The Mosaic Genome of Warm-Blooded Vertebrates

Giorgio Bernardi, Birgitta Olofsson, Jan Filipski
Marino Zerial, Julio Salinas, Gerard Cuny
Michele Meunier-Rotival, Francis Rodier

Summary. Most of the nuclear genome of warm-blooded vertebrates is a mosaic of very long (>>200 kilobases) DNA segments, the *isochores*; these isochores are fairly homogeneous in base composition and belong to a small number of major classes distinguished by differences in guanine-cytosine (GC) content. The families of DNA molecules derived from such classes can be separated and used to study the genome distribution of any sequence which can be probed. This approach has revealed (i) that the distribution of genes, integrated viral sequences, and interspersed repeats is highly nonuniform in the genome, and (ii) that the base composition and ratio of CpG to GpC in both coding and noncoding sequences, as well as codon usage, mainly depend on the GC content of the isochores harboring the sequences. The *compositional compartmentalization* of the genome of warm-blooded vertebrates is discussed with respect to its evolutionary origin, its causes, and its effects on chromosome structure and function.

Density gradient centrifugation in the presence of certain DNA ligands—such as silver ion, Ag^+; BAMD [3,6-bis-(acetatomercurimethyl)dioxane] (*1–3*)—results in the separation of nuclear DNA from warm-blooded vertebrates into four major components and several satellite and minor (such as ribosomal DNA) components (*4–9*). The former include: (i) two light components, L1 and L2, poorly or not resolved in some genomes (*5*); and (ii) two or three heavy components, H1, H2, H3. Figure 1 shows the relative amounts and the buoyant densities of the major components of the chicken, mouse, and human genomes (*8, 9*). The heavy components account for the strong heterogeneity and marked asymmetry of main-band DNA's from warm-blooded vertebrates (*4–8*). In contrast, main-band DNA's from most cold-blooded vertebrates show (Fig. 1) weak heterogeneities, only slightly skewed CsCl peaks, and major components that have buoyant densities which are only or mainly in the same range as the light components of warm-blooded vertebrates (*5, 10, 11*). 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, 12*) (Fig. 2), which have an average size well above 200 kilobases (kb) (*6, 7*), and are fairly homogeneous in base composition (*6, 8, 13–16*).

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 (*17–21*).

Distribution of Genes, Interspersed Repeats, and Integrated Viral Sequences

The sequences investigated and the major components in which they were found are shown in Table 1. The main findings (described below), concern three issues: (i) some properties of isochores, as judged from the localization of specific sequences; (ii) the relation between isochores and chromosomes; and (iii) the genomic distribution of the sequences investigated.

Single-copy genes are located in single major components (Fig. 3). This indicates that the separation of major components corresponds to a real fractionation of the genome; and that large segments around the genes tested are compositionally fairly homogeneous. Indeed, if either point were incorrect, a given single-copy gene would be found in more than one component. The same conclusions were drawn earlier as a result of different experimental approaches (*4–9, 14–16*). The only exception to these conclusions is the c-myc gene which seems to be located at an H1-H2 border.

Clustered genes are located in the same major component (Fig. 3), as expected if isochore size is large compared to gene cluster size, from 4 to 40 kb in the cases under consideration (Table 1). In contrast, scattered genes belonging to the same family may be located in different major components. For instance, the actin genes and pseudogenes are scattered over all DNA components (*22*).

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 microchromosome (*23*); therefore, the major component in which all these genes are located, H2, is present on several chromosomes. 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.

The distribution of genes and gene clusters within different major components is highly nonuniform. The data of Table 1 were obtained from study 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 percent, respectively, of the DNA.

Families of interspersed repeated sequences are concentrated in some major components (*15*). For instance, the Bam HI family and the CR-1 (Alu-like) family are almost only present in the two light components of mouse (*14*) and in the heaviest component of chicken (*16*), respectively.

Integrated viral sequences are only or mainly located in a given major component. The integrated sequences of bovine-leukemia virus (BLV) and hepatitis

The authors are members of the Laboratoire de Genetique Moleculaire, Institut Jacques Monod, 2 Place Jussieu, 75005 Paris, France.

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 mainly found in component L2 of mice (*24*).

The distribution of genes and interspersed repeats in the major components seems 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 (*14–16*).

Gene Composition and Codon Usage

The GC (G, guanine; C, cytosine) contents of genes, exons, and introns are linearly related to those of the major components in which they are located (Fig. 4, A to C). 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 content than light components, "heavy" genes have increasingly more GC than the corresponding heavy components (Fig. 4A). An increasing deviation from the unit slope is also exhibited by introns (Fig. 4B). In contrast, exons show a unit slope, but are about 10 percent higher in GC, on the average, than the components in which they are located (Fig. 4C). The larger scatter of points exhibited by exons compared to introns and genes are probably due to their smaller sizes, but a few exons deviate from the common relationship. 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. 4A).

The higher GC content of heavy relative to light exons is due to a different codon usage and not to the amino acid composition of the corresponding proteins. Indeed, if the codons used in heavy exons (53 to 67 percent GC) were replaced with the synonymous codons lowest in GC also used in the same exons, the GC contents of heavy exons would decrease to about 40 percent, a value as low as that of the lowest of light exons (40 to 55 percent GC), without any amino acid change. In fact, the lowest "allowed" GC level for heavy exons is practically identical to that of light exons. A striking example of large differences in the GC content of exons not accompanied by changes in amino acids is that of cardiac and skeletal mouse α-actin genes. These are located in L2 and H2 components, respectively, and differ by 8 and 16 percent in overall and third-position GC content, respectively (see below); yet, the corresponding proteins show only a 1 percent difference in amino acids (*25*).

Since most of the synonymous codons differ in third positions, we 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 to 69 percent to 61 to 90 percent for the light and the heavy genes, respectively (Fig. 5A). A few genes show, however, a deviation from the general trend (see next section).

Genes located in heavy components show a decreased discrimination against CpG doublets, which tend to be avoided in vertebrate genomes (*26*). In most cases CpG is strongly discriminated against in light exons, but only slightly in heavy exons (Fig. 1, right). As would be expected, most of the differences between heavy and light exons concern intercodon CpG, namely doublets in which third-position C is followed by first-position G. Intracodon differences, namely preferential usage of codons containing CpG instead of synonymous co-

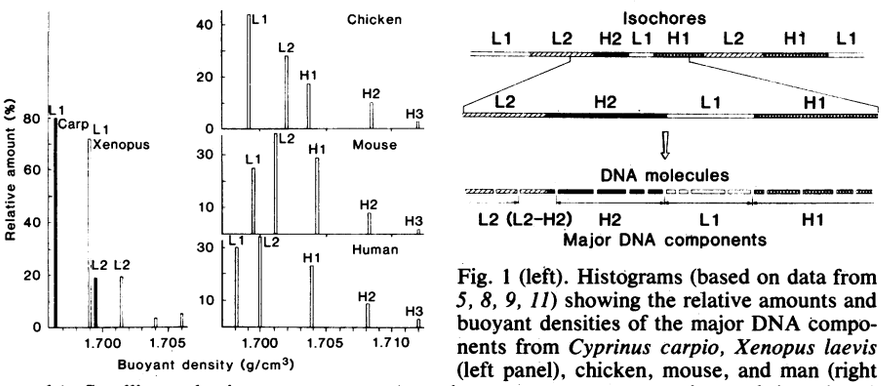


Fig. 1 (left). Histograms (based on data from *5, 8, 9, 11*) showing the relative amounts and buoyant densities of the major DNA components from *Cyprinus carpio, Xenopus laevis* (left panel), chicken, mouse, and man (right panels). Satellite and minor components (namely, components representing each less than 3 percent of DNA) (*8*) are not shown, with the exception of the minor components from mouse and chicken which have the buoyant density of H3; no genes have been localized so far in these minor components. Carp and *Xenopus* genomes represent extreme cases of low and high heterogeneity among cold-blooded vertebrates. Notice that even in *Xenopus*, DNA having a density higher than 1.704 represents less than 10 percent of the genome, as compared with 30 to 40 percent for warm-blooded vertebrates. Fig. 2 (right). Scheme depicting the mosaic organization of nuclear DNA from warm-blooded vertebrates. When the very long DNA segments, fairly homogeneous in base composition, the isochores, undergo breakage during DNA preparation, four major families of molecules having different GC contents are generated. These major DNA components can be resolved from each other by preparative density gradient centrifugation in the presence of certain DNA ligands. Component H3, minor, and satellite components are neglected in this scheme. If isochores correspond, as suggested [(*8*) and present work], to the DNA segments present in Giemsa and Reverse chromosomal bands as obtained at high resolution (*42*), their average size is ~1250 kb (*41*). This means that the 30- to 100-kb DNA molecules from the preparations used in this work are 12 to 40 times smaller in size than isochores; DNA molecules bridging contiguous isochores are, therefore, as expected rare in our preparations (one such molecule, L2-H2, is shown).

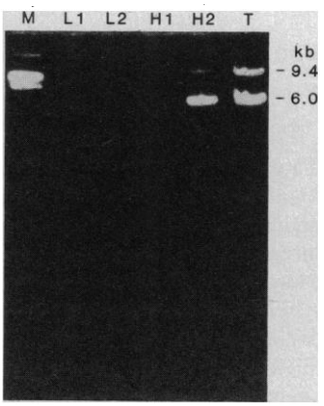


Fig. 3. A typical experiment for localizing a gene in the major components of a genome from a warm-blooded vertebrate. A chicken β-globin probe was hybridized to Eco RI digests of unfractionated chicken DNA (T) and its major components L1, L2, H1, and H2. The probe revealed not only the 6-kb fragments carrying the β-globin gene, but also 9.4 kb-fragment carrying the ρ-globin gene; both genes, which belong to the same cluster (Table 1), are located in the H2 component. M is a size-marker restriction digest (*48*). An alternative approach for gene localization consisted in hybridizing appropriate cloned complementary DNA probes to restriction digests from DNA fractions obtained by preparative density gradient centrifugation in the presence of DNA ligands; fractions were then analyzed in CsCl density gradients in order to assess their buoyant densities and assign them to major components. In every case, the sizes of the hybridizing restriction fragments were in agreement with those already published (references in Table 1). The DNA preparations submitted to fractionation had molecular sizes between 30 and 100 kb. ^{32}P-labeling of probes was done according to Rigby *et al.* (*49*). Fragment transfers from 0.8 to 1 percent agarose gels onto nitrocellulose, hybridization, and autoradiography were done as described (*48*).

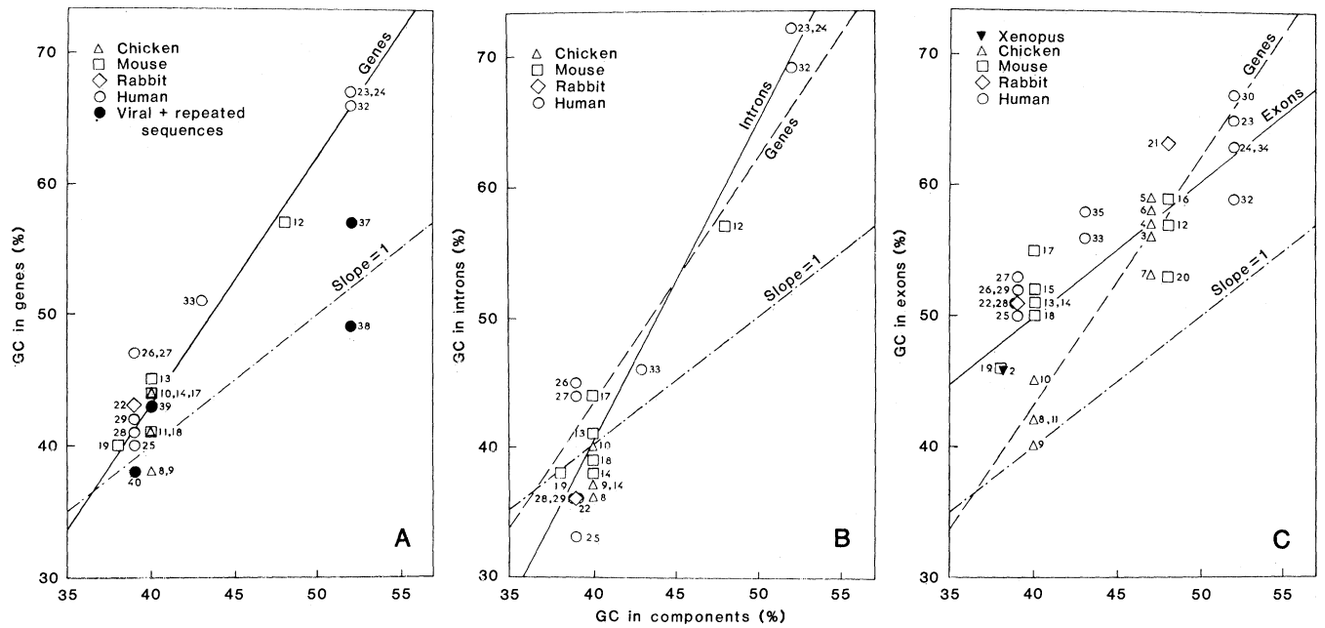


Fig. 4. Plot of the GC contents of (A) genes, viral and long interspersed repeated sequences, (B) introns, and (C) exons against the GC levels and the buoyant densities of DNA components in which they are located. The numbers indicate genes (see Table 1). The line was drawn using the least-square method. The unit slope line corresponds to the coincidence in GC contents of genes and major components in which genes are located.

dons not containing CpG, are, however, also found in heavy genes (not shown). Differences in CpG levels similar to those of exons are also found in the introns of the corresponding genes and in the untranslated regions (data not shown).

If the composition of genes and the codon usage rules (as discussed above) are generally valid, genes from any warm-blooded vertebrate should fall into compositional classes such as those found for genes located in different components (Fig. 4A), and these classes should, in turn, largely determine codon usage. Both the first and the second prediction are fulfilled (Fig. 6B), proving the general validity of the "rules." Two additional points made by Fig. 6 are that human genes (as well as those of other warm-blooded vertebrates tested) are predominantly heavy (Fig. 6B) although less so than indicated by the smaller gene sample of Table 1, and that, in contrast, light genes predominate in the light genomes of cold-blooded vertebrates, as would be expected (Fig. 6A). In this second case too, GC content in third positions of codons and CpG/GpC ratios are correlated with overall GC content (not shown).

Implications

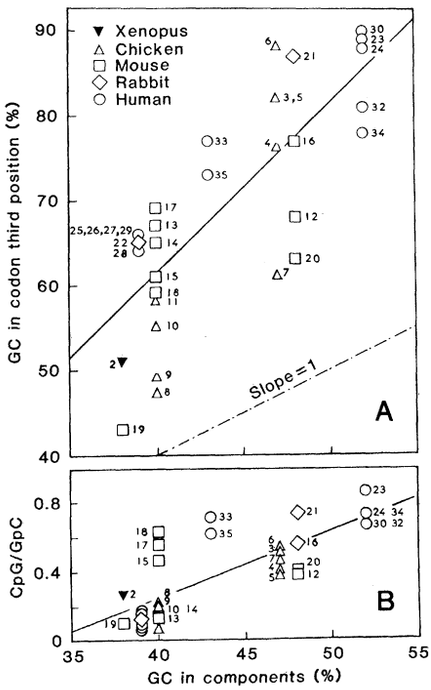
The mosaic genome organization discussed so far is typical of warm-blooded vertebrates. When the genomes of cold-blooded and warm-blooded vertebrates are compared with each other, it is clear that the main difference concerns the presence of abundant, heavy components in the latter (Fig. 1). As was just mentioned, this difference is accompanied by 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 being due to (i) regional increases in GC content of preexisting light sequences; (ii) amplification of preexisting heavy sequences; (iii) de novo formation of sequences. While there is no evidence, so far, in favor of the latter process, (which probably is operational in the generation of the clustered short repeats of satellite DNA's), the other two are well-documented. The first one is supported by our finding that light genes, ancestrally present in the light genomes of cold-blooded vertebrates (Figs. 1 and 6), have become heavy and are found in the heavy components of warm-blooded vertebrates (Fig. 6). For instance, the β-globin gene is light in *Xenopus* but heavy in chicken (Fig. 4C), and the insulin gene is light in hagfish (45 percent GC) but heavy in man (64 percent GC). The second process is exemplified by the amplification of heavy Alu sequences in mammalian genomes. The large difference in copy number of mouse and human Alu sequences (*27*) indicates that the amplifications of such interspersed repeats were recent events compared to the formation of heavy isochores.

The molecular mechanisms underlying these processes are different. The amplification of preexisting heavy Alu repeats, like that of interspersed repeats in general, implies rounds of duplication and insertion events. The specific genome distribution of different families of interspersed repeats indicates that such insertion events were targeted toward isopycnic isochores, namely isochores of matching GC content, or that the insertion stability was dependent upon such a correlation (or both). It should be mentioned here that the viral sequences explored exhibit the same phenomenon.

In contrast, the main process responsible for the formation of heavy isochores, namely the regional increase in GC, was brought about by (i) point mutations in coding sequences, mainly in third positions (Fig. 5A); or (ii) point mutations, deletions, and additions in introns (*28*). This process raises two questions concerning, respectively, its causes and its effects.

The causes for the regional GC increases are unknown at present, but they might only or mainly be of a structural nature, since they affect both coding and noncoding sequences to comparable extents. These causes might be related to the requirements of chromosome structure at the temperature prevailing in the cells of warm-blooded vertebrates. Addi-



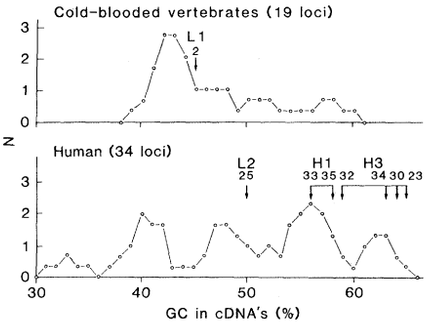


Fig. 5 (left). Plot of GC levels of third positions of codons (A) and ratios of CpG to GpC (B) for genes and exons against the GC contents of DNA components harboring the genes. Other indications as in Fig. 5. The cluster of unnumbered genes in B comprises genes 22 and 25 to 29. Fig. 6 (right). Plot of number of genes or gene clusters in 3 percent GC intervals (N) against GC of corresponding complementary DNA's. Primary data were from the EMBL (European Molecular Biology Laboratories) library (April 1984). Average GC values were taken for gene clusters. Short DNA sequences (<100 bp) were not taken into account. Arrows refer to GC contents of exons studied (Fig. 4C; numbers refer to genes; see Table 1). Results for loci from cold-blooded vertebrate (A) and from human genome (B) are shown. This latter plot tests the prediction that genes from warm-blooded vertebrates fall into compositional classes shown in (B). Since the information on genes available in data banks is too limited, this test was done with exons. A comparison of Fig. 4, A and C, indicates that such a choice has two drawbacks, namely (i) that exons corresponding to a given component show a larger GC scatter than genes and (ii) that exons corresponding to different components show a smaller difference in GC than genes. In other words, the use of exons instead of genes minimizes differences between compositional classes.

Table 1. Localization of some sequences in the major DNA components of warm-blooded vertebrates. References for gene sequences and hybridization results are given in parentheses; GC content of genes, exons, introns, codon third bases, and ratios of CpG to GpC (see Figs. 4 and 5) were calculated from these data. Nonstandard abbreviations: βM, β major; βm, β minor; αc, α cardiac; αs, α skeletal; p-omc, pre-pro-opiomelanocortin. Sequences were localized in separated major components or in preparative density gradients as indicated (see Fig. 3 legend).

Sequences*	Major component	Sequences*	Major component*
Xenopus		*Rabbit*	
1. α-globin (*50*)	L1	21. α-globin (*69*)	H2†
2. β-globin (*51*)	L1	22. β-globin (*70–73*)	L2†
Chicken		*Man*	
3. α^A-globin (*52*)	H2	23. α_1-globin (*74*)	H3†
4. α^D-globin (*52*)	H2	24. α_2-globin (*75, 76*)	H3
5. β-globin (*53*)	H2	25. β-globin (*77, 78*)	L2
6. ρ-globin (*54*)	H2	26. Aγ-globin (*79*)	L2
7. conalbumin (*55*)	H2	27. Gγ-globin (*79, 80*)	L2
8. ovalbumin (*56, 57*)	L2	28. δ-globin (*81*)	L2
9. Y (*58*)	L2	29. ϵ-globin (*82, 83*)	L2
10. X (*59*)	L2	30. p-omc (*84, 85*)	H3†
11. vitellogenin (*60*)	L2	31. vimentin (*65*)‡	H1
Mouse		32. c-Ha-ras 1 (*86*)	H3†
12. α-globin (*61*)	H2	33. c-myc (*87, 88*)‡	H1, H2
13. β_M-globin (*62, 63*)	L2	34. c-sis (*89*)	H3†
14. β_m-globin (*63*)	L2	35. c-mos (*90*)	H1†
15. α_c-actin (*64*)	L2†	36. c-abl (*68*)	H3†
16. α_s-actin (*64*)	H2†	*Viral sequences*§	
17. vimentin (*65*)‡	L2	37. BLV (*91*)	H2†
18. Ig^k constant (*66*)	L2	38. HBV (*92, 93*)	H3†
19. Ig^k variable (*67*)	L1	39. MMTV (*94*)	L2†
20. c-abl (*68*)	H2†	*Long repeated sequences*	
		40. Mouse Bam HI (*15, 95*)	L1, L2

**Xenopus* α- and β-globin genes are clustered; chicken α^A- and α^D-globin, β- and ρ-globin, ovalbumin, Y, and X genes are clustered within 4.4, 10, and 40 kb, respectively (*96–98*); mouse β-major and β-minor globin genes are clustered within 16 kb (*99*), and human β,γ,δ, and ϵ globin are clustered within 42.8 kb (*70*). †From preparative BAMD-Cs_2SO_4 density gradients. ‡In the case of vimentin a hamster DNA probe was used and sequence data are for the hamster vimentin gene. The localization of c-myc will be discussed elsewhere. §See text.

tional explanations are needed, however, to account (i) for the preferential distribution of genes in the heavy DNA components (Table 1 and Fig. 6), namely for a distribution which requires the largest GC change in ancestrally light genes and also the highest increase in CpG doublets; and (ii) for the higher GC level of exons and introns compared to that of the components in which they are located (Fig. 4, B and C). These features might be associated with the best protection of genes against DNA "breathing" and mutability in warm-blooded vertebrates (*29*), but there may be other causes as well.

The effects of the regional GC increases are twofold. First, a different codon strategy is used for different genes located in the same genome. This has been previously noted for rabbit (*30*) and human (*31*) α- and β-globin genes, and for human and mouse α-actin genes (*32*). What is shown here for the first time, however, is that a different codon usage (i) is the rule for different classes of genes from the same warm-blooded vertebrate genome; and (ii) is mainly determined by the location of genes in heavy or light isochores. This compositional constraint predominates over other constraints (*33–38*), which may also be operational, and be responsible for the deviations of some genes from the general relationship (Figs. 4C and 5A).

Second, 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 (*39*) that G bands correspond to AT-rich, late-replicating DNA and R bands to GC-rich early-replicating DNA; (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; and (iii) the parallel evolutionary conservation of isochores (as judged by the location of specific sequences), and of chromosomal bands (*40*). It should be recalled here (i) that, as expected, different isochores are represented on the same chromosome and the same isochore is represented on different chromosomes; and (ii) that the estimated average size of isochores (>>200 kb) is compatible with that of chromosomal bands (~1250 kb) (*41*) for the more than 2000 bands obtained at high resolution (*42*). This notion is effectively reinforced by (i) the confirmation of the first two points mentioned above by recent results obtained both in our and in another laboratory (*41*); (ii) the fact that gene amplifi-

cation leads to the appearance of homogeneous staining regions in chromosomes (*43*), 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 (*44–46*) which are located in the heaviest component (human c-Ha-ras 1 and α-globin genes and the mouse α-globin gene) and the presence in late-replicating DNA, namely in G bands, of genes (*46*) that are located in the lightest components (human β-globin gene); (iv) the parallelism between the preferential distribution of genes in GC-rich isochores found here and the paucity of genes found in G bands (*41*).

Conclusions

The investigations reported in this article 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 (*47*); (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 ratios of CpG to GpC, 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.

References and Notes

1. G. Corneo, E. Ginelli, C. Soave, G. Bernardi, *Biochemistry* **7**, 4373 (1968).
2. J. Cortadas, G. Macaya, G. Bernardi, *Eur. J. Biochem.* **76**, 13 (1977).
3. G. Macaya, J. Cortadas, G. Bernardi, *ibid.* **84**, 179 (1978).
4. J. Filipski, J. P. Thiery, G. Bernardi, *J. Mol. Biol.* **80**, 177 (1973).
5. J. P. Thiery, G. Macaya, G. Bernardi, *ibid.* **108**, 219 (1976).
6. G. Macaya, J. P. Thiery, G. Bernardi, *ibid.*, p. 237 (1976).
7. J. Cortadas, B. Olofsson, M. Meunier-Rotival, G. Macaya, G. Bernardi, *Eur. J. Biochem.* **99**, 179 (1979).
8. G. Cuny, P. Soriano, G. Macaya, G. Bernardi, *ibid.* **115**, 227 (1981).
9. B. Olofsson and G. Bernardi, *ibid.* **130**, 241 (1983).
10. A. P. Hudson, G. Cuny, J. Cortadas, A. E. V. Haschemeyer, G. Bernardi, *ibid.* **112**, 203 (1980).
11. V. Pizon, G. Cuny, G. Bernardi, *ibid.* **140**, 25 (1984); G. Cuny, V. Pizon, G. Bernardi, A. E. V. Haschemeyer, G. Bernardi, in preparation.
12. From the Greek isos, equal; and choros, region.
13. P. Soriano, G. Macaya, G. Bernardi, *Eur. J. Biochem.* **115**, 235 (1981).
14. M. Meunier-Rotival, P. Soriano, G. Cuny, F. Strauss, G. Bernardi, *Proc. Natl. Acad. Sci. U.S.A.* **79**, 355 (1982).
15. P. Soriano, M. Meunier-Rotival, G. Bernardi, *ibid.* **80**, 1816 (1983).
16. B. Olofsson and G. Bernardi, *Biochem. Biophys. Acta* **740**, 339 (1983).
17. Preliminary reports on some of these results have been published (*18–21*).
18. G. Bernardi, in *Mutations, Biology and Society*, D. N. Walcher, N. Kretschmer, H. L. Barnett, Eds. (Masson, New York, 1978), p. 327.
19. G. Cuny, G. Macaya, M. Meunier-Rotival, P. Soriano, G. Bernardi, in *Genetic Engineering*, H. W. Boyer and S. Nicosia, Eds. (Elsevier-North-Holland, Amsterdam, 1978), p. 109.
20. G. Bernardi, in *Recombinant DNA and Genetic Experimentation*, J. Morgan and W. J. Whelan, Eds. (Pergamon, London, 1979), p. 15.
21. G. Bernardi, in *Genetic Manipulation: Impact on Man and Society*, W. Arber, K. Illmensee; W. J. Peacock, P. Starlinger, Eds. (Cambridge Univ. Press, London, 1984), p. 171.
22. P. Soriano, P. Szabo, G. Bernardi, *EMBO J.* **1**, 579 (1982).
23. S. H. Hughes *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **76**, 1348 (1979).
24. R. Kettman *et al.*, *ibid.*, p. 4822; M. Fezial *et al.*, in preparation; J. Salinas *et al.*, in preparation.
25. J. Vanderkerckhove and K. Weber, *J. Mol. Biol.* **126**, 783 (1979).
26. M. N. Swartz, T. A. Trautner and A. Kornberg, *J. Biol. Chem.* **237**, 1961 (1962).
27. C. W. Schmid and W. R. Jelinek, *Science* **216**, 1065 (1982).
28. F. Crick, *ibid.* **204**, 264 (1979).
29. L. Orgel, personal communication.
30. W. Salser, *Cold Spring Harbor Symp. Quant. Biol.* **412**, 985 (1977).
31. B. G. Forget *et al.*, in *Eukaryotic Gene Regulation*, R. Axel, T. Maniatis, C. F. Fox, Eds. (Academic Press, New York, 1979), p. 367.
32. S. Alonso, A. Minty, M. Buckingham, in preparation; A. Hanauer *et al.*, *Nucleic Acids Res.* **11**, 3503 (1983).
33. T. Ikemura, *J. Mol. Biol.* **151**, 389 (1981).
34. M. Hasegawa, T. Yasunaga, T. Miyata, *Nucleic Acids Res.* **7**, 2073 (1979).
35. G. Modiano, G. Battistuzzi, A. G. Motulski, *Proc. Natl. Acad. Sci. U.S.A.* **78**, 1110 (1981).
36. D. J. Lipman and W. J. Wilbur, *J. Mol. Biol.* **163**, 363 (1983).
37. H. Grosjean and W. Fiers, *Gene* **18**, 199 (1982).
38. R. Grantham, C. Gautier, M. Gouy, M. Jacobzone, R. Mercier, *Nucleic Acids Res.* **9**, 43 (1981).
39. D. E. Comings, *Annu. Rev. Genet.* **12**, 25 (1978).
40. H. N. Seuanez, *The Phylogeny of Human Chromosomes* (Springer, Berlin, 1979).
41. G. Holmquist, M. Gray, T. Porter, J. Jordan, *Cell* **31**, 121 (1982); L. Medrano *et al.*, in preparation.
42. J. J. Yunis, *Hum. Genet.* **56**, 293 (1981).
43. R. T. Schimke, Ed., *Gene Amplification* (Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y., 1982).
44. A. Furst, E. H. Brown, J. D. Braunstein, C. L. Schildkraut, *Proc. Natl. Acad. Sci. U.S.A.* **78**, 1023 (1981).
45. R. E. Calza, L. A. Eckhardt, T. Del Giudice, C. L. Schildkraut, *Cell* **36**, 689 (1984).
46. M. A. Goldman, G. P. Holmquist, M. C. Gray, L. A. Caston, A. Nag, *Science* **224**, 686 (1984).
47. G. Bernardi, *Trends Biochem. Sci.* **4**, 197 (1979); in *Mitochondrial Genes*, P. P. Slonimski *et al.*, Eds. (Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y., 1982), p. 269; *Folia Biologica* **29**, 82 (1983).
48. M. Meunier-Rotival, J. Cortadas, G. Macaya, G. Bernardi, *Nucleic Acids Res.* **6**, 2109 (1979).
49. P. W. J. Rigby, M. Dieckmann, C. Rhodes, P. Berg, *J. Mol. Biol.* **113**, 237 (1977).
50. J. G. Williams, R. M. Kay, R. K. Patient, *Nucleic Acids Res.* **8**, 4247 (1980).
51. R. K. Patient, J. A. Elkington, R. M. Kay, J. G. Williams, *Cell* **21**, 565 (1980).
52. R. I. Richards and J. R. E. Wells, *J. Biol. Chem.* **225**, 9306 (1980); J. B. Dodgson, K. C. McCune, D. J. Rusling, A. Krust, J. D. Engel, *Proc. Natl. Acad. Sci. U.S.A.* **78**, 5998 (1981).
53. R. I. Richards, J. Shine, A. Ullrich, J. R. E. Wells, H. M. Goodmann, *Nucleic Acids Res.* **7**, 1137 (1979).
54. J. B. Roninson and V. M. Ingram, *Proc. Natl. Acad. Sci. U.S.A.* **78**, 4782 (1981).
55. J. M. Jeltsch and P. Chambon, *Eur. J. Biochem.* **122**, 291 (1982).
56. K. O'Hare, R. Breathnach, C. Benoist, P. Chambon, *Nucleic Acids Res.* **7**, 321 (1979).
57. S. L. C. Woo *et al.*, *Biochemistry* **20**, 6437 (1981).
58. R. Heilig, R. Murakowski, C. Kloepper, J.-L. Mandel, *Nucleic Acids Res.* **10**, 4363 (1982).
59. R. Heilig, F. Perrin, F. Gannon, J.-L. Mandel, P. Chambon, *Cell* **20**, 625 (1980).
60. B. Wieringa, thesis, Groningen University (1980).
61. Y. Nishioka and P. Leder, *Cell* **18**, 875 (1979).
62. D. A. Konkel, S. M. Tilghman, P. Leder, *ibid.* **15**, 1125 (1978).
63. D. A. Konkel, J. V. Maizel, Jr., P. Leder, *ibid.* **18**, 865 (1979).
64. A. J. Minty, S. Alonso, M. Caravatti, M. E. Buckingham, *ibid.* **30**, 185 (1982).
65. W. Quax, W. Vree Egberts, W. Hendriks, Y. Quax-Jeuken, H. Bloemendal, *ibid.* **35**, 215 (1983).
66. E. E. Max, J. V. Maizel, Jr., P. Leder, *J. Biol. Chem.* **256**, 5116 (1981).
67. Y. Nishioka and P. Leder, *ibid.* **255**, 3691 (1980).
68. J. Y. J. Wong, D. Baltimore, R. Lee, Y. Groner, F. Ledley, S. Goff, *Cell* **36**, 349 (1984).
69. H. C. Heindell, A. Lin, G. V. Paddock, G. M. Studnika, W. A. Salser, *ibid.* **15**, 43 (1978).
70. A. Efstratiadis *et al.*, *ibid.* **21**, 653 (1980).
71. R. C. Hardison, E. T. Butler III, E. Lacy, T. Maniatis, *ibid.* **18**, 1285 (1979).
72. A. Van Ooyen, J. Van Den Berg, N. Mantei, C. Weissmann, *Science* **206**, 337 (1979).
73. P. Dierks, A. Van Ooyen, N. Mantei, C. Weissmann, *Proc. Natl. Acad. Sci. U.S.A.* **78**, 1411 (1981).
74. A. Michelson and S. H. Orkin, *Cell* **22**, 371 (1980).
75. S. A. Liebhaber, J. Goossens, Y. W. Kan, *Proc. Natl. Acad. Sci. U.S.A.* **77**, 7054 (1980).
76. J. T. Wilson, L. B. Wilson, V. B. Reddy, C. Cavallesco, P. K. Ghosh, J. K. de Riel, B. G. Forget, S. M. Weissman, *J. Biol. Chem.* **255**, 2807 (1980).
77. R. M. Lawn, A. Efstratiadis, C. O'Connel, T. Maniatis, *Cell* **21**, 647 (1980).
78. C. A. Marotta, J. T. Wilson, B. G. Forget, S. M. Weissman, *J. Biol. Chem.* **252**, 5040 (1977).
79. J. L. Slighton, A. E. Blechl, O. Smithies, *Cell* **21**, 627 (1980).
80. C. Cavallesco, B. G. Forget, J. K. de Riel, L. B. Wilson, J. T. Wilson, S. M. Weissman, *ibid.*, p. 215.
81. R. A. Spritz, J. K. de Riel, B. G. Forget, S. M. Weissman, *ibid.*, p. 639.
82. F. E. Baralle, C. C. Shoulders, J. Proudfoot, *ibid.* p. 621.
83. F. E. Baralle, C. C. Shoulders, S. Goodbourn, A. Jeffreys, J. Proudfoot, *Nucleic Acids Res.* **8**, 4393 (1980).
84. M. Cochet, A. C. Y. Chang, S. N. Cohen, *Nature (London)* **297**, 335 (1982).
85. H. Takahashi, Y. Teranishi, S. Nakanishi, S. Numa, *FEBS Lett.* **135**, 97 (1981).
86. D. J. Capon, Y. E. Chen, A. D. Levinson, P. H. Seeburg, D. V. Goeddel, *Nature (London)* **302**, 33 (1983).
87. W. W. Colby, E. Y. Chen, D. H. Smith, A. D. Levinson, *ibid.* **301**, 722 (1983).
88. C. Gazin *et al.*, *EMBO J.* **3**, 383 (1984).
89. S. F. Josephs, C. Guo, L. Ratner, F. Wong-Staal, *Science* **223**, 487 (1984).
90. R. Watson, M. Oskarsson, G. F. Vande Woude, *Proc. Natl. Acad. Sci. U.S.A.* **79**, 4078 (1982).
91. D. Couez, J. Deschamps, R. Kettman, R. M. Stephens, R. V. Gilden, A. Burny, *J. Virol.* **49**, 615 (1984).
92. F. Galibert, F. Mandart, F. Fitoussi, P. Tiollais, P. Charney, *Nature (London)* **281**, 646 (1979).
93. R. Koshy, S. Koch, A. Freytag von Loringhoven, R. Kahmann, K. Murray, P. H. Hofschneider *Cell* **34**, 215 (1983).
94. M. Crépin, personal communication.
95. M. Meunier-Rotival and G. Bernardi, *Nucleic Acids Res.* **12**, 1593 (1984).
96. J. D. Engel and J. B. Dodgson, *Proc. Natl. Acad. Sci. U.S.A.* **77**, 2596 (1980).
97. J. B. Dodgson, J. Sommer, J. D. Engel, *Cell* **17**, 879 (1979).
98. A. Royal, A. Garapin, B. Cami, F. Perrin, J. L. Mandel, M. Le Meur, F. Bregegre, F. Gannon, J. P. Le Pennec, P. Chambon, P. Kourilsky, *Nature (London)* **279**, 125 (1979).
99. P. Leder, J. N. Hansen, D. Konkel, A. Leder, Y. Nishioka, C. Talkington, *Science* **209**, 1336 (1980).

100. We thank the Fogarty International Center for Advanced Study in the Health Sciences, National Institutes of Health, Bethesda, 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. We thank A. Huyard and P. Soriano for help in experiments; and M. Buckingham, I. Dawid, L. Orgel, and M. Singer for comments on this paper. Cloned DNA probes were obtained from S. Alonso and M. Buckingham (Paris); P. Chambon and J.-L. Mandel (Strasbourg); S. N. Cohen (Stanford); M. Crepin (Paris); C. M. Croce (Philadelphia); J. Ferrer and S. Kashmire (Philadelphia); M. Gruber, C. P. W. Meijlink and B. Wieringa (Gröningen); D. R. Lowy (Bethesda); K. M ray (Edinburgh); R. K. Patient (London); C Reynaud and K. Scherrer (Paris); M. Sir (Bethesda); P. Tambourin (Orsay); P. Tio (Paris); C. Weissman (Zurich); R. Willian (London).

6 August 1984; accepted 6 November 1984

RESEARCH ARTICLE

Expression of *Plasmodium falciparum* Circumsporozoite Proteins in *Escherichia coli* for Potential Use in a Human Malaria Vaccine

James F. Young, Wayne T. Hockmeyer, Mitchell Gross
W. Ripley Ballou, Robert A. Wirtz, James H. Trosper
Richard L. Beaudoin, Michael R. Hollingdale
Louis H. Miller, Carter L. Diggs, Martin Rosenberg

Abstract. *The circumsporozoite (CS) protein of the human malaria parasite* Plasmodium falciparum *may be the most promising target for the development of a malaria vaccine. In this study, proteins composed of 16, 32, or 48 tandem copies of a tetrapeptide repeating sequence found in the CS protein were efficiently expressed in the bacterium* Escherichia coli. *When injected into mice, these recombinant products resulted in the production of high titers of antibodies that reacted with the authentic CS protein on live sporozoites and blocked sporozoite invasion of human hepatoma cells in vitro. These CS protein derivatives are therefore candidates for a human malaria vaccine.*

The feasibility of immunization against the sporozoite stage of malaria has been established. Irradiated sporozoites have been used to immunize and protect both man and animals (*1*). This protection is correlated with antibody to a protein on the surface of the sporozoite—circumsporozoite (CS) protein (*2–6*). Monoclonal antibodies (Mab's) to the CS protein block infection with sporozoites in vitro and protect animals in vivo (*3, 4, 6*).

Recently, Dame *et al.* (*7*) cloned and sequenced the complete CS gene of the human malaria parasite *Plasmodium falciparum*. The gene encodes a protein of 412 amino acids. This protein has a sequence typical of a membrane protein with an NH_2-terminal signal peptide and a COOH-terminal anchor domain. The most striking feature of this polypeptide is a large central repeat domain composed of 37 tetrapeptides with the sequence Asn-Ala-Asn-Pro interspersed with four tetrapeptides with the sequence Asn-Asp-Val-Pro. This general structure is analogous to that of the CS protein of the simian malaria parasite *P. knowlesi* (*8*), although the overall sequence homology between the CS protein of *P. falciparum* and *P. knowlesi* is very low. In fact, only two regions of approximately 15 amino acids each, in the charged sequences flanking the repeat domain, are conserved (*7*).

Protection by Mab's to the CS protein is both species- and stage-specific and, in the case of *P. knowlesi*, Mab's react with the 12-amino-acid repeat region of the CS protein (*9*). These Mab's also bl the binding of polyclonal antisera to CS protein in a radioimmunometric say (*10*). Thus, Zavala *et al.* (*10*) p posed and Dame *et al.* (*7*) confirmed t the repeat domain was the immu dominant region of the CS protein. F different Mab's to the CS protein of *falciparum* recognized synthetic p tides of various lengths corresponding portions of the repeat region (*7*). T immunodominant repeat region mi thus form the basis for a malaria vacc (*7, 11*). That such a vaccine would be widespread use is indicated by the fi ing that the CS gene is highly conser in *P. falciparum* isolates from many g graphic areas (*12*). Here we describe efforts to develop a vaccine against sporozoite stage of *P. falciparum* us proteins containing tandem repeats the CS tetrapeptide sequence produ in *Escherichia coli*.

Expression of the P. falciparum *protein in* E. coli. A recombinant p mid (pUC8 clone 1) containing the I RI insert from λmPf1 (*7*) was the sou of the gene encoding the *P. falcipar* CS protein. This 2337 base pair (fragment contains the entire CS g (Fig. 1A). Since the sequence of the f 16 amino acids of the CS protein characteristic of a cleaved signal pept (*7*), these amino acids are presuma absent from the mature CS protein sporozoites. Restriction endonucle Stu I cleaves the CS gene in the 1 codon of the sequence. Thus, a 1216 Stu I–Rsa I fragment from pUC8 clon (Fig. 1A) should encode all but the f two amino acids of the mature CS p tein predicted from the sequence This fragment was isolated and liga into the λPL *E. coli* expression plasm pAS1 (Fig. 1B) (*13, 14*), which had b cut with Bam HI and treated with D polymerase to create a blunt-end. In resulting plasmid, pCSP, the coding gion of the CS protein is fused, in fra to the translation initiation codon a cent to the Bam HI site in pAS1 (*13,*

James F. Young, Mitchell Gross, and Martin Rosenberg are in the Department of Molecular Genet Smith Kline and French Laboratories, Philadelphia, Pennsylvania 19101. Wayne T. Hockmeyer, W. Rip Ballou, and Carter L. Diggs are in the Department of Immunology and Robert A. Wirtz is in the Departm of Entomology at Walter Reed Army Institute of Research, Washington, D.C. 20307. James H. Trosper Richard L. Beaudoin are in the Malaria Branch, Infectious Diseases Program Center, Naval Med Research Institute, Bethesda, Maryland 20014. Michael R. Hollingdale is at the Biomedical Resea Institute, Rockville, Maryland 20852. Louis H. Miller is in the Laboratory of Parasitic Diseases, Natic Institute of Allergy and Infectious Diseases, Bethesda, Maryland 20205.