GC Composition of the Human Genome: In Search of Isochores

Netta Cohen,[1] *Tal Dagan,*†[1] *Lewi Stone,*† *and Dan Graur*‡

*School of Computing, University of Leeds, Leeds, United Kingdom; †Department of Zoology, George S. Wise Faculty of Life Science, Tel Aviv University, Ramat Aviv, Israel; and ‡Department of Biology and Biochemistry, University of Houston

The isochore theory, proposed nearly three decades ago, depicts the mammalian genome as a mosaic of long, fairly homogeneous genomic regions that are characterized by their guanine and cytosine (GC) content. The human genome, for instance, was claimed to consist of five distinct isochore families: L1, L2, H1, H2, and H3, with GC contents of <37%, 37%–42%, 42%–47%, 47%–52%, and >52%, respectively. In this paper, we address the question of the validity of the isochore theory through a rigorous sequence-based analysis of the human genome. Toward this end, we adopt a set of six attributes that are generally claimed to characterize isochores and statistically test their veracity against the available draft sequence of the complete human genome. By the selection criteria used in this study: distinctiveness, homogeneity, and minimal length of 300 kb, we identify 1,857 genomic segments that warrant the label "isochore." These putative isochores are nonuniformly scattered throughout the genome and cover about 41% of the human genome. We found that a four-family model of putative isochores is the most parsimonious multi-Gaussian model that can be fitted to the empirical data. These families, however, are GC poor, with mean GC contents of 35%, 38%, 41%, and 48% and do not resemble the five isochore families in the literature. Moreover, due to large overlaps among the families, it is impossible to classify genomic segments into isochore families reliably, according to compositional properties alone. These findings undermine the utility of the isochore theory and seem to indicate that the theory may have reached the limits of its usefulness as a description of genomic compositional structures.

Introduction

How are nucleotides distributed along genomes? Is there a functional significance to the organization of bases along noncoding DNA, and do compositional structures reflect fundamental laws underlying evolutionary processes? Nonuniformity of nucleotide composition within genomes from a variety of taxa ranging from phages to mammals was revealed several decades ago by thermal melting and gradient centrifugation (Inman 1966; Filipski, Thiery, and Bernardi 1973). Later, on the basis of findings concerning buoyant densities of melted DNA fragments, Bernardi and coworkers (Macaya, Thiery, and Bernardi 1976; Thiery, Macaya, and Bernardi 1976; Bernardi et al. 1985) proposed a theory for the structure of the genomes of warm-blooded vertebrates (for a comprehensive review, see Bernardi 2000). The theory has since been known as the isochore theory (Cuny et al. 1981).

Isochores were defined as long genomic segments that are fairly homogeneous in their guanine and cytosine (GC) composition. The human genome was described as a mosaic of isochores of alternating low and high GC contents (defined as the fraction of G and C nucleotides along a sequence). Human isochores have since been classified into five families, L1, L2, H1, H2, and H3, whose corresponding ranges of GC contents were said to be <37%, 37%–42%, 42%–47%, 47%–52%, and >52%, respectively (Bernardi 2000). These isochore families are customarily modeled as Gaussian distributions of the GC contents of their member isochores. Figure 1 illustrates the GC-content distributions of isochore families (Bernardi 2001). The combined distribution of the five families (fig. 1, dashed line) characterizes the GC content of the entire genome.

Since the proposition of the isochore theory, many studies have revealed profound effects of GC content on various genomic properties. For example, the distributions of genes and repetitive elements were found to be associated with particular GC contents. Thus, small- and medium-sized genes were found to be more abundant in GC-rich regions of the human genome, whereas long genes (typically, genes with long introns) were found to be scarce in GC-rich regions (Duret, Mouchiroud, and Gautier 1995; Zoubak, Clay, and Bernardi 1996; Lander et al. 2001). The distributions of *Alu*s, *SINE*s, *LINE*s, and other remnants of transposition and retroposition were also found to be correlated with GC content (Smith and Higgs 1999; Lander et al. 2001). Similar correlations were claimed to characterize integration sites of retroviruses (Salinas et al. 1987; Zoubak et al. 1994) as well as patterns of CpG methylation (Caccio et al. 1997). In human and mouse genomes, GC content was found to be associated with particular chromosomal bands. For example, GC-rich isochores were found to coincide with the light reverse bands on metaphase chromosomes (Saccone et al. 1997, 2001; Lander et al. 2001).

The publication of the draft human genome (Lander et al. 2001) and of several completely sequenced chromosomes (Dunham et al. 1999; Hattori et al. 2000; Deloukas et al. 2001; Heilig et al. 2003; Hillier et al. 2003; Mungall et al. 2003), as well as the further accumulation of completely sequenced genomes from other species, has brought to the surface many objections to the isochore theory (e.g., Nekrutenko and Li 2000; Häring and Kypr 2001). Some of the objections concerned definitions, while others involved the methodology of inferring global compositional structures from buoyant-density data. The most telling criticism, however, touched upon the very existence of isochores and claimed that the "strict notion of isochores as compositionally homogeneous" could be ruled out, and hence, isochores do not merit the prefix "iso" (Lander et al. 2001). Notwithstanding these criticisms, it is widely accepted that the human genome contains large regions of distinctive GC

[1] These authors contributed equally to this work.

Key words: isochores, GC content, human genome, Jensen-Shannon entropic divergence, genome organization.

E-mail: dgraur@uh.edu.

Mol. Biol. Evol. 22(5):1260–1272. 2005
doi:10.1093/molbev/msi115
Advance Access publication February 23, 2005

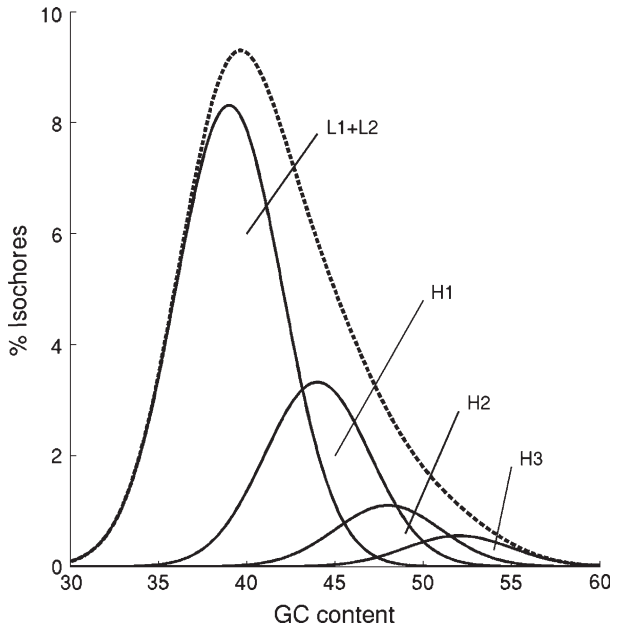

FIG. 1.—Illustration of the traditional five-Gaussian description of isochore families in the human genome. The Gaussians corresponding to the two GC-poor families (L1 and L2) are customarily merged into a single Gaussian. The superposition of the four remaining Gaussians is plotted in a dashed line. Modified from Pavlicek et al. (2002).

content (Lander et al. 2001). Thus, whether isochores exist, whether isochore families exist, and whether isochores can be assigned unambiguously to such families are all questions that deserve to be addressed in a rigorous manner.

In this study, we are interested in two issues: (1) do isochores exist and, if so, (2) are isochores a useful (or practical) concept. Stated differently, we ask whether it is possible to provide a rigorous definition of isochores such that the majority of the human genome can be classified as isochoric, and, if so, to what extent (i.e., with what confidence) can each isochore be classified into a particular isochore family. An important premise of this study is that all of our analyses and conclusions are based solely on compositional properties of the sequences studied. In other words, we intentionally restrict our attention to compositional properties of the genome and put aside the large body of literature that correlates GC content with evolutionary or functional properties (e.g., repeats and genes). In so doing, we are avoiding any circular reasoning that might arise from the definition of isochores.

Definitions of isochores abound, some "hard" (Clay et al. 2001) and others "soft" (Bernardi 2001). In this paper, we spell out six attributes that have been claimed to characterize isochores and their classification into a small number of isochore families; we then test these attributes against human genomic data.

Isochore Attributes

The following isochore attributes will be tested:

(A1) Characteristic GC content: An isochore is a DNA segment that has a characteristic GC content that differs significantly from the GC content of adjacent isochores (Bernardi 2001; Li et al. 2002; Oliver et al. 2002). Segments with characteristic GC contents were obtained by using a specially suited DNA-segmentation algorithm (see *Methods*).

(A2) Homogeneity: An isochore is more homogeneous in its composition than the chromosome on which it resides (Bernardi 2001). In the literature, various measures and criteria for relative homogeneity are used (Nekrutenko and Li 2000; Li 2002; Li et al. 2003). Here we compare GC-content variability along segments with that along chromosomes (see *Methods*). By referring back to the chromosome, we are treating the organization of genomic composition as a process occurring within the context of the genomic environment.

(A3) Minimum length: The length of an isochore typically exceeds a certain cutoff value. In the literature, the most commonly mentioned value is 300 kb (Macaya, Thiery, and Bernardi 1976; Cuny et al. 1981; Bernardi et al. 1985; Bernardi 2000; Clay et al. 2001; Pavlicek et al. 2002). Here we adopt this cutoff value of 300 kb and compare the results with those obtained with other possible cutoffs (50, 100, 200, 300, and 500 kb and no cutoff).

(A4) Genome coverage: The overwhelming majority of the human genome consists of segments obeying the criteria set out in A1–A3. Nonisochoric DNA takes up only a small fraction of the genome (Pavlicek et al. 2002).

(A5) Isochore families: The human genome comprises five isochore families, each described by a particular Gaussian distribution of GC content (Bernardi et al. 1985; Bernardi 2001; Pavlicek et al. 2002).

(A6) Isochore assignment into families: It is possible to classify each isochore into its isochore family based solely on its compositional properties (Bernardi 2001; Pavlicek et al. 2002).

In what follows, we apply attributes A1–A3 as selection criteria for identifying putative isochores and then determine whether and to what extent attributes A4–A6 hold for the collections of putative isochores so obtained.

Methods

Data

Human genomic sequences were obtained from the October 2003 version of the National Center for Biotechnology Information database (http://www.ncbi.nlm.nih.gov/). The contigs of each chromosome were concatenated together in the proper order to form long sequences.

Partition of Genomic Sequences into Segments that Have Characteristic GC Contents and Differ Significantly from the GC Contents of Adjacent Segments

Several methods have been proposed in the literature for identifying segments with characteristic GC content (Wen and Zhang 2003; Zhang and Zhang 2003). In this study, we partitioned the genomic sequences into segments by the binary recursive segmentation procedure, D_{JS}, proposed

by Bernaola-Galván, Róman-Roldán, and Oliver (1996). In this procedure, the chromosomes are recursively segmented by maximizing the difference in GC content between adjacent subsequences. The process of segmentation is terminated when the difference in GC content between two neighboring segments is no longer statistically significant.

Briefly, a chromosome of length L, GC content F_{GC}, and AT content $F_{AT} = 1 - F_{GC}$ is divided into two contiguous segments ($i = 1, 2$) of length l_i, GC content f^i_{GC}, and AT content f^i_{AT}. These segments are chosen to maximize the Jensen-Shannon entropic divergence measure, D_{JS}, defined as the difference between the overall Shannon entropy H^{tot} and the sum of segment Shannon entropies H^i:

$$D_{JS} \equiv \max\left[H^{tot} - \sum \frac{l_i}{L} H^i\right],$$

where $H^i = -f^i_{GC}\log_2 f^i_{GC} - f^i_{AT}\log_2 f^i_{AT}$ and $H^{tot} = -F_{GC}\log_2 F_{GC} - F_{AT}\log_2 F_{AT}$. The segmentation is then repeated recursively for each segment until a halting criterion $D_{JS} \geq D_C$ is met for all segments. In order to estimate D_C, 100,000 random sequences, each 1 Mb long, were generated from a uniform distribution. Each of these 100,000 sequences was segmented into two at a random point, and the D_{JS} value for each segmentation was calculated. We followed the procedure of Bernaola-Galván, Róman-Roldán, and Oliver (1996) and chose a D_C value corresponding to the lower 5% of the cumulative D_{JS} distribution. This method of selecting the halting parameter was chosen both for consistency with the existing literature and as a conservative choice, generating a large number of isochore-like segments longer than 300 kb. For convenience only, the algorithm was implemented on coarse-grained sequences of 32 base pair (bp) nonoverlapping windows.

Homogeneity Test

To test the compositional homogeneity of a segment versus that of the chromosome on which it resides, we used the F-test (Zar 1999) to compare the variance in GC content between the two. To this end, each chromosome was divided into 2,048-bp-long nonoverlapping windows (see below), and the GC-content values for each window were calculated for the entire chromosome and for the segment in question. Because the F-test assumes normal distributions, we applied the arcsine-root transformation to the GC-content values of the windows within each segment (and chromosome) before calculating the variance. A one-tailed F-test with a null hypothesis $H_0 : \sigma^2_{segment} \geq \sigma^2_{chromosome}$ and an alternative hypothesis $H_1 : \sigma^2_{segment} < \sigma^2_{chromosome}$ was applied with $n_1 - 1$ and $n_2 - 1$ degrees of freedom, where n_1 and n_2 are the numbers of windows in the segment and in the corresponding chromosome, respectively. If the variance over a segment was found to be significantly smaller ($P < 0.05$) than that of the corresponding chromosome, then the segment was deemed to be more homogeneous than the chromosome. Segments that are shorter than 10 kb, i.e., contain less than 5 windows, were excluded from the analysis. The choice of window size above was tested for a wide range of values (1, 2, 4, and 8 kb). For 2- to 8-kb windows, results were found to be robust to the choice of window size for sufficiently long segments (at least 5 windows long). For smaller window sizes (1 kb long), borderline-homogeneous segments sometimes failed the homogeneity test. The choice of 2,048-kb windows thus ensures robust results with the lowest segment length cutoff (10 kb).

Fitting a Collection of Gaussian Distributions to an Empirical Distribution

Isochore families are customarily defined as Gaussian distributions of the GC contents of their member isochores. Let N be the number of families. Each of the N distributions is fully described by three parameters: mean GC content, standard deviation (SD), and relative weight, i.e., the portion of all isochores that belong to this family. By normalizing the weights of the Gaussians so that their sum is 1, the number of independent parameters that are required to describe the full N-Gaussian distributions of isochore GC content is $3N - 1$.

We used a genetic algorithm (Price and Storn 1997) to obtain a maximum-likelihood fit of a fixed number of Gaussians (N) to the GC-content distribution of segments, putative isochores, or fixed-size windows in the human genome. Once an optimal fit was found for a given N, a one-sided goodness-of-fit χ^2 test with $\alpha = 0.05$ was applied to determine the statistical significance of the fit. Multi-Gaussian fits to the data were tested for models of two to eight Gaussians for seven different datasets: (i) putative isochores (homogeneous, $\geq$300 kb, N = 1,857), (ii) alternative putative isochores longer than 100 kb (homogeneous, $\geq$100 kb, N = 6,383), (iii) alternative putative isochores longer than 50 kb (homogeneous, $\geq$50 kb, N = 11,390), (iv) all homogeneous segments (N = 23,572), (v) all segments (N = 47,561), (vi) fixed-length windows 2^{16} bp long (65,536 bp, N = 43,664), and (vii) fixed-length windows 2^{19} bp long (524,288 bp, N = 5,458). The choice of window sizes here was aimed at roughly reproducing fragment sizes in gradient centrifugation experiments (dataset vi) and typical isochore sizes (dataset vii).

Isochore Classification and Classification Errors

The issue of isochore classification is nontrivial. Simply put, the existence of two (or more) statistically distinct distributions does not, in general, imply that an individual data point may reliably be classified into its corresponding group. The extent to which classification is possible depends on the degree of overlap between distributions. In particular, for largely overlapping distributions (as in a five-Gaussian fit to the GC-content distribution of putative isochores), an individual data point (here, an individual segment) might have come from two or more of these distributions (here, isochore families). Thus, for example, for five isochore families with five distinct sets of means, variances, and weights, we must still check the error associated with the classification of an individual segment into a particular family. Naturally, the entire discussion of classification assumes, for the sake of argument, that the distributions of isochore families are known.

Consider an N-Gaussian model of isochore families. Suppose we would like to classify a particular segment with a GC content x. At this value, each Gaussian is associated

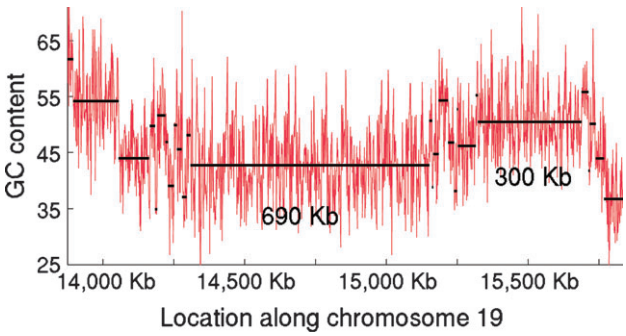

FIG. 2.—An illustration of the spatial distribution of GC content of nonoverlapping 1,024-bp windows along a fragment, approximately 1.4 Mb in length, from chromosome 19 (red). The segmentation algorithm yielded 27 segments for this fragment (superimposed in black), including two segments that are longer than 300 kb. Of these, the longer segment (690 kb) is relatively GC poor (42%), and the shorter segment (300 kb) is relatively GC rich (50%).

with a certain Gaussian amplitude. Clearly, the segment in question should be classified into whichever family has the greatest Gaussian amplitude at the GC-content value of x. The associated probability $p(x)$, i.e., the probability that this particular segment indeed belongs to that family, is the corresponding Gaussian amplitude, normalized by the sum of amplitudes of all N families at this value x (so that the sum of probabilities is 1). The associated classification error is then $\varepsilon(x) = 1 - p(x)$. This classification protocol is obviously optimal in the sense that it minimizes the classification error. Having classified all segments in this way, we define the mean classification error, $\bar{\varepsilon}$ as the average error over the entire range of GC-content values, weighted by the corresponding frequency of occurrence.

Results

An example of the D_{JS} segmentation results is illustrated in figure 2. The figure presents a 1.4-Mb-long DNA sequence from chromosome 19 for which the D_{JS} protocol yielded 27 segments. Most of the segments are relatively short and clustered together within three compact regions, characterized by large fluctuations in local GC content. The remainder of the sequence consists of two long (300 and 690 kb) and comparatively homogeneous segments, which may therefore be considered putative isochores. The 690-kb-long segment is relatively GC poor (42 ± 6%, i.e., borderline between L2 and H1), whereas the 300-kb-long segment is GC rich (50 ± 5%, with a GC content corresponding to H2).

Segment Numbers and Lengths

Application of the D_{JS} segmentation procedure to the human genome yielded 47,561 segments (segmentation results are given in the Supplementary Material; contig and segment information are given in table D1; segmentation statistics by chromosome are given in table S1). The number of segments per chromosome ranges from 592 (chromosome Y) to 4,136 (chromosome 2) and was found to be positively correlated with chromosome length ($r = 0.85$). Segment density, defined as the number of segments per megabase, ranges from 10 segments/Mb (chromosome 4) to 37 segments/Mb (chromosome 22) and was found to correlate negatively with chromosome length ($r = -0.60$). The mean segment length ranges from 27 kb in chromosome 22 to 96 kb in chromosome 4 and is positively correlated with chromosome length ($r = 0.59$). Furthermore, segment density and mean segment length are negatively correlated ($r = -0.94$).

Figure 3*a* shows that the distribution of segment lengths (plotted on a log-log scale) follows a heavy-tail distribution with a power-law decay exponent of −2.38

$$P \propto L^{-2.38}, \tag{1}$$

where P is the frequency of segments of length L. In other words, the segments do not have a characteristic length scale as might be expected from the isochore theory. Rather, there is an abundance of short segments and only a small number of long ones. Indeed, out of a total of 47,561

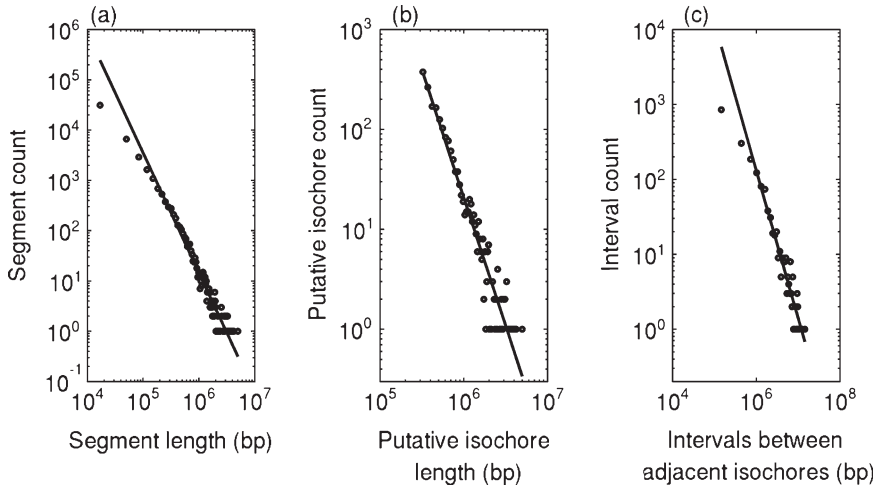

FIG. 3.—Length distributions of (*a*) D_{JS} segments, (*b*) putative isochores, and (*c*) intervals between adjacent putative isochores, plotted on a log-log scale (dots). The fitted regression lines (solid lines) indicate that the tails of the three distributions exhibit power-law decays with exponents (*a*) −2.38, (*b*) −2.55, and (*c*) −1.97.

segments in the complete human genome, approximately 17,000 (36%) are shorter than 10 kb and only a tiny proportion of segments (234 or ~1%) are longer than 1 Mb. The distribution of nucleotides in the human genome has been previously shown to exhibit power-law behavior over a more limited range of length scales (e.g., Peng et al. 1994). However, to the best of our knowledge, this is the first time that such power-law behavior is demonstrated quantitatively for the entire human genome.

Segment Homogeneity

To gauge segment homogeneity, we compared the variance in GC content within each segment to that of the chromosome on which it resides. Overall, 49.5% of segments longer than 10 kb (i.e., 23,572 segments) was found to be more homogeneous than the chromosomes on which they reside. These homogeneous segments span about 90% of the human genome.

Segment Length and Genome Coverage

The isochore model posits that only long segments (typically hundreds of thousands of bases or longer) deserve the label isochores. We handled this attribute by introducing a length cutoff, i.e., a minimal length required for a segment to be considered as a putative isochore. Note, however, that the power-law decay in segment-length frequency (fig. 3*a*) implies that different cutoff choices would strongly influence the proportion of long segments (fig. S1). In particular, of the 47,561 D_{JS} segments, only 1,863 segments (3.9%) are longer than the 300-kb value put forward by the proponents of the isochore theory (table 1).

Interestingly, the proportion of segments that was found to be homogeneous seems to depend on the length cutoff as well. The higher the length cutoff, the greater the proportion of homogeneous segments. However, we note that for a sufficiently high choice of cutoff, the homogeneity criterion loses significance. Thus, even for a relatively low length cutoff of 50 kb, approximately 95% of the segments is homogeneous. This proportion increases to almost 100% for a cutoff of 500 kb (table 1). Thus, in practical terms, any D_{JS} segment long enough to qualify as an isochore will almost certainly be found to satisfy the homogeneity criterion as well.

A further claim of the isochore theory is that the vast majority of the human genome is spanned by isochoric regions. We found that segments longer than 300 kb span only 41% of the human genome. Moreover, the percent coverage drops significantly for higher isochore cutoffs (fig. S1*b*). This is a direct result of a power-law length distribution with a decay exponent steeper than −2. While segments longer than 50 kb span 81% of the genome, segments longer than 500 kb span only 27% of the genome. Segments that are longer than 1 Mb cover merely 13% of the genome (table 1). As expected, the same trend is observed for homogeneous segments with different length cutoffs.

Table 1
Segment Statistics for the Complete Human Genome as a Function of Length Cutoff (l_c)

	Segments Longer than l_c		Homogeneous Segments Longer than l_c	
Length Cutoff (l_c)	Number (%)	Genome Coverage	Number (%)	Genome Coverage
50 kb	11,993 (25.2)	81	11,390 (23.9)	79
100 kb	6,520 (13.7)	68	6,383 (13.4)	67
200 kb	3,067 (6.5)	51	3,042 (6.4)	50
300 kb	1,863 (3.9)	41	1,857 (3.9)	41
500 kb	843 (1.7)	27	843 (1.7)	27
1 Mb	234 (0.5)	13	234 (0.5)	13

NOTE.—For each length cutoff, the table lists the number of segments that are longer than the length cutoff, their percentage from the total segments, and the corresponding genome coverage. The same data are also presented subject to an additional homogeneity constraint.

Putative Isochores

By the selection criteria used in this study, isochores should be compositionally distinct (A1), homogeneous in GC content (A2), and longer than 300 kb (A3). All D_{JS} segments obeying these criteria will, henceforth, be referred to as "putative isochores." The number of putative isochores in the human genome is 1,857. In other words, only 3.9% of the D_{JS} segments satisfies the length and homogeneity criteria (table 1). To determine the significance of this figure, we segmented the genome at random locations while keeping constant the number of segments and the segment-length distribution. By applying the same length and homogeneity criteria, we found that only 0.09% of these "control" segments qualified as putative isochores, spanning 5% of the genome. In other words, D_{JS} segmentation based on GC content yields significantly more putative isochores than random, composition-independent segmentation.

In what follows, some chromosome-by-chromosome statistics are presented, based on the data in table S2. The abundance of putative isochores ranges from 0.7% of all segments (chromosome 22) to 7.3% (chromosome 4). As expected, the number of putative isochores is correlated with chromosome length ($r = 0.95$). While the number of putative isochores is small, perhaps a more meaningful measure is their coverage of respective chromosomes, which varies between 10% in chromosome 22 and 58% in chromosome 4 (table S2) and correlates positively with mean segment length ($r = 0.85$). Similar to the distribution of D_{JS} segments, the distribution of putative isochores also follows a heavy-tail length distribution with a power-law decay exponent of −2.55 (fig. 3*b*).

Unlike the negative correlations seen above for all D_{JS} segments, the mean length of putative isochores correlates positively with their density along the chromosome ($r = 0.84$), suggesting that longer isochores tend to be less scattered than shorter ones. In fact, the distances between nearest-neighbor putative isochores range from 0 Mb (adjacent putative isochores) to 14.5 Mb. These nearest-neighbor intervals follow a heavy-tail distribution with a power-law tail asymptotically approaching that of the corresponding segment-length distribution (fig. 3c). Only 216 putative isochores (11.6%) have an abutting neighbor. The largest cluster of abutting putative isochores consists of five segments and is found on chromosome 6. The lengths of its constituent putative isochores are 469, 386, 569, 495,

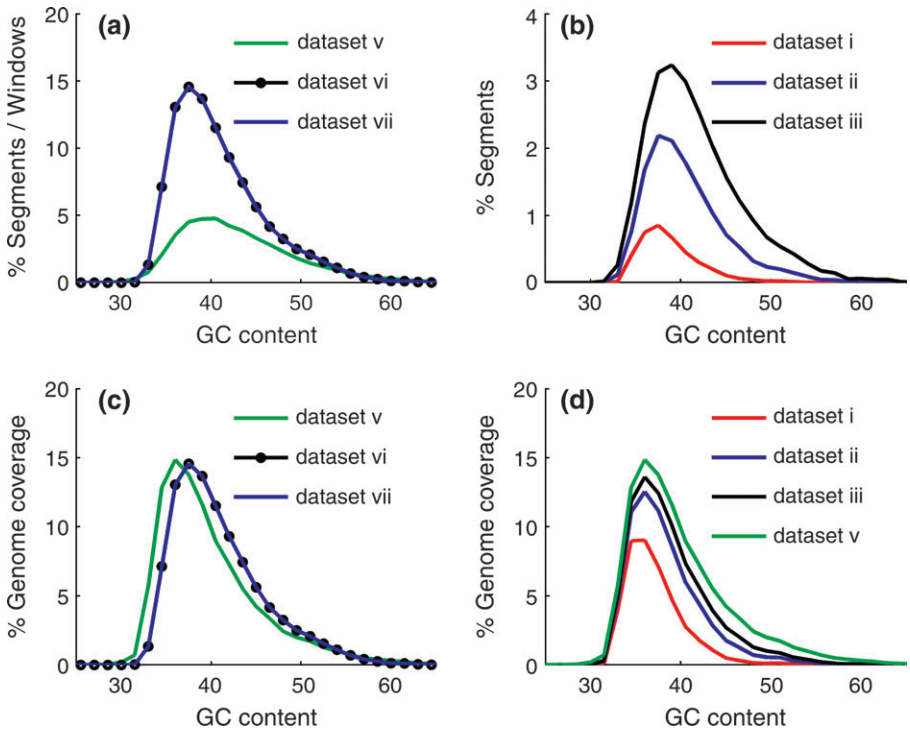

FIG. 4.—Probability density histograms as a function of GC content of the different datasets (see *Methods*). Subplots (*a*) and (*b*) show the relative fraction of segments or windows per unit GC. Datasets v, vi, and vii in subplot (*a*) span the entire genome. Histograms of windowed data are indistinguishable on this plot, whereas segments have a significantly broader GC-content distribution. Datasets i–iii (homogeneous segments with length cutoffs of 300, 100, and 50 kb, respectively) constitute only a small fraction of segments and are therefore shown separately in subplot (*b*). Lower length cutoffs correspond to broader and GC-richer distributions. Subplots (*c*) and (*d*) show genome-coverage distributions. Datasets spanning the entire genome are shown in subplot (*c*) and are effectively indistinguishable from one another. Subplot (*d*) shows successive length cutoffs and the corresponding subsets of the GC-content distribution they span. The higher the length cutoff, the smaller the subset. This effect is observable on the GC-rich tail of the distributions; in contrast, the GC-poor tail of the histograms (below 35% GC) is roughly the same for all four datasets. For purposes of comparison, dataset v is included in both subplots (*c*) and (*d*).

and 622 kb. All of these putative isochores are relatively GC poor with mean GC contents of 35%, 37%, 39%, 37%, and 35%, respectively.

GC Content, Genome Coverage, and Length Correlations in Segments and Fixed-Length Windows

A comparison of GC contents of segments, putative isochores, and fixed-length windows is shown in figure 4. Figure 4*a* shows that the distribution of fixed-length windows of lengths 65 and 525 kb are virtually indistinguishable. This has been noted in the literature (Pavlicek et al. 2002). However, these distributions are markedly different (both in the mean and in the width) from the distribution of D_{JS} segments. Still, as might be expected, the probability distributions of genome coverage by fixed-length windows and genome coverage by segments are very similar (fig. 4*c*).

Putative isochores and their alternatives (with different length cutoffs) constitute only a minute fraction of the segments and are presented separately in figure 4*b*. The figure demonstrates that increasingly longer putative isochores constitute narrower and narrower subsets of all D_{JS} segments with lower and lower GC contents. Figure 4*d* presents the corresponding genome coverage by putative isochores and alternatives with different length cutoffs. We emphasize that, for all chromosomes except 19 and 22, the mean GC content of putative isochores is significantly lower ($P < 0.05$) than the GC content of the corresponding chromosome. (In chromosomes 19 and 22, the GC content of putative isochores equals that of the chromosome; data not shown.)

Although GC content is only weakly correlated with segment length ($r = -0.19$), this is a consequence of the limited range of GC contents relative to the many orders-of-magnitude range of segment lengths. A more instructive correlation measure is therefore between the GC content and the logarithm of the segment length ($r = -0.33$). Therefore, the mean GC content of segments longer than 300 kb is lower by about 6% than the mean GC content of all D_{JS} segments (fig. 4*a*).

Classification of Segments and Putative Isochores into Families

The isochore theory defines five isochore families within the human genome, each described by the GC-content range of its constituent isochores. Thus, L1 and L2 are GC-poor families, and H1, H2, and H3 are GC-rich families. While each family is described by a symmetric (typically Gaussian) distribution of GC contents (fig. 1), an approximate range was provided for purposes of isochore classification (see *Introduction* for the GC-content ranges for each family). We initially classified all segments into the five families according to these specified ranges. GC-poor families (L1, L2) account for ~40% of the segments, while the GC-rich families comprise the remaining ~60%

Table 2
Distribution of Segments (Dataset v), Putative Isochores (Dataset i), and Fixed-Length Windows (Datasets vi and vii) Within the Five Traditional Isochore Families

	L1 (GC < 37%)	L2 (37% < GC < 42%)	H1 (42% < GC < 47%)	H2 (47% < GC < 52%)	H3 (GC > 53%)
Segments					
Percent (%)	17.22 (41.42)	24.20	21.54	15.32	21.72
Genome coverage	26.85 (67.40)	40.61	19.00	8.04	5.50
Putative isochores					
Percent (%)	33.49 (84.76)	51.27	12.44	2.37	0.43
Weighted by length	43.10 (88.57)	45.47	9.57	1.48	0.37
Genome coverage	17.60 (26.20)	18.60	3.92	0.60	0.15
65-kb windows (%)	24.01 (66.20)	42.19	21.06	8.77	3.97
525-kb windows (%)	24.02 (66.20)	42.18	21.07	8.76	3.98

NOTE.—Classification by GC ranges was according to Pavlicek et al. (2002). The distribution statistics are given as percentages of coverage. Percent coverage is obtained by weighting each segment by length. Note that for putative isochores, this is distinct from genome coverage (the latter also accounting for nonisochoric regions). The entries in parentheses sum over the two GC-poor families (L1 and L2).

of the segments. However, when these proportions are calculated by genome coverage (i.e., weighted by the segment lengths), the combined share of L1 and L2 increases to almost 70%, and the remaining 30% is of the H1, H2, and H3 families (table 2).

When considering only putative isochores (rather than all D_{JS} segments), significantly different results are obtained. Classification of putative isochores into the five isochore families according to their GC content yields a large proportion of L1 and L2 GC-poor isochores (~85%). The corresponding genome coverage by these GC-poor segments is almost 90% (table 2).

Gaussian Description of Segments versus Fixed-Length Windows

According to the isochore theory, the distribution of isochore GC contents is a collection of five so-called isochore families (attribute A5). Each family is described by a Gaussian distribution, and each isochore can be classified into one of these families (attribute A6). The distribution of GC content for the collection of all isochores is described as a superposition or weighted sum of these five Gaussians (Bernardi et al. 1985; Pavlicek et al. 2002). Our analysis of the Gaussian-family description of isochore GC contents is given in the remainder of the *Results*.

First, we tested whether putative isochores could be described by multi-Gaussian models. We further tested whether segments with varying length cutoffs and fixed-length windows of comparable length could also be described by such models. Toward this end, a maximum-likelihood fitting algorithm was implemented and χ^2 tested (see *Methods*). Model results for all possible fits are listed in table S3.

We found considerable variability between GC-content distributions of the different datasets, which was also reflected in the respective multi-Gaussian models. In particular, results varied widely for datasets i–v (putative and alternative putative isochores). Even the minimal number of Gaussians needed to obtain a statistical fit varied between datasets. For both putative isochores (dataset i) and alternative putative isochores with a length cutoff of 50 kb (dataset iii), significant fits were found for models of four or more Gaussians. For dataset iv (no length cutoff), only models of five Gaussians or more were obtained. However, for an intermediate length cutoff (dataset ii), all models (two to eight Gaussians) were rejected. In addition, no model was found for the dataset derived by the D_{JS} segmentation of the entire genome (dataset v).

For fixed window sizes (datasets vi and vii), the GC-content distributions were found to be virtually indistinguishable; accordingly, the fits obtained for these distributions were equivalent. Nonetheless (due to the data size dependence of the statistical test used), three Gaussians were sufficient to fit the distribution of large-sized windows (dataset vii), whereas at least four Gaussians were required for the smaller windows of data set vi (see *Discussion*). We note that the multi-Gaussian models obtained for these datasets once again differed considerably from fits found for datasets i–v.

In what follows, we restrict our discussion to statistically viable fits (datasets i, iii, iv, vi, and vii). Selected fits (for models of five Gaussians or less) are presented in figure 5. One notable observation that is common to the datasets for which both four- and five-Gaussian models were found is the striking similarity between the four- and five-Gaussian models (datasets i, iii, and vii, see Table S3 in the Supplementary Material). All five-Gaussian models appear to consist of four dominating Gaussian families and a minor one. In all corresponding data, the fifth Gaussian has negligible weight and/or width and is therefore hardly noticeable in the figures.

Models for Putative Isochores

Putative isochores have a broad distribution of GC content, ranging over 33.0%–54.8% (mean GC content, 38.8%; see fig. 4*b*). This population of segments can be modeled by multi-Gaussian distributions containing four to eight Gaussians. In the four-Gaussian model (fig. 5*a*), Gaussian widths (measured by 1 SD) increase with GC content from 0.8% at low GC content to a very broad 3.1% at the high GC-content tail of the distribution. In comparison, the SD of the total population of putative isochores is close to 3.3%. The five-Gaussian model is by and large similar, with an even wider Gaussian in the GC-rich tail (3.6% SD) and with a small (low weight) and particularly narrow fifth Gaussian centered at 39.3% (close to the mean of the parent

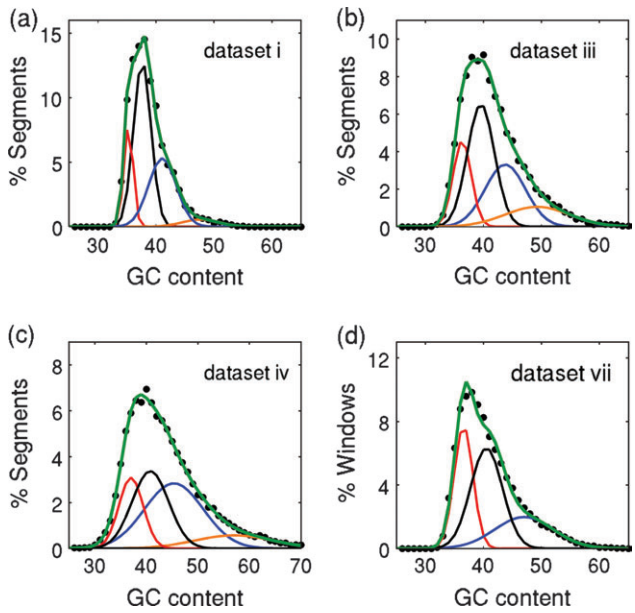

FIG. 5.—Selected multi-Gaussian fit results for the indicated data sets, superimposed on probability density histograms of the data as a function of GC content. The fits shown are the most parsimonious statistically valid multi-Gaussian models. Each subplot shows the Gaussian components (red, blue, black, and orange lines), the multi-Gaussian fit to the data (green line), and the underlying histogram of the raw data (dots). The number of bins is set to the square root of the number of data points. Subplots (*a*) and (*b*) show four-Gaussian fits. Subplot (*c*) shows a five-Gaussian fit; the smallest Gaussian is barely visible on the low GC-content tail of the distribution. Subplot (*d*) shows a three-Gaussian fit to the windowed data.

distribution). This fifth Gaussian is barely noticeable by eye and is fully contained within other much larger Gaussians. Hence, we note that all the Gaussians in this model are markedly different from those specified in Bernardi (2001). In particular, none of the Gaussian ranges in this model correspond to the H2 (47%–52%) or H3 (>52%) isochore families (fig. 6*a*).

As the number of Gaussians increases in successive models, the distributions tend to consist of narrower Gaussians, with some Gaussians fully contained within other larger Gaussians.

Robustness of Models

One immediate result that can be drawn by inspecting and comparing the model fits is the lack of robustness of models to variations in isochore definitions. For instance, one might ask whether five-Gaussian models obtained from segments with a length cutoff of 50 kb (dataset iii) are similar to those obtained for a 100-kb cutoff (dataset ii). The large discrepancy in the results for these examples (namely that no viable model could be found for dataset ii and yet four- to eight-Gaussian models were found for dataset iii) demonstrates that seemingly minor changes in definitions can yield significant differences in results. Clearly, the discrepancy arises from differences in the parent GC-content distributions for these datasets. For example, it can be seen that the shorter the cutoff, the higher the proportion of high GC-content segments (fig. 5*b*).

Ideally, we might hope that such differences in distribution would alter only the relative weights of the different families. In this case, one could claim that the families are robust to preprocessing of the data. However, the results indicate that this is not the case. In fact, we find that the means and SD (widths) of the different families vary significantly between different models. Because significantly different families were obtained for datasets that vary only in their length cutoff, it is difficult to determine which collection of families constitutes a preferred model.

Overlap Among Isochore Families

All models exhibited large overlaps between Gaussians. In general, the large overlap between adjacent Gaussians helps reproduce the smooth unimodal parent distribution of GC contents. Without such overlaps, the overall GC-content distribution would contain observable dips. Extreme cases of overlap, where one Gaussian family is fully contained within another, also occur (for example, in models of three Gaussians or more for dataset vii, see Supplementary Material). Families that are fully contained within others pose technical difficulties (see below) as well as interpretational difficulties within the scope of the existing isochore theory (see *Discussion*).

Segment and Isochore Classification

The ability to classify individual isochores into particular families is an important premise of the isochore theory (attribute A6). Such a classification would be trivial had isochore families been uniquely defined by their GC content. However, because overlapping Gaussian distributions are used to define isochore families, classification necessarily involves some error. To assess the viability of attribute A6 for the multi-Gaussian models obtained above, classification of D_{JS} segments and fixed-length windows into Gaussian families was performed and corresponding classification errors were calculated (see *Methods*). These classification results were obtained for all statistically valid models. Figure 6 shows the classification errors associated with putative isochores for the five-Gaussian model. As expected, peaks in the classification errors correspond to maximal overlap between two or more Gaussians. When only two adjacent Gaussians overlap, the classification error can rise to 50% at any GC-content value. Where three or more Gaussians contribute to a given GC content, the error can be higher (up to 62.1% in the five-Gaussian model). The mean classification error when fitting four Gaussians to putative isochores is 24 ± 13%. Fitting more than four Gaussians to the D_{JS} segments is possible, with similar classification errors. The mean classification error in the five-Gaussian model of the same data is 26 ± 14% (fig. 6*c*). Classification errors for datasets iii and vi are even larger (e.g., 36 ± 10% error for the five-Gaussian model over dataset vi).

To demonstrate the impact of errors of this magnitude, we calculated the proportion of segments or windows that can be classified with 95% confidence (i.e., with an expected error smaller than 0.05). For a four-Gaussian model, only 5.3% of all putative isochores can be classified with 95% confidence, and a mere 1.8% can be confidently

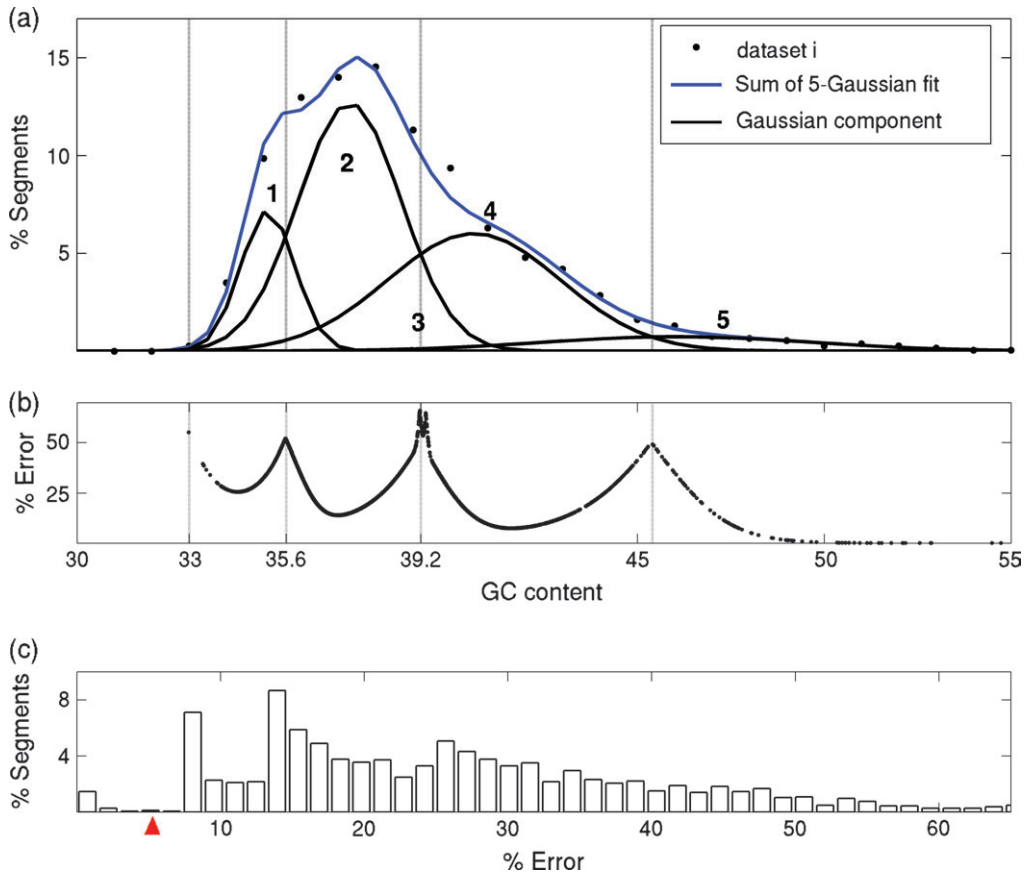

FIG. 6.—Expected classification errors for a five-Gaussian fit to putative isochores (dataset i). Subplot (*a*) shows the probability density histogram of the data (dots), the corresponding fit (blue), and the Gaussian components (numbered solid lines). Vertical dotted lines mark the intersection points between Gaussian components (i.e., the points at which classification changes between families). The classification errors are shown along the same GC-content axis in subplot (*b*). Intersection points between Gaussian families are shown to correspond to maximal classification errors. Subplot (*c*) shows the distribution of errors over segments (i.e., the fraction of segments that can be classified at the specified expected error). Classification errors range from 0% to almost 70% for this dataset (with a median error of 34%). Only a small fraction of segments can be classified with an expected error under 5% (red triangle).

classified for five-Gaussian models. Similar results are obtained for all multi-Gaussian models (and for all datasets). Thus, the classification of the vast majority of segments or fixed-length windows is ambiguous.

Discussion

The isochore theory, proposed nearly three decades ago (Macaya et al. 1976), is, to the best of our knowledge, the only attempt in the scientific literature to come to terms with the long-range compositional structures of metazoan genomes within an evolutionary framework. This theory stimulated hundreds of studies in which various biological phenomena were correlated with compositional features and were consequently interpreted within a dynamic evolutionary context. From the staining of chromosome bands (Saccone et al. 2001) through methylation patterns (Caccio et al. 1997) and gene localization (Duret, Mouchiroud, and Gautier 1995) to retroposition patterns (Smith and Higgs 1999) and tissue specificity of gene products (Vinogradov 2003), dozens of genetic and genomic features have been found to be associated with nucleotide composition.

The initial impetus for the isochore theory came from studies in which genomic DNA was randomly sheared into large fragments (~100 kb long), which were then sorted according to their GC content by centrifugation on CsCl or Cs_2SO_4 gradients. Later, higher resolution techniques (in particular gene sequencing) led to the discovery that the nucleotide composition of protein-coding genes is correlated with the GC levels of the nongenic portions flanking the genes (e.g., Bernardi and Bernardi 1985). Thereafter, it became generally accepted to assume that isochores harbor within them genes with corresponding GC contents. In fact, this interpretation laid the foundation for the development of proxy measures for GC contents that were used when the isochore GC content was unknown or when the family affiliation of an isochore could not be determined. In particular, most of the studies on isochores and their association with genomic elements as well as studies on evolutionary mechanisms responsible for the creation of isochores did not use isochoric data. Instead, they used the GC composition at third-codon positions of protein-coding genes. Thus, for many years, the common practice in the literature has been to take the existence of isochores

for granted and to interpret associations between various genetic traits and the GC content at second- and third-codon positions of protein-coding genes as isochoric associations (Zoubak, Clay, and Bernardi 1996; Galtier et al. 2001; Alvarez-Valin, Lamolle, and Bernardi 2002; Duret et al. 2002; Piganeau et al. 2002; Daubin and Perriere 2003; Cruveiller et al. 2004).

We note, however, that protein-coding genes constitute less than 5% of the human genome. Hence, the genic fraction of the genome may not be representative of the human genome as a whole. The sequencing of the complete human genome (Lander et al. 2001) paved the way for a sequence-based reinspection of the isochore theory. The results of this re-examination were somewhat controversial. Whereas Lander et al. (2001) suggested that isochores do not merit the prefix "iso," Li et al. (2003) entitled their rebuttal "Isochores merit the prefix 'iso'." Subsequently, the debate between the pro- and anti-isochore factions became mired in semantics characterized by clashes over the definition of words such as "roughly" and "approximately."

In this study, we tried to veer away from ambiguity. Instead, we identified six attributes of isochores and tested their veracity against human genomic sequences. We intended our definitions to be rigorous enough to eliminate misinterpretation. Of course, other definitions are possible. Nonetheless, we believe that the work presented here is an important and constructive step in addressing the isochore question and provides a better understanding of the compositional structures in the human genome.

Attribute Definitions (A1–A3)

The first attribute we considered was that isochores are genomic segments with a characteristic GC content (attribute A1). There may be two interpretations to this statement. One interpretation of the term characteristic involves the notion of stationarity or flatness (Grosse et al. 2002). An alternative interpretation invokes the notion of distinctiveness. In other words, we may interpret the term characteristic to imply that the fluctuations in GC content along an isochore are small enough to distinguish it from the GC content of adjacent segments. In this study, we used the D_{JS} segmentation algorithm (Bernaola-Galván, Róman-Roldán, and Oliver 1996) that implements the distinctiveness criterion. This procedure unravels the well-documented mosaic of compositionally distinct structures along the genome.

Statements of the isochore theory often describe isochores as "fairly" or "relatively" homogeneous. These claims have been challenged in the literature. For instance, based on a comparison of GC content between adjacent windows in human fixed-length DNA windows, it was claimed that the human genome is much more heterogeneous than anticipated by the isochore model (Nekrutenko and Li 2000). A similar conclusion was reached by comparing the GC variability within 300-kb windows to the expectations derived from a uniform GC distribution (Lander et al. 2001). A closer inspection of chromosomes 21 and 22, in which the variability of 10-kb, 100-kb, and 1-Mb windows was compared to a randomized sequence of these chromosomes, again yielded identical conclusions (Häring and Kypr 2001). However, all these challenges (Nekrutenko and Li 2000; Häring and Kypr 2001; Lander et al. 2001) were based on analyses of fixed-length windows rather than sequences obtained through a composition-based segmentation procedure.

Because the isochore theory is based firstly on the existence of homogeneous regions of characteristic GC content within the human genome (as verified here), a proper examination of the theory has to begin with the identification and quantification of these regions. This cannot be achieved using fixed-length windows to divide the DNA as this method does not map out the isochores themselves. Indeed, subsequent studies that used segmentation methods did find compositionally homogeneous regions within the human genome (Li 2002; Li et al. 2003). These studies, however, provided little detail as to the prevalence of isochores. In this study, we used a statistical test similar to that in Li (2002) to compare the variance of a segment to the variance of the chromosome on which it resides. In attribute A2, we interpret "relative homogeneity" as a criterion for isochores to be more homogenous in their composition than the chromosome on which they reside.

An additional attribute of isochores is their "characteristic length" (of the order of many hundreds to a few thousand kilobases, attribute A3). In fact, in line with previous studies (Buldyrev et al. 1998), we found that the distribution of segment lengths in all chromosomes resembles a power-law distribution (see fig. 3*a*). Hence, these segments have no characteristic length scale. This feature in itself is problematic for the isochore theory because isochores are frequently claimed to have characteristic sizes (Bernardi 2001). Note that the power-law distribution of segment lengths is not an artifact of the D_{JS} segmentation procedure; a variety of alternative segmentations and other statistical analyses yield similar scale-invariant behavior (not shown). The scale-free distribution of segment lengths suggests that any choice of length cutoff for putative isochores would be arbitrary and lacks biological meaning. However, in the primary and secondary literature, it is customary to describe isochores as much longer than 300 kb (Macaya, Thiery, and Bernardi 1976; Cuny et al. 1981; Bernardi et al. 1985; Bernardi 2000; Clay and Bernardi 2001; Pavlicek et al. 2002), and we have adopted this cutoff in this paper. Naturally, the choice of length cutoff is important and affects a variety of conclusions.

Putative Isochores

The results presented here suggest that if isochores truly possess attributes A1–A3, they span less than half of the sequenced part of the human genome. Nonetheless, alternative putative isochores (with lower length cutoffs) would satisfy this criterion. Even a modest compromise in the length cutoff to 100 kb results in 67% coverage of the human genome by alternative putative isochores (table 1).

Similar to genes and repetitive elements, the number of putative isochores correlates with the length of the corresponding chromosome. Interestingly, the density of putative isochores is negatively correlated with the density of genes ($r = -0.68$) and *Alu*s ($r = -0.71$) on the different chromosomes (data from Dagan et al. 2004). A possible explanation for this negative correlation might be that

the isochoric regions we found are by and large GC poor, while genes and *Alu*s are known to reside primarily in GC-rich regions (Lander et al. 2001). This finding invalidates the proposed relationship between isochores and gene expression (Bernardi et al. 1985; Bernardi 2000). If most genes are not located on isochores, then isochores cannot influence the mode of gene expression.

Compositional Families: Segments, Fragments, and Fixed-Length Windows

The original description of isochore families within the human genome was based on a histogram of buoyant densities of DNA fragments 50–100 kb in length (Filipski, Thiery, and Bernardi 1973). It was suggested that the observed heterogeneity in GC content could be explained by five symmetric (Gaussian) distributions, each representing an isochore family (attribute A5). We found that four Gaussians are sufficient to capture the underlying empirical distribution of GC content, and under William of Occam's *entia non sunt multiplicanda praeter necessitatem*, the more parsimonious model should be preferred. Thus, unless further justification is provided, a five-Gaussian model is unwarranted.

To place these findings in context, it is worthwhile to examine possible Gaussian-family descriptions of alternative putative isochores. This comparison is motivated by a search for a robust solution, i.e., a model that tolerates changes in isochore selection criteria. A priori, one would not expect a multi-Gaussian model of putative isochores to apply to alternative data sets. The incompatibility arises from the differences in underlying GC-content distributions (in particular, the relative abundance of GC-poor segments). Indeed, there does not appear to be any evidence of a robust multi-Gaussian description of alternative sets of putative isochores. This intense sensitivity to the choice of length cutoff calls into question the suitability and utility of multi-Gaussian models for describing isochore families.

Traditionally, multi-Gaussian models of isochore families have been derived, not from isochores directly but from gradient centrifugation fragments (Bernardi 2000) or, more recently, from fixed-length windows of the sequenced genome (Pavlicek et al. 2002). Both fragments and fixed-length windows were assumed to preserve isochoric compositional properties. The fact that fixed-length windows of a range of different lengths have been shown to possess very similar GC-content distributions (Clay and Bernardi 2001; fig. 4*c*) has been used as evidence for this claim. In what follows, we address the question, to what extent GC-content families obtained from fixed-size windows are equivalent to isochore families. For this discussion, we need to recall that, according to this theory, isochores are much longer than the fragments used in the gradient centrifugation studies, are homogeneous in their GC content, and span most of the genome (attributes A2–A4).

If isochores indeed span the majority of the genome and are much longer than DNA fragments obtained from gradient centrifugation, then the vast majority of fragments should lie entirely within isochores (rather than outside them or across isochore boundaries). This should hold for fragment sizes of 50–100 kb and isochores longer than 300 kb (at least for the fraction of the genome spanned by isochores). What is required for these fragments to have the same GC-content distribution as isochores?

First, the fragments should be sufficiently long: short sequences do not preserve the GC content of the corresponding isochore because they may only represent a local fluctuation in GC content. Furthermore, short sequences have a very broad GC-content distribution (much broader than isochore families, not to mention individual isochores). Are 50- to 100-kb-long fragments long enough to possess a GC content similar to their corresponding isochore? The answer is no: as we have seen, fixed-length windows have debatable compositional homogeneity and cannot in general be attributed to a characteristic GC content (Häring and Kypr 2001; Lander et al. 2001).

Second, even if isochores are perfectly homogeneous, so that the GC content of each fragment is the same as that of the corresponding isochore, the weighting of the distributions will be different: distributions of segment GC content give equal weight to short and long segments, whereas distributions of fragment GC content give larger weights to longer isochores (because long isochores contain many fragments). The two distributions can only be statistically equivalent if the GC content in isochores is statistically independent of isochore lengths. In fact, we have shown that this does not hold for D_{JS} segments because longer segments tend to be GC poor.

Finally, the segmentation performed here yields only 41% coverage of the genome by putative isochores. Thus, fixed-length windows contain an additional 59% of non-isochoric segments. Based on all these considerations, it is evident that GC-content families of genome fragments do not, in any way, correspond to families of putative isochores. Let us, therefore, discuss GC-content families derived from fixed-length windows, as opposed to putative isochores.

To generate genomic segments compatible with the centrifugation experiments, we used window sizes of about 65 kb. Indeed, we were able to fit five Gaussians or more to the empirical data. This naively confirms the suitability of a five-Gaussian model to describe fixed-length windows, with a length scale that roughly corresponds to the original experiments.

Furthermore, because previous papers in the literature have claimed that the size of the window is immaterial for the GC distribution (Clay et al. 2001; Pavlicek et al. 2002), we repeated the analysis for larger window sizes of about 525 kb (roughly corresponding to traditional quotes for isochore lengths). Indeed, the GC-content distribution of these larger windows is virtually indistinguishable from that of the shorter (65 kb) windows. However, due to our choice of statistical test for Gaussian-mixture models, three Gaussians were sufficient to model the distribution of GC contents in windows of 525 kb. In effect, the more appropriate conclusion is that the distribution of windows (of either size) may be modeled by a Gaussian-mixture model of as few as three Gaussians, up to a resolution of 1% GC; for a finer resolution of 0.35%, a minimum of five Gaussians is required. The number of Gaussians is thus reduced to the choice of resolution of the model used to describe the empirical data. Once again, we find that the utility of the

multi-Gaussian model is called into question as it does not appear robust to minor changes in criteria.

Why, then, is the distribution of GC content among alternative putative isochores so sensitive to the choice of length cutoff, while GC-content distributions of fixed-length windows appear rather stable? In contrast to putative (or alternative putative) isochores, fixed-length windows do not include wide ranges of contig sizes. Therefore, the stable GC-content distributions for fixed-length windows are consistent with our finding for compositionally distinct segments, namely that longer segments have a narrower GC-content distribution with a lower mean GC content.

Putting these findings aside, we may still wish to consider the viability of the five-Gaussian model presented in the literature for describing putative isochores. The Gaussian families obtained from our statistical models do not correspond to the descriptions of the isochore families found in the literature. In fact, throughout the years, the description of the five isochore families has somewhat wavered in terms of number of families, mean GC content, and relative weight. The most up-to-date description of the isochore families (Pavlicek et al. 2002) is defined to a resolution of 1% GC and lists the contributions of different isochore families to be 63% (for L1 and L2) and 25%, 8%, and 4% (for H1, H2, and H3, respectively). By applying these GC-content categories to D_{JS} segments, putative and alternative putative isochores, as well as to fixed-length windows, we found, as expected, that only the fixed-length window description was consistent with the findings of Pavlicek et al. (2002).

Classification by Family

A basic, though unstated, premise of the isochore theory is that the theory is useful. In other words, it is understood that given a genomic segment, it is possible to classify it (with reasonable statistical confidence) into a particular isochore family (attribute A6). However, for any of the multi-Gaussian models we have found, the error rates appear unacceptably high. Ambiguous classification of segments is a direct result of overlaps between candidate families. In fact, it seems that the long GC-rich tail of the distribution is responsible for most of the overlaps. Describing such a biased distribution by the sum of symmetric Gaussians is problematic, resulting in our inability to reliably classify a putative isochore into one of the families. Unfortunately, this ascertainment failure deems the Gaussian-mixture model of isochore families of dubious practical benefit.

The isochore model has been one of the most useful theories in molecular evolution for the last 30 years. Its main historical importance is in highlighting a fundamental compositional difference between eukaryotic and prokaryotic genomes and identifying the nonuniformity of nucleotide composition within eukaryotic genomes. These two observations remain valid. Nevertheless, in light of our analysis, the particulars of the theory must either be modified or be discarded. Nekrutenko and Li (2000) suggested a looser definition of isochores, whereby "any genomic fragment longer or equal to 100 kb such that when it is divided into a series of overlapping 10 kb windows, no two windows can differ by >7% GC." This definition was based on an analysis of the yeast genome (*Saccharomyces cerevisiae*), which is thought to be devoid of isochores, as a "control" genome. However, by relaxing the definitions of isochores, we ignore other regularities in the human genome, such as the length distribution of compositionally distinct segments. Moreover, one might ask, as more genomes become available for analysis, would following this path lead to further and further relaxations in the definition of isochores. In fact, a preliminary analysis of the genome of *Schizosaccharomyces pombe* indicates that the genome of *S. cerevisiae* is exceptional even among unicellular organisms and as such may be taxonomically unrepresentative (Dagan and Graur, unpublished data).

An alternative approach would suggest that the isochore theory has reached the limit of its usefulness as a description of genomic compositional structures. Isochores will remain a good approximate description suitable for didactic purposes, but for more purist aims, we may need to approach the issue of composition and heterogeneity in a different manner. We are now in a position that affords a look at the human genome with arbitrary resolution. This should enable us to find a more useful metaphor for the evolutionary dynamics of GC content within the human genome.

Supplementary Material

Tables D1, S1–S3, and figure S1 are available at *Molecular Biology and Evolution* online (www.mbe.oupjournals.org).

Acknowledgments

The authors appreciate helpful feedback from Oliver Clay. We thank Samuel Braunstein and Giddy Landan for numerous discussions and for their critical reading of the manuscript. T.D. was supported in part by a scholarship in Complexity Science from the Yeshaia Horvitz Association. L.S. and N.C. gratefully acknowledge the support of the James S. McDonnell Foundation.

Literature Cited

Alvarez-Valin, F., G. Lamolle, and G. Bernardi. 2002. Isochores, GC(3) and mutation biases in the human genome. Gene **300**:161–168.

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

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

———. 2001. Misunderstanding about isochores: part 1. Gene **276**:3–13.

Bernardi, G., and G. Bernardi. 1985. Codon usage and genome composition. J. Mol. Evol. **22**:363–365.

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

Buldyrev, S. V., N. V. Dokholyan, A. L. Goldberger, S. Havlin, C. K. Peng, H. E. Stanley, and G. M. Viswanathan. 1998. Analysis of DNA sequences using methods of statistical physics. Physica A **249**:430–438.

Caccio, S., K. Jabbari, G. Matassi, F. Guermonprez, J. Desgres, and G. Bernardi. 1997. Methylation patterns in the isochores of vertebrate genomes. Gene **205**:119–124.

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

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.

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

Cuny, G., P. Soriano, G. Macaya, and G. Bernardi. 1981. The major components of the mouse and human genomes. 1. Preparation, basic properties and compositional heterogeneity. Eur. J. Biochem. **115**:227–233.

Dagan, T., R. Sorek, E. Sharon, G. Ast, and D. Graur. 2004. *Alu*Gene: a database of *Alu* elements incorporated within protein-coding genes. Nucleic Acids Res. **32**:D489–D492.

Daubin, V., and G. Perriere. 2003. G+C3 structuring along the genome: a common feature in prokaryotes. Mol. Biol. Evol. **20**:471–483.

Deloukas, P., L. H. Matthews, J. Ashurst et al. (127 co-authors). 2001. The DNA sequence and comparative analysis of human chromosome 20. Nature **414**:865–871.

Dunham, I., N. Shimizu, B. A. Roe et al. (217 co-authors). 1999. The DNA sequence of human chromosome 22. Nature **402**:489–495.

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

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

Filipski, J., J. P. Thiery, and G. Bernardi. 1973. An analysis of the bovine genome by Cs2SO4-Ag density gradient centrifugation. J. Mol. Biol. **80**:177–197.

Galtier, N., G. Piganeau, D. Mouchiroud, and L. Duret. 2001. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics **159**:907–911.

Grosse, I., P. Bernaola-Galvan, P. Carpena, R. Roman-Roldan, J. Oliver, and H. E. Stanley. 2002. Analysis of symbolic sequences using the Jensen-Shannon divergence. Phys. Rev. E **65**:041905.

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

Hattori, M., A. Fujiyama, T. D. Taylor et al. (64 co-authors). 2000. The DNA sequence of human chromosome 21. Nature **405**:311–319.

Heilig, R., R. Eckenberg, J. L. Petit et al. (99 co-authors). 2003. The DNA sequence and analysis of human chromosome 14. Nature **421**:601–607.

Hillier, L. W., R. S. Fulton, L. A. Fulton et al. (107 co-authors). 2003. The DNA sequence of human chromosome 7. Nature **424**:157–164.

Inman, R. B. 1966. A denaturation map of the lambda phage DNA molecule determined by electron microscopy. J. Mol. Biol. **18**:464–476.

Lander, E. S., 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., P. Bernaola-Galvan, P. Carpena, and J. L. Oliver. (171 co-authors). 2003. Isochores merit the prefix 'iso'. Comput. Biol. Chem. **27**:5–10.

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

Li, W. T., P. Bernaola-Galvan, F. Haghighi, and I. Grosse. 2002. Applications of recursive segmentation to the analysis of DNA sequences. Comput. Chem. **26**:491–510.

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.

Mungall, A. J., S. A. Palmer, S. K. Sims et al. (171 co-authors). 2003. The DNA sequence and analysis of human chromosome 6. Nature **425**:805–811.

Nekrutenko, A., and W. H. Li. 2000. Assessment of compositional heterogeneity within and between eukaryotic genomes. Genome Res. **10**:1986–1995.

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

Pavlicek, 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.

Peng, C. K., S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger. 1994. Mosaic organization of DNA nucleotides. Phys. Rev. E **49**:1685–1689.

Piganeau, G., D. Mouchiroud, L. Duret, and C. Gautier. 2002. Expected relationship between the silent substitution rate and the GC content: implications for the evolution of isochores. J. Mol. Evol. **54**:129–133.

Price, K., and R. Storn. 1997. Differential evolution. Dr Dobbs J. **22**:18–24.

Saccone, S., S. Caccio, P. Perani, L. Andreozzi, A. Rapisarda, S. Motta, and G. Bernardi. 1997. Compositional mapping of mouse chromosomes and identification of the gene-rich regions. Chromosome Res. **5**:293–300.

Saccone, S., A. Pavlicek, C. Federico, J. Paces, and G. Bernardi. 2001. Genes, isochores and bands in human chromosomes 21 and 22. Chromosome Res. **9**:533–539.

Salinas, J., M. Zerial, J. Filipski, M. Crepin, and G. Bernardi. 1987. Nonrandom distribution of MNTV proviral sequences in the mouse genome. Nucleic Acids Res. **15**:3009–3022.

Smith, Z. E., and D. R. Higgs. 1999. The pattern of replication at a human telomeric region (16p13.3): its relationship to chromosome structure and gene expression. Hum. Mol. Genet. **8**:1373–1386.

Thiery, J. P., G. Macaya, and G. Bernardi. 1976. An analysis of eukaryotic genomes by density gradient centrifugation. J. Mol. Biol. **108**:219–235.

Vinogradov, A. E. 2003. Isochores and tissue-specificity. Nucleic Acids Res. **31**:5212–5220.

Wen, S. Y., and C. T. Zhang. 2003. Identification of isochore boundaries in the human genome using the technique of wavelet multiresolution analysis. Biochem. Biophys. Res. Commun. **311**:215–222.

Zar, J. H. 1999. Biostatistical analysis. Prentice Hall, Upper Saddle River, N.J.

Zhang, C. T., and R. Zhang. 2003. An isochore map of the human genome based on the Z curve method. Gene **317**:127–135.

Zoubak, B., J. H. Richardson, A. Rynditch, P. Hollsberg, D. A. Hafler, E. Boeri, A. M. L. Lever, and G. Bernardi. 1994. Regional specificity of HTLV-I proviral integration in the human genome. Gene **143**:155–163.

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

Takashi Gojobori, Associate Editor

Accepted January 25, 2005