

Expression patterns and gene distribution in the human genome

Giuseppe D'Onofrio*

Laboratorio di Evoluzione Molecolare, Stazione Zoologica Anton Dohrn, Naples 80121, Italy

Received 15 March 2002; received in revised form 29 April 2002; accepted 23 September 2002

Received by T. Gojobori

Abstract

Genes are non-uniformly distributed in the human genome, reaching the highest concentration in GC-rich isochores. This is one of the fundamental aspects of the human genome organization (see Bernardi, 2000. *Gene* 241, 3 for a review). In the present paper the gene distribution was analyzed in relationship to the gene expression pattern and levels. In this study evidence is produced showing that a biased gene distribution towards GC-rich isochores applies to both tissue-specific and housekeeping genes. Moreover, the analyses of recently published data on the distribution of the gene expression levels along the chromosomes, and their base composition, further support that highly transcribed genes are localized in GC-rich isochores. Since gene density and transcriptional levels are correlated with each other and both are correlated with the GC level of the isochores, the biased gene distribution in the human genome presumably is the result of selection at the gene expression levels. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: Gene expression levels; Human genome; Isochores

1. Introduction

The localization of 40 human genes in isochore families first showed that genes were not uniformly distributed in the human genome, being more concentrated in GC-rich isochores (Bernardi *et al.*, 1985). Thereafter, *in silico* localization of ~ 1400 human genes (D'Onofrio *et al.*, 1991) led to the same conclusions (Mouchiroud *et al.*, 1991), further confirmed by larger sets of genes (Zoubak *et al.*, 1996; Saccone *et al.*, 2001).

The biased gene distribution in the human genome raised a question about the correlation between gene distribution and gene expression pattern or, in other words, about the distribution of tissue-specific and widely expressed genes according to the GC level of the isochores.

The first attempt to answer the above question, summarizing independent experimental results on chromatin structure and gene composition, as well as gene distribution, led up to the hypothesis that: 'the H3 isochore

family presumably has the highest level of transcription because of its very high concentration of genes – especially housekeeping genes'. (Bernardi, 1993; and references therein). The high expression levels of the genes localized in H3 were further supported by *in silico* investigation on the sequence context of the AUG start codon. The results showed that genes located in GC-rich isochores required highly efficient translation (Pesole *et al.*, 1999).

Subsequent analyses on the correlation between gene distribution and gene expression levels showed that the majority of the widely expressed genes were localized mainly in GC-poor isochores, whereas tissue-specific genes were localized in the GC-rich ones (Gonçalves *et al.*, 2000). The authors drew these conclusions by analyzing the base composition and the distribution of genes with or without retroseudogenes, the former being more widely expressed than the latter. However, in the human genome, using a different algorithm, the propensity for retrotransposition was found to be unaffected by the GC content of the genes (Venter *et al.*, 2001).

Studying the origin of CpG islands, tissue-specific genes were confirmed to be mainly localized in GC-rich isochores (64% in H3), whereas widely expressed genes distribution was independent of the isochore context (Ponger *et al.*, 2001). The finding led Galtier *et al.* (2001) to argue, '...

Abbreviations: GC level, molar ratio of guanine plus cytosine; GC3, GC level at the third codon positions; CsCl, cesium chloride; L, H1, H2, H3, isochores families of the human genome; T-bands, telomeric chromosomal bands.

* Tel.: +39-81-583-3311; fax: +39-81-746-3155.

E-mail address: donofrio@sunev.szn.it (G. D'Onofrio).

selection, if any, must be unrelated to gene expression level or pattern'.

However, in both papers dealing with the correlation between gene distribution and expression patterns (Gonçalves et al., 2000; Ponger et al., 2001), the gene partition was performed according to the criteria defined in Mouchiroud et al. (1991). In the last decade, however, several results based on theoretical and experimental approaches led to an improvement of the gene partition criteria (Saccone et al., 1993, 1996, 1999; Zoubak et al., 1996; Federico et al., 2000).

In the present paper the distribution of widely expressed genes in the human genome was revisited by analyzing the dataset of human genes collected in the CpG islands database (Larsen et al., 1992; database 4.0, 1996), as well as a dataset of human genes orthologous to those of *Xenopus*, calf and murids. The results from the two independent datasets led to the conclusion that widely expressed genes: (i) are mainly localized in GC-rich isochores; (ii) are not the majority of the genes in the GC-rich isochores; and (iii) are not GC3 poorer than tissue-specific genes.

2. Materials and methods

The human CpG-islands database (Larsen et al., 1992; release 4.0, 1996; retrieved from <http://bioinformatics.weizmann.ac.il/databases/cpgisle/>), contains 1711 entries. After removing those with no information on the expression level (15%), grouping those belonging to the same gene (22%), and taking off redundancy (14%), 882 complete coding sequences (CDS) were recovered, hereafter on referred to as dataset A.

A second CDS dataset was obtained by pooling available sequences of human genes orthologous to those: (i) of *Xenopus laevis* and *Bos taurus* from Cruveiller et al. (1999); and (ii) of murids (rat and mouse) from Duret and Mouchiroud (2000). After removing overlaps, the expression pattern of 2423 CDS was estimated by retrieving from TIGR database (<http://www.tigr.org>) the number of tissues in which each gene was expressed. After discarding CDS without information on the expression pattern (5.5%), or expressed in pathological tissues only (20.3%), 1797 genes were retained and divided in two categories. Genes expressed in one to three tissues, 54.4%, were defined as tissue-specific, whereas those expressed in at least eight tissues, 17.5%, were defined as widely expressed. The criteria were stringent enough to reduce the probabilities of gene misclassification. Moreover, considering that hybridization techniques can give more false negative than false positive results, tissue-specific genes could be overestimated compared to widely expressed ones. The final dataset, hereafter referred to as dataset B, was made of 1291 CDS. Base composition was calculated using the program Codon W 1.3 (J. Peden; <http://molbiol.ox.ac.uk/Win95.codonW.zip>).

3. Results and discussion

The degree of overlap between datasets A and B was checked. The number of shared genes was 149, 28 of them were widely expressed genes. Therefore, widely expressed gene from the two datasets, accounting for the 3.2 and 2.2% of A and B dataset, respectively, can be considered as completely independent sets of genes.

In order to compare the results from the present datasets with those previously reported, the gene distribution of widely expressed genes was analyzed according to the criteria defined in Mouchiroud et al. (1991). The histograms of widely expressed genes from datasets A and B showed two different gene distributions (Fig. 1). In the top plot of Fig. 1, from dataset A, gene frequencies, 29.1, 37.8 and 33.2%, were very similar in the three GC3 ranges. On the contrary, in the bottom plot of Fig. 1, from dataset B, gene frequencies decreased as the GC3 level decreased, 42.7%, 37.6%, and 19.7%, respectively, as the GC₃ level increased. These results deserve several comments.

Ponger et al. (2001) stressed the advantages of their approach since in Larsen's dataset: (i) different methods and different tissues samples were used to infer the gene expression patterns; (ii) *galectin 1* and *E-apolipoprotein* were classified as restrictedly expressed genes, contrary to experimental data; and (iii) all housekeeping or widely expressed genes were reported to have a CpG islands at the transcription start, but this was confirmed for only 90% of them. In spite of that, the analysis of Larsen's dataset (Fig. 1, top) led to the same conclusions of Ponger and colleagues: 'the frequency of housekeeping genes is

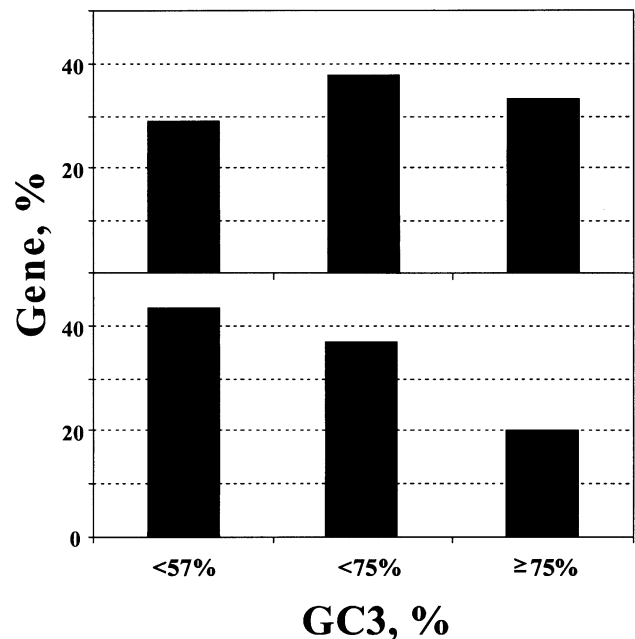


Fig. 1. Histogram of widely expressed genes from datasets A (top) and B (bottom), see also Section 2. The distribution of the genes in the three compositional classes was performed according to Mouchiroud et al. (1991).

independent of the isochore context'. Indeed, χ^2 statistical test showed no significant differences between GC-poor and GC-rich genes.

At the same time, from the histogram based on the dataset B (Fig. 1, bottom) we could reach the conclusion: 'genes with a wide tissue distribution are GC poor' (Gonçalves et al., 2000). Indeed, the χ^2 test was significant, $p < 10^{-8}$.

Therefore, using the same partitioning criteria on different sets of human genes, the same non-congruent conclusions of other authors were reached. The criteria used by Mouchiroud et al. (1991) were based on: (i) a threshold at 57% GC3 as boundary for L and H1 isochores; and (ii) grouping H1 and H2 in a middle compositional class. However, both criteria have been improved on the basis of theoretical and experimental results.

Gaussian decomposition of the CsCl profile of human DNA showed that the previous estimation of the L/H1 boundary 'was higher than the present one {47%}', and thus assigned some genes to L that by present criteria are more likely to belong the H1' (Zoubak et al., 1996). *In situ* hybridization of DNA fractions on human chromosomes, at 850 band resolution, showed that all the Giemsa bands were labeled by a DNA fraction with modal buoyant density value of 1.7003 g/cm³ in a CsCl gradient (Federico et al., 2000). The corresponding GC3 value was 45.8%, which was obtained by converting the buoyant density value into the GC level according to the Schildkraut et al. (1962) equation, and the last one into the GC3 level according to Zoubak et al. (1996).

Compositional partitions should include both H2 and H3 isochores in the GC-rich category. Indeed, *in situ* hybridization of a very GC-rich fraction, 85.4% GC3, showed that the two isochores co-localized in the chromosomal T-bands, even at high resolution banding, R850 (Saccone et al., 1999). By Gaussian decomposition, the lower limit for H2 + H3 was determined at 61% GC3 (Zoubak et al., 1996). By *in situ* hybridization the limit was found at 68.7% GC3 (Saccone et al., 1993). Indeed, T-bands were 'strongly and sharply labeled', at a 400 band resolution, by a DNA fraction with a modal buoyant density of 1.7081 g/cm³ (Saccone et al., 1993). Moreover, T-bands were barely contaminated by GC-poor isochores (Saccone et al., 1993, 1996).

From there, averaging theoretical and experimental results, the boundaries came out at: (i) 46.5% GC3 as the upper limit for the GC-poor isochores, little contaminated by the GC-rich ones; and (ii) 64.8% GC3 as the lower limit for the GC-rich isochores, little contaminated by the GC-poor ones. It is worth to note that in a study of Xenopus/human orthologous genes very close values (45% and 65% GC3, respectively) were used (Cruveiller et al., 1999). Using those boundaries values, it turned out that in the major transition, i.e. the compositional transition from cold- to warm-blooded vertebrates (see Bernardi, 2000b, for a review), Xenopus and human genes GC3 levels

<45%, and the encoded proteins, showed: (i) no differences in base compositions; (ii) similar amino acid frequencies; and (iii) comparable average hydrophobicity. On the contrary, significant differences were found for human genes having GC3 levels $\geq 65\%$ (Cruveiller et al., 1999).

Fig. 2 shows the distributions of widely expressed genes according to the above-defined boundaries. The gene frequencies, in this case, increased at increasing GC3 levels in both plots, being 14.7%, 28.6% and 59.7% from dataset A (Fig. 2, top) and 21.9%, 37.3%, and 40.8% from dataset B (Fig. 2, bottom). The gene frequencies at the extreme GC3 ranges were statistically significant by χ^2 test ($p < 10^{-11}$ and $p < 3 \times 10^{-4}$, for datasets A and B, respectively). Incidentally, the differences were significant also using 68.7% GC3 as lower limit for the GC-rich isochores, $p < 10^{-6}$ and $p < 10^{-2}$, for datasets A and B, respectively (data not show).

The finding that widely expressed genes were localized mainly in GC-rich isochores is in such a strong contradiction with the results of Gonçalves et al. (2000) that it deserves a more detailed discussion. Indeed, starting from an analysis of retroseudogenes, that are known to be produced mainly by retrotranscription of housekeeping genes, the authors found that the corresponding homologs were mainly localized in GC-poor isochores. Retroseudogenes sequences were retrieved using BLASTN2 program, disregarding those

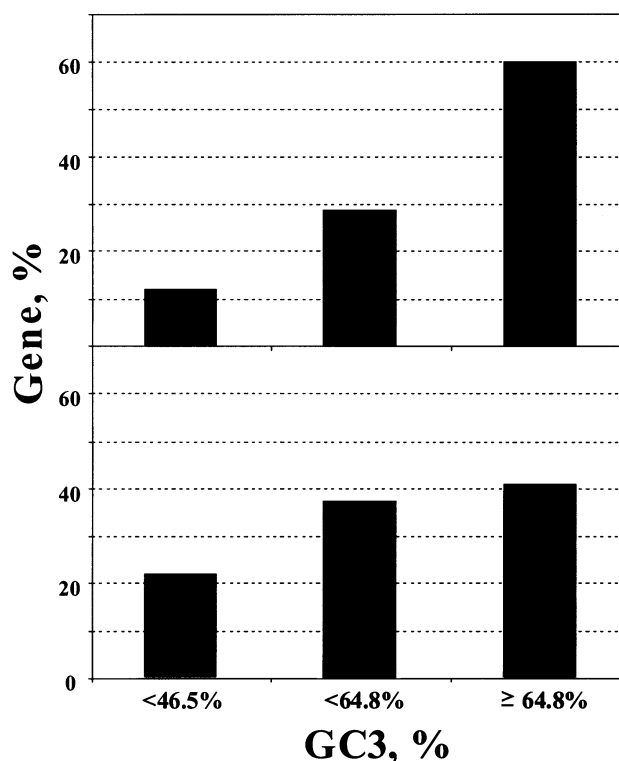


Fig. 2. Histogram of widely expressed genes from datasets A (top) and B (bottom). The distribution in the three compositional classes was performed averaging the boundaries defined by theoretical (Zoubak et al., 1996) and experimental results (Saccone et al., 1993, 1996, 1999; Federico et al., 2000).

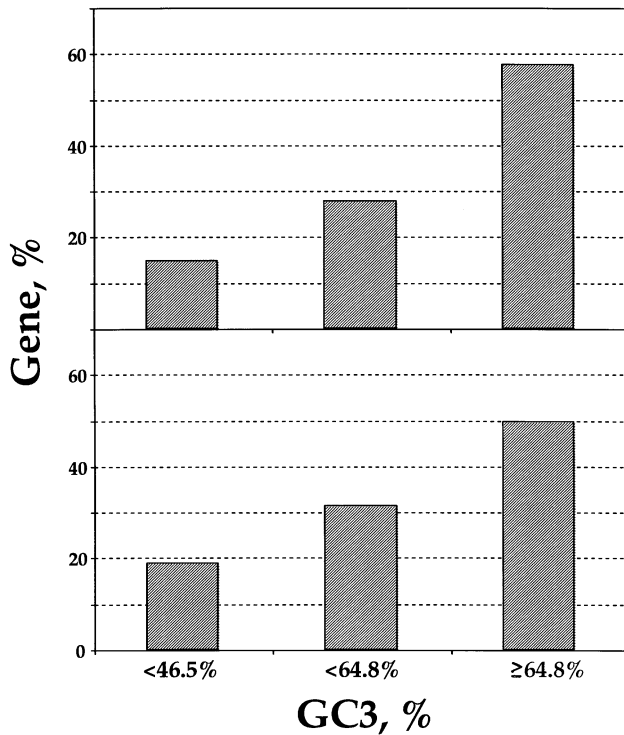


Fig. 3. Histogram of tissues-specific genes from datasets A (top) and B (bottom), see also legend Fig. 2.

sequences showing less than 80% identity over 100 nucleotides with functional genes (Gonçalves et al., 2000). However, one can question whether the final set of genes was not biased, since the described procedure was based on the implicit assumption that the base composition was not one of the factors affecting the divergence between functional genes and corresponding pseudogenes. Indeed, reports that: (i) pseudogenes with a GC content higher than the corresponding flanking regions also showed a higher substitution rate (Casane et al., 1997); and (ii) the synonymous substitution rates were positively correlated with the GC levels of the genes (Bielawsky et al., 2000), made the above assumption unjustified. Moreover, more recent results showed that silent sites of pseudogenes derived from GC-poor genes are evolving at neutral mutation rate, whereas the same sites in pseudogenes derived from GC-rich genes are evolving at a faster rate (Bustamante et al., 2002). Therefore, running algorithms in both GC-poor and GC-rich regions with fixed threshold value for similarity could lead to a biased recovery of processed pseudogenes, missing the GC-richest ones because of increasing divergence with the corresponding functional genes.

The distribution of tissue-specific genes was checked as well (Fig. 3). The results clearly support the view that the distribution of tissue-specific genes was also biased reaching the highest frequency in the GC-rich isochores. The gene frequencies were 14.7%, 27.8% and 57.4% in the dataset A (Fig. 3, top), and 18.9%, 31.4% and 49.6% in that

of the dataset B (Fig. 3, bottom), in agreement with previous reports (Gonçalves et al., 2000; Ponger et al., 2001).

Finally, tissue-specific and widely expressed genes showed slight differences at GC3 level (Table 1). The highest delta was 2.3% ($p < 4 \times 10^{-2}$), by comparing all the tissue-specific and the widely expressed genes in dataset B. No significant differences were found in dataset A, or by comparing the GC3 levels of the two kinds of genes in each GC3 range.

4. Conclusions

Re-examination of experimental and theoretical data from publications spanning almost 10 years (Saccone et al., 1993, 1996, 1999; Zoubak et al., 1996; Federico et al., 2000) allowed us to refine the gene partition criteria first used by Mouchiroud et al. (1991). Two independent datasets were analyzed: one was the updated Larsen's database (Larsen et al., 1992; database 4.0, 1996), the other a set of available human genes orthologous to genes from *Xenopus*, calf and murids. The analysis of the gene frequencies in GC-poor and GC-rich regions led us to reach the conclusions that tissue-specific and widely expressed genes:

- show no compositional differences at GC3 level;
- are in a ratio higher than one in GC-rich isochores, that is in these isochores tissue-specific genes are more abundant than widely expressed ones;
- follow the general gene distribution, reaching the highest frequency in the GC-rich isochores.

The last point deserves some comments. The gene density and the GC level of isochores were found to be correlated, and both were hypothesized to be correlated with the gene expression levels (Bernardi, 1993). The hypothesis was one of the key points supporting that the human genome organization and evolution are under selective forces (Bernardi, 2000a, for a review). Galtier et al. (2001) rejected the hypothesis arguing 'such a selective pressure

Table 1
Average GC3 levels and variance of tissue-specific (Ts) and widely expressed genes (W)

| Range | GC3, % | | | | | | | |
|-------|------------|--------|------------|--------|------------|--------|------------|--------|
| | dataset A | | | | dataset B | | | |
| | Ts | | W | | Ts | | W | |
| Mean | σ^2 | Mean | σ^2 | Mean | σ^2 | Mean | σ^2 | |
| <46.5 | 39.41 | 26.31 | 40.96 | 20.07 | 39.85 | 21.36 | 40.43 | 16.15 |
| <64.8 | 56.48 | 26.45 | 55.33 | 26.87 | 55.55 | 27.68 | 55.53 | 29.43 |
| ≥64.8 | 76.20 | 50.76 | 77.13 | 65.68 | 76.62 | 46.12 | 75.43 | 46.75 |
| All | 65.38 | 229.41 | 66.66 | 229.45 | 62.60 | 262.56 | 60.32 | 222.46 |

without apparent correlation with gene expression appeared quite speculative'. The statement was based essentially on the observation that 'the GC content of the genes does not correlate positively with their expression level or pattern (Gonçalves et al., 2000)'. In contrast, our present results support the view that the GC level of the genes is correlated with their expression patterns, indeed both tissue-specific and widely expressed genes are concentrated in the GC-rich isochores. Can we answer the second question, whether a correlation between the GC content and the expression levels of the genes also holds?

An analysis of gene transcription levels revealed 'a high order organization of the genome' (Caron et al., 2001). Indeed, regions of increased gene expression (RIDGE) were also characterized by high gene density, a correlation found for 50–60% of the RIDGES (Caron et al., 2001). Unexpectedly, it was also reported, 'about 40–50% of RIDGES are not gene dense. These RIDGES preferentially maps to telomeres' (Caron et al., 2001). Experimental results on human metaphase chromosomes showed the telomeric localization of H3, the isochore family with the highest gene concentration and GC content (Saccone et al., 1999). Unfortunately, the figures representing the RIDGES distribution (Caron et al., 2001) did not allow a precise localization of those regions along the chromosomes. Indeed, chromosomes were represented without correlation with the corresponding banding or length. Therefore, at present it is difficult to assess the exact correspondence of the telomeric region described in Caron et al. (2001), with those described in Saccone et al. (1999).

However, it was reported that chromosomes 4, 13, 18 and 21 were completely devoid of RIDGES, showing also "low gene expression and low gene density" (Caron et al., 2001). The same chromosomes turned out: (i) to have, on the average, very low GC levels: 38%, 38%, 40% and 41%, respectively (Venter et al., 2001); and (ii) to have low amount of GC-rich isochores, from a compositional profile obtained by a sliding window analysis (100Mb) (Pavlicek et al., 2002). From there, it could be argued that RIDGES distribution correlates not only with the gene density but also with the GC content of the isochores.

The original hypothesis of Bernardi (1993) that the GC rich 'isochore family presumably has the highest level of transcription because of its very high concentration of genes' was fundamentally correct. However, it should be expected that the transcriptional levels affect the gene concentration. Therefore, we can conclude that the force driving the non-uniform gene distribution in the human genome is the expression level of the genes.

Acknowledgements

Thanks are due to: Giorgio Bernardi, for critical reading and discussions, Oliver Clay, for the Larsen's database, Stephan Cruveiller, for sets of vertebrates orthologous

genes, and Luigi De Martino, for retrieving gene expression patterns from TIGR database.

References

- Bernardi, G., 1993. The vertebrate genome: isochores and evolution. *Mol. Biol. Evol.* 10, 186–204.
- 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, 31–43.
- Bernardi, G., Olofson, 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.
- Bustamante, C.D., Nielsen, R., Hartl, D.L., 2002. A maximum likelihood method for analyzing pseudogene evolution: implications for silent site evolution in humans and rodents. *Mol. Biol. Evol.* 19, 110–117.
- Caron, H., van Schaik, B., van der Mee, M., Baas, F., Riggins, G., van Sluis, P., Hermus, M.C., van Asperen, R., Boon, K., Voute, P.A., Heisterkamp, S., van Kampen, A., Versteeg, R., 2001. The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* 291, 1289–1292.
- Casane, D., Boissinot, S., Chang, B.H., Shimmin, L.C., Li, W., 1997. Mutation pattern variation among regions of the primate genome. *J. Mol. Evol.* 3, 216–226.
- Cruveiller, S., Jabbari, K., D'Onofrio, G., Bernardi, G., 1999. Different hydrophobicities of orthologous proteins from *Xenopus* and human. *Gene* 238, 15–21.
- D'Onofrio, G., Mouchiroud, D., Aissani, B., Gautier, C., Bernardi, G., 1991. Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. *J. Mol. Evol.* 32, 504–510.
- Duret, L., Mouchiroud, D., 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Mol. Biol. Evol.* 17, 68–74.
- Federico, C., Andreozzi, L., Saccone, S., Bernardi, G., 2000. Gene density in the Giemsa bands of human chromosomes. *Chromosome Res.* 8, 737–746.
- Galtier, N., Piganeau, G., Mouchiroud, D., Duret, L., 2001. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics* 159, 907–911.
- Gonçalves, I., Duret, L., Mouchiroud, D., 2000. Nature and structure of human genes that generate retropseudogenes. *Genome Res.* 10, 672–678.
- Larsen, F., Gundersen, G., Lopez, R., Prydz, H., 1992. CpG islands as gene markers in the human genome. *Genomics* 13, 1095–1107.
- 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.
- Pavlicek, A., Paces, J., Clay, O., Bernardi, G., 2002. A compact view of isochores in the draft human genome sequence. *FEBS Lett.* 511, 165–169.
- Pesole, G., Bernardi, G., Saccone, C., 1999. Isochore specificity of AUG initiator context of human genes. *FEBS Lett.* 464, 60–62.
- Ponger, L., Duret, L., Mouchiroud, D., 2001. Determinants of CpG islands: expression in early embryo and isochore structure. *Genome Res.* 11, 1854–1860.
- Saccone, S., De Sario, A., Della Valle, G., Bernardi, G., 1992. The highest gene concentrations in the human genome are in telomeric bands of metaphase chromosomes. *Proc. Natl. Acad. Sci. USA* 89, 4913–4917.
- Saccone, S., Sario, A.D., Weigant, J., Raap, A.K., Valle, G.D., Bernardi, G., 1993. Correlation between isochores and chromosomal bands in the human genome. *Proc. Natl. Acad. Sci.* 90, 11929–11933.
- Saccone, S., Caccio, 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., Pavlicek, A., Federico, C., Paces, J., Bernardi, G., 2001. Genes, isochores and bands in human chromosomes 21 and 22. *Chromosome Res.* 9, 533–539.
- Schildkraut, C., Marmur, J., Doty, P., 1962. Determination of the base composition of deoxyribonucleic acids from its buoyant density in CsCl. *J. Mol. Biol.* 4, 1039–1043.
- Venter, C., et al., 2001. The sequence of the human genome. *Science* 291, 1304–1351.
- Zoubak, S., Clay, O., Bernardi, G., 1996. The gene distribution of the human genome. *Gene* 174, 95–102.