ELSEVIER

Gene 276 (2001) 47–56

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

www.elsevier.com/locate/gene

Isochore chromosome maps of eukaryotic genomes

José L. Oliver[a,*], Pedro Bernaola-Galván[b], Pedro Carpena[b], Ramón Román-Roldán[c]

[a]*Departamento de Genética, Instituto de Biotecnología, Universidad de Granada, E-18071 Granada, Spain*
[b]*Departamento de Física Aplicada II, Universidad de Málaga, Malaga, Spain*
[c]*Departamento de Física Aplicada, Universidad de Granada, Granada, Spain*

Received 26 March 2001; received in revised form 2 May 2001; accepted 26 July 2001
Received by G. Bernardi

Abstract

Analytical DNA ultracentrifugation revealed that eukaryotic genomes are mosaics of isochores: long DNA segments ($\gg$ 300 kb on average) relatively homogeneous in G + C. Important genome features are dependent on this isochore structure, e.g. genes are found predominantly in the GC-richest isochore classes. However, no reliable method is available to rigorously partition the genome sequence into relatively homogeneous regions of different composition, thereby revealing the isochore structure of chromosomes at the sequence level. Homogeneous regions are currently ascertained by plain statistics on moving windows of arbitrary length, or simply by eye on G + C plots. On the contrary, the entropic segmentation method is able to divide a DNA sequence into relatively homogeneous, statistically significant domains. An early version of this algorithm only produced domains having an average length far below the typical isochore size. Here we show that an improved segmentation method, specifically intended to determine the most statistically significant partition of the sequence at each scale, is able to identify the boundaries between long homogeneous genome regions displaying the typical features of isochores. The algorithm precisely locates classes II and III of the human major histocompatibility complex region, two well-characterized isochores at the sequence level, the boundary between them being the first isochore boundary experimentally characterized at the sequence level. The analysis is then extended to a collection of human large contigs. The relatively homogeneous regions we find show many of the features (G + C range, relative proportion of isochore classes, size distribution, and relationship with gene density) of the isochores identified through DNA centrifugation. Isochore chromosome maps, with many potential applications in genomics, are then drawn for all the completely sequenced eukaryotic genomes available.

Keywords: Sequence heterogeneity; Compositional segmentation; Comparative genomics

1. Introduction

The genomes of warm-blooded vertebrates (Bernardi et al., 1985), and also of many other eukaryotes (Bernardi, 1995, 2000; Gautier, 2000), are mosaics of isochores, i.e. long DNA segments ($\gg$ 300 kb on average) relatively homogeneous in GC content (above a size of 3 kb) when compared to the pronounced heterogeneity throughout the entire genome. Isochores belong to five families covering a wide GC range. GC-poor isochores of the L1-L2 families (GC% $<$ 44) are poor in genes, while GC-rich isochores of the H1 (44 $\leq$ GC% $<$ 47), H2 (47 $\leq$ GC% $<$ 52) and H3 (GC% $\geq$ 52) families are increasingly rich in genes (cf. Zoubak et al., 1996). To date, only one clear isochore boundary has been experimentally found and fully characterized at the sequence level, that separating classes II and III of the human major histocompatibility complex (MHC) (Fukagawa et al., 1995, 1996; Tenzen et al., 1997; Stephens et al., 1999). The absence of a reliable computational method has hampered the systematic characterization of other isochore boundaries in the long genome regions now being generated by sequencing genome projects.

The G + C homogeneity within, and the differences between, isochores have been deduced mainly from CsCl profiles revealed by analytical ultracentrifugation (Macaya et al., 1976; Thiery et al., 1976), from probe hybridization to the different compositional fractions obtained by preparative density gradient centrifugation of bulk DNA (see Bernardi et al., 1985, and references therein), as well as from compositional mapping (Gardiner et al., 1990; Bettecken et al., 1992; De Sario et al., 1996). In some cases, as in the human dystrophin gene, a remarkably uniform (±0.5%) GC content within isochores was

Abbreviations: LHGR, long homogeneous genome regions; MHC, major histocompatibility complex

* Corresponding author. Fax: +34-958-244073.
E-mail address: oliver@ugr.es (J.L. Oliver).

0378-1119/01/$ - see front matter
PII: S0378-1119(01)00641-2

reported, whereas adjacent isochores are separated by compositional discontinuities of about 2% GC (Bettecken et al., 1992).

At the shorter sequence scale, however, this compositional structure is hard to reveal, mainly because of the complex heterogeneity of eukaryotic DNA (Li and Kaneko, 1992; Peng et al., 1992; Bernaola-Galván et al., 1996; Li et al., 1998; Román-Roldán et al., 1998). A moving-window graphic plot of GC content routinely accompanies the description of every new genomic sequence that appears in the literature, but, unfortunately, the long-range patterns appearing on these plots are usually identified only by eye. Examples of this practice are the 'isochores' tentatively identified on the complete sequences of human chromosomes 21 (Hattori et al., 2000) and 22 (Dunham et al., 1999). In other cases, DNA domains loosely defined by moving windows of arbitrary length are taken as a reference to test the correlation between GC content and different biological properties (Lander et al., 2001; Venter et al., 2001). Some more elaborate approaches to locate compositional patterns at the genome sequence level are available (Churchill, 1989; Frank and Lobry, 2000; Ramensky et al., 2000), but none was suitable for the detection of isochore boundaries. A specifically designed technique attempts to identify homogeneous regions by keeping the heterogeneity of overlapping moving windows below a given fluctuation limit (Nekrutenko and Li, 2000), but it produced a listing of overlapping regions, instead of a true partition of the sequence into separate regions of different composition. Another recently published approach (Häring and Kypr, 2001), also based on moving windows, fails to detect isochores in the human chromosomes 21 and 22. A choice of narrow fluctuation limits in the last two studies may have been partly responsible for their findings. More rigorous and specifically designed techniques are therefore needed for the accurate mapping and characterization of the long-range chromosome structures appearing in genomic DNA sequences.

The tool of choice for this task may be compositional segmentation, a proven method able to divide a DNA sequence into non-overlapping, relatively homogeneous domains at a given level of statistical confidence (Bernaola-Galván et al., 1996, 1999, 2000; Román-Roldán et al., 1998; Oliver et al., 1999). However, the lengths of the domains generated by an early version of this algorithm were far below the typical isochore size. We now show that an improved segmentation method, designed specifically to determine the most statistically significant partition of the sequence at each scale, is able to identify the boundaries between long homogeneous genome regions, or LHGRs. It is worth mentioning that since sequence homogeneity is only a relative concept, the term homogeneous in 'LHGR' refers to the high homogeneity within a region, when compared to the large heterogeneity in the entire sequence being analyzed. Many of the LHGRs we found in large human contigs display the typical features of isochores.

2. Data and methods

Only completely sequenced chromosomes or large genome contigs were analyzed. Human genome contigs larger than 500 kb were extracted from *.gbk files in GenBank (ftp://ncbi.nlm.nih.gov/genbank/genomes/H_sapiens/) containing genomic sequences assembled from finished (phase 3) high throughput genomic sequence data (October, 2000). The size of these contigs ranged from 512 kb to 28.7 Mb, with an average length of 1.5 Mb. We also analyzed a second set of contigs larger than 1000 kb from the Human Genome Project Working Draft, University of California at Santa Cruz (http://genome.ucsc.edu/, 7th October, 2000 freeze). The complete sequences of *Drosophila melanogaster*, *Caenorhabditis elegans*, *Arabidopsis thaliana* (chromosomes II and IV) and *Saccharomyces cerevisiae* were also retrieved from GenBank. The provisional assemblies of *Arabidopsis* chromosomes I, III and V were retrieved from the MIPS *Arabidopsis thaliana* group ftp server (ftp://ftp.mips.biochem.mpg.de/pub/cress/). The consensus sequence for the human MHC was produced by the Human Chromosome 6 Sequencing Group at the Sanger Centre and can be obtained from http://www.sanger.ac.uk/HGP/Chr6/published_consensus.fasta. A complete list of the sequences being analyzed is available at our website (http://bioinfo2.ugr.es/isochores), and will be continuously updated as more chromosome sequences appear in nucleotide databases. Isochore chromosome maps and detailed segmentation results on every analyzed sequence are also available at this site.

2.1. An improved segmentation algorithm

The algorithm used here is a modification of the entropic segmentation method for DNA sequences described elsewhere (Bernaola-Galván et al., 1996, 1999, 2000, 2001; Román-Roldán et al., 1998; Oliver et al., 1999). Several improvements are introduced here to deal with the isochore mapping problem. The most important improvement was that the cuts are now chosen in an ordered way: at each step, the cut maximizing the overall compositional complexity of the entire sequence is chosen (for a definition of sequence compositional complexity, see Román-Roldán et al., 1998). This procedure is equivalent to maximizing the statistical significance of each cut. This strategy may be more adequate in searching for homogeneous segments within the long-range correlated, fractal landscape of eukaryotic DNA (Peng et al., 1992; Román-Roldán et al., 1998). Within this multi-scale landscape, the statistical significance of isochore boundaries may vary with the length scale being considered. By cutting the sequence in the ordered way described above, the choice of the most significant cut at each scale is guaranteed. The theoretical fundamentals of the improved segmentation method are presented elsewhere (Román-Roldán et al., 2001). In brief, the new algorithm works as follows:

1. A sliding border is moved along the entire sequence,

computing at each point a distance measure between the left and right subsequences defined by the border. We used here the discrepancy D, defined as $D = JS_2 \times n$, where JS_2 is the Jensen–Shannon divergence between the left and right subsequences defined by the border (for details about computing JS_2, see Bernaola-Galván et al., 1996), and n is the length of the segment to be split. D takes into account not only the GC differences, but also the domain sizes, i.e. two very large adjacent LHGRs may be recognized as distinct even when they differ only slightly in GC content.

2. The position showing the maximum D value is used to split the sequence into two segments.
3. Steps 1 and 2 are iterated to the end of the process (see below for the stop segmentation criterion we used). It is worth mentioning that at each step all the existing segments are explored for discrepancy values. However, only one of these segments is split into two subsegments, that within which D reaches its maximum value for this step.

Binary (S/W, R/Y) as well as quaternary (A,T,C,G) alphabets can be used in segmenting DNA sequences (Bernaola-Galván et al., 1996; Román-Roldán et al., 1998; Oliver et al., 1999). Since isochores are patches of different G + C content, we used the S/W binary alphabet throughout.

2.2. Stop segmentation criterion

As mentioned above, the cuts given by the improved segmentation algorithm are ordered from highest to lowest contribution to the overall sequence complexity. Therefore, the first cuts may be expected to be more statistically significant than the last ones. As segmentation proceeds, less and less significant domains appear. We need, therefore, a standard criterion to stop the process at any reasonable intermediate step. The adoption of such a standard is also useful for the comparison of the segmentation results for different sequences.

Here, we take advantage of the current definition of isochores: long DNA segments homogeneous in GC levels (above a size of 3 kb) and differing from adjacent isochores by some compositional discontinuity (Bettecken et al., 1992; Bernardi, 2000). Consequently, at each step of the segmentation process, we check whether the resulting adjacent segments differ in GC levels. Whenever all the pairwise differences were statistically significant (i.e. $P \leq 0.05$), the segmentation process continued; otherwise it was stopped.

In testing for GC differences between adjacent segments, we divided each segment into non-overlapping tracts of 3 kb, and computed the tract GC content. A t-test was then used to compare the tract GC differences between each pair of adjacent segments. In this way, and in compliance with experimental observations, we are filtering out the short-range variability at the nucleotide level, and restricting ourselves to the GC patterns appearing above a size of 3 kb. The tract size seems not to be a critical parameter, as the isochore structure of the MHC region shown in Fig. 2 can be obtained with tract sizes ranging from 2 to 30 kb.

Since the statistical distribution of tract GC levels is unknown, a randomization test was used to validate the P values obtained in t-tests. In comparing two putative isochores, we first computed a true t value as described above. Then, 10,000 randomized data sets were produced. In each random data set, we took all the tract GC values and randomly reallocated them to the left or to the right of the potential isochore boundary. For each randomized set, we calculated a t value. The proportion of the 10,000 random data sets with t values higher than the true t value is a direct estimate of P.

It should be emphasized that isochore boundaries are predicted exclusively by the segmentation algorithm. The t-tests on tract GC content are used only to stop the process at some intermediate step.

2.3. Error in boundary determination

Given the statistical nature of segmentation methods, as well as the fact that we are using sequence compositional features to search for boundaries originated by biological processes not necessarily aimed to produce such a compositional differentiation, some fluctuation in boundary determination may be expected. We use numerical simulation to estimate this error. We generate a random (white noise) sequence built from two subsequences of lengths N_1 and N_2, P_1 and P_2 being the probabilities of obtaining C or G in the first and second subsequence, respectively. Next, by using our improved segmentation algorithm, we look for the location of the boundary (N_b) between these two subsequences. The distance (measured in bp) from the real boundary (N_1) to N_b gives us the error for this particular sequence. We repeat this experiment on a set of random sequences of several lengths and compositions (1000 sequences for each combination of length and composition) and, finally, we average all the errors. We find that the average error for typical differences in composition ($|P_1 - P_2| \sim 5\%$) is about 300 bp, independently of N_1 and N_2 (Fig. 1). This means that although the absolute error does not depend on N_1 and N_2, the relative error decreases as N_1 and N_2 increase. For example, for typical isochore sizes ($N_1, N_2 \sim 300$ kb) the relative error ranges from 0.15 to 0.05%. Further experiments are needed to determine the error rate in long-range correlated sequences.

3. Results

3.1. The isochores of the human MHC

Class II (in an L2 isochore) and class III (in an H3 isochore) regions of the human MHC are the only fully characterized isochores determined at the sequence level

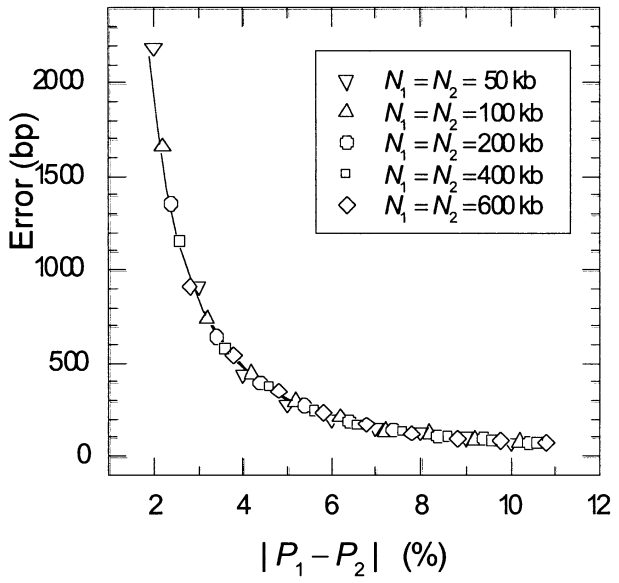

Fig. 1. The error in boundary determination plotted against the compositional differences between the two subsequences involved. A total of 1000 random sequences for each combination of length and composition were used in the simulation. Isochore sizes (N_1, N_2) of 50, 100, 200, 400, and 600 kb were considered.

to date (Fukagawa et al., 1995; Stephens et al., 1999; The MHC Sequencing Consortium, 1999). The isochore boundary separating the two regions contains sequences similar to the pseudoautosomal boundaries of the human sex chromosomes (Fukagawa et al., 1995, 1996), but no similar motifs have been found at the centromeric isochore boundary. Switching of DNA replication timing occurs at the isochore boundary, from 'later' replication in the class II region to 'earlier' replication in class III (Tenzen et al., 1997).

Previous predictions of isochore boundaries in the consensus sequence of this region (Stephens et al., 1999) were based on a moving-window plot of G + C content, and thus were only approximate: the reported 'sharp' G + C transitions at the isochore boundaries span between 50 and 100 kb. Our segmentation algorithm now enables a more precise location of MHC isochore boundaries (Fig. 2). The first, more statistically significant cut was given at position 2,483,966, which is within the sequence junction separating L2 and H3 isochores (Fukagawa et al., 1995). The second cut occurs at position 3,384,907, thus identifying the centromeric end of the isochore L2, while the third one was given at position 1,841,871, which marks the telomeric end of the isochore H3. The four additional cuts on this sequence all fell outside of the two isochores, thus defining other homogeneous regions within the MHC sequence. Given the sizes and compositional differences between MHC isochores, the estimated errors in the determination of these boundaries were all below 500 bp (see Section 2, Fig. 1).

3.2. Other human isochores

Next, we investigated the entire collection of human putative isochores in two separate sets of long human sequences: (1) contigs larger than 500 kb assembled from finished (phase 3) high throughput genomic sequence data (GenBank); and (2) contigs larger than 1000 kb from the human genome draft sequence (University of California at Santa Cruz, 7th October freeze). As an example, the isochore chromosome maps for the larger contigs of chromosomes 21 (NT_002836) and 22 (NT_001454) are shown in Fig. 3. Detailed segmentation results for the remaining contigs can be found at our website.

Since the ordering and orientation of the overlapping fragments in the University of California at Santa Cruz assembly are still tentative, we focus only on the GenBank contig data set for further statistical analyses. A total of 188 chromosome contigs, amounting to more than 210 Mb of DNA, were scanned by our segmentation algorithm. In segmenting such partial chromosome sequences, incomplete as well as complete LHGRs were produced. Three situations were possible. First, some contigs remained unsegmented; we considered them incomplete LHGRs,

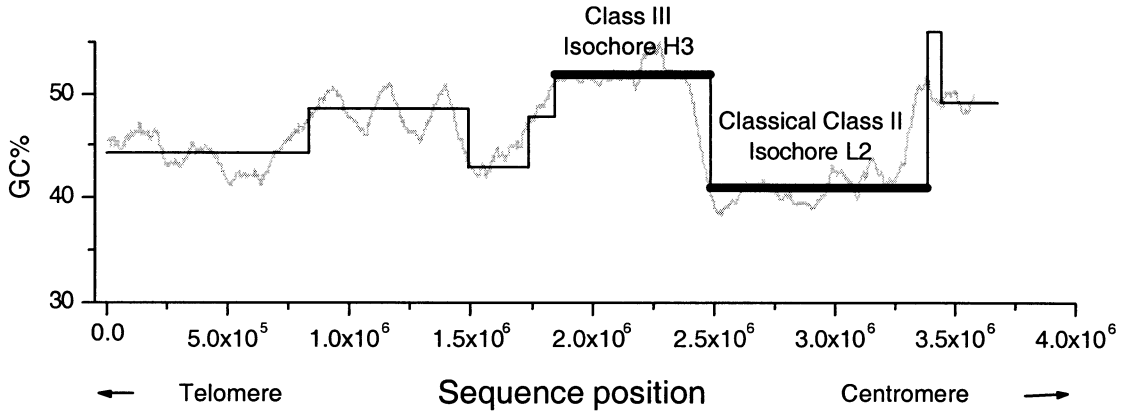

Fig. 2. Isochore chromosome map (straight lines) of the human MHC region ($P \leq 0.05$). Bold lines indicate the two experimentally determined isochores. For comparison, the GC% in a moving window (length, 100 kb; step, 1 kb) across the 3,673,800-base-long consensus MHC sequence is plotted (rough line). The estimated errors in boundary determination were all below 500 bp.

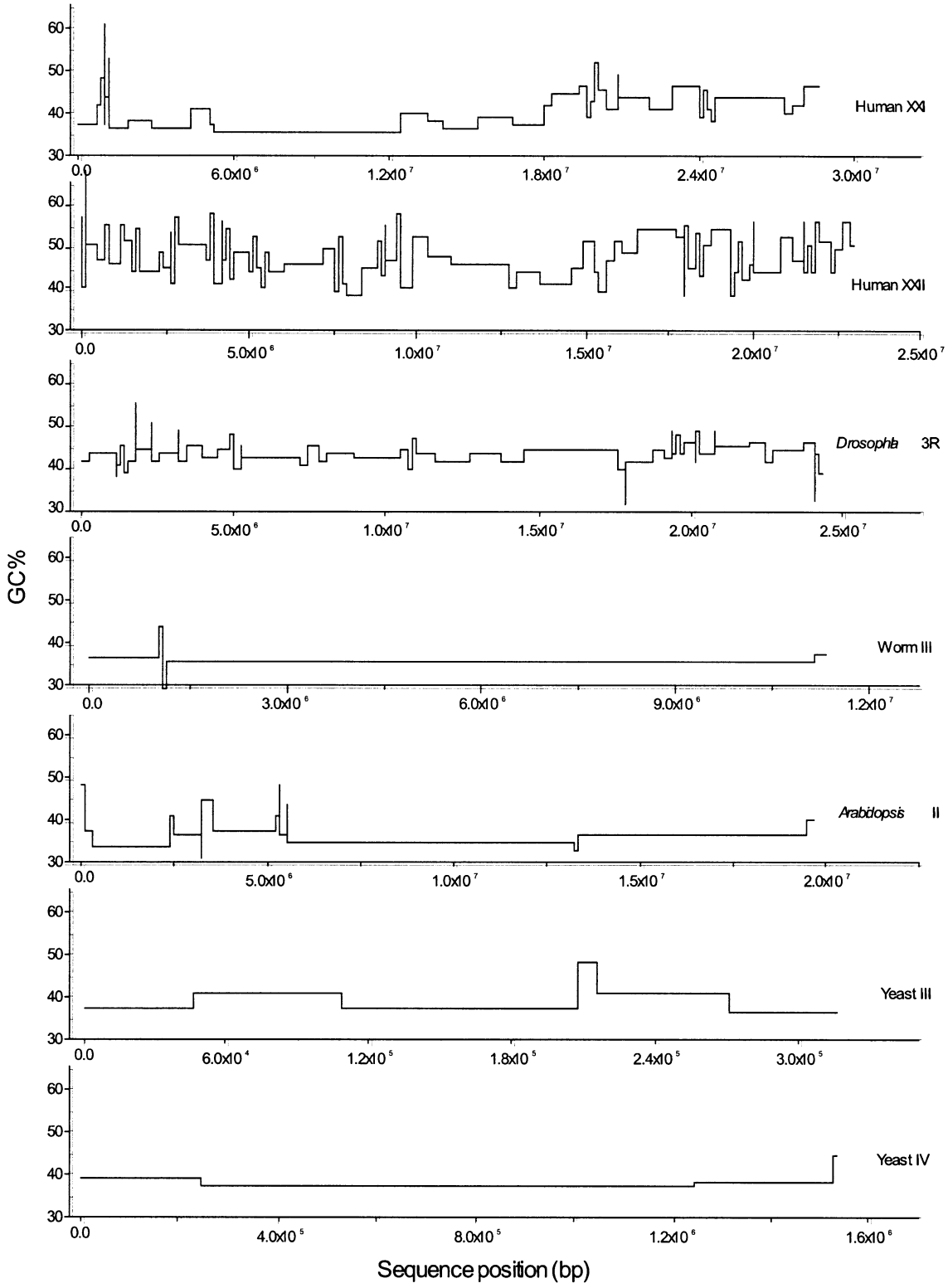

Fig. 3. Isochore maps of some representative eukaryotic chromosomes. For human chromosomes 21 and 22, the largest available contigs were used. Note that the horizontal scale is different for each map.

probably embedded within larger LHGRs with lengths exceeding the contig limits. Second, in some other contigs, only one cut was given. These unique cuts probably correspond to LHGR boundaries, but the flanking regions may also be considered as incomplete LHGRs. Third, in the remaining contigs, two or more cuts were given. In these

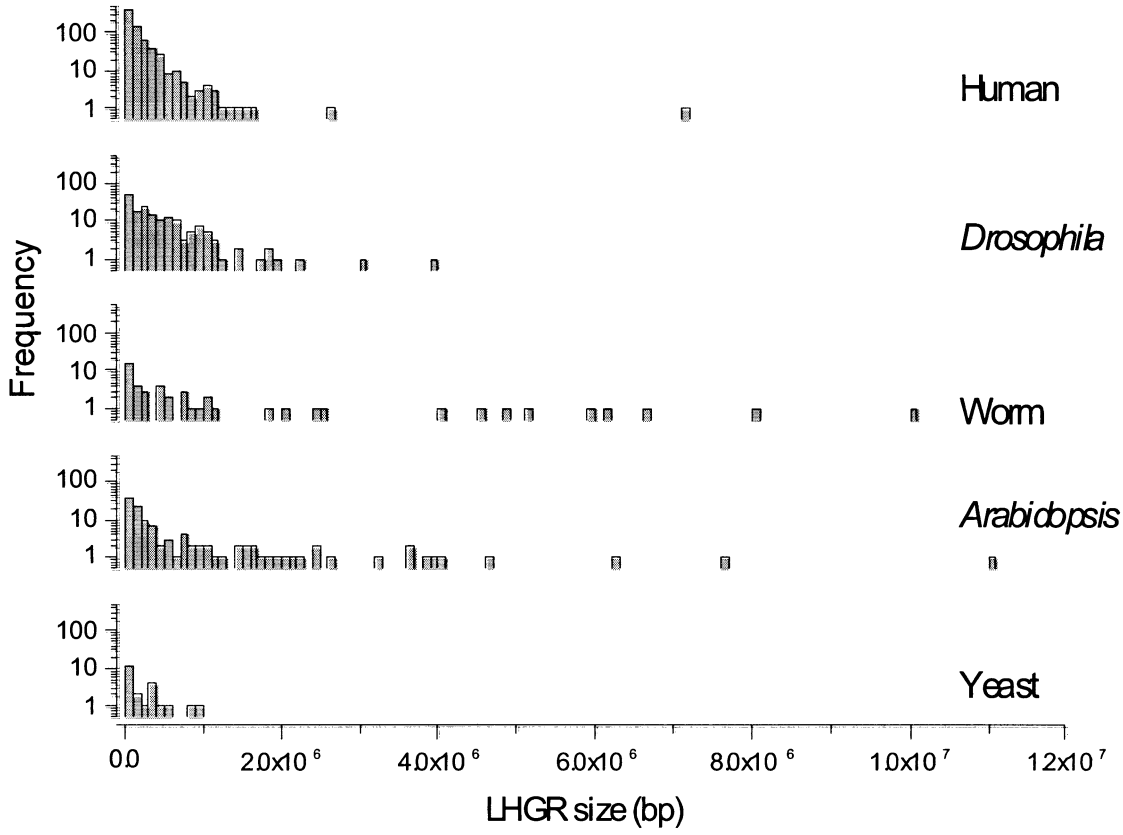

Fig. 4. Size distribution of the LHGRs found by segmenting different genomes.

cases, the ends of the contigs chop off part of the 'outer' LHGRs, and the complete LHGRs may be only the internal LHGRs, the boundaries of which lie within the contig.

Incomplete LHGRs were excluded from further analyses. This left 771 complete internal LHGRs, which may be tentatively considered as isochore-like regions or putative isochores. Human LHGRs show a strongly skewed size distribution (Fig. 4). A statistical analysis of these regions is shown in Table 1.

3.3. Isochore-like regions in other eukaryotic genomes

Isochore chromosome maps were also drawn for other completely sequenced eukaryotic genomes: *D. melanogaster*, worm (*C. elegans*), *Arabidopsis* and yeast (*S. cerevisiae*). A representative sample of isochore chromosome maps is shown in Fig. 3 and a summary of the statistics is given in Table 1. Detailed segmentation results for all the completely sequenced eukaryotic chromosomes can be found at our website.

A highly variegated compositional structure was found in *Drosophila* chromosomes, in agreement with a recent study (Jabbari and Bernardi, 2000) pointing to the compositional compartmentalization of this genome.

The largest LHGRs were found in the worm and *Arabidopsis* genomes (Fig. 4). Both the central part of worm chromosome III and the *Arabidopsis* chromosome I harbor the largest homogeneous regions (more than 10 Mb) found so far.

The isochore-like structure obtained for yeast chromo-

Table 1
Summary statistics of the LHGRs found in different genomes

Genome	*N*	Size (kb)			GC% level			GC% differences between adjacent LHGRs		
		Mean	Min.	Max.	Mean	Min.	Max.	Mean	Min.	Max.
Human	771	165	4	7104	44.3	29.7	68.0	7.2	1.4	23.0
Drosophila	172	444	4	3979	43.5	31.0	57.6	5.6	1.1	18.6
Worm	50	1528	16	10,006	37.5	29.7	68.3	5.3	0.5	30.4
Arabidopsis	121	863	6	11,041	36.6	22.7	47.9	6.1	0.7	13.7
Yeast	23	227	5	999	39.3	34.9	47.7	4.0	0.7	11.0

some 3 (Fig. 3) resembles the regional base composition variation described by Sharp and Lloyd (1993) (see also Bradnam et al., 1999) on the basis of gene GC content, although our segmentation algorithm identified an additional region of 8 kb between positions 206,888 and 214,873 harboring ten extremely GC-rich genes. Yeast chromosomes 7 and 15 remained unsegmented, while the short terminal regions of a few kilobases found in some other yeast chromosomes may correspond to telomeric or subtelomeric regions. Thus, striking differences appear in the long-range patterns shown by different yeast chromosomes, in strong contrast to the inter-chromosomal homogeneity at shorter length scales previously reported by our group (Li et al., 1998).

3.4. GC range and isochore families

The range of GC levels we found for human LHGRs (29.7–68.0%), with the modal class being between 40 and 45% GC for a bin size of 5% GC, agrees with the 30–60% range found experimentally (Bernardi, 1995). We classified the 771 human LHGRs into compositional families on the basis of their respective GC content (Zoubak et al., 1996). The relative amounts of DNA in L, H1, H2 and H3 putative isochore families within the GenBank human genome sample (Fig. 5) were similar to the proportions experimentally found in the entire human genome by DNA centrifugation (62.9, 24.3, 7.5 and 4.7%, respectively; Zoubak et al., 1996).

We used the same GC level classification to partition the LHGRs in the remaining eukaryotic genomes analyzed here (Fig. 5). The isochores in the GC range corresponding to

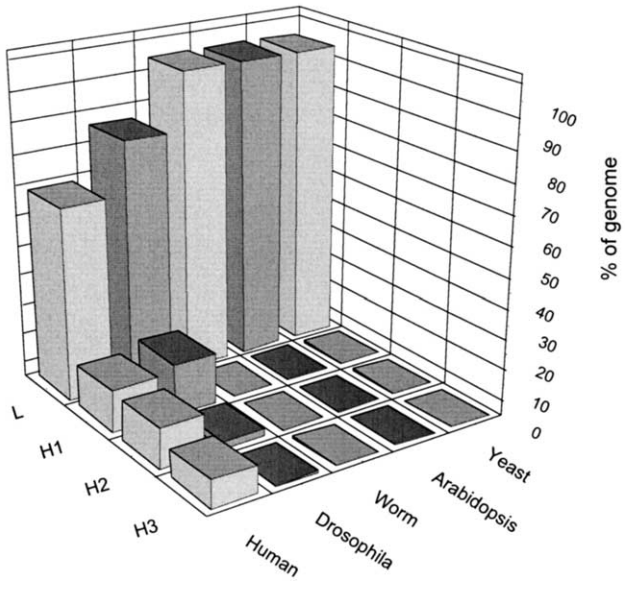

Fig. 5. The distribution of compositional classes in the different genomes. The GC ranges of the four compositional classes shown are those of the L, H1, H2, and H3 isochore families in the human genome (Zoubak et al., 1996).

that of the L family in human predominate by far in all the genomes; these GC-poor regions constitute 60–80% of the human or *Drosophila* genomes, but reach almost 100% in the worm, *Arabidopsis* or yeast genomes. In human, the H1, H2 and H3 isochore classes occupy a substantial fraction, while in *Drosophila* the GC-richest class comprises only 0.13% of the genome.

3.5. Isochore size and GC content

The 771 human LHGRs show a heavily skewed distribution (Fig. 4); the average size for the entire collection was 165 kb (Table 1), but reached 240 kb when only the 504 LHGRs larger than 50 kb were taken into account. Some of the shorter LHGRs may correspond to CpG islands, GC-rich repetitive elements, transposons, etc. The largest human LHGR (7.1 Mb) was in human chromosome 21, its coordinates matching the 'gene desert' of 7 Mb with very low GC content (35%) and also with a paucity of both *Alu* sequences and genes (Hattori et al., 2000).

The large sizes we found for human LHGRs are consistent with the experimental findings of some very long human isochores, such as those in the dystrophin gene (Bettecken et al., 1992) or the cystic fibrosis locus (Krane et al., 1991), as well as with recent estimates of several megabases for the largest size of human isochores (De Sario et al., 1996; Bernardi, 2000). It should be emphasized that the mean size we found for human LHGRs is surely a conservative estimate, since it may be constrained by the shorter lengths of the contigs retrieved from GenBank. In fact, when the longer contigs in the human genome draft sequence of the University of California at Santa Cruz were analyzed, a larger average length resulted.

The average LHGR size varied in the different compositional families (Table 2). GC-poor isochores with the GC levels of L isochores in human showed the largest sizes in all the genomes, while GC-rich isochores proved to be the shortest ones, thus confirming previous observations in the isochores identified by DNA centrifugation (Bettecken et al., 1992; Pilia et al., 1993; De Sario et al., 1996, 1997). Nevertheless, the size differences between isochores in the

Table 2
Average size of LHGR classes in different genomes

Genome	*N*	Size (kb)				L vs. H comparison (corresponding to L vs. H1 + H2 + H3 in human)	
		L	H1	H2	H3	*t*	*P*
Human	771	193	169	129	99	2.54	≤ 0.02
Drosophila	172	588	348	71	15	4.59	≤ 0.00001
Worm	50	1622	39	–	52	1.11	0.27
Arabidopsis	121	929	75	12	–	1.61	0.11
Yeast	23	248	5	8	–	1.20	0.24

GC classes of L and H proved statistically significant only in human and *Drosophila*, but not in the remaining genomes.

3.6. Variations in gene density

Gene distribution in vertebrate genomes is strikingly non-uniform, gene concentration increasing from a very low average level in L isochores to a 20-fold higher level in H3 isochores (Bernardi et al., 1985; Mouchiroud et al., 1991; Zoubak et al., 1996; Bernardi, 2000). Recently, the relative strength of this correlation has been questioned (Venter et al., 2001), since a higher proportion of genes seem located in the GC-poor regions than had been expected. However, when our segmentation algorithm is used to precisely draw up the boundaries of the regions with different composition, a close relationship emerges between GC content and gene density, as observed by Bernardi and coworkers. Fig. 6 illustrates the relationship between LHGR GC content and gene density (number of genes per kilobase) in the MHC region ($r = 0.86$, $P \leq 0.03$). A close relationship was also found in the largest contig of human chromosome 21 ($r = 0.72$, $P \leq 10^{-4}$), as well as in the entire genome of *Drosophila* ($r = 0.41$, $P \leq 10^{-6}$). The latter result agrees with both the observation of a substantial variation in gene density in *Drosophila* (Adams et al., 2000) and the finding that gene concentration increases with increasing GC of the regions embedding the genes (Jabbari and Bernardi, 2000).

However, and in agreement with the fairly constant gene density observed across its chromosomes (The *C. elegans* Sequencing Consortium, 1998), no relationship between gene density and GC level was found in the worm genome. The same occurs in *Arabidopsis* chromosomes II and IV, despite that a correlation, although weak, has been previously reported in this genome (Carels and Bernardi, 2000).

Lastly, a striking situation occurs in yeast: no relationship was found when the overall genome was considered, but a strong positive relationship appeared in chromosome 3 when analyzed separately ($r = 0.99$, $P < 10^{-4}$). This may explain some of the conflicting results previously reported for this genome (Dujon, 1996; Bradnam et al., 1999).

Table 3
Number (%) of neighborhoods among 771 human LHGRs

	L	H1	H2	H3
L	154 (29)			
H1	81 (15)	–		
H2	108 (20)	24 (5)	2 (0.4)	
H3	75 (14)	33 (6)	43 (8)	8 (2)

3.7. GC discontinuities and isochore neighborhood

The last three columns in Table 1 show the discontinuities in GC content found between adjacent LHGRs. In the human genome these range from 1.4% (between a pair of adjacent L isochores) to 23% (corresponding to an L-H3 boundary), the average discontinuity being 7.2%. Similar observations can be made for the different genomes.

We have also compiled the frequencies of the different neighborhoods among human LHGRs (Table 3). No restrictions appear to exist on the isochore classes that can be neighbors on the human chromosomes, and thus any two isochores can be adjacent to one another. In fact, neighborhood frequencies seem to depend only on the relative abundance of each isochore family. Thus, the higher frequencies are for neighborhoods in which L isochores are involved, and the lower ones are for HH isochore boundaries.

4. Discussion

We have shown here that an improved version of the entropic segmentation algorithm, specifically designed to determine the most statistically significant partition of a DNA sequence at each scale, was able to precisely identify the boundaries of the two isochores experimentally determined and characterized to date, the class II and class III regions of the human MHC region. In contrast, the algorithm of Nekru-

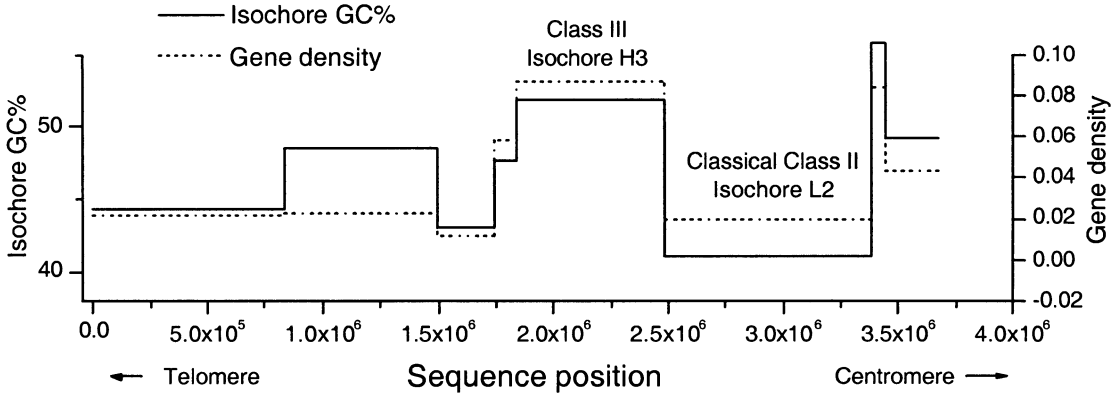

Fig. 6. Relationship between gene density (number of genes by kilobase) and the LHGR G + C content in the MHC region ($r = 0.86$, $P \leq 0.03$).

tenko and Li (2000) only finds a series of overlapping fragments in this region, none of which correspond to class II or class III. Our segmentation algorithm also appears to be successful when applied to a wide collection of large human chromosome contigs, identifying LHGRs that show the typical lengths and compositional heterogeneities of isochores. The relatively homogeneous regions we found show many of the features (G + C range, proportion of isochore classes, size distribution, and relationship with gene density) of the isochores identified through the centrifugation of vertebrate DNA fragments. Lastly, the method was also able to find isochore-like regions in other eukaryotic genomes. We conclude that the improved segmentation algorithm presented here may truly identify isochore boundaries in the long genomic sequences now being generated by large-scale sequencing projects.

The computational prescreening of isochore boundaries may have many applications in genomics. The changes in replication timing known to occur at isochore boundaries (Tenzen et al., 1997) could be exhaustively investigated at such computationally identified boundaries. LHGR prescreening would also be useful in searching for gene-rich genome regions, as we found here that gene density closely varies with the GC content of the LHGRs in both the human and the *Drosophila* genomes. In this way, regions for laborious experimental procedures such as exon trapping could be prioritized. Another use of the present approach may be in the field of computational gene identification. It is known that the performance of gene-finding programs may be improved by taking into account isochore compositional properties (Burge and Karlin, 1997). Isochore prescreening of the very large genomic sequences with which such programs are now being faced may help in choosing the appropriate set of compositional parameters for each segment. The analysis of the distribution across the genome of the different families of repeat sequences and transposable elements could also benefit from the isochore maps presented here. Finally, isochore chromosome maps may also be useful in comparative genomics. The comparison of genomes above the gene level is usually restricted to the study of chromosome-band patterns. The comparative analysis of isochore chromosome maps may now allow new insights in the field.

Two processes have been invoked to account for the evolutionary origin of isochores: mutational pressure and natural selection (for recent reviews on this topic, see Bernardi, 2000; Gautier, 2000). This subject lies beyond the scope of the present paper, but the detailed compositional maps presented here may be the first step in clarifying the functional and evolutionary interpretations of genome heterogeneity.

Acknowledgements

Special thanks are due to Oliver K. Clay, who significantly contributed with many discussions and suggestions to improve this work. We also acknowledge the critical reading of Manuel Ruiz Rejón and Antonio Marín. The help with the manuscript of David Nesbitt is also appreciated. This work is supported by grant BIO99-0651-CO2-01 from the Spanish Government.

References

Adams, M.D., et al., 2000. The genome sequence of *Drosophila melanogaster*. Science 287, 2185–2195.

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

Bernaola-Galván, P., Oliver, J.L., Román-Roldán, R., 1999. Decomposition of DNA sequence complexity. Phys. Rev. Lett. 83, 3336–3339.

Bernaola-Galván, P., Grosse, I., Carpena, P., Oliver, J.L., Román-Roldán, R., Stanley, H.E., 2000. Finding borders between coding and noncoding regions by an entropic segmentation method. Phys. Rev. Lett. 85, 1342–1345.

Bernaola-Galván, P., Grosse, I., Carpena, P., Román-Roldán, R., Oliver, J.L., Stanley, H.E., 2001. Analysis of symbolic sequences using the Jensen-Shannon divergence measure. Phys. Rev. E submitted.

Bernardi, G., 1995. The human genome: organization and evolutionary history. Annu. Rev. Genet. 29, 445–476.

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

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

Bettecken, T., Aissani, B., Müller, C.R., Bernardi, G., 1992. Compositional mapping of the human dystrophin-encoding gene. Gene 122, 329–335.

Bradnam, K.R., Seoighe, C., Sharp, P.M., Wolfe, K.H., 1999. G + C content variation along and among *Saccharomyces cerevisiae* chromosomes. Mol. Biol. Evol. 16, 666–675.

Burge, C., Karlin, S., 1997. Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 268, 8–94.

Carels, N., Bernardi, G., 2000. The compositional organization and the expression of the *Arabidopsis* genome. FEBS Lett. 472, 302–306.

Churchill, G.A., 1989. Stochastic models for heterogeneous DNA sequences. Bull. Math. Biol. 51, 79–94.

De Sario, A., Geigl, E.-M., Palmieri, G., D'Urso, M., Bernardi, G., 1996. A compositional map of human chromosome band Xq28. Proc. Natl. Acad. Sci. USA 93, 1298–1302.

De Sario, A., Roizes, G., Allegre, N., Bernardi, G., 1997. A compositional map of the cen-q21 region of human chromosome 21. Gene 194, 107–113.

Dujon, B., 1996. The yeast genome project: What did we learn? Trends Genet. 12, 263–270.

Dunham, I., Shimizu, N., Roe, B.A., Chissoe, S., et al., 1999. The DNA sequence of human chromosome 22. Nature 402, 489–495.

Frank, A.C., Lobry, J.R., 2000. Oriloc: prediction of replication boundaries in unannotated bacterial chromosomes. Bioinformatics 16, 560–561.

Fukagawa, T., Sugaya, K., Matsumoto, K., Okumura, K., Ando, A., Inoko, H., Ikemura, T., 1995. A boundary of long-range G + C% mosaic domains in the human MHC locus: pseudoautosomal boundary-like sequence exists near the boundary. Genomics 25, 184–191.

Fukagawa, T., Nakamura, Y., Okumura, K., Nogami, M., Ando, A., Inoko, H., Saitou, N., Ikemura, T., 1996. Human pseudoautosomal boundary-like sequences: expression and involvement in evolutionary formation of the present-day pseudoautosomal boundary of human sex chromosomes. Hum. Mol. Genet. 5, 123–132.

Gardiner, K., Aissani, B., Bernardi, G., 1990. A compositional map of human chromosome 21. EMBO J. 9, 1853–1858.

Gautier, C., 2000. Compositional bias in DNA. Curr. Opin. Genet. Dev. 10, 656–661.

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

Hattori, M., et al., 2000. The DNA sequence of human chromosome 21. Nature 405, 311–319.

Jabbari, K., Bernardi, G., 2000. The distribution of genes in the *Drosophila* genome. Gene 247, 287–292.

Krane, D.E., Hartl, D.L., Ochman, H., 1991. Rapid determination of nucleotide content and its application to the study of genome structure. Nucleic Acids Res. 19, 5181–5185.

Lander, E.S., Waterston, R.H., Sulston, J., et al. International Human Genome Sequencing Consortium, 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.

Li, W., Kaneko, K., 1992. Long range correlation and partial $1/f^{\alpha}$ spectrum in a noncoding DNA sequence. Europhys. Lett. 17, 655.

Li, W., Stolovitzky, G., Bernaola-Galván, P., Oliver, J.L., 1998. Compositional heterogeneity within, and uniformity between, DNA sequences of yeast chromosomes. Genome Res. 8, 916–928.

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

Mouchiroud, D., D'Onofrio, G., Aissani, B., Macaya, G., Gautier, C., Bernardi, G., 1991. The distribution of genes in the human genome. Gene 100, 181–187.

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

Oliver, J.L., Román-Roldán, R., Pérez, J., Bernaola-Galván, P., 1999. SEGMENT: identifying compositional domains in DNA sequences. Bioinformatics 15, 974–979.

Peng, C.-K., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Sciortino, F., Simons, M., Stanley, H.E., 1992. Long-range correlations in nucleotide sequences. Nature 356, 168–170.

Pilia, G., Little, R.D., Aissani, B., Bernardi, G., Schlessinger, D., 1993. Isochores and CpG islands in YAC contigs in human X26.1-qter. Genomics 17, 456–462.

Ramensky, V., Makeev, V., Roytberg, M., Tumanyan, V., 2000. DNA segmentation through the Bayesian approach. J. Comp. Biol. 7, 215–231.

Román-Roldán, R., Bernaola-Galván, P., Oliver, J.L., 1998. Sequence compositional complexity of DNA through an Entropic segmentation method. Phys. Rev. Lett. 80, 1344–1347.

Román-Roldán, R., et al., 2001. Information-theoretic symbolic sequence segmentation by maximum discrepancy ordering, manuscript in preparation.

Sharp, P.M., Lloyd, A.T., 1993. Regional base composition variation along yeast chromosome III: evolution of chromosome primary structure. Nucleic Acids Res. 21, 179–183.

Stephens, R., Horton, R., Humphray, S., Rowen, L., Trowsdale, J., Beck, S., 1999. Gene organisation, sequence variation and isochore structure at the centromeric boundary of the human MHC. J. Mol. Biol. 291, 789–799.

Tenzen, T., Yamagata, T., Fukagawa, T., Sugaya, K., Ando, A., Inoko, H., Gojobori, T., Fujiyama, A., Okumura, K., Ikemura, T., 1997. Precise switching of DNA replication timing in the GC content transition area in the human MHC. Mol. Cell. Biol. 17, 4043–4050.

The *C. elegans* Sequencing Consortium, 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science 282, 2012–2018.

The MHC Sequencing Consortium, 1999. Complete sequence and gene map of a human major histocompatibility complex. Nature 401, 921–923.

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

Venter, J.C., Adams, M.D., Myers, E.W., et al., 2001. The sequence of the human genome. Science 291, 1304–1351.

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