Can GC Content at Third-Codon Positions Be Used as a Proxy for Isochore Composition?

Eran Elhaik, Giddy Landan, and Dan Graur

Department of Biology and Biochemistry, University of Houston

The isochore theory depicts the genomes of warm-blooded vertebrates as a mosaic of long genomic regions that are characterized by relatively homogeneous GC content. In the absence of genomic data, the GC content at third-codon positions of protein-coding genes (GC3) was commonly used as a proxy for the GC content of isochores. Oddly, in the postgenomic era, GC3 is still sometimes used as a proxy for the GC composition of isochores. Here, we use genic and genomic sequences from human, chimpanzee, cow, mouse, rat, chicken, and zebrafish to show that GC3 only explains a very small proportion of the variation in GC content of long genomic sequences flanking the genes (GCf), and what little correlation there is between GC3 and GCf was found to decay rapidly with distance from the gene. The coefficient of variation of GC3 was found to be much larger than that of GCf and, therefore, GC3 and GCf values are not comparable with each other. Comparisons of orthologous gene pairs from 1) human and chimpanzee and 2) mouse and rat show strong correlations between their GC3 values, but very weak correlations between their GCf values. We conclude that the GC content of third-codon position cannot be used as stand-in for isochoric composition.

Introduction

Isochores were first defined by Macaya et al. (1976) as long (>300 kb) genomic domains with homogeneous GC content. The genomes of warm-blooded vertebrates (mammals and birds) were described as a mosaic of isochores of alternating low and high GC contents, as opposed to the genomes of cold-blooded vertebrates (fishes and amphibians) that were supposed to lack GC-rich isochores (Bernardi et al. 1985; Bernardi 2000).

In the absence of genomic sequences, the GC composition at third-codon positions of protein-coding genes (GC3) was commonly used as a proxy for the GC composition of the isochore in which the gene resides (Bernardi et al. 1985; Aota and Ikemura 1986; Mouchiroud et al. 1991; Kadi et al. 1993; Duret et al. 1995; Zoubak et al. 1996; Bernardi et al. 1997; Robinson et al. 1997; Galtier and Mouchiroud 1998). In recent years, genomic sequences became available and various methods for genome segmentation into compositionally homogeneous segments have been proposed. Oddly, however, the practice of using GC3 as a proxy for the GC content of flanking isochores (GCf) still persists (Bernardi 2001; Ponger et al. 2001; Alvarez-Valin et al. 2002; D'Onofrio 2002; D'Onofrio et al. 2002; Scaiewicz et al. 2006; Costantini and Bernardi 2008), even though protein-coding regions, from which the value of GC3 is computed, comprise less than 5% of the human genome (IHGSC 2001) and about 10% of chicken genome (ICGSC 2004).

In support of this common practice, several small-scale analyses have been conducted (Aissani et al. 1991; Clay et al. 1996; Musto et al. 1999; Eyre-Walker and Hurst 2001). For example, Eyre-Walker and Hurst (2001) found a strong correlation between the GC3 values in 369 genes located on human chromosomes 21 and 22 and the GC content of upstream and downstream flanking regions of size of 25 kb. Moreover, it has been argued that GC3 is a more suitable indicator of flanking GC content than the mean GC content of all three codon positions (Bernardi 2000; Eyre-Walker and Hurst 2001).

The presumed relationship between GC3 and isochores has been used numerous times in the literature to study isochore function and evolution (Aota and Ikemura 1986; Kadi et al. 1993; Duret et al. 1995; Zoubak et al. 1996; Bernardi et al. 1997; Robinson et al. 1997; Galtier and Mouchiroud 1998; Eyre-Walker and Hurst 2001; Alvarez-Valin et al. 2002; Duret et al. 2002; Vinogradov 2003; Chojnowski et al. 2007). The purpose of this study is to test the appropriateness of GC3 as a stand-in for GC content of isochores.

Methods

Data Retrieval and Filtering

Coding sequences from RefSeq database are annotated as: "inferred," "model," "predicted," "provisional," "reviewed," or "validated." We included only genes that are annotated as predicted, provisional, reviewed, or validated (PPRV) to increase the reliability of our data. We used only fully sequenced eukaryotic genomes that have more than 3,000 PPRV coding sequences. We only used PPRV coding sequences larger than 300 bp, which had at least 200 kb upstream and 200 kb downstream. Six species met our criteria: *Homo sapiens* (build 36.2), *Bos taurus* (build 3.1), *Danio rerio* (build 1.1), *Gallus gallus* (build 2.1), *Mus musculus* (build 36.1), and *Rattus norvegicus* (build 3.4). The genomes were downloaded from the NCBI ftp web site (ftp://ftp.ncbi.nlm.nih.gov/genomes/). Using NCBI data file gene2accession (version 01/16/07), we retrieved for every genome all the coding sequences and their chromosomal location. We then extracted the coding sequences from RefSeq database and their flanking sequences from the downloaded genomic sequences. Introns were ignored because, to the best of our knowledge, they have not been used to predict "isochores." Our data set is shown in table 1.

The orthologous genes for *H. sapiens* (NCBI36) and *Pan troglodytes* (CHIMP2.1), and for *M. musculus* (NCBIM37) and *R. norvegicus* (RGSC3.4) were identified by the BioMart tool (http://www.biomart.org/biomart/martview/) using the Ensembl implementation (Kasprzyk et al. 2004). The genomes were downloaded from Ensembl and the flanking sequences were extracted as previously

Key words: isochores, GC3, GC content, flanking regions, genome composition, compositional patterns.

E-mail: dgraur@uh.edu.

Mol. Biol. Evol. 26(8):1829–1833. 2009
doi:10.1093/molbev/msp100
Advance Access publication May 14, 2009

Table 1
GC3 and GC123 for Six Vertebrate Taxa

		GC3			GC123		
Species	No. of Genes	Mean (%)	σ	Range (%)	Mean (%)	σ	Range (%)
Homo sapiens	17,451	60	17	22–97	45	6	32–80
Bos taurus	5,522	62	16	25–97	43	6	33–76
Mus musculus	17,009	59	11	21–96	43	5	27–76
Rattus norvegicus	8,983	59	11	23–96	42	6	33–73
Gallus gallus	3,036	56	15	28–99	42	5	36–80
Danio rerio	4,344	56	8	27–92	35	2	34–68

The mean, standard deviation (σ), and range are shown for each measure.

described. Our data set included 13,078 and 15,344 pairs of orthologous genes and their flanking regions for *Homo-Pan* and *Mus-Rattus*, respectively.

Statistical Tests

We employed three analyses to test the relationship between GC3 and the GC content of the flanking regions of the gene (GCf). In the first analysis, we calculated four genic measures: the GC content at each of the three codon positions (GC1, GC2, and GC3) and the average GC content (GC123). Next, we calculated the GC content of 40 nonoverlapping 5-kb windows upstream and downstream of the gene. For each genome, we calculated the coefficient of determination (r^2) between every genic measure and the GCf of every window. The significance of r^2 was tested with the Bonferroni correction (Sokal and Rohlf 1995, pp. 240, 702–703) to adjust for multiple comparisons.

To test the effect of window size on the correlations, we used windows ranging from 5 to 100 kb, but the results

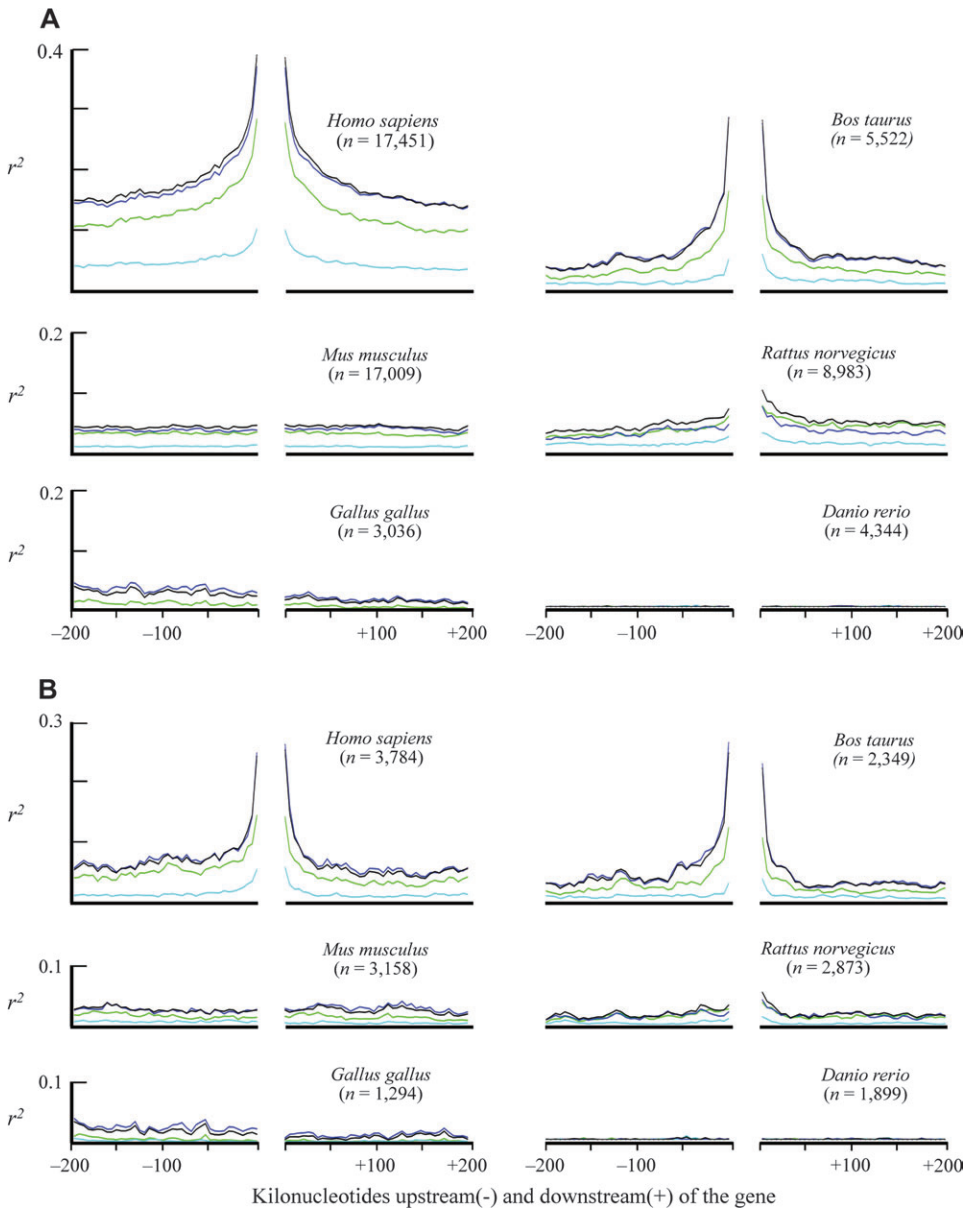

FIG. 1.—GC3 cannot predict GCf. Coefficients of determination (r^2) between GC1 (green), GC2 (turquoise), GC3 (blue), and GC123 (black), on the one hand, and GCf in 5-kb windows upstream and downstream of the gene, on the other. Calculations were carried (*a*) for all genes and (*b*) for genes that their 200-kb flanking regions do not overlap with other 200-kb flanking regions or with other genes. The number of genes is noted.

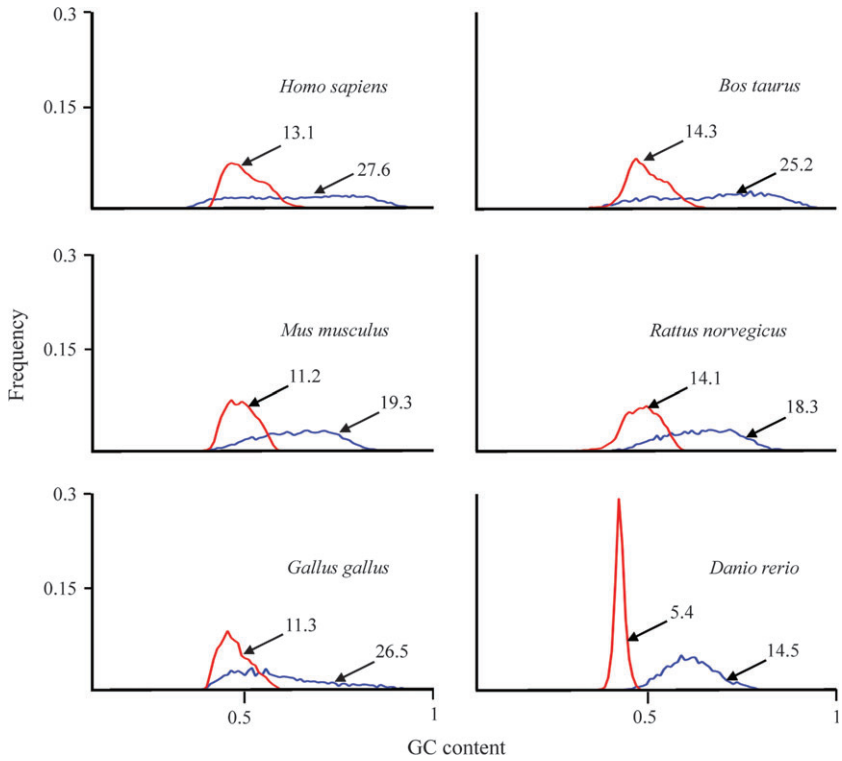

FIG. 2.—Frequency distribution of GC3 (blue) and 200-kb GCf (red). Coefficients of variation are shown.

did not change. We also repeated all calculations by using only genes that their 200-kb flanking regions did not overlap either with other 200-kb flanking regions or with other known genes. The results were unaffected.

In the second analysis, we compared the breadth of the distribution of GC3 and GCf by using coefficient of variation (Sokal and Rohlf 1995, pp. 57–59; Zar 1999, p. 40). We used flanking regions of size 200 kb upstream and downstream of the gene to estimate GCf.

In the third analysis, we compared the orthologous gene pairs from *Homo–Pan* and *Mus–Rattus* and calculated the relationship for GC1, GC2, GC3, GC123, and GCf pairs. We used nonoverlapping flanking regions of 5 kb up to 200 kb upstream and downstream of the gene. The significance of r^2 was tested with the Bonferroni correction (Sokal and Rohlf 1995, pp. 240, 702–703) to adjust for multiple comparisons.

Results

Means, standard deviations, and ranges of GC3 and GC123 for the different genomes are shown in table 1. We calculated the coefficient of determination between GC1, GC2, GC3, and GC123, on the one hand, and GCf, on the other, and found that for most genomes the GCf variation cannot be explained by any of these measures (fig. 1*a*).

The trend of decreasing r^2 values with increasing distance from the gene was observed in all genomes for both upstream and downstream directions. In the human and cow genomes, this trend was clearly observed as a sharp decrease in r^2 values for flanking regions within close range of the gene followed by a moderate decrease for the distant flanking regions. In these genomes, GC3 only explained a very small proportion of the variation in GC content of long genomic sequences flanking the genes (GCf). The GCf variation was not explained at all by any genic measure in mouse, rat, chicken, and zebrafish genomes. When we eliminated genes with overlapping flanking regions (fig. 1*b*), the coefficients of determination decreased but the overall trends remained the same.

When comparing the explanatory abilities of the four genic measures, we see that GC123 is a stronger predictor of GCf than GC3, although the difference is not significant. With the exception of cow and chicken, in all other genomes, the four measures follow the inequality $r^2_{(GC123,GCf)} > r^2_{(GC3,GCf)} > r^2_{(GC1,GCf)} > r^2_{(GC2,GCf)}$. Additionally, we did not observe any correlation between coding sequence size and GCf.

The distributions of GC3 and mean GCf for 200 kb upstream and downstream of the gene are shown in figure 2. The shape of the GCf distribution is not affected by the size of the flanking regions (results are not shown) and, therefore, we only present the distribution for 200 kb. We note that, on average, the coefficient of variation for GCf is considerably smaller than that for GC3.

The frequency distribution of human GC content at all codon positions as well as in flanking regions of size 200 kb is plotted in figure 3. Interestingly, the distributions of GC2 and GCf are very similar in all the genomes although GC2 only explains less than 9% of the variation in GCf. All other

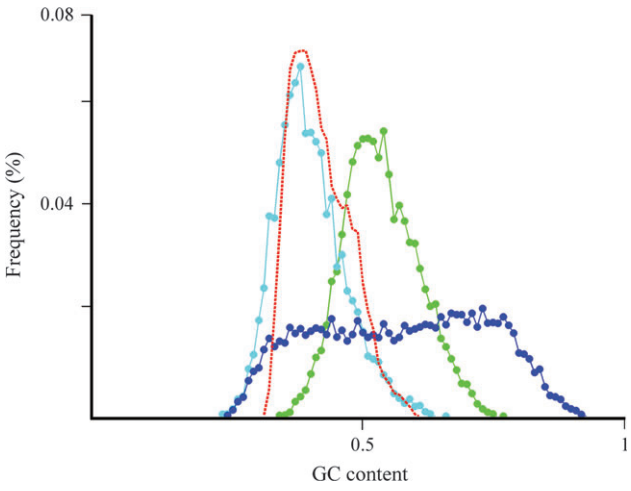

FIG. 3.—GC content in codon positions: GC1 (green), GC2 (turquoise), GC3 (blue), and 200-kb flanking regions (dashed red) in human.

genomes show a similar pattern of distributions, and are therefore, not shown.

Another way to test the evolutionary relationship between GC3 and GCf is to compare the GC3 and GCf of orthologous genes from two genomes. If the claim that the same natural processes occurred in both GC3 and GCf is true, then GC3 should be a good predictor of GCf and both GC3 and GCf would be highly correlated. We found a strong relationship between all genic measures (GC1, GC2, GC3, and GC123) of the orthologous genes. For the pair *Homo–Pan*: $r^2_{(GC1,GC1)}=0.91$, $r^2_{(GC2,GC2)}=0.91$, $r^2_{(GC3,GC3)}=0.94$, and $r^2_{(GC123,GC123)}=0.92$. For the pair *Mus–Rattus*: $r^2_{(GC1,GC1)}=0.56$, $r^2_{(GC2,GC2)}=0.58$, $r^2_{(GC3,GC3)}=0.57$, and $r^2_{(GC123,GC123)}=0.57$. In contrast, the correlation between GCf values was very weak. Figure 4 presents the r^2 between GC3 and GCf of orthologous genes for *Homo–Pan* and *Mus–Rattus*. For *Homo–Pan*, the range of $r^2_{(GCf,GCf)}$ is from 0.21 to 0.45 with a mean of 0.3. For *Mus–Rattus*, the range of $r^2_{(GCf,GCf)}$ is from 0.01 to 0.2 with a mean of 0.03. All results were significant at a 0.01 significance level. The decrease in r^2 values shows that GCf is not conserved among orthologous genes. The differences in r^2 values between the upstream and downstream directions were insignificant for all genomes.

Discussion

GC3 is routinely used as a proxy for the GC composition of isochores (Bernardi 2001; Ponger et al. 2001; Alvarez-Valin et al. 2002; D'Onofrio 2002; D'Onofrio et al. 2002; Scaiewicz et al. 2006; Costantini and Bernardi 2008), although to the best of our knowledge, the relationship between GC3 and the GC content of very long flanking regions (the presumed size of isochores) has never been tested on a large genomic or taxonomic scale. Previous analyses used few genes and flanking regions that were so short as to be completely irrelevant to the definition of isochores (Aissani et al. 1991; Clay et al. 1996; Musto et al. 1999; Eyre-Walker and Hurst 2001).

Our analyses tested the ability of four genic composition measures: GC1, GC2, GC3, and GC123 to predict the GC content in flanking regions 5′ and 3′ of the gene. Because GC3 is mostly unconstrained by functional requirements, that is, by the need to code specific amino acids, the third-codon position is a natural candidate for a predictive proxy of flanking GC content. We note, however, that a proxy must be able to explain most of the variation in GCf, not merely be correlated with it. Our analyses reveal that GC3 explains very little of the variation in GC content of large flanking regions. Moreover, we see that the predictive power either decreases rapidly the further one gets upstream and downstream of the gene or does not exist at all. Our orthologous gene pair analysis indicates that different evolutionary processes affect codon usage (GC3) and flanking regions (isochores) and, therefore GC3 cannot be used to predict GCf. Finally, we note that the predictive power of GC3 is almost nonexistent in nonhuman vertebrates.

We suggest that all associations between isochores and genic features (e.g., gene length, gene density, and chromosomal bands) that have been reported or suggested in the literature should be reevaluated if GC3 was used as a proxy for the GC content of isochores, as it was almost invariably done in the past.

Sometimes, GC3 is used when genomic sequences are not available (Galtier 2003; Hamada et al. 2003; Montoya-Burgos et al. 2003; Romero et al. 2003; Cruveiller et al. 2004; Federico et al. 2004; Gu and Li 2006; Chojnowski et al. 2007; Fortes et al. 2007; Chojnowski and Braun 2008). We show here that in all probability GC3 lacks predictable power as far as large flanking regions are concerned.

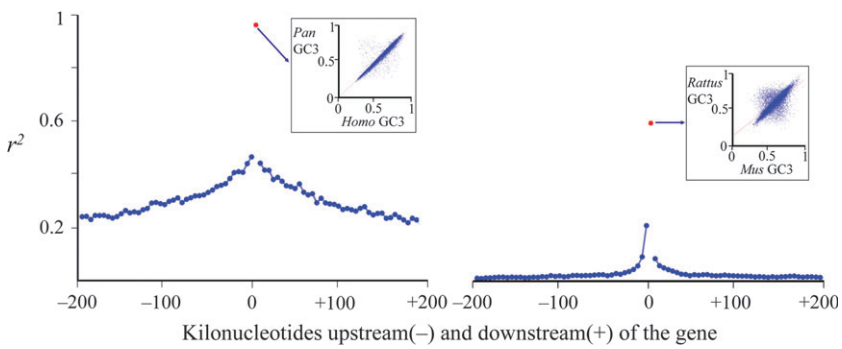

FIG. 4.—Coefficient of determination (r^2) between GCf values (circles) surrounding orthologous genes in *Homo-Pan* (left panel) and *Mus-Rattus* (right panel). The r^2 for GC3 is shown as a square at 0 on the *x*-axis. The correlation of GC3 values for orthologous genes is shown in the inset.

Acknowledgments

This work was supported in part by National Science Foundation grant DBI-0543342 to D.G.

Literature Cited

Aissani B, D'Onofrio G, Mouchiroud D, Gardiner K, Gautier C, Bernardi G. 1991. The compositional properties of human genes. J Mol Evol. 32:493–503.

Alvarez-Valin F, Lamolle G, Bernardi G. 2002. Isochores, GC3 and mutation biases in the human genome. Gene. 300:161–168.

Aota S, Ikemura T. 1986. Diversity in G + C content at the third position of codons in vertebrate genes and its cause. Nucleic Acids Res. 14:6345–6355.

Bernardi G. 2000. Isochores and the evolutionary genomics of vertebrates. Gene. 241:3–17.

Bernardi G. 2001. Misunderstandings about isochores. Part 1. Gene. 276:3–13.

Bernardi G, Hughes S, Mouchiroud D. 1997. The major compositional transitions in the vertebrate genome. J Mol Evol. 44(Suppl. 1):S44–S51.

Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, Meunier-Rotival M, Rodier F. 1985. The mosaic genome of warm-blooded vertebrates. Science. 228:953–958.

Chojnowski JL, Braun EL. 2008. Turtle isochore structure is intermediate between amphibians and other amniotes. Integr Comp Biol. 48:454–462.

Chojnowski JL, Franklin J, Katsu Y, Iguchi T, Guillette LJ Jr, Kimball RT, Braun EL. 2007. Patterns of vertebrate isochore evolution revealed by comparison of expressed mammalian, avian, and crocodilian genes. J Mol Evol. 65:259–266.

Clay O, Caccio S, Zoubak S, Mouchiroud D, Bernardi G. 1996. Human coding and noncoding DNA: compositional correlations. Mol Phylogenet Evol. 5:2–12.

Consortium, ICGS. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature. 432:695–716.

Consortium, IHGS. 2001. Initial sequencing and analysis of the human genome. Nature. 409:860–921.

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

Cruveiller S, Jabbari K, Clay O, Bernardi G. 2004. Compositional gene landscapes in vertebrates. Genome Res. 14:886–892.

D'Onofrio G. 2002. Expression patterns and gene distribution in the human genome. Gene. 300:155–160.

D'Onofrio G, Ghosh TC, Bernardi G. 2002. The base composition of the genes is correlated with the secondary structures of the encoded proteins. Gene. 300:179–187.

Duret L, Mouchiroud D, Gautier C. 1995. Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. J Mol Evol. 40:308–317.

Duret L, Semon M, Piganeau G, Mouchiroud D, Galtier N. 2002. Vanishing GC-rich isochores in mammalian genomes. Genetics. 162:1837–1847.

Eyre-Walker A, Hurst LD. 2001. The evolution of isochores. Nat Rev Genet. 2:549–555.

Federico C, Saccone S, Andreozzi L, Motta S, Russo V, Carels N, Bernardi G. 2004. The pig genome: compositional analysis and identification of the gene-richest regions in chromosomes and nuclei. Gene. 343:245–251.

Fortes G, Bouza C, Martínez P, Sánchez L. 2007. Diversity in isochore structure among cold-blooded vertebrates based on GC content of coding and non-coding sequences. Genetica. 129:281–289.

Galtier N. 2003. Gene conversion drives GC content evolution in mammalian histones. Trends Genet. 19:65–68.

Galtier N, Mouchiroud D. 1998. Isochore evolution in mammals: a human-like ancestral structure. Genetics. 150: 1577–1584.

Gu J, Li WH. 2006. Are GC-rich isochores vanishing in mammals? Gene. 385:50–56.

Hamada K, Horiike T, Ota H, Mizuno K, Shinozawa T. 2003. Presence of isochore structures in reptile genomes suggested by the relationship between GC contents of intron regions and those of coding regions. Genes Genet Syst. 78: 195–198.

Kadi F, Mouchiroud D, Sabeur G, Bernardi G. 1993. The compositional patterns of the avian genomes and their evolutionary implications. J Mol Evol. 37:544–551.

Kasprzyk A, Keefe D, Smedley D, London D, Spooner W, Melsopp C, Hammond M, Rocca-Serra P, Cox T, Birney E. 2004. EnsMart: a generic system for fast and flexible access to biological data. Genome Res. 14:160–169.

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.

Montoya-Burgos JI, Boursot P, Galtier N. 2003. Recombination explains isochores in mammalian genomes. Trends Genet. 19:128–130.

Mouchiroud D, D'Onofrio G, Aissani B, Macaya G, Gautier C, Bernardi G. 1991. The distribution of genes in the human genome. Gene. 100:181–187.

Musto H, Romero H, Zavala A, Bernardi G. 1999. Compositional correlations in the chicken genome. J Mol Evol. 49: 325–329.

Ponger L, Duret L, Mouchiroud D. 2001. Determinants of CpG islands: expression in early embryo and isochore structure. Genome Res. 11:1854–1860.

Robinson M, Gautier C, Mouchiroud D. 1997. Evolution of isochores in rodents. Mol Biol Evol. 14:823–828.

Romero H, Zavala A, Musto H, Bernardi G. 2003. The influence of translational selection on codon usage in fishes from the family Cyprinidae. Gene. 317:141–147.

Scaiewicz V, Sabbia V, Piovani R, Musto H. 2006. CpG islands are the second main factor shaping codon usage in human genes. Biochem Biophys Res Commun. 343:1257–1261.

Sokal RR, Rohlf FJ. 1995. Biometry, 3rd ed. New York: W.H. Freeman and Company.

Vinogradov AE. 2003. Isochores and tissue-specificity. Nucleic Acids Res. 31:5212–5220.

Zar JH. 1999. Biostatistical analysis. Upper Saddle River (NJ): Prentice-Hall.

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

Takashi Gojobori, Associate Editor

Accepted April 20, 2009