ELSEVIER

Gene 276 (2001) 3–13

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

www.elsevier.com/locate/gene

Review

Misunderstandings about isochores. Part 1

Giorgio Bernardi*

Laboratory of Molecular Evolution, Stazione Zoologica Anton Dohrn, Villa Comunale, 80121, Naples, Italy

Received 10 July 2001; accepted 21 August 2001
Received by T. Gojobori

Abstract

A few months ago the International Human Genome Sequencing Consortium (IHGSC) published a 61-page paper on the human genome (IHGSC, Nature 409 (2001) 860). Here comments will be presented on some points of the paper that were previously investigated in our laboratory, and some misunderstandings and misconceptions about the organization and the evolutionary history of the human genome will be discussed. A very recent article on the same subject (Eyre-Walker and Hurst, Nat. Rev. Genet. 2 (2001) 549) will also be addressed. The present paper is a complement to two review articles which were published last year (Bernardi, Gene 241 (2000) 3; Gene 259(1) (2000) 31).

Keywords: Genome organization; Genome evolution; Repeats; Mutational bias; Gene distribution

1. Introduction

The draft sequence paper published by the International Human Genome Sequencing Consortium (IHGSC) is different from previous sequence reports (including the paper published simultaneously by Venter et al., 2001), which presented data and addressed issues of sequence analysis and gene prediction, in that the IHGSC attempted to also present a general picture of a very broad and complex research area, that of the organization and evolution of the human genome. This attempt was apparently too ambitious, also in view of the time and space limitations imposed on the authors. As a consequence, some erroneous and controversial conclusions found their way into the paper. Since, in all likelihood, the IHGSC article will have a very wide circulation, it is important that such conclusions be corrected or critically discussed before they spread into the literature and become (at least temporarily) established truths.

This discussion of the IHGSC paper is, in fact, already underway on some of the topics addressed. For instance, the proposed horizontal transfer of bacterial genes to vertebrates was shown to be explained, in most cases, by descent through common ancestry (Stanhope et al., 2001; Roelofs and Van Haaster, 2001; DeFilippis et al., 2001). Here, the discussion will be focused on some subjects of the IHGSC paper previously dealt with in our laboratory, such as the broad genomic landscape, namely the isochore pattern of the human genome, the distribution of repeats and genes in the isochores, and the mutational bias, i.e. the non-randomness of the mutational input.

2. Broad genomic landscape (p. 875)[1]

2.1. Long-range variation in GC content (pp. 876–877)

According to the authors, “Bernardi and colleagues (Bernardi et al., 1985; Bernardi, 2000a) proposed that the long-range variations in GC content may reflect that the genome is composed of a mosaic of compositionally homogeneous regions that they dubbed ‘isochores’. They suggested that the skewed distribution is composed of five normal distributions, corresponding to five distinct types of isochore (L1, L2, H1, H2 and H3, with GC contents of <38%, 38–42%, 42–47%, 47–52%, respectively)”.

Even if it has occurred to us to sometimes mention isochores as ‘homogeneous regions’ for brevity, in our original paper, specifically dealing with the problem of the compositional heterogeneity of isochores, we defined isochores as ‘fairly homogeneous regions’ (Cuny et al.,

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

* Tel.: +39-081-5833300; fax: +39-081-2455807.
E-mail address: bernardi@alpha.szn.it (G. Bernardi).

[1] All citations of the International Human Genome Sequencing Consortium (IHGSC) paper were changed to fit the style of GENE; to facilitate cross-reference, the headings and relevant pages of the IHGSC paper are given.

0378-1119/01/$ - see front matter © 2001 Elsevier Science B.V. All rights reserved.
PII: S0378-1119(01)00644-8

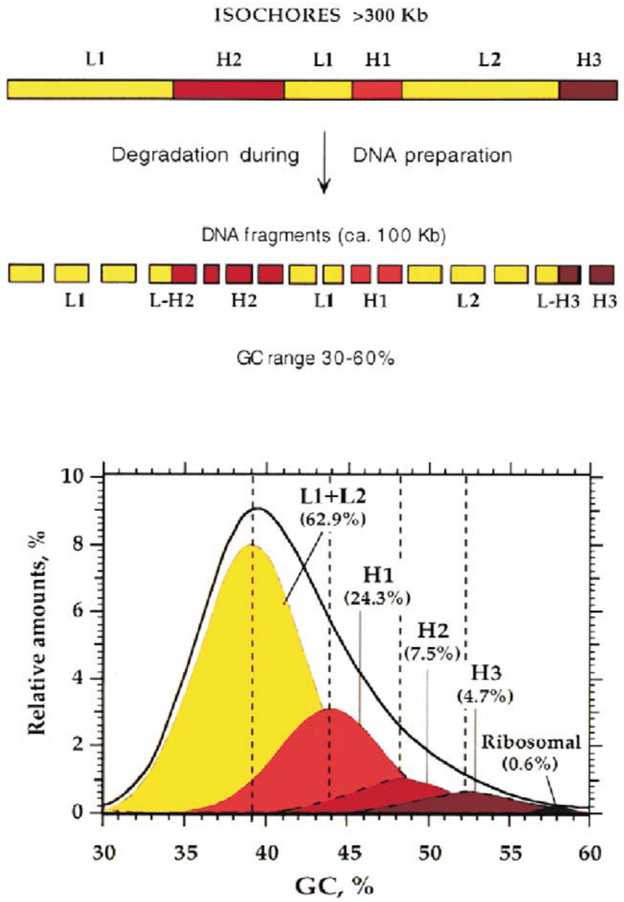

Fig. 1. (Top) Scheme of the isochore organization of the human genome. This genome, which is typical of the genome of most mammals, is a mosaic of large DNA segments, the isochores, which are compositionally fairly homogeneous and can be partitioned into a small number of families, 'light' or GC-poor (L1 and L2), and 'heavy' or GC-rich (H1, H2 and H3). Isochores are degraded during DNA preparation to fragments of 50–100 kb in size. The GC range of these DNA molecules from the human genome is extremely broad, i.e. 30–60%. (From Bernardi (1995).) (Bottom) The CsCl profile of human DNA is resolved into its major DNA components, namely DNA fragments derived from each one of the isochore families (L1, L2, H1, H2, H3). Modal GC levels of isochore families are indicated on the abscissa (broken vertical lines). The relative amounts of major DNA components are indicated. Satellite DNAs are not represented. (From Zoubak et al. (1996).)

1981). Indeed, the major DNA components (namely the compositional families of 50–100 kb DNA molecules derived from isochore families; see Fig. 1) are only about 30% more heterogeneous (by comparison of standard deviations) than bacterial DNAs having the same size and composition.

The IHGSC authors studied "the draft genome sequence to see whether strict isochores could be identified" and failed to find any. They concluded that their results "rule out a strict notion of isochores as compositionally homogeneous" and that "isochores do not appear to deserve the prefix 'iso'."

Since the terminology 'strict isochores' used by the authors denotes sequences that cannot be distinguished from random (uncorrelated) sequences in which every nucleotide is free to change, their failure to identify 'strict isochores' in the human genome could be predicted on two accounts: first, for over 40 years (Rolfe and Meselson, 1959) random sequences have been known to be much more homogeneous than the least heterogeneous genomic DNAs, namely bacterial DNAs (viral DNAs are not considered here), which are in turn much more homogeneous than eukaryotic DNAs; second, 'strict isochores' cannot exist in any natural DNA because non-coding sequences are compositionally correlated with the coding sequences that they embed (Bernardi et al., 1985; D'Onofrio et al., 1991; Clay et al., 1996) and coding sequences are made up of codons, in which the compositions of the three positions are correlated with each other (D'Onofrio and Bernardi, 1992). Detailed discussions of this problem are presented elsewhere (Clay and Bernardi, 2001a,b; Clay et al., 2001; Clay, 2001).

In summary, the conclusion of the authors that 'isochores' are not 'strict isochores' is correct, but it is something we have known for 20 years, since Cuny et al. (1981) quantified the heterogeneity of isochore families. Around the same time as the IHGSC paper, other laboratories also missed the point that 'strict isochores' cannot exist in natural DNA and also took random sequences as references for compositional homogeneity (Häring and Kypr, 2001; and also, in part, Nekrutenko and Li, 2000; these papers have been commented on in detail in the references quoted above). Moreover, perfectly homogeneous isochores separated by hyper-sharp boundaries were displayed by Eyre-Walker and Hurst (2001) in a figure entitled the 'classic isochore model', further confusing the issue for the non-specialist. Ironically, the 'classic isochores' of Eyre-Walker and Hurst (2001) managed to beat even 'strict isochores', as far as compositional homogeneity is concerned. Both 'strict isochores' and 'classic isochores' are misleading definitions that hopefully will have a short life-time in the literature, since they concern sequences that do not exist in natural DNAs.

Along another line, the authors do not seem to accept the idea that the human genome is a mosaic of isochores, namely that the large-scale compositional heterogeneity is discrete or discontinuous, rather than continuously drifting, as widely believed 30 years ago (until the publication of the work of Filipski et al., 1973) and as later proposed as a model by Fickett et al. (1992). The authors' tacit rejection of the discontinuous compositional heterogeneity neglected, however, the detailed investigations that led to this conclusion (Filipski et al., 1973; Thiery et al., 1976; Macaya et al., 1976), as well as the data concerning compositional discontinuities between isochores, as detected at the sequence level (Fukagawa et al., 1995) and at the chromosomal level (Saccone et al., 2001; see below). The authors also apparently overlooked the evidence that the heterogeneity arose by formation of GC-rich isochores at the transition between cold- and warm-blooded vertebrates, from regions that were much

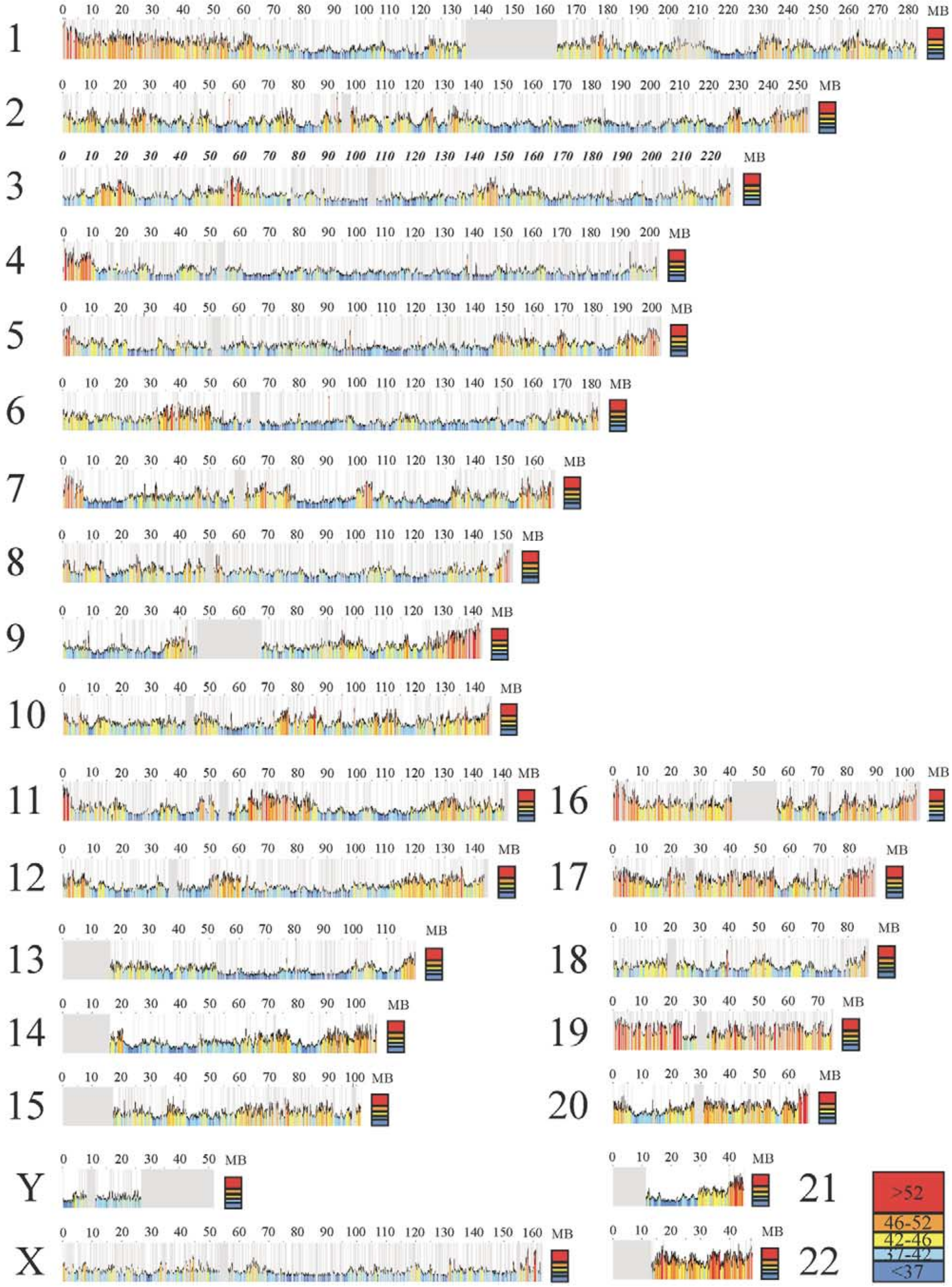

Fig. 2. A colour-coded compositional map of the chromosomes of the human genome, representing 100 kb moving window plots that scan the draft human genome sequence of International Human Genome Sequencing Consortium (2001). Colour codes span the spectrum of GC levels in five steps, from ultramarine blue (GC-poorest isochores) to scarlet red (GC-richest isochores). (Modified from Pavliček et al. (2001b).)

more homogeneous and lower in GC level and, in fact, very similar to the GC-poor isochores of the human genome (see Bernardi, 2000b, for a review). This emergence from the much more homogeneous compositional spectrum of the genome of cold-blooded vertebrates was the primary source for a discontinuous distribution.

A compositional map of human chromosomes derived from the IHGSC data (Fig. 2; see also Pavliček et al., 2001b, for more detailed maps) is very telling because it graphically displays the mosaic organization of the human genome. Indeed, apart from the abundance of gaps (grey bars) in the euchromatic regions of most chromosomes, the striking feature of the map is undoubtedly the large proportion of the genome represented by long GC-poor regions, uninterrupted by GC-rich regions. The next most notable observation is the scarcity of GC-poor regions in many of the blocks characterized by GC-rich regions. As far as these two points are concerned, the results fit with the previous

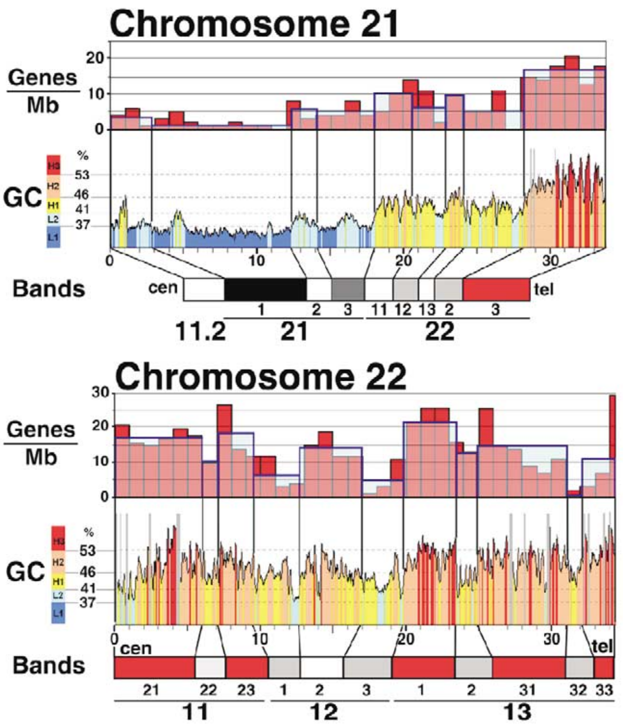

Fig. 3. Correlations between chromosomal bands, isochores, and gene concentration in human chromosomes 21 and 22. (Bottom to top) Bands: ideogram at a resolution of 850 bands showing the four classes of G bands staining with different intensities and the two classes ($H3^+$, red; $H3^-$, white) of R bands. The two chromosomes are represented according to their relative cytogenetic size. GC: the GC profiles were obtained using a window size of 100 kb; 37, 41, 46 and 53% GC were taken as the upper values of the L1, L2, H1 and H2 isochore families, respectively. The grey bars indicate the DNA sequences not yet available. Genes/Mb: gene density per Mb. The blue histogram concerns chromosomal bands, and the red histogram concerns 1 Mb segments. (From Saccone et al. (2001).)

estimates of the relative amounts of GC-poor and GC-rich isochores (see Fig. 1) and with the very high yields of physically separated major DNA components (Cuny et al., 1981). They contradict the suggestion (Eyre-Walker and Hurst, 2001), for which no evidence is quoted, that the isochore structure accounts for ‘only some parts’ of the genome.

The third observation is the increasing compositional fluctuations when moving from GC-poor to GC-rich isochores. This had already been noticed in previous work (Cuny et al., 1981; De Sario et al., 1996, 1997) and was confirmed by a detailed analysis of chromosomes 21 and 22 (Saccone et al., 2001) (see Figs. 3 and 4). A working hypothesis, currently being tested in our laboratory, is that the increase in compositional fluctuations may be due to a simultaneous requirement both for AT-rich regions to link DNA to the chromosome scaffold and for the very high GC levels of coding sequences and associated CpG islands in H2 and H3 isochores.

In their conclusion, the IHGSC authors mention (p. 860) “suggestions that large GC-poor regions are strongly correlated with ‘dark G-bands’ in karyotype” ignoring the fact that our present knowledge of the correlations between isochores and chromosomal bands goes much beyond that. Indeed, *in situ* hybridization of DNA from different isochore families on metaphase chromosomes at 400-band resolution showed, already several years ago, that H2 and H3 isochores hybridize on a small set of R(everse) bands that essentially correspond to the bands most resistant to heat-denaturation of Dutrillaux (1973), whereas GC-poor isochores are mainly concentrated on G(iemsa) bands (Saccone et al., 1992, 1993, 1996). More recently, *in situ* hybridization of the GC-poorest L1 isochores on prophase chromosomes at 850-band resolution showed that they were located on a number of G bands that largely correspond to the most intensely staining G bands of Francke (1994), and defined the R bands corresponding to the GC-richest H3 isochores (Saccone et al., 1999; Federico et al., 2000). These bands were called $L1^+$ and $H3^+$ bands, respectively, the remaining G and R bands ($L1^-$ and $H3^-$) being characterized (see Fig. 5) by an intermediate GC composition (Federico et al., 2000). Interestingly, $H3^+$ bands have a lower compaction of DNA compared with $L1^+$ bands, a point which will be commented upon later. Moreover, $L1^-$ bands were observed that exhibited higher GC levels compared to some $H3^-$ bands, the G or R banding depending upon the higher or lower GC levels, respectively, of flanking bands (Fig. 4). G and R bands therefore appear to depend largely on compositional

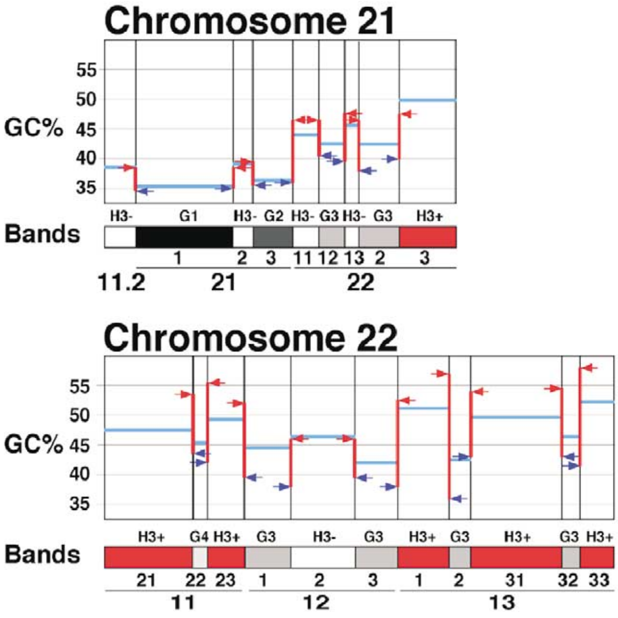

Fig. 4. (Bottom to top) Bands: band ideograms as in Fig. 3. G1–G4 (the four types of G bands showing decreasing staining intensities from G1 to G4; see Francke, 1994), $H3^+$ and $H3^-$ bands are indicated. GC%: average GC level of each chromosomal band (horizontal blue lines), and GC levels observed at band borders (vertical red lines indicate the GC difference over 300 kb regions around band borders; red and blue arrows indicate the GC level on the R and G band side, respectively). Note that all G bands showed lower GC levels than the adjacent R bands, and all R bands showed higher GC levels than the adjacent G bands. These differences were enhanced at band border regions (see above). (From Saccone et al. (2001).)

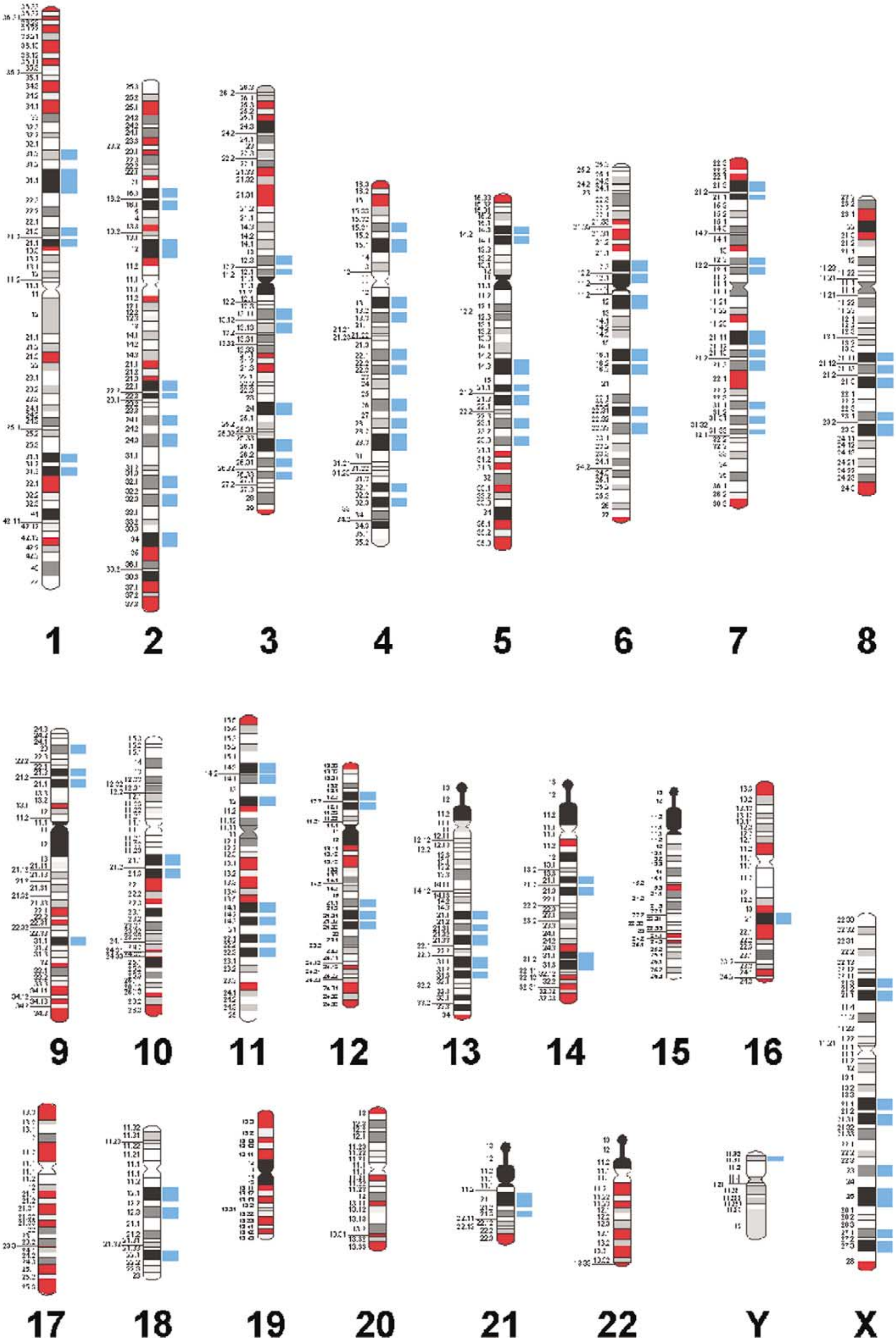

Fig. 5. Identification of the GC-poorest and the GC-richest chromosomal bands. Human karyotype at a resolution of 850 bands per haploid genome showing the chromosomal bands containing the GC-poorest (blue bars on the right of each chromosome) and the GC-richest isochores (red regions inside the chromosomes). The grey scale of the G bands is according to Francke (1994). The GC-richest isochore bands are from Saccone et al. (1999). (From Federico et al. (2000).)

contrasts between adjacent regions, rather than just on their absolute GC levels.

At this point, two observations should be mentioned: (i) the sharp discontinuities at band borders (Fig. 4); and (ii) the correlation between the GC level of bands and replication timing, the GC-richest bands replicating early and the GC-poorest bands replicating late in the cell cycle (Federico et al., 1998, 2000). Incidentally, the latter findings contradict early data suggesting a lack of correlation between GC levels of isochores and replication timing (Eyre-Walker, 1992), which are still considered by Eyre-Walker and Hurst (2001).

3. Repeat content of the human genome (p. 879)

3.1. Distributions by GC content (pp. 884–885)

The starting point of this section is our original findings (Meunier-Rotival et al., 1982; Soriano et al., 1983) on the preferential location of GC-poor LINEs and GC-rich Alus in GC-poor and GC-rich isochores, respectively.

According to the authors, "the preference of LINEs for AT-rich DNA seems like a reasonable way for a genomic parasite to accommodate its host, by targeting gene-poor AT-rich DNA and thereby imposing a lower mutational burden. Mechanistically, selective targeting is nicely explained by the fact that the preferred cleavage site of the LINE endonuclease is TTTT/A (where the slash indicates the point of cleavage), which is used to prime reverse transcription from the poly(A) tail of LINE RNA." Whether this interpretation involving a 'lower mutational burden' is correct is doubtful. Indeed, it is difficult to see why the same reasoning would not apply to the SINEs located in the gene-rich, GC-rich isochores, since in both cases integration practically only occurs in intergenic regions, and the 'mutational burden' cannot be too different.

As far as SINEs are concerned, the authors raise the question "How do SINEs accumulate in GC-rich DNA, particularly if they depend on the LINE transposition machinery (Jurka, 1997)?". According to the authors, "One possibility is that SINEs somehow target GC-rich DNA for insertion. The alternative is that SINEs initially insert with the same proclivity for AT-rich DNA as LINEs, but that the distribution is subsequently reshaped by evolutionary forces (Smit, 1999; Arcot et al., 1998)." Having tacitly ruled out possibility 1 (see our Table 1), and having shown that "recent Alus show a preference for AT-rich DNA resembling that of LINEs" (a result independently found by Pavliček et al., 2001a), they ask the question "What is the force that produces the great and rapid enrichment of Alus in GC-rich DNA?" and conclude that "this could be a higher rate of random loss of Alus in AT-rich DNA, negative selection against Alus in AT-rich DNA or positive selection in favour of Alus in GC-rich DNA (see Table 1). The first two possibilities seem unlikely because AT-rich DNA is gene-poor and tolerates the accumulation of other transposable elements. The third seems more feasible, in that it involves selecting in favour of the minority that lie in GC-rich regions rather than against the majority that lie in AT-rich regions. But positive selection for Alus in GC-rich regions would imply that they benefit the organism." The latter point was judged by the authors to be important enough to be included among the main conclusions of the paper (p. 860).

Table 1
Explanations for the accumulation of SINEs in GC-rich isochores

1. SINEs target GC-rich isochores for insertion
2. Young SINEs target GC-poor isochores for insertion but
   (a) there is a higher rate of random loss of SINEs from GC-poor isochores
   (b) SINEs are eliminated from GC-poor isochores by negative selection
   (c) SINEs deletions are more tolerated in GC poor isochores
3. SINEs are positively selected in GC-rich isochores

If one considers the very important, yet very rare role played by positive selection in evolution, this proposal (considered as controversial by the authors themselves) should be supported by evidence stronger than the hypothesis of Schmid (1998) that mention. This hypothesis postulates that the transcription of some SINEs under conditions of stress produces RNAs that specifically bind a particular protein kinase (PKR), which blocks the ability of PKR to inhibit protein translation. Such a promotion of protein translation under stress, while interesting in itself, should, however, concern the majority of Alus located in GC-rich isochores in order to account for the proposal, whereas the hypothesis of Schmid (1998) involves a small number of Alus and does not say anything about their isochore localization. "The idea that Alu correlates not with GC content but with actively transcribed genes" is contradicted by the finding that the highest density of Alu elements is in H2 isochores (Zerial et al., 1986; Jabbari and Bernardi, 1998; Pavliček et al., 2001a), a point not obvious in the graphs of the IHGSC paper which do not go beyond 54% GC, whereas the highest density of genes is in H3 isochores (Zerial et al., 1986; Mouchiroud et al., 1991; Zoubak et al., 1996).

The other two possibilities, ruled out either on no grounds at all (possibility 1 in Table 1), or "because GC-poor DNA is gene-poor and tolerates the accumulation of other transposable elements" (possibility 2 of Table 1), deserve to be considered more seriously, because they are in fact likely to account for the situation under discussion, even if their relative extents cannot be estimated at present. Concerning possibility 1, one should recall that even if lower than in GC-poor isochores, the frequency of potential acceptor sites TTT/A in GC-rich isochores is still much higher than the frequency of inserted Alus. There is, therefore, no shortage of available acceptor sites in GC-rich isochores, and the preference of young Alus to integrate in the GC-poor isochores does not mean that integration into GC-rich isochores does not occur. Indeed, Fig. 23 of the IHGSC paper and the results of Pavliček et al. (2001a) show only a two-fold higher frequency in L1 compared to H3 isochores. Concerning possibility 2, one should remember that the stability of an inserted sequence depends upon its compositional match with the isochore. Exclusion of inserted sequences which do not compositionally fit the isochores is well demonstrated in the case of proviruses, as is the lack of transcription of proviruses that deviate from a good compositional match (see Rynditch et al., 1998, for a review). The concept of the compositional match between the transposed sequence and the isochore in which it is inserted is missing (or tacitly rejected) in the reasoning of the authors. In contrast, they propose the explanation that repeated sequences may 'modulate' the overall GC content. This explanation is, however, contradicted by the evidence that the isochore composition is very similar whether one includes or excludes repeated sequences (Pavlíček et al., 2001a). In fact, after elimination of repeats, GC-rich sequences become, if anything, slightly GC-richer (Pavlíček et al., 2001a), a trend opposite to that expected from the putative 'modulation'.

3.2. Biases in human mutations (pp. 885–886)

"By studying sets of repeat elements belonging to a common cohort, one can directly measure nucleotide substitution rates in different regions of the genome." Using this approach, the authors find strong evidence that "the pattern of neutral substitution differs as a function of local GC content (Fig. 27 of the IHGSC paper)." The authors conclude that "because the results are observed in repetitive elements throughout the genome, the variation in the pattern of nucleotide substitution seems likely to be due to differences in the underlying mutational process rather than to selection."

In fact, the authors report two distinct observations. The first one is that "there is an absolute bias in substitution patterns resulting in directional pressure towards lower GC content throughout the human genome." This observation was previously made by Eyre-Walker (1999) and Smith and Eyre-Walker (2001) on single nucleotide polymorphisms (SNPs), and by Alvarez-Valin and Bernardi (2001) in a study of genes from the genetic disease mutation data set. In both cases, however, the conclusion was that the maintenance of base composition against the GC → AT mutational bias was due to selection (and/or to gene conversion, according to Smith and Eyre-Walker, 2001) adding one additional argument to many others accumulated over the years (see Bernardi, 2000a,b, for reviews). In their very recent paper, Eyre-Walker and Hurst (2001) examine the two remaining explanations and defend biased gene conversion against selection. Now, the arguments against the biased gene conversion that Eyre-Walker and Hurst (2001) quote, namely the positive correlation of K_s with GC, the ancient Y-linked GC-rich genes, and the high parameter sensitivity, are so serious that they make it difficult to accept the biased gene conversion as an explanation. In contrast, the arguments that Eyre-Walker and Hurst (2001) raise against selection have all been rebutted in papers (Chiusano et al., 1999, 2000; Cruveiller et al., 1999, 2000; Bernardi, 2000b) that they overlooked. We still consider, therefore, that the selection explanation is the only one which accounts for the formation and maintenance of isochores.

While the authors of the IHGSC paper take into consideration the selection explanation for GC-rich isochores, they prefer to account for them by "a constant influx of transposable elements" that "tend to increase the GC content". As already mentioned, this conclusion is contradicted by the finding of Pavlíček et al. (2001a) that transposable elements such as Alu sequences tend to slightly decrease the GC level of the GC-rich isochores in which they are inserted.

The authors' second observation is that GC base pairs are "more likely to mutate towards AT base pairs in AT-rich regions than in GC-rich regions." They conclude that "this bias could be due to a reported tendency for GC-rich regions to replicate earlier in the cell cycle than AT-rich regions and for guanine pools, which are limiting for DNA replication, to become depleted late in the cell cycle, thereby resulting in a small but significant shift in substitution towards AT base pairs (Wolfe et al., 1989; Mathews and Ji, 1992). Another theory proposes that many substitutions are due to differences in DNA repair mechanisms, possibly related to transcriptional activity and thereby to gene density and GC content (Sueoka, 1988; Holmquist and Filipski, 1994; Eyre-Walker, 1999)." It has been repeatedly stressed that the depletion hypothesis (still considered as a possibility also by Eyre-Walker and Hurst, 2001) is ruled out by the fact that the inactive X chromosome and the GC-rich satellite DNAs of mammals replicate at the very end of the cell cycle (see Bernardi et al., 1988, 1993; see also Graur and Li, 2000). This shows that any GC depletion at the end of the cell cycle would not be strong enough to alter the base composition of newly replicated DNA. The repair explanation also is notoriously unsatisfactory because it would essentially concern transcribed sequences, whereas the mutational bias under discussion concerns both transcribed and non-transcribed sequences.

In conclusion, the GC → AT mutational bias throughout the human genome (Eyre-Walker, 1999; Smith and Eyre-Walker, 2001; International Human Genome Sequencing Consortium, 2001; Eyre-Walker and Hurst, 2001; Alvarez-Valin and Bernardi, 2001) adds one more argument to several others raised over the years by our laboratory (see Bernardi, 2000a,b) and finally puts to rest the mutational bias hypothesis as an explanation for the formation and maintenance of isochores. At the same time, it contradicts the proposal of Francino and Ochman (1999) that mutation alone is responsible for the formation and maintenance of isochores, as well as the (already quoted) IHGSC conclusion that "the variation in the pattern of nucleotide substitution seems likely to be due to differences in the underlying mutational process rather than to selection."

4. Gene content of the human genome

4.1. Protein-coding genes (pp. 896–898)

This section of the IHGSC paper largely concentrates on

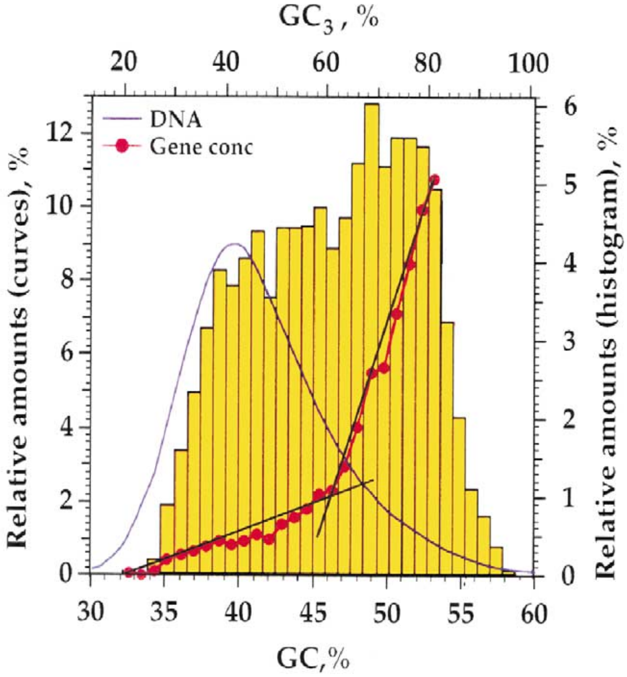

Fig. 6. Profile of gene concentration (red dots) in the human genome, as obtained by dividing the relative numbers of genes in each 1.5% GC_3 interval of the histogram of gene distribution (yellow bars) by the corresponding relative amounts of DNA deduced from the CsCl profile (blue line). The positioning of the GC_3 histogram relative to the CsCl profile is based on the correlation of GC_3 vs. GC of the isochores embedding the corresponding genes. (Modified from Bernardi (2000a).)

gene features like intron and exon size. Before discussing these results, it is appropriate to summarize the evidence for the existence of two classes of genes in vertebrates and plants. When, prompted by our early results (Bernardi et al., 1985), we investigated in detail the distribution of genes in the human genome (Mouchiroud et al., 1991; Zoubak et al., 1996), we found that, contrary to the general belief, this distribution is strikingly uneven, GC-poor isochores having a very low, and GC-rich isochores an increasingly high gene density. Moreover, as shown in Fig. 6, the plot of gene density vs. GC is characterized by two different slopes that cross each other at about 46% GC, a value that can be taken as the border between two 'gene spaces' (see Bernardi, 2000a, for a review), representing about 12% and 88% of the genome, respectively. These spaces were called the 'genome core' and the 'empty quarter' (from the classical name of the Arabian desert), respectively. The 'genome core' is endowed with several specific features: (i) a very high density of genes; (ii) genes with high GC levels, short introns and associated CpG islands; (iii) early replication; (iv) high recombination levels; and (v) high transcription levels and an open chromatin structure (which is reflected even in the lower compaction of $H3^+$ bands relative to $L1^+$ bands; see Fig. 3). In contrast, the 'empty quarter' is characterized by opposite features. It should be stressed that these different functional properties were present well before the two independent compositional transitions that took place in the reptilian ancestors of present day mammals and birds and concerned only one of the two gene spaces, namely the genome core.

A second line of evidence comes from the existence of two classes of genes in plants, a GC-rich class with no or few short introns, and a GC-poor class with numerous, long introns (Carels and Bernardi, 2000). The similarity of the properties of each class, as present in the genomes of maize and Arabidopsis, is particularly remarkable in view of the fact that these plants exhibit very large differences in genome size, average intron size, and DNA base composition. The functional relevance of the two classes of genes is stressed by the conservation in orthologous genes from maize and Arabidopsis not only of the number and location of introns, but also of the relative size of concatenated introns. Since housekeeping genes were found to be associated with GC-rich genes not only in Arabidopsis and maize (Chiapello et al., 1998), but also in vertebrates (see Larsen et al., 1992; Bernardi, 1995, for a review), shortage and small sizes of introns might be generally viewed as advantageous features for genes that are transcribed in a constitutive or at least in an extensive way. In the case of GC-poor genes, which are largely tissue-specific in vertebrates, the abundance and size of introns in these genes would be favourable for alternative splicing, an important mechanism of expression regulation of tissue-specific genes (Bell et al., 1998).

Although not mentioned by the authors, their results on intron size as a function of local GC level (see Fig. 7, a modification of Fig. 36c of the IHGSG paper) provide excellent support for the existence of two classes of genes in that they show a sharp transition in intron size, with a midpoint

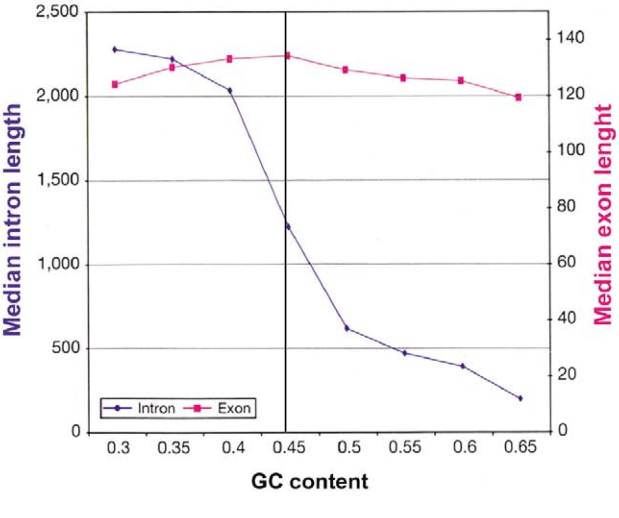

Fig. 7. Dependence of mean exon and intron lengths on GC content. For exons and introns, the local GC content was derived from alignments to finished sequence only, and was calculated from windows covering the feature or 10,000 bp centred on the feature, whichever was larger. The vertical straight line added in this paper to the original figure marks the midpoint of the transition. (Modified from Fig. 36c of International Human Genome Sequencing Consortium (2001).)

at about 45% GC, in agreement with the boundary between the genome core and the empty quarter of Fig. 6.

The plot of gene density as a function of GC in the human genome, as determined for 9315 genes (see Fig. 8, a modification of Fig. 36b of the IHGSG paper), is essentially identical to that published by us for 1610 and 4270 genes 10 and 5 years ago, respectively (Mouchiroud et al., 1991; Zoubak et al., 1996). Most surprisingly, the authors did not mention these results in spite of the fact that a review article presenting them (Bernardi, 2000a) is quoted in their reference list. Expectedly, the gene density values fall on two straight lines crossing each other at about 46% GC (see Fig. 6).

The explanations given by the authors for the results of Fig. 7 (their Fig. 36c) are astonishing. Indeed, they propose that "the variation in gene size and intron size can partly be explained by the fact that GC-rich regions tend to be gene-dense with many compact genes, whereas AT-rich regions tend to be gene-poor with many sprawling genes containing large introns." Obviously, this is not an explanation, but just a description of the situation. Also surprising is the statement that "the correlation between gene density and GC appears to be due primarily to intron size, which drops markedly with increasing GC content" (our Fig. 8; Fig. 36b of the IHGSG paper), since it ignores the fact that introns represent only 2–4% of the genome. Only at the end do the authors cautiously approach the obvious explanation and consider that "intergenic distance is also probably lower in high-GC areas". Their caveat that "this is hard to prove directly until all genes have been identified" suggests a high level of uncertainty in the estimate of the number of genes in the very section of the article providing such an estimate.

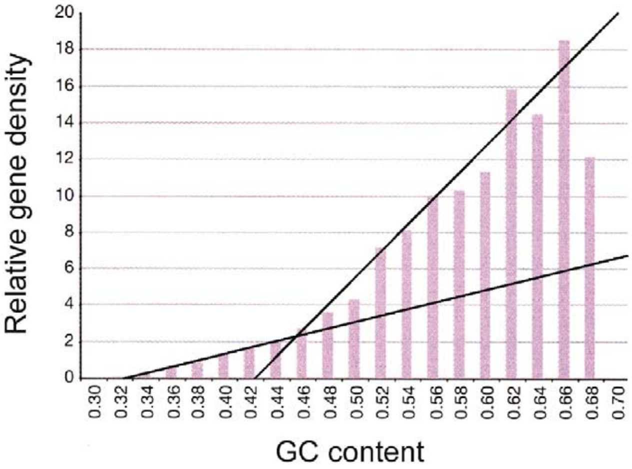

Fig. 8. Gene density as a function of GC content. Values are less accurate at higher GC levels because the denominator is small. The two slopes added here to the original figure concern the genes from GC-poor and GC-rich isochores, respectively. For 9315 known genes mapped to the draft genome sequence, the local GC content was calculated in a window covering either the whole alignment or 20,000 bp centred around the midpoint of the alignment, whichever was larger. Ns in the sequence were not counted. The GC content for the genome was calculated for adjacent non-overlapping 20,000 bp windows across the sequence. Both the gene and the genome distribution have been normalized to sum one. (Modified from Fig. 36b of International Human Genome Sequencing Consortium (2001).)

Fig. 9 shows an independent assessment of the two classes of genes as present in chromosomes 21 and 22. In this case, gene density was calculated on a Mb (megabase) scale. Again the two slopes cross each other at 45% GC.

5. Concluding remarks

The work of over 2000 people of the IHGSC can only be commended because of the amount of information and details that have been provided to researchers in many research areas spanning from medicine to evolution.

Two general remarks can, however, be made, in addition to the specific ones discussed above. The first one concerns the 11 points listed in the Introduction of the IHGSC paper. Although the authors do not make any explicit claim that these points correspond to new discoveries, nonetheless this is the impression the reader receives. A careful reading of the points shows, however, that none of them are original with one exception (on horizontal transfer), which raises, however, serious criticisms. The lack of any conceptual breakthrough is not surprising, since the themes approached in the paper were the subjects of active research in many laboratories around the world for the past three decades.

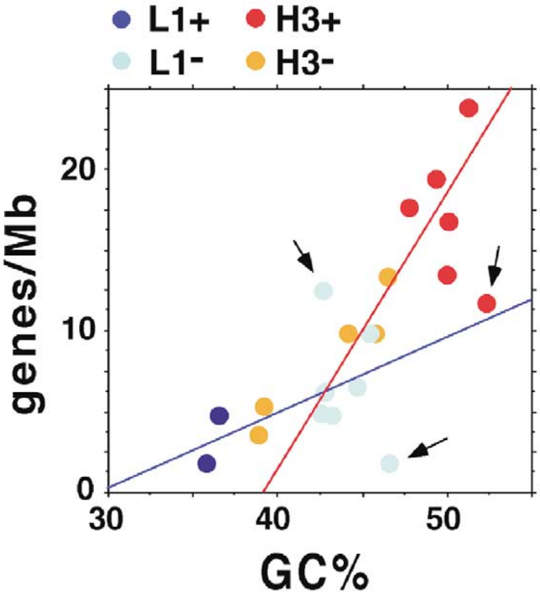

Fig. 9. The average GC level of each band of chromosomes 21 and 22 was plotted against its gene density. The highest and lowest gene densities were found in H3$^+$ and L1$^+$ bands, respectively, as expected. The remaining G and R bands (the L1$^-$ and H3$^-$ bands) showed gene densities that are correlated with their GC level, independently of their cytogenetic band type (G or R). Three points, indicated by the arrows, represent three outliers (two L1$^-$ bands and one H3$^+$ band) not taken into consideration when drawing the regression line. Inclusion of these points does not significantly change the lower slope and only slightly changes the higher slope (From Saccone et al. (2001).)

Obviously, this does not detract anything from the value of the vast amount of information presented in the paper.

The second remark concerns the denial of the very existence of isochores. While a mistake in itself, the major problems are its consequences which, apparently, were not realized by the authors. The first one is the tacit denial of a compositionally discontinuous sequence organization and the return to the continuous compositional spectrum for the human genome that was the predominant view until the work of Filipski et al. (1973). The second consequence is the denial of "an important level of genome organization, insofar as gene density (Zoubak et al., 1996), gene length (Duret et al., 1995), and patterns of codon usage (Sharp et al., 1995), as well as the distribution of different classes of repetitive elements (Soriano et al., 1983; Duret et al., 1995), are all correlated with GC content" (Fullerton et al., 2001). Other properties that could be added to the list are replication timing, recombination frequency (Fullerton et al., 2001), chromosomal banding, and stability and transcription of integrated sequences. In other words, the second consequence is the denial of "a fundamental level of genome organization" (Eyre-Walker and Hurst, 2001).

Acknowledgements

I wish to thank Fernando Alvarez-Valin, Giacomo Bernardi, Nicolas Carels, Maria Luisa Chiusano, Oliver Clay, Stéphane Cruveiller, Giuseppe D'Onofrio, Concetta Federico, Kamel Jabbari, Gabriel Macaya, Hector Musto, Adam Pavliček, Hector Romero, Salvatore Saccone and Serguei Zoubak for their contributions to the investigations mentioned here, and Giuseppe D'Onofrio, Giorgio Matassi, Adam Pavliček and especially Oliver Clay for their critical reading of this article.

References

Alvarez-Valin, F., Bernardi, G., 2001. Isochores and mutational bias. Genetics, submitted for publication.

Arcot, S.S., Adamson, A.W., Risch, G.W., LaFleur, J., Robichaux, M.B., Lamerdin, J.E., Carrano, A.V., Batzer, M.A., 1998. High-resolution cartography of recently integrated human chromosome 19-specific Alu fossils. J. Mol. Biol. 281, 843–856.

Bell, M.V., Cowper, A.E., Lefranc, M.P., Bell, J.I., Screaton, G.R., 1998. Influence of intron length on alternative splicing of CD44. Mol. Cell. Biol. 18, 5930–5941.

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

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

Bernardi, G., 2000b. The compositional evolution of vertebrate genomes. Gene 259 (1), 31–43.

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.

Bernardi, G., Mouchiroud, D., Gautier, C., Bernardi, G., 1988. Compositional patterns in vertebrate genomes: conservation and change in evolution. J. Mol. Evol. 28, 7–18.

Bernardi, G., Mouchiroud, D., Gautier, C., 1993. Silent substitution in mammalian genomes and their evolutionary implications. J. Mol. Evol. 37, 583–589.

Carels, N., Bernardi, G., 2000. Two classes of genes in plants. Genetics 154, 1819–1825.

Chiapello, H., Lisacek, F., Caboche, M., Henaut, A., 1998. Codon usage and gene function are related in sequences of Arabidopsis thaliana. Gene 209 (1–2), GC1–GC38.

Chiusano, M.L., D'Onofrio, G., Alvarez-Valin, F., Jabbari, K., Colonna, G., Bernardi, G., 1999. Correlations of nucleotide substitution rates and base composition of mammalian coding sequences with protein structure. Gene 238, 23–31.

Chiusano, M.L., Alvarez-Valin, F., Di Giulio, M., D'Onofrio, G., Ammirato, G., Colonna, G., Bernardi, G., 2000. Second codon positions of genes and the secondary structures of proteins. Relationships and implications for the origin of the genetic code. Gene 261, 63–69.

Clay, O., 2001. Standard deviations and correlations of GC levels in DNA sequences. Gene in press.

Clay, O., Bernardi, G., 2001a. The isochores in human chromosomes 21 and 22. Biochem. Biophys. Res. Commun. 285, 855–856.

Clay, O., Bernardi, G., 2001b. Compositional heterogeneity within and among isochores in mammalian genomes. II. Some general comments. Gene 276, 39–45.

Clay, O., Cacciò, S., Zoubak, S., Mouchiroud, D., Bernardi, G., 1996. Human coding and non-coding DNA: compositional correlations. Mol. Phylogenet. Evol. 5, 2–12.

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, 25–31.

Cruveiller, S., D'Onofrio, G., Jabbari, K., Bernardi, G., 1999. Different hydrophobicities of orthologous proteins from Xenopus and man. Gene 238, 15–21.

Cruveiller, S., D'Onofrio, G., Bernardi, G., 2000. The compositional transition between the genomes of cold- and warm-blooded vertebrates: codon frequencies in orthologous genes. Gene 261, 71–83.

Cuny, G., Soriano, P., Macaya, G., Bernardi, G., 1981. The major components of the mouse and human genomes: preparation, basic properties and compositional heterogeneity. Eur. J. Biochem. 115 (2), 227–233.

DeFilippis, V., Villarreal, L.P., Salzberg, S.L., Eisen, J.A., 2001. Lateral gene transfer or viral colonization? Science 293, 1040.

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

D'Onofrio, G., Bernardi, G., 1992. A universal compositional correlation among codon positions. Gene 110, 81–88.

D'Onofrio, G., Mouchiroud, D., Aissani, B., Gautier, C., Bernardi, G., 1991. Correlations between the compositional properties of human genes, codon usage and aminoacid composition of proteins. J. Mol. Evol. 32, 504–510.

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

Dutrillaux, B., 1973. Nouveau système de marquage chromosomique: les bandes T. Chromosoma 41, 395–402.

Eyre-Walker, A., 1992. Evidence that both G + C-rich and G + C-poor isochores replicate early and late in the cell cycle. Nucleic Acids Res. 20, 1497–1501.

Eyre-Walker, A., 1999. Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. Genetics 152, 675–683.

Eyre-Walker, A., Hurst, L.D., 2001. The evolution of isochores. Nat. Rev. Genet. 2, 549–555.

Federico, C., Saccone, S., Bernardi, G., 1998. The gene-richest bands of human chromosomes replicate at the onset of the S-phase. Cytogenet. Cell Cenet. 80, 83–88.

Federico, C., Andreozzi, L., Saccone, S., Bernardi, G., 2000. Gene density in the Giemsa bands of human chromosomes. Chromosome Res. 8 (8), 737–746.

Fickett, J.W., Torney, D.C., Wolf, D.R., 1992. Base compositional structure of genomes. Genomics 13, 1056–1064.

Filipski, J., Thiery, J.P., Bernardi, G., 1973. An analysis of the bovine genome by Cs_2SO_4-Ag^+ density gradient centrifugation. J. Mol. Biol. 80, 177–197.

Francino, M.P., Ochman, H., 1999. Isochores result from mutation not selection. Nature 400, 30–31.

Francke, W., 1994. Digitized and differentially shaded human chromosome ideograms for genomic applications. Cytogenet. Cell Genet. 6, 206–219.

Fukagawa, T., Sugaya, K., Matsumoto, K., et al., 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.

Fullerton, S.M., Carvalho, A.B., Clark, A.G., 2001. Local rates of recombination are positively correlated with GC content in the human genome. Mol. Biol. Evol. 18, 1139–1142.

Graur, D., Li, W.H., 2000. Fundamentals of Molecular Evolution, 2nd Edition. Sinauer Associates, Sunderland, MA.

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

Holmquist, G.P., Filipski, J., 1994. Organization of mutations along the genome: a prime determinant of genome evolution. Trends Ecol. Evol. 9, 65–68.

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

Jabbari, K., Bernardi, G., 1998. CpG doublets, CpG islands and Alu repeats in long human DNA sequences from different isochore families. Gene 224, 123–128.

Jurka, J., 1997. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. Proc. Natl. Acad. Sci. USA 94, 1872–1877.

Larsen, F., Gundersen, G., Lopez, R., Prydz, H., 1992. CpG islands as gene markers in the human genome. Genomics 13, 1095–1107.

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.

Mathews, C.K., Ji, J., 1992. DNA precursor asymmetries, replication fidelity, and variable genome evolution. Bioessays 14, 295–301.

Meunier-Rotival, M., Soriano, P., Cuny, G., Strauss, F., Bernardi, G., 1982. Sequence organization and genomic distribution of the major family of interspersed repeats of mouse DNA. Proc. Natl. Acad. Sci. USA 79, 355–359.

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.

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

Pavliček, A., Pačes, J., Clay, O., Bernardi, G., 2001b. A compact view of isochores in the draft human genome sequence. Genome Biol. submitted for publication.

Roelofs, J., Van Haaster, M., 2001. Genes lost during evolution. Nature 411, 1013–1014.

Rolfe, R., Meselson, M., 1959. The relative homogeneity of microbial DNA. Proc. Natl. Acad. Sci. USA 45, 1039–1043.

Rynditch, A.V., Zoubak, S., Tsyba, L., Tryapitsina-Guley, N., Bernardi, G., 1998. The regional integration of retroviral sequences into the mosaic genomes of mammals. Gene 222, 1–16.

Saccone, S., De Sario, A., Della Valle, G., Bernardi, G., 1992. The highest gene concentrations in the human genome are in T bands of metaphase chromosomes. Proc. Natl. Acad. Sci. USA 89, 4913–4917.

Saccone, S., De Sario, A., Wiegant, J., Rap, A.K., Della Valle, G., Bernardi, G., 1993. Correlations between isochores and chromosomal bands in the human genome. Proc. Natl. Acad. Sci. USA 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., Solovei, I., Croquette, M.F., Della Valle, G., Bernardi, G., 1999. Identification of the gene-richest bands in human prometaphase chromosomes. Chromosome Res. 7, 379–386.

Saccone, S., Pavliček, A., Federico, C., Pačes, J., Bernardi, G., 2001. Genes, isochores and bands in human chromosomes 21 and 22. Chromosome Res. in press.

Schmid, C.W., 1998. Does SINE evolution preclude Alu function? Nucleic Acids Res. 26, 4541–4550.

Sharp, P.M., Averof, M., Lloyd, A.T., Matassi, G., Peden, J.F., 1995. DNA sequence evolution: the sounds of silence. Philos. Trans. R. Soc. Lond. B Biol. Sci. 349, 241–247.

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

Smith, N.G.C., Eyre-Walker, A., 2001. Synonymous codon bias is not caused by mutation bias in G + C rich genes in humans. Mol. Biol. Evol. 18, 982–986.

Soriano, P., Meunier-Rotival, M., Bernardi, G., 1983. The distribution of interspersed repeats is non-uniform and conserved in the mouse and human genomes. Proc. Natl. Acad. Sci. USA 80, 1816–1820.

Stanhope, M.J., Lupas, A., Italia, M.J., Koretke, K.K., Volker, C., Brown, J.R., 2001. Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. Nature 411, 940–944.

Sueoka, N., 1988. Directional mutation pressure and neutral molecular evolution. Proc. Natl. Acad. Sci. USA 85, 2653–2657.

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

Venter, C., et al., 2001. The sequence of the human genome. Science 291, 1304–1351.

Wolfe, K.H., Sharp, P.M., Li, W.H., 1989. Mutation rates differ among regions of the mammalian genome. Nature 337, 283–285.

Zerial, M., Salinas, J., Filipski, J., Bernardi, G., 1986. Gene distribution and nucleotide sequence organization in the human genome. Eur. J. Biochem. 160, 479–485.

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