THE MOSAIC GENOME OF WARM BLOODED VERTEBRATES

G. BERNARDI, B. OLOFSSON, J. FILIPSKI, M. ZERIAL, J. SALINAS, G. CUNY, M. MEUNIER-ROTIVAL & F. RODIER

Laboratoire de Genetique Moleculaire, Institut Jacques Monod, 2 Place Jussieu, 75005 Paris, France

INTRODUCTION

Density gradient centrifugation in the presence of certain DNA ligands (Ag^+, BAMD; 1-3), allows the separation of nuclear DNA from warm-blooded vertebrates into four major components and several satellite and minor components (4-9). The former comprise: (i) two light components, L1 and L2, (5); and (ii) two heavy components, H1 and H2; a third heavy component, H3, is present at least in the human genome. The heavy components represent about one-third of the genome and account for the strong heterogeneity and marked asymmetry of main-band DNAs from warm-blooded vertebrates (4-8). In contrast, main-band DNAs from most cold-blooded vertebrates show weak heterogeneities, only slightly skewed CsCl peaks, and major components having buoyant densities which are only or mainly in the same range as the light components of warm-blooded vertebrates (5, 10, 11, and paper in preparation). The families of molecules forming the major components are derived, by the unavoidable breakage which accompanies DNA preparation, from much longer DNA segments, the isochores (8), which have an average size well above 200 kb (6,7), and are fairly homogeneous in base composition (6, 8, 12-15).

Here we have studied: (i) the distribution of several genes, of some families of interspersed repeats, and of some integrated viral sequences in the major components of genomes from warm-blooded vertebrates; and (ii) the correlation between this distribution and the base composition and codon usage of these sequences.

Proceedings of the 16th FEBS Congress
Part B, pp. 69–77

RESULTS

Genomic distribution of genes, interspersed repeats and integrated viral sequences

Table 1 lists the sequences which were investigated and the major components in which they were found. The main findings can be summarized as follows: a) Single-copy genes are located in single major components. This indicates, in agreement with previous conclusions, (4-9, 13-15): (i) that the separation of major components corresponds to a real fractionation of the genome; and (ii) that large segments around the genes tested are compositionally fairly homogeneous. b) Clustered genes are located in the same major component, as expected if isochore size is large compared to gene cluster size (4-40 kb in the cases under consideration). c) In contrast, scattered genes belonging to the same family may be located in different major components. The α and β globin gene clusters are located in the H2 (or H3) and in the L2 components of mammalian DNAs, respectively. Other gene families, like the actin genes and pseudogenes, are scattered over all DNA components (16). d) Genes present in a given major component may be located on different chromosomes. In chicken, α^A and α^D globin genes are located on the largest chromosomes, the conalbumin gene is located on a chromosome of intermediate size, and the β and ρ globin genes are present on a small macro- or on a micro-chromosome (17); therefore, the major component in which all these genes are located, H2, is present on several chromosomes. e) Conversely, genes present in different major components may be located on the same chromosome. For example, the human Ha-ras 1 and β globin genes, which belong to components H3 and L2 respectively, are both located on chromosome 11. f) The distribution of genes and gene clusters within different major components is highly non-uniform. The data of Table 1 concern a total of 34 genes corresponding to 24 "loci" (defined here as isolated genes or gene clusters), and to 14 functionally unrelated proteins. About half of the loci examined for each genome are present in the heaviest components (H2 or H3), which only represent 8% or 4%,

respectively, of the DNA. g) Families of interspersed repeated sequences are concentrated in some major components (14). For instance, the BamHI family and the CR-1 (Alu-like) family are almost only present in the two light components of mouse (13) and in the heaviest component of chicken (15), respectively. h) Integrated viral sequences are only or mainly located in a given major component. The integrated sequences of bovine leukemia virus (BLV) and hepatitis B virus (HBV) from the Alexander cell line were almost only found in components H2 and H3, respectively; those of mouse mammary tumor virus (MMTV) were only found in component L2 of Balb-c mice (18, and paper in preparation). i) The distribution of genes and interspersed repeats in the major components tends to be conserved in evolution. For instance, the α and β globin gene clusters, vimentin and c-abl genes are located in components identical or close in GC levels in different mammals. The same applies to specific families of interspersed repeats (13-15).

Gene composition and codon usage

a) The GC contents of genes, exons and introns are linearly related to those of the major components in which they are located. The slopes of the lines representing these relationships are equal to 1.9 for genes, to 3.0 for introns, and to 1.0 for exons. While "light" genes are, on the average, only slightly higher in GC than light components, "heavy" genes are increasingly higher in GC than the corresponding heavy components (Fig. 1). An increasing deviation from the unit slope is also exhibited by introns and by 5' and 3' untranslated sequences (not shown). In contrast, exons have a unit slope, but are about 10% higher in GC, on the average, than the components in which they are located (not shown). Finally, integrated viral sequences and long interspersed repeats seem to show a closer match in composition with the major components in which they are located, compared to genes (Fig. 1).

b) The higher GC level of "heavy" relative to "light" exons is due to a different codon usage and not to the amino acid composition of the corresponding proteins.

TABLE 1 : Localization of some sequences in the major DNA components of warm blooded vertebrates (a).

Sequences	Major component	Sequences	Major component
Xenopus		Rabbit	
1. α globin	L1	21. α globin	H2*
2. β globin	L1	22. β globin	L2*
Chicken		Man	
3. α_D^A globin	H2	23. α_1 globin	H3*
4. α^A globin	"	24. α_2 globin	"
5. β globin	H2	25. β^2 globin	L2
6. ρ globin	"	26. Aγ globin	"
7. conalbumin	H2	27. Gγ globin	"
8. ovalbumin	L2	28. δ globin	"
9. Y	"	29. ε globin	"
10. X	"	30. p-omc	H3*
11. vitellogenin	L2	31. vimentin	H1
		32. c-Ha ras 1	H3*
Mouse		33. c-myc	H1
		34. c-sis	H3*
12. α globin	H2	35. c-mos	H1*
13. β_M globin	L2	36. c-abl	H3*
14. β_m globin	"		
15. α_c actin	L2*	Viral & repeated seqs.	
16. α_s actin	H2*		
17. vimentin	L2	37. BLV	H2*
18. Ig_k const	L2	38. HBV	H3*
19. Ig_k var	L1	39. MMTV	L2*
20. c-abl	H2*	40. Mouse Bam H1	L1,L2

(a) Sequences were localized in separated major components or, (asterisks), in preparative BAMD-Cs_2SO_4 density gradients. References for gene sequences will be given elsewhere. Non-standard abbreviations: βM, β major; βm, β minor; α_c, α cardiac; α_s, α skeletal; p-omc, pre-pro-opiomelanocortin. Inverted commas refer to clustered genes.

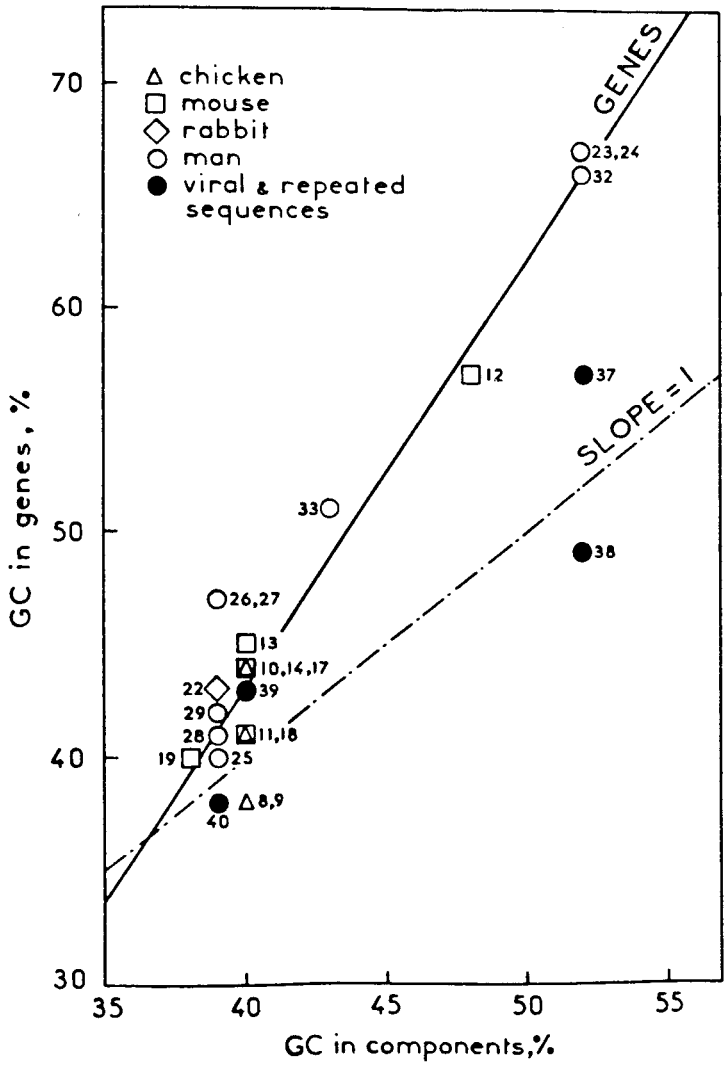

Figure 1. Plot of GC levels of genes and of viral and long interspersed repeated sequences against the GC levels and the buoyant densities of DNA components in which they are located. Numbers indicate sequences (see Table 1).

Indeed, if the codons used in "heavy" exons (53-67% GC) were replaced with the synonymous codons lowest in GC also used in the same exons, the GC levels of "heavy" exons would decrease to about 40%, a value as low as the lowest of "light" exons (40-55% GC), without any amino acid change. c) Since the vast majority of synonymous codons differ in third positions, one should expect that GC contents in codon third positions are different for "heavy" and "light" exons. This expectation is borne out, the GC level of codon third positions ranging from 43-69% to 61-90% for the "light" and the "heavy" genes, respectively. d) Genes located in "heavy" components exhibit a decreased discrimination against CpG doublets, (not shown) which tend to be avoided in vertebrate genomes. e) If the gene composition and codon usage "rules" (a-d, above) are generally valid, genes from any warm blooded vertebrate (i) should fall into compositional classes (such as those found for genes located in different components; Fig. 1); and (ii) these classes should, in turn, largely determine codon usage. Both the first and the second prediction are fulfilled (data not shown), proving the general validity of the "rules". Moreover, "light" genes predominate in the "light" genomes of cold-blooded vertebrates, as expected.

DISCUSSION

The mosaic genome organization discussed so far is typical of warm-blooded vertebrates. When the genomes of cold-blooded and warm-blooded vertebrates are compared with each other, it is clear that the main differences concern the presence of abundant, heavy components in the latter, as well as a predominance of "heavy" genes in warm-blooded vertebrates, and of "light" genes in cold-blooded vertebrates. These findings raise the question of the evolutionary origin of the heavy components present in the genome of warm-blooded vertebrates.

The evolutionary origin of the heavy components of the genome of warm-blooded vertebrates may be visualized as due to: (i) regional increases in GC levels of pre-existing "light" sequences; (ii) amplification of pre-existing "heavy" sequences. The first process is

predominates over other constraints, which may also be operational. b) The heterogeneity in DNA composition is associated with chromosomal G or R banding. The identification of isochores with the DNA segments present in G or R bands was previously suggested (8) on the basis of: (i) indications that G bands correspond to AT-rich, late-replicating DNA and R bands to GC-rich early replicating DNA; and (ii) the observation (8) that the increase in the heterogeneity of DNA composition when moving from cold-blooded to warm-blooded vertebrates (5) is paralleled by an increased G and R banding. This notion is now reinforced beyond reasonable doubt by: (i) the confirmation of the first two points mentioned above by recent results (paper in preparation); (ii) the fact that gene amplification leads to the appearance of homogeneous staining regions in chromosomes, as expected if the genome segments which are amplified are smaller than isochores; (iii) the presence in early replicating DNA, namely in R bands, of genes (human c-Ha ras 1 and α globin genes; mouse α globin gene) which are located in the heaviest component and the presence in late-replicating DNA, namely in G bands, of genes (human β globin gene) which are located in the lightest components (19).

CONCLUSIONS

The investigations reported here show that the compositional compartmentalization of the genome of warm-blooded vertebrates (i) largely dictates the base composition of genes and their codon usage; and (ii) plays a role in the timing of DNA replication and in the targeting of integration of mobile and viral sequences. From a more general viewpoint, it should be stressed that compositional compartmentalization (i) has an extremely wide evolutionary range, going as far as the mitochondrial genome (20); (ii) shows different patterns in different organisms, as exemplified here by cold-blooded and warm-blooded vertebrates; and (iii) plays a general role in genome structure and function; indeed, the different GC levels of isochores, their different CpG/GpC ratios, and the accompanying

differences in potential methylation sites are bound to be associated with differences in DNA and chromatin structure, and possibly, with differences in the regulation of gene expression.

Preliminary reports on some parts of this work were published previously (21-24).

ACKNOWLEDGEMENTS

We thank the Fogarty International Center for Advanced Study in the Health Sciences, National Institutes of Health, Bethesda, Md. 20205, for a scholarship to G.B., the Institut National de la Sante et de la Recherche Medicale, Paris, France , the Associazione Italiana per la Ricerca sul Cancro, Milan, Italy, the Ministerio de Educacion y Ciencia, Madrid, Spain, for Fellowships to J.F., M.Z. and J.S., respectively, and the Association pour la Recherche Contre le Cancer, Villejuif, France, for financial support.

REFERENCES

1. G. Corneo, E. Ginelli, C. Soave and G. Bernardi, Biochemistry, 7, 4373 (1968).
2. J. Cortadas, G. Macaya and G. Bernardi, Eur. J. Biochem. 76, 13 (1977).
3. G. Macaya, J. Cortadas and G. Bernardi, Eur. J. Biochem. 84, 179 (1978).
4. J. Filipski, J.P. Thiery and G. Bernardi, J. Mol. Biol. 80, 177 (1973).
5. J.P. Thiery, G. Macaya and G. Bernardi, J. Mol. Biol. 108, 219 (1976).
6. G. Macaya, J.P. Thiery and G. Bernardi, J. Mol. Biol. 108, 237 (1976).
7. J. Cortadas, B. Olofsson, M. Meunier-Rotival, G. Macaya and G. Bernardi, Eur. J. Biochem. 99, 179 (1979).
8. G. Cuny, P. Soriano, G. Macaya and G. Bernardi, Eur. J. Biochem. 115, 227 (1981).
9 B. Olofsson and G. Bernardi, Eur. J. Biochem. 130, 241 (1983).

10. A.P. Hudson, G. Cuny, J. Cortadas, A.E.V. Haschemeyer and G. Bernardi, Eur. J. Biochem. 112, 203 (1980).
11. V. Pizon, G. Cuny and G. Bernardi, Eur. J. Biochem. 140, 25 (1984);
12. P. Soriano, G. Macaya and G. Bernardi, Eur. J. Biochem. 115, 235 (1981).
13. M. Meunier-Rotival, P. Soriano, G. Cuny, F. Strauss and G. Bernardi, Proc. Natl. Acad. Sci. U.S.A. 79, 355 (1982).
14. P. Soriano, M. Meunier-Rotival and G. Bernardi, Proc. Natl. Acad. Sci. U.S.A. 80, 1816 (1983).
15. B. Olofsson and G. Bernardi, Biochem. Biophys. Acta 740, 339 (1983).
16. P. Soriano, P. Szabo and G. Bernardi, EMBO J. 1, 579 (1982).
17. S.H. Hughes, E. Stubblefield, F. Payver, J.D. Engel, J.B. Dodgson, D. Spector, B. Cordell, R.T. Schimke and H. E. Varmus, Proc, Natl. Acad. Sci. U.S.A. 76, 1348 (1979).
18. R. Kettman, M.Meunier-Rotiva, J. Cortadas, G. Cuny, J. Ghysdael, M. Mammerickx, A. Burny and G. Bernardi, Proc. Natl. Acad. Sci. U.S.A. 76, 4822 (1979).
19. M.A. Goldman, G.P. Holmquist, M.C. Gray, L.A. Caston, A. Nag, Science 224, 686 (1984).
20. G. Bernardi, Folia Biologica 29, 82 (1983).
21. G. Bernardi in Mutations, Biology and Society, D. N. Walcher, N. Kretschmer, H. L. Barnett, Eds., (Masson, New York, 1978) p. 327.
22. C. Cuny, G. Macaya, M. Meunier-Rotival, P. Soriano and G. Bernardi, in Genetic Engineering, H.W. Boyer and S. Nicosia, Eds. (Elsevier-North Holland Biomedical Press, Amsterdam, 1978), p. 109.
23. G. Bernardi, in Recombinant DNA and Genetic Experimentation, J. Morgan and W. J. Whelan, Eds. (Pergamon Press, London, 1979), p. 15.
24. G. Bernardi, in Genetic Manipulation: Impact on Man and Society, W. Arber, et al., Eds. (Cambridge University Pres, Cambridge, 1984), p. 171.