GC3 of Genes Can Be Used as a Proxy for Isochore Base Composition: A Reply to Elhaik et al.

Oliver K. Clay[1,2,3,*] and Giorgio Bernardi[3,†]

[1]School of Medicine and Health Sciences, Universidad del Rosario, Bogotá, Colombia

[2]Cellular and Molecular Biology Unit, Corporación para Investigaciones Biológicas, Medellín, Colombia

[3]Laboratory of Molecular Evolution, Stazione Zoologica Anton Dohrn, Naples, Italy

[†]Present address: Biology Department, Rome University 3, Rome, Italy.

***Corresponding author:** E-mail: oliver.clay@gmail.com.

Associate editor: Takashi Gojobori

Abstract

In an article published in these pages, Elhaik et al. (Elhaik E, Landan G, Graur D. 2009. Can GC content at third-codon positions be used as a proxy for isochore composition? *Mol Biol Evol*. 26:1829–1833) asked if GC3, the GC level of the third-codon positions in protein-coding genes, can be used as a "proxy" to estimate the GC level of the surrounding isochore. We use available data to directly answer this simple question in the affirmative and show how the use of indirect methods can lead to apparently conflicting conclusions. The answer reasserts that in human and other vertebrates, genes have a strong tendency to reside in compositionally corresponding isochores, which has far-reaching implications for genome structure and evolution.

Key words: isochores, GC3, GC level, genome composition, long-range correlations, chromatin.

In a paper published last year, Elhaik et al. (2009) offered readers of this journal the title question: "Can GC content at third-codon positions be used as proxy for isochore composition?" Their own unexpected conclusion was that "the GC content of third-codon position cannot be used as stand-in for isochoric composition."

Today, the "proxying" addressed by Elhaik et al. is hardly used anymore. In the past, it was used as a successful tool for exploring genomes and, in particular, for predicting the gene density distribution of the human genome before it was sequenced (Zoubak et al. 1996 and references therein; see also supplementary material, Supplementary Material online). Now large chromosomal sequences are available, and one can calculate the GC of an isochore by counting G's and C's along the assembled isochore sequence, so one no longer needs an indirect proxy or "stand-in" to estimate an isochore's GC. The question posed by the authors has, however, far-reaching implications for genome structure and evolution, so it is important to clarify it.

In figure 1, we show a scatterplot and a corresponding contour plot of the GC3 of human genes versus the GC of the embedding isochores (Costantini et al. 2006). The correlation coefficient r is 0.64 (see also supplementary material, Supplementary Material online). Unless one categorically considers any correlation coefficient below 0.7 void (see below), these plots already answer the question of Elhaik et al. in the affirmative.

To support their opposite conclusion, Elhaik et al. present plots for six vertebrate species, showing r^2 values between genes' third-codon positions and successively distant regions of size 5 kb, for intervening distances ranging from 0 to 200 kb on either side of the gene. The authors comment on the "sharp decrease" in compositional correlation coefficients r with increasing distance from the gene, for example, in human and cow. For example, in human, r^2 is about 0.4 for the 5-kb region immediately flanking (adjacent to) the gene but drops to $r^2 < 0.2$ when the 5-kb region is farther from the gene than 50 kb. The results for immediately flanking regions are in moderate agreement with a previous report (Costantini and Bernardi 2008, supplementary table T2), although the previous study found much higher r values for chicken and zebrafish (see supplementary material, Supplementary Material online). Elhaik et al. voice caution that, as a consequence of their observations, "all associations between isochores and genic features (e.g., gene length, gene density, chromosomal bands) that have been reported or suggested in the literature should be re-evaluated if GC3 was used as a proxy for the GC content of isochores, as it was almost invariably done in the past." We point out that even if their own data had conclusively answered their question, this caution would be unnecessary. The GC3 method was not almost invariably used in the past, as direct experimental estimates were used in parallel and gave concordant results; furthermore, many if not most of the associations between isochores and genic features have since been directly confirmed using whole-genome sequences without proxying (Costantini et al. 2006, 2009; Costantini, Auletta, et al. 2007; Costantini, Di Filippo, et al. 2007; Costantini and Bernardi 2008).

Clearly, two conceptually distinct types of correlation are under consideration and they should not be confused.

Letter

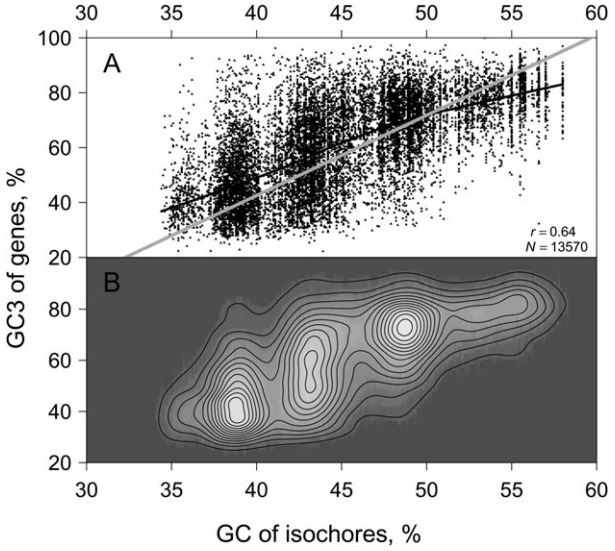

Fig. 1 (*A*) Scatterplot of GC3 of human genes versus GC of the isochore in which the genes are located. Isochore boundaries and GC levels are from the isochore map of Costantini et al. (2006, hg17); corresponding human genes' coding sequence (CDS) co-ordinates were taken from the Consensus CDS (CCDS; Pruitt et al. 2009) track of the UCSC Genome Browser Database ($N = 13{,}570$; $r = 0.64$). Repeats were not masked. Straight gray line: estimator used by Zoubak et al. (1996) to predict the gene density distribution of the human genome; bent black line: lowess fit of the data. (*B*) Corresponding contour plot. Clusters/peaks correspond to isochore families.

The object of the title and conclusion of Elhaik et al. is the correlation between the GC3 of genes and the GC of the full isochores that embed them. The object of the calculations they offer is the correlation between the GC3 of genes and the GC of (possibly small and possibly distant) subregions of the isochores that embed them. In the second case, one is just correlating one proxy of isochoric GC against another potential proxy, not against the isochore itself, so any answers one obtains will be for a different question: Can a gene's GC3 be used as a proxy for the GC of possibly small and possibly distant subregions of the isochore? As the figures of Elhaik et al. indicate, the answer is generally "no."

How can the two questions yield such different answers? The full explanation leads into statistically deep waters, but can be summarized by saying that the difference is just what one expects from the long-range correlations or long memory, that are particularly obvious in human and other warm-blooded vertebrate genomes. Processes of long-range dependence have been eloquently summarized by Cox (1984): "From a direct statistical point of view, their salient feature is perhaps that the variance of a mean decreases more slowly than the reciprocal of sample size, with implications for choice of sample size, for instance. From an intuitive point, possibly the most enlightening property is that the averaged process [asymptotically] takes a nondegenerate correlational structure." A complex coexistence of mosaicism with long-range correlation is found in human and other vertebrate genomes (Carpena et al. 2007). Thus, GC heterogeneities are consistently higher than one expects for "textbook" scenarios where no or only short-range (e.g., Markov chain) serial autocorrelations are present (for asymptotic formulae see Beran 1994, theorem 2.2 and section 4.4). This observation on heterogeneity is not new. It dates back to early ultracentrifuge experiments (Macaya et al. 1976, figure 8; Hudson et al. 1980, figure 2; Cuny et al. 1981, figure 5; see also Bernardi 2001), can be inferred a posteriori from even earlier ones (Meselson et al. 1957, figure 4), and can now be easily confirmed using sequences of well-characterized isochores in human and other genomes (see, e.g., Clay et al. 2001, figure 2). The higher the GC of the isochore, the higher will be the heterogeneity and, in particular, the average standard deviation among the isochore's 5-kb segments (in the sequence set used for fig. 1, the heterogeneity regression slope is 0.16 and its r is 0.67). The notable variability among 5-kb segments within a human isochore remains, however, substantially lower than that observed within the entire genome, so the original definition of isochores stays valid.

What about longer fragments than 5 kb? Elhaik et al. used fixed-length fragments of up to 100 kb, located as far away from the gene as 200 kb. Although one might correctly expect less heterogeneity among the larger segments of an isochore, this advantage is offset by a growing risk that the gene's segment can land in, or straddle, a different isochore with a different GC level, for example, if the gene is near the border of its isochore. The problem is that the authors did not use any of the isochore maps available to them (and that they have recently reviewed in Elhaik et al. 2010). One might argue that so far no isochore map is perfect, but one cannot logically answer the authors' question if one uses only fixed-length windows and does not try to see where the isochore boundaries are. If one takes this route, one always reaches an impasse: either the segments are too short for their GC to serve as a substitute for the isochore's GC and/or they are not safely within the same isochore as the gene and can, therefore, again not be expected to substitute for it.

A further reason for the discordant conclusion of Elhaik et al. is their unusually strict criterion for proxy, that is, for what they think GC3 must be able to do: "a proxy must be able to explain most of the variation in GCf, not merely be correlated with it." In their interpretation, as we understand it, this means that r^2 values below 0.5 would be unacceptable. We have not seen such a strict criterion elsewhere, applied to DNA in a similar context (see supplementary material, Supplementary Material online).

Elhaik et al. also raise a deeper concern: "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." It would seem that a statistical predictor is validated or invalidated by pragmatic success not by ease of interpretation. In this case, however, the interpretation is not difficult: Different evolutionary processes act on a genome's coding and/or noncoding DNA, but in such a concerted or concordant way that the compositional organization of the genome and its corresponding correlations are maintained (Bernardi, 2005). For example, mobile elements occupy

a formidable percentage of the human genome's noncoding DNA, but apparently only a very small percentage of the genome's DNA encoding functional human proteins (Pavlícek et al. 2002; Gotea and Makałowski 2006; Piriyapongsa et al. 2007), so the evolution of such interspersed repeats would be a process differing between coding and noncoding DNA. This difference does not prevent the correlation shown in figure 1. (A visual summary of how repeats contribute to GC-poor and GC-rich regions, and their base compositions in human and mouse is given in Paces et al. 2004, figure 3.)

In conclusion, human and other vertebrate genes do show a marked tendency to be found in isochores of corresponding GC, as is illustrated for human in figure 1. The GC3 of a gene can, therefore, be used as a proxy for the GC of the isochore that embeds it, or, more precisely, as the key variable for constructing an estimator of the isochore's GC. Linear estimators are a simple choice and allowed accurate estimates of the human genome's gene density distribution (cf. Zoubak et al. 1996, figure 4*a* with Lander et al. 2001, figure 36*b*; the line used by Zoubak et al., $GC_{isochore,est.} = 0.342\ GC3 + 25.45$, is shown also here in fig. 1*A* to document consistency). For some other purposes, a nonlinear choice may be best, as suggested by the lowess fit in figure 1*A*, or a compound estimator incorporating also genes' dinucleotide information. Indeed, recent leaps in our understanding of mechanisms by which GC-rich and CpG-rich regions intrinsically facilitate open chromatin (see, e.g., Blackledge et al. 2010; Thomson et al. 2010) reaffirm that the question posed by Elhaik et al. is likely to have deep functional correlates.

Supplementary Material

Supplementary material is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Maria Costantini and Stelios Arhondakis for discussions and two anonymous referees for helpful suggestions.

References

Beran J. 1994. Statistics for long-memory processes. Boca Raton (FL): Chapman & Hall.

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

Bernardi G. 2005. Structural and evolutionary genomics: natural selection in genome evolution. Amsterdam: Elsevier.

Blackledge NP, Zhou JC, Tolstorukov MY, Farcas AM, Park PJ, Klose RJ. 2010. CpG islands recruit a histone H3 lysine 36 demethylase. *Mol Cell*. 38:179–190.

Carpena P, Bernaola-Galván P, Coronado AV, Hackenberg M, Oliver JL. 2007. Identifying characteristic scales in the human genome. *Phys Rev E*. 75:032903.

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

Costantini M, Auletta F, Bernardi G. 2007. Isochore patterns and gene distributions in fish genomes. *Genomics* 90:364–371.

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

Costantini M, Cammarano R, Bernardi G. 2009. The evolution of isochore patterns in vertebrate genomes. *BMC Genomics*. 10:146.

Costantini M, Clay O, Auletta F, Bernardi G. 2006. An isochore map of human chromosomes. *Genome Res*. 16:536–541.

Costantini M, Di Filippo M, Auletta F, Bernardi G. 2007. Isochore pattern and gene distribution in the chicken genome. *Gene* 400:9–15.

Cox DR. 1984. Long-range dependence: a review. In: David HA, David HT, editors. Statistics: an appraisal. Aimes (IA): Iowa State Univ Press. p. 55–74.

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

Elhaik E, Graur D, Josic K. 2010. Comparative testing of DNA segmentation algorithms using benchmark simulations. *Mol Biol Evol*. 27:1015–1024.

Elhaik E, Landan G, Graur D. 2009. Can GC content at third-codon positions be used as a proxy for isochore composition? *Mol Biol Evol*. 26:1829–1833.

Gotea V, Makałowski W. 2006. Do transposable elements really contribute to proteomes? *Trends Genet*. 22:260–267.

Hudson AP, Cuny G, Cortadas J, Haschemeyer AE, Bernardi G. 1980. An analysis of fish genomes by density gradient centrifugation. *Eur J Biochem*. 112:203–210.

Lander ES, Linton LM, Birren B, et al. (255 co-authors). 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860–921.

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.

Meselson M, Stahl FW, Vinograd J. 1957. Equilibrium sedimentation of macromolecules in density gradients. *Proc Natl Acad Sci USA*. 43:581–588.

Paces J, Zika R, Paces V, Pavlícek A, Clay O, Bernardi G. 2004. Representing GC variation along eukaryotic chromosomes. *Gene* 333:135–141.

Pavlícek A, Clay O, Bernardi G. 2002. Transposable elements encoding functional proteins: pitfalls in unprocessed genomic data? *FEBS Lett*. 523:252–253.

Piriyapongsa J, Rutledge MT, Patel S, Borodovsky M, Jordan IK. 2007. Evaluating the protein coding potential of exonized transposable element sequences. *Biol Direct*. 2:31.

Pruitt KD, Harrow J, Harte RA, et al. (49 co-authors). 2009. The consensus coding sequence (CCDS) project: identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res*. 19:1316–1323.

Thomson JP, Skene PJ, Selfridge J, et al. (13 co-authors). 2010. CpG islands influence chromatin structure via the CpG-binding protein Cfp1. *Nature* 464:1082–1086.

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