Does biased gene conversion influence polymorphism in the circumsporozoite protein-encoding gene of Plasmodium vivax?

Proc. Natl. Acad. Sci. USA
Vol. 85, pp. 8102–8106, November 1988
Genetics

Does biased gene conversion influence polymorphism in the circumsporozoite protein-encoding gene of *Plasmodium vivax*?

(malaria parasite protein-encoding genes/tandemly repeated sequences/immunity to sporozoites)

D. E. ARNOT, J. W. BARNWELL, AND M. J. STEWART

Department of Medical and Molecular Parasitology, New York University School of Medicine, 341 East 25th Street, New York, NY 10010

Communicated by William Trager, June 21, 1988 (received for review May 5, 1988)

ABSTRACT **Variation between North Korean and Latin American isolates in the circumsporozoite (CS) protein encoding gene of the human malaria parasite *Plasmodium vivax* was studied. Polymorphic positions are confined to the central tandemly repeated sequences. Nucleotide substitutions in the tandem repeats produce variants; these substituted positions within the repeat array tend to be conserved between genes. The North Korean CS gene has a short insertion after the repeats encoding a 4-amino acid repeat (Ala-Gly-Gly-Asn) not found in the New World *P. vivax* genes. This sequence is found both flanking and within the tandem repeats of the CS genes of several strains of the Southeast Asian simian malaria parasite, *Plasmodium cynomolgi*. The intraspecific conservation of positions of variants within tandem repeat arrays and the interspecific conservation of probably ancestral repeat motifs at the end of these arrays are consistent with the occurrence of nonreciprocal genetic exchanges between the tandem repeats of these genes. However, a striking asymmetry in strand nucleotide composition within the tandem repeats of all CS genes leads us to suggest that biased correction of heteroduplexes formed during recombination plays a role in the evolution of these genes.**

Mosquito-borne infectious sporozoites of malaria parasites are covered with an immunodominant protein, the circumsporozoite (CS) protein (for review, see ref. 1). In all species of *Plasmodium*, a single copy of the CS gene encodes a protein with a central domain of tandemly repeated sequences flanked by predominantly nonrepeated sequences (2). Within the flanking amino- and carboxyl-terminal domains are two short sequences found in all *Plasmodium* CS genes and more extensive sequences conserved between closely related species, such as *Plasmodium vivax, Plasmodium knowlesi*, and *Plasmodium cynomolgi*. However, the length and sequence of the tandem repeats vary between species and, in some instances, between strains of the same species. The selective forces driving the evolution of CS protein repeats and the molecular mechanisms mediating this process are not well understood.

In the course of studies aimed at delineating the extent and rate of divergence of the CS protein of *P. vivax*, we found a pattern of substitutions in the CS gene of a North Korean (NK) strain of this parasite* that clarifies certain evolutionary relationships and provides some indication as to the genetic mechanism driving the generation of different forms of this gene.

EXPERIMENTAL METHODS

Parasite DNA. The North Korean (NK) strain of *P. vivax* was obtained from infected *Saimiri sciureus* monkeys. The gene was cloned by precipitating a *Bgl* II digest of *P. vivax* bloodstream parasite DNA with 4.5% (vol/vol) polyethelene glycol/1 M NaCl overnight at 37°C. Under these conditions, only molecules >9 kilobases (kb) precipitate. This fraction includes a 14-kb *Bgl* II fragment containing most of the CS-encoding gene that was ligated into the *Bam*HI site of λ EMBL3 and then packaged *in vitro*. The CS gene of the Belem strain of *P. vivax* was used as a hybridization probe to isolate the analogous NK strain gene from this library. Other DNA protocols followed standard techniques (3).

Subcloning and DNA Sequence Analysis. A 3.8-kb *Hpa* II fragment, containing the sequence hybridizing to the probe, was subcloned into the *Acc* I site of pUC19. All sequences were from M13mp18 and mp19 subclones of fragments isolated from the pUC19 subclone and were obtained using the dideoxy chain-termination technique using T7 DNA polymerase, ^{35}S-labeled dATP, and the comparison of dGTP and dITP incorporation in ambiguous regions. Sequences were obtained on both strands of the subclones, except for a short region 3′ to the repeats that was sequenced from two separate subclones. To determine the sequence of the repeats T7 DNA polymerase reactions were carried out with dNTP/ddNTP mixes in which the dNTP concentration was increased to 1 μM. These reactions were separated on a 50-cm electrophoresis apparatus for 16 hr at 1300 V. On these gels, 500–600 base pairs (bp) could be read giving extensive overlaps of the internal repeats of the NK CS gene and permitting direct counting of repeat number and variant positions from both ends of the repeat domain. Restriction enzyme fragment sizes predicted from the analysis of the subclones were confirmed in the parasite DNA by electrophoresing the relevant genomic digests through agarose gels, blotting onto nitrocellulose, and hybridizing to CS gene probes.

RESULTS

CS Gene Sequence. DNA fragments derived from clone pUCNK1 were subcloned into M13mp18 and mp19 for sequence analysis. The sequencing strategy is presented in Fig. 1. The nucleotide and deduced amino acid sequence of the CS gene of the NK strain is shown in Fig. 2 in comparison with the sequences of the Belem (Brazil) and Sal-1 (El Salvador) (4, 5) strains. The NK strain gene contains a *Bgl* II site near the 5′ end of the gene, which excludes from this clone the 5′ end of the gene encoding the first 26 amino acids of the CS protein as compared with the original Belem strain sequence. This site is present in the Belem strain genomic clone, but was uncut, probably because of incomplete digestion of the genomic DNA before ligation into the *Bam*HI sites of λ EMBL3. At every comparable nucleotide position

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: CS, circumsporozoite; NK, North Korean.

*The sequence reported in this paper is being deposited in the EMBL/GenBank data base (IntelliGenetics, Mountain View, CA, and Eur. Mol. Biol. Lab., Heidelberg) (accession no. J04090).

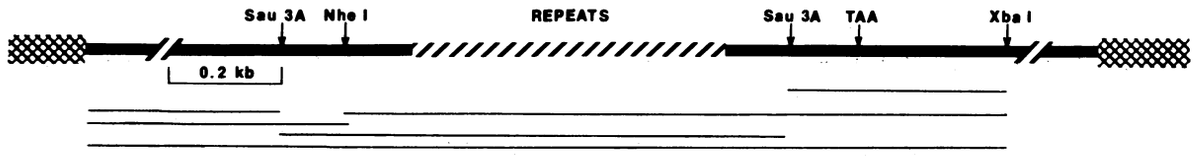

FIG. 1. Restriction map of the CS gene of the NK strain of *P. vivax*. Hatched regions indicate plasmid vector sequence, and each bar represents 1 of 20 repeats. Sites used to construct subclones and the stop codon are indicated to scale. Thin lines indicate M13 subclones used to determine this sequence.

outside of the internal tandem repeats, the NK, Belem and Sal-I strains are identical. This comparison can be extended into the 3′-untranslated regions, which are identical for at least a further 250 nucleotides, except for a cytosine residue insertion in the Belem and Sal-1 sequences 197 nucleotides from the stop codon (data not shown).

Immediately 3′ to the repeated sequences, the NK strain differs sharply from the two New World sequences in having an insertion encoding 16 amino acids (Figs. 2 and 3). This sequence has several interesting features. It has two copies of a repeat with the sequence Ala-Gly-Gly-Asn (AGGN), which is also present at the end of the 3′-terminal repeat of both the NK and Belem strain genes. This Ala-Gly-Gly-Asn repeat is a common feature in the strain-specific tandem repeats of the *P. cynomolgi* CS proteins that are closely related to the *P. vivax* and *P. knowlesi* CS proteins. In fact, the tandem repeats of the NIH and Mulligan strains of *P. cynomolgi* maintain 44 repeats of the sequence Ala-Gly-Gly-

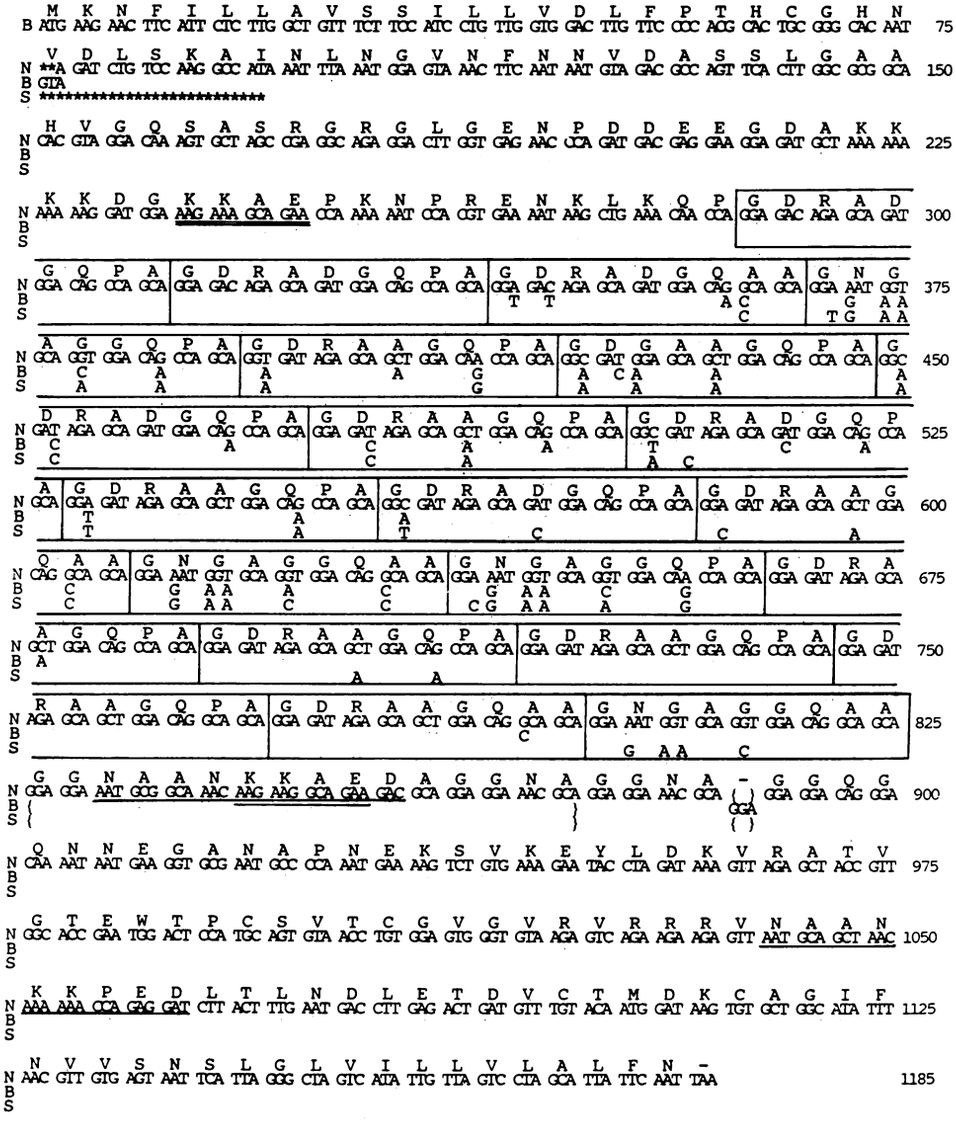

FIG. 2. The gene sequences of three strains of *P. vivax*. A complete gene sequence exists only for the Belem (B) strain (4). This is therefore the only sequence shown for the first 75 nucleotides. NK (N) and Sal-1 (S) sequences commence after the asterisks and, thereafter, the sequences are aligned relative to the NK strain with differences indicated at the relevant positions. The sequences encoding the 9-amino acid repetitive epitopes are boxed. Sequences not present in an isolate are bracketed. The postrepeat variability region found in the NK strain gene at two positions is underlined; a subset of this sequence also found at the 5′-terminus of the repeats is underlined twice. The previously published Belem sequence contains several mistakes—notably a sequence (111–125 in this figure) that was omitted and single-nucleotide substitutions at positions 99, 183, 192, 279, 286, 932, 960, 961, and 967; the Belem sequence in this figure corrects these sequencing errors.

FIG. 3. Comparison of the sequences following the 3′-terminal repeat of nine CS genes from the human–simian group of *Plasmodium*. Sequences are aligned relative to the NK strain gene, conserved nucleotides are in uppercase letters, and substitutions are in lowercase letters. Gombak, Mulligan (M) and NIH (N), London, Ceylon, and Berok are isolates from Malaysia and Sri Lanka of *P. cynomolgi* (6). The H strain of *P. knowlesi* (2) was isolated from an American who had visited Malaysia. One-letter amino acid code is beneath each gene sequence.

Asn (6). Immediately before the Ala-Gly-Gly-Asn repeat of the NK strain gene is a sequence of 9 amino acids, Asn-Ala-Ala-Asn-Lys-Lys-Ala-Glu-Asp (NAANKKAED), which is repeated further into the carboxyl-terminal domain (underlined in Fig. 2) with 8/9 amino acids conserved and 77% nucleotide conservation. A subset of this repeat encoding the sequence Lys-Lys-Ala-Glu (KKAE) is also found preceding the repeats. A further peculiarity of this NK strain-specific insertion is that fragments of this sequence occur in the same position in strains of *P. cynomolgi* and *P. knowlesi*, but are absent in the other *P. vivax* strains (Fig. 3). As the alignment of sequences immediately after the repeats of *P. cynomolgi*, *P. knowlesi*, and *P. vivax* demonstrates, the 16 additional amino acids present in the NK strain have many similarities to the sequences found in this postrepeat variability region in the *P. cynomolgi* and *P. knowlesi* CS genes (2, 6, 7). Many sequences found in this region are subsequences of the repeats found in these genes, notably Ala-Gly-Gly-Asn (AGGN) (*P. cynomolgi* Mulligan, NIH, Berok; *P. knowlesi* H; *P. vivax* NK), Ala-Gly-Gly-Gln (AGGQ) (London; Berok; *P. vivax* NK; and Ala-Gly-Gly-Gly (AGGG) (*P. cynomolgi* Gombak). As noted above, some of the sequences that follow the repeats occasionally precede them. The sequence encoding Gln-Gln-Pro-Pro-Ala-Gly-Gly (QQPPAGG) in the postrepeat variable region of the Gombak strain matches at 17/21 positions a prerepeat sequence encoding Gln-Pro-Pro-Pro-Ala-Asp-Gly (QPPPADG) of the London strain (6).

The Central Tandem Repeats. The nucleotide sequence comparison of the three strains of *P. vivax* presented in Fig. 2 demonstrates the accelerated rate of accumulation of mutations in the repeats. Excluding the postrepeat variable region, essentially all of the divergence between these genes occurs within the tandem repeats. In Fig. 4, the repeats of these strains are contrasted. The number of positions at which the tandem repeat arrays differ from each other is given in Table 1. The alignment illustrates that the DNA sequence of the repeats within a gene is well conserved. Less obvious relationships also become apparent. Within a gene the most divergent repeats are at opposite ends of the array, whereas adjacent repeats tend to be more similar. Variant repeats are found in approximately the same relative position within their respective arrays; this is most apparent with the terminal variants, where the 20th variant of the Belem alignment occurs as variant 20, 14, 13, and 4 in the NK repeats. At the 5′ end of the array, variants 1 and 2 of NK occur at positions 1, 2, and 6 of the Belem and 1, 2, 3, 7, 8, and 9 of the *Sal*-1 repeats. The most common variant in these strains occurs as NK variant 8, 10, 15, 16, 17, and 18; Belem variant 12, 16, 17, 18, and 19; and El Salvador variant 5, 13, 15, 17, and 18—again conforming to the tendency of different

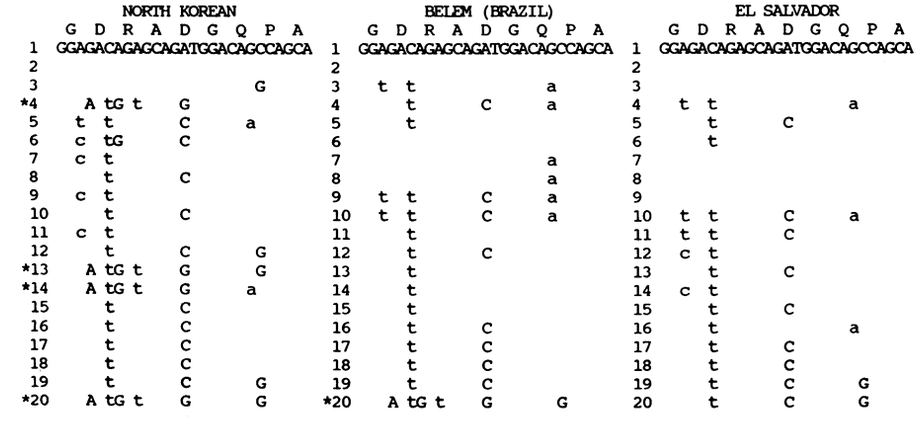

FIG. 4. The nucleotide sequences encoding the repetitive epitopes of *P. vivax* CS proteins. Sequences are given 5′–3′ from the first base of the first tandem repeat to the last codon in the last repeat. Synonymous substitutions are in lowercase letters; nonsynonymous substitutions are in uppercase letters. Repeats are numbered, and the most divergent variants are marked with asterisks.

Table 1. Number of nucleotide positions differing between the repeat arrays of the different strains

	Strains		
	NK	Belem	Sal-1
Strains			
NK	(38)	37	40
Belem		(30)	27
Sal-1			(27)

Positions within individual repeat domains diverging from a hypothetical homogenous array consisting of only the most common variant are shown in parenthesis. The most common variant is exemplified by NK repeat 10, Belem repeat 12, and Sal-1 repeat 13.

genes to maintain the same relative position of variants within an array of tandem repeats. It is notable that even in the strain with the least variation within the repeats, the Sal-I strain, mutations that change the amino acid sequence are as common as synonymous substitutions.

Nucleotide Composition Biases in the CS Gene Tandem Repeats. A frequently noted feature of all CS genes is that the use of alternative synonymous codons is not random (5, 8). For example, CCT is not used as one of the four possible proline codons in *P. vivax* CS genes, and both alanine (CGN) and glycine (GGN) codons have a strong bias towards adenine in the third position. However, this apparent codon usage bias is probably mainly a derivative effect of the maintenance of DNA sequence homogeneity in the tandem repeats. CS gene codon usage is inevitably influenced by the presence of many copies of short, well-repeated sequences. An alternative description of the situation in *P. vivax* CS genes is that the mRNA-like strand encoding the repeats has a low thymidine composition. Furthermore, this discrimination against thymidine residues on one strand is common in all CS genes and is independent of the size or amino acid composition of the repeat. The bias against thymidine residues appears limited to the repeats themselves and regions immediately adjacent to them, particularly at the 5'-end of the array. The strand composition biases of four CS genes are plotted in Fig. 5. The genes indicate different ends of the range, *Plasmodium falciparum* having comparatively thymidine-rich repeats. The outstanding example of this phenomenon is the mRNA-like strand of the CS gene of the *P. knowlesi* Nuri strain (7). This strand has one thymidine residue in the 432 nucleotides encoding 12 tandem repeats—the sole thymidine residue occurring in the 3' terminal repeat. The graph clearly illustrates that the strand composition bias is limited to the repeats and their flanking sequences, particularly those immediately 5' of the repetitive region. The "variant" repeats—that is, those differing from the majority consensus sequence, and the "thymidine-rich" repeats, are almost always the same thing and exhibit the tendencies that have been noted above. This is particularly clear in the case of the *P. falciparum* gene, where the thymidine content of the mRNA-like strand in the repeats forms a symmetrical distribution with the "thymidine-rich" variants encoding the terminal repeats.

DISCUSSION

Essentially all polymorphisms so far found in *P. vivax* CS proteins are associated with the tandemly repeated units and the sequences immediately flanking this region. This finding contrasts with the additional variation found in the nonrepeated region of the *P. falciparum* CS sequence that has been proposed to occur in functional domains, such as a T-cell recognition epitope, and to be maintained by selection (8).

Mutations in the tandem repeats of *P. vivax* CS genes appear to be spread between units, giving rise to a shifting

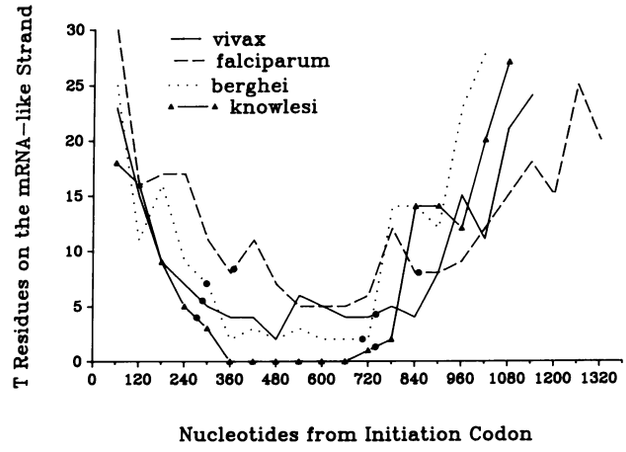

FIG. 5. Graph plotting thymidine residues per 60 nucleotides on the mRNA-like strand of four CS genes. *P. vivax* and *P. knowlesi* CS genes are similar outside the tandem repeats, although there is less extensive conservation of nucleotide sequence between these isolates and the *P. berghei* or *P. falciparum* CS genes. The first and last blocks plotted for each gene contain the initiation and termination codons. The black circles on each line indicate the boundaries of the tandemly repeated sequences. Thymidine composition of the mRNA-like strand of these isolates in their amino-terminal domain, tandem repeats, and carboxyl-terminal domain, respectively, are as follows: *P. vivax* (Belem) (4) 26%, 7%, and 25%; *P. berghei* (NK65) (9) 25%, 3%, and 30%; *P. falciparum* (Thai T4) (10) 27%, 10%, and 14%; and *P. knowlesi* (Nuri) (7) 17%, 0.2%, and 24%.

pattern of variation between the repeat arrays of individual isolates. Because meiotic recombination is probably responsible for this diffusion of mutations, the process will tend to maintain the homogeneity of a tandemly repeated sequence between individuals of the same species while leading to rapid divergence between homologous sequences in related species. This homogenizing process was originally observed in satellite DNA sequences, where it has been referred to as concerted or parallel evolution (11). Although occurring within a small single-copy gene, the concerted differentiation of CS genes must also be the result of spreading mutations through tandem repeats by exchange of sequence between variants.

Two recombinational processes can homogenize tandemly repeated DNA sequences: unequal crossing over after misalignment of repeats and gene conversion. Conversion involves the correction of mismatches in favor of one of the input strands after misalignment of repeats in hybrid DNA molecules formed during the interaction between chromatids before recombination. Such mechanisms have been considered to account for the observed polymorphisms in the CS (11) and other *Plasmodium* genes, particularly those encoding the *P. falciparum* S-antigens (12). These exchanges probably also explain our observation that the sequences flanking the repeats of the *P. vivax, P. cynomolgi*, and *P. knowlesi* CS genes are somewhat repetitive, occasionally conserved between species, and sometimes found at both ends of the tandem repeats. We propose that these sequences are the remnants of ancestral repeats that are rarely or never involved in the homology-dependent exchanges that maintain the present tandem repeat arrays.

Other than their restricted amino acid *compositions* (glycine-, alanine-, proline-, and asparagine-rich sequences predominate), CS protein *sequences* appear to have few constraints (6). It has been noted that purifying phenotypic selection that acts to maintain the amino acid sequence cannot explain both the conservation of CS protein repeats within a species and their rapid evolution between species (6, 13). Therefore the mechanism maintaining the repeats must

act at the DNA level (14). However, if unequal crossing over or unbiased gene conversion operate on these sequences, nucleotide substitutions (few of which are probably deleterious under these circumstances) should have the same chance of being eliminated or fixed. This does not explain how the pronounced asymmetries in strand nucleotide composition would arise. It seems unlikely that several *Plasmodium* species, the genomic base composition of which are A+T rich [*P. vivax* and *P. knowlesi* are ≈65–70% A+T; *P. falciparum* and *P. berghei* are ≈82% A+T (15)] and whose codon usage is correspondingly biased toward A+T-rich codons, should have amplified a unit with a fortuitously low-thymidine content on the mRNA-like strand.

A possible explanation for the discrimination against thymidine residues on the mRNA-like strand is that the strand bias results from the correction of heteroduplex DNA formed during recombination between tandem repeats. In biased gene conversion events the correction of heterozygosity can have a preferred direction, and the probability of a mutation being spread or corrected in favor of the currently fixed variant is not random but reflects the way in which heteroduplex repair operates in *Plasmodium*. Such a process implies that the CS tandem repeats and the regions flanking them constitute a conversion domain, which may also explain the tendency of the sequences immediately flanking CS repeats to manifest asymmetries in strand-nucleotide composition. The genetic mechanism causing the proposed conversion bias is unknown, although it probably does not simply remove mismatched thymidine residues, as such removal would ultimately deplete adenosine from the tandem repeats. Discrimination against thymidine on one strand may be the eventual result of cumulative slight imbalances, such as a combination of asymmetric recombination events involving only one strand of each sister chromatid (16) with a preferential correction of mismatches of the thymidine–thymidine type to the Watson–Crick base pair. An interesting precedent for the phenomenon of strand imbalances in base composition is the report that the tandemly repeated sequences within the gene encoding the human keratinocyte protein involucrin also display a marked paucity of thymidine residues on the mRNA-like strand (17).

Disparity in the direction of gene conversion is frequently found in yeast and fungi, where the correction of heteroduplex DNA is known to be influenced by the structure of the mismatch as well as operating by different rules in different experimental models (18). Gene conversion biases have also been quantitatively demonstrated to have significant effects on the evolution of tandemly repeated sequences (19). We envision that biases in gene conversion can "drive" (20) the homogenization of CS gene tandem repeats in such a way as to ensure their maintenance and concerted differentiation, even though variant repeats that have a conversion advantage have no selective advantage.

The *P. vivax* CS protein is currently being tested as an antimalaria vaccine candidate (21). A potential problem is whether vaccination-induced immune responses might select sporozoites with CS protein variants to which a vaccinated population could not respond. We do not know if the polymorphisms seen in these CS proteins represent the range of possible *P. vivax* CS proteins or if more drastic variation exists and could be selected. It would be interesting to know allelic frequencies in regions in which malaria transmission exists and to what extent these frequencies change during endemic and epidemic malaria.

We thank Ruth Nussenzweig and Jerry Vanderberg for much advice on malariology. D.E.A. thanks Enzo Enea for many discussions on gene evolution. We thank David Keeney for excellent technical assistance and Mary-Anne Griffith for typing the manuscript. This work was funded by the Agency for International Development (DPE0453-A-00-5012-00), the National Institutes of Health (Public Health Services grant 5P01-A121642) and a Whitehead Foundation Fellowship to D.E.A.

1. Nussenzweig, V. & Nussenzweig, R. S. (1985) *Cell* **42,** 401–403.
2. Ozaki, L. S., Svec, P., Nussenzweig, R. S., Nussenzweig, V. & Godson, G. N. (1983) *Cell* **34,** 815–822.
3. Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Lab., Cold Spring Harbor, NY).
4. Arnot, D. E., Barnwell, J. W., Tam, J. P., Nussenzweig, V., Nussenzweig, R. S. & Enea, V. (1985) *Science* **230,** 815–818.
5. de la Cruz, V. F., Lal, A. A., Welsh, J. A. & McCutchan, T. F. (1987) *J. Biol. Chem.* **262,** 6464–6467.
6. Galinski, M. R., Arnot, D. E., Cochrane, A. H., Barnwell, J. W., Nussenzweig, R. S. & Enea, V. (1987) *Cell* **48,** 311–319.
7. Sharma, S., Svec, P., Mitchell, G. H. & Godson, G. N. (1985) *Science* **229,** 779–782.
8. de la Cruz, V. F., Lal, A. A. & McCutchan, T. F. (1987) *J. Biol. Chem.* **262,** 11935–11939.
9. Eichinger, D. J., Arnot, D. E., Tam, J. P., Nussenzweig, V. & Enea, V. (1986) *Mol. Cell. Biol.* **6,** 3965–3972.
10. del Portillo, H. A., Nussenzweig, R. S. & Enea, V. (1987) *Mol. Biochem. Parasitol.* **24,** 289–294.
11. Brutlag, D. L. (1980) *Annu. Rev. Genet.* **14,** 121–144.
12. Cowman, A. F., Saint, R. B., Coppel, R. L., Brown, G. V., Anders, R. F. & Kemp, D. J. (1985) *Cell* **40,** 775–783.
13. Enea, V., Galinski, M., Schmidt, E., Gwadz, R. & Nussenzweig, R. S. (1986) *J. Mol. Biol.* **188,** 721–726.
14. Enea, V. & Arnot, D. E. (1988) in *Current Communications in Molecular Biology: The Molecular Genetics of Parasitic Protozoa*, eds. Turner, M. J. & Arnot, D. E. (Cold Spring Harbor Lab., Cold Spring Harbor, NY), pp. 5–12.
15. McCutchan, T. F., Dame, J. B., Miller, L. H. & Barnwell, J. (1984) *Science* **225,** 808–811.
16. Meselson, M. S. & Radding, C. M. (1975) *Proc. Natl. Acad. Sci. USA* **72,** 358–361.
17. Eckert, R. L. & Green, H. (1986) *Cell* **46,** 583–589.
18. Fogel, S., Mortimer, R., Lusnak, K. & Tavares, F. (1978) *Cold Spring Harbor Symp. Quant. Biol.* **43,** 1325–1341.
19. Nagylaki, T. & Petes, T. D. (1982) *Genetics* **100,** 315–337.
20. Dover, G., Brown, S., Coen, E., Dallas, J., Strachan, T. & Trick, M. (1982) in *Genome Evolution*, eds. Dover, G. A. & Flavell, R. B. (Academic, London), pp. 343–372.
21. Barr, P. J., Gibson, H. L., Enea, V., Arnot, D. E., Hollingdale, M. R. & Nussenzweig, V. (1987) *J. Exp. Med.* **165,** 1160–1171.