ENCYCLOPEDIA OF THE HUMAN GENOME

2000

Isochores

Genome

Chromosomes

Compositional genomics

Contents list:

Bernardi, Giorgio

Giorgio Bernardi

Stazione Zoologica Anton Dohrn, Naples, Italy

Chromosomes of warm-blooded vertebrates are mosaics of isochores. These are long DNA segments that are fairly homogeneous in base composition. Isochores can be assigned to five families characterised by increasing levels of guanine and cytosine and by increasing gene densities.

1. Sequence organisation of the human genome

How complex eukaryotic genomes such as the human genome are organised is a long-standing problem. Bacterial genomes have small sizes, 2-5 Mb (megabases, millions of base pairs), in which coding sequences represent about 85% of the genome and one gene corresponds, on the average, to 1 kb (kilobase, thousand base pairs) of DNA. In contrast, eukaryotic genomes cover a very broad size spectrum, ranging from 12 Mb in yeast to 3,200 Mb in human. (Much larger sizes are found in some other eukaryotes, in which the genome has been amplified by gene and/or genome duplication.) In yeast 6,000 genes represent 70% of the genome, and one gene corresponds, on the average, to 2 kb of DNA, whereas in the human genome 30,000 coding sequences represent only about 1% of the genome and one gene corresponds, on the average, to 100 kb. At least half of the 99% of the human genome which is not coding is formed by repeated sequences derived from the insertion and amplification (by many rounds of duplication) of transposons, namely of mobile sequences which have invaded the eukaryotic genome at some point in evolution.

In spite of the difficulties of the problem under consideration here, we have now a good understanding of the organisation of eukaryotic genomes, including the most complex ones, 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 re-association studies, as analysed 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.

2. The compositional approach

The major steps that led to our approach can be described as follows. In 1966, we were interested in isolating satellite DNAs (sequences that are made up of short tandem repeats) from the genomes of mouse and guinea pig. The separations obtained in CsCl density gradients not being satisfactory, we used a Cs_2SO_4 gradient (by itself having a much lower resolving power than CsCl) in the presence of a DNA ligand, Ag^+. Fractionations worked out beautifully but, in spite of both satellites being GC-poorer than "main band" (i.e. non-satellite, non-ribosomal) DNAs, the mouse satellite was "lighter" and the guinea pig satellite was "heavier" than the corresponding main band DNAs. Since both main-band DNAs were 40% GC, whereas the mouse and guinea pig satellites were 35.2% and 38.5% GC, respectively, this result demonstrated that the ligand Ag^+ was sequence-specific and that fractionation in Cs_2SO_4/Ag^+ gradient was due to the different frequency of sites able to bind the ligand on the DNA molecules from the two satellites. When we applied this approach to the numerous satellite DNAs present in the bovine genome, not only did we obtain a good resolution of four out of the eight satellite DNAs (we isolated later the other ones) but, much more interestingly, we discovered a striking and unexpected compositional heterogeneity of high molecular weight, "main-band" (see above) bovine DNA. Indeed, this was shown to comprise a broad compositional range of molecules that were distributed in a small number of families characterised by different base compositions. The demonstration of a discrete heterogeneity was an important finding, since it was at variance with the then widespread view that the bulk of the

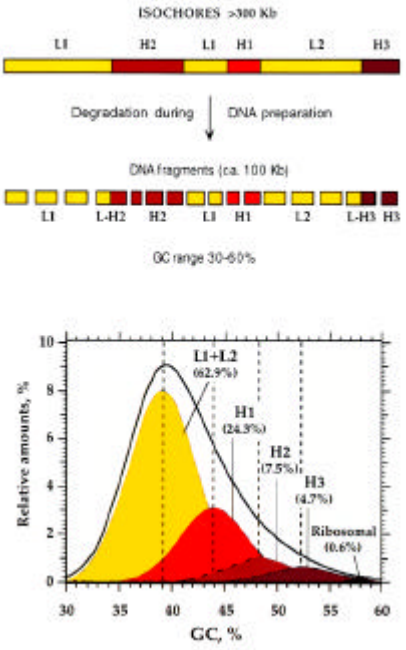
genome of higher organisms was formed by DNA molecules showing a continuous variation in GC level. On the other hand, the existence of eight GC-rich satellite DNAs accounted for the extremely large standard deviation previously estimated for bovine DNA.

3. The isochores

Further investigations showed that the compositional heterogeneity of the bovine main-band DNA was also found, with relatively minor differences, in the main bands of DNAs from all warm-blooded vertebrates. In contrast, the vast majority of the genomes from cold-blooded vertebrates exhibited a narrow compositional spectrum, noticeably lacking the GC-rich components. This provided the first indication that compositional patterns could be considered "genome phenotypes" and had an evolutionary relevance.

To sum up, in sharp contrast with bacterial DNAs that are homogeneous in base composition at fragment sizes of about 50 kb, mammalian and avian main-band DNAs (satellite DNAs are neglected here) are strikingly heterogeneous. A gaussian analysis of Cs_2SO_4/Ag^+ DNA fractions confirmed the existence of a small number of discrete DNA components and quantified their relative amounts.

In the human genome, which is typical of most mammalian genomes, DNA fragments of about 50 kb covered a 30-60% GC range, yet they could be partitioned into a small number of fairly homogeneous fragment families. Since the relative amounts of such families did not vary with increasing fragment size, it was concluded that the DNA fragments of each family derive from regions much longer than 50 kb (initially estimated as >>300 kb) and that chromosomes are mosaics of compositionally similar DNA regions, which were termed isochores. In the human genome, the isochore pattern is characterised by two GC-poor, "light" isochore families, L1 and L2, that represent about 30% and 33% of the genome, and by three GC-rich, "heavy" isochore families, H1, H2 and H3, that represent about 24%, 7.5% and 4.7%, respectively, of the genome (Fig. 1). The remaining DNA corresponds to satellite and ribosomal sequences.



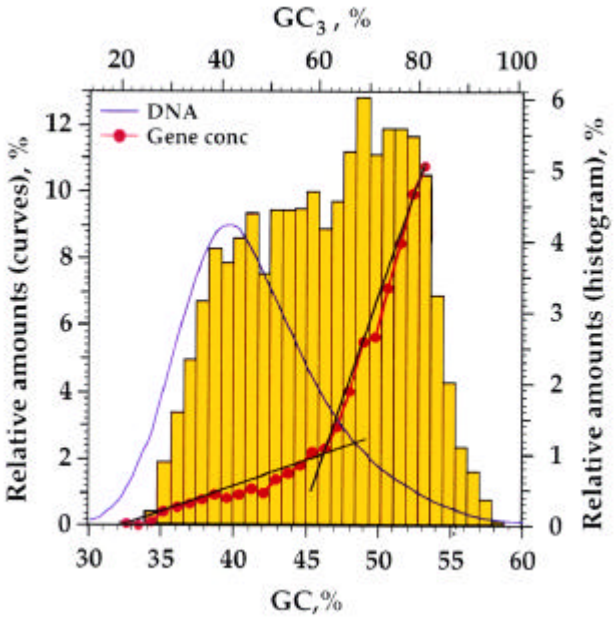
It should be pointed out that the isochore pattern of DNA is not the only compositional pattern of a genome. Indeed, another type of compositional pattern is that of coding sequences. In this case, either their GC levels or, more informatively, GC_3, the GC levels of third codon positions, define the pattern (see Fig. 2 for the 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 that the "genome equations", and linking coding and non-coding sequences amounting to a "genomic code" provided a strong evidence against the concept of non-coding DNA being "junk DNA" and in favour of compositional constraints affecting the genome as a whole.

4. Gene distribution

The linear correlation between GC_3 levels of coding sequences and GC levels of isochores is important from two viewpoints. On the one hand, it indicated, as just mentioned, that the constraints that operate on coding sequences also operate on the non-coding sequences strongly suggesting that the latter could not be considered a "junk DNA". On the other hand, it allowed the positioning of the distribution profile of coding sequences relative to that of DNA fragments (namely, the CsCl profile). In turn, this allowed estimating 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 the gene distribution in the human genome (and, for that matter, in the genomes of all vertebrates) is strikingly non-uniform (Fig. 2), gene concentration increasing from a very low average level in L1 isochores up to an about 20-fold higher level 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 Fig. 2) defines two "gene spaces" in the human genome. In the "genome core", formed by isochore families H2 and H3 (which make up 12% of the genome), gene concentration is very high, while in the "empty quarter" (a term derived from the classical name of the Arabian desert), formed by isochore families L1, L2 and H1 (which make up 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. Indeed, the latter confirmed not only the gradient of gene concentration paralleling the GC concentration, but also showed the two different slopes of Fig. 2. Moreover, it confirmed earlier results indicating that the genes located in the GC-poor empty quarter were characterised by long introns, those from the GC-rich genome core by short introns. 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. Incidentally, the two slopes of Fig. 2 were also found when plotting gene density against GC levels for Giemsa and Reverse bands from chromosomes 21 and 22. Very roughly, about half of human genes are located in the small genome core, the remaining half being located in the large empty space.

The two gene spaces are characterised 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, and correspond to an open chromatin structure. (CpG islands are regions that contain regulatory sequences, and that are about 1kb in size, rich in non-methylated CpG doublets and located upstream of the coding sequence.) The open chromatin structure is characterised by the scarcity or absence of histone H1, acetylation of histones H3 and H4, and a larger nucleosome

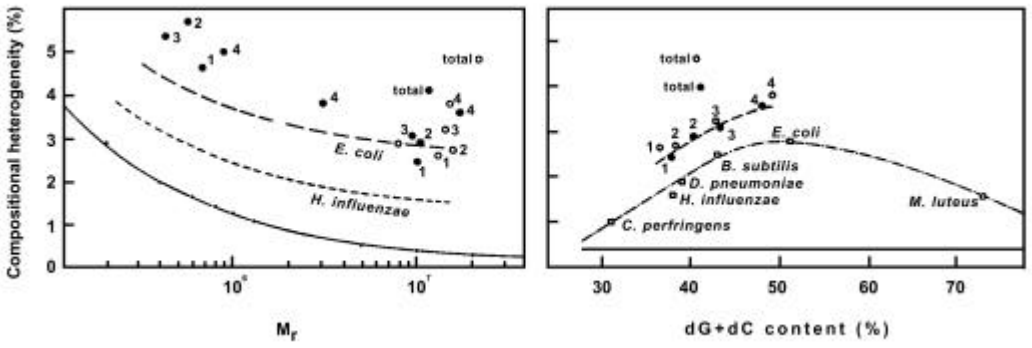
spacing. In contrast, the "empty space" corresponds to a closed chromatin structure. Interestingly, a different degree of DNA compaction was found when comparing, in chromosomes 21 and 22, the $H3^+$ and $L1^+$ bands, namely the bands hybridising the GC-richest, gene-richest isochores of the H3 family and those hybridising the GC-poorest, gene-poorest isochores of the L1 family, the former being less compact, more open, compared to the latter.

5. The compositional features of isochores.

Very recently, the International Human Genome Sequencing Consortium (IHGSC) 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.

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" simply do not exist in any natural DNA in the GC range of interest here. 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. Non-coding 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 cannot satisfy the condition because they have their own specific sequences.

2. "Strict isochores" are characterised 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 only about 30% higher than those shown by bacterial DNAs having the same size and the same GC level (Fig. 3). Since bacterial DNAs are the most homogeneous among natural DNAs (with the exception of satellite DNAs), although still much more heterogeneous than random DNAs, our original definition of isochore families as "fairly homogeneous" still seems to be an appropriate one.



3. As far as the mosaic organisation of isochores is concerned, multimodalities can be observed on human genome sequences already using an overlapping 100 kb window analysis. This is remarkable because such analyses average out all compositional

discontinuities corresponding to isochore borders. The problem of the rigorous partitioning of the genome into compositional regions is still open, but is being solved by using new algorithms that can identify isochore borders.

In conclusion, the findings just obtained are of interest in that the compositional patterns (the "genome phenotype"), the genome equations (the "genomic code"), and the gene distribution define a eukaryotic genome in terms of its structural and functional properties. This replaces the original, purely operational definition of the genome as the haploid chromosome set, which still is the only one presented, in an explicit or implicit form, in current textbooks. For this reason, these results also represent a break-through in the long-standing problem of genome organization of vertebrates.

Finally, and most importantly, these results raise the question of the evolutionary background of the compositional properties and of the gene distribution of the human genome. These points are presented and discussed in two other articles in this Encyclopedia.

Figure Legends

Fig. 1. (Top). 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 DNA segments, the isochores, which are compositionally 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, 1995).

(Bottom). 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 (broken vertical lines). The relative amounts of major DNA components are indicated. Satellite DNAs are not represented. (From Zoubak et al., 1996).

Fig. 2. Profile of gene concentration (red dots) 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 (yellow bars) by the corresponding relative amounts of DNA deduced from the CsCl profile (blue line). The positioning of the GC_3 histogram relative to the CsCl profile is based on the correlation of Fig. 2. (Modified from Bernardi, 2000a).

Fig. 3. Experimentally measured compositional heterogeneities, quantified by the standard deviation among GC levels of DNA fragments from bacterial genomes (open squares) and from isochore families and total genome DNA of human (open circles) and mouse (closed circles; in this case, total DNA is "main-band" DNA, the satellite DNA being neglected). The standard deviations are shown plotted against mean fragment size (left panel) and against mean GC content (right panel), and compared with the expectation for random sequences (bottom curve in each panel). A molecular weight M_r of 2×10^6 corresponds to approximately 3 kb. 1-4 correspond to isochores families L1, L2, H1 and H2 (from Cuny et al., 1981, where data sources and further details are given).

Further reading

Bernardi,G (1995) The human genome :organization and evolutionary history *Annu. Rev. Genet.* **29**: 445-476.

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

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

Clay O, Macaya G, Carels N, Douady C, Bernardi G (2001) Quantifying compositional heterogeneity within and among mammalian genomes *Gene* (in press).

Cuny G, Soriano P, Macaya G, Bernardi G (1981) The major components of the mouse and human genomes: preparation, basic properties and compositional heterogeneity *Eur. J. Biochem*. **115**(2): 227-233

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

Macaya G, Thiery JP, Bernardi G (1976) An approach to the organization of eukaryotic genomes at a macromolecular level *J. Mol. Biol.* **108**: 237-254.

Saccone S, Pavlicek A, Federico C, Paces J, Bernardi G (2001). Genes, isochores and bands in human chromosomes 21 and 22. *Chromosome Research* (in press)

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

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

