Gene, 100 (1991) 181–187
© 1991 Elsevier Science Publishers B.V. 0378-1119/91/$03.50

GENE 04053

The distribution of genes in the human genome

(Gene composition; isochores; codon positions; introns; coding sequences)

Dominique Mouchiroud[b], Giuseppe D'Onofrio[a]*, Brahim Aïssani[a], Gabriel Macaya[a], Christian Gautier[b] and Giorgio Bernardi[a]**

[a] Laboratoire de Génétique Moléculaire, Institut Jacques Monod, 75005 Paris (France); and [b] Laboratoire de Biométrie, Génétique et Biologie des Populations, U.R.A 243, Université Claude Bernard Lyon I, 69600 Villeurbanne (France)

Received by G. Roizès: 29 January 1991
Accepted: 8 February 1991

SUMMARY

Previous investigations on the human genome determined: (*i*) the base compositions (GC levels) and the relative amounts of its isochore families; (*ii*) the compositional correlations (i.e., the correlations between GC levels) between third codon positions of a set of genes and the DNA fractions in which the genes were localized; and (*iii*) the compositional correlations between (a) third and first + second codon positions, as well as that between (b) introns and exons from the set of 'localized genes' and from all the coding sequences and genes (genomic sequences of exons + introns) available in gene banks. Here, we have shown that the correlations (*iii*, a and b) for 'localized genes' and genes from the bank are in full agreement, indicating that the former set is representative of the latter. We haven then used the data (*i*) and the correlation (*ii*) to estimate the distribution of genes in isochore families. We have found that 34% of the genes are located in the GC-poor isochores (which represent 62% of the genome), 38% in the GC-rich isochores (31% of the genome) and 28% in the GC-richest isochores (3% of the genome). There is, therefore, a compositional gradient of gene concentration in the human genome. The gene density in the GC-richest 3% of the genome is about eight times higher than in the GC-rich 31%, and about 16 times higher than in the GC-poorest 62%.

INTRODUCTION

The human genome, like all vertebrate genomes, is made up of isochores. These are long (over 300 kb, on the average) DNA segments that are homogeneous in base composition (or GC levels). Isochores belong to a small number of families characterized by different GC levels. Compositional correlations (i.e., the correlations between GC levels) exist between coding sequences (or their third codon positions) and the isochores (in fact, the large DNA fragments) in which the corresponding genes are located (Bernardi et al., 1985; see also Bernardi, 1989; Aïssani et al., 1991; D'Onofrio et al., 1991). In principle, this allows inferring the location of genes in different isochore families on the basis of the GC levels (a) of coding sequences (or their third codon positions) as available in gene banks, and (b) of isochore families (in fact, of the large DNA fragments characterized by similar GC levels). This approach was originally used to demonstrate the existence of compositional classes in human coding sequences, that appeared to correspond to isochore families, and of differences in the compositional distributions of coding sequences between

Correspondence to: Dr. G. Bernardi, Laboratoire de Génétique Moléculaire, Institut Jacques Monod, 2, Place Jussieu, Paris 75251 Cedex 05 (France) Tel. 33-1-43 29 58 24; Fax 33-1-46 33 23 05.

* On leave from Stazione Zoologica, Villa Comunale, 80121 Naples (Italy)

** On leave from Universidad de Costa Rica, Ciudad Universitaria Rodrigo Facio, San José (Costa Rica)

Abbreviations: GC, % of guanine + cytosine; kb, kilobase(s).

man and cold-blooded vertebrates (Bernardi et al., 1985). Subsequent investigations (Mouchiroud et al., 1987; 1988; Bernardi et al., 1988) provided histograms of GC levels of third-codon positions of increasing numbers of human genes, and made use of the original localization (Bernardi et al., 1985) of nine human loci (defined as isolated genes or gene clusters) in compositional DNA fractions to estimate the isochore location of the genes under consideration. This was done under the tacit assumption that the nine loci were representative of human loci in general. Although nine loci admittedly are a very small gene sample, data on 15 loci from the genomes of other warm-blooded vertebrates appeared to obey the same rules of correlation, so providing some additional support for the proposed representativity.

Very recent work increased the number of human loci localized in compositional fractions to 21 and showed that the compositional correlations defined with this set of genes were indistinguishable from those established for a second set of 32 loci (four of which were common with the first set) localized in long (over 10 kb) sequences available in gene banks (Aïssani et al., 1991). This provided us with a reference set of 49 (21 + 32 − 4) human genes localized in compositional DNA fractions, or in known extended sequences. On the other hand, the number of human sequences which were recently studied (D'Onofrio et al., 1991) included about 1400 coding sequences and 238 genes (genomic sequences of introns + exons).

We decided, therefore, to re-assess the isochore location of human genes on the basis of the data presently available, and to analyze in detail all the factors involved. A precise assessment is of importance for at least three major reasons; (*i*) no detailed study of the gene distribution in the human genome is available so far, although it is known that such distribution is highly non-uniform, the highest concentration being found in the GC-richest isochores (Bernardi et al., 1985); (*ii*) the assessment under consideration should shed light on the distribution of genes in chromosomal bands, because the distribution of human genes in isochores families of increasing GC levels is paralleled by the distribution of genes in Giemsa$^+$, Giemsa$^-$ and telomeric (T) chromosomal bands (Bernardi et al., 1985; Bernardi, 1989; Gardiner et al., 1990); (*iii*) the results obtained for the human genome have a broad relevance, because the gene distribution found in the human genome is very similar to that present in the genomes of other mammals (Bernardi et al., 1988).

Here, we tried to obtain the best possible assessment of the distribution of genes in the isochores of the human genome. The location of human genes in chromosomal bands, particularly in T-bands (Dutrillaux, 1973) will be discussed elsewhere.

MATERIALS AND METHODS

Most of the data used here are derived from recent investigations by Aïssani et al. (1991) and D'Onofrio et al. (1991). More specifically, the data of Figs. 1, 2A and 3A are derived from Figs. 3B, 5 and 4A, respectively, of Aïssani et al. (1991), whereas those of Figs. 2B and 3B are derived from Figs. 1C and 5A, respectively, of D'Onofrio et al. (1991).

The data from Aïssani et al. (1991) concern: (*i*) the correlation between GC levels of third codon positions and GC levels either of compositional fractions from the human genome in which the corresponding genes were localized, or of long (over 10 kb) sequenced continuous segments from the human genome, as available in gene banks (Fig. 1); numerical data and further information can be found in Table 1 of Aïssani et al. (1991); (*ii*) the correlation of GC levels of third codon positions and of first + second codon positions for the genes just referred to (Fig. 2A); and (*iii*) the correlation between the GC levels of introns and of the corresponding exons (Fig. 3A), again for the genes referred to.

The data from D'Onofrio et al. (1991) concern: (*i*) the correlation between GC levels of third-codon positions and GC levels of first + second codon positions for about 1400

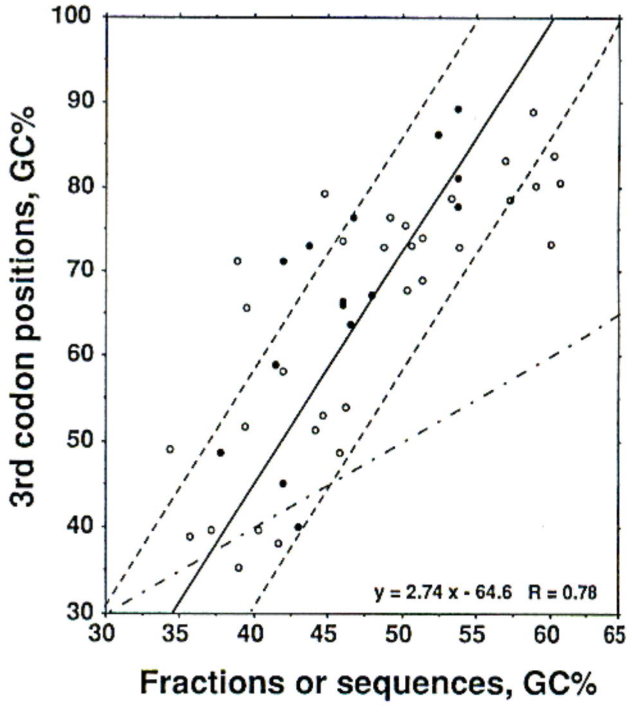

Fig. 1. GC levels of third codon positions from human genes are plotted against the GC levels of DNA fractions (blackened circles) or extended sequences (open circles) in which the genes are located. The correlation coefficient and the slope are indicated. The slope was calculated using an orthogonal regression (see MATERIALS AND METHODS). The dash-and-point line is the diagonal line (slope = 1). The dashed lines indicate the ± 5% GC limits of the slope. (Modified from Fig. 3B of Aïssani et al., 1991; see that reference for further details).

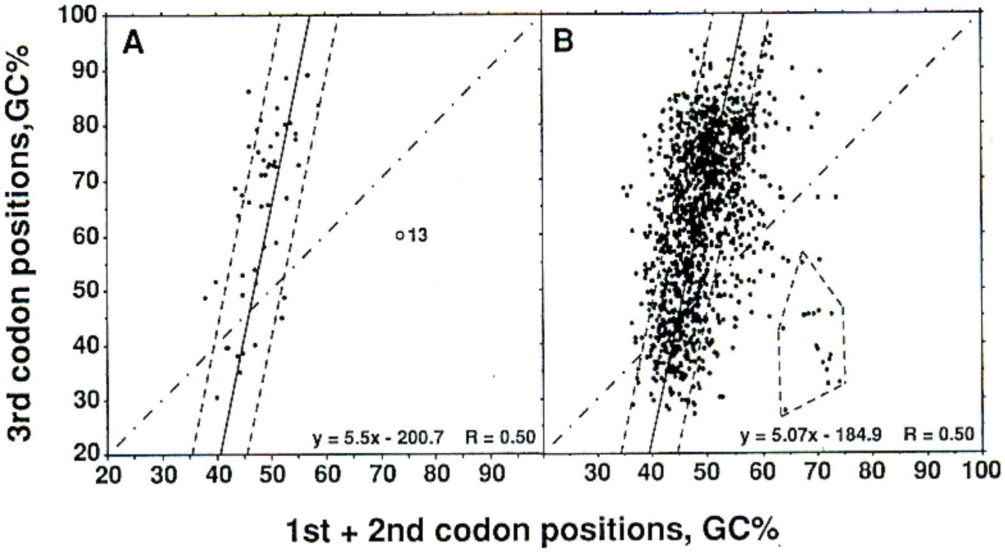

Fig. 2. GC levels of third vs. first + second codon positions of (A) human genes localized in compositional DNA fractions and extended sequences (modified from Fig. 5 of Aïssani et al., 1991) are compared with a similar plot (B) concerning approx. 1600 human coding sequences available in gene banks (modified from Fig. 1C of D'Onofrio et al., 1991). Orthogonal slopes are shown in both cases, as well as (dashed lines) the ±5% GC limits of the slopes. Point 13 (α VI collagen genes) of Fig. 2A and the deviating points surrounded by a dashed line in Fig. 2B are shown, but were not used in calculating the correlation coefficients and the slope, because of the strongly biased amino acid compositions of the corresponding proteins.

coding sequences as available in gene banks; these were expanded here to the 1610 sequences as available from Release 65 of GenBank (Bilofsky et al., 1990) in October 1990 (Fig. 2B); and (*ii*) the correlation between GC levels of introns and exons from a set of 238 sequenced genes available in gene banks (Fig. 3B).

Two regression procedures were used. The orthogonal regression (or Principal Component Analysis, in the case of two variables) was used when the purpose was to find the best representation of scatter plots (Figs. 1 and 2). This approach (see D'Onofrio et al., 1991) minimizes the sum of square distances between points and regression lines. The other procedure, linear regression, was used in Fig. 3, and also in estimating the GC levels of DNA isochores corresponding to the GC levels of genes in third codon positions (Fig. 4).

Finally, the data of Fig. 4 concern the GC levels of third-codon positions from the 1610 coding sequences just mentioned.

RESULTS

(a) Correlations and assumptions

We have previously tackled the problem of assessing the distribution of genes in the isochores that make up the

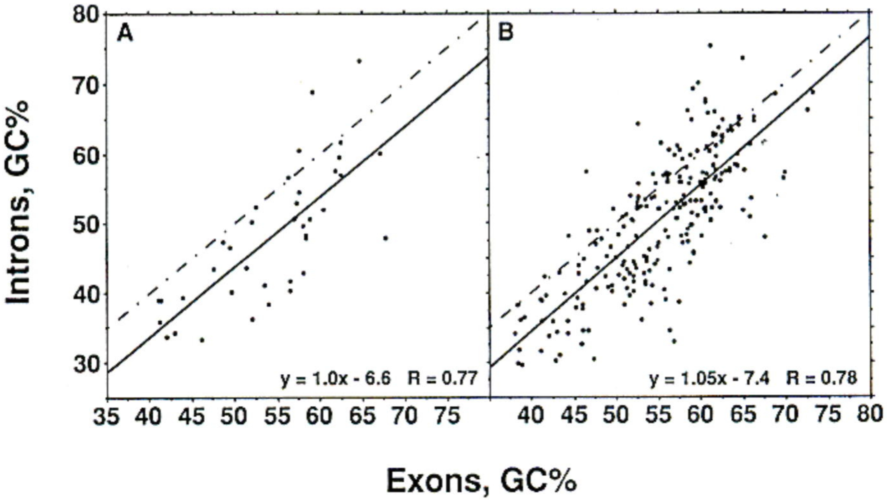

Fig. 3. GC levels of introns vs. exons of (A) human genes localized in DNA fractions and extended sequences (from Fig. 4A of Aïssani et al., 1991) are compared with a similar plot (B) for the 238 human genes available for this purpose (from Fig. 5A of D'Onofrio et al., 1991).

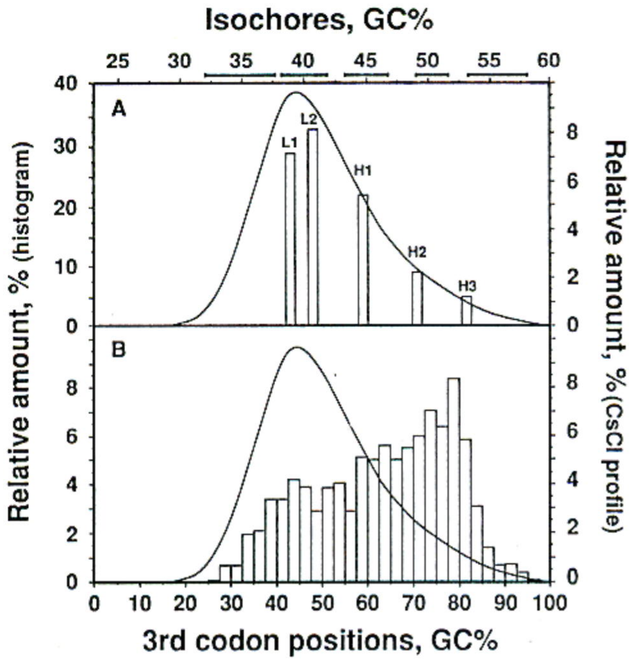

Fig. 4. Histograms of isochore families and human genes. (A) Histogram of relative amounts of isochore families (or 'major DNA components' L1, L2, H1, H2, H3), as obtained from the data of Cuny et al. (1981) and Zerial et al. (1986). (B) Histogram of relative amounts of human genes divided in classes according to GC levels of third codon positions. Bars correspond to 2.5% GC intervals. The upper scale concerns the GC levels of the isochores, and was obtained from the lower scale through the correlation of Fig. 1. Confidence intervals at 95% probability are also shown by horizontal bars under the upper scale. CsCl profiles of human DNA are superimposed on both histograms.

human genome (Bernardi et al., 1985; 1988; Bernardi, 1989; Mouchiroud et al., 1987; 1988) by (*i*) establishing the compositional correlations that exist between GC levels of coding sequences, third codon positions and introns on the one hand, and the GC levels of the DNA fragments in which the genes had been localized, on the other hand; (*ii*) assuming that the very small set of localized genes was representative of all human genes, and that all human genes obeyed the same rule of compositional correlation with isochores; (*iii*) using the compositional distribution of third codon positions of all human genes available in data banks, the compositional correlations mentioned above, and previous estimates of GC levels and relative amounts of isochore families to reconstruct the distribution of human genes in the isochores of the human genome and evaluate gene concentrations in them. Here we have critically re-examined the steps just outlined and used the most recent data so as to obtain the best possible assessment.

(1) The compositional correlations

The compositional correlation of particular interest here concerns third codon positions and isochores. Indeed, third codon positions exhibit a much wider GC range than exons, introns, or the other two codon positions (which might also be used for the same purpose), as well as a better fit with isochore composition (as judged from the correlation coefficient) compared with other codon positions and introns. The correlation coefficient is practically identical with that obtained with exons (Aïssani et al., 1991). In principle, the correlation concerning genes localized in compositional fractions from the human genome should be used, because this is the correlation that links GC levels of third codon positions and of intergenic sequences, which make up the vast majority of human DNA. It has been shown, however, that this correlation is indistinguishable from the correlation that links GC levels of third codon positions and of introns, as obtained for genes present in extended (over 10 kb), continuously sequenced segments of human DNA (Aïssani et al., 1991). Under these circumstances, we considered that the correlation cumulating the two sets of data just mentioned (as presented in Fig. 3B of Aïssani et al., 1991) was to be preferred because it was based on a larger number of loci (49 instead of 21). Moreover, it has already been argued elsewhere (Aïssani et al., 1991) that the correlation concerning only genes localized in compositional fractions might have a slightly lower slope than that reported because of the underestimation of GC levels for the GC-richest isochores. This can be taken care of, although in an indirect way, by using the cumulative data from genes localized in compositional fractions and in long, sequenced DNA segments, because the latter show a lower slope.

Using an orthogonal regression, a slope of 2.74 was found (Fig. 1) when plotting the GC levels of third codon positions of the 49 'localized genes' (as we will call this reference set, alluding to their localization in compositional fractions, or in extended, continuous sequences) against the GC levels of the large DNA fragments, or of the extended, sequenced DNA segments in which the genes under consideration were localized. Only ten points were beyond a ± 5% GC distance from the line through the points, five of them being very close to the limits. All these points, with only one exception, were in the 65–85% GC range of third-codon positions. It should be noted that, if a linear regression is used in a plot of GC levels of fractions or sequences against GC levels of third codon positions, an identical slope (2.77) was found, the correlation coefficient being 0.78. In this case, confidence intervals at 95% probability could be estimated.

(2) The representativity of the localized genes

Instead of being tacitly assumed as was previously done (Bernardi et al., 1985; 1988; Bernardi, 1989; Mouchiroud et al., 1987; 1988), the representativity of the 'localized genes' was checked in two different ways.

(*i*) The compositional correlation between third and first + second codon positions obtained with 'localized genes' was found to show a slope only 11% higher (5.5 instead of 5.1) than that exhibited by the 1600 genes or so from gene banks (Fig. 2). This difference is small and not statistically significant. It should be stressed that the similarity of the correlations found is far from being an obvious result, because the scatter of points in the diagram concerning the 1600 genes is such that it would be quite possible to find a set of 49 loci showing a completely different correlation. Such would be the case, for instance, if the localized genes belonged to particular areas of the scatter diagram.

(*ii*) Compositional correlations between exons and introns from the 'localized genes' and from the ensemble of exon-intron units present in the bank were identical (Fig. 3). In the first case, the correlation coefficient was 0.77 and the slope was 1.0, in the second, values were 0.78 and 1.05, respectively. An identical result was obtained by comparing orthogonal slopes (not shown).

In conclusion, the set of 'localized genes' showed the same properties, in terms of the two correlations used as criteria, as the large sets of 1600 coding sequences and 238 genes from the gene banks.

(b) Reconstruction of the isochore distribution of human genes

Fig. 4A displays a histogram showing the GC levels and relative amounts of the isochore families (i.e., of the 'major DNA components' L1, L2, H1, H2, H3) of the human genome (Thiery et al., 1976; Macaya et al., 1976; Cuny et al., 1981; Zerial et al., 1986). The data used to construct the histogram were those of Cuny et al. (1981). These authors left aside 7% of DNA of 'extreme pools' (representing the GC-poorest and GC-richest DNA fragments), as formed by satellite and minor components. The subsequent work of Zerial et al. (1986) showed, however, that approx. 3% out of that 7% corresponded to a very GC-rich fraction containing abundant, protein-coding genes which was called H3; this fraction is, therefore, represented in the histogram.

Fig. 4B displays a histogram of GC levels of third-codon positions for all the 1610 human genes presently known in their coding sequences. The histogram is characterized by a very wide range, 25–97.5%, almost as broad as that from all bacterial genes, and by a multimodality.

The representativity of the 'localized genes' (as far as the correlations presented in the preceding section are concerned) suggests that the correlation of Fig. 1 is generally valid for human genes. Under such circumstances, the distribution of genes in isochores characterized by different GC levels can be calculated, using the correlation of Fig. 1, from the GC levels of the third codon positions of genes. This allowed us to transform the third codon position GC scale of Fig. 4 into an isochore GC scale and to estimate the confidence limits at 95% probability.

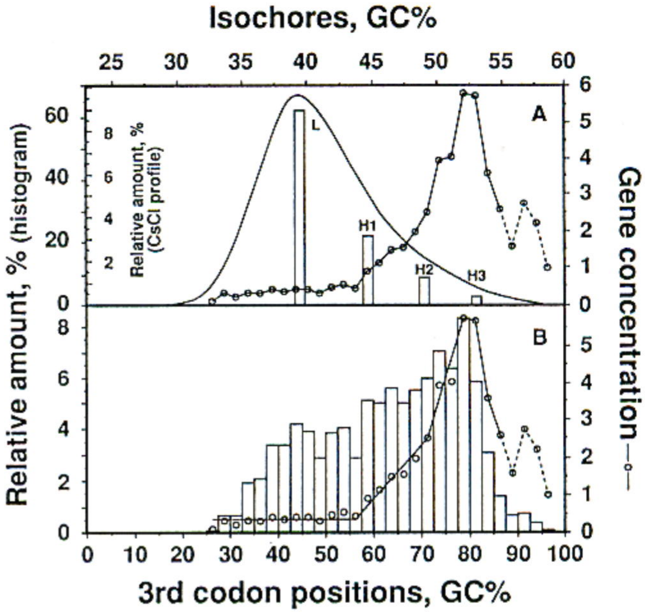

Fig. 5. Profiles of gene concentration in the human genome (obtained by dividing the relative amounts of genes in each 2.5% GC interval of the histogram of Fig. 4B by the corresponding relative amounts of DNA as deduced from the CsCl profile) are superimposed on the histograms of isochore families (A) in which major components L1 and L2 have been pooled together and of third codon positions (B). The dashed line concerns a region in which gene concentration is very uncertain (see DISCUSSION, section c). The CsCl profile of human DNA is also shown in A. The gene concentration profile of part B is represented by a series of straight lines, to show slope discontinuities.

CsCl profiles of human DNA were superimposed on the histograms of Fig. 4,A and B. In the first case, this showed the fit between the CsCl profile and the histogram of isochore families. In the second, it was used to calculate the gene concentration across the DNA peak by dividing the percentage of genes corresponding to each bar of the histogram of Fig. 4 by the corresponding percentage of DNA, as deduced from the CsCl profile.

The resulting gene-concentration profile is superimposed on the histograms of isochore families and third codon positions, respectively, in Fig. 5,A and B. It should be noted that in Fig. 5A, major components L1 and L2 have been pooled together, and in Fig. 5B, the concentration profile has been traced as a series of straight lines, to show slope change more clearly.

DISCUSSION

(a) The compositional correlation

The linear correlation of Fig. 1 is very important in that it was used to determine the isochore GC scale of Fig. 4, and, as a consequence, the positioning of the histogram of isochore families and of the CsCl profile relative to the histogram of GC levels of third codon positions. A change in its slope would lead, therefore, to a change in the isochore assignment of genes. The estimation of confidence intervals has shown that maximum shifts of the scale are less than ± 2.5% GC in the 40–55% GC region (Fig. 4A). Such shifts would not seriously alter any of the conclusions drawn in the present paper.

Another point to be considered concerns the fact that there is, indeed, just one single correlation for all human genes, independent of their GC levels, as tacitly assumed in previous work (Bernardi et al., 1985). In other words, the points fit a single line and not, for instance, two lines with different slopes.

It should be noted, however, that GC levels of third-codon positions (as well as of exons and introns) are identical to those of intergenic sequences (represented by the diagonal line) for GC-poor genes, whereas they are increasingly higher for GC-rich genes. A similar, but weaker trend is shown by the compositional correlations of exons (Fig. 1 of Aïssani et al., 1991). If the latter is confirmed, it would mean that GC-poor coding sequences match compositionally the surrounding intergenic sequences, whereas GC-rich coding sequences stand out as GC-richer islands. This point is not apparent if a linear regression is used (compare Fig. 1 with Fig. 3B of Aïssani et al., 1991), since the latter relationship indicates, for all genes, a constant, higher GC level relative to flanking intergenic sequences. Moreover, a horizontal scatter of points is only noticed for GC-rich coding sequences. Such scatter, which is only partially correlated with those shown in Figs. 2A and 3A, is important when assessing gene concentrations in isochores (see section c, below).

(b) The representativity of the localized genes

The correlation coefficient and slopes of Figs. 2 and 3 indicate a good representativity for the set of 'localized genes'. In other words, the sample of localized genes does not comprise points which show serious deviations in the scatter diagram of third vs. first + second codon positions concerning the ensemble of human genes, nor in the intron vs. exon plots.

(c) Reconstruction of the isochore distribution of human genes

The histogram of Fig. 5B deserves two main comments. (*i*) Three regions can be distinguished in the histogram using the profile of gene concentration and the histogram of the 'major DNA components' (as defined by Thiery et al., 1976; Macaya et al., 1976; Cuny et al., 1981; Bernardi et al., 1985; Zerial et al., 1986).

The first region is characterized by a constant, low gene concentration and corresponds to DNA components L1 and L2, two components so close in modal buoyant densities that they were originally considered to be 'subcomponents' (Thiery et al., 1976) and should now be pooled in an L component (Fig. 5A). This region ends at about 57% GC in third codon positions, a value 10% lower than that previously estimated on the basis of far more limited data (Mouchiroud et al., 1987; 1988; Bernardi et al., 1988) likewise, what was considered to be L1 by Mouchiroud and Gautier (1990) definitely is L1 + L2.

The second region is characterized by an increasing gene concentration and the third region by the highest gene concentration. The border between the second and the third regions is rather uncertain. If account is taken of the discontinuity in the slope of gene concentration (Fig. 5B), the border should be set at about 72% GC in third-codon position. If the modal GC levels of isochore families H2 and H3 are considered, the border would be set at about 77% GC in third-codon position (a value identical with previous estimates). At present, an intermediate limit, at 75% GC, seems to be the best choice. As far as the upper end of the gene concentration profile is concerned, this should be considered as highly imprecise, owing to two main problems: (a) it is difficult to define the terminal part of the CsCl profile, a slight difference in the baseline causing quite a change in the estimate of DNA; (b) the end of DNA distribution on the high GC side is characterized by the presence of ribosomal DNA, which should be subtracted from the profile.

The regions defined above match some discontinuities in the gene histogram. Some fine readjustments may be needed in the future, but the assignments proposed here are unlikely to show important changes.

(*ii*) Fig. 5 shows that there is a compositional gradient of gene concentration in the genome. If the borders between the GC-poor isochores (major components L1 and L2, now pooled in a single major component L) and the GC-rich isochores (major components H1 and H2) and between the latter and the GC-richest isochores (major component H3) are set at 57.5 and 75% GC in third-codon position, the first region comprises 34% of genes, the second 38% and the third 21%. Now, the first region corresponds to 62% of DNA, the second to 31% and the third to 3%, as judged from the relative amounts of the corresponding major components. This leads to gene/DNA % ratios of 0.55, 1.23 and 9. In other words, the GC-richest region has a gene concentration eight times higher than the GC-rich region and 16 times higher than the GC-poor region.

Two comments on the gene concentration estimates are the following: (a) housekeeping genes are currently underrepresented in gene banks. If these genes tend to be GC-rich, as suggested, they would cause a further increase in the already very large concentration of genes present in the GC-richest isochores; (b) the horizontal scatter of points shown in Fig. 1 indicates that a given GC level in third codon positions may correspond to a range of GC levels in the isochores containing the corresponding genes. Since points are symmetrically distributed about the orthogonal regression line, genes which should be moved to the left or to the right on the isochore GC scale, are compensated by genes which should be moved in the opposite direction in neighboring sections of the distribution. Because of this compensation, the gene distribution shown in Fig. 4B should reflect fairly well the real gene distribution in isochores. At a finer level of analysis, one may argue, however, that the broader horizontal scatter of third-codon positions in the 65–85% GC range leads to a situation in which a number of genes belonging to the GC-rich class are in fact present in the GC-poor isochores and in the GC-richest isochores. This phenomenon is not compensated by GC-poor and GC-richest genes which are located in GC-rich isochores, as it can be judged from the points of Fig. 1. Therefore, the number of genes present in GC-rich isochores (major components H1 + H2) tends to be overestimated by the histogram of third codon positions.

In conclusion, the present work has led to an assessment of gene distribution in the isochores of human DNAs. This assessment may certainly be improved in its details by further work, but it is highly unlikely that it will be considerably changed in the future.

ACKNOWLEDGEMENTS

We thank the Association Française contre les Myopathies (AFM) for a Fellowship to B.A., FEBS and the Di Capua Foundation, Naples, Italy, for fellowships to G.D'O., the European Economic Community (EEC) for a Fellowship to G.M., the Association pour la Recherche contre le Cancer (ARC), the Ministry for Research and Technology (MRT) and AFM for financial support.

REFERENCES

Aïssani, B., D'Onofrio, G., Mouchiroud, D., Gardiner K., Gautier, C. and Bernardi, G.: The compositional properties of human genes. J. Mol. Evol. (1991) in press.

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

Bernardi, G. and Bernardi, G.: Compositional properties of nuclear genes from cold-blooded vertebrates. J. Mol. Evol. (1991) in press.

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

Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M. and Rodier, F.: The mosaic genome of warm-blooded vertebrates. Science 228 (1985) 953–968.

Bilofsky, H.S., Burks, C., Fickett, J.W., Goad, W.B., Lewitter, F.L., Rindone, W.P., Swindell, C.D. and Tung, C.S.: The GenBank genetic sequence data bank. Nucleic Acids Res. 14 (1986) 1–4.

Cuny, G., Soriano, P., Macaya, G. and Bernardi, G.: The major components of the mouse and human genomes, 1. Preparation, basic properties and compostional heterogeneities. Eur. J. Biochem. 115 (1981) 227–233.

D'Onofrio, G., Mouchiroud, D., Aïssani, B., Gautier, C. and Bernardi, G.: Correlations between the compositional properties of human genes, codon usage and aminoacid composition of proteins. J. Mol. Evol. (1991) in press.

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

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

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

Mouchiroud, D. and Gautier, C.: Codon usage changes and sequence dissimilarity between human and rat. J. Mol. Evol. 31 (1990) 81–91.

Mouchiroud, D., Fichant, G. and Bernardi, G.: Compositional compartmentalization and gene composition in the genome of vertebrates. J. Mol. Evol. 26 (1987) 198–204.

Mouchiroud, D., Gautier, C. and Bernardi, G.: The compositional distribution of coding sequences and DNA molecules in humans and murids. J. Mol. Evol. 27 (1988) 311–320.

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

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