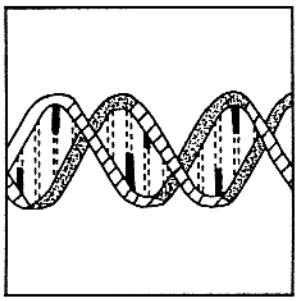

# Human Genome and Its Evolutionary Origin

GIORGIO BERNARDI, *Institut Jacques Monod*



## Glossary

**Recombination** Any process that gives rise to cells or individuals (recombinants) associating in new ways two or more hereditary determinants (genes) by which their parents differed

**Sister chromatids** Two chromatids derived from one and the same chromosome during its replication in interphase

**Translocation** Chromosomal structural change characterized by the change in position of chromosome segments (and the genes they contain)

THE BASIC QUESTION this article addresses is of the organization of nucleotide sequences in the human genome and the evolutionary origin of such organization. It shows that, far from being a "bean bag," the genome is highly ordered from the chromosomes down to the nucleotide level. Nucleotide sequences, whether in genes or in the very abundant nongenic segments, obey precise compositional rules. On the other hand, compositional rules also are evident in chromosome structure in that DNA composition is responsible for chromosome bands. These structural properties are associated with functional properties and shed a new light on genome evolution.

## I. Introduction

### A. The Genome

Every living organism contains, in its genome, all the genetic information that is required to produce its proteins and that is transmitted to its progeny. The genome consists of deoxyribonucleic acid (DNA), which is made up of two complementary strands wound around each other to form a double helix. The building blocks of each DNA strand are deoxyribonucleotides. These are formed by a phosphate ester of deoxyribose (a sugar), linked to one of four bases: two purines (adenine [A] and guanine [G]) and two pyrimidines (thymine [T] and cytosine [C]). In the DNA double helix, purines pair with pyrimidines (A with T, G with C) and the phosphates bridge the paired building blocks of the two strands to form the double helix. [*See* DNA AND GENE TRANSCRIPTION; GENOME, HUMAN.]

During cell replication, the two strands of the double helix are unwound, and a complementary copy of each is made, producing two identical copies (except for rare mistakes, or mutations) of the parental double helix. The two strands are also unwound at the time when one strand, the sense-strand carrying the genetic information, is copied into a complementary ribonucleic acid (RNA), which differs from the DNA master copy in having ribose instead of deoxyribose and uracil instead of thymine. RNA transcripts of genes are used as templates for the synthesis of proteins. The translation of each RNA transcript into the corresponding protein involves a very complex machinery that makes use of ribosomes (particles made up of two subunits, each containing a ribosomal RNA) and trans-

  

fer RNAs specific for different amino acids. It will suffice here to mention that subsequent sets of three adjacent nucleotides (or triplets, also referred to as codons) of the transcript specify amino acids that follow each other in the protein chain. Because there are 64 triplets (three of which are termination codons, marking the end of translation) and only 20 amino acids, all amino acids, except for methionine and tryptophan, are encoded by more than one codon. In other words, several synonymous codons may be used to specify the same amino acid; therefore, the genetic code is said to be degenerate, which means that alternative possibilities (synonymous codons) exist for the same amino acid. Differences among synonymous codons are mainly at the third codon positions.

The genome of living organisms greatly differs in size, from 4.2 Mb (megabases, or millions of base pairs [bp]) for a typical bacterium such as *Escherichia coli* to 3,000 Mb or 3 Gb (gigabases, or billions of bp) for a eukaryote such as humans. While prokaryotes (bacteria) are characterized by small genome sizes, clustering around the value given above for *E. coli*, eukaryotes exhibit larger genome sizes that cover a wide range—from 20 Mb for the yeast *Saccharomyces cerevisiae* to 3 Gb for mammals.

The much larger genome size of higher eukaryotes is not only due to the presence of a larger number of genes (see below). In fact, the increase in size is mainly due to noncoding sequences that are both intergenic and intragenic. The latter sequences, called introns, separate different coding stretches, or exons, of most eukaryotic genes. The intron part of the primary transcript is eliminated by splicing; leaving the mature transcript or messenger RNA that encodes a given protein.

Eukaryotes differ from prokaryotes in other respects as well. They have a nucleus that is separated from the cytoplasm by a nuclear membrane. In addition to the nuclear genome, eukaryotic cells also have organelle genomes, which are located in mitochondria and, in the case of plants, also in chloroplasts. Organelle genomes contain a very limited yet an essential amount of genetic information (genes) encoding organelle-specific proteins and RNAs. Organelle genomes apparently originated from symbiotic bacteria, which entered procukaryotic cells. Like the bacterial genomes, organelle genomes are physically organized in a rather simple way. In contrast, in the nuclear genome of eukaryotes DNA is wrapped around histone octamers to form nucleosomes, which are packaged into chromatin fibers. In turn, these fibers are folded into chromatin loops, consisting of 30–100 Kb (kilobases, thousands of bp) of DNA, which are, in turn, packaged into chromosomes.

### B. The Human Genome

The nuclear genome of humans consists of about 3 billion base pairs, whereas the mitochondrial genome is only 16,000 bp. Estimates of the number of (nuclear) human genes range from 30,000 to 100,000. If coding sequences average 1,000 bp, they would represent 1–3% of the human genome; 97–99% of which, therefore, is made up of noncoding sequences. It should be noted that the larger number of genes in humans compared with bacteria is mainly due to the fact that many human genes exist as multigene families, the result of gene duplications during evolution.

Our present knowledge of human coding sequences, in terms of primary structures (or nucleotide sequences), is limited to about 1,000 of them. In other words, we only know about 1–3% of our coding sequences, or 0.03% of our genome. Our knowledge of noncoding sequences in terms of primary structure is even more limited. Nevertheless, we know that a sizable part of intergenic noncoding sequences are formed by repeated sequences that belong in several families. Two important families are called LINES and SINES (the long and short interspersed sequences), which are present in about 100,000 and 900,000 copies, respectively. These families of repeated sequences have been largely studied by reassociation kinetics. This experimental approach is based on the fact that, if small DNA fragments (having a size of 300–400 bp) are separated into their complementary strands, the reassociation of the latter proceeds at a rate that depends on the frequency in the genome of the sequences present in the fragments. Single strands from sequences that are present very many times in the genomes will find their complementary strands faster than single strands from genes that are present a few times or even only once in the genome. This technique allows estimating the relative amounts of repetitive sequences and single-copy sequences, the latter being present only once or a small number of times in the genome.

The other level of knowledge of the human genome concerns chromosomes. Each human germ cell (sperm or oocyte) contains 23 chromosomes. In

haploid cells, 22 chromosomes (1–22 in order of decreasing size) are autosomes, which are identical in both sexes. The 23rd chromosome, the sex chromosome, is an X chromosome in females and a Y chromosome in males. Somatic cells are diploid; they have two haploid chromosome sets. Female diploid cells have two X chromosomes (one of which is inactive), whereas males have one X and one Y chromosome. During mitosis, chromosomes condense and, at metaphase, they are characterized by specific staining properties. G-bands (Giemsa-positive, or Giemsa dark bands; equivalent to Q-bands, or Quinacrine bands) and R-bands (reverse bands; equivalent to Giemsa-negative or Giemsa light bands) are produced by treating metaphase chromosomes with fluorescent dyes, proteolytic digestion, or differential denaturing conditions. [*See* CHROMOSOMES.]

Under standard conditions, Giemsa staining produces a total of about 400 bands that comprise, on the average, 7.5 Mb of DNA. If staining is applied to prophase chromosomes, which are more elongated, up to 2,000 bands can be visualized. At this high resolution, one chromosomal band contains, on the average, 1.5 Mb of DNA. Prophase chromosomes can also be studied at meiosis, the process leading to haploid germ cells. Staining meiotic chromosomes (usually at pachytene) produces results that are similar to those just mentioned. Indeed, a pattern of chromatin condensations, the chromomeres, are visualized. Chromomeres and the interchromomeres separating them correspond to high-resolution Giemsa-positive and Giemsa-negative bands, respectively. A number of approaches, ranging from purely genetical to molecular ones have allowed assigning genes not only to individual chromosomes but also to chromosome bands.

## II. Isochores and Genome Organization

Our present knowledge of the human genome is clustered around two levels of organization: genes, which are in a DNA size range of a few Kb, and chromosome bands, which are in a size range of a few Mb. We know very little about the intermediate size range.

Recent discoveries, however, not only have linked the gene level with the chromosome level of genome organization but have also shed light on the functional and evolutionary implications of genome organization. Indeed, the human genome is made up of isochores, large DNA domains (>300 Kb) that are compositionally homogeneous and belong to a small number of families ranging from 35 to 55% GC (see Fig. 1; GC is the molar percentage of the two complementary bases G and C in DNA). This point was demonstrated in two ways. (1) The relative amounts of isochore families, as judged by fractionating human DNA fragments according to their base composition, is independent of molecular size between 3 and >300 Kb. The fact that breaking down DNA fragments to small sizes does not change the relative ratios of the compositional families obviously indicates compositional homogeneity in the larger fragments. (2) Hybridization of single-copy sequences to compositional fractions of DNA fragments about 100 Kb in size occurs within a very narrow GC range, about 1%. This means that all DNA fragments that carry a given gene are extremely close in composition despite the fact that they were derived from intact chromosomal DNA by a random breakdown due to the unavoidable physical and enzymatic degradation that occurs during DNA preparation.

The discovery of an isochore organization in the human genome (and in the eukaryotic genomes in

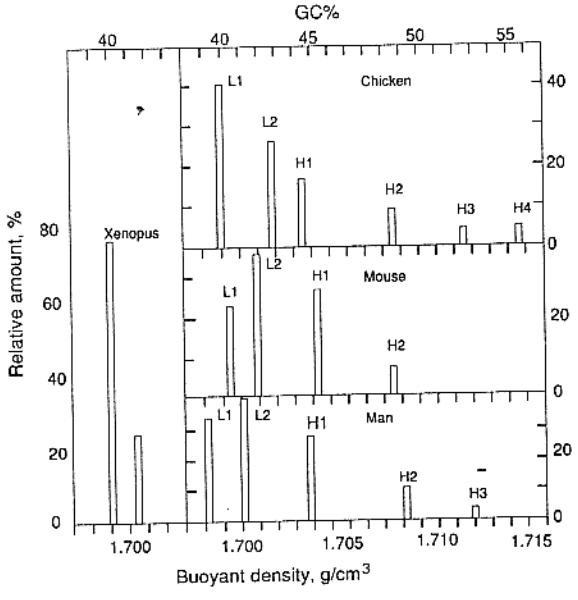


**FIGURE 1** Histograms showing the relative amounts and modal buoyant densities in CsCl and GC levels of the major DNA components (L1, L2, H1, H2, H3, H4) from *Xenopus*, chicken, mouse, and man. Satellite (clustered repeated sequences) and minor components (like ribosomal DNA) are not shown. [Reproduced, with permission, from G. Bernardi (1989). The isochore organization of the human genome. *Annu. Rev. Genet.* **23**, 637–661.]

general) was accompanied by the discovery of compositional correlations (1) between genes and the isochores containing them, and (2) between chromosomal bands and isochores. These correlations will be discussed in the two following sections.

### A. Isochores and Genes

The localization of a number of genes in compositional fractions of human DNA has shown a very important point, namely that a linear relationship exists between GC levels of coding sequences (and of their first, second, and third codon positions) and GC levels of the DNA fragments (about 100 Kb in size) containing them (see Fig. 2). Because the DNA fragments are mainly composed of noncoding intergenic sequences, the compositional correlation just described is in fact a correlation between coding sequences and the noncoding sequences that embed them. This suggests that both coding sequences representing <5% of the genome and the flanking noncoding sequences representing >95% of the genome are subject to the same compositional constraints.

The existence of this compositional relationship is also important from another viewpoint. Indeed, it allows us to know the genomic distribution of genes over isochores having a different GC content. This can be deduced from histograms showing the compositional distribution of coding sequences or of their different codon positions. In fact, because the coding sequences are compositionally correlated with the vast DNA regions surrounding them, the composition of coding sequences themselves indicates the location of these genes in different isochore families. For instance, a coding sequence that is GC-rich will be present in a GC-rich isochore, whereas a gene that is GC-poor is located in a GC-poor isochore. If one examines the composition of all human coding sequences studied so far (or, even better, of their third codon positions, which are free to change with less alteration of the corresponding amino acids), one discovers (Fig. 3) a strong predominance of GC-rich coding sequences, which are mainly located in the GC-richest isochores. These results point to an extremely nonuniform distribution of genes in the human genome because the GC-richest isochores are the least represented in the human genome.

Very interestingly, this gene distribution is mimicked (1) by the distribution of CpG doublets, the only potential sites of methylation (CpG doublets in GC-poor genes are underrepresented relative to statistical expectations, whereas in GC-rich genes, CpG doublets have the statistical frequency), and (2) by the distribution of CpG islands, which are sequences >0.5 Kb in size and characterized by high GC levels, by clustered unmethylated CpG's, by G/C boxes (e.g., GGGGCGGGGC or closely related sequences), and by clustered sites for rare-cutting restriction enzymes (which recognize GC-rich sequences comprising one or two unmethylated CpG doublets). CpG islands are associated with the 5′ flanking sequences, exons, and introns of all housekeeping genes (namely of the genes whose activity is required by every cell) and many tissue-specific genes (namely of genes that are expressed only in some specific tissues, such as liver cells, and with the 3′ exons of some tissue-specific genes.

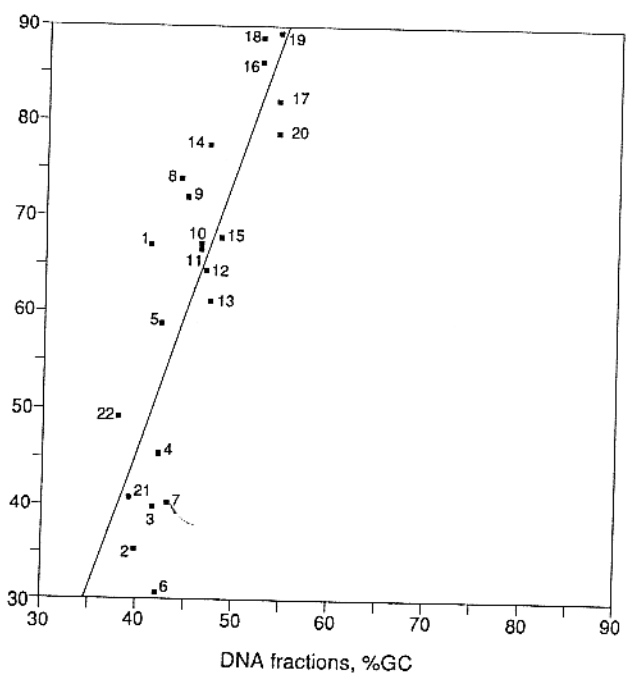


**FIGURE 2** Plot of GC levels of third codon positions against the GC levels of compositional DNA fractions in which they were localized. The numbers indicate different coding sequences.

### B. Isochores and Chromosome Bands

A number of lines of evidence show that GC-poor and GC-rich isochores largely correspond to the DNA of G- and R-bands, respectively. However, G- and R-bands not only differ in their overall isochore makeup, but also in their internal isochore

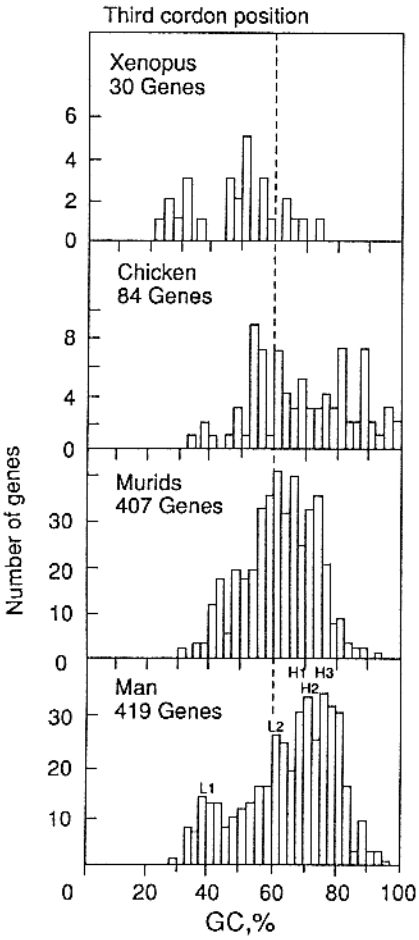


**FIGURE 3** Compositional distribution of third codon position from vertebrate genes. (This distribution is the most informative because of its wider spread compared with coding sequences and first or second codon positions.) The number of genes under consideration is indicated. A 2.5% GC window was used. The vertical broken line at 60% GC is shown to provide a reference. Approximate identifications of different compositional classes of coding sequences corresponding (see Fig. 1) to the major components of the human genome (L1, L2, H1, H2, and H3) are indicated. The borders between L1-L2, H1-H2, and H3 can be tentatively estimated as 67.5% and 77.5% GC, respectively. [Reproduced, with permission, from G. Bernardi (1989). The isochore organization of the human genome. *Ann. Rev. Genet.* **23**, 637–661.]

structure as indicated by several lines of evidence, and, more recently, by the compositional mapping of the long arm of chromosome 21. Compositional maps may be constructed whenever long-range physical maps are available by assessing GC levels around landmarks (localized on the physical map) that can be probed. This simply requires the hybridization of the probes on DNA fractionated according to base composition, because this establishes the GC level of vast regions (≥200 Kb) around the sequence probed. This approach has provided a direct demonstration for the compositional homogeneity of G-bands, for the compositional heterogeneity of R-bands, and for the highest GC levels and highest gene concentrations in the telomeric region of the long arm of chromosome 21 (a similar situation is possibly present in the telomeric regions of many other chromosomes).

## III. Isochores and Genome Functions

While the functional correlates of the isochore organization of the human genome are still largely open problems, a number of points, as outlined below, are already well established.

### A. Isochores and Integration of Mobile and Viral DNA Sequences

Stable integration of mobile and viral DNA sequences is mostly found in isochores of matching composition. Mobile sequences that have been amplified by retrotranscription and translocated to numerous loci of the human genome during mammalian evolution, such as LINES and SINES (see Section I) are predominantly located in isochores of matching GC levels. This indicates that reinsertion is targeted to matching genome environments and/or that integration is more stable within such environments. Incidentally, such reinsertion may be a cause of mutation, if it occurs in genes that are thereby disrupted.

Needless to say, these observations are of interest in connection with the integration of foreign DNA into the genome of transgenic mammals. In this case, the important, yet unresolved, question concerns the effect of genomic compositional context on the expression of integrated sequences.

### B. Isochores, Translocation Breakpoints, and Fragile Sites

Translocation breakpoints are not randomly located on chromosomes. R-bands and G/R borders are the predominant sites of exchange processes, including spontaneous translocations, spontaneous or induced sister-chromatid exchanges, and the chromosomal abnormalities seen after X-ray and chemical damage. Likewise, fragile sites tend to be more fre-

quent in R-bands or near the border of R- and G-bands. Moreover, cancer-associated chromosomal aberrations are also nonrandom, because a limited number of genomic sites are consistently involved and frequently associated with cellular oncogenes or fragile sites. These observations indicate that R-bands and G/R borders are particularly prone to recombination. They also raise the question of the role played in these phenomena by the compositional discontinuities at G/R borders and within R bands as well as by the genomic distribution of SINES, CpG islands, and other recombinogenic sequences, such as minisatellites, which are predominantly located in R-bands.

Chromosomal rearrangements have two important consequences: the activation of oncogenes by strong promotors that could be put upstream of them by the rearrangement, and the formation, in evolutionary time, of reproductive barriers and speciation.

### C. Other Functional Aspects of Isochores

Isochores are related to replication and condensation timing in the cell cycle. G-bands and genes located in GC-poor isochores replicate late in the cell cycle, whereas R-bands and genes located in GC-rich isochores replicate early. In contrast, condensation timing during mitosis occurs early in the cell cycle for G-bands and late for R-bands.

The distribution of genes over the isochores also has some correlation with gene function. Housekeeping genes (including oncogenes) are preferentially distributed in GC-rich isochores and R-bands, whereas tissue-specific genes are more abundant in GC-poor isochores and G-bands.

The range of GC values (30–100%) in codon third positions of human genes practically is as wide as that exhibited by the genes of all prokaryotes. This very extended range implies very large differences in codon usage (namely in the differential use of synonymous codons for the same amino acid) for GC-poor and GC-rich genes present in the same genome. In particular, at the high-GC end of the range, an increasing number (up to 50%) of codons are simply absent. In turn, GC values in codon third positions are paralleled (although, expectedly, to a more limited extent) by the GC values in first and second positions. The range of the latter values leads to very significant differences in the frequency of certain amino acids in GC-rich and GC-poor genes. For instance, the ratio of alanine+arginine to serine+lysine (namely, of the two amino acids that contribute most to the thermodynamic stability of proteins over the two that do so the least) increases by a factor of four between proteins encoded by the GC-poorest and the GC-richest coding sequences in the human genome.

The distributions of CpG doublets and of CpG islands suggests that the distribution of methylation in the genome of humans and other vertebrates is highly nonuniform, a point of interest in view of the role of DNA methylation in gene function and of the distribution of housekeeping and tissue-specific genes.

The results on CpG islands have an additional functional relevance. In GC-poor isochores, genes are usually endowed with a TATA or a CCAAT box and an upstream control region, whereas in the GC-rich isochores there is no TATA box but promotors containing properly positioned "G/C boxes" that bind transcription factor Spl, a protein that activates RNA polymerase II transcription. These GC-rich promotors apparently are associated with all genes located in GC-rich isochores.

## IV. Isochores and Genome Evolution

As shown in Figures 1 and 3, the compositional patterns of the human genome are characterized by a wide GC range of both isochore families and coding sequences and by a predominance of GC-rich genes. These patterns are found in all mammals explored so far, with only slightly narrower distributions in three families of myomorpha (a suborder of rodents that includes mouse, rat, and hamster). Very similar, yet not identical, patterns are found in birds, a vertebrate class that arose from reptiles independently of mammals. In the case of birds, the compositional distributions of both DNA fragments and coding sequences attain, however, higher GC values than in mammals (see Figs. 1 and 3). In sharp contrast, the genomes of the vast majority of cold-blooded vertebrates exhibit compositional patterns that are narrower and do not reach the GC levels of the GC-rich DNA fragments of warm-blooded vertebrates or of the coding sequences contained in them (see Figs. 1 and 3).

The compositional patterns of vertebrate genomes define two modes of genome evolution. In the conservative mode, which prevails in mammals (and probably in birds), the composition of DNA

fragments and coding sequences is maintained during evolution despite a very high degree of nucleotide divergence (which may be >50% in third codon positions). This compositional conservation appears to require negative selection, operating at the isochore level, to eliminate any strong deviation from a presumably functionally optimal composition. It is clear that if, during mammalian evolution, a number of genes retain a GC level of >90% in their third codon position, the mutations that would have decreased this level must somehow have been eliminated.

In contrast, in the transitional or shifting mode, parallel compositional changes are seen in both isochores and coding sequences. The two major compositional transitions found in vertebrate genomes are the GC increases that occurred between the genomes of reptiles on the one hand and those of birds and mammals on the other. (Apparently, these increases were accompanied by the replacement of TATA and CCAAT boxes by GC-rich promotors.) These compositional changes are due to a directional fixation of point mutations largely caused by both negative and positive selection at isochore levels. Selection appears to be for the higher thermal stabilities of proteins, RNA and DNA, which are required by the higher body temperature of warm-blooded vertebrates. Of course, selection implies functional differences and, therefore, supports the idea that isochores are functionally relevant structures. Moreover, the compositional relationships between coding and noncoding (particularly intergenic) sequences indicate that the same compositional constraints apply to both kinds of sequences. The selection pressures underlying such constraints cannot be understood if noncoding sequences are "junk DNA," with no biological functions.

## V. Conclusions

Isochores represent a new structural level in the organization of the genome of warm-blooded vertebrates that bridges the enormous size gap between the gene level, including both their exon–intron systems and the corresponding regulatory sequences, and the chromosome level, with its banding patterns. These three levels are correlated with each other, because genes match compositionally the isochores in which they are harbored, while GC-poor and GC-rich isochores are DNA segments located in G- and R-bands, respectively.

The investigations that led to the discovery of isochores and of these two correlations have firmly established the existence of differences in the base composition and gene concentration of DNA segments present in G- and R-bands. Moreover, they have revealed that these segments are characterized by strikingly different complexities. The isochores present in G-bands are GC-poor, very close in composition, and characterized by a low gene concentration, whereas the isochores present in R-bands belong to different GC-rich compositional families, including those of the GC-richest family that have the highest concentration of genes and CpG islands. R-bands also comprise, however, a number of GC-poor isochores (as shown by the compositional mapping of chromosome 21) that may correspond to internal "thin" G-bands only seen at high resolution. In conclusion, isochores correspond to a chromosome organization level lower than standard chromosomal bands, possibly to chromomeres and interchromomeres.

While isochores from the genomes of warm-blooded vertebrates belong to a number of families characterized by large differences in base composition, this is not true for cold-blooded vertebrates, in which case isochores are characterized by much smaller differences in composition, which correspond to much weaker banding patterns in metaphase chromosomes.

Isochores are, however, not only structural units but also appear to play functional roles. Some of these, like the integration of mobile and viral sequences or recombination and chromosome rearrangements, are well established. The observations on the gene distribution of the genome, the relationships of such distribution with gene functions (housekeeping, tissue-specific), with codon usage, and with different kinds of regulatory sequences are also indicative of functional roles for isochores. In contrast, DNA replication timing and chromosome condensation timing at mitosis seem to be rather correlated with the chromomere–interchromomere organization of chromosomes, independently of the composition of the corresponding DNA stretches, because they are also found in cold-blooded vertebrates.

Isochores are evolutionary units of vertebrate genomes. Their composition may be conserved in spite of enormous numbers of point mutations or may undergo dramatic changes after more modest numbers of point mutations. In the case of the two independent evolutionary transitions from cold-

**TABLE I** The Human Genome

| Paleogenome (G-bands, GC-poor isochores) | Neogenome (R-bands, (GC-rich isochores) |
|---|---|
| Chromomeres | Interchromomeres |
| Late replication | Early replication |
| Early condensation | Late condensation |
| Abundance of LINES | Abundance of SINES |
| Compositional homogeneity | Compositional heterogeneity |
| Scarcity of genes | Abundance of genes (esp. in H3) |
| GC-poor gene (esp. tissue-specific) | GC-rich genes (esp. housekeeping) |
| Scarcity of CpG islands | Abundance of CpG islands |
| TATA box promotors | G/C box promotors |
| Less frequent recombination | More frequent recombination |

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

blooded vertebrates to mammals and birds, the compositional transitions occurring in the genome seem to be largely associated with the optimization of genome functions following environmental body temperature changes. Interestingly, these transitions appear to be accompanied by very conspicuous changes in promotor sequences.

In conclusion, two large compositional compartments can be distinguished in the human genome and, more generally, in the genomes of warm-blooded vertebrates (see Table I). The first compartment, the paleogenome, is characterized by its similarity to what it was, and still is, in cold-blooded vertebrates: the late-replicating, compositionally homogeneous, GC-poor isochores of early-condensing chromomeres contain relatively rare, GC-poor (largely tissue-specific) genes having TATA box promotors (CpG islands are scarce). The second compartment, the neogenome, is characterized by the fact that it changed its compositional features compared with what it was in cold-blooded vertebrates. In the neogenome, the ancestral, early-replicating, GC-poor isochores of late-condensing interchromomeres were changed into compositionally heterogeneous, GC-rich isochores that contain abundant genes (perhaps including most housekeeping genes) having G/C box promotors (genes and CpG islands are particularly abundant in the GC-richest isochores).

## Bibliography

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

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.

Gardiner, K., Aissani, B., and Bernardi, G. (1990). A compositional map of human chromosome 21. *EMBO J.* (in press).