LETTER

How Not to Search for Isochores: A Reply to Cohen et al.

Oliver Clay and Giorgio Bernardi

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

In a recent paper in these pages, Cohen et al. search for isochores in the human genome, based on a system of attributes that they assign to isochores. The putative isochores that they find and choose for presentation are almost all below 45% GC and cover only about 41% of the genome. Closer inspection reveals that the authors' methodology systematically loses GC-rich isochores because it does not anticipate the considerable fluctuations and corresponding long-range correlations that characterize mammalian DNA and that are highest in GC-rich DNA. Thus, they over-fragment GC-rich isochores (and also many GC-poor isochores) beyond recognition.

In a recent paper published in these pages, Cohen et al. (2005) describe their search for isochores in the human genome. In their opinion, their failed attempts to find what they were looking for "undermine the utility of the isochore theory" and suggest "that the isochore theory has reached the limit of its usefulness as a description of genomic compositional structures." The reader is left with the impression that the "isochore model has been one of the most useful" ones "in molecular evolution for the last 30 years" but that the isochore concept is now outdated; while it may still retain the role of a relic "suitable for didactic purposes," the authors doubt that it is satisfactory "for more purist aims." The authors therefore call for "a more useful metaphor" that will adequately describe megabase-scale base compositional variation. They do not indicate where or how it will be found. They also do not comment on the fact that nobody, including themselves, has found this elusive new metaphor over the past 30 years.

Similar contentions to those of Cohen et al. (2005) (and often similarly phrased) appeared in print very soon after the first human chromosomes were sequenced (Eyre-Walker and Hurst 2001; Häring and Kypr 2001; Lander et al. 2001; Nekrutenko and Li 2001). The most memorable edict appeared in the paper presenting the public draft sequence of the human genome: Lander et al. (2001) claimed that they were able to "rule out a strict notion of isochores as compositionally homogeneous," so that isochores would "not appear to deserve the prefix 'iso'." These criticisms were soon shown to be incorrect or irrelevant. The logical flaws were slightly different in each of the papers, but most could be traced back to misunderstandings and/or statistical oversights, some of them quite elementary (Bernardi 2001; Clay and Bernardi 2001*a*, 2001*b*; Clay et al. 2001; Li 2002; Li et al. 2003). Now, years later, Cohen et al. (2005) revive old contentions and add some new criticisms of their own, which we address here. Before doing so, we comment on the authors' reference to an "isochore theory"; we do not know of any such theory that was proposed in the past. The basic facts were observed or discovered by ultracentrifugation not by theorizing. They have since been confirmed by genome sequences and are not controversial.

Key words: base composition, evolution, heterogeneity, long-range correlations.

E-mail: bernardi@szn.it.

Mol. Biol. Evol. 22(12):2315–2317. 2005
doi:10.1093/molbev/msi231
Advance Access publication August 10, 2005

Cohen et al. (2005) propose that isochores are, or should be, well characterized by three "attributes" or properties, to which they assign a kind of axiomatic status. These three basic attributes or "selection criteria" for isochores are that they should differ (significantly) in GC from the isochores that flank them (A1), that they should be less heterogeneous than the chromosome on which they reside (A2), and that they should be longer than a cutoff value such as 300 kb (A3). The authors then go on to check whether these properties lead to human isochores satisfying three other properties: the human isochores should cover most of the genome (A4), they should form, on the basis of their GC alone, five isochore families (A5), and strict assignment of isochores to their families should be possible just by looking at the compositional properties of the sequences (A6).

We acknowledge the attempt of Cohen et al. (2005) to propose a minimal set of attributes. To prevent further misunderstandings, however, we also think that it is important to identify where they may have been inappropriately chosen or applied.

The first of the attributes, namely, contrast with adjacent isochores (A1), is appropriate. The authors are, however, wrong in assuming that the DNA segmentation algorithm they choose is "specially suited." It is the original, earliest version of an algorithm (Bernaola-Galván, Román-Roldán, and Oliver 1996) that has since been updated with many improvements. Cohen et al. (2005) cite a publication that explicitly describes a more recent version (Oliver et al. 2002; see Oliver et al. 2004, for the most recent presentation), but they then use the oldest variant. They afterward seem surprised to find a huge number of very, very short segments. The superposition of the segmentation of Cohen et al. (2005) on the plots of chromosome 21 according to the newer algorithm (Oliver et al. 2004) shows, after some searching for mislabeled contigs, that the two coincide largely where segments are long but never where they are very short: Cohen et al. (2005) do not coarse-grain at 3 kb prior to segmentation, as is done in Oliver et al. (2002, 2004). A limit of 3 kb was implicit (Macaya et al. 1976; Cuny et al. 1981) and later explicit (G. Bernardi and G. Bernardi 1986) in the definition of isochores and corresponds to a practical limit of resolution in density gradient experiments.

More generally, segments will tend to be systematically short when one neglects compositional fluctuations, which increase from GC-poor to GC-rich isochores (Cuny et al. 1981; for explanations, see Bernardi 2004). The fluctuations in GC-rich isochores are much higher than if the nucleotides

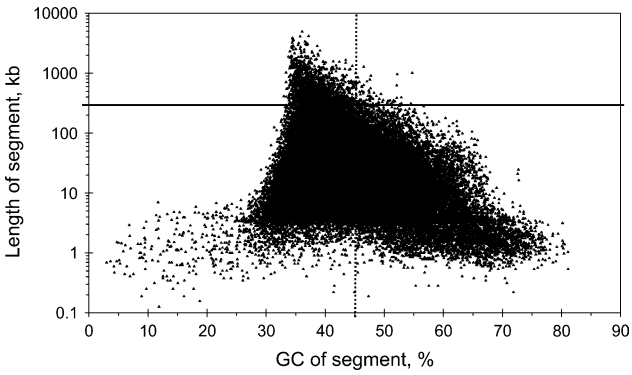

FIG. 1.—Compositions and lengths of the human genome segments obtained by Cohen et al. (2005). The points indicate segments that the authors found to satisfy their requirement (A1); those above the horizontal line are those that survived the subsequent application of their requirement (A3), i.e., that exceed 300 kb. The vertical dotted line indicates 45% GC, beyond which there are relatively few of the long segments that cover 41% of the genome. Data are taken directly from the authors' table S1.

(or short windows) in those isochores were independent. This property of mammalian DNA, known for a long time from experiments and confirmed by the corresponding sequence studies (Clay et al. 2001), is responsible for the very large number of short segments reported in GC-rich isochores. Short segments of this type may often correspond to compositionally prominent features at the gene and subgene scales, such as interspersed repeats, matrix attachment regions, CpG islands, exons, or introns. Almost all standard statistical contrast tests rely on independent and identically distributed nucleotides for their validity and will therefore report a large proportion of false positive contrasts in real DNA (see, however, Beran 1989, 1994, section 8.6, for tests that take long-range dependence into account).

The second attribute, homogeneity relative to the chromosome (A2), is a novel misunderstanding that is strange because the authors cite a paper (Bernardi 2001; itself about misunderstandings) that never mentions or suggests such a criterion. The criterion may happen to work when a chromosome is as heterogeneous as the entire genome, but it is certainly wrong for those chromosomes that contain practically no GC-poor or no GC-rich regions (see Pavlícek et al. 2002, for the chromosomes' GC distributions). Karyotypic changes are frequent in vertebrates, so an attribute that artificially restricts attention to individual chromosomes would not be evolutionarily robust (as well as difficult to justify). By closing one's eyes to the heterogeneity of the genome, one will artificially inflate the percentage of the genome that cannot be assigned to isochores. In a vertebrate such as chicken, which has many of its GC-rich regions in mini/micro-chromosomes, this inflation will be even worse.

The third attribute, "minimum length" (A3), is roughly appropriate where the authors refer to the literature and use the word "typically" but inappropriate where they subsequently hard-wire a cutoff of 300 kb (the one they emphasize most in the text and abstract) and then discard all shorter segments as "nonisochoric" DNA. This sharp cutoff is apparently not inspired by the literature, and there is no precedent of its use of which we are aware. Isochores should not be confused with the largest DNA fragments that experimentally revealed their presence and order of magnitude. Both are longer than 300 kb but that is about all their length distributions have in common.

At this point, the authors are already chaining three criteria, with inevitable propagation of errors: where the outdated criterion (A1) was inappropriate, this affects the significance found in the intrachromosomal comparison (A2), and these in turn affect the number of segments left that exceed 300 kb. Echoing the words of Lander et al. (2001), Cohen et al. (2005) deem these the only ones "that warrant the label 'isochore'" and report that they cover only about 41% of the human genome, thus apparently falsifying attribute (A4). That GC-rich DNA is more heterogeneous than GC-poor DNA has been repeatedly shown in ultracentrifugation experiments and sequence analyses (Cuny et al. 1981; Clay et al. 2001). This fact is visible also in the output from the authors' analysis, and figure 1 shows two columns of their table S1 as a scatterplot. The DNA that does not obey the serial application of criteria (A1–A3) and is therefore discarded (below the horizontal line in fig. 1) contains almost all the GC-rich DNA, so the remaining "putative isochores" will collectively be too GC poor, as will the Gaussian components of the isochores and of their DNA (data set i, figs. 4–6 of Cohen et al. 2005). It is not a surprise, then, that the authors find components (isochore families) that are markedly and systematically "different from those specified in Bernardi (2001)," namely, with GC levels that are systematically too low. The authors' statement refers only to the 41% of the DNA that survives their compositionally biased "selection" procedure, not to the human genome.

The authors' "findings" on isochore families and their comments on attributes (A5) and (A6) are hardly new because most of them were already discussed in Zoubak, Clay, and Bernardi (1996). As mentioned above, the GC levels of the Gaussian components found by Cohen et al. (2005) are systematically too low because the distributions being decomposed systematically lack GC-rich DNA.

In conclusion, Cohen et al. (2005) have been misled by using an inadequate methodology that is unable to deal with the increasing compositional fluctuations associated with increasing GC levels of isochores. Although it is reasonable to expect that isochores should be identifiable from the genome sequence alone, the criteria used to determine them should result in isochores that concord, for example, with replication timing, banding, compositional syntenies, interphase looping, and gene density. In our opinion, the authors have therefore also been wrong in neglecting the "hundreds of studies" and the "dozens of genetic and genomic features … associated with nucleotide composition" that they mention (reviewed in Bernardi 2004), which show that the GC-rich isochores of mammals and birds emerged by GC increases of genome regions that were much less GC rich in fishes and amphibians, thus giving rise to the mosaic organization of the genomes of the former.

Literature Cited

Beran, J. 1989. A test of location for data with slowly decaying serial correlations. Biometrika **76**:261–269.

———. 1994. Statistics for long-memory processes. Chapman & Hall/CRC, Boca Raton, Fla.

Bernaola-Galván, P., R. Román-Roldán, and J. L. Oliver. 1996. Compositional segmentation and long-range fractal correlation in DNA sequences. Phys. Rev. E **53**:5181–5189.

Bernardi, G. 2001. Misunderstandings about isochores: part 1. Gene **276**:3–13.

———. 2004. Structural and evolutionary genomics: natural selection in genome evolution. Elsevier, Amsterdam.

Bernardi, G., and G. Bernardi. 1986. Compositional constraints and genome evolution. J. Mol. Evol. **24**:1–11.

Clay, O., and G. Bernardi. 2001*a*. Compositional heterogeneity within and among isochores in mammalian genomes—II. Some general comments. Gene **276**:25–31.

———. 2001*b*. Isochores: dream or reality? Trends Biotechnol. **20**:237.

Clay, O., N. Carels, C. Douady, G. Macaya, and G. Bernardi. 2001. Compositional heterogeneity within and among isochores in mammalian genomes—I. CsCl and sequence analyses. Gene **276**:15–24.

Cohen, N., T. Dagan, L. Stone, and D. Graur. 2005. GC composition of the human genome: in search of isochores. Mol. Biol. Evol. **22**:1260–1272.

Cuny, G., P. Soriano, G. Macaya, and G. Bernardi. 1981. The major components of the mouse and human genomes. I. Preparation, basic properties and compositional heterogeneity. Eur. J. Biochem. **115**:227–233.

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

Häring, D., and J. Kypr. 2001. No isochores in the human chromosomes 21 and 22? Biochem. Biophys. Res. Commun. **280**:567–573.

Lander, E. A., L. M. Linton, B. Birren et al. (256 co-authors). 2001. Initial sequencing and analysis of the human genome. Nature **409**:860–921.

Li, W. 2002. Are isochore sequences homogeneous? Gene **300**:129–139.

Li, W, P. Bernaola-Galván, P. Carpena, and J. L. Oliver. 2003. Isochores merit the prefix 'iso'. Comput. Biol. Chem. **27**:5–10.

Macaya, G., J. P. Thiery, and G. Bernardi. 1976. An approach to the organization of eukaryotic genomes at a macromolecular level. J. Mol. Biol. **108**:237–254.

Nekrutenko, A., and W. H. Li. 2001. Assessment of compositional heterogeneity within and between eukaryotic genomes. Genome Res. **10**:1986–1995.

Oliver, J. L., P. Carpena, M. Hackenberg, and P. Bernaola-Galván. 2004. IsoFinder: computational prediction of isochores in genome sequences. Nucleic Acids Res. **32**:W287–W292.

Oliver, J. L., P. Carpena, R. Román-Roldán, T. Mata-Balaguer, A. Mejias-Romero, M. Hackenberg, and P. Bernaola-Galván. 2002. Isochore chromosome maps of the human genome. Gene **300**:117–127.

Pavlícek, A., J. Paces, O. Clay, and G. Bernardi. 2002. A compact view of isochores in the draft human genome sequence. FEBS Lett. **511**:165–169.

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

Takashi Gojobori, Associate Editor

Accepted August 3, 2005