ELSEVIER

Gene 174 (1996) 95–102

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

The gene distribution of the human genome

Serguei Zoubak *[1], Oliver Clay, Giorgio Bernardi

Laboratoire de Génétique Moléculaire, Institut Jacques Monod, 2 Place Jussieu, 75005 Paris, France

Received 17 April 1996; accepted 10 June 1996

Abstract

Linear correlations exist between the GC levels of third codon positions (GC_3) of individual human genes and the GC levels of long genomic sequences and DNA molecules (50–100 kb in size) embedding the genes. These linear relationships allow the positioning of the GC_3 histogram of cDNA sequences from the databases relative to the CsCl profile of human DNA. In turn, this allows an estimate of the relative concentrations of genes in genomic regions of different GC content. An estimate obtained by using current sequence data and Gaussian decompositions of the GC_3 histogram and of the CsCl profile indicates that the GC-richest (non-ribosomal) component of the human genome is at least 17 times as gene-rich as the GC-poor regions. Moreover, our results suggest that the most recent physical maps of the human genome consisting of overlapping YACs cover less than 50% of the genes.

Keywords: Chromosomes; Isochores; Physical and genetic mapping

1. Introduction

The human genome is a mosaic of isochores, long DNA segments (>300 kb, on the average), which are characterized by a high compositional homogeneity (above a 3 kb size level). Isochores belong to five families covering a wide GC range (GC is the molar fraction of guanine + cytosine in DNA). GC-poor isochores of the L1 and L2 families, which make up about 62% of the human genome, are poor in genes. GC-rich isochores of the H1, H2 and H3 families, which represent about 22%, 9% and 4% of the human genome, respectively (the rest corresponding to satellite and rDNA), are increasingly rich in genes (see Bernardi, 1995, for a recent review).

The parallelism between GC-richness and gene richness, first noticed over ten years ago (Bernardi et al., 1985), has been put on a quantitative basis more recently (Mouchiroud et al., 1991), and can be understood on the basis of the compositional evolution of the vertebrate genome (Bernardi, 1993, 1995). In previous work (Mouchiroud et al., 1991), estimating of gene concentrations in different isochore families was achieved by positioning a histogram of the compositional distribution of GC_3 (the GC levels of third codon positions) from 1600 human genes relative to the CsCl profile of human DNA, which corresponds to the compositional distribution of human DNA molecules, and by dividing gene numbers by DNA amounts in corresponding bins. The positioning of the GC_3 histogram relative to the CsCl profile was obtained via the correlation between the GC_3 levels of individual human genes and the GC levels of the DNA fragments or long segments in which the corresponding genes were located (Aïssani et al., 1991). This correlation has recently been reinvestigated in more detail (Clay et al., 1996).

In the present work, the positioning of the GC_3 histogram relative to the CsCl profile was based also on the Gaussian decompositions of both CsCl profile and GC_3 histogram of 4270 human genes, and on the resulting correlation between GC_3 levels of families of coding sequences (GC_3 histogram components) and GC levels of isochore families (CsCl profile components). We have used the gene concentration curve obtained in this way to assess the coverage of genes by the most recent physical map of the human genome.

* Corresponding author. Tel. +33 1 44278172; Fax +33 1 44277977; e-mail: bernardi@citi2.fr

[1] Permanent address: Institute of Molecular Biology and Genetics, National Ukrainian Academy of Sciences, 150 Zabolotnogo St., 252627 Kiev, Ukraine.

Abbreviations: bp, base pairs; EST(s), expressed sequence tag(s); GC, molar fraction of guanine + cytosine in DNA; GC_3, GC of third codon positions of genes; kb, kilobases; Mb, megabases; rDNA, ribosomal DNA.

0378-1119/96/$15.00 © 1996 Elsevier Science B.V. All rights reserved
PII S0378-1119(96)00393-9

2. Materials and methods

2.1. Histogram of GC_3 values

The database of 4270 genes and their GC_3 levels was extracted from GenBank Release 89.0 (July 1995; Benson et al., 1994), and is the same as that used in a recent study (Clay et al., 1996). This sample of genes contained no obvious redundancies and no T-cell receptor, MHC and immunoglobulin sequences, since the presence of immunologically relevant polymorphisms at these loci (especially at the MHC locus) has led to their increasing overrepresentation in the database, often in the form of different alleles or variants of the same gene (cf. e.g. Fields et al., 1994). The resulting number of GenBank sequences is similar to that recently obtained by Adams et al. (1995) after redundancy elimination. Histograms of GC_3 levels (percentages) were constructed by partitioning the values into bins of width 2.5%.

2.2. CsCl profile

The CsCl profile used was obtained in previous studies (Macaya et al., 1976; Thiery et al., 1976). Conversions between buoyant density and GC level were obtained using the relation of Schildkraut et al. (1962).

2.3. Calculation of Gaussian components

Gaussian components were calculated independently for the CsCl curve and the GC_3 histogram. In each case, the decomposition was found by minimizing the residual sum of squares using the Levenberg-Marquard algorithm as implemented in the program Igor (WaveMetrics; Press et al., 1988), for five (CsCl curve) or four (GC_3) components of equal widths. The decompositions were robust with respect to small perturbations of the data, except for the relative amplitude and mean GC values of the two smallest, GC-richest components in the CsCl profile decomposition (H3, rDNA), which were stabilized by fixing the position of the rDNA component prior to decomposition. The decomposition of the GC_3 histogram was repeated twice, with almost identical results, (a) placing no constraints on either amplitudes or positions of the components (shown in Fig. 2), and (b) fixing positions (but not amplitudes) of the components at those of the corresponding components of the CsCl profile (not shown). To ensure that the GC_3 decomposition was not significantly dependent on the histogram bar width (2.5% GC_3), it was repeated for bin widths of 1%, 2% and 4% GC_3, with very similar results.

2.4. Orthogonal regression analyses

Regression lines were calculated by the orthogonal (major axis) regression formula $y-\langle y\rangle = s(x-\langle x\rangle)$, $s=(b+\sqrt{b^2+4c^2})/(2c)$. Here, x and y are GC and GC_3, respectively, s is the slope, $\langle x\rangle$, $\langle y\rangle$ are the coordinates of the center of mass of the scatterplot, i.e. the mean GC and GC_3 levels, $b=v_y-v_x$ is the difference of the sample variances, and c is the covariance. In the absence of any errors in the individual data points, the confidence intervals for the slopes would be around $s\pm 0.3$, as calculated by the formula of Jolicoeur (1990). Here, however, the confidence intervals for slope and intercept are larger than in the ideal case, due to measurement errors for modal buoyant densities and local fluctuations in GC_3 values. The propagation of such errors to the slope and intercept (Roe, 1992) can be estimated only roughly, since the measurement errors and local fluctuations are not easily quantifiable. If we assume errors of 3–5% in both GC and GC_3 values, the total errors in the slopes are around 0.4–0.5, and those in the intercept are around 15–20.

2.5. Classification of genes according to CpG island status

To obtain an idea of the influence of CpG islands in the scatterplot regressions, these regressions were repeated for only the 'non-CpG island genes' among them, i.e. for those sequences listed in the CpGIsle database (Release March 1995; see Larsen et al., 1992) as genes in which no CpG islands had been found, or (for sequences with experimentally determined embedding GC levels) which contained no CpG island region when viewed with the EGCG program CPGPLOT (R. Lopez; see Larsen et al., 1992) using window sizes 100 and 300 nt.

3. Results and discussion

3.1. Relative positioning of GC_3 histogram and CsCl profile

In a first attempt to find the equation for conversion between GC_3 values of genes and the GC values of the 50–100 kb DNA molecules embedding them that determine the CsCl profile, orthogonal regressions obtained from recent scatterplots (Clay et al., 1996) were used, as in previous work (Mouchiroud et al., 1991; Aïssani et al., 1991). The linear relationship between the GC_3 values of genes and the GC levels of the DNA embedding them was almost identical whether the latter were measured experimentally (see Aïssani et al., 1991 for details) or estimated from the GC levels of long (>1 kb) 3′ flanking sequences located far (>500 nt) from the genes' stop codons (Fig. 1, dashed line). A second relationship, independent of the scatterplots, was obtained (Fig. 1, solid line) from a regression of modal (or mean) GC_3 values of the Gaussian components in the histogram of the GC_3 values of genes against modal (or mean) buoyant

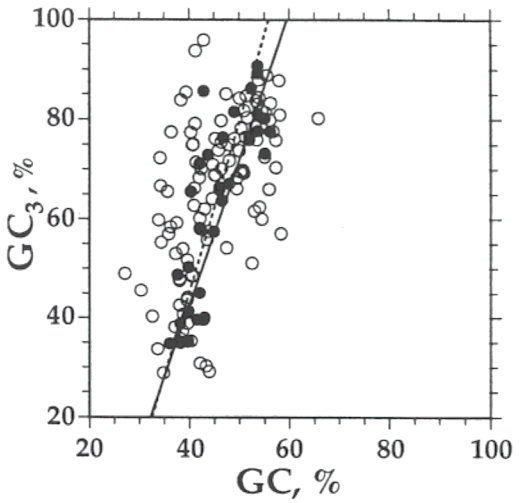

Fig. 1. Correlations between GC_3 of coding regions of genes and the GC levels of fractions or YACs in which the genes were localized (filled circles), and of 3′ flanking sequences farther than 500 bp from the stop codon (open circles). The corresponding, essentially coincident orthogonal regression lines, indicated here by a single dashed line, are given by $GC_3 = 3.45\ GC - 94.6$ ($R = 0.82$, $N = 32$; fractions and YACs) and $GC_3 = 3.39\ GC - 88.2$ ($R = 0.56$, $N = 103$; far 3′ sequences). The relation obtained by comparing the independent Gaussian decompositions of GC and GC_3 distributions (see text and Fig. 2), is indicated by a solid line, and given by $GC_3 = 2.92\ GC - 74.3$. This solid line happens to be almost indistinguishable from the regression lines for those genes shown here that were known to contain no CpG islands, $GC_3 = 2.89\ GC - 71.9$ ($R = 0.87$, $N = 13$; fractions and YACs) and $GC_3 = 2.90\ GC - 66.4$ ($R = 0.67$, $N = 30$; far 3′ sequences).

density values of the Gaussian components in the CsCl profile of total human DNA (see Fig. 2, and below). This regression line (2.92 GC − 74.3, after converting buoyant densities to GC values) is intermediate in slope between the one obtained in previous work ($GC_3 = 2.74\ GC - 64.6$; see Mouchiroud et al., 1991; Aïssani et al., 1991) and the one just described, and happens to be almost indistinguishable from scatterplot regression lines for genes known to contain no CpG islands (see the legend to Fig. 1).

3.2. Compositional distribution of gene frequencies

The five families of isochores in the human genome (Macaya et al., 1976; Cuny et al., 1981; Bernardi et al., 1985; Zerial et al., 1986) are those belonging to the five (non-ribosomal) components of the CsCl profile, as obtained by analytical ultracentrifugation of human DNA to sedimentation equilibrium, and are denoted in order of increasing mean (or modal) GC as L1, L2, H1, H2 and H3 (Bernardi et al., 1985). The mean values of L1 and L2 are very close, and although these two components are distinguishable when the CsCl profiles of fractions of DNA of different modal GC values are compared (Macaya et al., 1976), they cannot be resolved accurately from the CsCl curve of total human DNA using numerical techniques, and were treated as one double-component (L) in the following.

Fig. 2A shows the Gaussian decomposition of the CsCl mainband into five equal-width components, of which the first four correspond to L, H1, H2 and H3, and the fifth to ribosomal DNA. Although the widths of the CsCl components increase slightly with buoyant density and with GC (Cuny et al., 1981), the equal-width approximation has the benefit of simplicity and of allowing stability of the decomposition, even in the presence of noise in the CsCl curve, which is inevitable under the experimental conditions of analytical ultracentrifugation. To increase the numerical reliability of the decomposition in the DNA-poor GC-richest region of the CsCl curve, the position of the fifth, very GC-rich component, formed by the 280–400 repeats of the 43-kb rDNA unit, was fixed at 58.4% GC (Zerial et al., 1986;

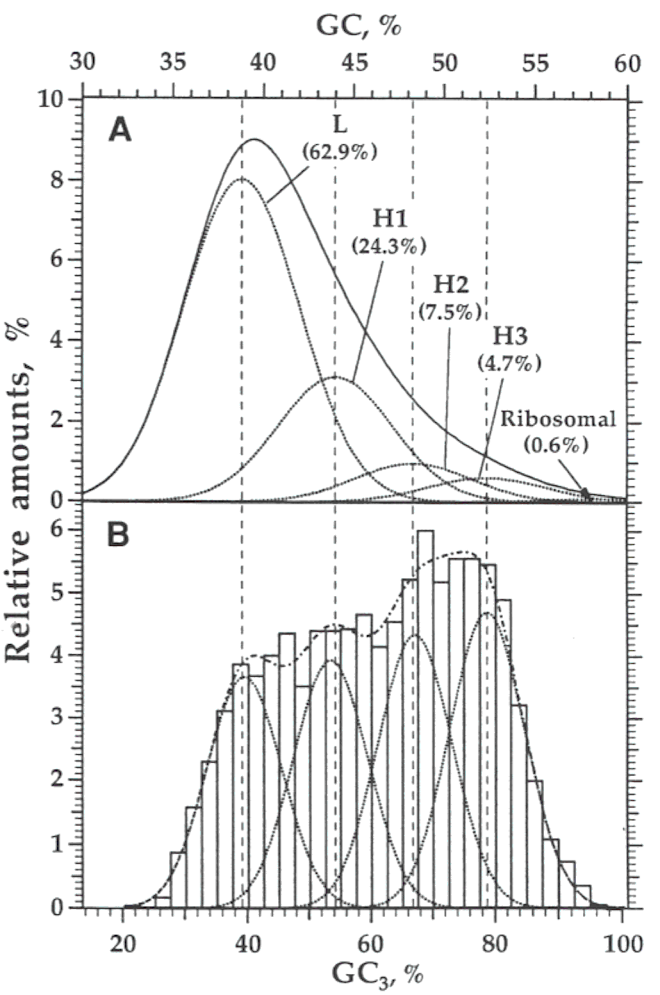

Fig. 2. **(A)** Decomposition of the CsCl profile of human DNA into 4 major Gaussian components and a small ribosomal DNA component. **(B)** Independent decomposition of the GC_3 histogram of 4270 human genes into 4 Gaussian components. Vertical dashed lines show the peak (mean) GC_3 values of the CsCl curve's components; the conversion relation used here to display the distributions is G $C_3 = 2.92\ GC - 74.3$.

Gonzalez and Sylvester, 1995). The resulting decomposition allocated approximately 0.6% of the total human DNA to this rDNA component, a value close to accepted estimates (≈0.5–0.6%; cf. Gonzalez and Sylvester, 1995).

The Gaussian analysis of DNA components presented here is in close agreement with previous estimates (Macaya et al., 1976; Thiery et al., 1976; Cuny et al., 1981; Zerial et al., 1986) obtained via other approaches, which were all based on experimental fractionation of DNA followed by recentrifugation in the presence of DNA-specific ligands. The present decomposition accurately resolved all of the major components of the human genome, except for the relative contributions of L1 and L2 to L, directly from the CsCl curve of total human DNA, using only the Levenberg-Marquard algorithm (see Materials and Methods). Although the genomes of other species will not always allow such stable resolution of major genomic components directly from the CsCl profile of total DNA, especially if there are substantial contributions from satellite components, the convergence of experimental and Gaussian decomposition methods for the human genome suggests that further development of the direct decomposition approach could provide a powerful tool for the analysis of genomes (see also Thiery et al., 1976; Macaya et al., 1976).

Fig. 2B shows the histogram of GC_3 values of a non-redundant collection of 4270 human cDNA and genomic DNA sequences, and its decomposition, independent of any information from the CsCl profile's decomposition, into four equal-width Gaussian components, with no restraints on positions or amplitudes of the components. The maxima of the first four components of the CsCl curve and those of the GC_3 histogram are related by a remarkably linear relationship (see below) with regression equation $GC_3 = 2.92\ GC - 74.3$ (indicated by a solid line in Fig. 1), which has been used to position the two independent decompositions in Fig. 2 so that they can be compared. The linearity of this relationship, its identity with the relationship obtained for genes known not to contain CpG islands, and its closeness to the regressions obtained from the scatterplots, strongly suggest that the GC_3 histogram components represent the (approximate) distributions of the GC_3 levels of genes from the corresponding CsCl components, and that if larger data sets were available for a scatterplot they would yield a similar regression equation.

Based on these considerations, we chose the above equation to align the GC and GC_3 scales calibrating human DNA in the following. The linearity of the relation between the component positions in the DNA GC and genic DNA GC_3 distributions ($R = 0.9995$) can be seen by comparing Fig. 2B with Fig. 2A, in that the peaks of the GC_3 histogram components coincide almost exactly, after alignment of scales using the linear equation, with those of the DNA components (vertical dashed lines). As a check, a second decomposition of the GC_3 histogram was performed by fixing the positions of the four components in advance so that they align exactly with those in Fig. 2A. This 'fixed' decomposition yielded relative amplitudes and results (see below) almost identical to those for the 'free' decomposition of Fig. 2B, and will not be discussed further here.

The decompositions provide the first of two possible ways of estimating ratios of gene concentrations in GC-rich and GC-poor regions of the genome (Fig. 3), since the gene concentration in each component can be calculated by dividing the component of the GC_3 decomposition by the corresponding component of the GC distribution. For example, the gene concentration in the highest GC components (H3 + rDNA) is 17 times higher than that in the lowest GC components (L1 + L2, i.e. L). A slightly different way of estimating this ratio is to compare the average value of the gene density curve (see below) in the region $74\% < GC_3 < 82\%$, i.e. in the part of H3 in which the gene density curve is presumably not a large underestimate, with that in the region $GC_3 < 47\%$, i.e. in L. This again gives a ratio of about 17 (≈10/0.59). This value is similar to a previous estimate, 16, obtained for a smaller number of sequences

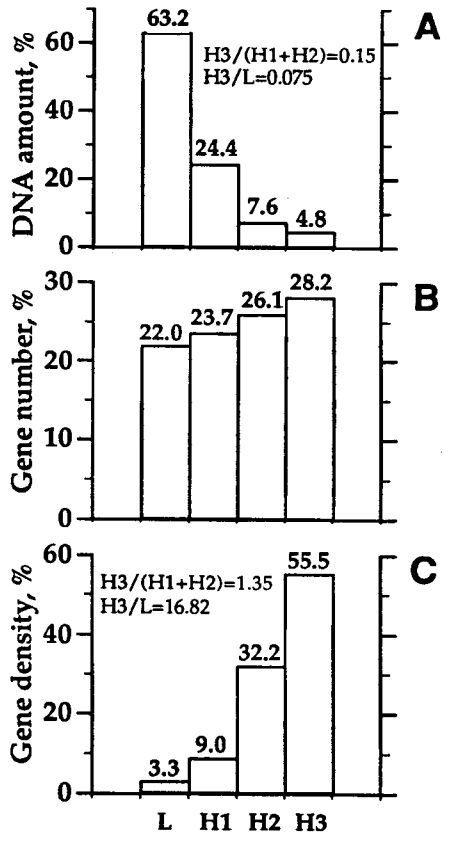

Fig. 3. Relative amounts of human DNA (A), relative numbers of genes (B) and relative gene densities (C) in the 4 Gaussian components corresponding to the L, H1, H2, and H3 isochore families (see Fig. 2). The slight differences between the values in panel A and those in Fig. 2A, amounting to 0.6%, are due to exclusion of rDNA in this calculation.

and using essentially this latter method (Mouchiroud et al., 1991).

The concordance between GC and GC_3 decompositions, i.e. the linearity of the relation between corresponding peaks, suggests that the partitioning of human (non-ribosomal) DNA into four components can also be used to reliably assign genes to their expected DNA components, based on coding sequence information (GC_3 data) alone (cf. also Mouchiroud et al., 1991). The approximate GC_3 boundaries, i.e. the values of GC_3 where the probability of a gene belonging to a given isochore family becomes higher than that of it belonging to an adjacent isochore family, are near 47% (L/H1), 61% (H1/H2) and 74% (H2/H3) GC_3, according to the decompositions shown here.

The first and third of these boundaries have been estimated previously, using different criteria, and using a different estimate of the GC_3 vs. GC correspondence ($GC_3 = 2.74\,GC - 64.6$), to lie at 57.5% and 75% GC_3, respectively (Mouchiroud et al., 1991). While the previous H2/H3 boundary estimate agrees with the present one, the previous L/H1 boundary estimate was higher than the present one, and thus assigned some genes to L that by the present criteria are more likely to belong to H1. The new L/H1 boundary estimate is the most notable of the few differences from the results of the previous study (Mouchiroud et al., 1991), the present estimates of the gene concentrations and their ratios being very similar to the previous ones. This confirms the stability of the results in spite of a large increase in the sample (4270 vs. 1600 genes) and the use of a different method. It should also be mentioned that the gene concentrations, their ratios, and the GC_3 boundaries obtained via the present method are essentially independent of the relation of Schildkraut et al. (1962) for converting buoyant densities to GC levels, which in the case of the human genome may be of limited accuracy (Cuny et al., 1981).

3.3. Gene concentration curve

Fig. 4A shows the GC_3 histogram and CsCl curve from Fig. 2, together with the gene concentration curve obtained by division of these distributions at GC intervals corresponding to 2.5% GC_3. The gene concentration starts with undetectable levels in the GC-poorest 5% of the genome, rises linearly over the next 25% of the genome, and remains low until the beginning of H1. In other words, gene concentration increases slowly over the L1 isochore family and remains relatively constant over the L2 isochore family (which represent 29% and 33% of the genome, respectively; Cuny et al., 1981). It then becomes steeper within the H1 compartment, and is again almost linear, although steeper yet, throughout H2 and possibly also throughout H3.

Two remarks should be made about the ends of the

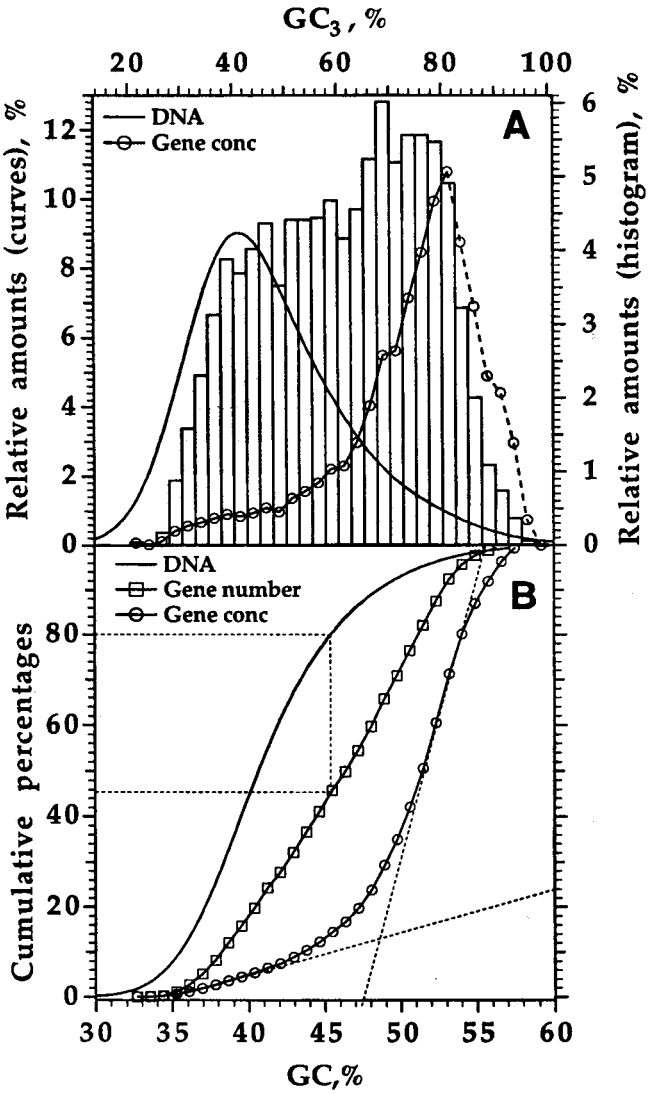

Fig. 4. (A) Histogram of GC_3 levels of 4270 human genes superimposed on the CsCl profile of total human DNA, together with the profile of gene concentration (circles), showing the ratio of these two distributions for each bar of the histogram. Profiles and histogram are each normalized to an area of 100%. The conversion between GC_3 (top scale) and the corresponding GC levels of the isochores (bottom scale) is given by the relation $GC_3 = 2.92\,GC - 74.3$. The gene concentration profile was calculated using GC_3 intervals of 2.5% (bar widths). (B) Cumulative plots of the CsCl profile, gene number histogram and gene concentration profile shown in (A). The broken lines on the left of the figure indicate that a YAC coverage of the GC-poor 80% of the human genome will correspond to the coverage of only about 45% of the genes. The broken lines on the right of the figure indicate the two slopes in cumulative gene concentration.

gene distribution. On the low GC side, Fig. 4A indicates the absence of genes in a sizable fraction of the 50–100 kb DNA molecules forming the CsCl band. This is understandable in terms of the scarcity of genes in the GC-poorest isochores, as exemplified by the low gene score of Giemsa bands (Gardiner et al., 1990; Pilia et al., 1993; Xu et al., 1995), and in terms of the corresponding large size of intergenic sequences and even of introns

(see Bettecken et al., 1992; Duret et al., 1995). On the high GC side, it should be pointed out that the many (280–400) very GC-rich 43-kb rDNA repeats contribute to the main-band CsCl profile, yet these genes have not been included in the GC_3 histogram. If such genes are taken into account, the concentration of genes should continue to increase or remain high into the GC-richest regions of human DNA (e.g. 58–59% GC for all long fragments consisting entirely of rDNA; the calculated curve is therefore shown with dashed lines). This suggests that the ratio of gene concentration in isochore families H3 and L is higher than 17, the value estimated above for H3+rDNA over L. The presence of high gene concentrations continuing well beyond 53% GC (Fig. 4A, dashed lines) is supported also by the 36 very long (>50 Kb) human genomic DNA sequences containing genes that are currently available in GenBank (May 1996), of which nine sequences are expected, or have been confirmed (De Sario et al., 1996), to be in H3 isochores. Five of these H3 sequences have GC levels higher than 53%. One of these is exceptional in that it has an average GC level of 62.5% and covers a single long gene (≈50 kb, for polycistic kidney disease-associated protein PKD1; HUMPKD1GEN), with 50 exons, in a GC-rich H3* band (16p13.3; Saccone et al., 1996, Table 2). Another of these sequences (HUMXPDG1, with an average GC level of 54%) is located completely or largely in a GC-rich $H3^+$ band (19q13.3; Saccone et al., 1996, Table 2). The remaining sequences (HSU52111/2, each >150 kb, and HUMFLNG6PD) are from H3 isochore(s) in the band Xq28 (DeSario et al., 1996), and have average GC levels of 55-57%.

An additional complicating factor working in the same direction is a possible slight overestimate of the end of the CsCl curve due to a faint upward slope of the baseline, which is however difficult to estimate accurately. Finally, another independent reason for considering 17 an underestimate of the gene concentration ratio is (a) that housekeeping genes are currently underrepresented in data banks; (b) that these genes are almost always associated with CpG islands, whereas this is not the case for other genes (Larsen et al., 1992); and (c) that the distribution of CpG island genes attains a more pronounced maximum in the GC-richest isochores than does that of genes in general (Aïssani and Bernardi, 1991a; Aïssani and Bernardi, 1991b; and paper in preparation).

Fig. 4B shows the three cumulative frequency distributions corresponding to the distributions (probability densities) represented in Fig. 4A. While the sigmoid shape for DNA is expected because of the bell shape of the CsCl curve, the linearity of the cumulative gene number curve, or, equivalently, the flatness of the central part of the GC_3 histogram (i.e. the almost equal amplitudes of this histogram's Gaussian components; see Fig. 2) is a novel, remarkable feature. Moreover, the cumulative gene concentration curve defines two slopes, which correspond to the L+H1 and to the H3 isochores. It may be asked if the flatness in the GC_3 histogram, and the correspondingly high gene concentration in the H3 component, could be partly due to an overrepresentation of GC rich genes in the nucleotide database. Although this possibility cannot be absolutely ruled out, we know of no clear indication for such a representational bias in GenBank. As a double check, we plotted GC histograms of 3′-directed cDNA sequences in the BodyMap database (see Okubo et al., 1992; Kawamoto et al., 1996) that were available in GenBank (HUMGS0*), as well as GC histograms of 3′ ESTs (expressed sequence tags; Adams et al., 1991) from two releases of the nonredundant EST database UniGene (formerly UniGene/UniEST; Boguski and Schuler, 1995). On the basis of known compositional correlations between the GC_3 levels of human genes and the GC levels of their noncoding 3′ flanks (Clay et al., 1996), we expected a histogram of the GC levels of purely noncoding 3′ ESTs to have a similar shape to that of the GC_3 levels of the genes in GenBank, although possibly with a lower resolution of the peaks, i.e. with a larger variance of the components, since ESTs tend to be short (typically 300–500 nt or less). The 3′-directed cDNA sequence sets analyzed (1040 sequences ≥200 bp, 516 sequences ≥300 bp) indeed showed flat-topped GC histograms, extending somewhat further than those for the genes in Fig. 2. For the larger, but not 3′-directed, UniGene database, the 'most reliable' sequences among those in an earlier release that were designated as 3′ sequences (September 1995; upper-case nucleotides only, 5701 out of 22088 3′ sequences) gave a GC histogram, and thus a contribution of H3 genes, similar to the one shown here. In the 3′ ESTs of a more recent release (April 1996, 27155 3′ sequences), however, as well as in the 'less reliable' 3′ sequences of the earlier release, this was no longer the case. The difference could be due to an increased contribution of coding regions. This is suggested also by the narrowing of the length distribution to a 450 nt peak, in recent releases, a length that in many genes would exceed the length of the 3′ untranslated region. Since coding regions have higher GC levels than noncoding regions, GC histograms for mixtures of short (<500 nt) coding and noncoding sequences in unknown proportions can unfortunately not allow a reliable independent check of the GC distribution of genes, at this stage.

If one accepts current estimates for the size of the human genome (3.4×10^9 bp; cf. Cavalier-Smith, 1989) and for the number of genes (50 000–80 000; cf. Fields et al., 1994), the average gene concentrations shown in Fig. 3C will be approximately one gene per 150 kb in L; one gene per 54 kb in H1; one gene per 15 kb in H2; and one gene per 9 kb in H3. These values correspond to density estimates of 6.7, 19, 67 and 111 genes per Mb,

respectively. It should however be emphasized that these are only average values, and that the steep slope of the gene concentration curve in the GC-rich regions (Fig. 4A) implies a large variation in gene densities within H3, and within most of H2, the GC-richest regions of H3 reaching average densities as high as one gene per 4 kb.

A recently sequenced region of 220 Kb of very GC-rich (57%) genomic DNA (Chen et al., 1996), which is known to be located in an H3 isochore (De Sario et al., 1996), and which would also be expected to be in an H3 isochore according to Fig. 2B, allows the predictions of gene-finding programs such as GRAIL (Uberbacher and Mural, 1991) to be compared with the gene concentration predictions given here. Based on analyses using GRAIL programs, Chen et al. (1996) predict 19 genes in this region (13 known genes and 6 additional genes), i.e. a gene density of one gene per 11 or 12 $\bar{K}$b. This value is only slightly lower than the average gene density suggested here for H3 (one gene per 9 $\bar{K}$b), and within the bounds allowed by the range of gene densities in H3 and by the estimate of 50000 to 80000 genes in the human genome.

3.4. Gene coverage by recent physical maps of the human genome

The data of Fig. 4 can be used to estimate gene coverage by the mapping (and sequencing) of chromosomal regions in a given range of GC levels. The GC-richest and gene-richest chromosomal bands correspond to 28 T bands and 31 T′ bands, the former showing strong, the latter medium or weak hybridization signals with H3 DNA (Saccone et al., 1992,1993,1996) as well as with CpG island probes (Saccone et al., 1996), as judged from the results of Craig and Bickmore (1994). Practically all the T bands and the majority of the T′ bands are not covered by the most recent YAC contig map of the human genome (Chumakov et al., 1995), a situation explained in part by the instability of YACs derived from these bands (De Sario et al., 1996). These GC-richest regions correspond to about 20% of the human genome (see Saccone et al., 1996), namely to the majority of the gaps, which are estimated to represent 25% of the map (Chumakov et al., 1995). As shown in Fig. 4B, under these circumstances only about 45% of the genes are covered by the latest YAC contig map of the human genome. It is obvious that future mapping and sequencing efforts should be directed, as already stressed before (Gardiner et al., 1990; Saccone et al., 1992; Bernardi, 1989, 1994), towards these genome regions which, although the most important ones in terms of gene densities, still are largely *terra incognita.*

Acknowledgement

We thank Gabriel Macaya for valuable discussions and expert advice, Dominique Mouchiroud for the non-redundant database of human genic GC_3 levels, and Christian Gautier for discussions on regression analysis, and David Schlessinger for valuable comments. Serguei Zoubak acknowledges the award of a long-term fellowship from the Federation of European Biochemical Societies.

References

Adams, M.D. et al. (1995) Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. Nature (London) 377(suppl.), 3–174.

Adams, M.D. et al. (1991) Complementary DNA sequencing: Expressed sequence tags and human genome project. Science 252, 1651–1656.

Aïssani, B. and Bernardi, G. (1991a) CpG islands: features and distribution in the genome of vertebrates. Gene 106, 173–183.

Aïssani, B. and Bernardi, G. (1991b) CpG islands, genes, isochores in the genome of vertebrates. Gene 106, 185–195.

Aïssani, B., D'Onofrio, G., Mouchiroud, D., Gardiner, K., Gautier, C. and Bernardi, G. (1991) The compositional properties of human genes. J. Mol. Evol. 32, 497–503.

Benson, D.A., Boguski, M., Lipman, D.J. and Ostell, J. (1994) GenBank. Nucleic Acids Res. 22, 3441–3444.

Bernardi, G. (1989) The isochore organization of the human genome. Annu. Rev. Genet. 23, 637–661.

Bernardi, G. (1993) The isochore organization of the human genome and its evolutionary history – a review. Gene 135, 57–66.

Bernardi, G. (1994) Questions about genomes and genome projects. Biotechnology 12, 840.

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

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

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

Boguski, M.S. and Schuler, G.D. (1995) ESTablishing a human transcript map. Nature Genetics 10, 369–371.

Cavalier-Smith, T. (1985) Eukaryote gene numbers, non-coding DNA and genome size. In: Cavalier-Smith, T. (Ed.), The Evolution of Genome Size, Wiley, London, pp. 69–103.

Chen, E.Y., Zollo, M., Mazzarella, R., Ciccodicola, A., Chen, C., Zuo, L., Heiner, C., Burough, F., Repetto, M., Schlessinger, D. and D'Urso, M. (1996) Long-range sequence analysis in Xq28: Thirteen known and six candidate genes in 219.4 kb of high GC DNA between the RCP/GCP and G6PD loci. Hum. Mol. Genet. 5, 659–668.

Chumakov, I.M. and 60 other authors (1995) A YAC contig map of the human genome. Nature (London) 377(suppl.), 175–297.

Clay, O., Cacciò, S., Zoubak, S., Mouchiroud, D. and Bernardi, G. (1996) Human Coding and Noncoding DNA: Compositional Correlations. Mol. Phylogenet. Evol. 5, 2–12.

Craig, J.M. and Bickmore, W.A. (1994) The distribution of CpG islands in mammalian chromosomes. Nature Genet. 7, 376–382.

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

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

Duret, L., Mouchiroud, D. and Gautier, C. (1995) Statistical analysis

of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. J. Mol. Evol. 40, 308–317.

Fields, C., Adams, M.D., White, O. and Venter, J.C. (1994) How many genes in the human genome? Nature Genet. 7, 345–346.

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

Gonzalez, I.L. and Sylvester, J.E. (1995) Complete sequence of the 43-kb human ribosomal DNA repeat: Analysis of the intergenic spacer. Genomics 27, 320–328.

Jolicoeur, P. (1990) Bivariate allometry: Interval estimation of the slopes of the ordinary and standardized normal major axes and structural relationship. J. Theor. Biol. 144, 273–285.

Kawamoto, S., Matsumoto, Y., Mizuno, K., Okubo, K. and Matsubara, K. (1996) Expression profiles of active genes in human and mouse livers. Gene, in press.

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

Macaya, G., Thiery, J.P. and 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., Aïssani, B., Macaya, G., Gautier, C. and Bernardi, G. (1991) The distribution of genes in the human genome. Gene 100, 181–187.

Okubo, K., Hori, N., Matoba, R., Niiyama, T., Fukushima, A., Kojima, Y. and Matsubara, K. (1992) Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. Nature Genetics 2, 173–179.

Pilia, G., Little, R.D., Aïssani, B., Bernardi, G. and Schlessinger, D. (1993) Isochores and CpG islands in YAC contigs in human Xq26.1-qter. Genomics 17, 456–62.

Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. (1988) Numerical Recipes in C. Cambridge University Press, New York.

Roe, B.P. (1992) Probability and Statistics in Experimental Physics. Springer, New York.

Saccone, S., De Sario, A., Della Valle, G. and 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, C., De Sario, A., Wiegant, J., Rap, A.K., Della Valle, G. and 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. and Bernardi, G. (1996) Identification of the gene-richest bands in human chromosomes. Gene 174, 85–94.

Schildkraut, C.L., Marmur, J. and Doty, P. (1962) Determination of the base composition of deoxyribonucleic acid from its buoyant density in CsCl. J. Mol. Biol. 4, 430–443.

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

Uberbacher, E.C. and Mural, R.J. (1991) Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. Proc. Natl. Sci. Acad. USA 88, 11261–11265.

Xu, H., Wei, H., Tassone, F., Graw, S., Gardiner, K. and Weissman, S.M. (1995) A search for genes from the dark band regions of human chromosome 21. Genomics 27, 1–8.

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