Segmenting the Human Genome into Isochores

Paolo Cozzi[1,2], Luciano Milanesi[1] and Giorgio Bernardi[1,3]

[1]National Research Council, Institute for Biomedical Technologies, Segrate, Milan, Italy. [2]Parco Tecnologico Padano, Lodi, Italy. [3]Science Department, Rome 3 University, Rome, Italy.

ABSTRACT: The human genome is a mosaic of isochores, which are long (>200 kb) DNA sequences that are fairly homogeneous in base composition and can be assigned to five families comprising 33%–59% of GC composition. Although the compartmentalized organization of the mammalian genome has been investigated for more than 40 years, no satisfactory automatic procedure for segmenting the genome into isochores is available so far. We present a critical discussion of the currently available methods and a new approach called isoSegmenter which allows segmenting the genome into isochores in a fast and completely automatic manner. This approach relies on two types of experimentally defined parameters, the compositional boundaries of isochore families and an optimal window size of 100 kb. The approach represents an improvement over the existing methods, is ideally suited for investigating long-range features of sequenced and assembled genomes, and is publicly available at https://github.com/bunop/isoSegmenter.

KEYWORDS: bioinformatics, comparative genomics, evolution

CITATION: Cozzi et al. Segmenting the Human Genome into Isochores. *Evolutionary Bioinformatics* 2015:11 253–261 doi: 10.4137/EBO.S27693.

TYPE: Methodology

RECEIVED: April 15, 2015. **RESUBMITTED:** August 25, 2015. **ACCEPTED FOR PUBLICATION:** August 31, 2015.

ACADEMIC EDITOR: Jike Cui, Associate Editor

PEER REVIEW: Five peer reviewers contributed to the peer review report. Reviewers' reports totaled 1888 words, excluding any confidential comments to the academic editor.

FUNDING: This investigation was funded by the Institute for Medical Biotechnologies of the National Research Council. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: gbernardi@uniroma3.it

Introduction

A compositional strategy was developed almost 50 years ago in order to understand the complex organization of the eukaryotic genome. This strategy relied on the most elementary, yet the most fundamental, property of DNA, the frequency of short sequences (3–5 nucleotides in size), and, as a proxy, the base composition. The rationale was that the properties of the genome basically depend upon the composition of its coding and noncoding nucleotide sequences. Originally, the compositional approach was based on the high resolution of DNA preparations (10–20 kb in size), as obtained by preparative ultra centrifugation in Cs_2SO_4 density gradients run in the presence of sequence-specific ligands, such as silver ions[1] and, later, 3,6-bis-(acetatomercurimethyl)-dioxane (BAMD).[2] This approach fractionated DNA fragments according to the density of the short nucleotide sequences that were binding the ligand (Supplementary Fig. 1). This fractionation cannot be achieved by approaches (such as CsCl ultracentrifugation) that rely only on GC levels. Since short sequences determine the fine structure of DNA, as well as its interaction with proteins such as histones and transcription factors, the compositional strategy leads, in fact, to a fractionation of the genome on the basis of its structure and function.

Using the original ultracentrifugation approach, the compositional strategy led to a breakthrough, namely the demonstration that the genomes of vertebrates (neglecting satellite DNAs) are compartmentalized in terms of base composition, such that they can be resolved into a small number of "major components"[3] characterized by different frequencies of short nucleotide sequences.[4] In fact, the DNA molecules of the major components are derived (by degradation during preparation) from DNA stretches that were originally estimated to be more than 300 kb in average size,[5,6] and have a "fairly homogeneous" GC (guanine+cytosine) level, which were called "isochores".[7] Isochores belong to a small number of families that correspond to the above-mentioned "major components" and cover a very broad GC range (33%–59%) in the human genome. Each isochore family covers a range of 4%–7% GC (Supplementary Table 1). Needless to say, as soon as chromosomal and genomic sequences became available, the compositional approach was applied to DNA sequences using GC levels as a proxy for the frequencies of short nucleotide sequences. Indeed, these frequencies are different in different isochore families.[8,9]

Since all structural/functional properties of the genome that could be tested are correlated with the base composition of nucleotide sequences,[10] it is obvious that a map of isochores is of great interest in the study of the organization, function, and evolution of genomes. We will first briefly describe here the experimental approach developed a few years ago and then move on to the new automatic procedure.

The original approach consisted in assessing GC levels of 100-kb DNA stretches over chromosomal sequences.[10–14] More recently, Costantini et al.[15], using UCSC release hg17, partitioned the entire chromosomal sequences of the human genome[16–18] assembly into nonoverlapping 100-kb windows and calculated their GC levels using the program draw_chromosome_gc.pl (http://genomat.img.cas.cz).[13,14] The choice of the window size of 100 kb was justified by the fact that the high variances observed when using smaller windows (due to frequency variations in exons, introns, CpG islands, 3′- and 5′-untranslated regions, scaffold/matrix attachment regions, and, especially, interspersed repeats) decrease with increasing window size, reaching a plateau value at 100 kb (Fig. 2 of Ref. 15).

The profile of GC levels of the 100-kb windows in each chromosome was scanned for steps that were detectable on the basis of GC differences between contiguous windows. In other words, isochore borders were identified on the basis of marked compositional differences of contiguous isochores that belonged to different families (Fig. 3B of Ref. 15). The results obtained via this simple procedure, which involved only properties of bulk DNA and no annotated features, demonstrated a complete coverage of the human euchromatic genome sequence by 3,159 isochores having an average size of 900 kb and totaling 2,854 Mb.[15] As already mentioned, isochores belong to families characterized by GC levels that are comparable with those initially detected by ultracentrifugation experiments.[6,19]

Putting isochore size into bins of 1% GC revealed local maxima (peaks) of isochore families at 35.5%, 38.7%, 43.0%, 48.5%, and 55.0% GC for L1, L2, H1, H2, and H3 isochores and local minima (valleys) at 37%, 41%, 46%, and 53% (Supplementary Table 1 and Supplementary Fig. 2A).[15] The existence of isochore families in the human genome, originally based on the ultracentrifugation experiments and short sequence frequencies already mentioned, is also supported by the multimodal distribution of coding sequences (Supplementary Fig. 2B),[20] as well as by the presence of genes that largely belong to different functional classes,[10,21] by different chromatin states[22] and by different chromatin structures.[23]

A remarkable discovery was that the GC level ranges of isochore family borders are essentially conserved during evolution.[24–26] Indeed, among vertebrates, compositional genome differences barely concern the isochore family ranges, except the DNA amounts present in isochore families, as shown for fish (four species, zebrafish, medaka, stickleback, and pufferfish, that belong to four distant orders and cover almost the entire GC range of fish genomes), chicken, mammals (chimpanzee, dog, mouse, opossum, platypus), and xenopus. In fact, similar results were also obtained in invertebrates,[27] in which case, however, only a relatively small number of genomes were investigated.

In the human genome, five isochore families were estimated (Supplementary Table 1) to represent 19%, 36%, 31%, 11%, and 3% (from L1 to H3), respectively,[15,19,28] of the total genome size. While these values are essentially conserved among eutherian mammals, they differ in other vertebrates, as already mentioned, GC-rich families being even absent in a number of cold-blooded vertebrates.

Material and Methods

Although the approach of Costantini et al.[15] was satisfactory for investigating a number of genome properties, it had two minor problems and one major practical problem. The first minor problem was that a fixed, nonoverlapping window approach obviously does not cut isochores at their borders and therefore integrates flanking sequences into isochores. However, this problem is less serious than one may imagine because (1) the "wrong" starts and ends of isochores are <100 kb away from the "right" ones; (2) the fraction of the affected isochore is notable only for the smallest isochores, which represent a very small minority of all isochores[15]; and (3) the "wrong" starts and ends of isochores are located in flanking isochores that, as a rule, belong to the compositionally closest families, a fact that minimizes their impact on isochore composition and family assignment. The second minor problem was that in a small number of cases, contiguous stretches characterized by slightly different GC levels were separated despite belonging to the same isochore family.

Although the approach of Costantini et al.[15] led to a definition of isochore families and to a mapping of isochores, despite the above-mentioned minor problems, a serious practical problem remained, namely, the fact that the approach is very labor intensive and time consuming. This hinders a large-scale application of the procedure. Moreover, a small level of subjective decisions about isochore borders is unavoidable.

An automatic compositional approach has, therefore, become an absolute necessity because of the exponentially increasing number of genome sequences that are currently produced and also because of the need to rapidly explore the results obtained when changing parameters. Therefore, we developed a fast approach that is very flexible, which again is a necessity because of the variety of compositional situations in eukaryotic genomes and the different problems that may be investigated.

The approach follows the same logic used in previous investigations in that the compositional (GC) profiles of chromosomes are scanned through fixed, nonoverlapping windows, the optimal value of 100 kb being used as the window size, and the upper and lower GC boundaries of isochore families being those defined by Ref. 15 (Supplementary Table 1). Indeed, 50-kb windows lead to oversegmentation and 200- or 400-kb windows to a loss of resolution (Supplementary Fig. 3).

In order to define isochores on chromosomal DNA sequences, we developed an in-house script in Python,[29] which was applied to the entire sequence of the completed human genome assembly hg38.[18] The approach will now be described as applied to the smallest human chromosome, chromosome 21, which was chosen for the sake of convenience. Indeed, even if only 33 Mb were considered (because the initial stretch, essentially represented by repeated sequences and ribosomal DNA, was neglected), it still comprises the full range from L1 to H3 of isochore families.

The algorithm starts by segmenting the chromosome sequence into nonoverlapping 100-kb DNA segments (as originally done in Ref. 13) and assigns each window to an isochore family (L1, L2, …) relying on the percentage of GC calculated on the sequence itself. The family assignment on the basis of the GC level was that adopted in Ref. 15.

As expected, the compositional profile obtained when plotting all the segments of 100 kb within the euchromatic part of the chromosome long arm (Fig. 1A) shows that there are large regions, well above the window size of 100 kb, in which the GC level shows only very small variations, within the limits of the compositional boundaries of isochore families. Expectedly, this profile is completely different from the profile obtained when DNA segments are randomly reshuffled (not shown).

The procedure, isoSegmenter, developed here for genome segmentation into isochores may be described by presenting the main decision points as algorithm steps. In each case, we merged adjacent windows by relying on the average and standard deviation of their GC levels. Once the step was completed, we assigned each group to an isochore family (L1, L2,.) by relying on the final average GC value.

In the first step (Fig. 1B), contiguous 100-kb segments that belong to the same family are merged and averaged in GC in order to provide an initial definition of isochores.

In the second step (Fig. 1C), isolated, single 100-kb DNA segments from a different, yet compositionally closest, family are merged with the nearest similar adjacent isochore. More precisely, the algorithm evaluates three different cases, which consist in merging the isolated window with the right- or left-hand segments or with both of them. Then, it evaluates the standard deviation between the GC levels of the windows in each of these different cases and chooses the merging that leads to the lowest standard deviation of GC. This step is best illustrated by an example, for istance the region between 17 and 18 Mb of chromosome 21 (Fig. 1B). Starting from the right end, the program will try to merge the isolated H1 segment with the right-hand L2 segments; such a group is then compared with the group composed by the same window and the left-hand L2 segment and with the group composed by the windows comprising both flanking segments. Since the latter combination leads to the lowest standard deviation, the third group is chosen. This step in particular decreases the final isochore number leading to larger homogeneous regions. However, after this step, it is possible that two adjacent groups have the same family classification but are still separated because of slightly different GC levels.

In the final third step (Fig. 1D), groups with the same isochore family classification are merged and the GC levels of contiguous regions are averaged to produce the final isochore representation. Figure 2 presents the results obtained for all human chromosomes.

It should be noted that the sequences were first analyzed in order to identify gap positions, and the gaps >5 kb were excluded from the calculations by resizing the window dimensions. Two windows separated by a gap >5 kb were not assigned to the same isochore, and so when a gap is encountered in the sequence, the current isochore terminates and the next isochore starts after the gap.

The output of our program is a Comma Separated Table in which isochore sizes and positions are reported along with the average GC level of the isochores, their standard deviation, and the difference of GC levels of adjacent isochores (see https://github.com/bunop/isoSegmenter). The program also displays isochore distribution in a graph in which average values of isochores are presented by colored boxes along with their genomic positions, while gaps are shown as gray boxes (Figs. 1 and 2).

Results and Discussion

A comparison of isoSegmenter with previous segmentation methods. Since the past 15 years, a number of genome segmentation approaches were proposed for the human genome.[30–36] Four such approaches, BASIO,[32] GC profile,[33,34] least-square,[35] and isoFinder,[36] were carefully compared by Schmidt and Frishman.[37] These authors stressed the highly different segmentation results (Supplementary Table 2). Indeed, the number of isochores varied from 1,206 (GC-profile) and 1,252 (least-squares) to 38,823 (isoFinder) and 76,833 (BASIO), and the average isochore size varied from 40 kb (BASIO) and 72 kb (isoFinder) to 2,385 kb (GC-profile) and 2,459 kb (least-squares). Therefore, the two extreme cases, BASIO and GC-profile, showed ~60-fold differences in the number and size of isochores, while the other two, isoFinder and least-squares, exhibited ~30- to ~34-fold differences. If one compares all these values with the ~3,200 isochores having an average size of ~900 kb obtained by Costantini et al.[15], one can see that the GC profile and least-square strongly undersegment, whereas isoFinder and BASIO strongly oversegment the genome (Supplementary Table 2).

Despite such striking differences, Schmidt and Frishman[37] realized that “the total amount of genomic DNA classified into the same isochore families is very large, with all methods being in perfect agreement for more than two-thirds of the human genome”. These authors proposed a consensus approach, isoBase (Fig. 3), which implied, however, averaging very different data. Indeed, isoBase comprises 31,176 isochores with an average size of 99 kb, two values very different, roughly by a factor of 10, from the estimates of Costantini et al (Supplementary Table 2).[15] Despite these problems, a number of common features undoubtedly appear among all the patterns of Figure 3, which concerns the telomeric 100 Mb of the short arm of human chromosome 1. Moreover, the estimate of the relative amounts of isochores in the L1 to H3 families in the consensus was not too far from the original estimates of Costantini et al (Supplementary Table 1).[15]

Figure 4 extends the comparisons of Figure 3 to include the results of Costantini et al.[15], those obtained by using

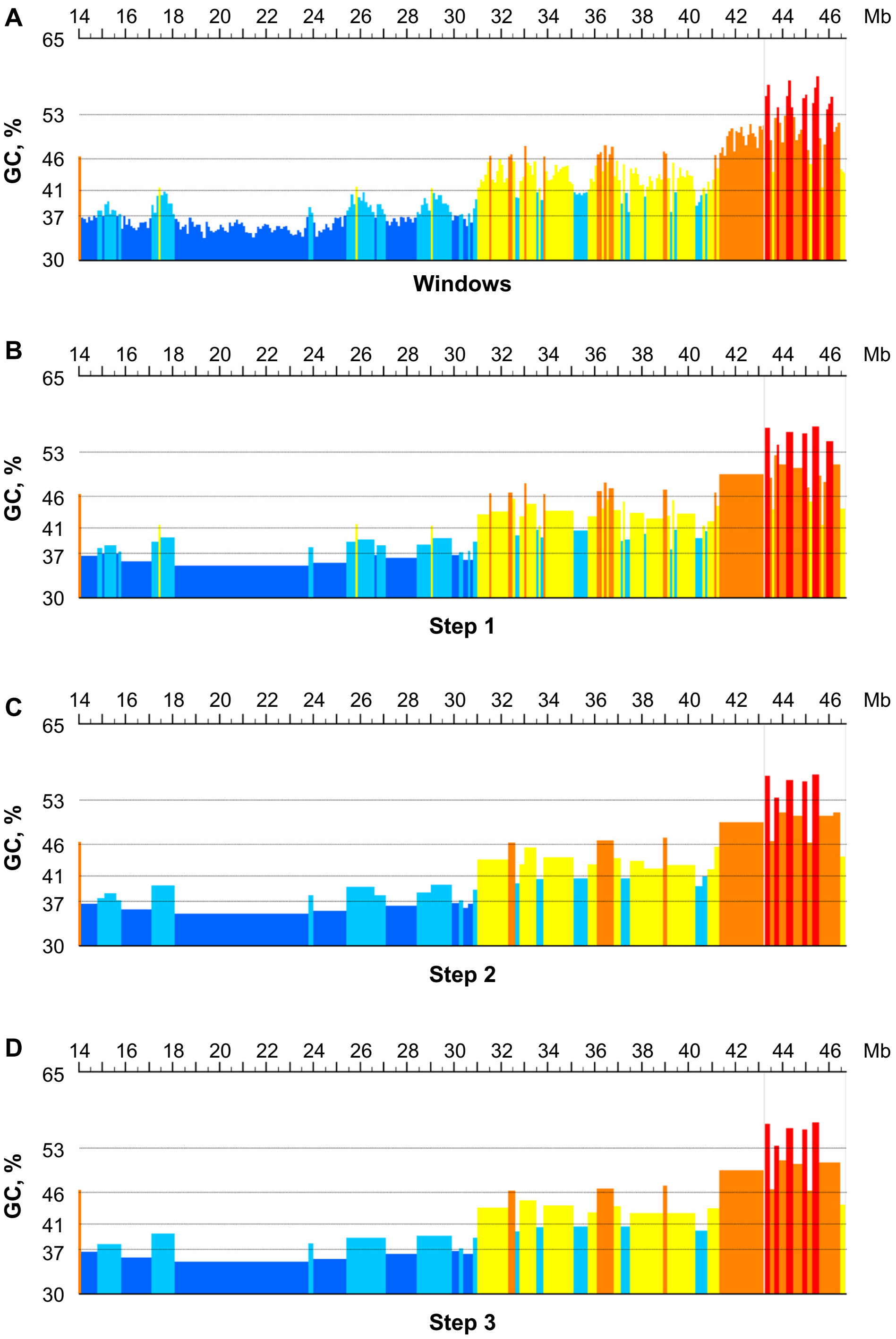

Figure 1. The top frame shows the compositional profile of the human chromosome 21 (release hg38) as seen through nonoverlapping fixed windows of 100 kb. The following frames are described in the Materials and Methods section. The left-hand scale is a GC scale in which minimal GC values (valleys between the peaks of isochore families) are indicated. L1 to H3 isochores are represented here in different colors, deep blue, light blue, yellow, orange, and red. For the following frames, see the text.

isoSegmenter, isoBase, isoFinder, and a fourth program isoPlotter.[38–41] While the first four sets of data showed the results expected from Figure 3, isoPlotter shared no common feature with any of the other approaches. According to the most recent publication on isoPlotter (Table 1 of Ref. 39 and Supplementary Table 2), the human genome comprises 107,571 "compositional domains" with a mean domain size of 25,865 bp (~26 kb). At this size level, the standard deviation of GC level is very high because of the variable contribution especially of repeated sequences (Fig. 2 of Ref. 15), and the long regions belonging to different isochore families detected by all other approaches are missed. The majority of such domains, 74,579 (about 70% of the total), are "homogeneous domains" with a mean size of 29,668 bp (~30 kb). The minority, 32,992

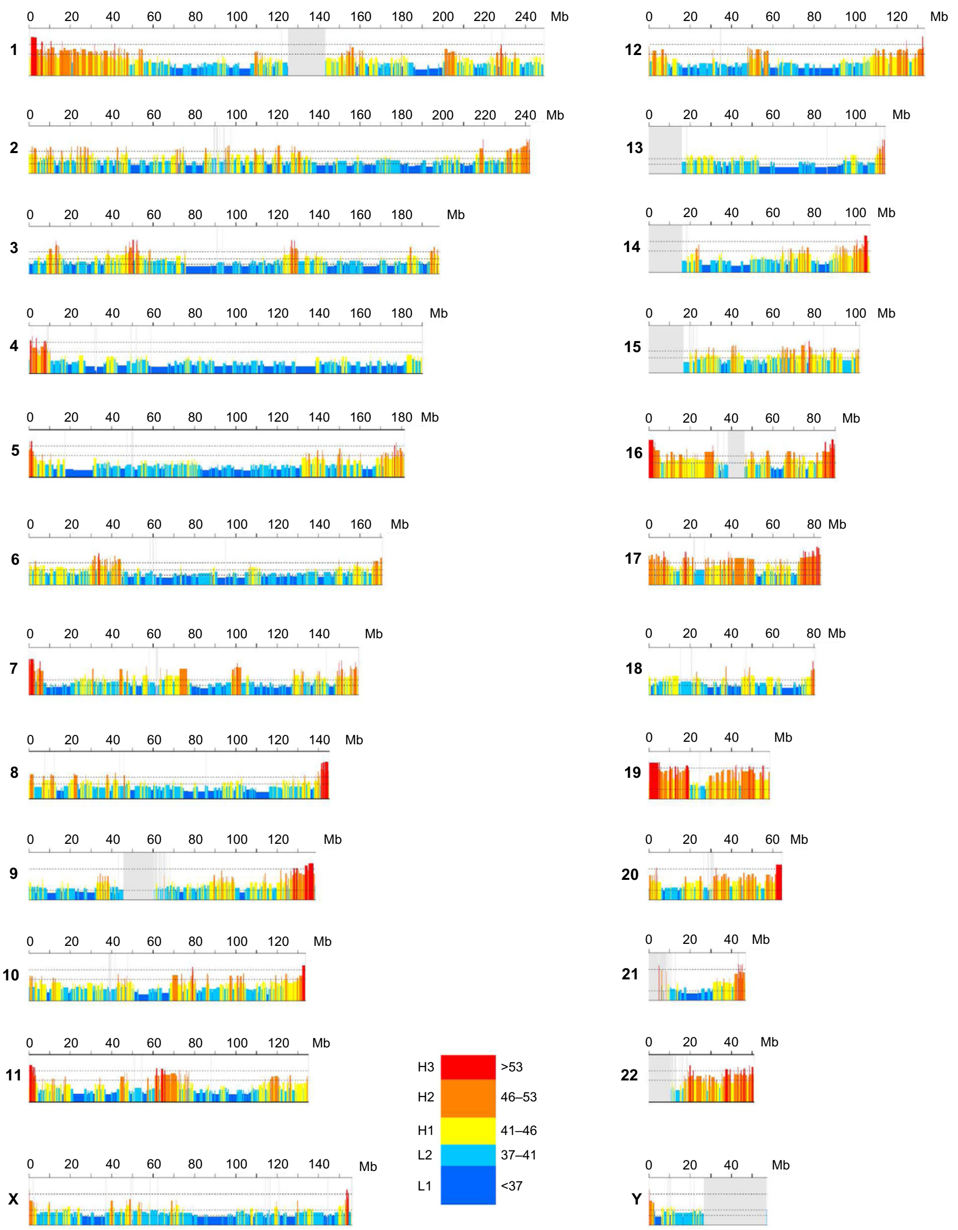

Figure 2. Isochore profiles of all human chromosomes from hg38 release as obtained following the procedure presented in this article. For the left-hand scale and horizontal lines, see the legend of Figure 1.

(about 30% of the total), are "nonhomogeneous domains" with a mean size of 17,269 bp (~17 kb). Finally, 1,071 "isochore domains" have a mean size of 652,778 bp (~653 kb) and represent only 1% of all compositional domains. As shown in Supplementary Table 2, the number and the sizes of "compositional compartments" are much higher and much smaller, respectively, compared to the corresponding figures for isochores of Costantini et al.[15]

Differences between the results of isoPlotter and all other results are extremely striking and call for an explanation, which must concern some fundamental issue(s). In our work, the approach relied on the demonstrated compositional compartmentalization of the mammalian genome into long, fairly homogeneous domains, the isochores that belong to a small number of compositional families. Most importantly, isochore families were found to be associated with all the structural and functional properties of the genome that could be tested (eg, gene density, DNA replication, etc.), a set of correlations called the genomic code.[10] In contrast, the results of isoPlotter are based on "a recursive segmentation algorithm that employs a dynamic threshold which takes into account the composition and length of each segment". In view of the results of Figure 4, isoPlotter can only be defined as an exercise in DNA sequence segmentation with no biological relevance (which, in fact, was not claimed), as indicated by the lack of correlation between the compositional and the structural/functional properties of the genome. Regrettably, Elhaik and Graur[39] made the same mistake of Lander et al.[16], 13 years before, by looking at very

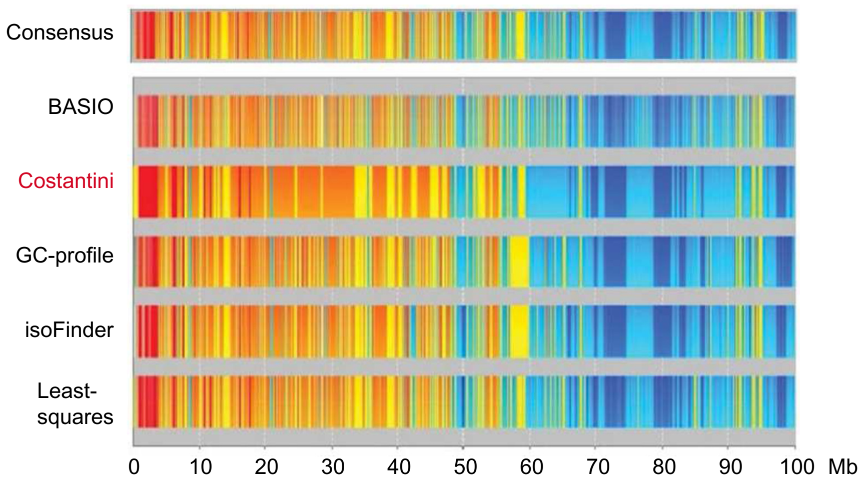

Figure 3. Graphical representation of the isochore assignments for the first 100 Mb of the telomeric end of the short arm of human chromosome 1. The results of the experimental approach of Costantini et al.[15] are compared with those obtained from several computational approaches and with their consensus (from Schmidt and Frishman[37]).

short sequences (20–25 kb) characterized by highly variable GC levels that led to the denial of the existence of isochores.

Finally, a Bayesian analysis of isochores,[42] as applied to the *Anolis* reptile genome,[43] led to the wrong conclusion that this genome is a genome without isochores, a conclusion in conflict with our finding of an isochore structure in all vertebrate genomes tested including *Anolis*.[24–26] In fact, previous results[26] identified in the scaffolds available at that time one major, L2, and one minor, H1, isochore family, as well as very small amounts of isochores from the L1 and H2 families,

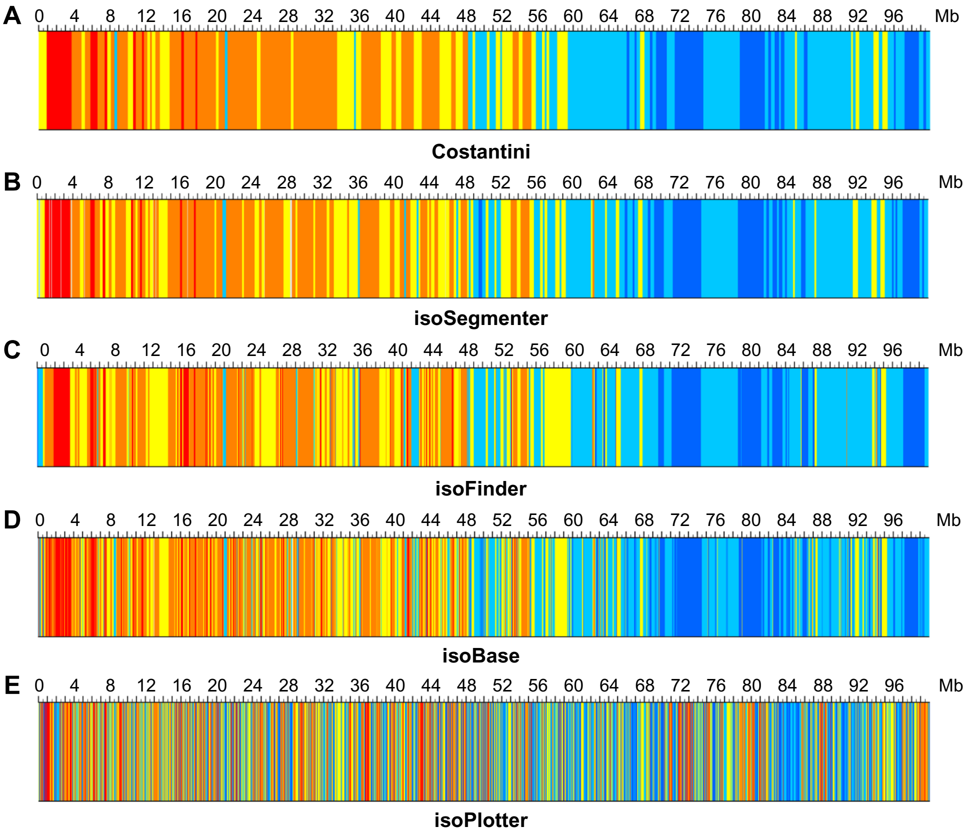

Figure 4. A comparison of the isochore profiles of the telomeric 100 Mb human chromosome 1 (short arm) release hg17, obtained using the approach of Costantini et al.[15] (top), our segmentation approach (isoSegmenter), isoFinder, isoBase, and isoPlotter. Pre-calculated data from isoBase (the consensus approach) were downloaded from hg17 isoBase mirror. For isoFinder and isoPlotter+, isochores were calculated with the default parameters suggested by the authors directly on hg17 sequences. Results were converted and used to produce Figure 4. For the color code, see the legend of Figure 1.

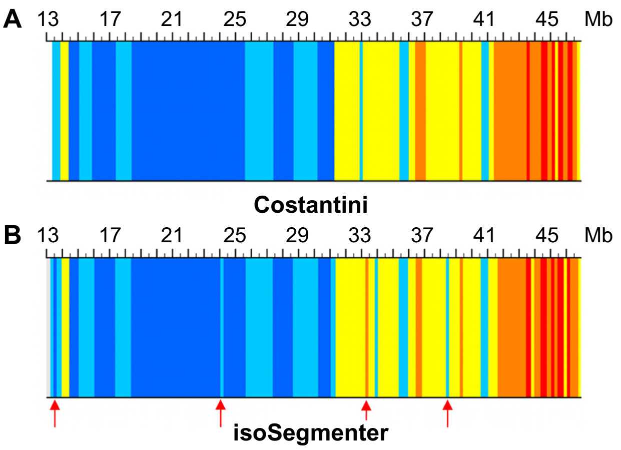

Figure 5. Comparison of the results of Costantini et al.[15] for the long arm of human chromosome 21 (release hg17) with the results obtained using isoSegmenter on this same release. Red arrows indicate some 200 kb isochores present in isoSegmenter and absent in Ref. 15.

all the isochores showing the dinucleotide frequencies typical of the isochore families under consideration. Clearly, these isochores belonging to different families were missed by the Bayesian analysis of isochores.

A comparison of isoSegmenter with the results of Costantini et al. IsoSegmenter is, by far, the computational approach that leads to results that are the closest ones to the manually curated data of Costantini et al.[15] Indeed, there is no stretch where isoFinder, or any other program, is closer to that of Costantini et al.[15] than isoSegmenter is. Among all other data, even the nearest ones, those of isoBase gave 7- to –8-fold higher and lower values for isochore sizes and numbers, respectively (see Supplementary Table 2). However, isoSegmenter results show some differences when compared with those of Ref. 15. The main source of such differences is the different way of analyzing isochore boundaries. Indeed, isoSegmenter uses the strict isochore boundaries presented in Supplementary Table 1, instead of accepting boundaries with a small compositional range (as done in Ref. 15), which requires, however, a subjective choice. Moreover, a number of 200 kb isochores were merged with flanking isochores in Ref. 15, but not in the case of isoSegmenter. These are the major reasons for the higher number of isochores, 4,107, detected by isoSegmenter vs. the smaller number, 3,254, of Costantini et al.[15] As a consequence, the mean isochore size estimated by isoSegmenter, 715 kb, is lower than 903 kb of Ref. 15. A comparison of the two approaches is depicted in Figure 5, which shows the significantly larger number of 200 kb isochores detected by isoSegmenter.

These differences also led to differences in the estimates of relative amounts of isochores belonging to different families. Indeed, as shown in Figure 6 and Supplementary Table 1, L1 and L2 isochores are overestimated, whereas H1 isochores are underestimated. This result may be explained by the increasing level of compositional heterogeneity of isochore families characterized by increasing GC levels (Supplementary Table 1).

In conclusion, we have developed a computational method for genome segmentation into isochores, which is fast and flexible, a necessity at a time when the sequences of so many genomes are available. This method allows a quick view of large-scale features of human genomes, as well as of any other genome for which assembled sequences are available. It should be stressed that the method is now undergoing a further elaboration in order to move from the present fixed isochore borders to flexible ones that would much better reflect reality (avoiding, however, the subjective decisions of Ref. 15).

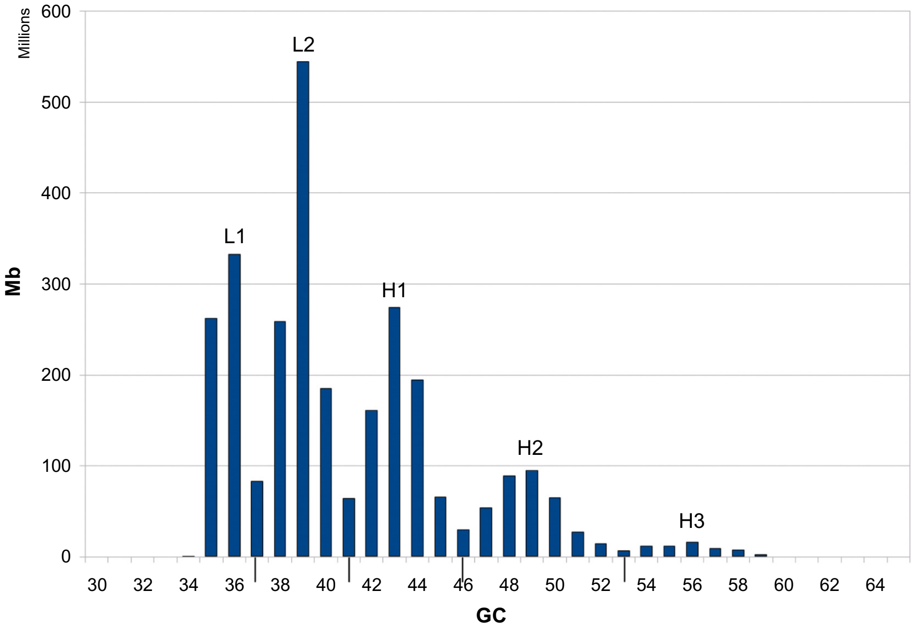

Figure 6. Histogram of isochore families in bins of 1% GC (from release hg38; Supplementary Table 1).

The compositional approach of the present automatic version opens up a vast array of applications that concern the large variations that exist not only in the genomes of different species but also in the genomes of the same species. It is well known that large-scale rearrangements, insertions/deletions, and translocations are frequent events, especially in GC-rich regions of the genome.[44] Needless to say, such changes essentially occur in noncoding sequences that not only represent the vast majority of the vertebrate genomes (98.5% in the human genome) but are also endowed with functional (regulatory) and structural properties.

General Conclusions

The general conclusions of this article concern (1) a critical assessment of existing genome segmenting approaches; this led to stressing that isoPlotter relies on very small DNA segments (25 kb on an average) that do not reflect the isochore structure of the genome and that do not have any biological relevance; (2) the development of a fully automatic program, isoSegmenter, that can segment the human genome (as well as other genomes) into isochores; this program supersedes existing programs in providing results that are closer to the compositional structure of the genome.

Acknowledgment

We thank our colleague Oliver Clay for very extensive and useful discussions, criticisms, and advice.

Author Contributions

Conceived and designed the experiments: GB, PC. Analyzed the data: GB, PC. Wrote the first draft of the manuscript: GB. Contributed to the writing of the manuscript: GB, PC, LM. Agree with manuscript results and conclusions: GB, PC, LM. Jointly developed the structure and arguments for the paper: GB. PC, LM. Made critical revisions and approved final version: GB, PC, LM. All authors reviewed and approved of the final manuscript.

Supplementary Material

Supplementary Table S1. Isochore families in the human genome (a).

Note: (a) The first four columns are from data of Costantini et al.[15]

Supplementary Table S2. Estimates of isochore numbers and sizes in the human genome.

Supplementary Figure S1. Scheme of the fractionation of complexes of DNA with sequence specific ligands. Binding of ligand molecules (red boxes) on DNA molecules depends upon the frequency of binding sites (oligonucleotides; blue boxes). Two DNA fragments are represented, which are characterized by different frequencies of such sites (modified from ref. 10).

Supplementary Figure S2A. The histogram shows the isochores from the human genome as pooled in bins of 1% GC. The Gaussian profile shows the distribution of isochore families that are represented by different colors as in Fig. 1. Gene densities and all other properties of the isochore families define two genome spaces (separated by a vertical broken red line): the genome desert and the genome core (modified from Ref. 15).

Supplementary Figure S2B. Smoothed contour plot of the gene landscape produced by plotting GC2 vs. GC3 (the GC levels of second and third codon positions, respectively) of 10,218 curated human genes. The GC3-richest cluster of genes is bimodal, even if this is only faintly visible in the figure. The vertical broken red line crosses the wide gap between the genes belonging to the L1, L2, H1 and those belonging to the H2, H3 isochore families (modified from Ref. 20).

Supplementary Figure S3. Isochores of chromosome 21 (release hg38) as obtained using 50-Kb to 400-Kb nonoverlapping windows.

REFERENCES

1. Corneo G, Ginelli E, Soave C, Bernardi G. Isolation and characterization of mouse and guinea pig satellite deoxyribonucleic acids. *Biochemistry*. 1968;7:4373–9.
2. Cortadas J, Macaya G, Bernardi G. An analysis of the bovine genome by density gradient centrifugation: fractionation in Cs_2SO_4/3,6-bis (acetato-mercurimethyl) dioxane density gradient. *Eur J Biochem*. 1977;76:13–9.
3. Filipski J, Thiery JP, Bernardi G. An analysis of the bovine genome by Cs_2SO_4/Ag^+ density gradient centrifugation. *J Mol Biol*. 1973;80:177–97.
4. Devillers-Thiery A. *Utilisation des endonucleases dansl'etude des sequences des ADN* [thesis]. Paris: Université Paris VII; 1974. [Reviewed in Bernardi, 2004].
5. Thiery JP, Macaya G, Bernardi G. An analysis of eukaryotic genomes by density gradient centrifugation. *J Mol Biol*. 1976;108:219–35.
6. Macaya G, Thiery JP, Bernardi G. An approach to the organization of eukaryotic genomes at a macromolecular level. *J Mol Biol*. 1976;108:237–54.
7. Cuny G, Soriano P, Macaya G, Bernardi H. The major components of the mouse and human genomes: preparation, basic properties and compositional heterogeneity. *Eur J Biochem*. 1981;111:227–33.
8. Costantini M, Bernardi G. Replication timing, chromosomal bands and isochores. *Proc Natl Acad Sci U S A*. 2008;105:3433–7.
9. Arhondakis S, Auletta F, Bernardi G. Isochores and the regulation of gene expression in the human genome. *Genome Biol Evol*. 2011;3:1080–9.
10. Bernardi *G. Structural and Evolutionary Genomics: Natural Selection in Genome Evolution*. Amsterdam: Elsevier; 2004. [Reprinted in 2005].
11. Bernardi G. Misunderstandings about isochores. *Gene*. 2001;276:3–13.
12. Saccone S, Pavliček A, Federico C, Pačes J, Bernardi G. Genes, isochores and bands in human chromosomes 21 and 22. *Chromosome Res*. 2001;9:533–9.
13. Pavliček A, Pačes J, Clay O, Bernardi G. A compact view of isochores in the draft human genome sequence. *FEBS Lett*. 2002;511:165–9.
14. Paces J, Zíka R, Paces V, Pavlícek A, Clay O, Bernardi G. Representing GC variation along eukaryotic chromosomes. *Gene*. 2004;333:135–41.
15. Costantini M, Clay O, Auletta F, Bernardi G. An isochore map of human chromosomes. *Genome Res*. 2006;16:536–41.
16. Lander ES, Linton LM, Birren B, et al. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*. 2001;409:860–921.
17. Kent WJ, Sugnet CW, Furey TS, et al. The human genome browser at UCSC. *Genome Res*. 2002;12:996–1006.
18. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*. 2004;431:931–45.
19. Bernardi G, Olofsson B, Filipski J, et al. The mosaic genome of warm-blooded vertebrates. *Science*. 1985;228:953–8.
20. Cruveiller S, Jabbari K, Clay O, Bernardi G. Compositional gene landscapes in vertebrates. *Genome Res*. 2004;14:886–92.
21. Berná L, Chaurasia A, Angelini C, Federico C, Saccone S, D'Onofrio G. The footprint of metabolism in the organization of mammalian genomes. *BMC Genomics*. 2012;13:174.
22. Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol*. 2010;28:817–25.
23. Frenkel ZM, Bettecken T, Trifonov EN. Nucleosome DNA sequence structure of isochores. *BMC Genomics*. 2011;12:203.

24. Costantini M, Di Filippo M, Auletta F, Bernardi G. Isochore pattern and gene distribution in the chicken genome. *Gene*. 2007;400:9–15.
25. Costantini M, Auletta F, Bernardi G. Isochore patterns and gene distributions in fish genomes. *Genomics*. 2007;90:364–71.
26. Costantini M, Cammarano R, Bernardi G. The evolution of isochore patterns in vertebrate genomes. *BMC Genomics*. 2009;10:146.
27. Cammarano R, Costantini M, Bernardi G. The isochore patterns of invertebrate genomes. *BMC Genomics*. 2009;10:538.
28. Zoubak S, Clay O, Bernardi G. The gene distribution of the human genome. *Gene*. 1996;174:95–102.
29. van Rossum G. *Python tutorial, Technical Report CS-R9526*. Amsterdam: Centrum voor Wiskunde en Informatica (CWI); 1995.
30. Oliver JL, Bernaola-Galván P, Carpena P, Román-Roldán R. Isochore chromosome maps of eukaryotic genomes. *Gene*. 2001;276:47–56.
31. Li W. Delineating relative homogeneous G+C domains in DNA sequences. *Gene*. 2001;276:57–72.
32. Ramensky VE, Makeev VJ, Roytberg MA, Tumanyan VG. Segmentation of long genomic sequences into domains with homogeneous composition with BASIO software. *Bioinformatics*. 2001;17:1065–6.
33. Zhang CT, Wang J, Zhang R. A novel method to calculate the G+C content of genomic DNA sequences. *J Biomol Struct Dyn*. 2001;19:333–41.
34. Oliver JL, Carpena P, Hackenberg M, Bernaola-Galvan P. IsoFinder: computational prediction of isochores in genome sequences. *Nucleic Acid Res*. 2004;32:W287–92.
35. Zhang CT, Gao F, Zhang R. Segmentation algorithm for DNA sequences. *Phys Rev E Stat Nonlin Soft Matter Phys*. 2005;72:041917.
36. Haiminen N, Mannila H. Discovering isochores by least-squares optimal segmentation. *Gene*. 2007;394:53–60.
37. Schmidt T, Frishman D. Assignment of isochores for all completely sequenced vertebrate genomes using a consensus. *Genome Biol*. 2008;9:R104.
38. Elhaik E, Graur D. IsoPlotter$^+$: a tool for studying the compositional architecture of genomes. *ISRN Bioinformatics*. 2013;2013:725434.
39. Elhaik E, Graur D. A comparative study and a phylogenetic exploration of the compositional architectures of mammalian nuclear genomes. *PLOS Comp Biol*. 2014;10:e1003925.
40. Elhaik E, Graur D, Josic K. Comparative testing of DNA segmentation algorithms using benchmarck simulations. *Mol Biol Evol*. 2010;27:1015–24.
41. Elhaik E, Graur D, Josic K, Giddy L. Identifying compositionally homogeneous and nonhomogeneous domains within the human genome using a novel segmentation algorithm. *Nucleic Acid Res*. 2010;38(15):e158.
42. Fearnhead P, Vasileiou D. Bayesian analysis of isochores. *J Am Stat Assoc*. 2009;104:132–41.
43. Fujita MK, Edwards SV, Ponting CP. The anolis lizard genome: an amniote genome without isochores. *Genome Biol Evol*. 2011;3:974–84.
44. Costantini M, Bernardi G. Mapping insertions, deletions and SNPs on Venter's chromosomes. *PLoS One*. 2009;4(6):e5972.