Isochores

Advanced article

Giorgio Bernardi, *Stazione Zoologica Anton Dohrn, Naples, Italy*

Chromosomes of warm-blooded vertebrates are mosaics of isochores. These are long deoxyribonucleic acid segments that are fairly homogeneous in base composition. In the human genome, isochores can be assigned to five families characterized by increasing levels of guanine and cytosine and by increasing gene densities.

Article contents

doi: 10.1038/npg.els.0005003

Sequence Organization of the Human Genome

The organization of complex eukaryotic genomes such as the human genome is a long-standing problem. Bacterial genomes are small, 2–5 Mb (megabase pairs, 1 million base pairs (bp)); coding sequences represent about 85% of the genome and one gene corresponds, on average, to 1 kb (kilobase pairs, 1000 bp) of deoxyribonucleic acid (DNA). In contrast, eukaryotic genomes cover a very broad size spectrum, ranging from 12 Mb in yeast to 3200 Mb in humans. (Much larger genomes are found in some other eukaryotes, in which the genome has been amplified by polyploidization.) In yeast 6000 genes represent 70% of the genome, and one gene corresponds, on average, to 2 kb of DNA; in the human genome 30 000 or so coding sequences represent only about 1% of the genome (assuming an average coding sequence size of 1000 nucleotides) and one gene corresponds, on average, to 100 kb. Obviously, apart from some exceptional cases, genes are much smaller than this theoretical average value, essentially because of the large amounts of noncoding DNA forming the intergenic regions. The latter comprise interspersed repeated sequences derived from the insertion and amplification (by many rounds of duplication) of transposons, namely mobile sequences that have invaded the eukaryotic genome at some point in evolution. The two major families of such repeated sequences are long interspersed nuclear elements (LINEs: 6–8 kb in size, 850 000 copies, 21% of the genome) and short interspersed nuclear elements (SINEs: 100–300 bp in size, 1 500 000 copies, 13% of the genome).

Compositional Approach

In spite of the complexity of the problem under consideration here, we now have a good understanding of the organization of mammalian genomes, thanks to a molecular approach based on the most elementary property of DNA, namely its base composition. This approach was initially applied to DNA molecules and later, when they became available, to DNA sequences. Previous attempts based on DNA reassociation kinetics, as analyzed by separating single- and double-stranded DNA on hydroxyapatite columns, could not go beyond the important, yet limited, finding of the existence of repeated sequences in eukaryotic genomes.

We will now consider how base composition varies along the chromosomes of different genomes. In bacteria, DNA preparations are formed by fragments derived from the corresponding circular chromosomes and show a certain degree of compositional heterogeneity at molecular sizes of 50–100 kb. This degree of heterogeneity is not the same for different bacterial genomes, but even in the most heterogeneous cases, heterogeneity is low compared with that of mammalian DNAs. In the latter case, even if one neglects satellite DNAs (which are formed by short tandem repeats), the genome is strikingly heterogeneous. For example, in the case of the human genome, DNA molecules in the 50–100 kb size range cover a 30–60% GC spectrum (GC being the molar ratio (percentage) of guanine plus cytosine in the DNA), not far from the whole spectrum (25–70% GC) covered by all bacterial genomes.

Fractionation methods based on density gradient centrifugation to equilibrium in the presence of sequence-specific DNA ligands have shown that in the human genome, DNA fragments of 50–100 kb can be partitioned into a small number of fairly homogeneous families, in fact as homogeneous as bacterial DNAs.

Isochores

Since the relative amounts of families of DNA molecules from vertebrate DNAs do not vary with increasing fragment size (up to over 300 kb), it was concluded that the DNA fragments of each family derive from regions much longer than 50–100 kb and that chromosomes are mosaics of compositionally similar DNA regions, termed isochores (for compositionally ‘equal landscapes’). In the human genome, the isochore pattern is characterized by two GC-poor, 'light' isochore families, L1 and L2, which represent about 30% and 33% of the genome, and by three GC-rich, 'heavy' isochore families, H1, H2 and H3, which represent about 24%, 7% and 5% respectively, of the genome (**Figure 1**). The remaining DNA corresponds to satellite and ribosomal sequences. It should be stressed that the isochore pattern of human DNA is a good representative of the patterns of mammals and birds, and that the patterns of fishes and amphibians are characterized by a narrower compositional range, because their GC-rich isochores are less GC-rich and less abundant than those of the DNA of warm-blooded animals. The transition between these two patterns is discussed elsewhere in this book. (***See*** Evolutionary History of the Human Genome; Genome Organization of Vertebrates.).

The isochore pattern is not, however, the only compositional pattern of a genome. Indeed, another type of compositional pattern is that of coding sequences. In this case, their GC levels or, more informatively, their GC_3 levels, the GC levels of third-codon positions, define the pattern. **Figure 2** presents the compositional pattern of human coding sequences.

An obvious question is whether there is any correlation between the compositional patterns of coding sequences (which may represent as little as 1% of the genome in vertebrates) and the compositional patterns of DNA fragments (99% of which may be formed by intergenic sequences and introns). Another question is whether there is any correlation within genes between the base composition of exons and that of introns. The answer to both questions is yes. Needless to say, the 'genome equations', linking coding and noncoding sequences and amounting to a 'genomic code' (Bernardi, 2000, 2001), provide a strong evidence against noncoding DNA being 'junk DNA'.

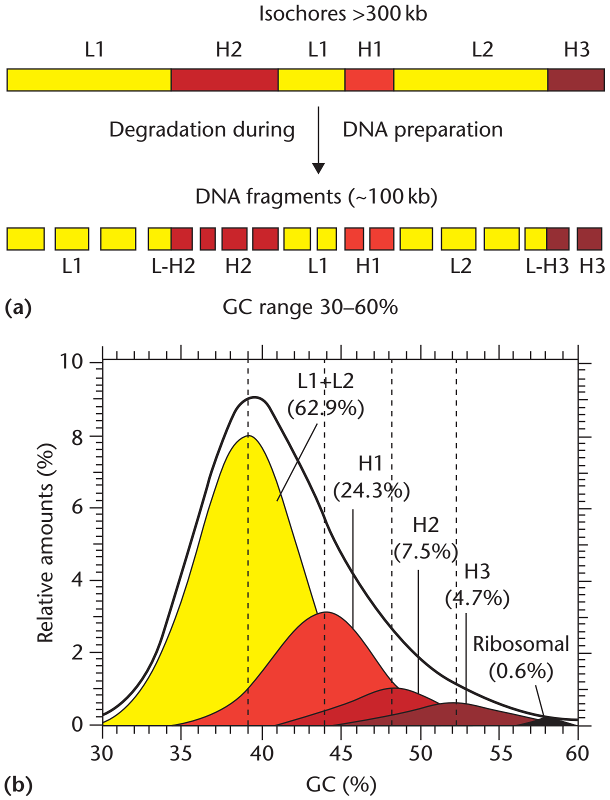

Figure 1 (a) Scheme of the isochore organization of the human genome. This genome, which is a good representative of the genome of mammals and birds, is a mosaic of large DNA segments, the isochores, which are compositionally fairly homogeneous and can be partitioned into a small number of families, 'light' or GC poor (L1 and L2) and 'heavy' or GC rich (H1, H2 and H3). Isochores are degraded during DNA preparation to fragments of approximately 100 kb in size. The GC range of these DNA molecules from the human genome is extremely broad, 30–60%. (From Bernardi, 2000.) (b) The CsCl profile of human DNA is resolved into its major DNA components, namely the families of DNA fragments derived from the isochore families (L1, L2, H1, H2, H3). Modal GC levels of isochore families are indicated on the abscissa (red). The relative amounts of major DNA components are indicated. Satellite DNAs are not represented. (From Zoubak *et al*., 1996.)

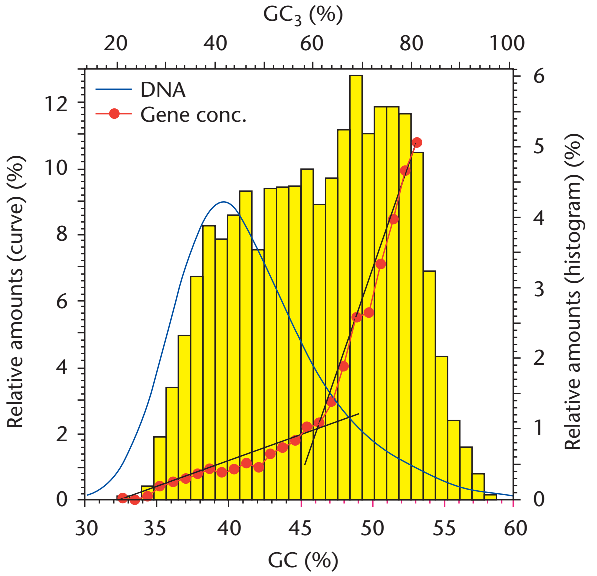

Figure 2 Profile of gene concentration (full circles) in the human genome, as obtained by dividing the relative numbers of genes in each 2% GC_3 interval of the histogram of gene distribution (bars) by the corresponding relative amounts of DNA deduced from the CsCl profile (full curve). The positioning of the GC_3 histogram relative to the CsCl profile is based on the correlation between GC_3 and the GC level of the isochores in which the genes were embedded. (Modified from Bernardi, 2000.)

Gene Distribution

The linear correlation between GC_3 levels of coding sequences and GC levels of isochores is important, not only because it indicates that the compositional constraints affect the genome as a whole, as just mentioned, but also because it allows the positioning of the distribution profile of coding sequences relative to that of DNA fragments (namely the CsCl profile). In turn, this permits estimation of the relative gene density by dividing the percentage of genes located in given GC intervals by the percentage of DNA located in the same intervals.

Since it had been tacitly assumed that genes were uniformly distributed in eukaryotic genomes, it came as a big surprise that gene distribution in the human genome (and, for that matter, in the genomes of all vertebrates) is strikingly nonuniform (**Figure 2**), gene concentration increasing from a very low average level in L1 isochores up to a level about 20-fold higher in H3 isochores.

The existence of a break in the slope of gene concentration at 60% GC_3 of coding sequences and at 46% GC of isochores (see **Figure 2**) defines two 'gene spaces' in the human genome (Bernardi, 2000, 2001). In the 'genome core', formed by isochore families H2 and H3 (which make up about 12% of the genome), gene concentration is very high, while in the 'empty quarter' (a term derived from the classical name for the Arabian desert), formed by isochore families L1, L2 and H1 (which make up about 88% of the genome), gene concentration is very low. It should be noted that the existence of the two genome spaces has been confirmed by the draft human genome sequence published by the International Human Genome Sequencing Consortium (IHGSC) (2001). Indeed, the latter confirmed not only the gradient of gene concentration paralleling the GC concentration, but also showed (**Figure 3**) the two different slopes of **Figure 2**. Moreover, it confirmed earlier results indicating that the genes located in the GC-poor empty quarter were characterized by long introns and those from the GC-rich genome core by short introns (**Figure 4**). These features correspond to very different transcriptional regulations, alternative splicing being frequent in the genes with long introns and rare or absent in those with short introns. Very roughly, about half of human genes are located in the small genome core, the other half being located in the large empty space.

The two gene spaces are characterized by a number of other different structural and functional properties. Indeed, most genes located in the genome core are associated with CpG islands, are actively transcribed, are replicated early in the cell cycle, are located in the distal parts of chromosomes and correspond to an open chromatin structure, whereas the genes of the empty quarter are endowed with the opposite properties. The open chromatin structure is characterized by the scarcity or absence of histone H1, acetylation of histones H3 and H4 and a larger nucleosome spacing (Tazi and Bird, 1990). Interestingly, the two major classes of interspersed repeated sequences show a different distribution in the human genome, the concentration of LINEs being highest in GC-poor isochores and that of SINEs in GC-rich isochores.

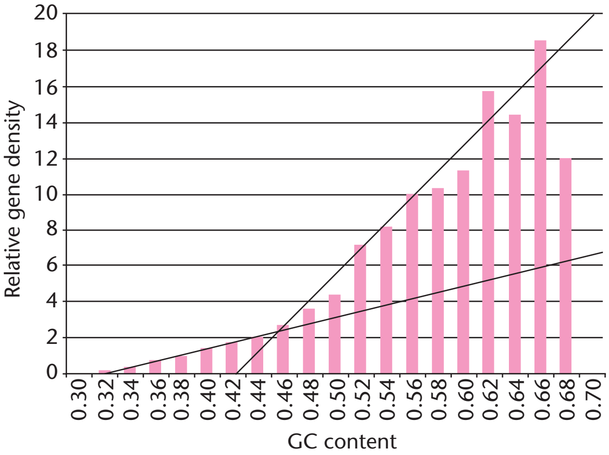

Figure 3 Gene density as a function of GC content. The local GC content was calculated for 9315 known genes mapped to the draft genome sequence. The two slopes added here to the original figure concern the genes from GC-poor and GC-rich isochores respectively. (From Bernardi (2001); modified from Figure 36b of International Human Genome Sequencing Consortium (2001).)

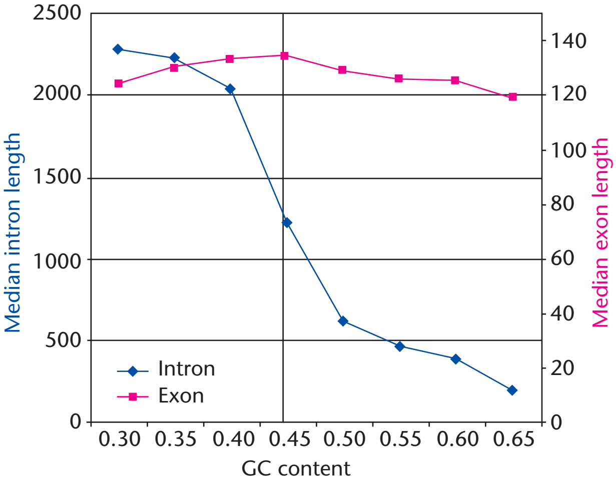

Figure 4 Dependence of mean exon and intron lengths on GC content. The vertical straight line added to the original figure marks the midpoint of the transition. (From Bernardi (2001); modified from Figure 36c of International Human Genome Sequencing Consortium (2001).)

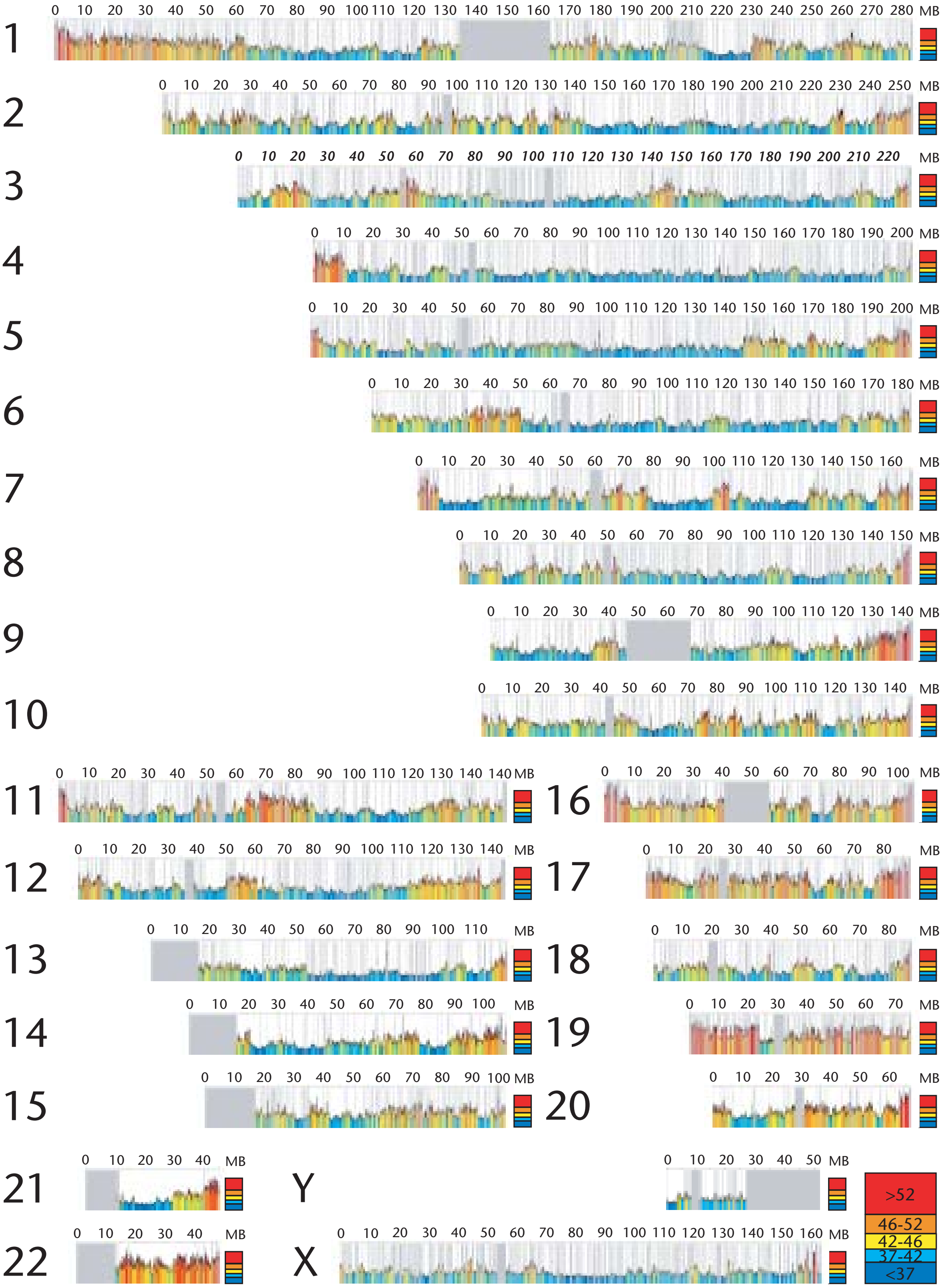

Figure 5 A color-coded compositional map of the chromosomes of the human genome, representing 100 kb moving window plots that scan the draft human genome sequence of International Human Genome Sequencing Consortium (2001). Color codes span the spectrum of GC levels in five steps, from ultramarine blue (GC-poorest isochores) to scarlet red (GC-richest isochores). The gray bars correspond to 5000 gaps present in the euchromatic regions of most chromosomes. (Modified from Pavliček *et al*., 2002.)

Compositional Features of Isochores

Very recently, the International Human Genome Sequencing Consortium (2001) studied the draft genome sequence to see whether 'strict isochores' could be identified, concluding that their results rule out a strict notion of isochores as compositionally homogeneous and that isochores do not appear to merit the prefix 'iso'. These conclusions deserve three comments (Bernardi, 2001).

1. 'Strict isochores' as defined by the IHGSC are indistinguishable from random DNA sequences in which nucleotides are independent and uncorrelated with each other. It is not surprising, therefore, that strict isochores could not be identified in the human genome. Indeed, 'strict isochores' or random sequences simply do not exist in any natural DNA. This can be easily understood by considering that a coding sequence can never satisfy the condition of nucleotide independence, because of the very existence of codons and of the compositional correlations that hold among different codon positions. Noncoding sequences, which represent the vast majority of complex eukaryotic genomes such as the human genome, cannot satisfy the condition of independence either, because they are compositionally correlated with the coding sequences that they embed. Finally, interspersed repeats (such as SINEs and LINEs) cannot satisfy the condition because they have their own specific sequences.
2. 'Strict isochores' are characterized by extremely small standard deviations of GC level. In contrast, families of DNA fragments from real isochores, such as those in the human genome, exhibit relatively large standard deviations. These are, however, much lower than those of total nuclear DNA and are in the same range shown by bacterial DNAs having the same size and the same GC level. Since bacterial DNAs are the most homogeneous among natural DNAs (with the exception of satellite DNAs), although more heterogeneous than random DNAs, the original definition of isochore families as 'fairly homogeneous' (Cuny *et al*., 1981) still seems to be an appropriate one.
3. As far as the mosaic organization of isochores is concerned, this can easily be seen in human genome sequences already using an overlapping 100-kb window analysis (**Figure 5**), in spite of the fact that such analysis averages out all compositional discontinuities corresponding to isochore borders.

A final remark concerns the denial of the very existence of isochores. While a mistake in itself, the major problems are its consequences which, apparently, were not realized by the IHGSC authors. The first one is the tacit denial of a compositionally discontinuous sequence organization and a return to the continuous compositional spectrum for the human genome that was the predominant view until the early 1970s. The second consequence is the denial of an important level of genome organization, insofar as gene density, intron size and patterns of codon usage, as well as the distribution of different classes of repetitive elements, replication timing, recombination frequency, chromosomal banding and stability and transcription of integrated sequences are all correlated with GC content. In other words, the second consequence is the denial of what has been called 'a fundamental level of genome organization' (Eyre-Walker and Hurst, 2001).

See also

Evolutionary History of the Human Genome
GC-rich Isochores: Origin
Gene Distribution on Human Chromosomes
Genome Organization of Vertebrates
L Isochore Map: Gene-poor Isochores

References

Bernardi G (2000) The compositional evolution of vertebrate genomes. *Gene* **259**: 31–43.

Bernardi G (2001) Misunderstandings about isochores. Part I. *Gene* **276**: 3–13.

Cuny G, Soriano P, Macaya G and Bernardi G (1981) The major components of the mouse and human genomes: preparation, basic properties and compositional heterogeneity. *European Journal of Biochemistry* **115**: 227–233.

Eyre-Walker A and Hurst LD (2001) The evolution of isochores. *Nature Reviews Genetics* **2**: 549–555.

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.

Pavlíček A, Pačes J, Clay O and Bernardi G (2002) A compact view of isochores in the draft human genome sequence. *FEBS Letters* **511**: 165–169.

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

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

Further Reading

Bernardi G (2000) Isochores and the evolutionary genomics of vertebrates. *Gene* **241**: 3–17.

Bernardi G (2002) *Structural and Evolutionary Genomics. Natural Selection and Random Drift in Genome Evolution*. Amsterdam: Elsevier.

Venter JC, *et al*. (2001) The sequence of the human genome. *Science* **291**: 1304–1351.