The short-sequence designs of isochores from the human genome

Maria Costantini and Giorgio Bernardi*

Laboratory of Molecular Evolution, Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy

Edited by Francisco J. Ayala, University of California, Irvine, CA, and approved July 14, 2008 (received for review April 24, 2008)

The human genome, a typical mammalian genome, is made up of long (≈1-Mb, on average) regions, the isochores, that are fairly homogeneous in base composition and belong in five families characterized by different GC levels. An analysis of di- and tri-nucleotide densities in the isochores from the five families has shown large differences. These different "short-sequence designs:" (*i*) account for the fractionation of human DNA (and vertebrate DNA in general) when using sequence-specific ligands in density gradients, (*ii*) are very similar in whole isochores and in the corresponding intergenic sequences and introns, (*iii*) are reflected in different codon usages, (*iv*) lead to amino acid differences that increase the thermal stability of the proteins encoded by genes located in increasingly GC-rich isochore families, and (*v*) correspond to different chromatin structures.

amino acids | chromatin structure | codon usage | dinucleotides | trinucleotides

Forty years ago, the complete separation of major satellites from the "main-band" DNAs of mouse and guinea pig was achieved by ultracentrifugation in Cs_2SO_4/Ag^+ density gradient (1). The two satellites differed by only 3% GC (molar fraction of guanine and cytosine in DNA), yet the mouse satellite was very "light" (ρ° = 1.456 g/cm^3) and the guinea pig satellite very "heavy" (ρ° = 1.534 g/cm^3) in Cs_2SO_4/Ag^+. In contrast, both main bands were centered at an intermediate density, 1.500 g/cm^3, and were very broad. Because the resolving power of Cs_2SO_4 density gradient *per se* is even lower than that of CsCl (where both satellites appear only as shoulders on the main bands), this was a clear indication that the basis for the wide separation of the two satellites from the main bands was the differential binding of silver ions to their short internal repeats. Moreover, the large spreads of the main bands in Cs_2SO_4/Ag^+ suggested that they were compositionally complex. Indeed, when the Cs_2SO_4/Ag^+ approach was used to investigate the bovine genome, this not only led to a separation of four satellites, but also to the fractionation of three "major DNA components" that formed the main band (2).

The observations concerning the main band of the bovine genome were then shown to be valid for most mammalian DNAs (3). Using Cs_2SO_4 density gradient and another sequence-specific ligand, bis(acetato mercury methyl)dioxane (BAMD), the human genome, a typical mammalian genome, could be fractionated in a DNA size range of 25–100 kb. This led to the identification of five major components, L1, L2, H1, H2, and H3, in order of increasing GC levels (only three major components were isolated originally, because L1 and L2 had not been separated in the Cs_2SO_4/Ag^+ gradients, and H3 had not been identified as a major component, because it was present in small amounts). Moreover, it was discovered that the 25- to 100-kb DNA molecules that were fractionated derived from isochores, the compositionally fairly homogeneous chromosomal regions that were initially estimated as >300 kb (4) and are now known to have an average size of ≈1 Mb (megabase; see ref. 5).

A few years ago, contiguous human DNA sequences having a size of 20 kb and derived from 300-kb stretches were shown to be characterized by large standard deviations of GC, which were well above the standard deviation of random sequences, a finding purported to put in question the reality of isochores (6). In fact, random sequences were well known (7–10) to be much

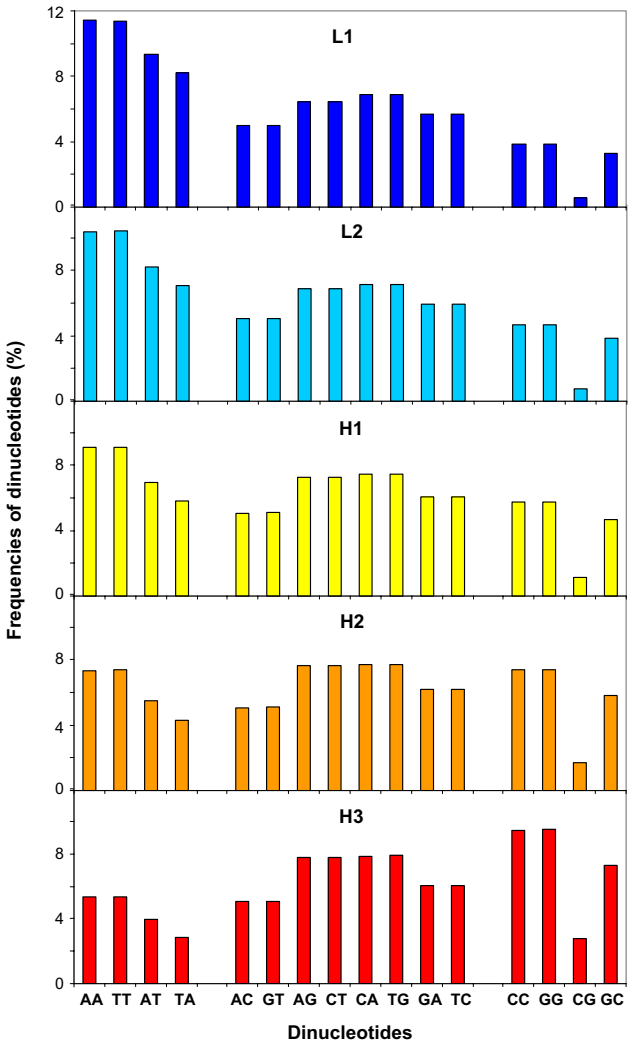

Fig. 1. Frequency of dinucleotides per 100-kb DNA sequences from the five isochore families. Frequencies are calculated as percentages of the total per family in Figs. 1–3.

Author contributions: M.C. performed research; M.C. and G.B. analyzed data; and G.B. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

*To whom correspondence should be addressed. E-mail: bernardi@szn.it.

This article contains supporting information online at www.pnas.org/cgi/content/full/0803916105/DCSupplemental.

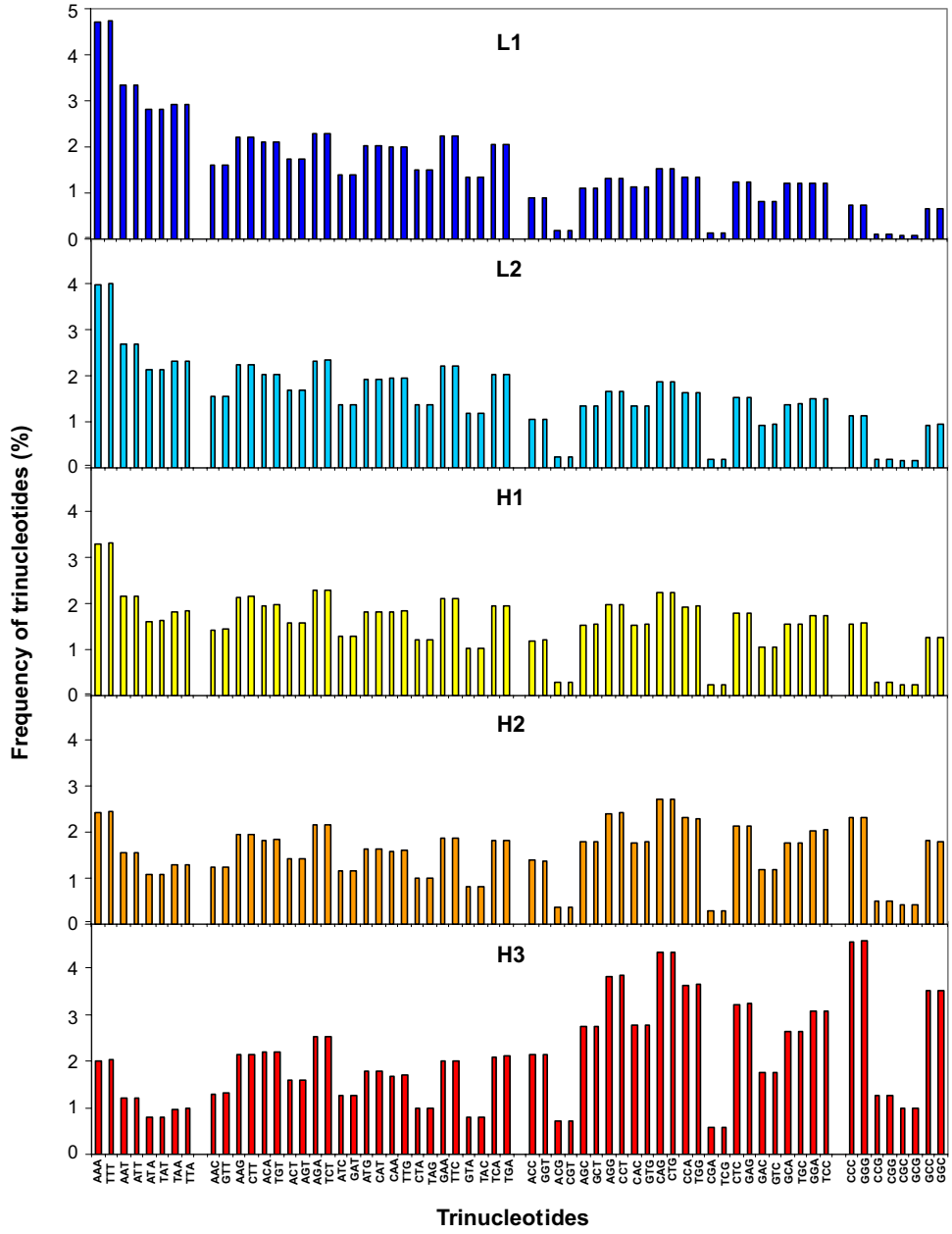

Fig. 2. Frequency of trinucleotides per 100-kb DNA sequences derived from the five isochore families.

more homogeneous than the least heterogeneous natural DNAs, those of prokaryotes (which are, in turn, much less heterogeneous than eukaryotic DNAs). The large standard deviations reported (6) are now explained by the small size, 20 kb, of the DNA sequences investigated. Indeed, Costantini *et al.* (5) found that the standard deviation of GC in DNA sequences could reach a (low) plateau region only above a size of 100 kb. Below this value, standard deviations increase with decreasing sequence size because of the increasing contributions of coding sequences and, especially, of interspersed repeats (see figure 2 of ref. 5), which are characterized by their own compositional properties.

Although the above observations dissipated the doubts raised by Lander *et al.* (6) about the very existence of isochores, they did not explain how the approach used could fractionate five families of DNA fragments in the 25- to 100-kb size range (see refs. 11 and 12). A possible explanation, based on our previous investigations on satellite DNAs (see above), was that the gradient fractionation occurred because of different sequence-specific ligand densities on DNA fragments from the main band. In turn, the different ligand densities were likely due to different distributions of short sequences on DNA fragments from different isochore families. This led us to explore the di- and tri-nucleotide densities on 100-kb DNA sequences derived from the five isochore families. This work not only solved the puzzle of main band DNA fractionation but, more importantly, provided information on the ”short-sequence designs“ (8) of isochores and on their implications.

Results

Densities of di- and tri-nucleotides were assessed on human DNA sequences 100 kb in size as derived from different isochore families. The comparison of such densities (Figs. 1 and 2) showed a number of differences. Indeed, among dinucleotides, the ‘‘AT set,’’ ApA, TpT, ApT, and TpA, showed a remarkable decrease

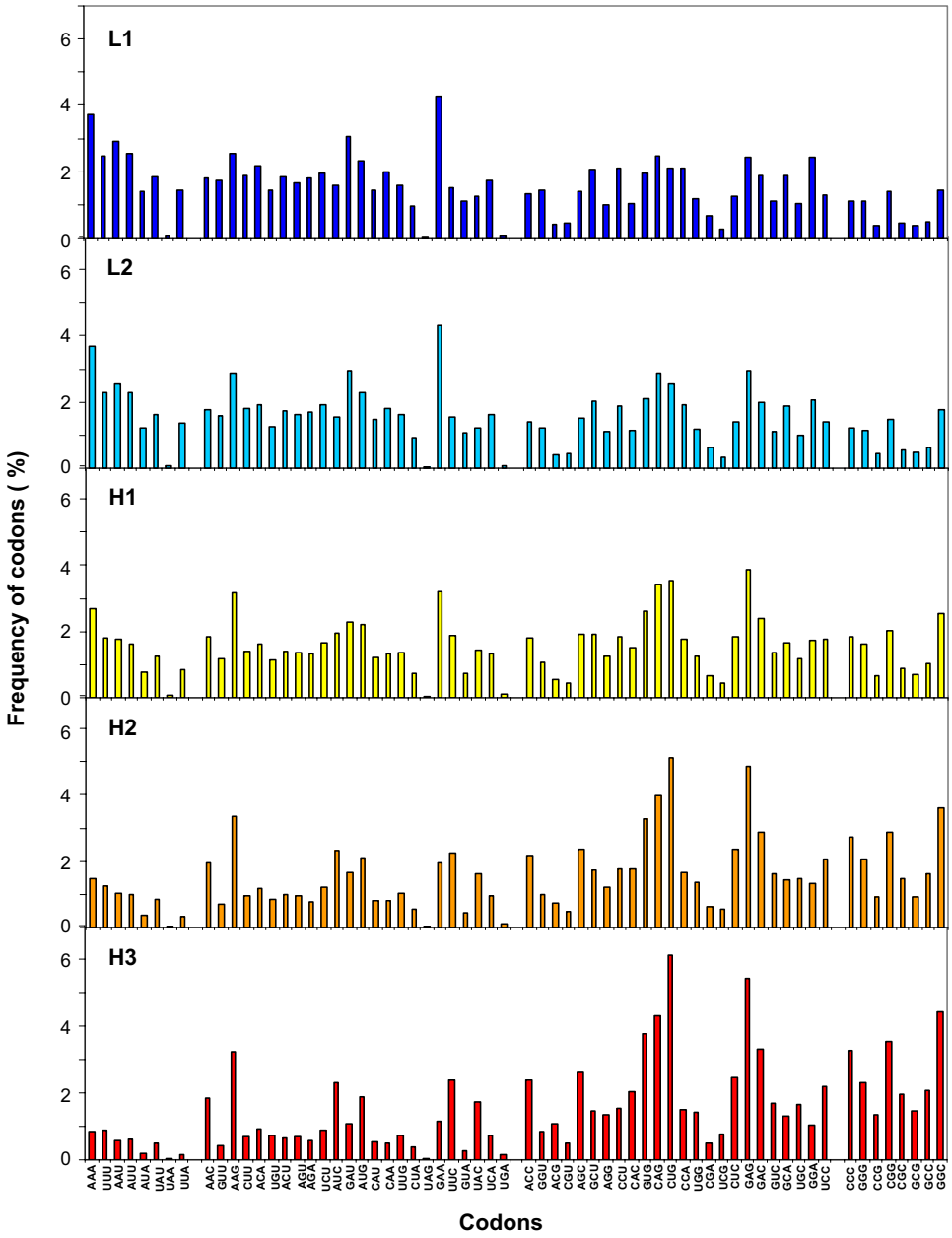

Fig. 3. Frequency of codons in genes located in different isochore families.

from the L1 to H3 families. In contrast, the "GC set," CpC, GpG, CpG, and GpC, showed an increase, the CpG density reaching a 5-fold higher level in H3 compared with L1 isochores. In the case of trinucleotides, those containing the "AT set" of dinucleotides also showed a decrease when moving from GC-poor to -rich isochores, whereas those comprising the "GC set" showed the opposite trend. For example, the CGC density was >12-fold higher in H3 compared with L1 isochores.

The results just presented prompted an analysis of di- and tri-nucleotides in intergenic sequences, introns, and exons from different isochore families. In the case of intergenic sequences, frequency patterns were practically identical to those just reported for sequences from whole isochore families [see supporting information (SI) Fig. S1 *A* and *B*], as expected from their abundance in isochores (Figs. 1 and 2). In the case of introns, some small differences were observed, such as ApA<TpT, ApC<GpT, AAA<TTT, ACA<TGT, etc. (see Fig. S2 *A* and *B*), possibly due to the biased representation of some dinucleotides in the small-size introns.

In the case of exons, the codon frequency distribution was expected to be different from that of trinucleotides from the corresponding families and again different for different isochore families (Fig. 3). When individual codon positions were assessed in terms of nucleotide composition for isochore families of increasing GC, one could see, however, a strong decrease in A and T and an increase in G and C in third codon positions. At a progressively lesser extent, such changes could also be seen in first and second codon positions (see Fig. 4).

Because changes in second codon positions are strongly correlated with changes in the encoded amino acids, we also analyzed the frequencies of amino acids corresponding to genes located in different isochore families. This showed (see Fig. 5) that some amino acids, especially alanine and arginine but also glycine and proline (all corresponding to codons with G or C in

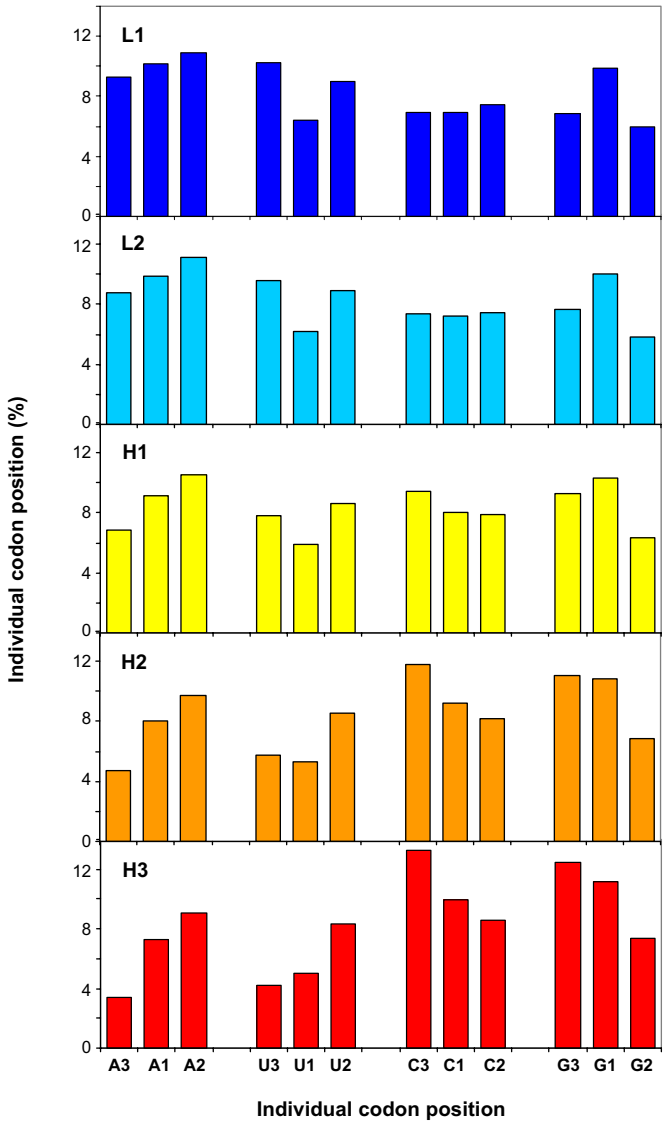

Fig. 4. Individual codon positions assessed in terms of nucleotide composition for isochore families of increasing GC.

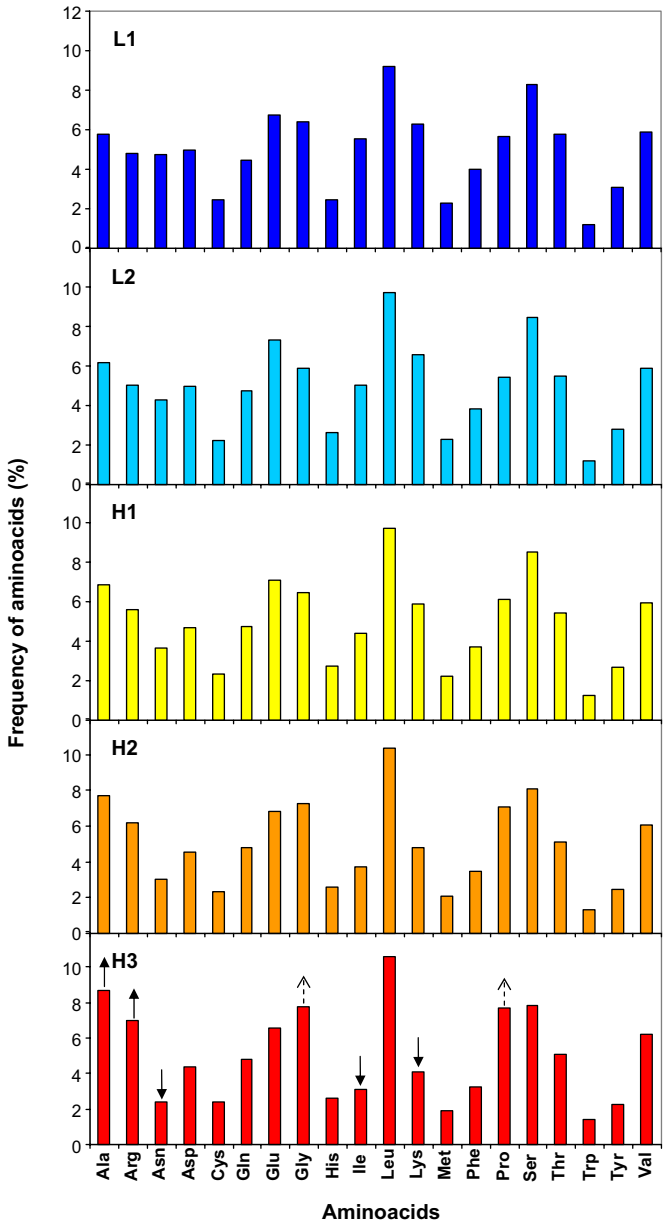

Fig. 5. Frequencies of amino acids encoded by genes located in different isochore families. The black or dashed upward or downward arrows in the H3 image indicate amino acids that increase or decrease in frequency from L1 to H3 (≥30% and ≥15%, respectively).

their second positions) showed an increase, whereas others, especially lysine, isoleucine, and asparagine (all corresponding to codons with A in their second positions), showed a decrease in proteins encoded by genes located in isochore families of increasing GC.

Discussion

The nearest-neighbor analysis showed that the frequencies of nucleotide doublets were usually close to those expected from a random distribution of nucleotides in prokaryotes and in most eukaryotes. Remarkable exceptions were the dinucleotides CpG and TpA, which showed a strong and a moderate shortage, respectively, in the genomes of vertebrates (13–15) and were discussed elsewhere (16–20).

Our basic observation that the densities of di- and trinucleotides from different isochore families of the human genome are different provides much more information (see below) than the results on whole genomes just mentioned. Incidentally, the first indication of such differences was obtained by finding different frequencies of A, T, G, and C in the terminal nucleotides of DNA fragments as released by spleen and snail DNAses from the major components of the human genome (ref. 21; see figure 3.8 of ref. 17).

The main conclusions reached in this work can be summed up and commented on as follows: (*i*) The di- and tri-nucleotide densities of Figs. 1 and 2 do account for the observation that vertebrate DNA can be fractionated in a Cs_2SO_4/BAMD density gradient. Indeed, although the BAMD-binding oligonucleotides have not been identified, we know that the GC-poor DNA molecules bind more BAMD and become "heavy" in the Cs_2SO_4 gradient. In other words, the specific oligonucleotide frequencies of DNA segments from different isochore families are indeed responsible for the fractionation achieved by using the sequence-specific ligand BAMD in a Cs_2SO_4 density gradient. Because the critical factor is the density of binding sites on DNA, it is understandable that fractionation is independent of sequence size in the 25- to 100-kb range.

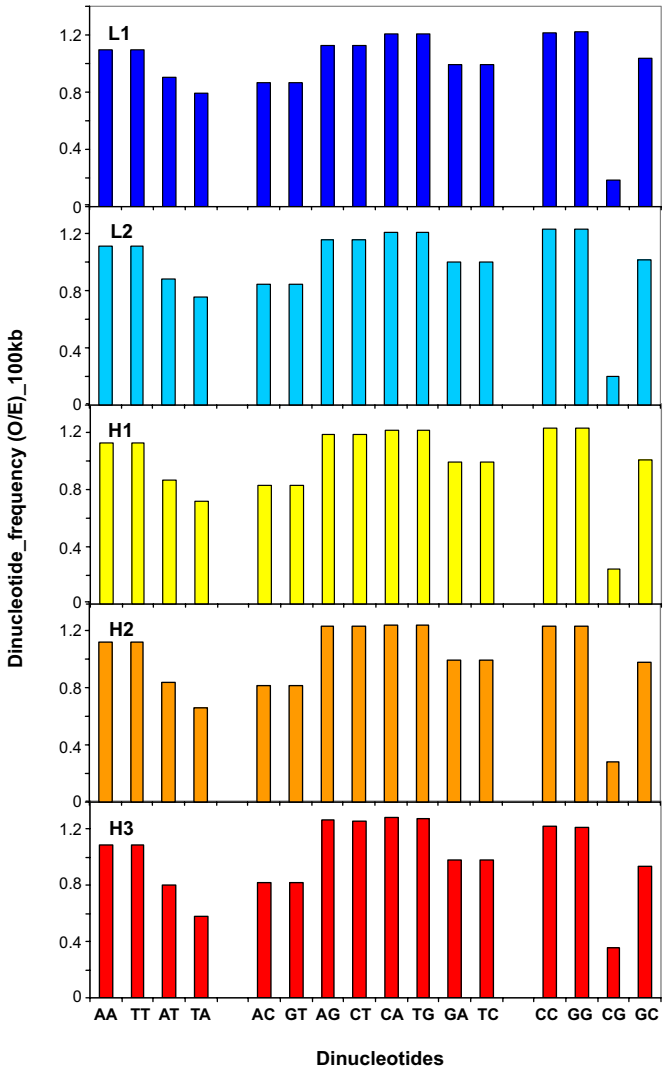

Fig. 6. Observed/expected frequencies for dinucleotides in 100-kb DNA segments.

(*ii*) The point just made has a general relevance, because it stresses that the short-range interactions of DNA (e.g., the interactions with a sequence-specific ligand) essentially depend on the actual frequency of short sequences in the isochore belonging to different families and not on observed/expected frequency ratios, even if the latter show some significant changes from one family to the next, the 2-fold increase in CpG (already observed in ref. 11 and further studied in refs. 19–21), the 1.4-fold decrease of TpA (see Fig. 6) being particularly striking. Incidentally, the observed/expected ratios for the dinucleotides from a given genome, which have been called "general design" (see ref. 15) and "genome signature" by Karlin and Burge (22), average the CpG and ApT ratios obtained for different isochore families. In addition, the conclusion of Gentles and Karlin (23) that "with minor exceptions, all dinucleotide biases are clearly invariant both across and between chromosomes" not only is incorrect but also is contradicted by the finding (23) that the CpG and TpA densities on chromosome 21 do show variations that match the isochore map of this chromosome based on GC levels (5).

(*iii*) The frequencies of di- and tri-nucleotides provide a characteristic pattern not only for different isochore families but also for individual isochores belonging to different families, as shown by di- and tri-nucleotide analyses and by the very low standard deviations exhibited by each dinucleotide within a family (see Table S1). In other words, the "short-sequence designs" do characterize isochores originating from different families. Moreover, the weight average of the di- and tri-nucleotide frequencies, as assessed for different isochore families, perfectly matches the patterns of the human genome as shown by reconstruction experiments (data not shown).

(*iv*) The large differences in codon frequency distribution in different isochore families (Fig. 3) and the differences in the observed frequencies of a codon divided by the frequencies expected if all synonymous codons are used equally, Relative Synonymous Codon Usage [(RSCU); see Fig. S3] indicate strong changes in codon usage for genes located in different isochore families. These are the "compositional constraints" on codon usage first reported by Bernardi *et al.* (ref. 11; see also ref. 17 for a review). Moreover, they lead to the expectation (which is fulfilled; see Fig. 5) that the frequencies of amino acids encoded by the genes located in different isochore families are different. It is remarkable that amino acids that are supposed to confer thermal stability to proteins (24, 25), alanine, arginine, and, to a lesser extent, glycine and proline, increase, whereas those leading to lower stability, like lysine, asparagine, and isoleucine, decrease when encoded by genes located from L1 to H3. This point, already made by Bernardi and Bernardi (26) in their "thermodynamic stability hypothesis," which linked the higher stability at increasing GC of DNA and RNA with that of the encoded proteins, is now confirmed on the basis of all of the proteins encoded by the 18,796 human genes analyzed here.

(*v*) Although the importance of dinucleotide properties, in particular stacking energies, for local DNA structure has been known for a long time (27, 28), that of periodicities of ApA/TpT/TpA and GpC in connection with position and stability of nucleosomes has been stressed only recently (29). The different densities of di- and tri-nucleotides suggest that chromatin structure may be different at the level of isochores belonging to different families. This has been demonstrated by mapping DNase-I hypersensitive sites and showing that the density of these sites on the human genome increases with increasing GC of isochores (30). At a larger scale, we already know that the GC-richest and -poorest chromosomal regions have a very different compaction, the former ones corresponding to "open chromatin," the latter ones to "closed chromatin" (31).

In conclusion, the different short-sequence designs of the isochores from the human genome, which is a good model for all warm-blooded vertebrates (see ref. 17), has a strong effect (*i*) on codon usage (first reported in ref. 11); (*ii*) on the properties of the encoded proteins; (*iii*) on chromatin structure and, as a consequence, on gene expression (see ref. 32 for a review), and possibly, on replication timing (33).

Methods

Analysis of Di- and Tri-Nucleotides. The entire chromosomal sequences of the finished human genome assembly (UCSC release hg17; refs. 34 and 35) were partitioned into nonoverlapping 100-kb windows. The number of di- and tri-nucleotides was calculated with a script implemented by us for each segment of 100 kb. The frequency of each di- and tri-nucleotide was evaluated from its percentage in each isochore family: 36.0% in L1, 38.9% in L2, 43.1% in H1, 48.7% in H2, and 54.5% in H3 (see ref. 5). The same analysis was performed on intergenic and intronic sequences. The ratio of observed versus expected frequency was also calculated.

Analysis of Codons and Amino Acid Residues. The number of codons in the genes located in the five isochore families were calculated by using the CodonW program. The RSCU value, namely the observed frequency of a codon divided by the frequency expected if all synonymous codons for that amino acid were used equally, was calculated according to Sharp and Li (36).

The number of all amino acids was calculated by using a script implemented

by us for the genes located in different isochore families. The frequency of each amino acid was evaluated from its percentage in the amino acids encoded in each isochore family.

ACKNOWLEDGMENTS. We thank Oliver Clay for very helpful discussions and Kamel Jabbari for comments. We thank also Fabio Auletta and Giuseppe Torelli for bioinformatic support.

1. Corneo G, Ginelli E, Soave C, Bernardi G (1968) Isolation and characterization of mouse and guinea pig satellite DNAs. *Biochemistry* 7:4373–4379.
2. Filipski J, Thiery JP, Bernardi G (1973) An analysis of the bovine genome by $Cs_2SO_4{}^-Ag^+$ density gradient centrifugation. *J Mol Biol* 80:177–197.
3. Thiery JP, Macaya G, Bernardi G (1976) An analysis of eukaryotic genomes by density gradient centrifugation. *J Mol Biol* 108:219–235.
4. 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.
5. Costantini M, Clay O, Auletta F, Bernardi G (2006) An isochore map of human chromosomes. *Genome Res* 16:536–541.
6. Lander ES, *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409:860–921.
7. Rolfe R, Meselson M (1959) The relative homogeneity of microbial DNA. *Proc Natl Acad Sci USA* 45:1039–1043.
8. Hudson AP, Cuny G, Cortadas J, Haschemeyer AEV, Bernardi G (1980) An analysis of fish genomes by density gradient centrifugation. *Eur J Biochem* 112:203–210.
9. 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* 111:227–233.
10. Bernardi G (2001) Misunderstandings about isochores. *Gene* 276:3–13.
11. Bernardi G, *et al.* (1985) The mosaic genome of warm-blooded vertebrates. *Science* 228:953–958.
12. Zoubak S, Clay O, Bernardi G (1996) The gene distribution of the human genome. *Gene* 174:95–102.
13. Josse J, Kaiser AD, Kornberg A (1961) Enzymatic synthesis of deoxyribonucleic acid. *J Biol Chem* 236:864–875.
14. Swartz MN, Trautner TA, Kornberg A (1962) Enzymatic syntesis of deoxyribonucleic acid. *J Biol Chem* 237:1961–1967.
15. Russell GJ, Walker PMB, Elton RA, Subak-Sharp JH (1976) Doublet frequency analysis of fractionated vertebrate nuclear DNA. *J Mol Biol* 108:1–23.
16. Jabbari K, Cacciò S, Pais de Barros J-P, Desgrès J, Bernardi G (1997) Evolutionary changes in CpG and methylation levels in vertebrate genomes. *Gene* 205:109–118.
17. Bernardi G (2004) *Structural and Evolutionary Genomics. Natural Selection in Genome Evolution* (Elsevier, Amsterdam).
18. Aïssani B, Bernardi G (1991) CpG islands, genes and isochores in the genome of vertebrates. *Gene* 106:185–195.
19. Jabbari K, Bernardi G (1998) CpG doublets, CpG islands and Alu repeats in long human DNA sequences from different isochore families. *Gene* 224:123–128.
20. Jabbari K, Bernardi G (2004) Cytosine methylation and CpG, TpG (CpA) and TpA frequencies. *Gene* 333:143–149.
21. Devillers-Thiery A (1974) PhD thesis (Université Paris VII, Paris).
22. Karlin S, Burge C (1995) Dinucleotide relative abundance extremes: A genomic signature. *Trends Genet* 11:283–290.
23. Gentles AJ, Karlin S (2001) Genome-scale compositional comparisons in eukaryotes. *Genome Res* 11:540–546.
24. Argos P, *et al.* (1979) Thermal stability and protein structure. *Biochemistry* 18:5698–5703.
25. Nishio Y, *et al.* (2003) Comparative complete genome sequence analysis of the amino acid replacements responsible for the thermostability of *Corynebacterium efficiens. Genome Res* 13:1572–1579.
26. Bernardi G, Bernardi G (1986) Compositional constraints and genome evolution. *J Mol Evol* 24:1–11.
27. Dickerson RE (1992) DNA structure from A to Z. *Methods Enzymol* 211:67–111.
28. Travers AA (1993) *DNA-Protein Interactions* (Chapman and Hall, New York).
29. Segal E, *et al.* (2006) A genomic code for nucleosome positioning. *Nature* 442:772–778.
30. Di Filippo M, Bernardi G (2008) Mapping Dnase I-hypersensitive sites on human isochores. *Gene* 419:62–65.
31. Saccone S, Federico C, Andreozzi L, D'Antoni S, Bernardi G (2002) Localization of the gene-richest and the gene-poorest isochores in the interphase nuclei of mammals and birds. *Gene* 300:169–178.
32. Felsenfeld G, Groudine M (2003) Controlling the double helix. *Nature* 421:448–453.
33. Costantini M, Bernardi G (2008) Replication timing, chromosomal bands and isochores. *Proc Natl Acad Sci USA* 105:3433–3437.
34. International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature* 431:931–945.
35. Kent WJ, *et al.* (2002) The human genome browser at UCSC. *Genome Res* 12:996–1006.
36. Sharp PM, Li WH (1986) An evolutionary perspective on synonymous codon usage in unicellular organisms. *J Mol Evol* 24:28–38.