Nucleic Acids Research, 2004, Vol. 32, Web Server issue ***W287–W292***
DOI: 10.1093/nar/gkh399

IsoFinder: computational prediction of isochores in genome sequences

José L. Oliver*, Pedro Carpena[1], Michael Hackenberg and Pedro Bernaola-Galván[1]

Departamento de Genética, Instituto de Biotecnología, Facultad de Ciencias, Universidad de Granada and [1]Departamento de Física Aplicada II, Universidad de Málaga, Spain

Received January 22, 2004; Revised March 4, 2004; Accepted March 25, 2004

ABSTRACT

Isochores are long genome segments homogeneous in G+C. Here, we describe an algorithm (IsoFinder) running on the web (http://bioinfo2.ugr.es/IsoF/isofinder.html) able to predict isochores at the sequence level. We move a sliding pointer from left to right along the DNA sequence. At each position of the pointer, we compute the mean G+C values to the left and to the right of the pointer. We then determine the position of the pointer for which the difference between left and right mean values (as measured by the *t*-statistic) reaches its maximum. Next, we determine the statistical significance of this potential cutting point, after filtering out short-scale heterogeneities below 3 kb by applying a coarse-graining technique. Finally, the program checks whether this significance exceeds a probability threshold. If so, the sequence is cut at this point into two subsequences; otherwise, the sequence remains undivided. The procedure continues recursively for each of the two resulting subsequences created by each cut. This leads to the decomposition of a chromosome sequence into long homogeneous genome regions (LHGRs) with well-defined mean G+C contents, each significantly different from the G+C contents of the adjacent LHGRs. Most LHGRs can be identified with Bernardi's isochores, given their correlation with biological features such as gene density, SINE and LINE (short, long interspersed repetitive elements) densities, recombination rate or single nucleotide polymorphism variability. The resulting isochore maps are available at our web site (http://bioinfo2.ugr.es/isochores/), and also at the UCSC Genome Browser (http://genome.cse.ucsc.edu/).

INTRODUCTION

Mammalian genomes are made up of isochores, long DNA segments ($\gg$300 kb) fairly homogeneous in G+C which were first revealed by analytical ultracentrifugation of bulk DNA (1–3). The relevance of the isochore model for genome biology is based on observations of gene and SINE (short interspersed repetitive elements) densities, as well as recombination frequency, which are all higher in (G+C)-rich isochores, whereas LINEs (long interspersed repetitive elements) are denser in (G+C)-poor isochores (3). Besides compositional differences, genome segments separated by isochore boundaries also differ in replication timing, as in the isochores of the human major histocompatibility complex (MHC) locus (4), or in recombination rates, as in the human neurofibromatosis NF1 region (5). Isochores are found in a large variety of taxa, including plants and cold-blooded vertebrates, although they are more conspicuous in the genome of warm-blooded vertebrates [see (3) and references therein]. The isochore concept has increased our appreciation of the complexity and compositional variability of eukaryotic genomes (6), having been recently summarized as 'a fundamental level of genome organization' (7). The evolutionary origin and maintenance of isochores in today's genomes is currently the object of active debate (7–11).

The recent availability of the draft human genome sequence allowed for a direct test of the isochore model. A first analysis denied the existence of isochores in the human chromosomes 21 and 22 (12), while a second one ruled out a strict notion of isochores as compositionally homogeneous, concluding that isochores do not appear to deserve the prefix 'iso' (13). Both approaches, however, present serious drawbacks (14–17). First, denying the existence of isochores means denying the existence of compositional discontinuities in the human genome and going back to a genome organization characterized by a continuous compositional variation, a view shown to be wrong in the early 1970s (18). Second, the methodological problem common to the approaches denying isochores is that they take as a reference the random, uncorrelated model

*To whom correspondence should be addressed. Tel: +34 958243261; Fax: +34 958244073; Email: oliver@ugr.es

(in which every nucleotide is free to change) to test sequence homogeneity. This would lead to the absurd conclusion that only highly repetitive DNA sequences would be homogeneous. However, since 1981, when the heterogeneity within isochore families was quantified (19), we have known that the homogeneity of isochores is only relative—hence their definition as *fairly* homogeneous regions (3). Indeed, we have recently shown that when the appropriate scale and statistical test are chosen, isochores still merit the prefix 'iso' (20).

A third analysis (6), also based on the random, uncorrelated sequence model, reports a high degree of compositional heterogeneity in the human genome, but fails to detect the most conspicuous isochore boundary experimentally characterized to date, the one separating Class II and Class III regions of the human MHC locus (21–23).

In the past few years, we have been developing an algorithm, based on compositional segmentation, able to predict isochore boundaries at the sequence level (17,24–26). Most of the long homogeneous genome regions (LHGRs) predicted by this algorithm can be identified with Bernardi's isochores, given their correlation with biological features such as gene density, SINE and LINE densities, recombination rate and SNP (single nucleotide polymorphism) variability (17,26). Its efficacy as compared with other methods has also been proven (26,27). Now, after optimizing the algorithm by making several key improvements, we describe its implementation in a computer program (IsoFinder) running on the web (http://bioinfo2.ugr.es/IsoF/isofinder.html).

METHODS

The partition of a DNA sequence into fairly homogeneous, isochore-like segments is performed by means of a modified version of the entropic compositional segmentation. This algorithm was proposed in 1996 (24) and has proven useful in finding homogeneous regions (28,29) and measuring the compositional complexity or heterogeneity of DNA sequences (25).

While the original algorithm was designed to maximize the global difference in composition between adjacent segments and was intended for finding segments at all scales, the modified version presented here maximizes the G+C contrast between adjacent segments and looks for large-scale isochore-like segments.

The algorithm works as follows. Consider a DNA sequence of length L, along which we move a sliding pointer from left to right. At each position of the pointer, we compute the mean G+C values to the left and to the right of the pointer. To measure the difference between left and right mean values, we use the t-statistic. We next determine the position of the pointer for which t reaches its maximum value, t_{max}. Let us assume that this point divides the sequence into two subsequences of lengths L_{left} and L_{right}.

Next, we determine the statistical significance of the candidate to be a cutting point. As we wish to detect only isochore-like segments, we need to avoid the influence of short-scale heterogeneities on the statistical significance. Thus, we filter out the heterogeneity below a given minimum length (ℓ_0); i.e. we divide both subsequences into non-overlapping windows of length ℓ_0 and compute the G+C content in each window. In this way, we convert the subsequence of length L_{left} (L_{right}) into an array of L_{left}/ℓ_0 (L_{right}/ℓ_0) real numbers corresponding to the G+C content of each window of size ℓ_0. The online version of IsoFinder allows the user to choose among three different values for ℓ_0 (1, 2 and 3 kb) to perform the filtering procedure. However, the results presented in previous works (17,26) were obtained with $\ell_0 = 3$ kb, which corresponds to a homogeneity criterion for mammalian isochores, derived from ultracentrifugation of DNA at different molecular weights (2).

To determine the validity of the candidate to be a cutting point, we have to verify that the mean value of G+C of the left-hand-side windows is significantly different from the mean value of G+C in the right-hand-side windows. In doing so, we again use the t-Student statistics, but now computed on the two arrays of real numbers obtained after the filtering procedure, thus obtaining t_{filt}. In this way, the estimator of the difference in composition between left and right subsequences is not affected by short-scale heterogeneities (below ℓ_0).

Since the candidate to be a cutting point was found as the one maximizing the difference in composition, the statistical significance of t_{filt} cannot be measured by the standard Student's distribution. Thus, here we use the distribution of t_{filt} values $[P(\tau = t_{filt})]$ obtained by means of Monte Carlo simulations (30). The significance level $P(\tau)$ of a possible cutting point with $t_{filt} = \tau$ is defined as the probability of obtaining the value τ or lower values within a random sequence. Thus, a series of N random numbers of fixed mean would remain unsegmented with probability $P(\tau)$. Finally, we check whether this significance exceeds a selected threshold P_0, usually taken to be 95%. If so, the sequence is cut at this point into two subsequences; otherwise, the sequence remains undivided. If the sequence is cut, the procedure continues recursively for each of the two resulting subsequences created by each cut. All resulting segments have a statistically significant difference in their means. The process stops when none of the possible cutting points has a significance exceeding P_0, and we say that the sequence has been segmented at the 'significance level P_0'. Our method leads to the partitioning of a DNA sequence into LHGRs with well-defined mean G+C levels, each significantly different from the mean G+C level of the adjacent LHGRs. Other algorithms proposed later (27,31) do not provide the statistical significance of the resulting partition of the sequence.

IMPLEMENTATION AND WEB INTERFACE

IsoFinder core routines were developed using the Lahey/Fujitsu Fortran 95 compiler under Debian Linux. A graphical gnuplot routine (http://www.gnuplot.info/) was used to generate the isochore maps. Lastly, a Perl CGI script for the Apache web server was used to integrate data input/output. Text fields to choose the significance level and the coarse-graining tract length to compute the statistical significance of cutting points are provided in the launch page (http://bioinfo2.ugr.es/IsoF/isofinder.html). The sequence to be segmented can be uploaded from a file on the local machine or pasted into the appropriate text field. Raw sequences and some of the standard sequence formats (EMBL, GenBank or FASTA) are accepted.

The output web page provides links to the results of sequence segmentation in three formats: (i) coordinates, sizes and G+C contents of the predicted isochores as an HTML table (Figure 1); (ii) the same but in plain text; and (iii) the isochore map of the sequence in PNG format (Figure 2). These results

Isochore Predictions by IsoFinder

Sequence: out585990/1070968482__MHC -- Sig. level = 0.9500 -- Sig. method: Maximum -- Coarse graining: 3000 bp -- SCC: 0.6214E+01

LHGR	From	To	Size	GC%
1	1	299270	299270	46.08
2	299271	354226	54956	39.57
3	354227	364029	9803	53.95
4	364030	833239	469210	43.43
5	833240	1040717	207478	49.20
6	1040718	1168415	127698	46.24
7	1168416	1174527	6112	61.98
8	1174528	1230906	56379	52.47
9	1230907	1237030	6124	45.57
10	1237031	1244738	7708	49.73
11	1244739	1396980	152242	45.87
12	1396981	1469257	72277	52.32
13	1469258	1479458	10201	43.24
14	1479459	1490846	11388	51.70
15	1490847	1739420	248574	43.05
16	1739421	1841871	102451	47.73
17	1841872	2483966	642095	51.87
18	2483967	3054365	570399	40.19
19	3054366	3065923	11558	45.49
20	3065924	3074335	8412	51.84
21	3074336	3080554	6219	47.39
22	3080555	3088089	7535	51.36
23	3088090	3159420	71331	38.31
24	3159421	3384907	225487	42.95
25	3384908	3444780	59873	55.84
...	...	...	...	...
37	3661923	3673778	11856	52.85

Figure 1. Coordinates, sizes and GC contents of the predicted isochores as an HTML table.

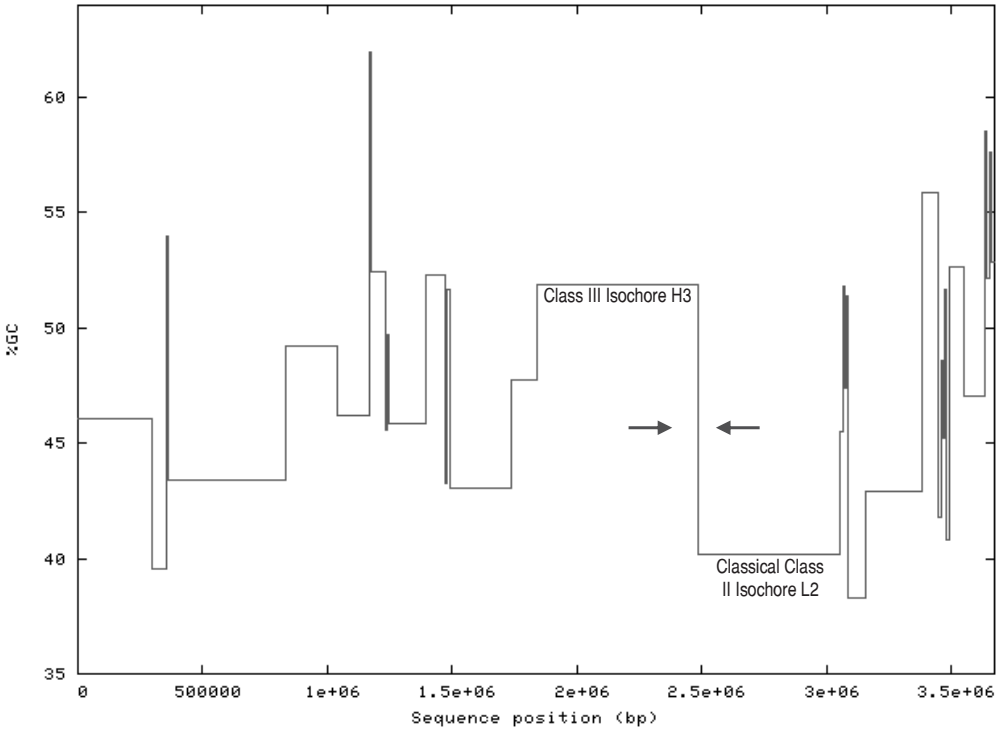

Figure 2. Isochore map of the human MHC region. The H3 and L2 isochores in this region are indicated. The arrows point to the pronounced isochore boundary between these two isochores.

remain available on the server for 24 h, and the user can access them through the hyperlinks provided in the output page.

EXAMPLES

MHC isochores

Class II (in an L2 isochore) and Class III (in an H3 isochore) regions of the human MHC are the only fully characterized isochores determined at the sequence level to date (21–23). Figure 2 shows the segmentation we made of the consensus sequence for the human MHC produced by the Human Chromosome 6 Sequencing Group at the Sanger Centre (http://www.sanger.ac.uk/HGP/Chr6/published_consensus.fasta). The IsoFinder algorithm enables precise location of the MHC isochore boundaries. The first three predicted boundaries were at positions 2483966, 3054365 and 1841872, corresponding to the sequence junction separating L2 and H3 isochores, the centromeric end of the isochore L2, and the telomeric end of the isochore H3, respectively. The remaining cuts on this sequence all fell outside of these two isochores, thus defining other homogeneous regions within the MHC sequence.

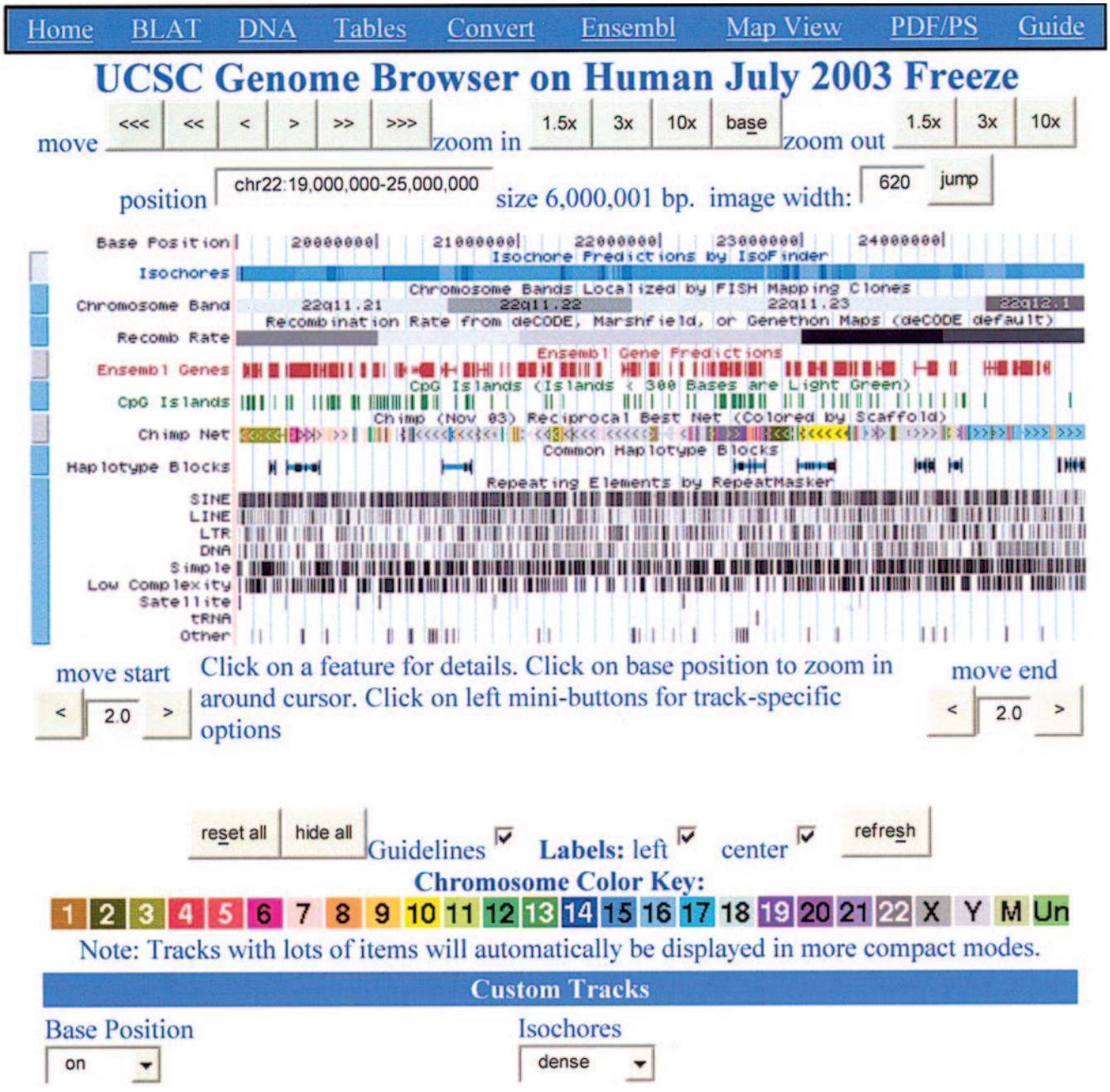

Figure 3. Isochore structure of a region of human chromosome 22 spanning 6 Mb between coordinates 19 000 000 and 25 000 000. Chromosome bands, recombination rate, gene density, CpG islands, best alignments with the chimpanzee genome, haplotype blocks and repeat densities are jointly shown with the isochore structure (top track) of this region.

Isochore visualization with the Genome Browser at UCSC

The Genome Browser at UCSC (http://genome.ucsc.edu/) stacks annotation tracks beneath genome coordinate positions, allowing rapid visual correlation of different types of information. The user can look at a whole chromosome to get a feel for gene density, open a specific cytogenetic band to see a positionally mapped disease-gene candidate, or zoom in to a particular gene to view its spliced ESTs and possible alternative splicing. Isochore predictions by IsoFinder have also been integrated with all this comprehensive genome information (http://bioinfo2.ugr.es/isochores/), being available also at the UCSC genome browser as a custom track (http://genome.cse.ucsc.edu/goldenPath/customTracks/custTracks.html). As an example, the top track in Figure 3 shows the isochore structure of a region of human chromosome 22 integrated with information about chromosome bands, recombination rate, gene density, CpG islands, best alignments with the chimpanzee genome, haplotype blocks and repeat densities.

CONCLUSIONS AND FUTURE DEVELOPMENTS

The IsoFinder web server allows accurate and reliable isochore predictions in genome sequences. Four key advantages of the isochores predicted by IsoFinder are that (i) the isochore heterogeneity at different genome scales is shown in the same plot; (ii) pair-wise compositional differences between adjacent isochores are all statistically significant; (iii) isochore boundaries are accurately defined to single base pair resolution and (iv) both gradual and abrupt isochore boundaries are simultaneously revealed.

Apart from the MHC isochores (see above), no other experimentally confirmed isochore dataset exists. Therefore, the algorithm has not been validated against experimental data, using, e.g. parameters such as specificity or sensitivity. Instead, we have used numerical simulations to estimate the error in isochore-boundary determination [see section 2.3 and figure 1 in (26)]. We found that for typical isochore sizes ($\approx$300 kb), the relative error ranges from 0.15 to 0.05%.

IsoFinder has also been compared with other available methods. In particular, an extensive comparison of IsoFinder against the wavelet multiresolution method (27) found a very good agreement between both algorithms, the differences being always lower than 1%. The remaining available methods either deny the existence of isochores (12,13) or fail in detecting even the clearer isochores experimentally identified to date (6).

Our algorithm has also been successfully used to relate isochore chromosome structure to gene density, SINE and LINE densities and SNP variability (17,26). We are now using IsoFinder to analyse the evolutionary factors driving the biased distribution of Alu retrotransposons in human isochores (M. Hackenberg and J. L. Oliver, submitted for publication). There are several other potential applications of the algorithm described here. First, one could scan the predicted LHGR boundaries searching for changes in replication timing known to occur at isochore boundaries (4). Second, anonymous, recently obtained genomic sequences can now be quickly and accurately scanned for gene-rich regions, as we found that gene density depends heavily on the G+C content of the LHGRs. Third, we have recently shown (32) that the prediction of the coding proportion in a sequence is better when LHGRs, instead of moving windows, are used. Therefore, improvements in computational gene identification are also expected, as the specific compositional parameters of the corresponding isochores can now be taken into account as input for gene-finding programs. Fourth, in the same way, other programs making use of local compositional parameters to predict sequence patterns, such as RepeatMasker, could be improved by considering LHGRs instead of moving windows. Finally, studies of comparative genomics could benefit from the detailed isochore chromosome maps provided by IsoFinder.

ACKNOWLEDGEMENTS

The help of David Nesbitt with the English version of the manuscript is appreciated. This work was supported by the Spanish Government (Grants Nos. BIO2002-04014-C03-01 and 02) and Plan Andaluz de Investigación (CVI-162). M.H. acknowledges a predoctoral grant from the University of Granada (Spain).

REFERENCES

1. Macaya,G., Thiery,J.P. and Bernardi,G. (1976) An approach to the organization of eukaryotic genomes at a macromolecular level. *J. Mol. Biol.*, **108**, 237–254.
2. Bernardi,G., Olofson,B., Filipski,J., Zerial,M., Salinas,J., Cuny,G., Meunier-Rotival,M. and Rodier,F. (1985) The mosaic genome of warm-blooded vertebrates. *Science*, **228**, 953–958.
3. Bernardi,G. (2000) Isochores and the evolutionary genomics of vertebrates. *Gene*, **241**, 3–17.
4. Tenzen,T., Yamagata,T., Fukagawa,T., Sugaya,K., Ando,A., Inoko,H., Gojobori,T., Fujiyama,A., Okumura,K. and Ikemura,T. (1997) Precise switching of DNA replication timing in the GC content transition area in the human MHC. *Mol. Cell. Biol.*, **17**, 4043–4050.
5. Eisenbarth,I., Vogel,G., Krone,W., Vogel,W. and Assum,G. (2000) An isochore transition in the Nf1 gene region coincides with a switch in the extent of linkage disequilibrium. *Am. J. Hum. Genet.*, **67**, 873–880.
6. Nekrutenko,A. and Li,W.H. (2000) Assesment of compositional heterogeneity within and between eukariotic genomes. *Genome Res.*, **10**, 1986–1995.
7. Eyre-Walker,A. and Hurst,W. (2001) The evolution of isochores. *Nat. Rev. Genet.*, **2**, 549–555.
8. Eyre-Walker,A. (1992) Evidence that both G+C rich and G+C poor isochores are replicated early and late in the cell cycle. *Nucleic Acids Res.*, **20**, 1497–1501.
9. Francino,M.P. and Ochman,H. (1999) Isochores result from mutation not selection. *Nature*, **400**, 31–32.
10. Fryxell,J.J. and Zuckerkandl,E. (2000) Cytosine deamination plays a primary role in the evolution of mammalian isochores. *Mol. Biol. Evol.*, **17**, 1371–1383.
11. Piganeau,G.I., Mouchiroud,D., Duret,L. and Gautier,Ch. (2002) Expected relationship between the silent substitution rate and the GC content: implications for the evolution of isochores. *J. Mol. Evol.*, **54**, 129–133.
12. Häring,D. and Kypr,J. (2001) No isochores in the human chromosomes 21 and 22? *Biochem. Biophys. Res. Commun.*, **280**, 567–573.
13. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (International human genome sequencing consortium) (2001) Initial sequencing and analysis of the human genome, *Nature*, **409**, 860–921.
14. Bernardi,G. (2001) Misunderstandings about isochores. Part 1. *Gene*, **276**, 3–13.
15. Clay,O. and Bernardi,G. (2002) Isochores: dream or reality? *Trends in Biotech.*, **20**, 237.

16. Li,W. (2001) Delineating relative homogeneous G+C domains in DNA sequences. *Gene*, **276**, 57–72.
17. Oliver,J.L., Carpena,P., Román-Roldán,R., Mata-Balaguer,T., Mejías-Romero,A., Hackenberg,M. and Bernaola-Galván,P. (2002) Isochore chromosome maps of the human genome. *Gene*, **300**, 117–127.
18. Filipski,J., Thiery,J.P. and Bernardi,G. (1973) An analysis of the bovine genome by Cs_2SO_4–Ag density gradient centrifugation. *J. Mol. Biol.*, **80**, 177–197.
19. Cuny,G., Soriano,P., Macaya,G. and Bernardi,G. (1981) The major components of the mouse and human genomes: preparation, basic properties and compositional heterogeneity. *Eur. J. Biochem.*, **115**, 227–233.
20. Li,W., Bernaola-Galvan,P., Carpena,P. and Oliver,J.L. (2003) Isochores merit the prefix 'Iso'. *Comp. Biol. Chem.*, **27**, 5–10.
21. Fukagawa,T., Sugaya,K., Matsumoto,K., Okumura,K., Ando,A., Inoko,H. and Ikemura,T. (1995) A boundary of long-range G+C% mosaic domains in the human MHC locus: pseudoautosomal boundary-like sequence exists near the boundary. *Genomics*, **25**, 184–191.
22. Stephens,R., Horton,R., Humphray,S., Rowen,L., Trowsdale,J. and Beck,S. (1999) Gene organisation, sequence variation and isochore structure at the centromeric boundary of the human MHC. *J. Mol. Biol.*, **291**, 789–799.
23. The MHC sequencing consortium (1999) Complete sequence and gene map of a human major histocompatibility complex. *Nature*, **401**, 921–923.
24. Bernaola-Galván,P., Román-Roldán,R. and Oliver,J.L. (1996) Compositional segmentation and long-range fractal correlation in DNA sequences. *Phys. Rev. E.*, **53**, 5181–5189.
25. Román-Roldán,R., Bernaola-Galván,P. and Oliver,J.L. (1998) Sequence compositional complexity of DNA through an entropic segmentation method. *Phys. Rev. Lett.*, **80**, 1344–1347.
26. Oliver,J.L., Bernaola-Galván,P., Carpena,P. and Román-Roldán,R. (2001) Isochore chromosome maps of eukaryotic genomes. *Gene*, **276**, 47–56.
27. Wen,S.Y. and Zhang,C.T. (2003) Identification of isochore boundaries in the human genome using the technique of wavelet multiresolution analysis. *Biochem. Biophys. Res. Comm.*, **311**, 215–222.
28. Azad,R.K., Bernaola-Galván,P., Ramaswamy,R. and Rao,J.S. (2002) Segmentation of genomic DNA through entropic divergence: power laws and scaling. *Phys. Rev. E*, **65**, 0519091–0519096.
29. Grosse,I., Bernaola-Galván,P., Carpena,P., Román-Roldán,R., Oliver,J.L. and Stanley,H.E. (2002) Analysis of symbolic sequences using the Jensen-Shannon divergence. *Phys. Rev. E*, **65**, 0419051–04190516.
30. Bernaola–Galván,P., Ivanov,P.Ch., Amaral,L.A.N. and Stanley,H.E. (2001) Scale invariance in the nonstationarities of human heart rate. *Phys. Rev. Lett.*, **87**, 168105: 1–4.
31. Zhang,C.T. and Zhang,R. (2004) Isochore structures in the mouse genome. *Genomics*, **83**, 384–394.
32. Carpena,P., Bernaola-Galván,P., Román-Roldán,R. and Oliver,J.L. (2002) Simple and species-independent coding measure. *Gene*, **300**, 97–104.