Available online at www.sciencedirect.com
SCIENCE DIRECT®
Gene 333 (2004) 135–141
GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES
www.elsevier.com/locate/gene

Representing GC variation along eukaryotic chromosomes

Jan Pačes[a], Radek Zíka[a], Václav Pačes[a], Adam Pavlíček[a], Oliver Clay[b], Giorgio Bernardi[b,*]

[a] *Institute of Molecular Genetics, Academy of Sciences of the Czech Republic, Flemingovo 2, Prague CZ-16637, Czech Republic*
[b] *Laboratorio di Evoluzione Molecolare, Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy*

Received 28 August 2003; accepted 10 February 2004

Available online 23 April 2004

Abstract

Genome sequencing now permits direct visual representation, at any scale, of GC heterogeneity along the chromosomes of several higher eukaryotes. Plots can be easily obtained from the chromosomal sequences, yet sequence releases of mammalian or plant chromosomes still tend to use small scales or window sizes that obscure important large-scale compositional features. To faithfully reveal, at one glance, the compositional variation at a given scale, we have devised a simple scheme that combines line plots with color-coded shading of the regions underneath the plots. The scheme can be applied to different eukaryotic genomes to facilitate their comparison, as illustrated here for a sample of chromosomes chosen from seven selected species. As a complement to a previously published compact view of isochores in the human genome sequence [FEBS Lett. 511 (2002a) 165], we include here an analogous map for the recently sequenced mouse genome, and discuss the contribution of repetitive DNA to the GC variation along the plots. Supplementary information, including a database of color-coded GC profiles for all recently sequenced eukaryotes and the program draw_chromosomes_gc.pl used to obtain them, are available at http://genomat.img.cas.cz.

Keywords: Genome organization; Mammalian DNA; Analytical ultracentrifugation; Compositional homogeneity; Software

1. Introduction

Entire genome sequences permit for the first time the graphic portrayal of compositional heterogeneity, at the DNA sequence level, at any desired scale. A portrait or profile of GC level allows one to monitor variation along chromosomes. It can be a powerful tool when comparing different regions of the same genome, when comparing different genomes, or when comparing different draft assemblies of individual chromosomes. Statistical descriptions of the GC variation along chromosomes can be derived directly from CsCl gradient ultracentrifugation of DNA. This principle was used to demonstrate the presence of isochores in mammals well before sequences were available (Macaya et al., 1976). Furthermore, in the absence of sequence information, the large-scale variation of GC along individual chromosomes of vertebrates can be plotted by in situ hybridization of GC fractions taken after preparative ultracentrifugation of the DNA (Saccone et al., 1993, 1996, 2002). The availability of entire chromosome sequences now allows such variation to be displayed almost effortlessly at high resolutions, via fixed-length moving-window plots.

Published plots, accompanying releases of mammalian or plant chromosome sequences or follow-up analyses, still often use excessively small scales and/or window sizes, e.g. in attempts to economize space in journals or online supplements. Vertical scales representing GC level in line plots are sometimes only a few millimeters high with no guidelines, and color-coded GC/isochore tracks are typically considered a replacement of line plots, rather than their complement. Inappropriately small scales or window sizes have, in one or two cases, led authors to question the significance of clear intrachrosomal (inter-isochore) contrasts in relative GC or CpG frequencies (IHGSC, 2001; Gentles and Karlin, 2001) that become obvious (and can be confirmed statistically) as soon as one changes scale. We

Abbreviations: GC, molar fraction of guanine and cytosine in DNA; LINE, long interspersed repeat; SINE, short interspersed repeat; MAR, SAR, matrix, scaffold attachment region.

* Corresponding author. Tel.: +39-81-5833300; fax: +39-81-2455807.
E-mail address: bernardi@alpha.szn.it (G. Bernardi).

0378-1119/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2004.02.041

therefore felt that this technical point needed to be explicitly addressed. With future comparisons in mind, we devised a simple scheme for portraying large-scale variation that can be applied to most, if not all, eukaryotic chromosome sequences. We also used this approach, together with available annotation databases, to graphically summarize the genome-wide compositional variation of repetitive and non-repetitive DNA in human and mouse.

2. Materials and methods

The program draw_chromosome_gc.pl was written in Perl. It requires installation of Perl version 5 (freely available at http://www.perl.org) and the Perl GD module (freely available from http://www.cpan.org). The current version of the program works with GD module versions 1.19 and higher and produces files in the png (Portable Network Graphic) format. The program is freely distributed as source code under General Public License (GLP) and can be downloaded from http://genomat.img.cas.cz.

3. Results and discussion

3.1. GC mosaicism, its visual display, and the uses of compositional maps

Abrupt changes in GC level represent landmarks that naturally partition or calibrate chromosome sequences. They also correlate with key biological properties in many eukaryotes, such as changes in gene density (Mouchiroud et al., 1991; Zoubak et al., 1996; IHGSC, 2001; Venter et al., 2001), switches in replication timing (Tenzen et al., 1997; Stephens et al., 1999; MHCSC, 1999), and differences between the locations of adjacent regions in the interphase nucleus (Saccone et al., 2002; Mahy et al., 2002). Some of these properties have been shown primarily for mammals and birds, whereas other properties are apparently found in a wide range of eukaryotes. Thus, gene density has been found to correlate with GC also in *Drosophila* (Jabbari and Bernardi, 2000; Myers et al., 2000), *Arabidopsis* (Carels and Bernardi, 2000), and in the heterogeneous chromosome 3 of yeast (Oliver et al., 2001).

In addition to their biological relevance, GC level plots are also useful tools that can facilitate mapping, including synteny mapping (Pavlíček et al., 2002b). Window sizes are best determined by the question in which one is interested. Some genomes admit no obvious, intrinsic best choice of window size for representing large-scale GC variation (above the scale of genes). Other genomes such as those of mammals or birds reveal long, fairly homogeneous regions, or isochores, when window sizes exceed about 50 kb, yet when much smaller windows are chosen (e.g., 10–20 kb; IHGSC, 2001), local fluctuations obscure the large-scale structure and the unaided eye can no longer recognize it (Clay et al., 2001; Pavlíček et al., 2002a). In *Arabidopsis* chromosomes, the systematic increase in GC levels towards the telomeres (Carels and Bernardi, 2000) was similarly left undocumented by the primary annotators, presumably because of a small vertical scale and short windows. In the smaller chromosomes of prokaryotes, the variation present in some taxa at scales above genes (<70 kb) can also be recognized by correspondingly sized windows.

To reveal, at one glance, the important compositional features of a chromosome at any given scale, we have devised a simple scheme that combines line plots with color-coded shading of the region underneath the plot. The scheme is chosen so that it can be applied to all eukaryotic genomes and facilitate their comparison. A discrete color palette spans the range of GC levels encountered in eukaryotes, with changes in color every 2.5% GC. We have found that this gradation is a good tradeoff between the two requirements of high resolution and discrete, easily recognizable colors. If one follows this scheme, systematic long-range differences in GC between adjacent regions will typically be picked up by visible color differences. This can be seen in Fig. 1, for example, in *Arabidopsis* chromosome 5 (where the important color change is at 35% GC), or in telomeric regions of *Encephalitozoon cuniculi* (chromosome 1). In the case of mammalian genomes, furthermore, the intervals bounded by 37.5%, 42.5%, 47.5% and 52.5% GC largely indicate the isochore families in which the windows are most likely to be found (Bernardi, 2001 and references therein).

Plotting the GC levels for both 50 and 100 kb windows (e.g., with steps 1/10 of the window size) reveals the large-scale variation present above the genic scale in most higher eukaryotes (Fig. 1a–d). A plot of the human genome using this scheme, after pooling pairs of 2.5% intervals for clarity, is shown for 100 kb windows in Pavlíček et al. (2002a); 300 kb windows yield a very similar plot, whereas 10 or 20 kb windows yield a quite different picture (see above). For very short genomes, smaller window sizes are appropriate (Fig. 1e–h).

The scheme is particularly useful for interspecies studies of syntenic regions. We recently used essentially the same scheme to show a perfect correspondence between human and mouse isochores in the HLA/MHC region, where transitions between isochores are seen to be well conserved (Pavlíček et al., 2002b). We have also used a modification of the scheme for visualizing the clustering of repeated sequences in human chromosomes (Pavlíček et al., 2001).

During sequencing, assembly and mapping of contigs the scheme can be used as a tool for rapidly visualizing differences between candidate assemblies or successive drafts of a chromosomal sequence, since the human eye can quickly recognize large relocations and inversions when guided by the combination of colors and line plots. Positions and extents of remaining gaps can also be shown on the plots. Furthermore, cytogenetic maps and ideograms obtained by in situ fluorescence hybridization (FISH) of very GC-rich DNA

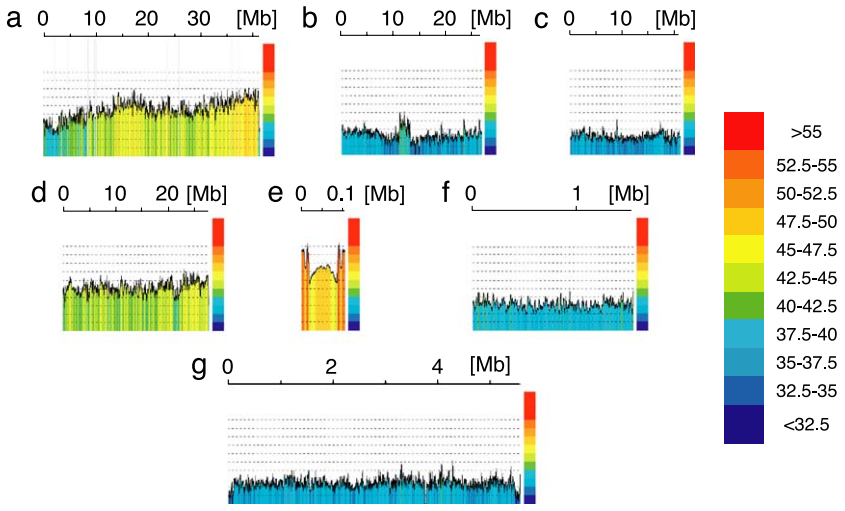

Fig. 1. Color-coded GC level plots of seven eukaryotic chromosomes. (a) *Anopheles gambiae* chromosome 3L (Holt et al., 2002), using a 100-kb window. (b) *Arabidopsis thaliana* chromosome 5 (AGI, 2000), 100 kb window. (c) *Caenorhabditis elegans* chromosome 5 (CSC, 1998), 100 kb window. (d) *Drosophila melanogaster* chromosome 3R (Myers et al., 2000), 100 kb window. (e) *E. cuniculi* chromosome 1, 10 kb window. (f) *Saccharomyces cerevisiae* chromosome 4 (Jacq et al., 1997), 10 kb window. (g) *Schizosaccharomyces pombe* chromosome 1 (Wood et al., 2002), 10 kb window. Line plots show the GC levels of overlapping 100 kb (a–d) and 10 kb (e–g) windows, colors indicate the 2.5% GC intervals to which these GC levels belong. Data and methods: automatization of plots was done in Perl (see Materials and methods). Sequences were obtained from the NCBI (http://www.ncbi.nlm.nih.gov) server. Unidentified nucleotides: if the number of consecutive N's in any window exceeded a pre-defined threshold (set at 1000 and 10,000 for the 10 and 100 kb windows, respectively) a gap was registered (grey vertical bars). In all other cases, GC% was calculated as $100\% \times (G+C)/(A+C+G+T)$.

show in red the GC-richest, gene-richest bands (see Saccone et al., 1993, 1996, 2002 for such maps of the human chromosomes). Candidate assemblies can therefore be checked for compatibility with such data by simply comparing the red/orange regions in the cytogenetic map with those in the color-coded GC plot.

3.2. A compact view of isochores in the mouse genome

Fig. 2 presents GC profiles of the recently sequenced mouse genome (MGSC, 2002). The plots show, in agreement with the early demonstration by Macaya et al. (1976), that this genome is composed of long stretches of compositionally similar segments, the isochores; in each of the long stretches (typically in the megabase range) there is only minor variation in color, yet most chromosomes span the full color range from deep blue to bright red. The contrasts between GC-poor and GC-rich isochores are not quite as pronounced in mouse and other myomorph rodents as they are in human (see Bernardi, 2000; Douady et al., 2000). Correspondingly, the mouse profile contains fewer bright-red regions than the analogous human profile (Fig. 2 and Pavlíček et al., 2002b).

3.3. Contributions of repetitive and unique DNA in GC poor and GC rich DNA

An obvious feature of both human and mouse chromosomes is the general preponderance of very long GC-poor isochores (shown in blue in the compact view). In accordance with the correlation between gene density and GC level of the isochore (see above), genes in GC poor regions are widely spaced, and the rare genes that are present in such regions often span several exons that are interrupted by long introns. One might be tempted to conclude that the large quantities of very GC-poor noncoding or intergenic DNA are likely to just consist of repetitive DNA, i.e., of highly repetitive satellite DNA and/or middle-repetitive interspersed DNA belonging to families such as LINEs and SINEs. The genomic sequences now allow one to conclude that this is not the case (Fig. 3).

The repetitive DNA that has been fixed in the human and mouse genomes does show a conspicuous preference for the vast expanses of GC-poor, mainly noncoding, DNA. It is indeed striking to see how high repeat densities (i.e., low proportions of unique DNA) are essentially absent from the sequenced GC-rich DNA in both genomes (the scatterplots shown in the middle row of Fig. 3 have a neat upper-triangular form). In GC-poor DNA, however, the situation is different: it is clear from the plots that many of the 100-kb regions of GC-poor DNA consist mainly of repetitive DNA, but also that many other regions consist mainly of unique DNA.

Other explanations must be sought, therefore, for the large expanses of noncoding (and, apparently, mostly intergenic), unique GC-poor DNA in mammalian genomes. Interestingly, the interspersed repeats in human and mouse are hardly ever as GC-poor as the GC-poorest DNA in these species. This fact can be seen from the histograms (top row) or from the scatterplots (bottom row) of Fig. 3 (see also Pavlíček et al., 2001 for similar plots). The latter show that the GC level of repetitive DNA does

correlate well with the GC level of the unique DNA in which it is found, a correlation that is explained by negative selection on repeats that differ compositionally, over long regions, from their genomic environment (Pavlíček et al., 2001). On the other hand, the correlation does not follow the diagonal but is less steep: in other words, contrasts between GC-poor and GC-rich isochores are diminished, not increased, when one includes repeats. It is therefore obvious that repetitive DNA, or its accumulated effect over a long time, cannot be held responsible for the large GC contrasts that are observed along the human and mouse chromosomes.

After excising the repetitive DNA, the remaining GC-poorest regions (GC < 35%) still account for perhaps a tenth of the DNA in the two genomes. It has been proposed that most regions of about kilobase scales and having a GC less

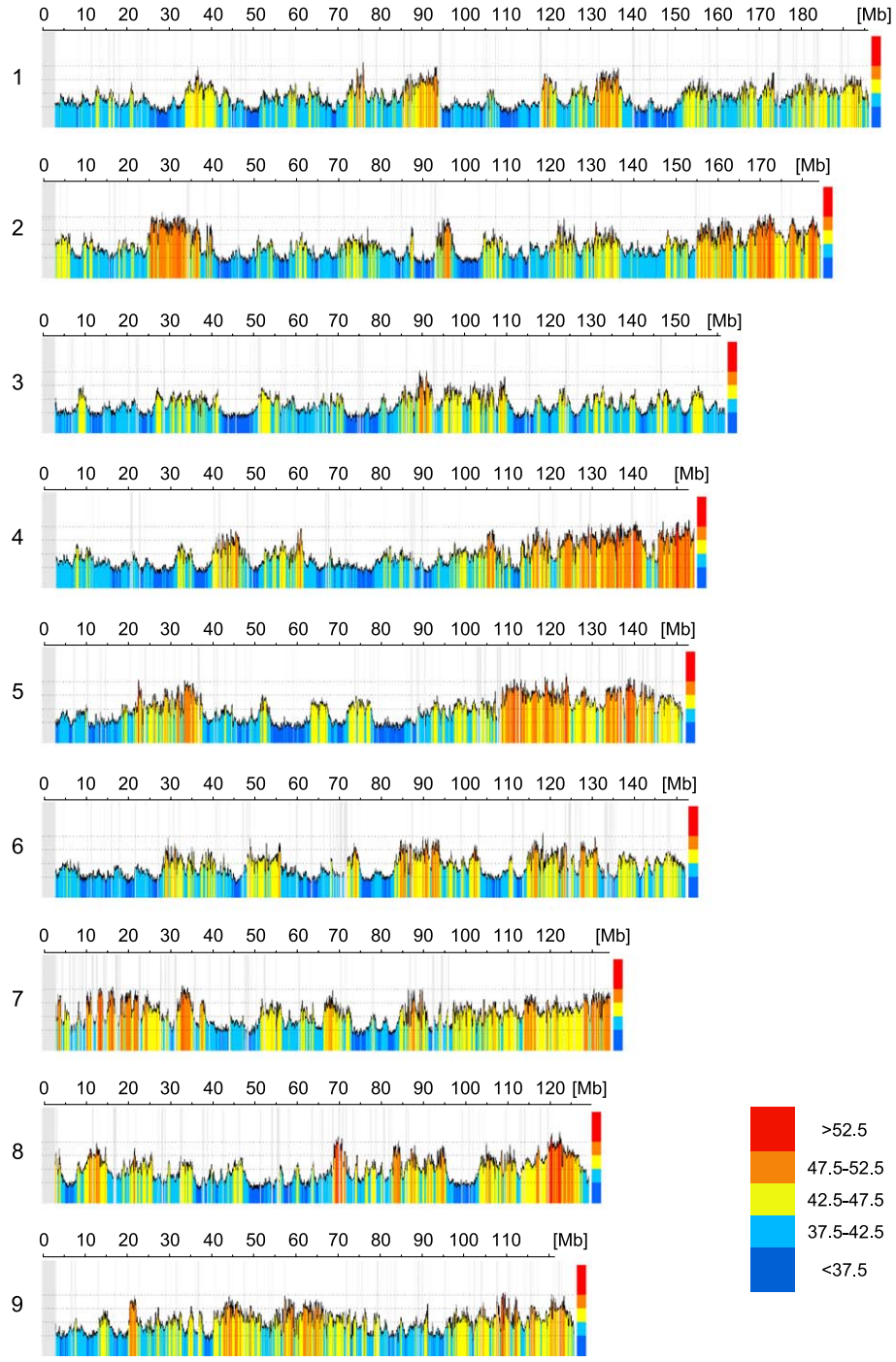

Fig. 2. A compact view of isochores in the mouse genome sequence (UCSC GoldenPath mouse genome draft of February 2003). Sequences were scanned using a 100-kb moving window. In this variant, adjacent 2.5% intervals were pooled, i.e., the line plots were filled with five colors encoding GC levels from deep blue (lowest) to red (highest). Grey bars show heterochromatic DNA and gaps in the sequence of the euchromatic DNA.

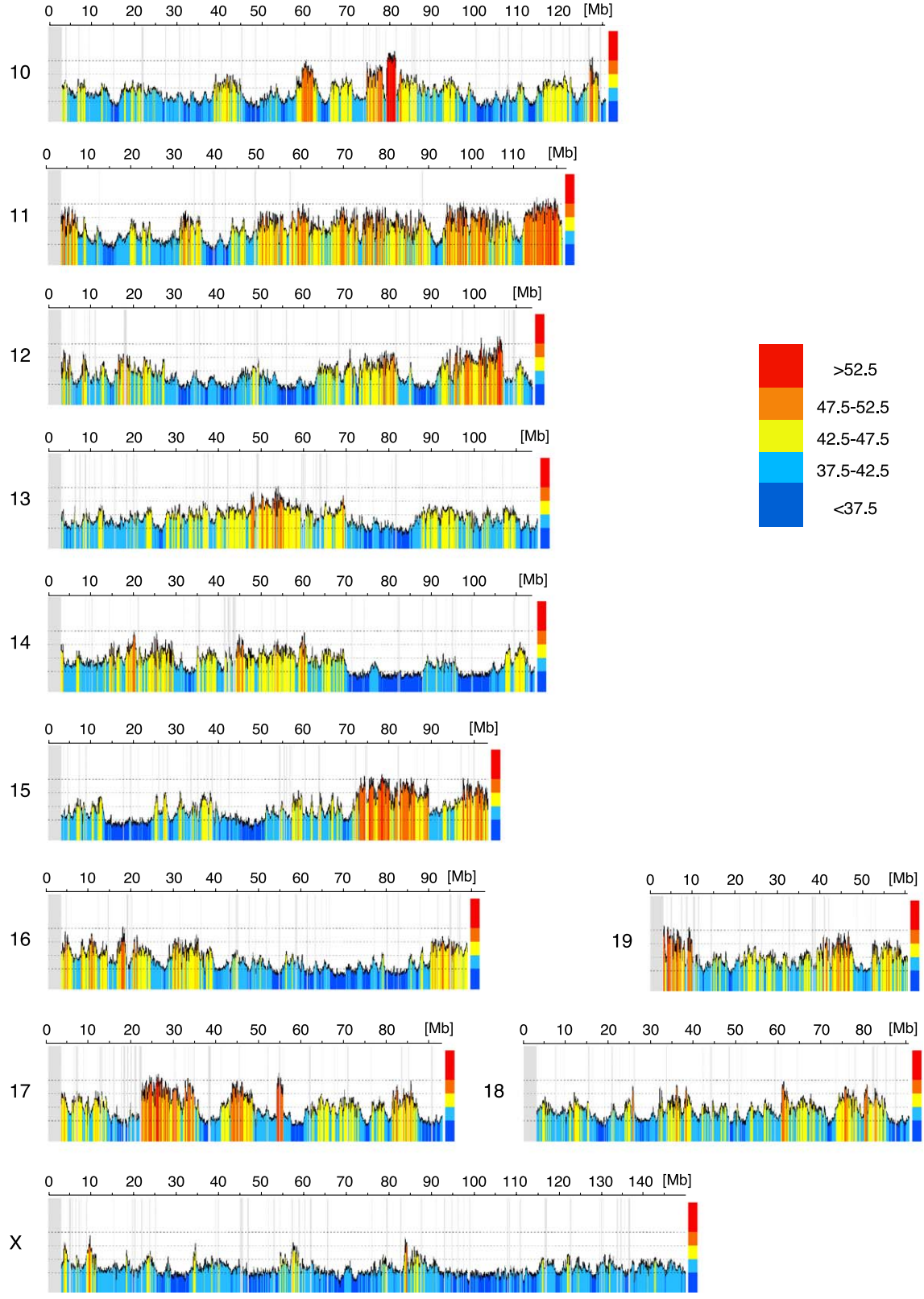

Fig. 2 (*continued*).

than 35% are likely to be potential matrix/scaffold attachment regions (MARs/SARs; Saitoh and Laemmli, 1994). According to our current understanding, however, MARs typically span no more than a few kb, and alternate with much longer regions of DNA that loop out and away from the nuclear matrix. The highly conserved noncoding regions found so far

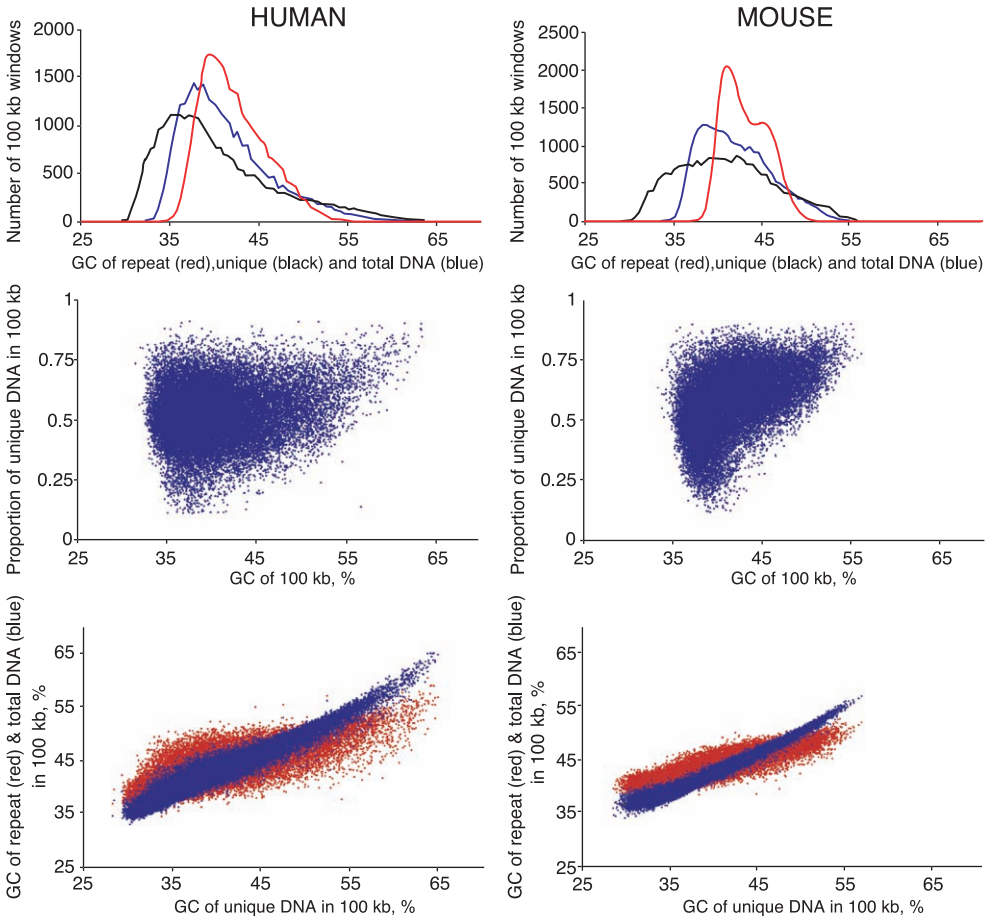

Fig. 3. Plots showing the contributions of interspersed repeats to 100 kb regions in the sequenced euchromatic portions of the human and mouse genomes. Top row: histograms for 0.5% GC intervals, showing the base compositions of repeat (red), unique or repeat-masked (black) and total DNA in each 100 kb segment of human (left) and mouse (right) DNA. Middle row: scatterplots showing the proportions of unique DNA in the 100 kb regions. Note the complete absence of high repeat densities in GC-rich regions. Bottom row: scatterplots showing how the GC levels of total (blue), repeat (red) and masked or nonrepetitive DNA covary. Repeat coordinates were obtained from the UCSC human and mouse annotation databases (Karolchik et al., 2003; http://genome.ucsc.edu, 'rmsk' files), where more detailed information can be extracted on the contributions of different repeat families. Repeats were located by RepeatMasker (A.F.A. Smit and P. Green, unpublished; Smit, 1999), using RepBase (Jurka, 2000). To eliminate noise (short effective window sizes), the very few windows containing less than 10% repetitive DNA, or less than 10% unique DNA, were excluded, as were windows containing over 50% N's (unidentified nucleotides).

by similarity searches can, similarly, account for only a small percentage of the GC-poor, and largely intergenic, DNA in human and mouse. A considerable quantity of very GC-poor, single-copy DNA therefore remains unexplained in these genomes, but its more detailed mapping by methods such as the one presented here may facilitate the search for its causes.

4. Conclusion

In summary, we have presented a scheme for mapping and presenting GC levels and their large-scale variation that should be applicable to most eukaryotic chromosome sequences, and in some cases the choices it implements appear nearly optimal. Applications range from the recognition of gene-dense regions (which are GC-rich, likely to replicate early, and preferentially extend away from the matrix in interphase) to comparisons between genomes or between draft assemblies of the same genome. The same scheme can also be applied (using smaller windows) to smaller regions, in order to recognize or highlight compositional features at smaller scales such as genes, exons, MARs, and CpG islands.

Acknowledgements

This work was supported by the Center for Integrated Genomics of the Czech Republic.

References

Arabidopsis Genome Initiative (AGI), 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408, 796–815.

Bernardi, G., 2000. The compositional evolution of vertebrate genomes. Gene 259, 31–43.

Bernardi, G., 2001. Misunderstandings about isochores: Part I. Gene 276, 3–13.

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

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

Douady, C., Carels, N., Clay, O., Catzeflis, F., Bernardi, G., 2000. Diversity and phylogenetic implications of CsCl profiles from rodent DNAs. Mol. Phylogenet. Evol. 17, 219–230.

Gentles, A.J., Karlin, S., 2001. Genome-scale compositional comparisons in eukaryotes. Genome Res. 11, 540–546.

Holt, R.A., Subramanian, G.M., Halpern, A., et al., 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. Science 298, 129–149.

International Human Genome Sequencing Consortium (IHGSC), 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.

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

Jacq, C., Alt-Morbe, J., Andre, B., et al., 1997. The nucleotide sequence of *Saccharomyces cerevisiae* chromosome IV. Nature 387, 75–78.

Jurka, J., 2000. Repbase update: a database and an electronic journal of repetitive elements. Trends Genet. 16, 418–420.

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., Weber, R.J., Haussler, D., Kent, W.J., 2003. The UCSC genome browser database. Nucleic Acids Res. 31, 51–54.

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.

Mahy, N.L., Perry, P.E., Bickmore, W.A., 2002. Gene density and transcription influence the localization of chromatin outside of chromosome territories detectable by FISH. J. Cell Biol. 159, 753–763.

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

Mouse Genome Sequencing Consortium (MGSC), 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520–562.

Myers, E.W., Sutton, G.G., Delcher, A.L., et al., 2000. A whole-genome assembly of Drosophila. Science 287, 2196–2204.

Oliver, J.L., Bernaola-Galván, P., Carpena, P., Román-Roldán, R., 2001. Isochore chromosome maps of eukaryotic genomes. Gene 276, 47–56.

Pavlíček, A., Jabbari, K., Pačes, J., Pačes, V., Hejnar, J., Bernardi, G., 2001. Similar integration but different stability of Alus and LINEs in the human genome. Gene 276, 39–45.

Pavlíček, A., Pačes, J., Clay, O., Bernardi, G., 2002a. A compact view of isochores in the draft human genome sequence. FEBS Lett. 511, 165–169.

Pavlíček, A., Clay, O., Jabbari, K., Pačes, J., Bernardi, G., 2002b. Isochore conservation between MHC regions on human chromosome 6 and mouse chromosome 17. FEBS Lett. 511, 175–177.

Saccone, S., De Sario, A., Wiegant, J., Raap, A.K., Della Valle, G., Bernardi, G., 1993. Correlations between isochores and chromosomal bands in the human genome. Proc. Natl. Acad. Sci. U. S. A. 90, 11929–11933.

Saccone, S., Cacciò, S., Kusuda, J., Andreozzi, L., Bernardi, G., 1996. Identification of the gene-richest bands in human chromosomes. Gene 174, 85–94.

Saccone, S., Federico, C., Bernardi, G., 2002. Localization of the gene-richest and the gene-poorest isochores in the interphase nuclei of mammals and birds. Gene 300, 169–178.

Saitoh, Y., Laemmli, U.K., 1994. Metaphase chromosome structure: bands arise from a differential folding path of the highly AT-rich scaffold. Cell 76, 609–622.

Smit, A.F., 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr. Opin. Genet. Dev. 9, 657–663.

Stephens, R., Horton, R., Humphray, S., Rowen, L., Trowsdale, J., Beck, S., 1999. Gene organization, 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 major histocompatibility complex. Mol. Cell. Biol. 17, 4043–4050.

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

The MHC sequencing consortium (MHCSC), 1999. Complete sequence and gene map of a human major histocompatibility complex. Nature 401, 921–923.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., et al., 2001. The sequence of the human genome. Science 291, 1304–1351.

Wood, V., Gwilliam, R., Rajandream, M.A., et al., 2002. The genome sequence of *Schizosaccharomyces pombe*. Nature 415, 871–880.

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