Genome Organization of Vertebrates

Giorgio Bernardi, *Laboratory of Molecular Evolution, Stazione Zoologica Anton Dohrn, Naples, Italy*

Introductory article

Article Contents

Online posting date: 15th July 2008

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

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 (**Figure 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 the 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 and 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 earlier base-pairing scheme), 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). This differs from the DNA master copy in having in its nucleotides ribose instead of deoxyribose, and uracil (U) instead of thymine. RNA transcripts of genes are used as templates for the synthesis of proteins, except for ribosomal RNA (rRNA) and transfer RNA (tRNA), which are used in the translation of proteins but are not themselves translated.

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 an rRNA) and a number of 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 (**Figure 2**). As there are 64 triplets (minus the three termination codons, which mark the end of translation, and the initiation codon, AUG, which also encodes the amino acid methionine) 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 in the nucleotides of third codon positions.

In summary, the two central roles played by the genome in living organisms are:

Faithful replication of itself and transmission of the genetic information to the organism's progeny (**Figure 1**). Mutations may occur, however, through mistakes in replication (the major factor), recombination and environmental factors; mutations undergo repair, and the nucleotide substitutions that survive repair and are fixed are subject to natural selection.

Coding for proteins using a genetic code (**Figure 2**) whose existence provides the ultimate evidence for the single origin of all living organisms.

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 approximately 3200 Mb or 3.2 Gb (gigabases, or billions of bp) for eukaryotes such as humans. Whereas prokaryotes (bacteria and archaea) are characterized by small genome sizes, clustering around the value given earlier for *E. coli*, eukaryotes exhibit larger genome sizes and a greater range of them from 12 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 unicellular eukaryotic genomes) in comprising enormous amounts of noncoding sequence.

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 later). In fact, the increase in size is mainly due to the existence in eukaryotes

ELS subject area: Genetics and Molecular Biology

How to cite:
Bernardi, Giorgio (July 2008) Genome Organization of Vertebrates. In: Encyclopedia of Life Sciences (ELS). John Wiley & Sons, Ltd: Chichester.
DOI: 10.1002/9780470015902.a0005001.pub2

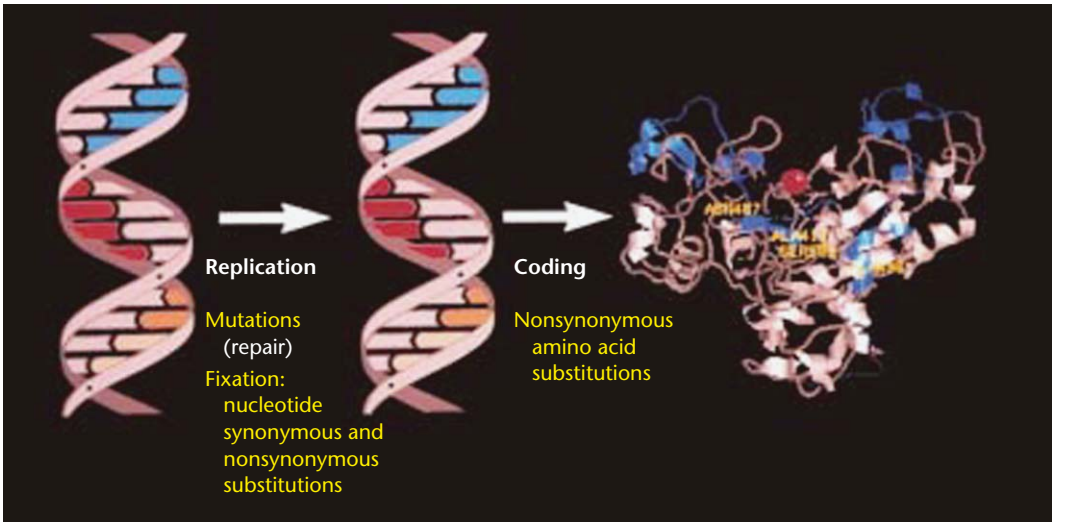

Figure 1 The two fundamental functions of DNA: replication and coding. During replication mistakes may occur, resulting in mutations. Some of these are repaired, but others persist and may spread into the progeny reaching 100% levels in the population: mutations are then said to be 'fixed' into nucleotide substitutions. The latter, after transcription of DNA into RNA and translation of RNA into proteins, may be silent (no amino acid change) or may appear as amino acid changes, which may be very rarely advantageous but more frequently are neutral or deleterious. Silent changes are also called synonymous, whereas nucleotide changes leading to aminoacid changes are called non-synonymous. The symbols on the protein chain on the right indicate specific amino acids.

of noncoding sequences (which are present only at a very low level in prokaryotes). 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 are very small (the size of the animal mitochondrial genomes is only 16 000 bp or 16 kb), yet they contain an essential amount of genetic information encoding organelle-specific proteins, rRNAs and tRNAs. Organelle genomes originated from symbiotic bacteria, which entered proeukaryotic cells. Like the bacterial genomes from which they derive, organelle genomes are physically organized in a rather simple way. By contrast, the nuclear genome of eukaryotic DNA is wrapped around octamers of histones (which are basic proteins) to form nucleoprotein bodies called nucleosomes, which are packaged into chromatin fibres. These fibres are folded into chromatin loops, consisting of 30–100 kb of DNA, which are, in turn, packaged into chromosomes.

Human Genome

Estimates of the number of (nuclear) human genes range from 25 000 to 30 000. If coding sequences average 1000 bp, they would represent approximately 1% of the human genome, 99% or so of which is, therefore, made up of noncoding sequences (see **Table 1**). It should be noted that the larger number of genes in humans (and eukaryotes in general) as compared with bacteria is mainly due to the fact that many eukaryotic genes exist as multigene families, which are 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 now essentially 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 earlier).

As far as intergenic sequences are concerned, a sizeable part is formed by repeated sequences that belong to several families. The two most important families are called LINEs and SINEs (the long and short interspersed sequences), which are present in approximately 850 000 and 1 500 000 copies, respectively. LINEs and SINEs are retroposons, genetic elements that are propagated in a process in which RNA transcripts are reverse-transcribed into DNA and reinserted at many different sites in the genome. Whereas 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 localized in centromeres. Because of the features of their sequences, these tandem repeats were recognized early on as satellite DNAs, namely DNA sequences that could be separated from the bulk of nuclear DNA by centrifugation in density gradients (see below).

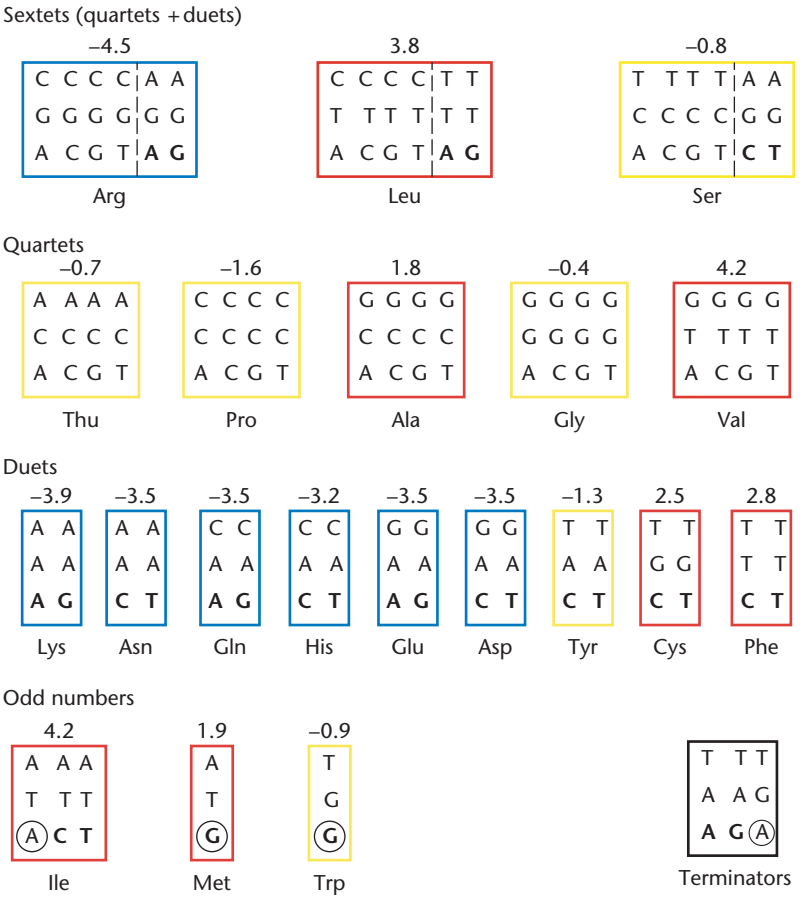

Figure 2 The genetic code. The 1980 Grantham's representation of the genetic code was modified in that codons rather than anticodons are shown, a distinction is made among third position nucleotides of quartet, duet and odd number codons, and hydropathy values for amino acids using the scale of Kyte and Doolittle are shown. Codons are displayed vertically, the first position being on the top. One can observe that most third codon position changes are silent or synonymous, they do not lead to a change in the corresponding aminoacids. Reproduced from D'Onofrio G, Jabbari K, Musto H and Bernardi G (1999) The correlation of protein hydropathy with the composition of coding sequences. *Gene* **238**: 3–14.

Table 1 Genome size, coding sequences and gene numbers in some representative organisms (approximate figures)

Organisms	Genome size (Mb)	Coding sequences (%)	Genes	kb/gene
Haemophilus	2	85	2000	1
Yeast	12	70	6000	2
Human	3200	1	32 000	100

Notes: kb, kilobases, or thousands of bp and Mb, megabases, or millions of base pairs (bp).

Sequence Organization of the Mammalian Genome

How complex eukaryotic genomes, such as the human genome, are organized is a long-standing problem. Previous attempts were based on DNA reassociation 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, whereas 'single-copy' sequences take a longer time and reassociate slowly. Analysis of the reassociation kinetics by estimating the amounts of single- (nonreassociated) and double-stranded (reassociated) DNAs as a function of time using hydroxyapatite chromatography 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 be solved, however, by an

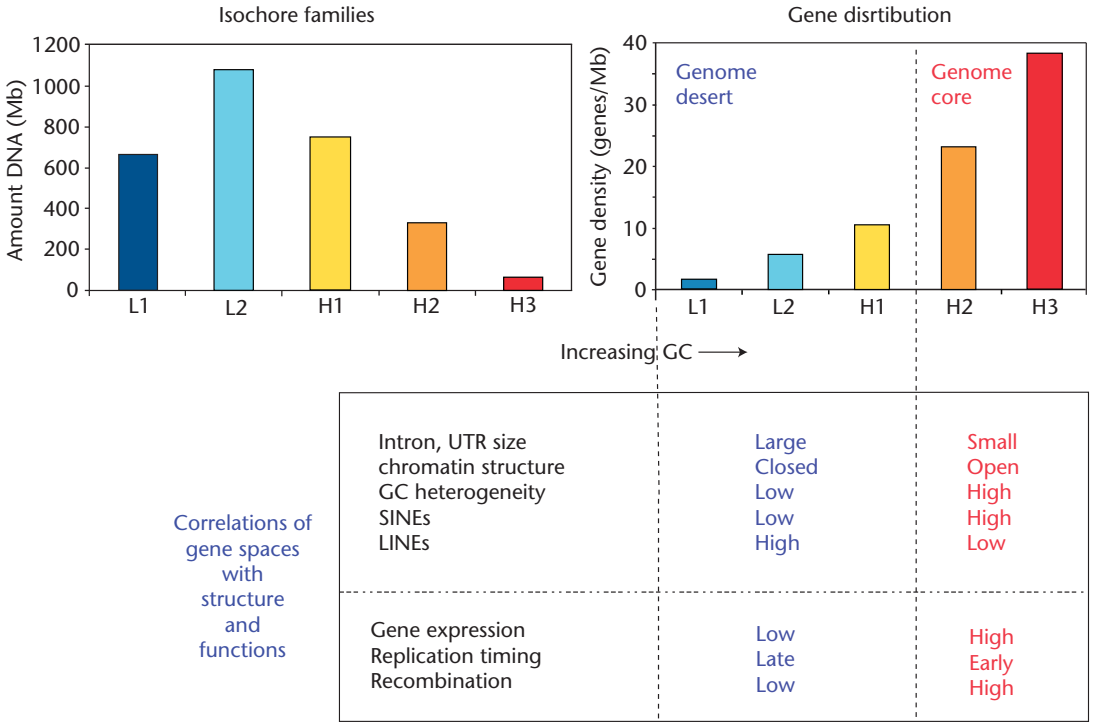

Figure 3 DNA and gene distribution in the isochore families of the human genome. The major structural and functional properties associated with each gene space are listed (in blue for the *genome desert* and in red for the *genome core*). Reproduced from Bernardi G (2007) The neo-selectionist theory of genome evolution. *Proceedings of the National Academy of Sciences of the USA* **104**: 8385–8390.

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)1-4-dioxane), bind to DNA 'molecules' (i.e. DNA fragments 25–100 kb in size) according 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. nonsatellite, nonribosomal) bovine DNA. In fact, this mammalian DNA was shown to comprise a broad spectrum of molecules that were distributed in a small number of families characterized by different base compositions. Further work showed that the genomes of mammals are compositionally compartmentalized, in that they are *mosaics of isochores*, fairly homogeneous regions, originally estimated as longer than 300 kb (isochore is derived from the Greek for compositionally 'equal landscape'). These regions could be assigned to five compositionally narrow families that cover a very wide GC (GC is the molar ratio, namely, the percentage of guanine + cytosine in DNA) range (34–59% GC in the human genome). This discontinuous compartmentalization was in sharp contrast with the then predominant view of a continuous compositional change of DNA along chromosomes.

Recent results have confirmed these conclusions at the sequence level by mapping isochores on human chromosomes, and by providing information on the number (approximately 3200), the average size (approximately 1 Mb) and the GC levels of isochores. The standard deviations of GC levels are around 1% GC within isochores covering 85% of the genome, and around 2% GC in the remaining, mostly GC-rich, isochores. If isochores are pooled in 1% GC bins, their distribution confirms that they belong to the five families previously described (see **Figure 3**). Isochores are, in fact, the ultimate chromosomal bands.

It is generally accepted that the molecular basis of cytogenetic bands is not well understood. Isochores do allow one, however, to define the standard Giemsa and Reverse bands at a 850- or 400-band resolution on purely compositional grounds, thus ending the debate on whether the chromosomal bands are due to DNA, as thought by Caspersson, or to the associated proteins. In addition, the isochore map allowed us to detect a nested structure which concerns not only contiguous isochores, but also contiguous high-resolution bands.

The assessment of gene density in compositional DNA fractions (see **Figure 3**) led to the discovery that genes are not uniformly distributed in the mammalian genome, contrary to the random distribution that was previously visualized. Indeed, in the human genome almost two-thirds of the protein-coding genes are concentrated in the GC-richest isochore families H2 and H3 (only representing 15% of the genome), which was called the *genome core*. The rest is spread over the vast (approximately 85%) GC-poor

genome desert, namely the GC-poor isochore families L1, L2 and H1. These two *gene spaces* are different not only in gene density, but also in a number of other basic properties, which are summarized in **Figure 3**. Recent data have revealed that the isochore map matches the replicon map, and that both isochore size and average GC of isochore families are conserved in vertebrates, reinforcing the concept that isochores represent 'a fundamental level of genome organization'.

Compositional Correlations: The Genomic Code

Positive compositional correlations were found to hold between coding and contiguous noncoding sequences, both intergenic and intragenic. Moreover, GC_1, GC_2 and GC_3 (the GC levels of the three codon positions) were shown to be correlated with each other. Such general rules were called *genomic code* (not to be confused with the *genetic code*), a definition later extended to the correlations between the GC levels of the three codon positions with amino acid composition, hydrophobicity and secondary structures of the encoded proteins. Interestingly, the genomic code allows predicting the composition of flanking sequences and introns from the composition of coding sequences (and vice versa), as well as predicting protein properties from the composition of the corresponding coding sequences (and vice versa). Along the same line, plots of GC_2 versus GC_3 revealed clusters of protein-coding genes corresponding to each isochore family and forming *gene landscapes* that have led to simple proteomic checks for detecting noncoding RNA. Needless to say, all these results provided strong evidence for the genome as an integrated ensemble with little or no room left for what was called *junk* DNA (the non-coding DNA).

In conclusion, the rationale of the compositional approach is that (i) because of their AT bias (namely the predominance of GC→AT over AT→GC changes), mutations and fixations tend to alter the base composition of DNA and (ii) that changes in base composition affect not only DNA structure and stability, but also DNA interactions with proteins and nucleic acids. 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 to the continuous compositional spectrum that prevailed until the 1970s. **See also**: Evolutionary History of the Human Genome; GC-rich Isochores in the Interphase Nucleus; Gene Distribution in Human Chromosomes; Isochores

Further Reading

Bernardi G (2004, reprinted in 2005) *Structural and Evolutionary Genomics. Natural Selection in Genome Evolution*. Elsevier: Amsterdam.

Bernardi G (2007) The neo-selectionist theory of genome evolution. *Proceedings of the National Academy of Sciences of the USA* **104**: 8385–8390.

Costantini M and Bernardi G (2008) Correlations between coding and contiguous non-coding sequences in isochore families from vertebrate genomes. *Gene* **410**: 241–248.

Costantini M and Bernardi G (2008). Replication timing, chromosomal bands and isochores. *Proceedings of National Academy of Sciences USA* (accepted for publication).