ENCYCLOPEDIA OF THE HUMAN GENOME

2000

©Nature Publishing Group

The genome organisation of vertebrates

Isochores

Base composition

Chromosomes

Compositional genomics

Contents list:
1. The genome
2. The human genome
3. Sequence organisation of the mammalian genome
4. Compositional correlations
5. Gene distribution and gene spaces
6. Isochores, genes and chromosomes

Bernardi, Giorgio

Giorgio Bernardi

Stazione Zoologica Anton Dohrn, Naples, Italy

Compositional genomics, an approach relying on the base composition of genomes, greatly helped to solve a long-standing problem, namely the sequence organisation of vertebrate genomes (and more generally, of eukaryotic genomes).

1.   The genome

Every living organism contains in its genome (a term coined in 1920 by the German botanist Hans Winkler) 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 (Fig. 1). The building blocks of each DNA strand are deoxyribonucleotides. These are formed by a phosphate ester of deoxyribose (a pentose 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.
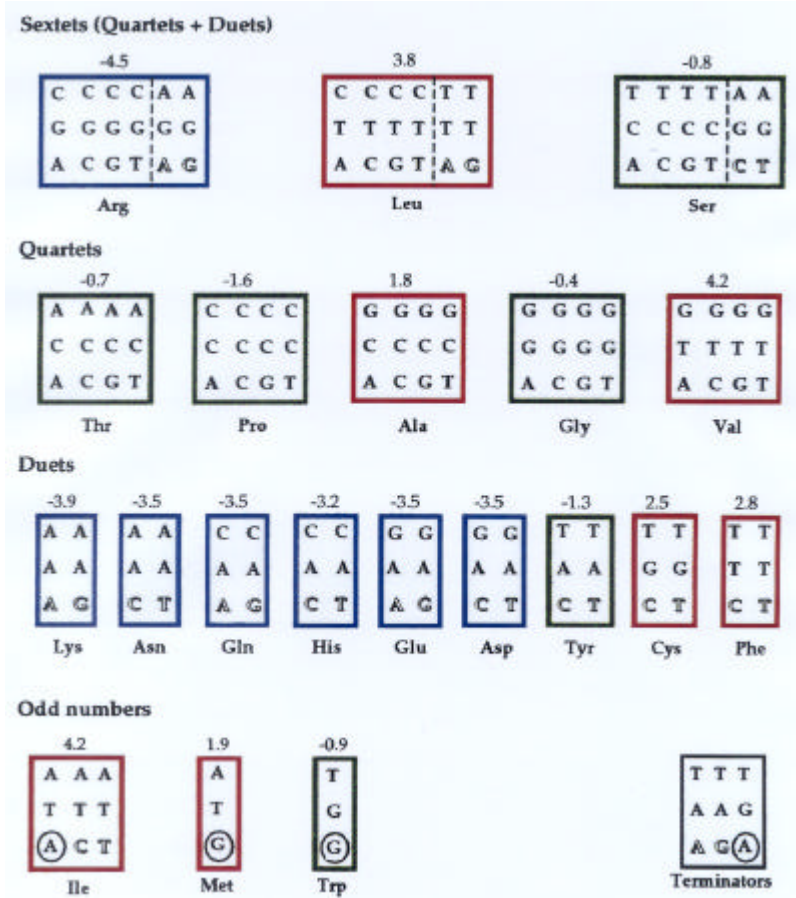
During cell replication, the two strands of the double helix are unwound, and a complementary copy of each is made (following the base pairing scheme mentioned above), producing two identical copies (except for rare mistakes, or mutations) of the parental double helix. The two strands are also unwound at the time when one strand, the "sense strand" carrying the genetic information, is copied into a complementary RNA. This differs from the DNA master copy in having in its nucleotides ribose instead of deoxyribose, and uracil instead of thymine. RNA transcripts of genes are used as templates for the synthesis of proteins, except for some RNAs (ribosomal and transfer RNAs; see below) which are used in the translation of proteins, but are not translated themselves.
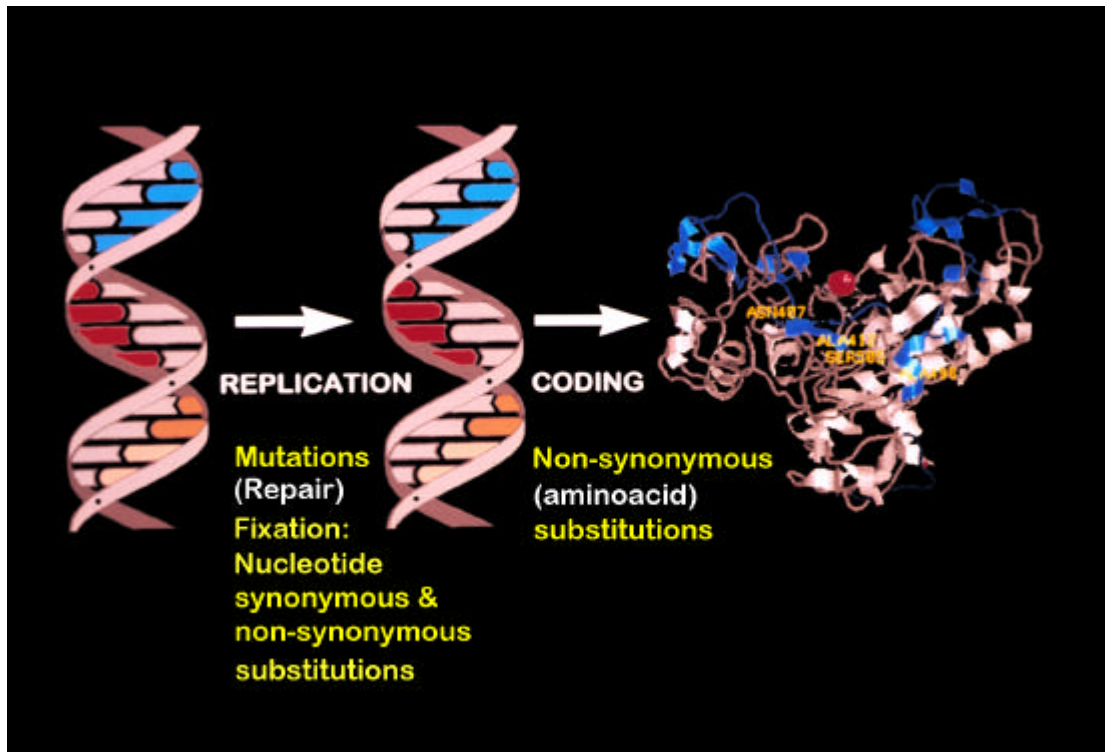
The translation of each RNA transcript into the corresponding protein involves a very complex machinery that makes use of ribosomes (particles made up of two subunits, each containing a ribosomal RNA, or rRNA) and a number of transfer RNAs (or tRNAs) that are specific for different amino acids. Subsequent 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 (Fig. 2). Because there are 64 triplets (three of which are termination codons, marking the end of translation, and one of which, AUG, is the initiation codon) 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. The genetic code is, therefore, said to be degenerate, which means that alternative possibilities exist for encoding the same amino acid. Differences among synonymous codons are mainly at the nucleotides in third codon positions.

In summary, the two central roles played by the genome in living organisms are: 1) faithful replication of itself and transmission of the genetic information to the progeny (Fig. 1); however, mutations may occur due to mistakes in replication (the major factor), to recombination, to environmental factors; mutations undergo repair; the nucleotide substitutions that survive repair and are fixed are subject to natural

selection; 2) coding for proteins using a genetic code (Fig. 2), whose existence provides the ultimate evidence for the single origin of all living organisms.

Sextets (Quartets + Duets)

-4.5	3.8	-0.8
C C C C A A	C C C C T T	T T T T A A
G G G G G G	T T T T T T	C C C C G G
A C G T A G	A C G T A G	A C G T C T
Arg	Leu	Ser

Quartets

-0.7	-1.6	1.8	-0.4	4.2
A A A A	C C C C	G G G G	G G G G	G G G G
C C C C	C C C C	C C C C	G G G G	T T T T
A C G T	A C G T	A C G T	A C G T	A C G T
Thr	Pro	Ala	Gly	Val

Duets

-3.9	-3.5	-3.5	-3.2	-3.5	-3.5	-1.3	2.5	2.8
A A	A A	C C	C C	G G	G G	T T	T T	T T
A A	A A	A A	A A	A A	A A	A A	G G	T T
A G	C T	A G	C T	A G	C T	C T	C T	C T
Lys	Asn	Gln	His	Glu	Asp	Tyr	Cys	Phe

Odd numbers

4.2	1.9	-0.9	
A A A	A	T	T T T
T T T	T	G	A A G
(A) C T	(G)	(G)	A G (A)
Ile	Met	Trp	Terminators

The genomes of living organisms differ greatly in size, from 4.2 Mb (megabases, or millions of base pairs [bp]) for a typical bacterium, such as *Escherichia coli*, to about 3,200 Mb or 3.2 Gb (gigabases, or billions of bp) for eukaryotes such as humans. While prokaryotes (bacteria) are characterised by small genome sizes, clustering around the value given above for *E.coli*, eukaryotes exhibit larger genome sizes that cover a wide range – from 13 Mb for the yeast *Saccharomyces cerevisiae* to 3 Gb for mammals (eukaryotes with larger genome sizes are also known). Table 1 stresses the fact that "complex" eukaryotic genomes, such as the human genome, are very different from the genomes of prokaryotes (and of 'simple' eukaryotic genomes) in comprising enormous amounts of non-coding sequences.

> Table 1 should be placed here

Indeed, the much larger genome size of eukaryotes (as compared with prokaryotes) is due only in small part to the presence of a greater number of genes (see below). In fact, the increase in size is mainly due to the existence in eukaryotes (but only at a

very low level in prokaryotes) of noncoding sequences. These can be both intergenic, between genes, and intragenic, within genes. The latter sequences, called introns, separate different coding stretches, or exons, of most eukaryotic genes. The intron parts of the primary RNA transcript are eliminated by splicing, leaving the mature transcript, or messenger RNA (mRNA), that encodes a protein.

Eukaryotes differ from prokaryotes not only in the features of their genome but in other respects as well. They have a nucleus that is separated from the cytoplasm by a nuclear membrane. Moreover, in addition to the nuclear genome, the only one discussed so far, eukaryotic cells also have organelle genomes, which are located in mitochondria and, in the case of plants, also in chloroplasts. Organelle genomes have a very small size (the size of the animal mitochondrial genomes is only 16 kb), yet they contain an essential amount of genetic information encoding organelle-specific proteins, rRNAs and tRNAs. Organelle genomes apparently originated from symbiotic bacteria, which entered proeukaryotic cells. Like the bacterial genomes from which they derive, organelle genomes are physically organised in a rather simple way. In contrast, the nuclear genome of eukaryotic DNA is wrapped around octamers of histones (which are basic proteins) to form nucleoprotein bodies called nucleosomes, that are packaged into chromatin fibers. 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.

2.   The human genome

Estimates of the number of (nuclear) human genes range from 30,000 to 40,000, figures in the lower range being supported by recent results. If coding sequences average 1,000 bp, they would represent about 1% of the human genome, 99% or so of which is, therefore, made up of non-coding sequences (see Table 1). It should be noted that the larger number of genes in humans (and eukaryotes in general) compared with bacteria is mainly due to the fact that many eukaryotic genes exist as multigene families, the result of genome and gene duplications during evolution.

Our present knowledge of human coding sequences, in terms of primary structures (or nucleotide sequences), is complete, at least in principle. Indeed, while a "draft" sequence of the human genome was obtained in year 2001, putting end-to-end exons

into coding sequences is still far from complete. Difficulties mainly arise from the frequent presence of very long introns and very short exons in mammalian genes. This accounts for the uncertainty in the number of human genes (see above).

As far as intergenic sequences are concerned, a sizable part is formed by repeated sequences that belong into several families. The two most important families are called LINES and SINES (the Long and Short Interspersed sequences), which are present in about 850,000 and 1,500,000 copies, respectively. LINES and SINES are retroposons, genetic elements that are propagated with a process in which RNA transcripts are reverse-transcribed into DNA and reinserted at many different sites in the genome. While the SINES (which are 300-bp long) and LINES (which cover a broad size range up to 10,000 bp) are scattered over the genome (mostly in intergenic regions), other repeated sequences consist of tandem oligonucleotides forming very long stretches typically localised in centromeres. Because of the features of their sequences, these tandem repeats were recognised early on as satellite DNAs, namely DNA sequences which could be separated from the bulk of nuclear DNA by centrifugation in density gradients (see below).

## 3.   Sequence organisation of the mammalian genome

How complex eukaryotic genomes such as the human genome are organised is a long-standing problem. Previous attempts were based on DNA re-association studies, in which DNA is fragmented into small pieces, denatured and reannealed for different times. Repeated sequences find their complementary strands easily and reassociate fast, 'single copy' sequence, take a longer time and reassociate slowly. Analysis of the reassociation kinetics by estimating the amounts of single- and (non-reassociated) double-stranded (reassociated) DNAs as a function of time allowed the discovery of the existence of abundant repeated sequences in eukaryotic genomes, but this approach could not proceed any further. The problem of the genome organization of complex eukaryotic genomes could, however, be solved by an experimental approach based on the most fundamental property of DNA, namely its base composition. Indeed, sequence-specific ligands, such as silver ions [or BAMD, 3,6-bis(acetato mercurimethyl)1-4dioxane], bind to DNA molecules proportionally to the frequencies

of the oligonucleotide binding sites making DNA molecules "lighter" or "heavier" in the density gradient and so allowing a high resolution fractionation. This approach led to the discovery of a striking and unexpected compositional heterogeneity of high molecular weight, "main band" (i.e. non-satellite, non-ribosomal) bovine DNA. In fact, this mammalian DNA was shown to comprise a large spectrum of molecules that were distributed in a small number of families characterised by different base compositions. Further work showed that the DNA fragments (or molecules) from vertebrates (routinely 100 kb in size) that form DNA preparations derive from much longer segments, the isochores (initially estimated as $\gg$300 kb), that are compositionally fairly homogeneous and belong to a small number of families characterised by different GC levels (GC is the molar ratio, namely, the percentage of guanine+cytosine in DNA). Isochore is a word derived from the Greek for "equal" (compositional) "landscape".

In the case of the human genome, the isochore pattern is characterised (Fig. 3) by GC-poor, "light", L1 and L2, isochores that represent about 30% and 33% of the genome, whereas GC-rich, "heavy", H1, H2 and H3 isochores make up about 24%, 7.5% and 4.7%, respectively, of the genome. The remaining DNA corresponds to satellite and ribosomal sequences.

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, GC3, the GC levels of their third codon positions, define the pattern (see Fig. 4 for the pattern of human coding sequences).

One should stress that the compositionally discontinuous isochore structure of the human genome (as well as of the genomes from warm-blooded vertebrates in general) is very different from the continuous compositional spectrum that was the prevailing view until the 70's.
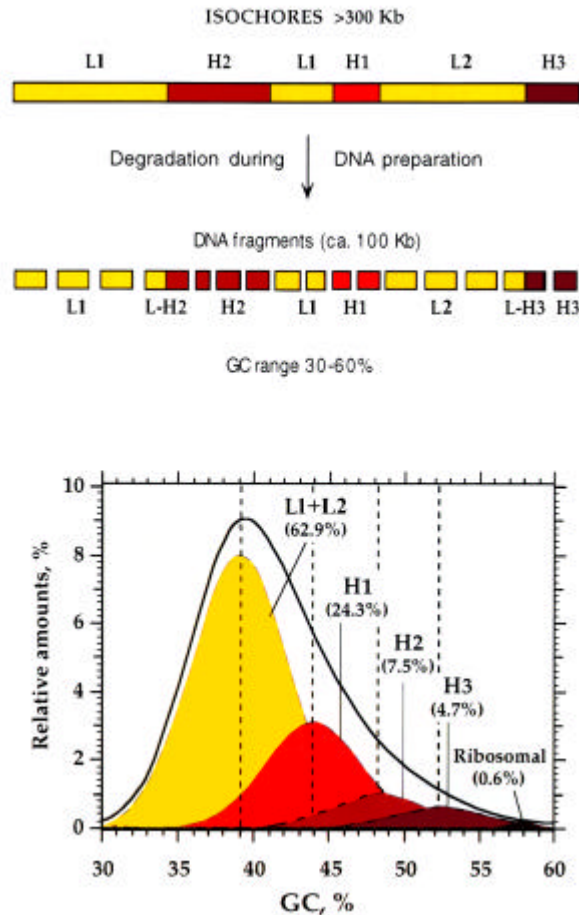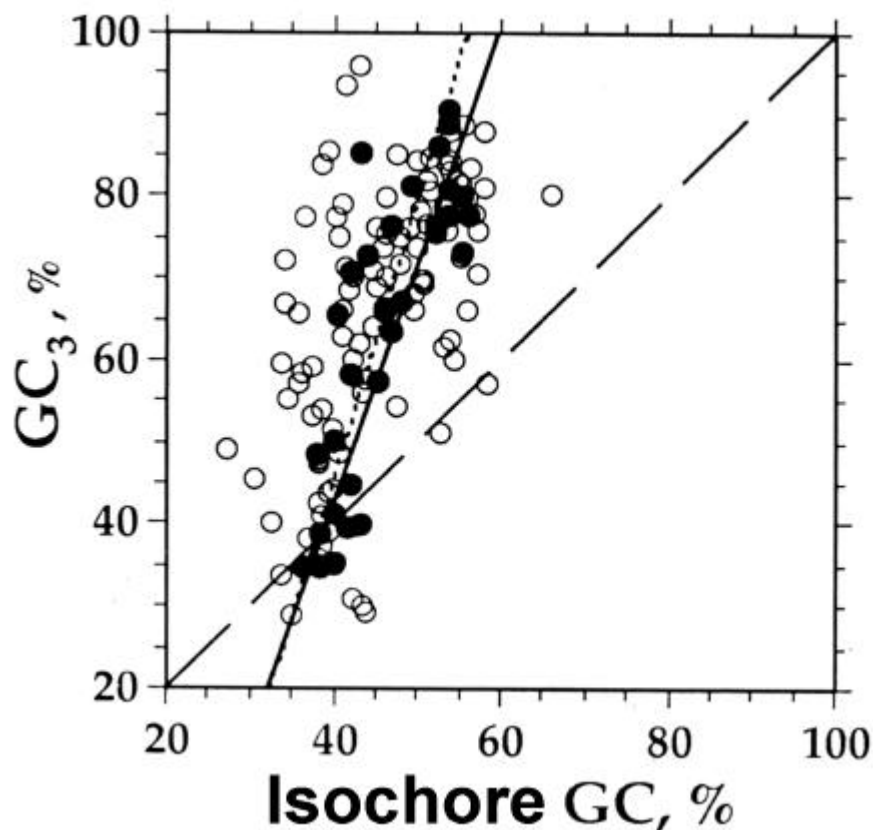
4. Compositional correlations

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.

Indeed, linear correlations hold between the GC levels (and the GC_3 levels; GC_3 is the average GC value of third codon positions of a given gene) of coding sequences and the GC levels of the isochores in which coding sequences are located (see Fig. 4). Interestingly, GC-poor coding sequences and their flanking sequences show very similar values, whereas GC-rich coding sequences are increasingly higher above the diagonal (that corresponds to equal values on the abscissa and on the ordinate), because GC_3 values depart more and more from the intergenic sequences. Linear

correlations also hold between the GC levels of coding sequences and the GC levels of the introns of the same genes, the GC levels of the former being higher than those of the latter (not shown).

Even if the evolutionary issues related to the isochore structure of the mammalian and avian genomes will be discussed elsewhere in this Encyclopedia (G. Bernardi 'The Evolutionary History of the Human Genome'; ms. Ref. #70), it is important to stress here that the compositional correlations that hold between coding and non-coding sequences indicate that the latter cannot simply be "junk DNA". Indeed, the correlations strongly suggest that the compositional constraints that apply to non-coding sequences are similar to those of coding sequences.


5.   Gene distribution and gene spaces

The correlation between GC_3 levels of coding sequences and GC levels of isochores (Fig. 4) is also important from a practical viewpoint. Indeed, it allows the positioning of the distribution profile of coding sequences relative to that of DNA fragments, the CsCl profile. In turn, this allows 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; see below) is strikingly non-uniform (Fig. 5), gene concentration increasing from a very low average level in L1 isochores up to an about 20-fold higher level in H3 isochores (more precisely, in the 100-kb DNA molecules derived from L1 and H3 isochores).
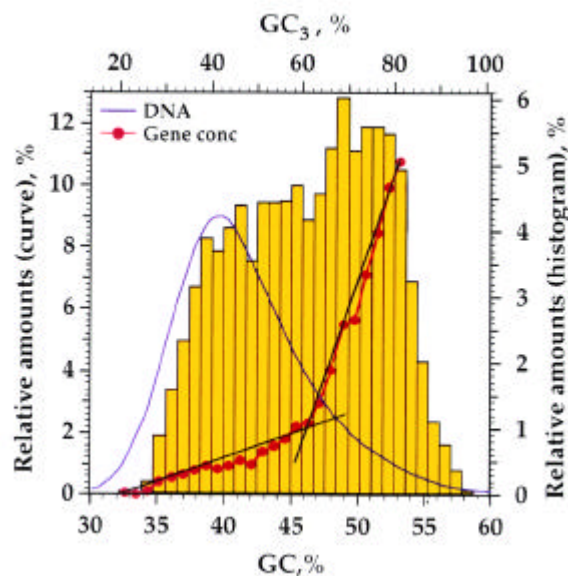
The existence of a break in the slope of gene concentration at 60% GC_3 of coding sequences and at 46% GC of isochores (see Fig. 5) defines two "gene spaces" in the human genome. In the "genome core", formed by isochore families H2 and H3 (which make up 12% of the genome), gene concentration is very high, in the "empty quarter", formed by isochore families L1, L2 and H1 (which make up 88% of the genome), gene concentration is very low. Note that the definition of "genome core" is prompted not only by the position of the break in the gene concentration curve of Fig. 5, but also by the similar gene concentrations in L1/L2/H1 and H2/H3 isochores, respectively. About half of human genes are located in the small genome core, the other half being located in the large empty quarter (the classical name for the Arabian desert).

The two gene spaces are characterised by a number of different structural and functional properties. Indeed, most genes located in the genome core are associated with CpG islands (regulatory sequences, about 1kb in size, rich in non-methylated CpG doublets and located upstream of the coding sequence), are characterised by short introns, are actively transcribed and correspond to an "open" chromatin structure, are replicated early in the cell cycle and are located in highly recombinogenic regions of the genome. This 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 quarter corresponds to a "closed" chromatin structure (see also the following section).

6. Isochores, genes and chromosomes

Another 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 to 22 in order of decreasing size in the standard karyotype) 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 characterised by specific staining properties. G bands (Giemsa-positive, or Giemsa dark bands; equivalent to Q-bands, or Quinacrine bands) and R bands (Reverse bands; essentially equivalent to Giemsa-negative or Giemsa light bands) are produced by treating metaphase chromosomes with fluorescent dyes, proteolytic digestion, or differential denaturing conditions.

When applied to metaphase chromosomes, Giemsa staining produces a total of about 400 bands, each of which comprises, on the average, 8.5 Mb of DNA. If staining is applied to prometaphase chromosomes, which are more elongated, 850 bands can be visualized. At this high resolution, one chromosomal band contains, on the average, 4 Mb of DNA. A number of approaches, ranging from purely genetical to molecular ones have allowed the assignment of genes not only to individual chromosomes but also to chromosomal bands.

In situ hybridisation of isochores on metaphase chromosomes showed high levels of the GC-richest H3 isochores in a number of R bands, and high levels of the GC-poorest L1 isochores in a number of G(iemsa) bands. These H3+ and L1+ bands (as they were called), largely corresponded to the bands most resistant to heat-denaturation, and to the most intensely staining sets of G bands, respectively. The remaining G and R bands are characterised by an intermediate GC composition (see Fig. 6 and Table 2). These intermediate bands, H3? and L1?, do not hybridise H3 or L1 isochores, respectively.
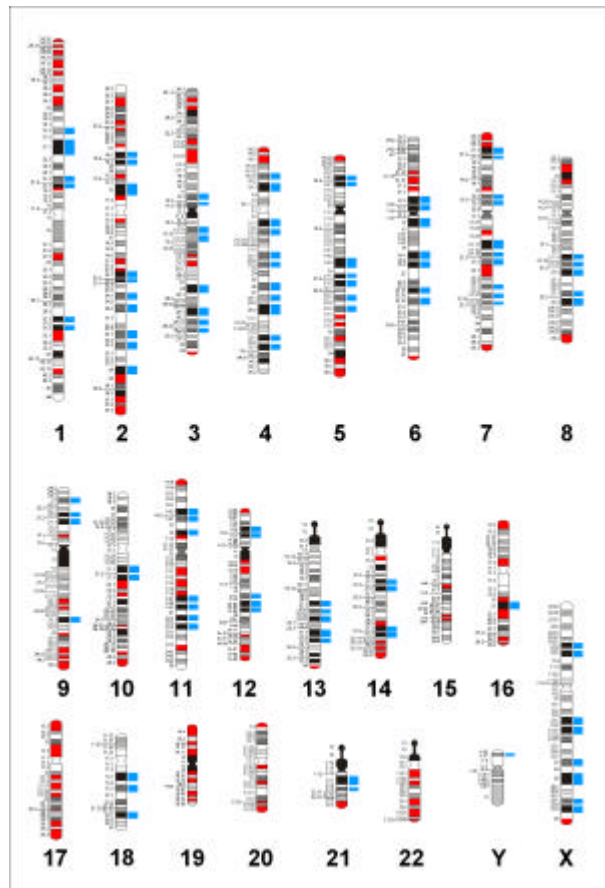
Thus, the human genome is composed by at least three compositionally different sets of chromosomal bands. The GC-richest one is characterised not only by the highest concentration of genes, but also by an "open" chromatin structure, a very high level of transcriptional activity, the highest recombination frequency and the earliest replication timing in the S phase of the cell cycle. The GC-poorest one is characterised by opposite features. The third set of bands (L1? and H3? bands; see Table 2) is characterised by properties intermediate between those of the two other sets.

Table 2 should be placed here

An analysis of the nucleotide sequences (Fig. 7) of the long arms of human chromosomes 21 and 22 not only provided a new quantification of the gene density/GC level relationships, which was in agreement with that of Fig. 3, but also allowed to show that H3$^+$ bands have a lower compaction of DNA compared with L1$^+$ bands (in agreement with their open chromatin structure). As far as intermediate bands were concerned, some G bands, L1?, were observed that exhibited higher GC levels compared to some R bands, H3?, the G or R banding appearance depending upon the higher or lower GC levels, respectively, of flanking bands. Finally, if the four different sets of bands, H3$^+$, H3?, L1?, L1$^+$, of chromosomes 21 and 22 are representative of the corresponding band types over the entire human karyotype, the gene density of Table 2 lead to an estimate of human gene number equal to 28,000.
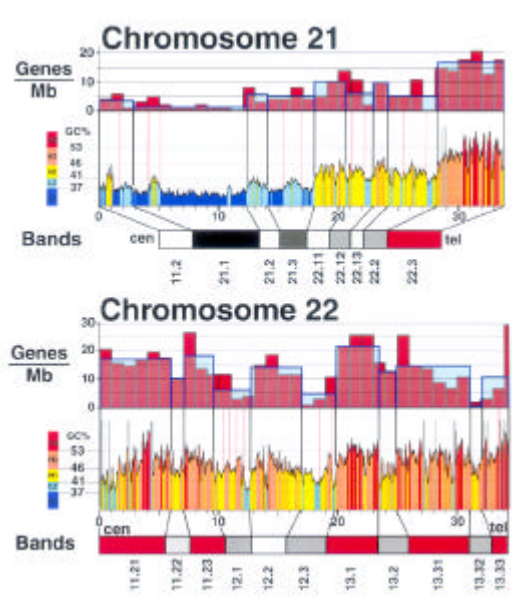
In conclusion, the findings just outlined are of interest in that the compositional patterns, the genome equations concerning compositional correlations, and the gene distribution define the human genome in terms of its structural and functional properties. This replaces the original, purely operational 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. Moreover, the compositionally discontinuous pattern of the genome is in sharp contrast with the continuous compositional spectrum that prevailed until the 70's.

Table 1. # Genome size, coding sequences and gene numbers in some representative organisms.

Organism	Genome size*	Coding sequences	Genes*	kb/gene
	Mb	%		
Haemophilus	2	85	2,000	1
Yeast	12	70	6,000	2
Human	3,200	1	32,000	100

* in round figures; Mb, megabases, or millions of base pairs, bp; kb, kilobases, or thousands of bp; Gb, gigabases, or billions of bp.

Table 2. Classification, relative amounts, gene numbers and gene densities of chromosomal bands from the entire human karyotype (1).

Bands	%	by staining properties (2)		by isochore content (3)		gene number (4)	gene density (5)
Giemsa	47%	G1	13.7%	L1+	26.4%	2,700	3.0
		G2	12.7%				
		G3	13.1%	L1-	20.7%	4,800	6.8
		G4	7.6%				
Reverse	53%			H3-	35.6%	10,400	8.6
		T	˜ 15%	H3+	17.4%	10,200	17.3

(1) Relative amounts of different band types were assessed on the basis of band sizes in the 850-band karyotype of Francke (1994). This Table is from Saccone et al., 2001.

(2) We call here G1-G4 bands the G bands characterised by different stainig intensities (black to pale grey in the ideogram of Francke). T bands are the most heat-denaturation resistant R bands. (Dutrillaux, 1973).

(3) See Federico et al., 2000.

(4) The estimated number of genes is 28,100 considering a genome size of 3,400 Mb.

(5) This is the average gene density per Mb found in the bands of chromosomes 21 and 22.

Figure Legends

Fig. 1. # Scheme of the two fundamental functions of DNA: replication and coding for aminoacids in proteins. During replication, mistakes may occur, resulting in mutations. Some of these are repaired, but other ones persist and may spread into the progeny reaching 100% levels: mutations are then said to be "fixed" into nucleotide substitutions. The latter, after transcription of DNA into RNA and translation of RNA into proteins (not shown), may be silent (no aminoacid change) or appear as aminoacid changes, which may be very rarely advantageous, but more frequently are neutral or deleterious. The symbols on the protein chain on the right indicate specific aminoacids.

Fig. 2. # Genetic code. The Grantham (1980) representation of the genetic code. was modified in that codons rather than anticodons are shown, a distinction is made among third position nucleotides of quartet, duet and odd number codons, and hydropathy values for aminoacids using the scale of Kyte and Doolittle (1982) are shown. Red, blue and green boxes indicate the most hydrophobic, the most hydrophilic and the intermediate aminoacids, respectively. (From D'Onofrio et al., 1999a,b).

Fig. 3. # (Top). Scheme of the isochore organisation 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 preparations 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 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. 4. # Correlation between GC_3 of human coding sequences and the GC levels of the DNA fractions in which the genes were localized (filled circles) and of 3' flanking sequences further than 500 bp from the stop codon (open circles; the solid and the broken lines are the regression lines through the two sets of points). (From Clay et al., 1996).

Fig. 5. # 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, 2000).

Fig. 6. # Identification of the GC-poorest and the GC-richest chromosomal bands. The human karyotype at a resolution of 850 bands per haploid genome shows the chromosomal bands containing the GC-poorest (blue bars on the right of each chromosome) and the GC-richest isochores (red regions inside the chromosomes). G bands are pale gray to black according to staining intensity. R bands are red (H3+) or white (H3-). (From Federico et al., 2000).

Fig. 7. # Correlations between chromosomal bands, isochores, and gene concentration in human chromosomes 21 and 22. (Bottom to top). Bands, ideogram at a resolution of 850 bands showing the four classes of G bands staining with different intensities and the two classes (H3+, red; H3-, white) of R bands; the two chromosomes are represented according to their relative cytogenetic size. GC, the GC profiles were obtained using a window size of 100 kb; 37%, 41%, 46%, and 53% GC were taken as the upper values of the L1, L2, H1 and H2, isochore families, respectively; the grey bars indicate the DNA sequences not yet available. Genes/Mb, gene density per Mb; the blue histogram concerns chromosomal bands; the red histogram 1 Mb segments. (From Saccone et al., 2001).

Further reading

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

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

Bernardi G (2000b). The compositional evolution of vertebrate genomes. *Gene* **259(1-2)**: 31-43.

Clay O, Cacciò S, Zoubak S, Mouchiroud D and Bernardi G (1996) Human coding and non-coding DNA : compositional correlations *Mol. Phylogenet. Evol.* **5:** 2-12.

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

Federico C, Andreozzi L, Saccone S and Bernardi G (2000) Gene density in the Giemsa bands of human chromosomes. *Chromosome Research* **8**(8):737-746.

Francke U (1994) Digitized and differentially shaded human chromosome ideograms for genomic applications *Cytogenet. Cell Genet*. **65**(3): 206 – 218.

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

Rynditch AV, Zoubak S, Tsyba L, Tryapitsina-Guley N and Bernardi G (1998) The regional integration of retroviral sequences into the mosaic genomes of mammals *Gene* **222:** 1-16.

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

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

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

Glossary

Alu element # A dispersed, intermediately repetitive, 300-bp DNA sequence, ~1,000,000 copies of which exist in the human genome.

Codon usage bias # Unequal frequencies, in a protein-coding sequence of DNA, of the alternative codons that specify the same amino acid

GC # The molar ratio, namely the percentage, of guanine + cytosine in DNA

GC_3 # GC at the third codon positions

Isochores # Long regions of DNA characterized by quasi-homogeneity of base composition. Isochore size is 0.2-2 Mb or more. Isochores belong to a small number of families having distinct base compositions

LINE element # Long, interspersed sequences, such as L1, generated by retrotransposition

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)

Silent codon positions # One at which nucleotide change is not accompanied by an amino acid change in the translation product

SINE element # Short, interspersed, repetitive sequences, such as Alu elements, generated by retrotransposition

Synonymous codon # One at which a nucleotide change does not alter the amino acid encoded