Isochores

Giorgio Bernardi, *Rome 3 University, Rome, Italy*

Advanced article

Article Contents

Online posting date: 15th November 2012

Chromosomes of warm-blooded vertebrates are mosaics of isochores, megabase deoxyribonucleic acid (DNA) stretches that are fairly homogeneous in base composition. In the human genome, isochores can be assigned to five families that cover a very broad range of guanine+cytosine (30–60% GC; this range is narrower in cold-blooded vertebrates). The isochore structure of the genome led to new insights into a number of important genome features such as (1) the distribution of genes, (2) the structure of chromatin, and (3) the localisation of insertions/deletions and of the integration of proviral sequences. Moreover it also led to (1) the discovery of a set of correlations (collectively called the genomic code) linking coding sequences and extended flanking noncoding sequences, as well as other genome features, (2) the visualisation of the isochore patterns of the genomes as genome phenotypes, that may be conserved or change in evolution, like the classical phenotypes, and (3) the development of the neoselectionist theory of evolution, which explains the two modes, conservative and shifting, of genome evolution.

The Genome

The term genome was proposed 92 years ago by Winkler (1920) to designate the haploid chromosome set. This definition, originally concerning eukaryotic cells, was forgotten for many years until it was justified by two discoveries; the constant amount of deoxyribonucleic acid (DNA) per haploid cell of a given species (Vendrely and Vendrely, 1948) and the identical functional potential of DNA in all the cells of an organism (Gurdon, 1962).

In the 50s and 60s of the last century, the word genome indicated the sum total of genes, an extrapolation from the prokaryotic to the eukaryotic genome. At the end of the 60s, however, the kinetics of reassociation of DNA, based to a large extent on the separation of double- from single-stranded DNA on hydroxyapatite (Bernardi, 1965), led to the discovery of repeated sequences (Britten and Kohne, 1968), and to the conclusion that genes only represented a part, and often a very small part, of the eukaryotic genome, which was then redefined as the sum total of coding and noncoding sequences. It is now known that in the human genome there are only approximately 21 000 protein coding sequences that represent approximately 1.5% of the genome (Clamp *et al*., 2007).

Although the term genome changed its meaning three times in the first 50 years of its life, it remained the operational definition which it was at the beginning and which still is proposed, often implicitly, in most current textbooks. This was in contrast with the conceptual definition of the gene (Johannsen, 1909), which was viewed as the ultimate unit of inheritance, phenotypic difference and mutation.

A crucial question then is whether the genome is satisfactorily described by an operational definition or is more than the sum of its parts. This dilemma may also be phrased in a different way, namely whether the component parts of the genome are endowed with simple additive properties or with cooperative properties. The first view could be called the operational view of the genome, according to which, although a large amount of details is available, no comprehensive rules about the organisation of genome have emerged so far. The second view would be that the genome is an integrated ensemble, because structural, functional and evolutionary interactions occur among different parts of the genome.

The solution to this problem required a strategy. This was achieved by the development of a compositional approach relying on the most elementary property of DNA, base composition. The rationale was the assumption that the properties of the genome depend upon its nucleotide sequences, both coding and noncoding. Originally, this compositional strategy was based on the fractionation (and characterisation) of large (up to 50 kb) DNA f ragments by preparative ultracentrifugation in density gradients of Cs_2SO_4 run in the presence of a sequence-specific ligand, Ag^+ (Corneo *et al*., 1968). More recently, this approach was based on the analysis of genome sequences.

The Isochores

The compositional strategy led to a breakthrough, namely the demonstration that the genomes of vertebrates (and

eLS subject area: Molecular Biology

How to cite:
Bernardi, Giorgio (November 2012) Isochores. In: eLS. John Wiley & Sons, Ltd: Chichester.
DOI: 10.1002/9780470015902.a0005003.pub2

other eukaryotes) were mosaics of isochores, originally defined as DNA stretches over 300 kb in size, that had a fairly homogeneous guanine + cytosine (GC) level and belonged in a few families endowed with remarkably different properties (Filipski *et al*., 1973; Thiery *et al*., 1976; Macaya *et al*., 1976). It is now known (Costantini *et al*., 2006) that in the human genome the average size of isochores is approximately 1 Mb (one megabase is one million base pairs). It is also known (Costantini *et al*., 2007) that isochores correspond to the ultimate, highest resolution chromosomal bands (approximately 3000 of them were seen in the human genome) and to replication units (Costantini and Bernardi, 2008a). A map of isochores on human chromosome 21 is shown in **Figure 1**. The isochore families of the human genome are shown in **Figure 2**.

In the case of the human genome, a typical mammalian genome, some major observations were made. The first observation was the demonstration that gene density was strikingly nonuniform in the genome, being very low in the GC-poorest isochore family L1 and increasingly higher in the GC-richer families, to reach a maximum in the GC-richest H3 family (Bernardi *et al*., 1985; Mouchiroud *et al*., 1991; Zoubak *et al*., 1996; **Figure 3**). The discontinuity indicated by the two different slopes of gene density versus GC levels among GC-poor families and between them and GC-rich families (**Figure 3**), as well as by the different sizes of isochores and introns (see **Table 1**) identified two genome spaces; a small, GC-rich, gene-rich genome core, consisting of isochore families H2 and H3 and a large GC-poor, gene-poor genome desert, represented by families L1, L2 and H1.

As shown in **Table 1**, a number of other basic properties were found to be associated with the genome core and to be strikingly different (in most cases, just opposite) in the genome desert. In the interphase nuclei of warm-blooded vertebrates, chromatin was 'open' in the GC-rich isochores of the genome, 'closed' in the GC-poor, gene-poor ones (Saccone *et al*., 2002; see **Figure 3**). Interestingly, the gene-rich and gene-poor regions of the genome have different locations, in the central region and at the periphery of the interphase nucleus, respectively.

The genome core also showed an early replication timing (Costantini and Bernardi, 2008a) and a high recombination rate (Fullerton *et al*., 2001; Kong *et al*., 2002). The open chromatin of the genome core is, indeed, the preferential location for the initial integration of retroviral sequences (see Bernardi, 2005, for a review) and Alu sequences (Costantini *et al*., 2012), for the transposition of duplicated genes (Jabbari *et al*., 2003; Rayko *et al*., 2005), for segmental duplications (Jurka *et al*., 2004), for deletions/insertions (Costantini and Bernardi, 2009) and for karyotypic changes (Bernardi, 1993). The frequencies of GC-rich (and, especially, CpG-containing) trinucleotides increases with the increasing GC levels of isochores, whereas those of GC-poor trinucleotides decreases (Costantini and Bernardi, 2008b; Arhondakis *et al*., 2011). In both cases frequencies are strikingly different from the expectations for random sequences having the same GC levels. DNA methylation increased with increasing GC, but less than the methylation acceptor sites, the dinucleotides CpG (Varriale and Bernardi, 2009). The structural and functional properties of isochores shown in **Table 1** indicate that isochores represent a fundamental level of genome organisation (Eyre-Walker and Hurst, 2001).

Isochores also led to new insights in genome structure, function and evolution.

The genomic code

A set of strong correlations (collectively called the genomic code) link (1) the GC levels of coding sequences (long interspersed repeated elements (LINES) and short interspersed repeated elements (SINES)) with those of their extended noncoding flanking sequences (Bernardi *et al*., 1985; Costantini and Bernardi, 2008c); (2) the GC levels of coding sequences with the thermodynamic stability (Bernardi and Bernardi, 1986), the secondary structures

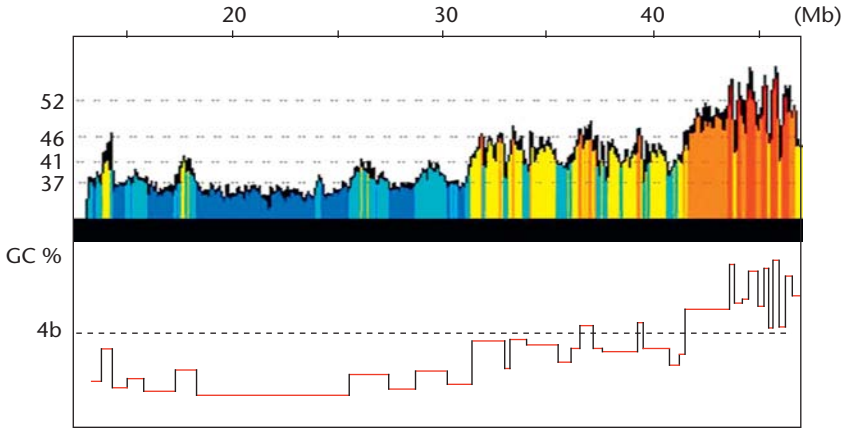

Figure 1 Overview of the isochores of human chromosome 21. The top frame shows the compositional profile at a 100 kb resolution. Broken horizontal guidelines indicate GC levels. In the bottom frame the horizontal red stretches represent isochores. (From data of Costantini *et al*., 2006.)

(Chiusano *et al*., 1999, 2000), and hydropathy (D'Onofrio *et al*., 1999) of proteins and (3) and the GC levels of noncoding and regulatory sequences with their short sequence frequencies (Bernardi *et al*., 1973; Costantini and Bernardi, 2008b). The existence of these correlations is very important because (1) it rules out the hypothesis of 'junk DNA' (Ohno, 1972) and of 'selfish DNA' (Doolittle and Sapienza, 1980; Orgel and Crick, 1980), (2) it connects the nucleotide coding sequence with the detailed structure and the stability of proteins and (3) it links the GC levels of noncoding and regulatory sequences with nucleosome positioning and transcription factor binding, that is, the regulation of gene expression (Arhondakis *et al*., 2011). The latter point is very important because changes in the organismal phenotype depend much more on changes in regulatory sequences, than on changes in protein coding genes as first proposed a long time ago (Zuckerkandl and Pauling, 1965; King and Wilson, 1975) and now well established (see, for example, Kasowski *et al*., 2010). We also know that *cis*-regulatory sequences and chromatin structure are two major factors that influence gene expression, and both act via protein/DNA interaction. Although on the protein side, transcription factors are concerned in the first case and histones in the second, on the DNA side, short nucleotide sequences are involved in both cases. Indeed, the different frequencies of trinucleotides in different isochore families (Costantini and Bernardi, 2009) and specifically in the corresponding regulatory sequences imply a different regulation of gene expression in different isochore families (Arhondakis *et al*., 2011). This is interesting because different functional gene classes are differentially distributed in isochore families (see D'Onofrio *et al*., 2007), and because it suggests possible coregulations of genes located in the same isochore family. In general, housekeeping genes tend to be located in GC-rich isochores, whereas genes associated with development are preferentially located in GC-poor isochore families (Hiratani *et al*., 2004; Ren *et al*., 2007; Kikuta *et al*., 2007; Navratilova and Becker, 2009; see Arhondakis *et al*., 2011, for a review) that are compacted into 'closed' chromatin at the end of development.

Table 1 The genome core: structural and functional properties[a]

GC level	High
Isochore size	Small
Gene density	High
Chromatin structure	Open
Nuclear location	Central
Intron size	Small
'GC-rich triplets'	+
'GC-poor triplets'	−
DNA methylation	−
LINES	−
SINES	+
DNA duplications	+
Insertions/deletions	+
Provirus integration	+
Recombination rate	+
Replication	Early

[a]'+' and '−' signs have obvious meanings.

The neoselectionist theory

The isochore patterns are genome phenotypes; like organismal phenotypes they may evolve according to a conservative or to a shifting mode (**Figure 4**). The conservation of the genome phenotype over extended evolutionary times (e.g. 100 million years in the case of mammalian genomes)

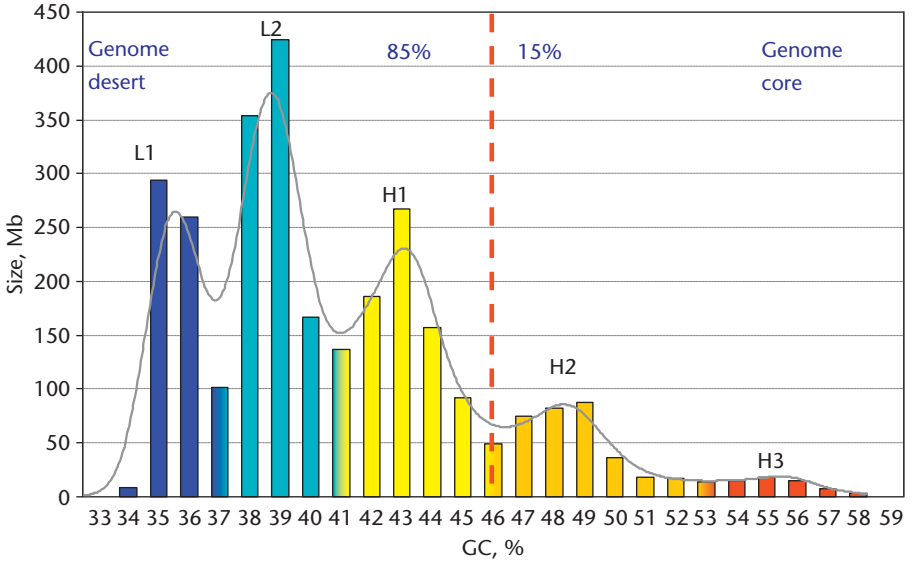

Figure 2 The histogram shows the isochores as pooled in bins of 1% GC. Colours represent isochore families. The Gaussian profile shows the distribution of isochore families, namely the genome phenotype. Reproduced from Costantini *et al.* (2006) with permission of Cold Spring Harbor Laboratory Press.

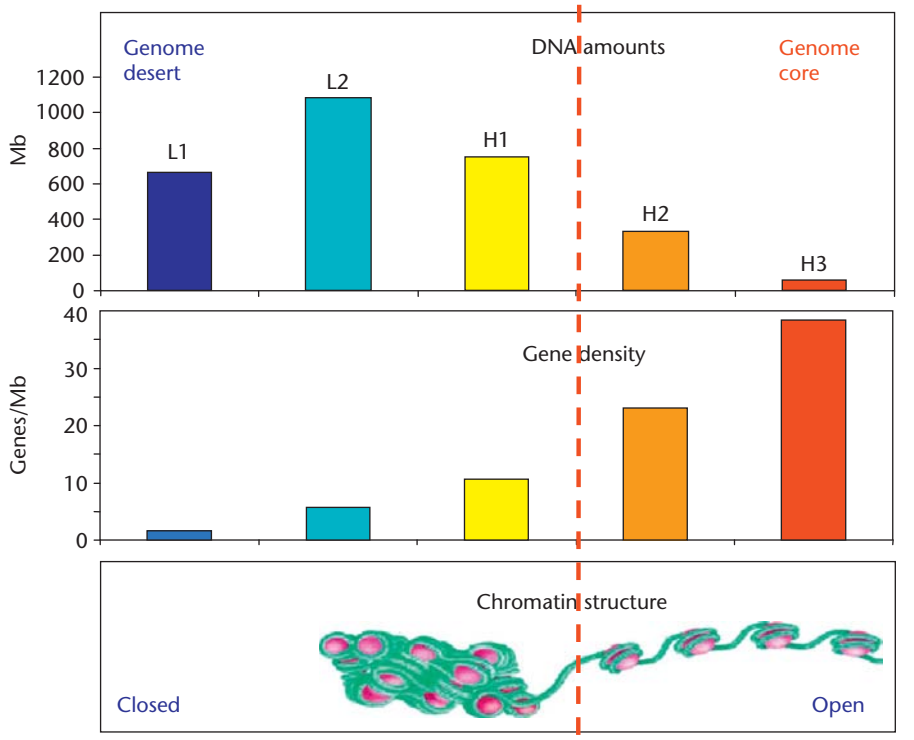

Figure 3 A scheme of the major properties (DNA amounts, gene density and chromatin structure) of the isochore families belonging in the two genome spaces: the genome core and the genome desert. Gene densities show two different slopes, a shallow one for GC-poor isochores and a steep one between GC-poor and GC-rich isochores. They cross each other between H1 and H2 isochores, so defining the two genome spaces.

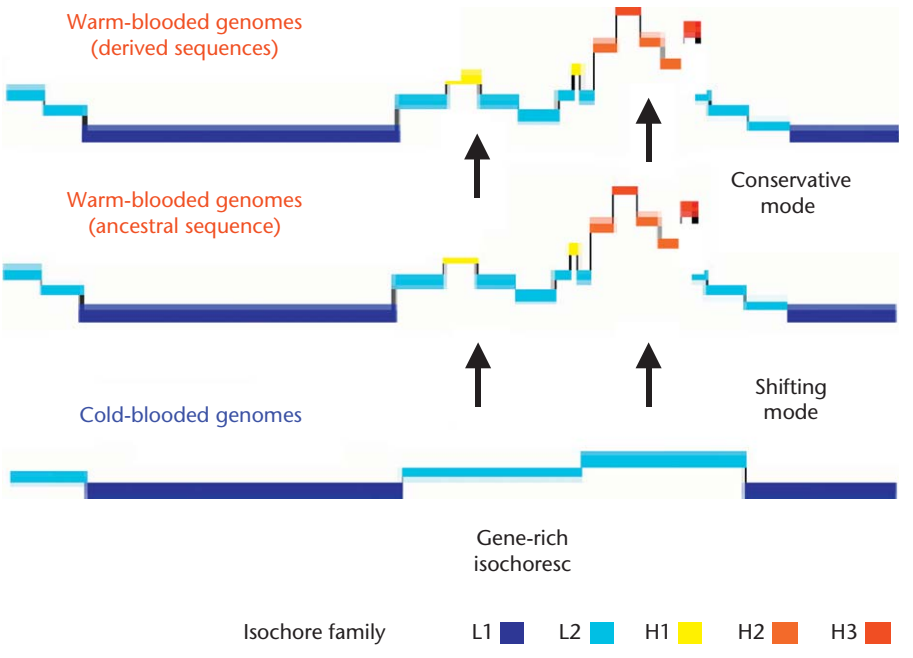

Figure 4 The two modes of compositional evolution of the vertebrate genome. In the shifting mode, the modestly GC-rich isochores of the genomes from cold-blooded vertebrates show a GC increase to become the GC-rich isochores of the warm-blooded vertebrates. The further evolution of the latter from the ancestral sequence is characterised by a conservative mode. The GC-poor regions of the genomes of cold-blooded vertebrates do not undergo major changes.

was explained by the neoselectionist theory (Bernardi, 2007) as due to the fact that insertions/deletions and the clustering of point mutations (that are always adenine + thymine (AT)-biased, GC→AT predominating over AT→GC changes) lead to chromatin alterations that affect gene expression and, as a consequence, to negative selection of the carriers. The shifting mode is exemplified by the genome changes that took place between cold- and warm-blooded vertebrates and led from a modestly heterogeneous genome to a strikingly heterogeneous one (**Figure 4**). Positive selection may be the main explanation for the shifting of the genome phenotype of cold-blooded to that of warm-blooded vertebrates.

Conclusions

In conclusion, the genome is an integrated ensemble which obeys a set of rules, the genomic code, and which encodes not only aminoacid sequences (according to the genetic code) and the secondary structure of proteins but also the regulation of gene expression. The high level of order implied by the genomic code is preserved by natural selection, mainly negative selection. This is Darwinism at the genomic level. **See also**: Base Composition Patterns; Biased Gene Conversion and its Impact on Human Genome Evolution; Evolutionary History of the Human Genome; GC-rich Isochores in the Interphase Nucleus; Gene Distribution in Human Chromosomes; Genome Organization of Vertebrates; The Neo-selectionist Theory of Genome Evolution

References

Arhondakis S, Auletta F and Bernardi G (2011) Isochores and the regulation of gene expression in the human genome. *Genome Biology and Evolution* **3**: 1080–1089.

Bernardi G (1965) Chromatography of nucleic acids on hydroxyapatite. *Nature* **206**: 779–783.

Bernardi G (1993) Genome organization and species formation in vertebrates. *Journal of Molecular Evolution* **37**: 331–337.

Bernardi G (2005) *Structural And Evolutionary Genomics: Natural Selection In Genome Evolution*. Amsterdam: Elsevier. This book is freely available at www.giorgiobernardi.it.

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

Bernardi G and Bernardi G (1986) Compositional constraints and genome evolution. *Journal of Molecular Evolution* **24**: 1–11.

Bernardi G, Ehrlich SD and Thiery JP (1973) The specificity of deoxyribonucleases and their use in nucleotide sequence studies. *Nature New Biology* **246**: 36–40.

Bernardi G, Olofsson B, Filipski J *et al*. (1985) The mosaic genome of warm-blooded vertebrates. *Science* **228**: 953–958.

Britten RJ and Kohne DE (1968) Repeated sequences in DNA. *Science* **161**: 529–540.

Chiusano ML, Alvarez-Valin F, Di Giulio M *et al*. (2000) Second codon positions of genes and the secondary structures of proteins. Relationships and implications for the origin of the genetic code. *Gene* **261**: 63–69.

Chiusano ML, D'Onofrio G, Alvarez-Valin F *et al*. (1999) Correlations of nucleotide substitution rates and base composition of mammalian coding sequences with protein structure. *Gene* **238**: 23–31.

Clamp M, Ben F, Kamal MX, *et al*. (2007) Distinguishing protein-coding and noncoding genes in the human genome. *Proceedings of the National Academy of Sciences of the USA* **104**: 19428–19433.

Corneo G, Ginelli E, Soave C *et al*. (1968) Isolation and characterization of mouse and guinea pig satellite DNAs. *Biochemistry* **7**: 4373–4379.

Costantini M, Auletta F and Bernardi G (2012) The distribution of 'new' and 'old' Alu sequences in the human genome: the solution of a 'mystery'. *Genome Biology and Evolution* **29**: 421–427.

Costantini M and Bernardi G (2008a) Replication timing, chromosomal bands and isochores. *Proceedings of the National Academy of Sciences of the USA* **105**: 3433–3437.

Costantini M and Bernardi G (2008b) The short sequence design of isochores from the human genome. *Proceedings of the National Academy of Sciences of the USA* **105**: 13971–13976.

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

Costantini M and Bernardi G (2009) Mapping insertions, deletions and SNPs on Venter's chromosomes. *PLoS ONE* **4**(6): e5972.

Costantini M, Clay O, Auletta F *et al*. (2006) An isochore map of human chromosomes. *Genome Research* **16**: 536–541.

Costantini M, Clay O, Federico C *et al*. (2007) Human chromosomal bands: nested structure, high-definition map and molecular basis. *Chromosoma* **116**: 29–40.

D'Onofrio G, Ghosh TC and Saccone S (2007) Different functional classes of genes are characterized by different compositional properties. *Federation of European Biochemical Societies Letters* **581**: 5819–5824.

D'Onofrio G, Jabbari K, Musto H *et al*. (1999) The correlation of protein hydropathy with the composition of coding sequences. *Gene* **238**: 3–14.

Doolittle WF and Sapienza C (1980) Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284**: 601.

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

Filipski J, Thiery JP and Bernardi G (1973) An analysis of the bovine genome by $Cs_2SO_4Ag^+$ density gradient centrifugation. *Journal of Molecular Biology* **80**: 177–197.

Fullerton SM, Carvalho AB and Clark AG (2001) Local rates of recombination are positively correlated with GC content in the human genome. *Molecular Biology and Evolution* **18**: 1139–1142.

Gurdon JB (1962) The developmental capacity of nuclei taken from intestinal epithelium cells of feeding tadpoles. *Journal of Embryology and Experimental Morphology* **10**: 622–640.

Hiratani I, Leskovar A and Gilbert DM (2004) Differentiation-induced replication-timing changes are restricted to AT-rich/long interspersed nuclear element (LINE)-rich isochores. *Proceedings of the National Academy of Sciences of the USA* **101**: 16861–16866.

Jabbari K, Cruveiller S, Clay O *et al*. (2003) The correlation between GC3 and hydropathy in human genes. *Gene* **317**: 137–140.

Johannsen W (1909) *Elemente der exakten Erblichkeitslehre*. Jena: Fischer.

Jurka J, Kohany O, Pavlicek A *et al*. (2004) Duplication, coclustering, and selection of human Alu retrotransposons. *Proceedings of the National Academy of Sciences of the USA* **101**: 1268–1272.

Kasowski M, Grubert F, Heffelginger C *et al*. (2010) Variation in transcription factor binding among humans. *Science* **328**: 232–235.

Kikuta H, Laplante M, Navratilova P *et al*. (2007) Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Research* **17**: 545–555.

King MC and Wilson AC (1975) Evolution at two levels in humans and chimpanzees. *Science* **188**: 107–116.

Kong A, Gudbjartsson DF, Sainz J *et al*. (2002) A high-resolution recombination map of the human genome. *Nature Genetics* **31**: 241–247.

Macaya G, Thiery JP and Bernardi G (1976) An approach to the organization of eukaryotic genomes at a macromolecular level. *Journal of Molecular Biology* **108**: 237–254.

Mouchiroud D, D'Onofrio G, Aissani B *et al*. (1991) The distribution of genes in the human genome. *Gene* **100**: 181–187.

Navratilova P and Becker TS (2009) Genomic regulatory blocks in vertebrates and implications in human disease. *Briefings in Functional Genomics* **8**: 333–342.

Ohno S (1972) So much 'junk' DNA in our genome. *Brookhaven Symposia in Biology* **23**: 366–370.

Orgel LE and Crick FHC (1980) Selfish DNA. *Nature* **284**: 604.

Rayko E, Jabbari K and Bernardi G (2005) The evolution of introns in human duplicated genes. *Gene* **365**: 41–47.

Ren L, Gao G, Zhao D *et al*. (2007) Developmental stage related patterns of codon usage and genomic GC content: searching for evolutionary fingerprints with models of stem cell differentiation. *Genome Biology* **8**: R35.

Saccone S, Federico C, Andreozzi L *et al*. (2002) Localization of the gene-richest and the gene-poorest isochores in the interphase nuclei of mammals and birds. *Gene* **300**: 169–178.

Thiery JP, Macaya G and Bernardi G (1976) An analysis of eukaryotic genomes by density gradient centrifugation. *Journal of Molecular Biology* **108**: 219–235.

Varriale A and Bernardi G (2009) Distribution of DNA methylation, CpGs, and CpG islands in human isochores. *Genomics* **95**: 25–28.

Vendrely R and Vendrely C (1948) La teneur de noyau cellulaire en AND à travers les organs, les individus et les espèces animales. *Experientia* **4**: 434–436.

Winkler H (1920) *Verbreitung und Ursache der Parthenogenesis im Pflanzen und Tierreich*. Jena: Fischer.

Zoubak S, D'Onofrio G, Cacciò S *et al*. (1996) Specific compositional pattern of synonymous positions in homologous mammalian genes. *Journal of Molecular Evolution* **40**: 293–307.

Zuckerkandl E and Pauling L (1965) Evolutionary divergence and convergence in proteins. In: Bryson V and Vogel HJ (eds) *Evolving Genes and Proteins*, pp. 97–166. New York: Academic Press.