehensive picture emerging from an approach based on molecular genetics and molecular evolution. This picture fully contradicts the view that comprehensive rules about the organization of genomes have not emerged, notwithstanding the availability of a large amount of details, or the view that biology is nothing more than detail upon detail with no universals or laws emergent. [1]

GENOME ORGANIZATION

The Compositional Properties of the Vertebrate Genome

ISOCHORES The genomes of vertebrates are mosaics of isochores (20), which were originally identified (73) as long (> 300) KbDNA segments homogeneous in base composition (above a 3-Kb level). This was shown in the most straightforward way by the analysis of long sequences from data banks (62, 64) and by compositional mapping (22, 55, 71a, 92; A De Sario et al, in preparation). Isochores can be divided into several families characterized by different base compositions, as initially shown by density gradient fractionation of DNA molecules that derive from isochores through the unavoidable enzymatic and mechanical breakage occurring during DNA preparation (Figure 1*a*). In the human genome, which is typical for the genomes of most mammals, "light", GC-poor (L1 and L2) isochore families represent 62% of the genome, whereas "heavy", GC-rich (H1 and H2) and very GC-rich (H3) isochore families correspond to 22, 9, and 3–4% of the genome, respectively (Figure 1*b*). The

[1]Abbreviations: BAMD, 3,6-bis (acetato mercuri methyl) 1-4 dioxane; bp, base pair; BP, before present; DAPI, 4′,6-diamino-2-phenylindole; CDS, coding sequence(s); GC, molar ratio of guanine + cytosine in DNA; G bands, Giemsa bands; GC, GC_1, GC_2, GC_3, GC level in first, second, and third codon positions; HPLC, high performance liquid chromatography; Mb, megabase(s); My, million years; Kb, kilobase(s); pg, picogram(s); R bands, reverse bands; R′ bands, reverse bands exclusive of T bands; R" bands, R′ bands exclusive of T′ bands; T bands, telomeric bands; T′ bands, R′ bands containing H2 and H3 isochores; YAC(s), yeast artificial chromosomes.

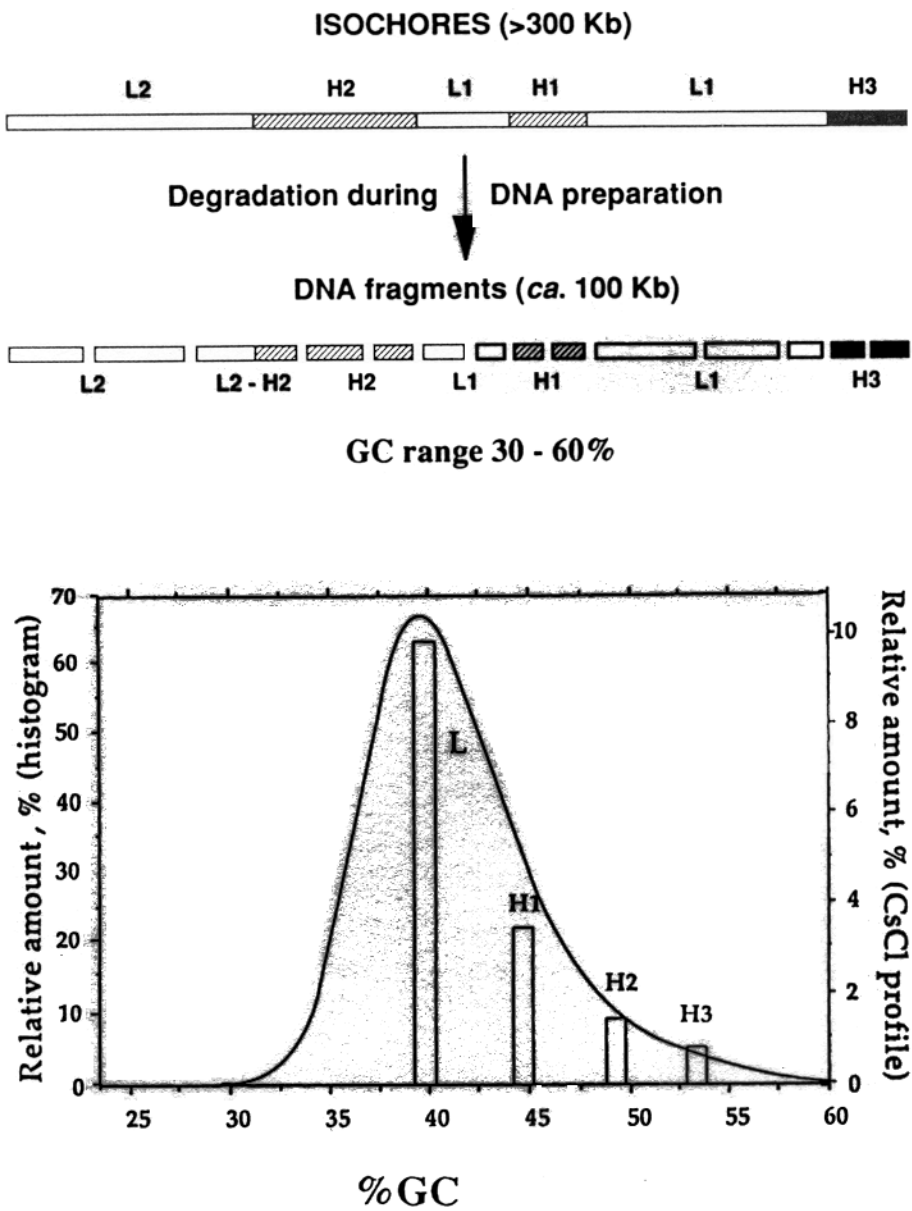


Figure 1 (*a*) Scheme of the isochore organization of the human genome. This genome, which is typical of the genome of most mammals, is a mosaic of large (>300 Kb on average) DNA segments, the isochores, which are compositionally homogeneous (above a size of 3 Kb) and can be divided into a small number of families, "light" or GC-poor (L1 and L2), and "heavy" or GC-rich (H1 and H2), and very GC-rich (H3). Isochores are degraded during DNA preparation to DNA fragments (~100 Kb in size). The GC-range of the isochores from the human genome is 30–60%. (Modified from Reference 20.)

(*b*) Histogram of the isochore families from the human genome. The relative amounts of major DNA components, the families of DNA fragments derived from isochore families L (i.e. L1+L2), H1, H2, H3 are superimposed on the CsCl profile of human DNA. Modal GC levels are positioned on the abscissa axis. Satellite and ribosomal DNA are not represented in either (*a*) or (*b*). (Modified from Reference 83.)

remaining 3–4% consists of satellite and ribosomal DNAs, which can also be viewed as isochores because of their homogeneous base composition.

COMPOSITIONAL PATTERNS AND GENOME PHENOTYPES The compositional patterns, namely, the compositional distributions (*a*) of large (~100 Kb)

genome fragments, such as those forming standard DNA preparations; (*b*) of coding sequences; (*c*) of each of the three codon positions (and particularly of third codon positions); and (*d*) of introns (20, 84) represent genome phenotypes (14), in that they differ characteristically not only between cold- and warm-blooded vertebrates, but also between mammals and birds, and even among mammals (Figures 2*a, b*).

The compositional patterns of DNAs from cold-blooded vertebrates are substantially different from those of warm-blooded vertebrates (15–17, 61, 93, 111). Indeed, the former are characterized by lower intermolecular compositional heterogeneities and CsCl band asymmetries, as well as by the fact that their buoyant densities do not reach the high values found in the GC-richest DNA components from warm-blooded vertebrates (see Figure 7, below).

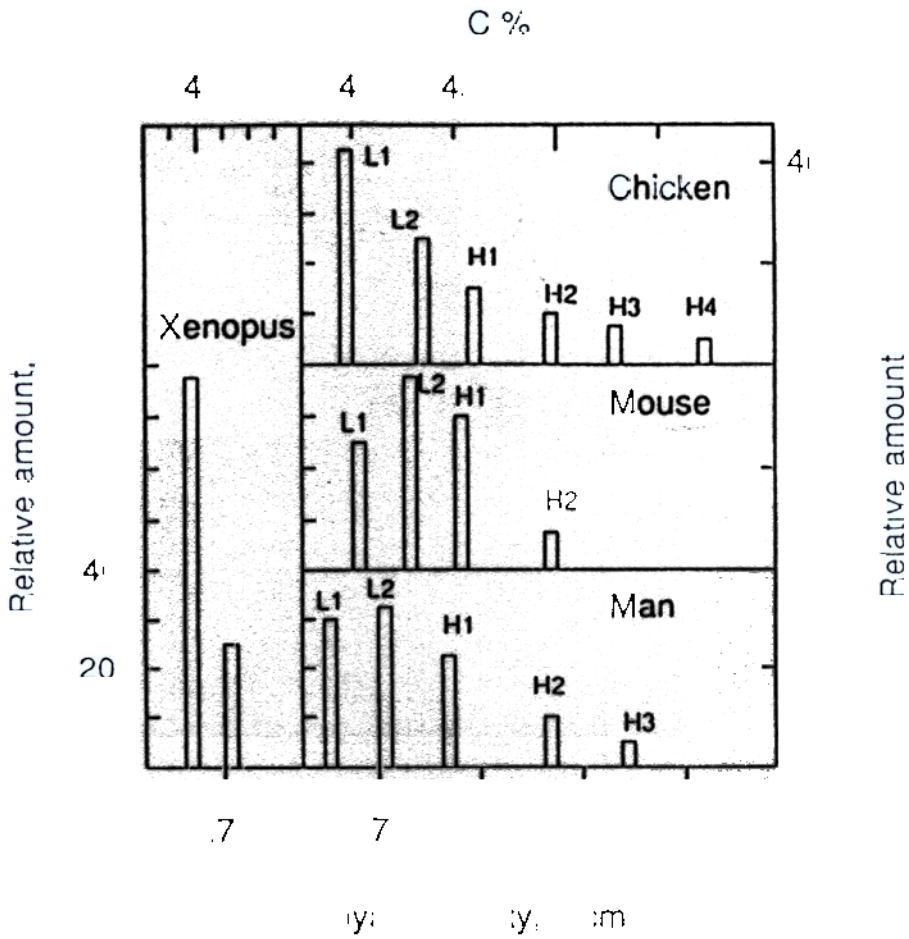
In mammalian genomes, a general pattern characterized by large amounts (6–10%) of very GC-rich isochores (defined here as DNA higher than 1.710 g/cm^3 in buoyant density) was found in species belonging to eight of the nine mammalian orders explored (97). Special patterns, different from each other, but all lacking the GC-richest isochores of the general pattern, were found in some infraorders or families of the orders Rodents (Myomorphs), Chiropters (Megachiropters), and Insectivores (Soricids), as well as in Pangolin, a species from the only genus of the order Pholidota (26, 85, 97, 102). Interestingly, other infraorders and families from the same orders (Sciuromorphs and Histricomorphs, as well as Microchiropters and other families of Insectivores) share the general pattern.

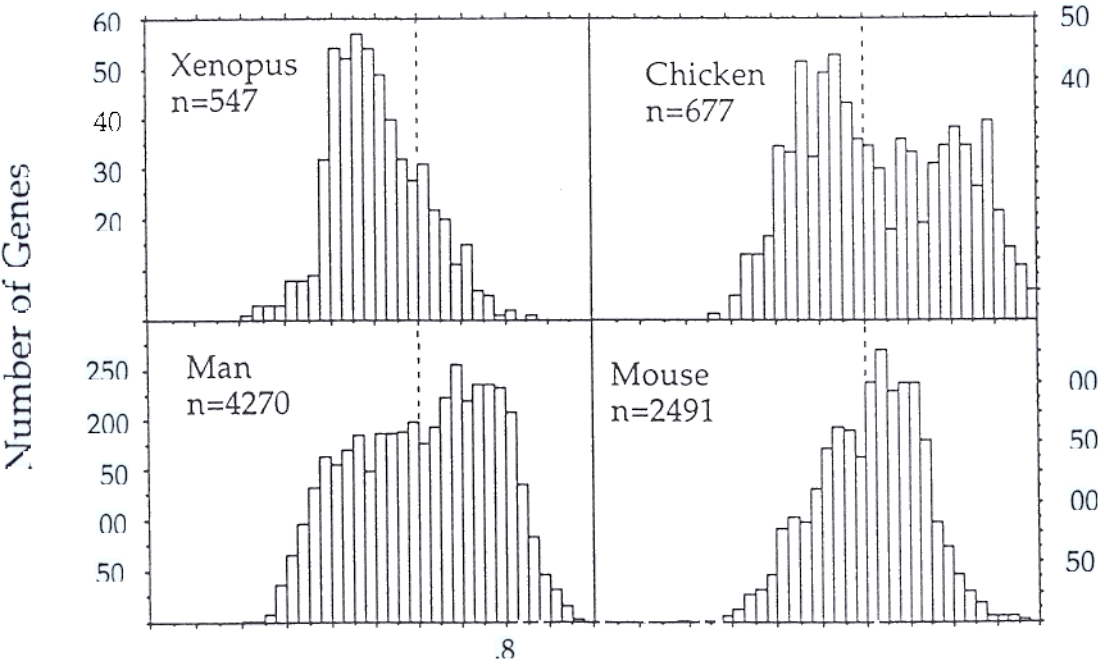
In contrast to mammals, the compositional patterns of DNAs from species belonging to eight avian orders (including both paleognathous and neognathous species) were extremely close to each other and were characterized by compositional distributions basically similar to those of the general mammalian pattern; there were, however, significantly larger amounts (9–13%) of GC-rich isochores (66).

In the compositional patterns of third codon positions (see 82, 84), large differences were found between the only cold-blooded vertebrate that could be studied, *Xenopus*, and warm-blooded vertebrates (see Figure 2*b*). Indeed,

Figure 2 (*a*) Compositional patterns of vertebrate genomes. Histograms showing the relative amounts, modal buoyant densities and modal GC levels of the major DNA components (the families of DNA fragments derived from different isochore families; see Figure 1*a*) 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). Satellite and minor DNA components (such as ribosomal DNA) are not shown. (Modified from Reference 20.)

(*b*) Compositional distribution of third codon positions from genes of *Xenopus*, chicken, man, and mouse. A 2.5% GC window was used. The broken line at 60% GC is shown to provide a reference. The number of genes, *n*, taken into account is indicated. This number is still small for *Xenopus* and chicken. (Updated from Reference 12.)





Third Codon Position GC

Xenopus showed a narrower distribution centered on a lower GC_3 level (GC_3 is defined as the GC level of third codon positions). Moreover, differences were found between chicken, the only avian species that could be tested, and mammals. In fact, coding sequences from chicken showed a distribution reaching GC_3 values as high as 100%. Finally, differences were found between the coding sequence distributions, for example, of man or calf, which exhibit the general DNA pattern, and those of mouse or rat, which are narrower (82, 85). The differences shown in Figure 2 are not due to differences in gene samples since they were also found in histograms of homologous genes. Compositional distributions of introns (not shown here because of their small sample sizes) parallel those of exons (see following section).

COMPOSITIONAL CORRELATIONS AND THE GENOMIC CODE The similarities of compositional patterns, as seen at the DNA and at the coding sequences levels, suggest that they are correlated with each other. Indeed, compositional correlations exist between coding sequences (and their codon positions) and isochores or flanking sequences, as well as between exons and introns from the same genes (3, 7, 13, 14, 20, 31, 61a; see Figure 3*a–g*). These correlations

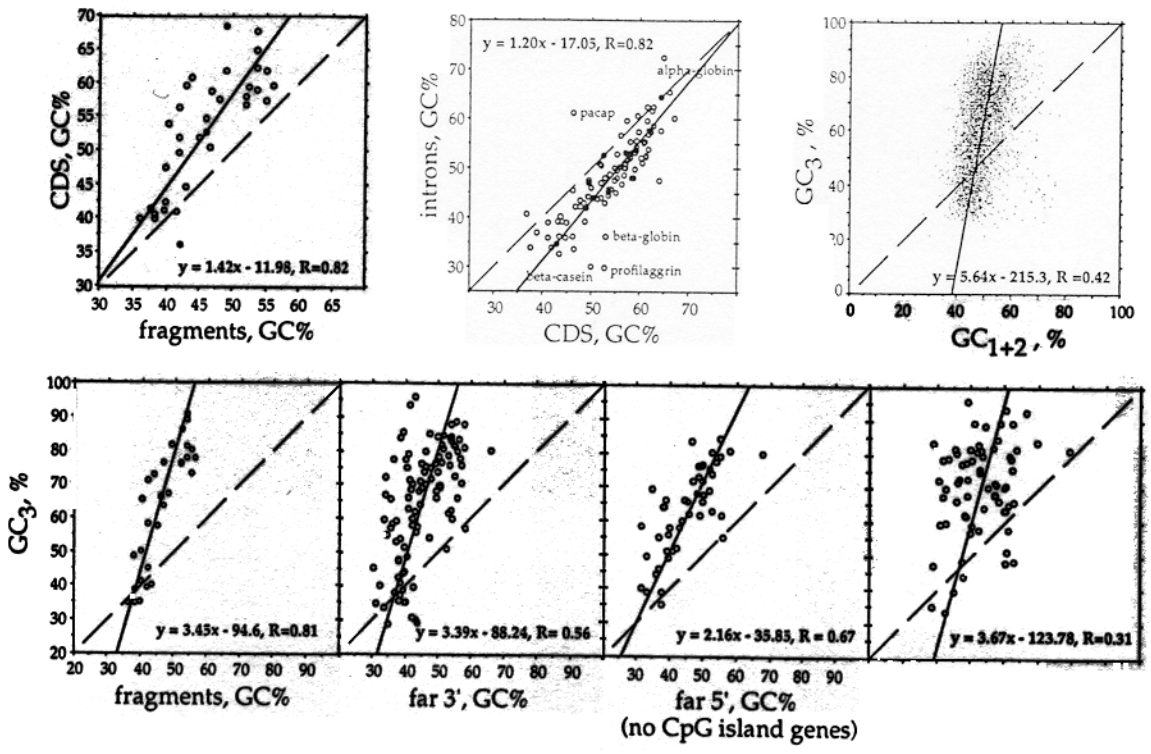


Figure 3 (*a*) GC levels of coding sequences (CDS) are plotted against GC levels of the isochores in which they were experimentally localized; (*b*) Intron GC is plotted against the GC levels of the corresponding coding sequences; (*c*) GC_3 is plotted against GC_{1+2}, (*d*) (*e*) (*f*) (*g*) GC_3 is plotted against the GC levels of the isochores containing the corresponding genes, of the far 3′ flanks (>500 bp from termination codon) and of the far 5′ flanks (>500 bp from initiation codon) from genes without and within CpG islands. In all plots, orthogonal (solid line) relationships are shown along with the diagonal (slope = 1), the equations, and the correlation coefficient. (From Reference 31.)

concern coding and noncoding sequences, because isochores are essentially made of noncoding sequences, and introns obviously are also noncoding sequences. The correlations, in fact, define genome equations (see Figure 3) that represent a genomic code (12).

The correlation of Figure 3*a* between coding sequences and the isochores embedding them (*a*) is not trivial because coding sequences only make up about 3% of the human genome, whereas noncoding sequences correspond to the remaining 97%; (*b*) shows 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*) implies 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 Figure 3*b* shows that introns increase in GC in parallel with the corresponding exons, but are systematically lower by about 5% GC, although this difference seems to decrease for GC-rich genes. The correlation of Figure 3*d* between GC_3 and GC levels of the isochores embedding the corresponding genes allows the study of the distribution of genes in the genome (see the following section). The correlations of Figures 3*e*, 3*f*, and 3*g* show that GC_3 is well correlated with 3′ flanks (in which case the slope is close to that of Figure 3*d*) and with 5′ flanks (in the latter case, however, only for genes that are not associated with CpG islands).

Finally, Figure 3*c* 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, a universal correlation that also holds for all eukaryotic and prokaryotic genes (17, 42).

Gene Distribution in Isochores

The first few genes localized in compositional DNA fractions provided a preliminary indication that genes are not distributed randomly in the human genome, as most genes tested were present in the scarce GC-richest isochores (20). Subsequent investigations with increasing numbers of localized genes (3, 114) pointed in the same direction. The compositional correlations between GC_3 and isochore GC (Figure 3*d*) or GC of 3′ far flanks (Figure 3*e*) allowed positioning the human coding sequence histogram relative to the CsCl profile (12, 83). The relative number of genes per histogram bar was then divided by the corresponding relative amount of DNA to provide a gene concentration profile. This profile indicates that 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 up

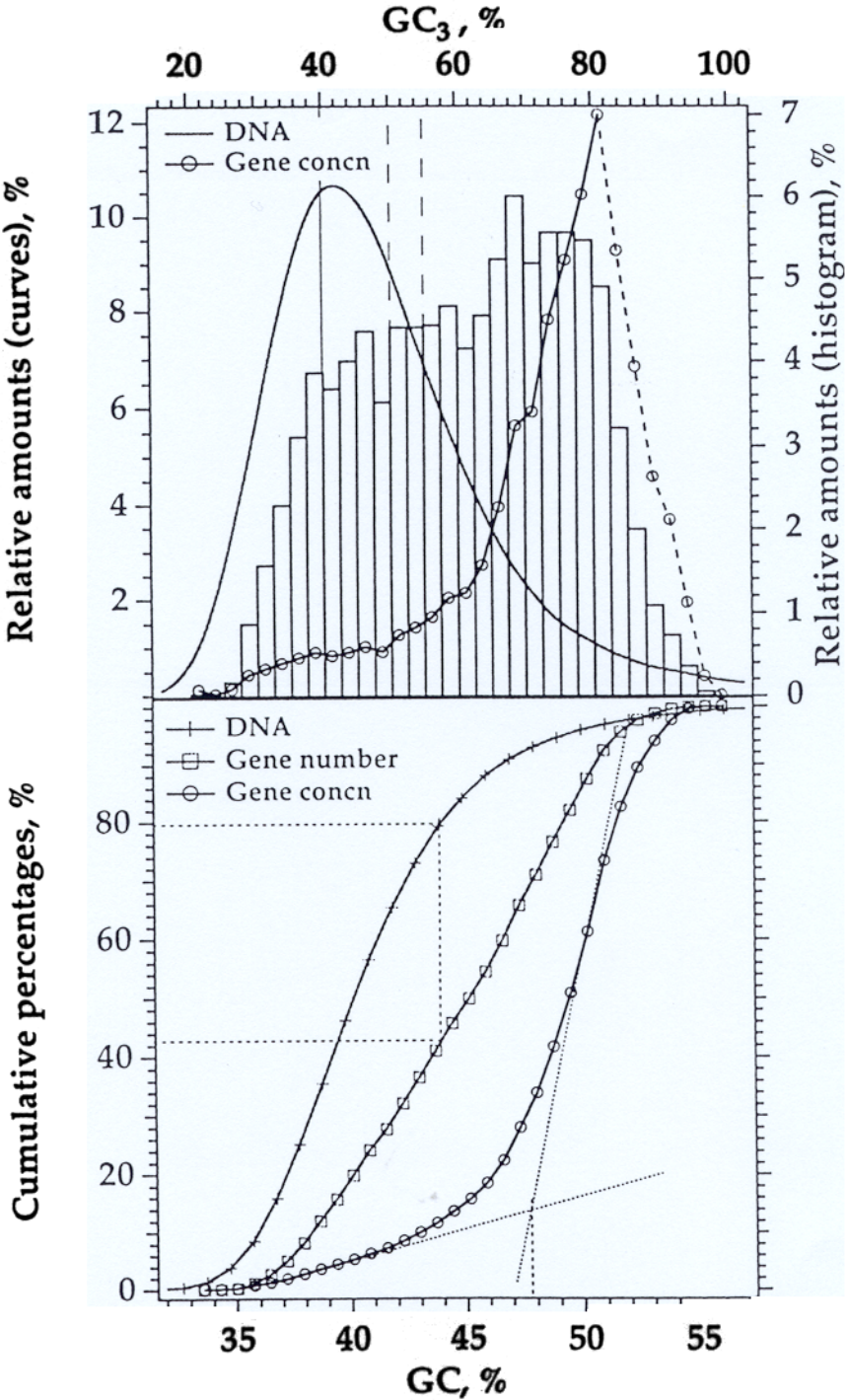
to a 20-fold higher gene concentration compared to GC-poor isochores (Figure 4*a*). Incidentally, this value is an underestimate, because of the presence of ribosomal DNA in the buoyant density range of H3 isochores (see Figure 9), and because housekeeping genes, which seem to be preferentially located in H3 isochores (see a following section), are currently underrepresented in gene banks.

If the cumulative percentage of DNA is plotted against GC levels of isochores, a sigmoid curve is obtained (Figure 4*b*), as expected from the bell shape of the CsCl profile (Figure 4*a*). If the cumulative percentage of genes is plotted against GC_3 or isochore GC, an essentially linear relationship is observed (Figure 4*b*). If gene concentration is calculated as the ratio of the two cumulative plots, two distinct slopes are found, a shallow one corresponding to the GC-poor isochores and a steep one corresponding to GC-rich isochores. These results (*a*) raise the problem of the reason(s) accounting for the correlation between GC level and gene concentration (see the sections on Genome Evolution); (*b*) indicate that compositional mapping (see the following section) is, in fact, a mapping of gene concentration; and (*c*) allow estimation of both the relative amounts of genes in different isochore families (see below) and the gene coverage by physical mapping (see a following section).

If GC-poor isochores make up 62% of the genome, GC-rich isochores 31%, and very GC-rich isochores 3–4%, the plots of Figure 4*b* of the H3 family lead to relative gene/DNA ratios of 0.45, 1.48, and 7.4. In other words, the GC-richest isochores have a gene concentration 7 times higher than GC-rich isochores and 16 times higher than the GC-poor isochores, in good agreement with previous conclusions (83). The H3 isochore family has been called the "genome core" (11); this definition should, however, be extended to also comprise the H2 isochore family, which is systematically clustered with it in chromosomes (see a following section).

Figure 4 (*a*) 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 of Figure 2*b* by the corresponding relative amounts of DNA deduced from the CsCl profile. The apparent decrease in the concentration of protein-encoding genes for very high GC values (broken line) is due to the presence of ribosomal DNA in that region. The last concentration values are uncertain because they correspond to very low amounts of DNA. (From Reference 117). The vertical line at 39% GC is the frontier below which no integrated HTLV-I sequences are found. The two vertical broken lines indicate the GC range of the host genome in which both transcribed and nontranscribed HTLV-I sequences were found. Only nontranscribed sequences were found below that range, only transcribed sequence above.

(*b*) Cumulative plot of relative amounts of DNA and genes as a function of GC of DNA or GC_3 of coding sequences. The ratio of the two curves is also plotted together with its two slopes. The plot also shows that if the GC-richest 20% of the genome is not mapped, only 42% of the genes are located in mapped regions. (From S Zoubak et al, in preparation.)


GC3, %
20
40
60
80
100
DNA
Gene concn
Relative amounts (curves), %
Relative amounts (histogram), %
DNA
Gene number
Gene concn
Cumulative percentages, %
GC, %
35
40
45
50
55

Compositional Mapping

An experimental approach, compositional mapping, which is, in fact, a gene concentration mapping (see preceding section), has provided detailed information on isochore sizes and borders, as well as on the correlations between isochores and chromosomal bands. Originally, compositional maps were constructed by hybridizing probes for landmarks (i.e. genes or single-copy anonymous sequences) localized on physical maps with compositional DNA fractions. Compositional maps allowed assessing of GC levels over 100–200 Kb around the landmarks (9). Another approach involved the analysis of the base composition of yeast artificial chromosomes (YACs) from physically mapped regions of the genome. Either YACs were simply isolated and analyzed (71a), or yeast clones were lysed and YACs were separated in shallow preparative gradients of CsCl, detected by hybridization with an appropriate probe and assessed for composition by comparison with internal buoyant density DNA markers (40). The latter procedure avoids YAC purification and can, therefore, be applied to large numbers of YACs.

In the human genome, compositional mapping has so far been applied to the long arm of chromosome 21 (55), to the cystic fibrosis locus on chromosome 7 (71a), to the dystrophin locus on the short arm of the X chromosome (22), as well as to the q26-q28 region (92) and to the q28 band of the X chromosome (A De Sario et al, in preparation). The essentially complete compositional map of this band (Figure 5) showed (*a*) that it comprises three regions formed mainly, in the proximal to distal direction, by H1 isochores, H2+H3 isochores and L isochores, respectively; (*b*) that GC-rich and very GC-rich regions are all separated, preceded, and followed by GC-poor isochores; and (*c*) that unstable YACs and unclonable DNA (as indicated by gaps in the physical map) are concentrated in the middle, H2–H3, region.

ISOCHORE SIZES AND BORDERS Isochore sizes, as determined by the compositional mapping results available so far, range from 0.2 to 1.3 Mb, in agreement with the original estimate of an average size of over 0.3 Mb (73). The isochore border between a L and a H2 isochore was recently sequenced and shown (54) to correspond to a sharp compositional discontinuity occurring between a 20 Kb LINE (L1 sequence) cluster associated with a 650-bp sequence homologous to the PAB1 motif described at the boundary of the pseudoautosomal region (47) on the GC-poor side and a 30 Kb SINE (Alu I) cluster on the GC-rich side (LINES and SINES are two families of long GC-poor and short GC-rich interspersed repeated sequences, respectively; 108). This very important result has provided the first direct evidence for the sharp compositional discontinuities among isochores predicted from fractionation studies (53, 111) and from the multimodal profile of mammalian DNA

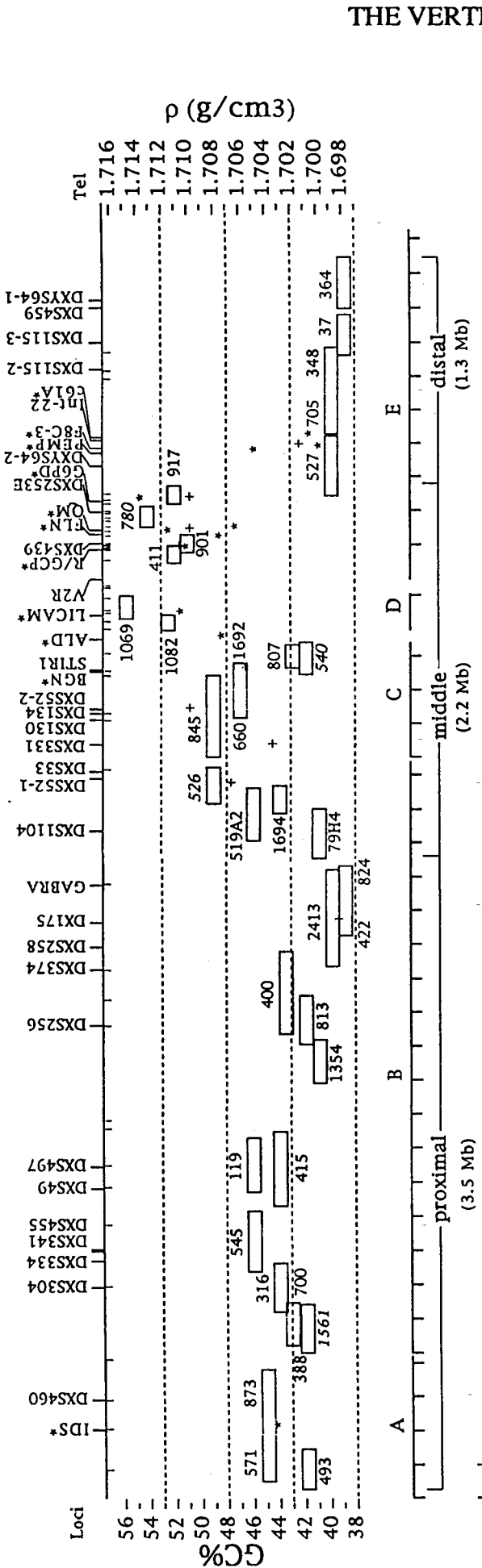


Figure 5 Compositional map of human chromosome band Xq28. YACs are shown as boxes positioned along the physical map on the horizontal scale and according to their buoyant density (right) and GC level (left) on the vertical scale. GC levels of isochores calculated from mapped genes or assessed by hybridization of probes on fractionated DNA (92) are represented by asterisks and crosses, respectively. Mapped genes and loci are also shown on the top of the Figure along with CpG islands, which are represented by small vertical bars. *A* to *E* indicate YAC contigs. The proximal, middle and distal regions are characterized by the predominance of H1, H2+H3, and L isochores, respectively (A De Sario et al, in preparation).

larger than 300 Kb in size (73). Over 100 pseudoautosomal-like sequences have been detected in the human genome (T Ikemura, personal communication). Two additional pseudoautosomal-associated isochore boundaries have very recently been found on human chromosome Y (112a) and 20 (T Ikemura, personal communication), respectively.

Correlations between Isochores and Chromosomal Bands

Compositional mapping may also be applied at the chromosomal level. In situ hybridization of DNAs from human compositional fractions (98, 99) showed the following results (Figure 6*a* and Table 1).

1. G bands, the Giemsa-positive bands, are formed almost exclusively by GC-poor isochores, with only a minor contribution of GC-rich isochores from the H1 family (99). The conclusion that G bands are made up of GC-poor DNA was already suggested by the earliest experiments in which G bands were stained with quinacrine, an AT-specific fluorochrome (see 33). The compositional uniformity of G bands is, however, a novel finding [also supported by all compositional mapping results on isochores from G bands (22, 55, 71a, 92)], which cannot be deduced from classical cytogenetic data.
2. T bands are formed mainly by isochores of the H2 and H3 families (99). Indeed, hybridization of DNA derived from the human H3 isochore family (98) produced a pattern of strong signals (Figure 6*a*) corresponding to two largely coincident subsets of R bands previously described, namely (*a*) the Telomeric bands (46), and (*b*) the chromomycin A-3 positive, DAPI-negative bands (4). The former are elusive bands that correspond to the most heat denaturation–resistant R bands about half of which are located in telomeric positions. The latter correspond to the GC-richest bands of the human karyotype. In neither case, however, was it investigated whether such bands corresponded to satellite or nonsatellite ("main band") sequences. The serious possibility was established by the hybridization results obtained with compositional DNA fractions (98, 99). Hybridization of DNA from the H2 isochore family was also predominantly found in T bands, although this family also made a lesser contribution to R′ bands (99; see also below).
3. R′ bands, definied as R bands exclusive of T bands, comprise both GC-rich isochores (mainly of the H1 family) and GC-poor isochores in almost equal amounts. The compositional heterogeneity of R′ bands was not only shown by the in situ hybridization experiments under discussion, but also by compositional mapping (55, 92; A De Sario et al, manuscript in preparation) and GC_3 levels of genes localized in R′ bands (39, 63). In turn, R′ bands

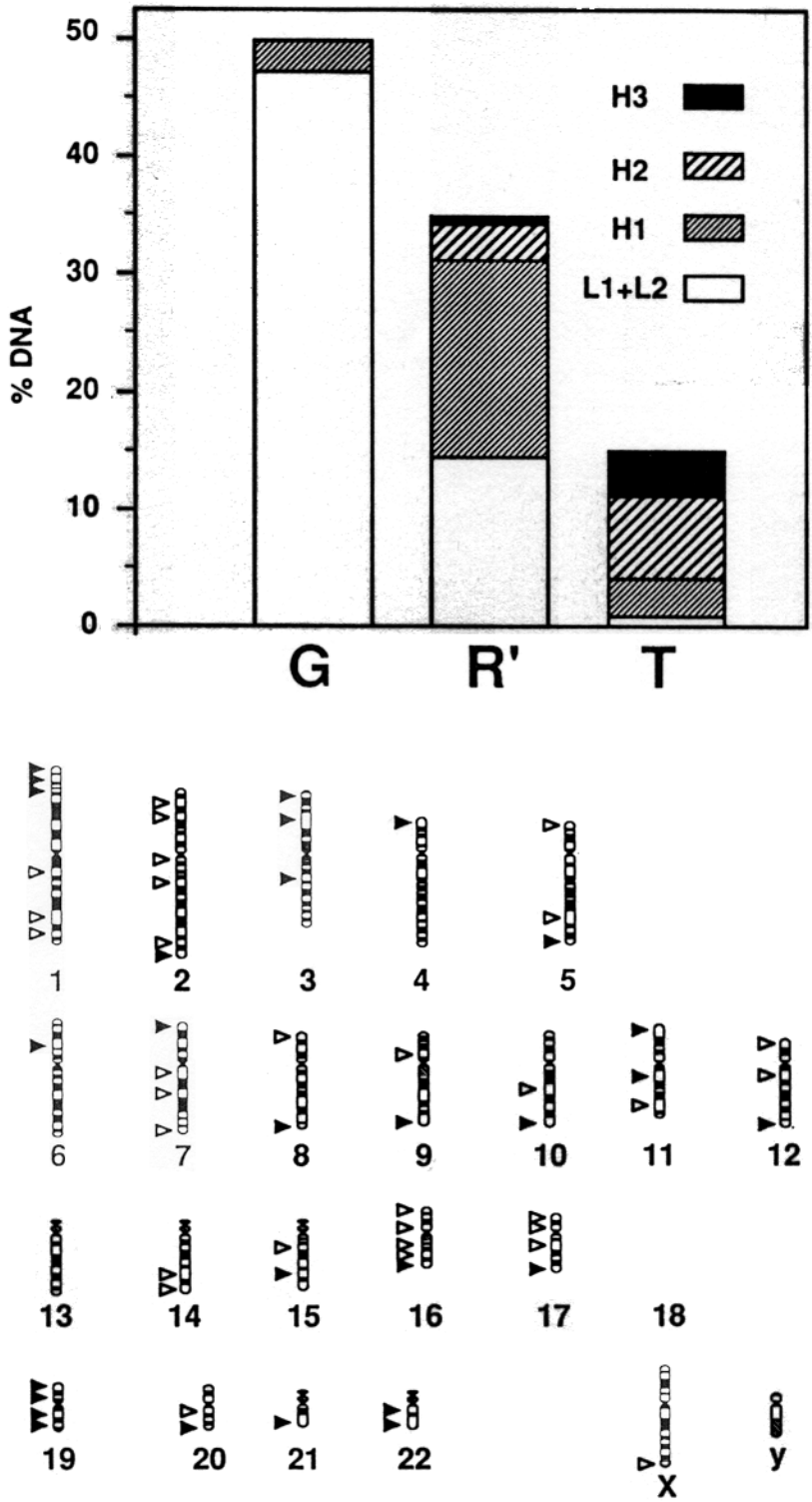


Figure 6 (*a*) A scheme of the relative amounts of isochore families L1 + L2, H1, H2, and H3 in the chromosomal bands of the human karyotype. R′ bands are R bands exclusive of T bands. (From Reference 99.)

(*b*) Distribution of sequences hybridizing DNA from H3 isochores on human chromosomes. $H3^+$ (T) bands (solid arrows) correspond to strong signals, H3 (T′) bands (open arrows) to medium signals. The remaining R bands correspond to $H3^-$ (R″) bands. (From Reference 98; S Saccone et al, manuscript in preparation.)

Table 1 Chromosomal Bands

	Isochores[a]	Gene concentration
R bands		
T-bands (**H3$^+$**)	**H3** + **H2** + *H1* + *L*	++++
R'-bands:		
T'-bands (**H3**)	H3 + H2 + H1 + L	+++
R''-bands (**H3$^-$**	H1 + L	++
G bands (**L**)	*H1* + **L**	+

[a] Bold-type indicates the predominant isochore family(ies), italics the minority isochore family(ies).

can be subdivided in two subclasses, provisionally called T′ bands, which contain clusters of H3 (and H2) isochores, and R″ bands, which do not (S Saccone et al, manuscript in preparation). In T′ bands, hybridization of DNA from H3 isochores (99) detected, next to about 28 strong signals corresponding to T, a set of about 31 medium signal bands due to the presence of H3 isochores in T′ bands (98; S Saccone et al, manuscript in preparation). The compositional map of Xq28 (A De Sario, E-M Geigl, M D'Urso, G Bernardi, manuscript in preparation; Figure 5) provided a precise example of a T′ band by revealing the existence of a 1.5-Mb region of clustered H3 and H2 isochores.

BASE COMPOSITION AND CHROMOSOMAL BANDS The results just presented indicate that R bands comprise at least three sets of bands, T, T′, and R″, which have a different average base composition. This is indicated not only by the isochores from different families that are present in the different sets of bands, but also by the fact that GC_3 values of genes located in those sets are remarkably different, ranging from values identical to those of G bands (for R″ bands) to increasingly higher values for genes located in T′ and T bands, respectively (S Saccone et al, manuscript in preparation). This leads to the important conclusion that G and R bandings are not due to differences in base composition because different subsets of R bands have different compositions, the lowest GC_3 levels found in R bands being identical to those found in G bands (this does not, however, imply an identical distribution of L and H1 isochores in those bands). An alternative explanation is that G bands are different from R bands in that they are compositionally homogeneous, endowed with a closed chromatin structure and, possibly (9), with a higher DNA packing. This explanation is compatible with a recent model (100) in which the basis of the banding patterns is the differential folding of the AT-rich scaffold and packing of DNA loops in G and R bands.

THE CLASSIFICATION OF CHROMOSOMAL BANDS The G bands and three sets of R bands, T, T′, and R″ bands discussed above could be more appropriately called $H3^+$, H3 and $H3^-$ bands, respectively. This proposed classification (see Table 1) has the advantage of being (*a*) based on the hybridization of compositional DNA fractions on metaphase chromosomes, as well as on compositional mapping data (which will become increasingly available in the future), and (*b*) directly related to the gene concentration of bands. The alternative classification of chromosomal bands in five "metaphase chromatin flavors" (60), corresponding to G bands, GC-poor R bands (Alu-poor or Alu-rich), and GC-rich R bands or T bands (again Alu-poor or Alu-rich) has a major weakness: it goes beyond the old distinction between R bands and their subset of T bands (46) only in also taking Alu concentration into account. However, this parameter is of limited significance, in view of its much higher level in the human genome compared to other mammalian (and even primate) genomes.

GENES AND TELOMERES When telomeric bands were tested by hybridization of the telomeric tandem repeat, TTAGGG, on compositional DNA fractions, in order to determine the isochore families forming the 100 Kb or so of DNA contiguous to the telomeric repeats that form the terminal 2–30 Kb of chromosomes (see 69), they were shown to correspond to isochore families H1, H2, and H3 (39). In agreement with this result, only four telomeric bands (on chromosome arms 3p, 18p, 19q, and Yp) correspond to (small) G bands. The other telomeric bands correspond to 16 $H3^+$ (T) bands, 8 H3 (T′) bands, 5 ribosomal bands, and 14 $H3^-$ (R″) bands. Thus, of 48 telomeric bands, 29 are very rich in genes, 14 moderately rich, and 4 are poor in genes, and one, the Yq telomere, which consists of satellite DNA, is lacking genes altogether.

The high concentration of genes in telomeric regions is a conspicuous property not only of the human genome, but also of other vertebrate genomes (P Perani et al, manuscript in preparation) and of genomes from other higher eukaryotes (see, for example, 29). This is of interest because telomeres are tightly associated with the nuclear envelope and with the nuclear matrix (37), which has recently been visualized as a system of internal nuclear channels (94). Under these circumstances, while transcription and splicing take place at chromosome domain boundaries (116), the production of a majority of mRNAs is located near the nuclear pores through which they are exported into the cytoplasm for translation.

GENE DISTRIBUTION AND THE HUMAN GENOME PROJECT The most comprehensive physical map of the human genome (32) covers only 80% of it. The regions not yet mapped practically coincide with T bands (which represent about 15% of the human genome) and with other regions largely corresponding to T′ bands (see above). In other words, the map lacks the GC-richest 20% of

the genome. Figure 4*b* indicates that, under these circumstances, only about 40% of human genes are located in physically mapped regions. This is a very optimistic view, however, because it neglects YAC rearrangements and deletions. Note that telomeric regions are more sparsely covered with markers in genetic maps (112) and that YACS from those regions are unstable, a property associated with compositional heterogeneity (A De Sario et al, in preparation). This problem is, however, being solved by using microsatellite markers from H2 and H3 isochores for genetic mapping (112).

Structural and Functional Properties Associated with Gene Distribution

CHROMATIN STRUCTURE AND TRANSCRIPTION The wide spectrum of gene concentrations in different isochore families is accompanied by distinct structural and functional properties. First, genes from GC-rich isochores (and more so their third codon positions) tend to stand out from their chromosomal compositional environment because of their higher GC levels, unlike genes from GC-poor isochores (Figures 3*a,d*).

Second, increasing gene concentration across isochore families of increasing GC is paralleled by an increasing concentration of CpG islands (1,2). Originally detected as "Hpa II tiny fragments" or HTF, due to the abundance of Hpa II sites, CCGG (see 23), CpG islands are regions about 1 Kb in size, and are characterized by high (65%) GC levels, by clustered, unmethylated CpGs (which are are approaching statistical frequencies, in contrast with their characteristic "shortage" elsewhere in the genome), by G/C boxes (GGGGCGGGGC and closely related sequences that can bind transcription factor Sp1), and by clustered sites for rare-cutting restriction enzymes that recognize GC-rich sequences containing one or two unmethylated CpG doublets. The above description applies to warm-blooded vertebrates, where CpG islands occupy mainly the 5′ flanks of genes and increasingly larger percentages of gene size as gene GC levels increase, until they cover whole genes in H3 isochores (1, 2), which stand out of their compositional context (see above). Expectedly, CpG islands increase in relative amounts in isochore families of increasing GC levels as genes do (1, 2). Since CpG islands are associated with all housekeeping genes and widely expressed genes, but only with 40% of tissue-specific genes and genes with a limited expression (72), housekeeping genes are probably especially abundant in H3 isochores. The very high concentration of genes, particularly housekeeping genes, in H3 isochores points to a very high level of transcription in the "genome core."

Third, different isochore families are associated with different chromatin structures (see 34). Chromatin structure is "open" in the GC-rich isochores of the H2 and H3 families, which are predominant in T bands and are represented

in T' bands, but not so in the GC-poor isochores that form G bands. The "open" chromatin structure is characterized by accessibility to DNases (67), as well as by the scarcity or absence of histone H1, by the acetylation of histones H3 and H4, and by a wider nucleosome spacing (110).

CODON AND AMINO ACID USAGE Genes located in H3 isochores exhibit an extreme codon usage (43). At 100% GC_3, a value approached by genes located in H3 (see Figure 2*b*), only 50% of codons can be present in any given coding sequence. Moreover, because GC_3 is correlated with GC_1 and GC_2 (Figure 3*c*), these genes encode proteins characterized by amino acid frequencies different from those encoded by GC-poor isochores (43). Incidentally, the importance of genome GC in determining codon and aminoacid usage is stressed by the great similarities of these properties in *Plasmodium falciparum* and *Staphylococcus aureus,* two distantly related organisms with only a very GC-poor genome in common (88).

DNA REPLICATION, DNA RECOMBINATION AND COMPOSITIONAL HETEROGENEITY OF RBANDS R bands replicate early and G bands replicate late in the cell cycle (see 33, 57). However, replication timing appears to be associated not with G/R banding, but with gene concentration and expression (see 9). Replication (BrdU) banding is also found in cold-blooded vertebrates that show no or poor G/R banding (105). This suggests that in R bands, which are compositionally heterogeneous and characterized by different gene concentrations, replication might be bimodal, with GC-rich isochores replicating early and GC-poor isochores replicating late. Alternatively, both GC-poor and GC-rich isochores from R bands might replicate early (49). These two possibilities are not mutually exclusive in that different R bands behave differently. While this would be hard to discriminate by standard cytogenetic techniques, it could be discriminated by using other approaches (106).

The correlation of recombination with G/R banding has been briefly reviewed (9). In short, R bands and G/R borders are the predominant locations of exchange processes (including spontaneous translocations, sister chromatid exchanges, chromosomal aberrations, mitotic chiasmata) and fragile sites. These observations point to a possible role played by compositional discontinuities at G/R borders and within R bands. Needless to say, recombinogenic sequences, like Alu sequences, CpG islands and minisatellites, which are specially abundant in R bands (30, 35, 52, 71, 95), also play a role.

At the DNA level, recombination frequencies (investigated as chiasmata densities) are higher in GC-rich isochores (50). Indeed, at the chromosomal level, $H3^+$ (T) bands and H3 (T') bands contain the preferred regions for recombination. Along the same line, compositional mapping of Xq28 has provided evidence for the instability of YACs that contain com-

positionally heterogeneous GC-rich sequences, and the stability of compositionally homogeneous GC-poor YACs (A De Sario et al, manuscript in preparation).

VIRAL SEQUENCE INTEGRATION IN MAMMALIAN GENOMES Especially in the case of expressed sequences, viral sequence integration preferentially takes place in host genome sequences that compositionally match the viral sequences. Indeed, integrated, expressed sequences of GC-rich bovine leukemia virus (68), hepatitis B virus (115), and Rous sarcoma virus (96) were found in the GC-richest isochores. In contrast, GC-poor mouse mammary tumor virus sequences were localized in the GC-poor compartments of the host genome (103). These results, by themselves, suggest that in order to be expressed, integrated viral sequences must be located in compositionally matching chromosomal environments as are host genes.

The compositional distribution of retroviral genomes is bimodal (119). The GC-rich class comprises the genomes of all oncoviruses, except for some oncoviruses of B and D types. The GC-poor class comprises lentiviruses and spumaviruses that have no oncogenes, but contain genes for regulatory proteins, as well as oncoviruses of B type and some of D type. On the other hand, the coding sequences present in the same viral genome and, to some degree, even the corresponding LTRs show a remarkable compositional homogeneity. Such compositional bimodality might well reflect the origin of viral sequences from different isochores and/or their permanence in them in integrated form.

The most recent and extensive work in this area concerns the integration of 40 HTLV-I sequences into the genome of immortalized human cell lines and T-cell clones and its correlation with transcription (118). HTLV-I sequences, which are GC-rich, were only found in the >39% GC range of the host genome (see Figure 4*a*). In other words, no HTLV-I sequences are present in the GC-poor 40% of the human genome, indicating either an inability to integrate or a high instability after integration. Moreover, the distribution of integrated proviral copies in the GC-rich part of the genome is different (*a*) for the trancriptionally active HTLV-I sequences, which were only found in the >44% GC regions of the host genome; in fact, most of the HTLV-I sequences that were integrated in the GC-rich compartment were located in regions above 47% GC, which correspond to the GC-richest 15% of the genome that make up T bands; and (*b*) for the transcriptionally inactive HTLV-I sequences, which were only found in the 39–42% GC regions. Interestingly, the locations of transcriptionally inactive and active proviral HTLV-I sequences correspond to the isochores associated with a closed and with an open chromatin structure, respectively.

There are two possibilities, not mutually exclusive, why some HTLV-I

sequences are transcribed and others are not: The proviral sequences integrated in GC-poor isochores may be transcriptionally inactivated by methylation and/or deletions, or they may be silenced by the "closed" chromatin structure associated with GC-poor isochores. In either case, these integrations may be responsible for the low level of expression of some retroviral constructs used in gene therapy (for example, see 58).

Mobile sequences in general as well as integrated viral sequences exhibit a compositional match with the isochores in which they are predominantly embedded. For example, GC-poor LINES are predominantly present in GC-poor isochores and GC-rich SINES in GC-rich isochores (109). The same applies to tobacco and maize mobile elements (28, 29, 76). Furthermore, some cases of gene silencing in plant transgenes are associated with compositional discrepancies with the host genome environment (80).

GENOME EVOLUTION

Compositional Genome Transitions in Vertebrates

Figure 7 recalls that the genomes of cold-blooded vertebrates are characterized by a small compositional heterogeneity; their GC-richest isochores are much lower in GC than the GC-richest isochores of warm-blooded vertebrates (15–17).

At some point in time, the genomes of species on the two ancestral lines leading from reptiles to mammals and birds, respectively, underwent compositional transitions (Figure 7). In a major part of the genome, isochores remained GC-poor, whereas in a minor part isochores became GC-rich. The former were called the paleogenome, the latter the neogenome (9). The compositional transition was paralleled by the appearance of R bands in metaphase chromosomes (105) and by compositional changes in coding sequences and especially in third codon positions. The latter changes can be conveniently studied in homologous genes (Figure 8) and provide important insights into the evolution of vertebrate genome organization.

First, a comparison of homologous genes from *Xenopus* and man (Figure 8*a*) showed that the majority of human genes exhibited a higher GC_3 level than their *Xenopus* homologs (most of the points are above the diagonal); this increase is more evident in increasingly GC_3-richer *Xenopus* genes. This behavior suggests that the neogenome of warm-blooded vertebrates essentially originated from the GC-richer isochores of reptiles. Moreover, this strongly points to a conserved gene distribution in vertebrates, in that the gene-richest isochores are also the GC-richest ones in both cold- and warm-blooded vertebrates. This conclusion is also supported by the finding that DNA from the isochore family H3 hybridizes (under conditions suppressing the contributions

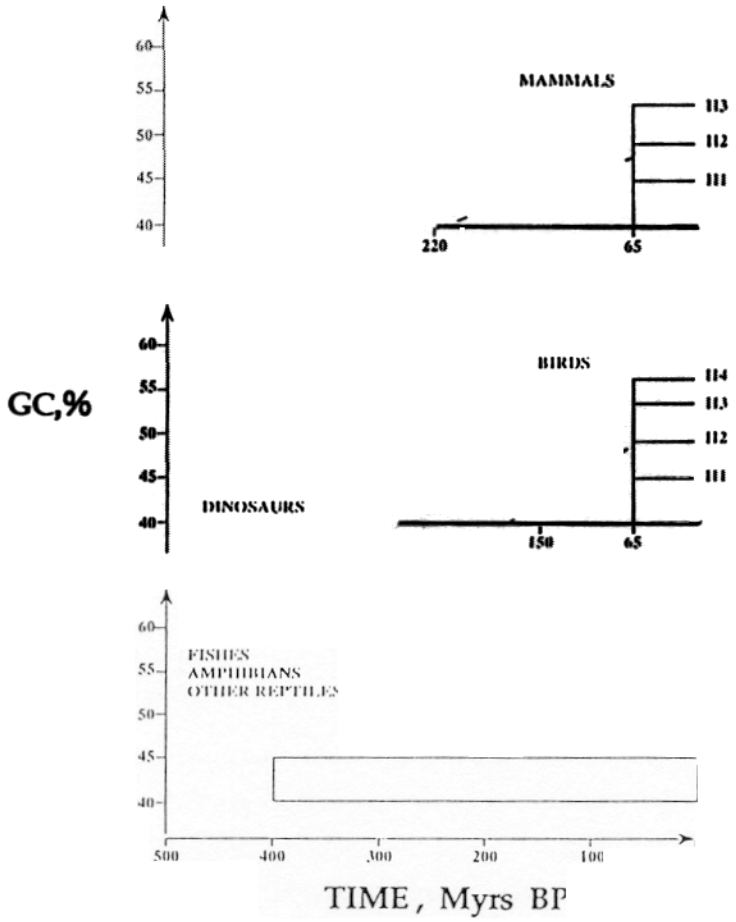


Figure 7 Scheme of the formation of GC-rich isochores in the genome of warm-blooded vertebrates. While the genomes of cold-blooded vertebrates are characterized by a narrow range of compositional heterogeneity, the genomes of lines leading to mammals and birds underwent strong GC increases in a minority of the isochore families, H1 to H4; the majority remained formed by L isochores. (Modified from Reference 11.)

from repeated sequences) with the GC-richest isochores of all other vertebrates, including cold-blooded vertebrates (26).

Second, changes in GC_3 were similar in homologous genes from mammals and birds, as shown by the significant correlation of Figure 8*b*. This result is especially interesting because these two classes of vertebrates had completely independent origins in evolution: Mammals arose about 220 My ago from therapsids, and birds about 150 My ago from dinosaurs. This result can again only be understood if gene distribution is conserved in vertebrates, and if GC increases preferentially affected the isochores characterized by higher gene concentrations in both transitions. Interestingly, the scatter of points is much lower for GC_3-poor genes that did not undergo compositional changes (see above) than for GC_3-rich genes that did. Finally, this result points to common causes being responsible for these parallel compositional changes that took place in mammalian and avian genomes.

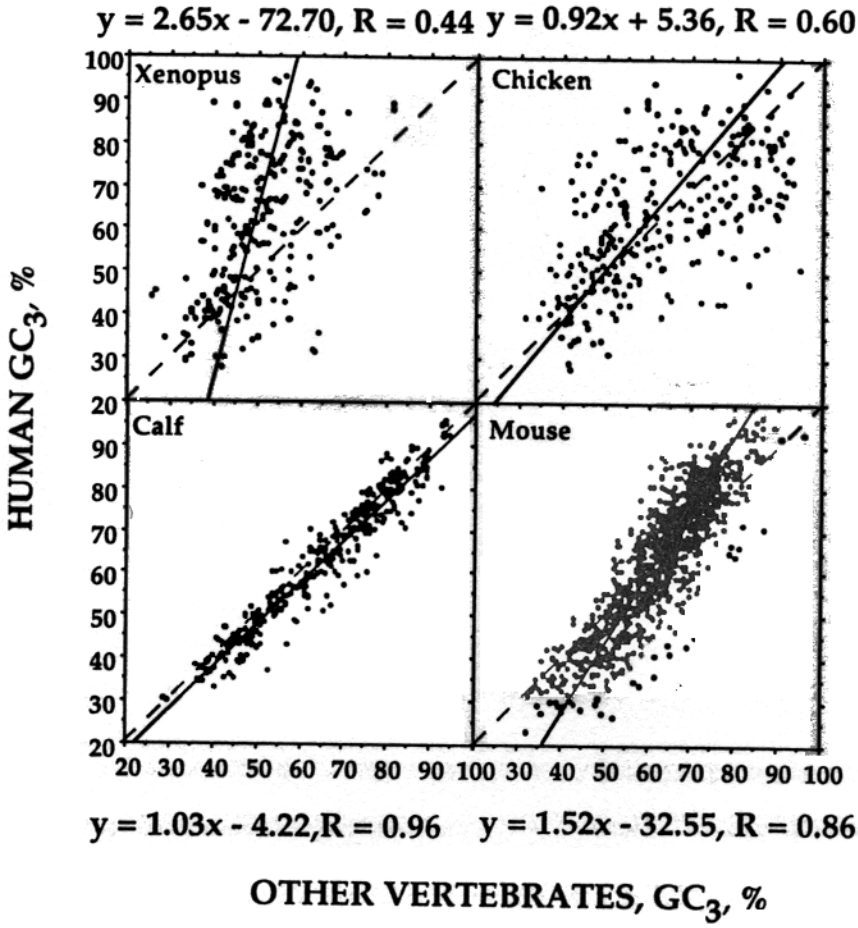


Figure 8 GC levels of third codon positions of human genes (ordinate) are plotted against those of their homologs from calf, mouse, bovine, chicken and *Xenopus* (abscissa). Orthogonal regression lines, diagonals, slopes (s) and correlation coefficients (R) are shown. (Updated from References 66, 82.)

Compositional Genome Conservation in Vertebrates

Major compositional changes such as those that took place in the two lines just discussed and that led to the strong heterogeneities of the genome of warm-blooded vertebrates, did not occur in cold-blooded vertebrates (Figure 7), nor did they take place over the past 65 My in most mammals and in all birds examined. Indeed, GC_3 values of homologous mammalian genes from genomes exhibiting the general pattern are extremely similar (Figure 8*c*). Because of the star phylogeny of mammalian orders from a common ancestor, one should conclude that such changes stopped, and a compositional equilibrium was reached, at the latest with the appearance of present-day mammals. The limited available data show that GC_3 of homologous genes from birds belonging to different orders also are very similar (66).

Two points should be stressed. First, the compositional equilibrium prevailing in the genome of warm-blooded vertebrates was accompanied by the maintenance of a very broad compositional spectrum, whereas the composi-

tional equilibrium of cold-blooded vertebrates concerned much less heterogeneous genomes.

The second point is that the genomes of myomorphs (as well as those of other infraorders and families of mammals) underwent further changes, leading to narrower compositional distributions that indicate a relaxation of the constraints that maintain strong heterogeneity in other mammals (see above). Moreover, if one accepts the monophyly of rodents (which is supported by paleontology; 89), the special pattern exhibited by only one of the three Rodent infraorders can only be a secondary pattern derived from the general pattern.

SYNONYMOUS SUBSTITUTIONS IN HOMOLOGOUS MAMMALIAN GENES Only some of the many interesting features of silent changes are dealt with here. First, the frequencies of synonymous substitutions cover a much wider range than previously thought and are gene specific: "fast" and "slow" genes in one mammal are fast and slow, respectively, in any other one (86).

Second, the frequencies of synonymous substitutions are correlated with the frequencies of nonsynonymous substitutions in the same genes (86, 90a, 90b).

Third, when synonymous substitutions were studied in quartet (fourfold degenerate) codons of homologous genes from four mammalian orders, their frequencies were clearly different from those expected as the result of a stochastic process in which nucleotide substitutions accumulated at random over time. This was especially true for GC-rich genes (27), which represent the majority of genes in warm-blooded vertebrates. Moreover, conserved positions exhibited significantly different base compositions (higher in G and C) compared to those expected from a random substitution process; again, this was much more evident for GC-rich than for GC-poor genes (117). In other words, this analysis revealed strong constraints in a number of synonymous positions of quartet codons, especially in GC-rich genes. Interestingly, the relaxation of such constraints undergone by the genomes of Myomorphs (see above) was accompanied by increased synonymous substitution rates (113a), as one would expect from the decrease of the number of G- and C-rich conserved synonymous positions.

The compositional correlations between GC_3, especially of GC-rich genes (which comprise large numbers of constrained, conserved positons), and GC of the isochores hosting the corresponding genes indicate the existence of similar constraints in the latter, even if the degree of the constraint is different (as it is between different codon positions). These constraints are also indicated by the sequence conservation of a number of noncoding sequences (48, 70), but not of others (38, 59). These results suggest (11, 14) that the intergenic noncoding sequences (especially in the case of the "genome core") are not "junk DNA," but play a functional role, which may have to do with chromo-

somal structure and gene expression, as suggested by the results on the transcription of integrated HTLV-I sequences (118).

Some Consequences of the Compositional Genome Transitions

SPECIATION IN VERTEBRATES Karyotypic change rate is tenfold higher and speciation rate is fivefold higher in mammals compared to cold-blooded vertebrates (25). This difference has been explained as due to the formation of small demes and social structuring (25). Alternatively, the compositional discontinuities of the chromosomes of warm-blooded vertebrates, which are characteristic of R bands, are responsible for an instability of the genome and may favor chromosomal rearrangements and speciation (10).

THE ORIGIN OF CpG ISLANDS In contrast to their abundance in warm-blooded vertebrates, CpG islands are extremely scarce or absent in the genomes of cold-blooded vertebrates (1, 2, 9, 36). When homologous genes from these two groups of vertebrates were compared (1, 2), only a certain number of genes from cold-blooded vertebrates exhibited "primitive CpG islands"; these showed (*a*) CpG doublets that, although characterized by a low frequency, tended to approach statistical expectations in these GC-poor genomes; and (*b*) shared none of the other features of CpG islands associated with genes from warm-blooded vertebrates (high GC levels, Hpa II sites, rare-cutter sites, and G/C boxes).

Since the features of CpG islands in cold-blooded vertebrates are the ancestral ones, two questions may be raised: How did these islands acquire the features present in warm-blooded vertebrates and how did they arise in the first place? Concerning the first question, it is clear that mammalian CpG islands are not "the last remnants of the long tracts of nonmethylated DNA that make up most of the invertebrate genome" as proposed (23). Indeed, these islands essentially arose as a consequence of the compositional genome transition intervening between reptiles and warm-blooded vertebrates. This is demonstrated by the fact that GC-rich coding sequences present in H3 isochores and classified as CpG islands were shown, by comparison of homologous sequences (Figure 8), to have undergone directional changes, especially in third codon positions, leading to G and C enrichment. The newly formed CpG doublets were obviously not methylated, and being located in the particular nucleotide sequences of CpG islands, were inherently resistant to methylation (21). Therefore, they did not undergo the 5-methyl C→T changes causing the CpG shortage elsewhere. Interestingly, when the GC-richest isochores of the general patterns of mammalian genome happened to disappear, as in Myomorphs, a number of CpG islands associated with the genes located in those isochores (such as the α-globin genes) concomitantly disappeared (1, 2, 78).

The explanation that links the formation of CpG islands of warm-blooded

vertebrates to the compositional transition from the genomes of cold-blooded vertebrates is further supported by the observation that the frequency of CpG doublets increases linearly with increasing GC of exons; these correlations are very similar for both cold- and warm-blooded vertebrates(1, 2). The origin of the primitive CpG islands may then be linked with the existence of relatively high local GC levels in the gene-rich isochores of cold-blooded vertebrates. This explanation can also account for the appearance of CpG islands in the genomes of plants (5), which share no common ancestor, carrying CpG islands, with vertebrates.

Contraction-Expansion Phenomena and Compositional Changes

The GC-poor, gene-poor isochores of vertebrates are not only characterized by a lack of compositional change upon the transition from cold- to warm-bloodedness, but also by phenomena of contraction and expansion, which appear to be much more frequent and/or extensive than in GC-rich, gene-rich isochores. This can be best observed in fish, where a number of genomes covering a wide range of genome sizes (*c*-values) were investigated (15–17). An extreme case of genome contraction is that of *Tetraodontiformes*, (one of the most recent fish orders), which are characterized by genome sizes (59a) as small as 0.4 pg (vs 3.3 pg for the human genome) and a striking scarcity of repeated sequences (93) compared to mammals (108).

If cases of polyploidy are neglected, a negative correlation is found between GC level and genome size, indicating an AT increase in the larger genomes (15). Such contraction-expansion phenomena appear to affect mainly the GC-poor, gene-poor isochores. This conclusion is supported by two findings: (*a*) genes located in the GC-poor isochores of mammals and *Xenopus* are characterized by larger introns than genes located in GC-rich isochores (45); and (*b*) while the GC-rich G6DP gene is only four times larger in humans than in a tetraodontiform, *Fugu* (74), the GC-poor Huntington gene disease is 7.4 larger (8), a value identical to the ratio between the two genome sizes. Under such circumstance, the interest of the *Fugu* genome model (24) mainly concerns GC-poor genes.

Isochores in Eukaryotes

ISOCHORES IN PLANTS The nuclear genomes of angiosperms are characterized by an isochore structure, as indicated by the compositional homogeneity of large regions (>100–200 Kb) surrounding the coding sequences tested; flanking regions and introns are typically lower in GC than exons (75–77, 81, 101). Interestingly, the compositional patterns of the genomes from *Gramineae* differ from those of the other monocots and dicots explored in that they exhibit

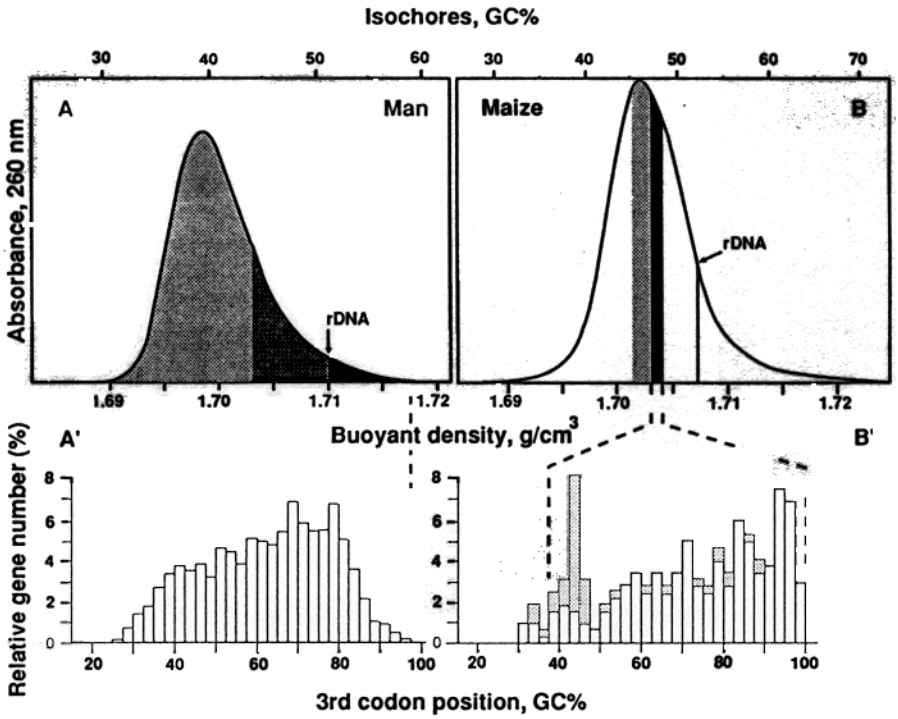


Figure 9 *Top*: The CsCl profile of human DNA (*A*) is compared with that of maize DNA (*B*). In human DNA, black, gray, and light gray areas correspond to high, medium, and low gene concentrations in the L1+L2, H1+H2, and H3 (i.e. the GC-poor, GC-rich, and very GC-rich) isochore families, respectively. In maize, the black area corresponds to the "gene space"; however, zein genes occupy not only the "gene space" but also DNA compartments corresponding to the light gray area to its left; the white areas represent genome regions where no protein-encoding genes were detected. Ribosomal genes were found to be centered at the buoyant densities marked rDNA.

Bottom: The relative number of genes from human (*A'*) and maize (*B'*) are plotted against GC_3 of the corresponding coding sequences. The histogram of 2762 human genes is from Reference 83. Maize data are from GenBank and concern 265 genes plus 41 zein genes (shown as light gray bars). The vertical broken lines correlate the gene distributions (bottom frames) with the DNA distributions (top frames). In maize, such vertical lines only bracket the GC_3 range of the genes tested. (From Reference 29.)

broader DNA fragment distributions centered around higher GC levels. In addition, GC_3 values of genes from *Gramineae* range from 30 to 100%, with most values in the high range, whereas those of other angiosperms are contained between 15 and 70%. Interestingly, while the genome of maize exhibits the same genome size, the same compositional distribution of DNA fragments and a similar distribution of coding sequences (Figure 9), its gene distribution is very different (29). Almost all the genes explored in maize (with the exception of some of the genes coding for storage proteins, which showed only a slightly wider distribution) were comprised in 10–20% of the genome, in isochores (called the "gene space") only covering a 1–2% GC range. These isochores are distributed over all chromosomes, preferentially in telomeric regions, where most of the recombination occurs; the rest of the genome contains only very few or no genes. The "gene space" of maize apparently

also comprises the only genome compartments in which a number of compositionally matching transcribed mobile elements are found, a result reminiscent of the findings on transcribed viral sequences in the mammalian genome (see a preceding section).

ISOCHORES IN TRYPANOSOMES AND PLASMODIA The nuclear genomes of *T. brucei* and *T. equiperdum* are so strongly compartmentalized as to exhibit a bimodal DNA distribution (65). Likewise, the GC_3 distribution is bimodal in *T. brucei* and *T. cruzii* (87). Preliminary hybridization experiments have shown that GC-poor expression site-associated genes (ESAG) are located in GC-poor compartments and that GC-rich housekeeping genes are typically found in GC-rich compartments. In plasmodia, isochores that are likely to average 100 Kb form the nuclear genome of *P. cynomolgi* (79).

ISOCHORES IN YEAST Three yeast chromosomes III (315 Kb), XI (666 Kb), and II (807 Kb) have been fully sequenced (44, 51, 91). In chromosome XI, four major GC-rich peaks were seen on the left arm and one major peak (plus a minor one) on the right arm (which is much shorter). Like chromosomes III and II, chromosome XI appears to be a mosaic of GC-rich and GC-poor segments of 50 Kb each. Interestingly, this mosaic structure is seen not only in silent positions, where it is sharper, as expected, and where it was first detected (107), but also in entire open reading frames [(ORFs) which are potentially coding sequences] and in intergenic sequences. Moreover, introns are systematically lower in GC than exons, as is the case in vertebrates and plants. Finally and most important, the alternating regional variations in average GC correlate with variations in local gene density.

CONCLUSIONS

The major conclusions of this review are as follows:
(*a*) As previously suggested (20), an isochore organization is very widespread, and possibly general, in the genomes of eukaryotes, in contrast to prokaryotes, where, as a rule, genomes are compositionally very homogeneous. The different compositional heterogeneity and the accompanying abundance of repeated sequences are the major distinctive features of prokaryotic and eukaryotic genomes.

The isochore organization of eukaryotic genomes is very different in the distantly related organisms explored so far, namely yeast, trypanosomes, plasmodia, plants, and vertebrates; and highly significant differences also exist at least within the last two phyla. This situation justifies the concept of compositional genome phenotype originally introduced for vertebrate genomes (14).

If narrow phylogenetic ranges are considered, genome phenotypes are, however, remarkably conserved in evolution.

(*b*) The compositionally conservative mode of genome evolution (18) is definitely the predominant evolutionary mode in vertebrates. In contrast, the transitional mode of genome evolution is rare in the case of major compositional shifts, like those between cold- and warm-blooded vertebrate, or between *Gramineae* and other monocots, but less so in that of minor compositional shifts, like those between the general mammalian pattern and the special myomorph pattern, or those detected among cold-blooded vertebrates (16). The causes and mechanisms underlying the two modes of genome evolution are issues of major importance in molecular evolution, but they are not discussed here (however, see References 14, 18, 19, 27, 86, 117).

(*c*) Gene distribution appears to be the most fundamental property of the eukaryotic genome. In higher eukaryotes, the distribution of genes in the genome is strikingly nonuniform, as stressed by the "genome core" of humans and the "gene space" of maize. Gene-rich isochores are characterized by distinct structural and functional features, like chromatin structure and levels of transcription and recombination. Moreover, gene distribution in the genomes of vertebrates and, probably, of plants is highly conserved in evolution.

Under these circumstances, it is not surprising that a remarkable conservation in gene order is found at the chromosomal level. Syntheny is extensive within mammals (90, 104) and still quite detectable even between vertebrate classes as separated in evolution as mammals and fishes.

While the "gene space" and the "genome core" are key features of the genomes of higher eukaryotes and may be the starting point for speculations about the evolution of genomes of higher vertebrates (GB, manuscript in preparation), they are also important from a practical point of view because their high gen concentrations indicate obvious priorities in the strategies of genome projects (see Reference 6).

(*d*) In situ hybridization of compositional DNA fractions corresponding to different isochore families has provided the most rigorous approach to a rational classification of chromosomal bands that, so far, has relied on empirical criteria. At the same time, it has provided a gene distribution map of chromosomes, which will be further refined by compositional maps at the molecular level. Finally, it has clarified the much debated correlations between base composition and chromosomal bands of metaphase chromosomes.

e) The correlations between coding and noncoding sequences (namely, the genome equations of Figure 3, which represent a genomic code) are due to compositional constraints that are strong in the gene-rich, GC-rich isochores, and weak in the gene-poor, GC-poor isochores. This indicates that the vertebrate genome is not only a structural mosaic of isochores and a functional mosaic of transcription, replication, and recombination levels, but also an

evolutionary mosaic characterized by different levels of constraints and substitution rates. The compositional constraints on noncoding sequences stress the fundamental unity of the genome and the functional role of "junk DNA."

ACKNOWLEDGMENTS

Thanks are due to Brahim Aïssani, Abdelali Barakat, Giacomo Bernardi, Thomas Bettecken, Simone Cacciò, Nicolas Carels, Oliver Clay, Albertina De Sario, Eva-Maria Geigl, Kamal Jabbari, Farida Kadi, Gabriel Macaya, Giorgio Matassi, Dominique Mouchiroud, Hector Musto, Paolo Perani, Helena Rodriguez-Maseda, Alla Rynditch, Georgette Sabeur, Salvatore Saccone, and Serguei Zoubak for their contribution to the investigations reported here.

Literature Cited

Aissani B, Bernardi G. 1991. CpG islands: features and distribution in the genomes of vertebrates. *Gene* 106:173–83

2. Aissani B, Bernardi G. 1991. CpG islands, genes and isochores in the genomes of vertebrates. *Gene* 106:185–95
3. Aissani B, D'Onofrio G, Mouchiroud D, Gardiner K, Gautier C, Bernardi G. 1991. The compositional properties of human genes. *J. Mol. Evol.* 32:497–503
4. Ambros PF, Sumner AT. 1987. Correlation of pachytene chromomeres and metaphase bands of human chromosomes, and distinctive properties of telomeric regions. *Cytogenet. Cell Genet.* 44:223–38
5. Antequera F, Bird AP. 1988. Unmethylated CpG islands associated with genes in higher plant DNA. *EMBO J.* 7:2295–99
6. Antonarakis SE. 1994. Genome linkage scanning: systematic or intelligent? *Nature Genet.* 8:211–12
7. Aota SI, Ikemura T. 1986. Diversity in G+C content at the third position of codons in vertebrate genes and its cause. *Nucleic Acids Res.* 14:6345–55
8. Baxendale S, Abdulla S, Elgar G, Buck D, Berks M, et al. 1995. Comparative sequence analysis of the human and pufferfish Huntington's disease genes. *Nature Genet.* 10:67–76
9. Bernardi G. 1989. The isochore organization of the human genome. *Annu. Rev. Genet.* 23:637–61
10. Bernardi G. 1993. Genome organization and species formation in vertebrates. *J. Mol. Evol.* 37:331–37
11. Bernardi G. 1993. The vertebrate genome: isochores and evolution. *Mol. Biol. Evol.* 10:186–204
12. Bernardi G. 1993. The human genome organization and its evolutionary history: a review. *Gene* 13:57–66
13. Bernardi G, Bernardi G. 1985. Codon usage and genome composition. *J. Mol. Evol.* 22:363–65
14. Bernardi G, Bernardi G. 1986. Compositional constraints and genome evolution. *J. Mol. Evol.* 24:1–11
15. Bernardi G, Bernardi G. 1990. Compositional patterns in the nuclear genomes of cold-blooded vertebrates. *J. Mol. Evol.* 31:265–81
16. Bernardi G, Bernardi G. 1990. Compositional transitions in the nuclear genomes of cold-blooded vertebrates. *J. Mol. Evol.* 31:282–93
17. Bernardi G, Bernardi G. 1991. Compositional properties of nuclear genes from cold-blooded vertebrates. *J. Mol. Evol.* 33:57–67
18. Bernardi G, Mouchiroud D, Gautier C, Bernardi G. 1988. Compositional patterns in vertebrate genomes: conservation and change in evolution. *J. Mol. Evol.* 28:7–18
19. Bernardi G, Mouchiroud D, Gautier C,

1993. Silent substitutions in mammalian genomes and their evolutionary implications. *J. Mol. Evol.* 37:583–89
20 Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, et al. 1985. The mosaic genome of warm-blooded vertebrates. *Science* 228:953–58
Bestor TH, Gundersen G, Kolsto A-B, Prydz H. 1992. CpG islands in mammalian gene promoters are inherently resistant to de novo methylation. *Gata* 9:48–53
22 Bettecken T, Aissani B, Müller CR, Bernardi G. 1992. Compositional mapping of the human dystrophin gene. *Gene* 122:329–35
23. Bird AP. 1987. CpG islands as gene markers in the vertebrate nucleus. *Trends Genet.* 3:342–47
24. Brenner S, Elgar G, Sandford R, Macrae A, Venkatesh B, Aparicio S. 1993. Characterization of the pufferfish (Fugu) genome as a compact model vertebrate genome. *Nature* 366:265–68
25 Bush GL, Case SM, Wilson AC, Patton JL. 1977. Rapid speciation and chromosomal evolution in mammals. *Proc. Natl. Acad. Sci. USA* 74:3942–46
26. Cacciò S, Perani P, Saccone S, Kadi F, Bernardi G. 1994. Single-copy sequence homology among the GC-richest isochores of the genomes from warm-blooded vertebrates. *J. Mol. Evol.* 39: 331–39
27. Cacciò S, Zoubak S, D'Onofrio G, Bernardi G. 1995. Nonrandom frequency patterns of synonymous substitutions in homologous mammalian genes. *J. Mol. Evol.* 40:280–92
28. Capel J, Montero LM, Martinez-Zapater JM, Salinas J. 1993. Non-random distribution of transposable elements in the nuclear genome of plants. *Nucleic Acids Res.* 21:2369–73
29. Carels N, Barakat A, Bernardi G. 1995. The gene distribution in the maize genome. *Proc. Natl. Acad. Sci. USA.* In press
30. Christmann A, Lagoda PJL, Zang KD. 1991. Non-radioactive in situ hybridization pattern of the M13 minisatellite sequences on human metaphase chromosomes. *Hum. Genet.* 86:487–90
31 Clay O, Cacciò S, Zoubak S, Mouchiroud D, Bernardi G. 1995. Human coding and noncoding DNA sequences: compositional correlations. *Mol. Phylogenet. Evol.* In press
32 Cohen D, Chumakov I, Weissenbach J. 1993. A first-generation physical map of the human genome. *Nature* 366:698–701
33 Comings DE. 1978. Mechanisms of chromosome banding and implications for chromosome structure. *Annu. Rev. Genet.* 12:25–46
34. Craddock CF, Vyas P, Sharpe JA, Ayyub H, Good WG, Higgs DR. 1995. Contrasting effects of α and β globin regulatory elements on chromatin structure may be related to their different chromosomal environments. *EMBO J.* 14:1718–26
35. Craig JM, Bickmore WA. 1994. The distribution of CpG islands in mammalian chromosomes. *Nat. Genet.* 7:376–81
36. Cross S, Kovarik P, Schmidtke J, Bird A. 1991. Non-methylated islands in fish genomes are GC-poor. *Nucleic Acids Res.* 19:1469–74
37 de Lange T. 1992. Human telomeres are attached to the nuclear matrix. *EMBO J.* 11:717–24
38 Den Dunnen JT, Van Neck JW, Cremers FPM, Lubsen NH, Schoenmakers JGG. 1989. Nucleotide-sequence of the rat gamma-crystallin gene region and comparison with an orthologous human region. *Gene* 78:201–13
39 De Sario A, Aïssani B, Bernardi G. 1991. Compositional properties of telomeric regions from human chromosomes. *FEBS Lett.* 295:22–26
40. De Sario A, Geigl E-M, Bernardi G. 1995. A rapid procedure for the compositional analysis of yeast artificial chromosomes. *Nucleic Acids Res.* In press
41 Deleted in proof
42 D'Onofrio G, Bernardi G. 1992. A universal compositional correlation among codon positions. *Gene* 110:81–88
43. D'Onofrio G, Mouchiroud D, Aïssani B, Gautier C, Bernardi G. 1991. Correlations between the compositional properties of human genes, codon usage and aminoacid composition of proteins. *J. Mol. Evol.* 32:504–10
44. Dujon B, Alexandraki D, Andre B, Ansorge W, Baladron V, et al. 1994. Complete DNA sequence of yeast chromosome XI. *Nature* 369:371–78
45. Duret L, Mouchiroud D, Gautier C. 1995. Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. *J. Mol. Evol.* 40:308–17
46. Dutrillaux B. 1973. Nouveau système de marquage chromosomique: les bandes T. *Chromosoma* 41:395–402
47. Ellis NA, Goodfellow PN. 1989. The mammalian pseudoautosomal region. *Trends Genet.* 5:406–10
48 Epp TA, Dixon IMC, Wang H-Y, Sole MJ, Liew C-C. 1993. Structural organization of the human cardiac alpha-

myosin heavy-chain gene. *Genomics* 18: 505–26
49. Eyre-Walker A. 1992. Evidence that both G+C rich and G+C poor isochores are replicated early and late in the cell cycle. *Nucleic Acids Res.* 20:1497–501
50. Eyre-Walker A. 1993. Recombination and mammalian genome evolution. *Proc. R. Soc. London Ser. B* 252:237–43
51. Feldmann H, Aigle M, Aljinovic G, Andre B, Baclet MC, et al. 1994. Complete DNA sequence of yeast chromosome II. *EMBO J.* 24:5795–809
52. Ferraro M, Predazzi V, Prantera G. 1993. In human chromosomes telomeric regions are enriched in CpGs relative to R-bands. *Chromosoma* 102:712–17
53. Filipski J, Thiery JP, Bernardi G. 1973. An analysis of the bovine genome by Cs2SO4-Ag+ density gradient centrifugation. *J. Mol. Biol.* 80:177–97
54. Fukagawa T, Sugaya K, Matsumoto K-i, Okumura K, Ando A, et al. 1995. Characterization of the boundary region of long-range G+C% mosaic domains in the human MHC locus; pseudoautosomal boundary-like sequence near the boundary. *Genomics* 25:184–91
55. Gardiner K, Aïssani B, Bernardi G. 1990. A compositional map of human chromosome 21. *EMBO J.* 9:1853–58
56. Goldman MA, Holmquist GP, Gray MC, Caston LA, Nag A. 1984. Replication timing of genes and middle repetitive sequences. *Science* 224:686–92
57. Gyapay G, Morissette J, Vignal A, Dib C, Fizames C, et al. 1994. The 1993–94 généthon human genetic linkage map. *Nat. Genet.* 7:246–339
58. Hafenrichter DG, Wu X, Rettinger SD, Kennedy SC, Flye MW, Ponder KP. 1994. Quantitative evaluation of liver-specific promoters from retroviral vectors after in vivo transduction of hepatocytes. *Blood* 84:3394–404
59. Hardison R, Miller W. 1993. Use of long sequence alignments to study the evolution and regulation of mammalian globin gene clusters. *Mol. Biol. Evol.* 10:73–102
59a. Hinegardner R, Rosen DE. 1972. Cellular DNA content in the evolution of teleostan fishes. *Am. Nat.* 100:621–644
60. Holmquist GP. 1992. Chromosome bands, their chromatin flavors, and their functional features. *Am. J. Hum. Genet.* 51:17–37
61. Hudson AP, Cuny G, Cortadas J, Haschemeyer AEV, Bernardi G. 1980. An analysis of fish genomes by density gradient centrifugation. *Eur. J. Biochem.* 112:203–10
61a. Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.* 2:13–34
62. Ikemura T, Aota SI. 1988. Global variation in G+C content along vertebrate genome DNA. *J. Mol. Biol.* 203:1–13
63. Ikemura T, Wada KN. 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 cytogeneric data. *Nucleic Acids Res.* 19:4333–39
64. Ikemura T, Wada KN, Aota SI. 1990. Giant G+C% mosaic structures of the human genome found by arrangement of genbank human DNA sequences according to genetic positions. *Genomics* 8:207–16
65. Isacchi A, Bernardi G, Bernardi G. 1993. Compositional compartmentalization of the nuclear genomes of *Trypanosoma brucei* and *Trypanosoma equiperdum*. *FEBS Lett.* 335:181–83
66. Kadi F, Mouchiroud D, Sabeur G, Bernardi G. 1993. The compositional patterns of the avian genomes and their evolutionary implications. *J. Mol. Evol.* 37:544–51
67. Kerem BS, Goitein R, Diamond G, Cedar H, Marcus M. 1984. Mapping of DNAase I sensitive regions of mitotic chromosomes. *Cell* 38:493–99
68. Kettmann R, Meunier-Rotival M, Cortadas J, Cuny G, Ghysdael J, et al. 1979. Integration of bovine leukemia virus DNA in the bovine genome. *Proc. Natl. Acad. Sci. USA* 76:4822–26
69. Kipling D. 1995. *The Telomere*. Oxford, UK: Oxford Univ. Press
70. Koop BF, Hood L. 1994. Striking sequence similarity over almost 100 kilobases of human and mouse T-cell receptor DNA. *Nat. Genet.* 7:48–53
71. Korenberg J, Rikowski M. 1988. Human molecular organization: Alu, Lines and the molecular structure of metaphase chromosome bands. *Cell* 53:391–400
71a. Krane DE, Hartl DL, Ochman H. 1991. Rapid determination of nucleotide content and its application to the study of genome structure. *Nucleic Acids Res.* 19:5181–85
72. Larsen F, Gundersen G, Lopez R, Prydz H. 1992. CpG islands as gene markers in the human genome. *Genomics* 13: 1095–107
73. Macaya G, Thiery JP, Bernardi G. 1976. An approach to the organization of eukaryotic genomes at a macromolecular level. *J. Mol. Biol.* 108:237–54
74. Mason PJ, Stevens DJ, Luzzatto L, Brenner S, Aparicio S. 1995. Genomic

structure and sequence of the Fugu rubripes glucose-6-phosphate dehydrogenase gene (G6PD). *Genomics* 26: 587–91

75. Matassi G, Melis R, Kuo KC, Macaya G, Gehrke CW, Bernardi G. 1992. Large-scale methylation patterns in the nuclear genomes of plants. *Gene* 122: 239–45
76. Matassi G, Melis R, Macaya G, Bernardi G. 1991. Compositional bimodality of the nuclear genome of tobacco. *Nucleic Acids Res.* 19:5561–67
77. Matassi G, Montero LM, Salinas J, Bernardi G. 1989. The isochore organization and the compositional distribution of homologous coding sequences in the nuclear genome of plants. *Nucleic Acids Res.* 17:5273–90
78. Matsuo K, Clay O, Takahashi T, Silke J, Schaffner W. 1993. Evidence for erosion of mouse CpG islands during mammalian evolution. *Somat. Cell Mol. Genet.* 19:543–55
79. McCutchan TF, Dame JB, Gwadz RW, Vernick KD. 1988. The genome of *Plasmodium cynomolgi* is partitioned into separable domains which appear to differ in sequence stability. *Nucleic Acids Res.* 16:4499–510
80. Meyer P. 1995. *Gene Silencing in Higher Plants and Related Phenomena in Other Eukaryotes*. Heidelberg: Springer-Verlag
81. Montero LM, Salinas J, Matassi G, Bernardi G. 1990. Gene distribution and isochore organization in the nuclear genome of plants. *Nucleic Acids Res.* 18:1859–67
82. Mouchiroud D, Bernardi G. 1993. Compositional properties of coding sequences and mammalian phylogeny. *J. Mol. Evol.* 37:109–16
83. Mouchiroud D, D'Onofrio G, Aïssani B, Macaya G, Gautier C, Bernardi G. 1991. The distribution of genes in the human genome. *Gene* 100:181–87
84. Mouchiroud D, Fichant G, Bernardi G. 1987. Compositional compartmentalization and gene composition in the genome of vertebrates. *J. Mol. Evol.* 26: 198–204
85. Mouchiroud D, Gautier C, Bernardi G. 1988. The compositional distribution of coding sequences and DNA molecules in humans and murids. *J. Mol. Evol.* 27:311–20
86. Mouchiroud D, Gautier C, Bernardi G. 1995. Frequencies of synonymous substitutions in mammals are gene-specific and correlated with frequencies of nonsynonymous substitutions. *J. Mol. Evol.* 40:107–13
87. Musto H, Rodriguez-Maseda H, Bernardi G. 1994. The nuclear genomes of African and American trypanosomes are strikingly different. *Gene* 141:63–69
88. Musto H, Rodriguez-Maseda H, Bernardi G. 1995. Compositional properties of nuclear genes from *Plasmodium falciparum*. *Gene* 152:127–32
89. Novacek MJ. 1992. Mammalian phylogeny: shaking the tree. *Nature* 356:121–25
90. O'Brien SJ, Womack JE, Lyons LA, Moore KJ, Jenkins NA, Copeland NG. 1993. Anchored reference loci for comparative genome mapping in mammals. *Nat. Genet.* 3:103–12

90a. Ohta T. 1995. Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. *J. Mol. Evol.* 40:56–63

90b. Ohta T, Ina Y. 1995. Variation in synonymous substitution rates among mammalian genes and the correlation between synonymous and nonsynonymous divergences. *J. Mol. Evol.* In press

91. Oliver SG, Vanderaart QJM, Agostoni ML, Aigle M, Alberghina L, et al. 1992. The complete DNA sequence of yeast chromosome III. *Nature* 357:38–46
92. Pilia G, Little RD, Aissani B, Bernardi G, Schlessinger D. 1993. Isochores and CpG islands in YAC contigs in human Xq26.1-qter. *Genomics* 17:456–62
93. Pizon V, Cuny G, Bernardi G. 1984. Nucleotide sequence organization in the very small genome of a tetraodontid fish, *Arothron diadematus*. *Eur. J. Biochem.* 140:25–30
94. Razin SV, Gromova II, Iarovaia JV. 1995. Specificity and functional significance of DNA interaction with the nuclear matrix: new approaches to clarify the old question. *Int. Rev. Cytol.* In press
95. Royle NJ, Clarkson RE, Wong Z, Jeffreys AJ. 1988. Clustering of hypervariable minisatellites in the proterminal regions of human autosomes. *Genomics* 3:352–60
96. Rynditch A, Kadi F, Geryk J, Zoubak S, Svoboda J, Bernardi G. 1991. The isopycnic, compartmentalized integration of Rous sarcoma virus sequences. *Gene* 106:165–72
97. Sabeur G, Macaya G, Kadi F, Bernardi G. 1993. The isochore patterns of mammalian genomes and their phylogenetic implications. *J. Mol. Evol.* 37:93–108
98. Saccone S, De Sario A, Della Valle G, Bernardi G. 1992. The highest gene concentrations in the human genome are in T bands of metaphase chromosomes. *Proc. Natl. Acad. Sci. USA* 89:4913–17
99. Saccone S, De Sario A, Wiegant J, Raap

AK, Della Valle G, Bernardi G. 1993. Correlations between isochores and chromosomal bands in the human genome. *Proc. Natl. Acad. Sci. USA* 90: 11929–33

Saitoh Y, Laemmli UK. 1994. Metaphase chromosome structure: bands arise from a differential folding path of the highly AT-rich scaffold. *Cell* 76: 609–22

101 Salinas J, Matassi G, Montero LM, Bernardi G. 1988. Compositional compartmentalization and compositional patterns in the nuclear genomes of plants. *Nucleic Acids Res.* 16:4269–85

Salinas J, Zerial M, Filipski J, Bernardi G. 1986. Gene distribution and nucleotide sequence organization in the mouse genome. *Eur. J. Biochem.* 160:469–78

Salinas J, Zerial M, Filipski J, Crépin M, Bernardi G. 1987. Non-random distribution of MMTV proviral sequences in the mouse genome. *Nucleic Acids Res.* 15:3009–22

Scherthan H, Cremer T, Arnason U, Weier H-U, Lima-de-Faria A, Frönicke L. 1994. Comparative chromosome painting discloses homologous segments in distantly related mammals. *Nat. Genet.* 6:342–47

105 Schmid M, Guttenbach M. 1988. Evolutionary diversity of reverse (R) fluorescent chromosome bands in vertebrates. *Chromosoma* 97:101–14

106 Selig S, Okumura K, Ward DC, Cedar H. 1992. Delineation of DNA replication time zones by fluorescence in situ hybridization. *EMBO J.* 11:1217–25

107 Sharp PM, Lloyd AT. 1993. Regional base composition variation along yeast chromosome III: evolution of chromosome primary structure. *Nucleic Acids Res.* 21:179–83

108. Soriano P, Macaya G, Bernardi G. 1981. The major components of the mouse and human genomes. 2. Reassociation kinetics. *Eur. J. Biochem.* 115:235–39

109. Soriano P, Meunier-Rotival M, Bernardi G. 1983. The distribution of interspersed repeats is non-uniform and conserved in the mouse and human genomes. *Proc. Natl. Acad. Sci. USA* 80:1816–20

110. Tazi J, Bird A. 1990. Alternative chromatin structure at CpG islands. *Cell* 60:909–20

111. Thiery JP, Macaya G, Bernardi G. 1976. An analysis of eukaryotic genomes by density gradient centrifugation. *J. Mol. Biol.* 108:219–35

112. Weissenbach J, Gyapay G, Dib C, Vignal A, Morissette J, et al. 1992. A second-generation linkage map of the human genome. *Nature* 359:794–801

112a. Whitfield LS, Hawkins TL, Goodfellow PN, Sulston J. 1995. 41 Kilobases of analyzed sequence from the pseudoautosomal sex-determining regions of the short-arm of the human Y chromosome. *Genomics* 27:306–11

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

113a. Wu CI, Li WH. 1985. Evidence for higher rates of nucleotide substitution in rodents than in man. *Proc. Natl. Acad. Sci. USA* 82:1741–45

114. Zerial M, Salinas J, Filipski J, Bernardi G. 1986. Gene distribution and nucleotide sequence organization in the human genome. *Eur. J. Biochem.* 160:479–85

115. Zerial M, Salinas J, Filipski J, Bernardi G. 1986. Genomic localization of hepatitis B virus in a human hepatoma cell line. *Nucleic Acids Res.* 14:8373–86

116. Zirbel RM, Mathieu UR, Kurz A, Cremer T, Lichter P. 1993. Evidence for a nuclear compartment of transcription and splicing located at chromosome domain boundaries. *Chromosome Res.* 1: 93–106

117. Zoubak S, D'Onofrio G, Cacciò S, Bernardi G, Bernardi G. 1995. Specific compositional patterns of synonymous positions in homologous mammalian genes. *J. Mol. Evol.* 40:293–307

118. Zoubak S, Richardson JH, Rynditch A, Höllsberg P, Hafler DA, et al. 1994. Regional specificity of HTLV-I proviral integration in the human genome. *Gene* 143:155–63

119. Zoubak S, Rynditch A, Bernardi G. 1992. Compositional bimodality and evolution of retroviral genomes. *Gene* 119:207–13