A compact view of isochores in the draft human genome sequence

Adam Pavlíček[a,b], Jan Pačes[b,c], Oliver Clay[d], Giorgio Bernardi[a,d,]*

[a]*Laboratoire de Génétique Moléculaire, Institut Jacques Monod, 2 place Jussieu, 75005 Paris, France*
[b]*Institute of Molecular Genetics, Academy of Sciences of the Czech Republic, Flemingovo 2, Prague CZ-16637, Czech Republic*
[c]*Center for Integrated Genomics, Flemingovo 2, Prague CZ-16637, Czech Republic*
[d]*Laboratory of Molecular Evolution, Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy*

Received 2 October 2001; revised 28 November 2001; accepted 3 December 2001

First published online 2 January 2002

Edited by Takashi Gojobori

Abstract Prior to genome sequencing, information on base composition (GC level) and its variation in mammalian genomes could be obtained using density gradient ultracentrifugation. Analyses using this approach led to the conclusion that mammalian genomes are organized into mosaics of fairly homogeneous regions, called isochores. We present an initial compositional overview of the chromosomes of the recently available draft human genome sequence, in the form of color-coded moving window plots and corresponding GC level histograms. Results obtained from the draft human genome sequence agree well with those obtained or deduced earlier from CsCl experiments. The draft sequence now permits the visualization of the mosaic organization of the human genome at the DNA sequence level. © 2002 Federation of European Biochemical Societies. Published by Elsevier Science B.V. All rights reserved.

Key words: Genome organization; Mammalian DNA; Analytical ultracentrifugation; Compositional homogeneity

1. Introduction

The draft human genome sequence permits, for the first time at the DNA sequence level, the visualization of the mosaic organization of mammalian genomes that was deduced 25 years ago from ultracentrifugation experiments [1]. Moreover, the mosaicism is seen to be characterized by fairly constant average GC levels (GC is the molar fraction of guanine and cytosine in DNA) persisting over long distances, and by abrupt jumps to higher or lower GC levels. Such a mosaic organization is at variance with an alternative hypothesis that has been considered by some authors [2,3], in which GC levels drift more or less continuously throughout the human genome. Other proposals in the recent literature that are refuted by the draft sequence and by ultracentrifugation experiments are discussed elsewhere [4,5].

2. Results and discussion

Fig. 1 shows a color-coded compositional map of the human chromosomes, representing 100 kb moving window plots that scan the recently published draft human genome sequence [3]. Color codes span the spectrum of GC levels in five steps, from ultramarine blue (GC-poorest isochores) to scarlet red (GC-richest isochores). In all chromosomes except for the previously sequenced chromosomes 21 and 22 [6,7], there is a somewhat startling abundance of gaps (gray bars) in the euchromatic regions, amounting to about 5000 gaps that have been estimated to cover nearly 300 Mb, i.e. nearly one tenth of the genome (not counting the many shorter sequence gaps or the heterochromatic regions [3]). Apart from such gaps, three features of the map are prominent. The most striking feature is the large proportion of the genome represented by long blue regions, uninterrupted by red. The next most notable observation is the scarcity of blue in many of the blocks characterized primarily by orange and/or yellow, or by red and/or orange, and the tendency of such regions to be much shorter than the expanses of uninterrupted blue. The longer GC-poor regions are in agreement with the much greater abundance of GC-poor isochore DNA in the human genome (63%; see [8], and references therein). The third observation is that compositional fluctuations increase as one moves from GC-poor to GC-rich isochores. Short (∼200 kb) GC-rich isochores, and a higher heterogeneity of GC levels in long, predominantly GC-rich regions, had also been observed earlier during yeast artificial chromosome compositional mapping of chromosomal bands [9,10]; the higher heterogeneity within GC-rich isochores had already been inferred and quantified previously via CsCl analyses [11].

Although fixed-size moving window plots that show GC level variation along a chromosome have well-known drawbacks (discussed in [12,13]) and, in particular, cannot show variation at scales smaller than the windows' lengths, they are conceptually very simple, and can be easily reproduced and explored using traditional software. Discrete color coding of windows according to non-overlapping GC ranges in which they are located, when used together with a line plot of the GC level, can quickly give an intuitive overview of the large-scale compositional homogeneities and heterogeneities in a chromosome. In order to do so, however, the window size must be large (100–300 kb), and the vertical scale must be large in relation to the horizontal scale. Such conditions, which are often not met in published GC level plots, are essential, or the compositional mosaicism of the chromosomes at the isochore level will not be easy to recognize. A particular advantage of the fixed-size window method is that histograms of the windows' GC values can be obtained either from DNA sequences or, experimentally, via CsCl density gradient centrifugation, a method for rapidly characterizing genomes that was in use already before the discovery of the genetic code, and well before the advent of genomic sequencing.

*Corresponding author. Fax: (39)-081-245 5807.
E-mail address: bernardi@alpha.szn.it (G. Bernardi).

Abbreviations: GC, molar ratio of G+C in DNA

PII: S0014-5793(01)03283-5

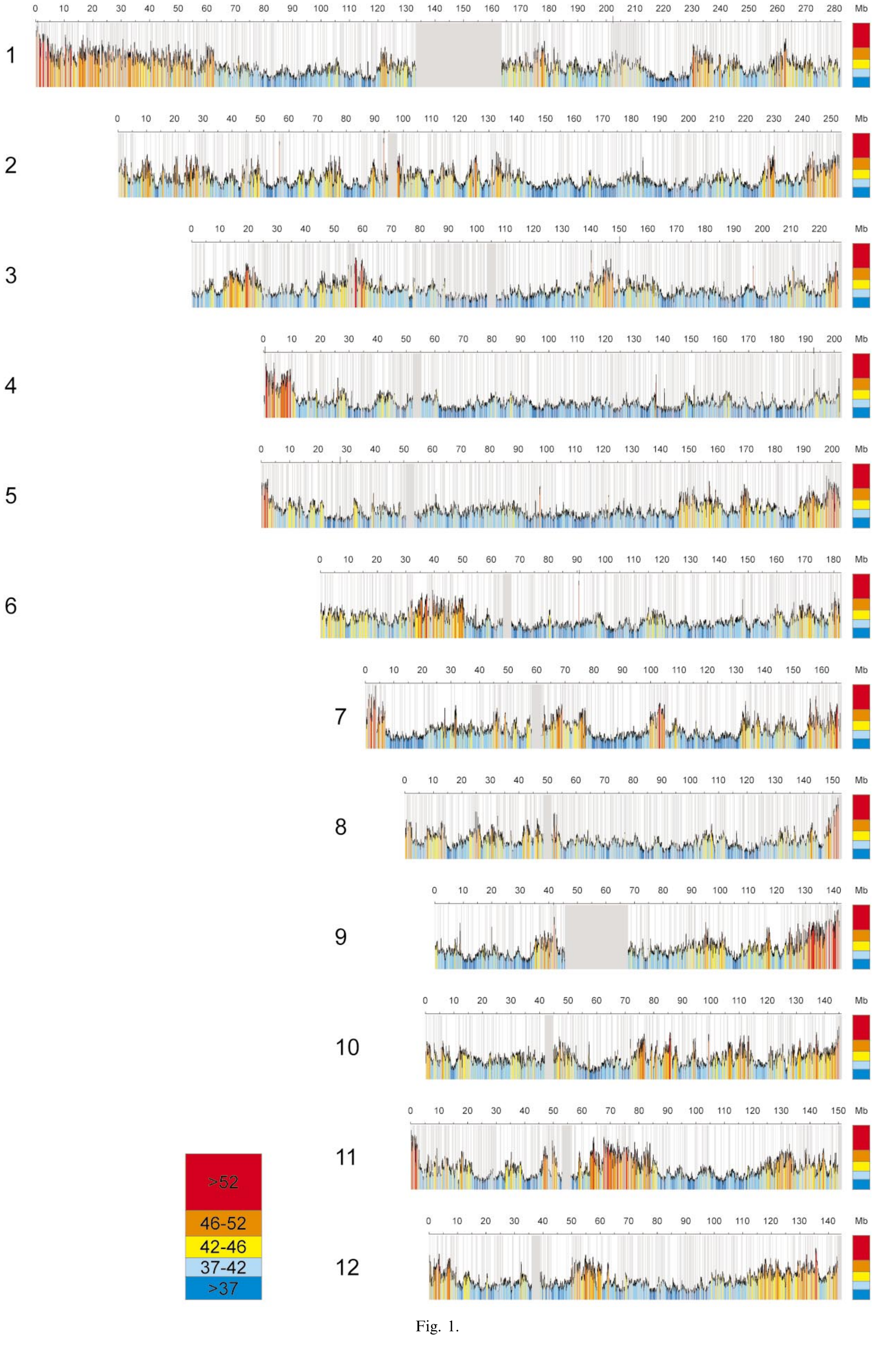
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250 260 270 280 Mb
1
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250 Mb
2
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 Mb
3
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 Mb
4
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 Mb
5
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 Mb
6
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 Mb
7
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 Mb
8
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 Mb
9
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 Mb
10
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 Mb
11
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 Mb
12
>52
46-52
42-46
37-42
>37

Fig. 1.

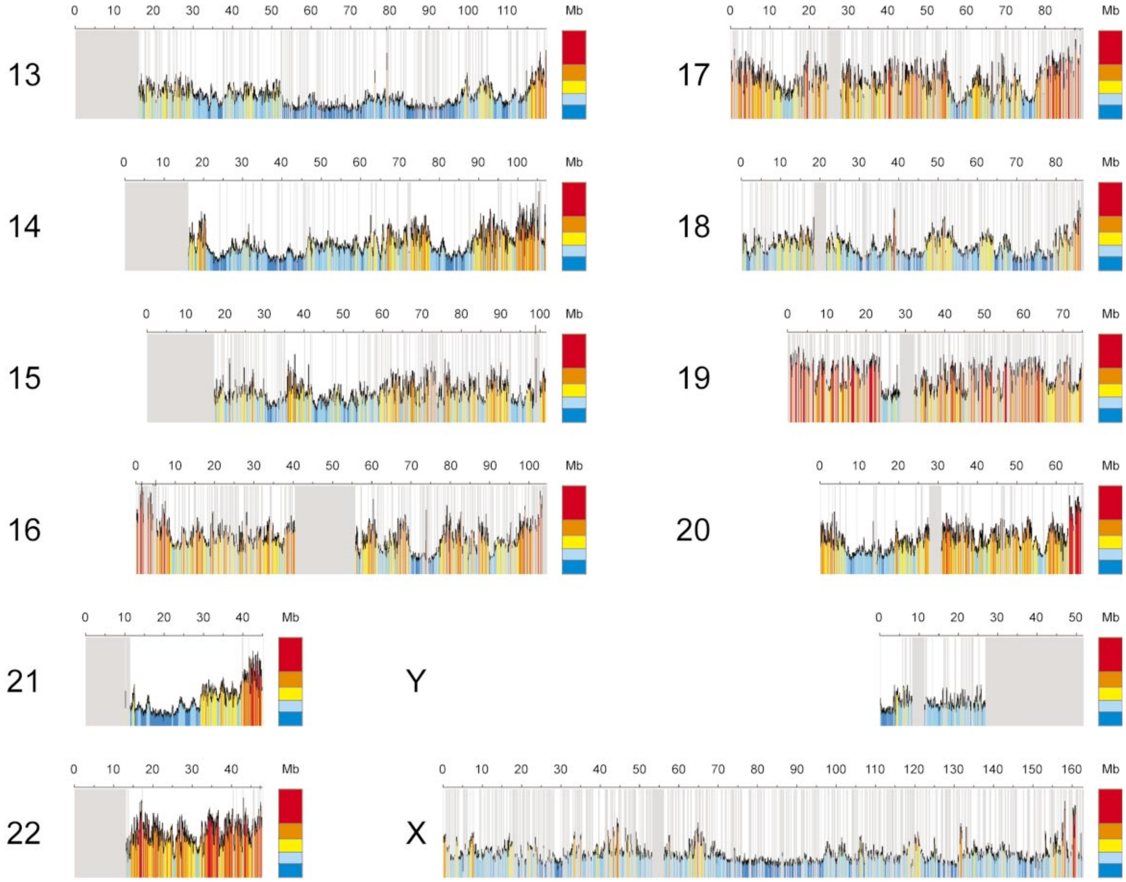

Fig. 1. (*continued*). Color-coded moving window plots, using overlapping 100 kb windows, of the chromosomes of the draft human genome sequence (from the draft assembly [14] described in [3]). The 100 kb windows were partitioned into five classes, representing the five major DNA components and the isochore families from which they derive (L1: <37% GC, L2: 37–42% GC, H1: 42–46% GC, H2: 46–52% GC, H3: >52% GC; see color bar). Gray vertical bars show gaps in the draft sequence; wide gray blocks show centromeric, telomeric and/or heterochromatic DNA not represented in the draft sequence.

Isochores were defined in 1976 [1], and their name introduced in 1981 [11], as long regions of DNA, initially estimated as much longer than 300 kb on average, that are 'fairly homogeneous' in base composition (i.e. in GC level), compared to the heterogeneity present in the main-band (non-satellite) DNA of a mammalian genome. The lower bound of around 300 kb for the length of most isochores was the highest estimate possible using the CsCl methodology at that time, which allowed the comparison of GC distributions at different fragment lengths (corresponding, in the plots of Fig. 1, to window lengths), ranging from a diffusion threshold at around 2 kb to a threshold of possible aggregation, between about 500 kb and 1 Mb.

Fig. 2 shows the GC distributions of the 100 kb windows from Fig. 1, superposed on the corresponding plot when a window length of 300 kb is used. In spite of the threefold difference in window size, the two profiles coincide almost exactly, except for small fluctuations in the 300 kb profile. This is a highly non-trivial behavior of the distributions, first observed 25 years ago: if an isochore organization were not present in most of the genome, the 300 kb distribution would be distinctly narrower than the 100 kb distribution. As expected, both curves match well the CsCl profiles as obtained by analytical ultracentrifugation for these different fragment lengths: the similarity among such CsCl profiles for different

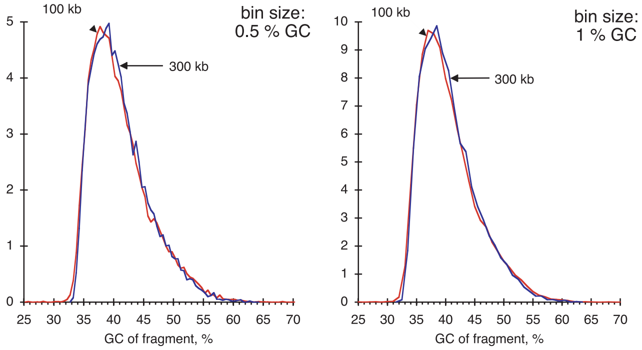

Fig. 2. GC level histograms of the 100 kb windows shown in Fig. 1 (red curves) and, superposed, of 300 kb windows taken over the same draft sequence data (blue curves), for bin sizes of 0.5% GC (left panel) and 1% GC (right panel). The histograms reproduce, using sequence data, the fragment distributions that are obtained experimentally, for these molecular weights, by CsCl density gradient ultracentrifugation. The histograms shown here are for the male (heterogametic, XY) genome; histograms for the female (XX) genome differ only negligibly (mean and standard deviation are approximately 0.6% and 0.03% GC lower than in the male, respectively).

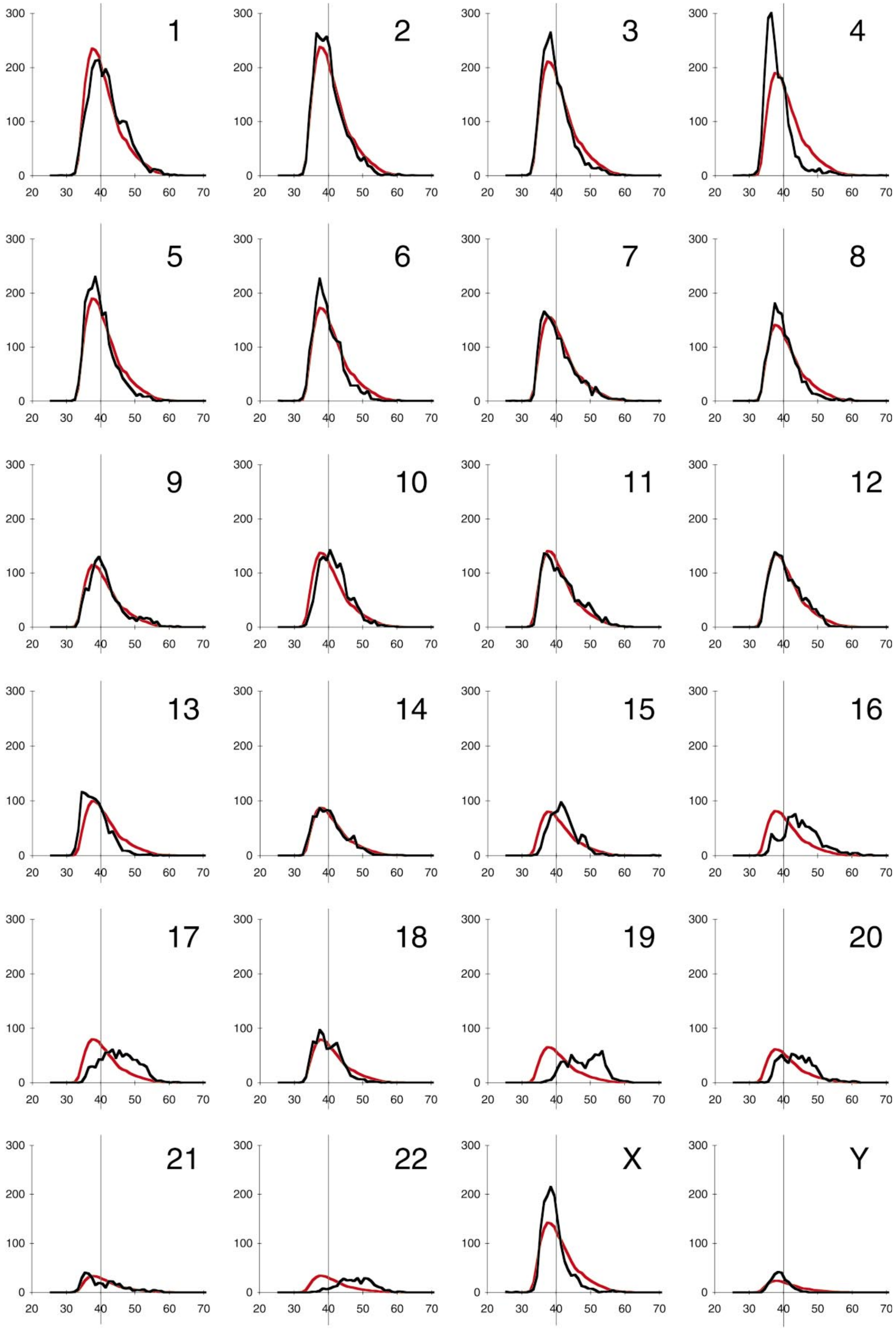

Fig. 3. GC level histograms of the 100 kb windows shown in Fig. 1, for the individual chromosomes of the human genome (black curves) and, superposed, the expected histogram if each chromosome had the same GC distribution as the entire genome (red curves). Bin sizes are 1% GC. Vertical axis: number of windows; horizontal axis: GC, %. Vertical reference bars indicate 40% GC.

molecular weights was an early, decisive indication that mammalian genomes are organized into isochores [1].

Given the draft sequence, we can now go further, and obtain the GC distribution of the euchromatic DNA in each individual chromosome. As can be seen in Fig. 3, not all chromosomes show similar GC distributions (black curves, shown here at the 100 kb level): several chromosomes deviate strongly from the distribution of the whole genome (red curves), in that either GC-poor or GC-rich DNA is underrepresented. In Fig. 1, the former case corresponds to a scarcity of blue, the latter case to an absence of red and orange in the moving-window plot. Such chromosomes are, as a simple consequence of their biased GC distributions, not prone to show the dramatic jumps in GC level that are characteristic of the chromosomes containing both GC-rich and GC-poor euchromatic DNA in normal proportions.

As a result of the strong asymmetry of the CsCl profile of human DNA, i.e. of the GC distributions of DNA fragments of lengths in the range from 70 to over 300 kb, it is inevitable that GC-poor regions will tend to be much longer than GC-rich regions, as is observed in the plots of Fig. 1. The inevitability of long GC-poor regions can be easily seen by using 24 cards to represent regions of DNA of a given length (e.g. of 300 kb), and then considering their possible configurations. Thus, if 15 cards represent the GC-poor L DNA (= L1 and L2; 62.5%; 13 spades/two jokers), six represent H1 (25%; clubs), two represent H2 (8.3%; hearts) and one represents the GC-richest, H3 DNA (4.2%; diamond), corresponding approximately to their ratios in the human genome [8], then many of the 15 cards representing L regions must inevitably occur in groups of two or more (the exact number will be between six, as in LHLHLHLLLLLLHLHLHLHLHLHL, and 15). In other words, much if not most of the L DNA will occur in long isochores.

The practical utility of the isochore concept and the functional relevance of compositional maps follow from the documented correlates of base composition. For example, in Fig. 1 the differences in GC among the long, chromatically homogeneous regions of the chromosomes correspond to differences in gene density, the GC-richest (red) isochores having much higher gene densities than the GC-poorest (blue) isochores, and to differences in intron length, patterns of codon usage and distributions of repetitive elements (see [8] for a review). In many regions of the genome, isochore borders can be recognized already from obvious changes in the GC level of the sequence, or by using appropriate segmentation algorithms [12,13]. In other regions, more work is still needed, which will include fine-tuning of algorithms, and taking into account the known systematic differences in GC level that correspond to genes and to other structural and functional features. Part of this work will involve a detailed study of the very few long, contiguously sequenced regions of the human genome that are accompanied by publicly available, dense annotations for experimentally verified features such as genes, promoters/enhancers, matrix/scaffold attachment regions, origins of replication, or other replication information, which could give further clues to the functional roles of the isochore organization of mammalian genomes.

References

[1] Macaya, G., Thiery, J.P. and Bernardi, G. (1976) J. Mol. Biol. 108, 237–254.
[2] Fickett, J.W., Torney, D.C. and Wolf, D.R. (1992) Genomics 13, 1056–1064.
[3] IHGSC (International Human Genome Sequencing Consortium) (2001) Nature 409, 860–921.
[4] Clay, O. and Bernardi, G. (2001) Gene 276, 25–31.
[5] Clay, O. and Bernardi, G. (2001) Biochem. Biophys. Res. Commun. 285, 855–856.
[6] Hattori, M., Fujiyama, A. and Taylor, T.D. et al. (2000) Nature 405, 311–319.
[7] Dunham, I., Shimizu, N. and Roe, B.A. et al. (1999) Nature 402, 489–495.
[8] Bernardi, G. (2000) Gene 241, 3–17.
[9] De Sario, A., Geigl, E.-M., Palmieri, G., D'Urso, M. and Bernardi, G. (1996) Proc. Natl. Acad. Sci. USA 93, 1298–1302.
[10] De Sario, A., Roizes, G., Allegre, N. and Bernardi, G. (1997) Gene 194, 107–113.
[11] Cuny, G., Soriano, P., Macaya, G. and Bernardi, G. (1981) Eur. J. Biochem. 115, 227–233.
[12] Oliver, J.L., Bernaola-Galván, P., Carpena, P. and Román-Roldán, R. (2001) Gene 276, 47–56.
[13] Li, W. (2001) Gene 276, 57–72.
[14] Haussler, D. et al. (2000) Human Genome Working Draft (URL: http://genome.ucsc.edu).