Proc. Natl. Acad. Sci. USA
Vol. 88, pp. 1751–1755, March 1991
Immunology

Structural diversity in the *Plasmodium falciparum* merozoite surface antigen 2

(malaria/protein polymorphisms/repeat sequences)

JASON A. SMYTHE*, ROSS L. COPPEL*, KAREN P. DAY*, RODGER K. MARTIN†, AYOADE M. J. ODUOLA‡, DAVID J. KEMP*, AND ROBIN F. ANDERS*§

*The Walter and Eliza Hall Institute of Medical Research, Melbourne, Victoria 3050, Australia; †Walter Reed Army Institute of Research, Washington, DC 20307; and ‡University of Ibadan, Ibadan, Nigeria

Communicated by Gustav J. V. Nossal, October 29, 1990 (received for review July 11, 1990)

ABSTRACT **Antigens associated with the surface of merozoites of the malaria parasite *Plasmodium falciparum* are directly accessible to immune attack and therefore are prime vaccine candidates. We have previously shown that one of the two known merozoite surface antigens (merozoite surface antigen 2; MSA-2) exhibits considerable sequence and antigenic diversity in different isolates. The sequences of MSA-2 from three isolates revealed a central domain composed of repeats that vary in number, length, and sequence, flanked in turn by nonrepetitive variable sequences and by conserved N- and C-terminal domains. We report here the sequences of a further four MSA-2 alleles, containing repetitive sequences that are related but not identical to each other. The seven alleles of MSA-2 can be divided into two distinct allele families on the basis of nonrepetitive sequences. Hybridization studies with repeat probes indicated that all of the 44 *P. falciparum* isolates examined contained repeat regions similar to those defined in known MSA-2 sequences.**

Two well-defined integral membrane proteins are associated with the surface of *Plasmodium falciparum* merozoites—namely, merozoite surface antigen 1 (MSA-1), previously referred to as P195, gp185, PSA, or PMMSA (1), and merozoite surface antigen 2 (MSA-2) (2–6). MSA-1 is a glycoprotein of M_r ≈195,000 that is processed at the time of schizont rupture to generate several smaller polypeptides on the mature merozoite surface (1). MSA-2 is a smaller polypeptide of M_r ≈45,000 that is not processed during parasite maturation but, like MSA-1, appears to be anchored in the merozoite membrane by a glycosylphosphatidylinositol moiety (2). There are many antigenically diverse forms of MSA-1 and MSA-2 (1–9). Much of the diversity in MSA-1 has arisen from limited intragenic recombination between dimorphic alleles (7–9). Repetitive sequences, antigenically prominent in many *P. falciparum* proteins (10), are only a minor structural feature of MSA-1.

Studies of the genes from three isolates of *P. falciparum* (2, 3) revealed that MSA-2 contains highly conserved and variable sequences and that repetitive sequences form a major component of each allele. The MSA-2 of isolate FCQ27/PNG (FC27) contains two copies of a 32-amino acid repeat, whereas the MSA-2 of Indochina 1 (IC1) and 3D7 contain multiple copies of a 4-amino acid sequence with no apparent relationship to the FC27 repeat. The sequences immediately flanking the repeats are also variant among these three alleles but are flanked by highly conserved N- and C-terminal domains that are considerably longer than the signal and anchor sequences. Oligonucleotide probes corresponding to sequences within the 4- and 32-amino acid repeats of MSA-2 hybridized with genomic DNA from 12 of 15 parasite isolates tested (3). This suggested that some isolates of *P. falciparum* express a form of MSA-2 with yet other types of repeat sequences.

We have now determined the sequences of four additional MSA-2 alleles from isolates IMR143, MAD71, K1, and Nigeria 32 (NIG32).¶ The new alleles contain repetitive sequences related but not identical to the previously reported MSA-2 sequences. The seven alleles of MSA-2 we have now sequenced can be divided into two distinct allele families on the basis of the nonrepetitive variable sequences. Each of these allele families has characteristic repeat types, one or more of which reacted with all of 44 field and cultured *P. falciparum* isolates examined by dot-blot hybridization.

MATERIALS AND METHODS

Parasites. The KF *P. falciparum* isolates were obtained from children aged 2–13 years presenting with acute symptoms of malaria in Madang, Papua New Guinea (11). The origins of all other isolates have been reported elsewhere (12, 13). Primary isolates were cultured as described (3) and DNA was prepared from asynchronously cultured parasites (3).

Polymerase Chain Reaction (PCR) and Sequencing. Oligonucleotides used as primers for the PCR (14) corresponded to the first and last 30 nucleotides of the translated FC27 MSA-2 gene sequence (nucleotides 94–123 and 859–888 in ref. 2) and the gene was amplified through 25 PCR cycles. The PCR products were gel purified, ligated into M13 vectors, and sequenced by the chain-termination method.

DNA Dot-Blot Hybridization. Samples of DNA amplified by the PCR were dot-blotted onto nitrocellulose and hybridized with oligonucleotide probes (3). The 5′ oligonucleotide (50 pM) probes were end-labeled by incubation with polynucleotide kinase (Boehringer Mannheim) in the presence of 150 μCi of [γ-^{32}P]ATP (>5000 Ci/mmol; 1 Ci = 37 GBq; Amersham). Probes used in the dot-blot hybridization were as follows: probe A, FC27 repeat-specific probe (ATCACAAACTACTACTC); probe B, 3D7/IC1 repeat-specific probe (GGTGGTAGTGCTGGTGGTAG); and probe C, IMR143 and MAD71 8-amino acid repeat-specific probe (CAGAACCAGCAACA).

RESULTS

Sequence of Additional Major Repeat Types in MSA-2. Eight of 44 isolates tested by DNA dot-blot hybridization

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: MSA-1, merozoite surface antigen 1; MSA-2, merozoite surface antigen 2; PCR, polymerase chain reaction.

§To whom reprint requests should be addressed at: The Walter and Eliza Hall Institute of Medical Research, P.O. Royal Melbourne Hospital, Melbourne, Victoria 3050, Australia.

¶The sequences reported in this paper have been deposited in the GenBank data base (accession nos. M59765, M59766, M59767, and M59768).

(Table 1) failed to react with probes to the previously described 4- or 32-amino acid MSA-2 repeats (2, 3). The MSA-2 genes from three of these isolates (IMR143, MAD71, and NIG32) were amplified by PCR, and the nucleotide sequences were determined. The MSA-2 sequences for two of these isolates are presented in Fig. 1 and the amino acid sequences are aligned with other MSA-2 sequences in Fig. 2. The IMR143 and MAD71 MSA-2 genes are 903 bp in length

Table 1. MSA-2 repeat types in *P. falciparum* isolates

Isolate	Origin	Repeat probe* A	B	C
KF1703	PNG	+	−	−
KF1776	PNG	+	−	−
KF1787	PNG	+	−	−
KF1898	PNG	+	−	−
KF1917	PNG	+	−	−
KF1918	PNG	+	−	−
KF1925	PNG	+	−	−
KF1933	PNG	+	−	−
KF1935	PNG	+	−	−
FC27	PNG	+	−	−
NIG14	Africa	+	−	−
NIG16	Africa	+	−	−
NIG61	Africa	+	−	−
NIG63	Africa	+	−	−
D6†	Africa	+	−	−
V1†	S.E. Asia	+	−	−
K1	S.E. Asia	+	−	−
HB3†	S. America	+	−	−
KF1905	PNG	−	+	−
KF1934	PNG	−	+	−
IMR144	PNG	−	+	−
Palo Alto	Africa	−	+	−
NF7	Africa	−	+	−
NIG2	Africa	−	+	−
NIG55	Africa	−	+	−
NIG59	Africa	−	+	−
FCS	Africa	−	+	−
W2†	S.E. Asia	−	+	−
Viet Smith	S.E. Asia	−	+	−
CSL2	S.E. Asia	−	+	−
IC1	S.E. Asia	−	+	−
ItG2†	S. America	−	+	−
3D7†	Unknown	−	+	−
KF1775	PNG	−	−	+
KF1785	PNG	−	−	+
KF1904	PNG	−	−	+
KF1930	PNG	−	−	+
IMR143	PNG	−	−	+
IMR147	PNG	−	−	+
MAD71	PNG	−	−	+
NIG32‡	Africa	−	−	−
KF1916§	PNG	−	+	+
311†§	S. America	−	+	+
KF1931§	PNG	+	+	+

PNG, Papua New Guinea.

*The sequences of the hybridization probes for the FC27 32-amino acid repeat (A), the IC1 4-amino acid repeat (B), and the MAD71 8-amino acid repeat (C) are given in the text.

†Cloned parasite line.

‡The MSA-2 of NIG32 failed to hybridize with probe A, B, or C and, as discussed in the text, was shown by sequencing to contain a different repeat sequence.

§Isolates KF1916 and KF1931 hybridizing with more than one probe may contain mixtures of parasites with different MSA-2 alleles, but recent results indicate that MSA-2 of the cloned line 311 has a repeat sequence that hybridizes with the 4- and 8-amino acid repeat-specific probes (B and C, respectively).

and differ from each other by only two point mutations: a first base mutation at base 367 results in the substitution of an S in IMR143 for a G in MAD71; the second mutation, a silent third base mutation, occurs at nucleotide 501. In contrast, the NIG32 MSA-2 gene is 1044 base pairs in length and has a central repeat region that differs significantly from the IMR143 and MAD71 sequences. Differences between these sequences and those of the known MSA-2 allele are shown schematically in Fig. 3*A*. In contrast to the FC27 (32-amino acid repeat) or the 3D7 and IC1 (4-amino acid repeat), MSA-2 alleles IMR143 and MAD71 contain an 8-amino acid repeat sequence. The nucleotide sequence of the repeat, which commences at base 151, is GGT GCT GTT GCT GGT TCT GGT GCT and encodes the sequence GAVAGSGA. In addition to six complete copies of this repeat sequence, the IMR143 and MAD71 genes contain two incomplete copies of the repeat (Figs. 1 and 2). The incomplete copies are the second and the last and both encode repeats lacking two of the eight amino acids (GAGSGA and VAGSGA, respectively).

Of the six isolates in addition to IMR143 and MAD71 that failed to hybridize with probes for the 4- or 32-amino acid repeats all, except NIG32, hybridized with a probe for the 8-amino acid repeat found in IMR143 and MAD71 (Table 1). The gene sequence for the NIG32 MSA-2 predicts a protein containing multiple copies of related 4- and 8-amino acid repeat sequences that are different from the IMR143/MAD71 and previously described MSA-2 repeats. The NIG32 repeat sequences, commencing at base 157, encode four tandem copies of the 8-amino acid sequence GAVASAGN (encoded by the sequence GGT GCT GTT GCT AGT GCT GGT AAT) followed by 15 copies of the 4-amino acid sequence GAGN (encoded by the sequence GGT GCT GGT AAT), then two more copies of the 8-amino acid repeat GAVASAGN separated by GAGN, and finally the sequence GAVA, which is half of the 8-amino acid repeat.

The MSA-2 Genes of Isolates K1 and FC27 Are Related. Previously, by using dot-blot hybridization, we established that the MSA-2 of isolate K1 contained a sequence related to the 32-amino acid repeat in FC27. However, the K1 MSA-2 reacted poorly with some antibodies to the FC27 MSA-2 polypeptide and for this reason the K1 gene was amplified by PCR and then sequenced. The K1 MSA-2 gene is 843 nucleotides in length and, as expected, is more closely related to the MSA-2 gene of FC27 than to the MSA-2 of 3D7/IC1, MAD71/IMR143, or NIG32. Nevertheless, there are some notable differences between the FC27 and K1 sequences. Whereas the FC27 antigen contains two copies of the 32-amino acid repeat, this sequence occurs only once in the K1 MSA-2 and has a third base mutation at nucleotide 207 that substitutes an S for the R in the FC27 sequence. At the C-terminal end of the 32-residue sequence in K1 there is a 12-amino acid sequence, ESNSRSPPITTT, which repeats in tandem five times (encoded by nucleotides 295–474). There is a closely related sequence (ESISPSPPITTT) found at the C-terminal end of the second 32-residue repeat in FC27 (residues 131–142). This FC27 sequence occurs only once and is encoded by a nucleotide sequence that differs from the K1 repeat sequence by second base mutations in the third and fifth codons, both of which result in amino acid differences (Fig. 3*C*).

The MSA-2 Alleles Contain Conserved N- and C-Terminal Domains. The seven known MSA-2 protein sequences are aligned in Fig. 2. All seven sequences are conserved at the N-terminal end for the first 43 residues and throughout the last 74 residues with the exception of an N to S change at residue 237 in the FC27 antigen (Fig. 2) resulting from a second base mutation in the corresponding codon. As the oligonucleotides used for priming in the PCR corresponded to sequences within the N-terminal signal sequence and the

```
MAD71
 M K V I K T L S I I N F F I F V T F N I K N E S K Y S N T F I N N A Y N M S I R      40
ATGAAGGTAATTAAAACATTGTCTATTATAAATTTCTTTATTTTTGTTACCTTTAATATTAAAAATGAAAGTAAATATAGCAACACATTCATAAACAATGCTTATAATATGAGTATAAGG     120
 R S M E E S K P P T G A V A G S G A G A G S G A G A V A G S G A G A V A G S G A      80
AGAAGTATGGAAGAAAGTAAGCCTCCTACTGGTGCTGTTGCTGGTTCTGGTGCTGGTGCTGGTTCTGGTGCTGGTGCTGTTGCTGGTTCTGGTGCT     240
 G A V A G S G A G A V A G S G A G A V A G S G A V A G S G A G N G A N P G A D A      120
GGTGCTGTTGCTGGTTCTGGTGCTGGTGCTGTTGCTGGTTCTGGTGCTGGTGCTGTTGCTGGTTCTGGTGCTGGTAATGGTGCTAATCCTGGTGCAGATGCT     360
 E R G P S T P A T T T T T T T T N D A E A S T S T S S E N R N H N N A E T N P K      160
GAGAGAGGTCCAAGTACTCCCGCTACTACCACAACTACCACAACTACTAATGATGCAGAAGCATCTACCAGTACCTCTTCAGAAAATCGAAATCATAATAATGCCGAAACAAATCCAAAA     480
 G K G E V Q K P N Q A N K E T Q N N S N V Q Q D S Q T K S N V P R T Q D A D T K      200
GGTAAAGGAGAAGTTCAAAAACCAAATCAAGCAAATAAAGAAACTCAAAATAACTCAAATGTTCAACAAGACTCTCAAACTAAATCAAATGTTCCACGCACTCAAGATGCAGACACTAAA     600
 S P T A Q P E Q A E N S A P T A E Q T E S P E L Q S A P E N K G T G Q H G H M H      240
AGTCCTACTGCACAACCTGAACAAGCTGAAAATTCTGCTCCAACAGCCGAACAAACTGAATCCCCCGAATTACAATCTGCACCAGAGAATAAAGGTACAGGACAACATGGACATATGCAT     720
 G S R N N H P Q N T S D S Q K E C T D G N K E N C G A A T S L L N N S S N I A S      280
GGTTCTAGAAATAATCATCCACAAAATACTTCTGATAGTCAAAAAGAATGTACCGATGGTAACAAAGAAAACTGTGGAGCAGCAACATCCCTCTTAAATAACTCTAGTAATATTGCTTCA     840
 I N K F V V L I S A T L V L S F A I F I *                                      300
ATAAATAAATTTGTTGTTTTAATTTCAGCAACACTTGTTTTATCTTTTGCCATATTCATATAA                                                          903

NIG32
 M K V I K T L S I I N F F I F V T F N I K N E S K Y S N T F I N N A Y N M S I R      40
ATGAAGGTAATTAAAACATTGTCTATTATAAATTTCTTTATTTTTGTTACCTTTAATATTAAAAATGAAAGTAAATATAGCAACACATTCATAAACAATGCTTATAATATGAGTATAAGG     120
 R S M A E S K P P T G D G A V A S A G N G A V A S A G N G A V A S A G N G A V A      80
AGAAGTATGGCAGAAAGTAAGCCTCCTACTGGTGATGGTGCTGTTGCTAGTGCTGGTAATGGTGCTGTTGCTAGTGCCGGTAATGGTGCTGTTGCT     240
 S A G N G A G N G A G N G A G N G A G N G A G N G A G N G A G N G A G N G A G N      120
AGTGCTGGTAATGGTGCTGGTAATGGTGCTGGTAATGGTGCTGGTAATGGTGCTGGTAATGGTGCTGGTAATGGTGCTGGTAATGGTGCTGGTAATGGTGCTGGTAAT     360
 G A G N G A G N G A G N G A G N G A G N G A G N G A V A S A G N G A G N G A V A      160
GGTGCTGGTAATGGTGCTGGTAATGGTGCTGGTAATGGTGCTGGTAATGGTGCTGGTAATGGTGCTGGTAATGGTGCCGTTGCTAGTGCTGGTAATGGTGCTGGTAATGGTGCTGTTGCT     480
 S A G N G A V A E R S S S T P A T T T T T T T T N D A E A S T S T S S E N S N H      200
AGTGCTGGTAATGGTGCTGTTGCTGAGAGAAGTTCAAGTACTCCCGCTACTACCACAACTACCACAACTACTAATGATGCAGAAGCATCTACCAGTACCTCTTCAGAAAATTCAAATCAT     600
 N N A E T N P K G N G E V Q P N Q A N K E T Q N N S N V Q Q D S Q T K S N V P R      240
AATAATGCCGAAACAAATCCAAAAGGTAATGGAGAAGTTCAACCAAATCAAGCAAATAAAGAAACTCAAAATAACTCAAATGTTCAACAAGACTCTCAAACTAAATCAAATGTTCCACGC     720
 T Q D A D T K S P T A Q P E Q A E N S A P T A E Q T E S P E L Q S A P E N K G T      280
ACTCAAGATGCAGACACTAAAAGTCCTACTGCACAACCTGAACAAGCTGAAAATTCTGCTCCAACAGCCGAACAAACTGAATCCCCCGAATTACAATCTGCACCAGAGAATAAAGGTACA     840
 G Q H G H M H G S R N N H P Q N T S D S Q K E C T D G N K E N C G A A T S L L N      320
GGACAACATGGACATATGCATGGTTCTAGAAATAATCATCCACAAAATACTTCTGATAGTCAAAAAGAATGTACCGATGGTAACAAAGAAAACTGTGGAGCAGCAACATCCCTCTTAAAT     960
 N S S N I A S I N K F V V L I S A T L V L S F A I F I *                        347
AACTCTAGTAATATTGCTTCAATAAATAAATTTGTTGTTTTAATTTCAGCAACACTTGTTTTATCTTTTGCCATATTCATATAA                                    1014

K1
 M K V I K T L S I I N F F I F V T F N I K N E S K Y S N T F I N N A Y N M S I R      40
ATGAAGGTAATTAAAACATTGTCTATTATAAATTTCTTTATTTTTGTTACCTTTAATATTAAAAATGAAAGTAAATATAGCAACACATTCATAAACAATGCTTATAATATGAGTATAAGG     120
 R S M A N E G S N T K S V G A N A P K A D T I A S G S Q S S T N S A S T S T T N      80
AGAAGTATGGCAAATGAAGGTTCTAATACTAAGAGTGTAGGTGCAAATGCTCCAAAAGCTGATACTATTGCTAGTGGAAGTCAAAGTAGTACAAATAGTGCAAGTACTAGTACTACTAAT     240
 N G E S Q T T T P T A A D T P T A T E S N S R S P P I T T T E S N S R S P P I T      120
AATGGAGAATCACAAACTACTACTCCTACCGCTGCTGATACCCCTACTGCTACAGAAAGTAATTCACGTTCACCACCCATCACTACTACAGAAAGTAATTCACGTTCACCACCCATCACT     360
 T T E S N S R S P P I T T T E S N S R S P P I T T T E S N S R S P P I T T T E S      160
ACTACAGAAAGTAATTCACGTTCACCACCCATCACTACTACAGAAAGTAATTCACGTTCACCACCCATCACTACTACAGAAAGTAATTCACGTTCACCACCCATCACTACTACAGAAAGT     480
 S S S G N A P N K T D G K G E E S E K Q N E L N E S T E E G P K A P Q E P Q T A      200
TCAAGTTCTGGCAATGCACCAAATAAAACAGACGGTAAAGGAGAAGAGAGTGAAAAACAAAATGAATTAAATGAATCAACTGAAGAAGGACCCAAAGCTCCACAAGAACCTCAAACGGCA     600
 E N E N P A A P E N K G T G Q H G H M H G S R N N H P Q N T S D S Q K E C T D G      240
GAAAATGAAAATCCTGCTGCACCAGAGAATAAAGGTACAGGACAACATGGACATATGCATGGTTCTAGAAATAATCATCCACAAAATACTTCTGATAGTCAAAAAGAATGTACCGATGGT     720
 N K E N C G A A T S L L N N S S N I A S I N K F V V L I S A T L V L S F A I F I *  280
AACAAAGAAAACTGTGGAGCAGCAACATCCCTCTTAAATAACTCTAGTAATATTGCTTCAATAAATAAATTTGTTGTTTTAATTTCAGCAACACTTGTTTTATCTTTTGCCATATTCATATAA 843
```

FIG. 1. Nucleotide and amino acid sequences of MSA-2 alleles from the MAD71, NIG32, and K1 isolates of *P. falciparum*. All nucleotide position references in the text are taken from this figure. The IMR143 sequence is identical to that of MAD71 except for two mutations shown in bold. At position 367, G is replaced by A in IMR143 and at position 501, A is replaced by G.

C-terminal hydrophobic sequence we do not know if there is any sequence diversity in these regions that would be lacking from the mature protein.

The Variable Regions of MSA-2 Define Two Families of Alleles. The conserved N- and C-terminal domains flank sequences that show considerable diversity among the different MSA-2 alleles. This central variable region includes the sequence repeats discussed above but also nonrepetitive sequences flanking the repeats. A comparison of the repetitive and nonrepetitive sequences indicates that the IC1, IMR143, MAD71, 3D7, and NIG32 MSA-2 alleles are more closely related to each other than to the FC27 and K1 alleles, which, in turn, are closely related. All of the 4- and 8-amino acid repeats found in the former group of alleles have similar amino acid compositions with only five different amino acids being used. Furthermore, there is a dipeptide periodicity within these repeats that reflects a hexanucleotide repeat underlying the higher order repeats in these MSA-2 genes. All of these repeats are encoded by a series of hexanucleotides, which, with one exception, are related to the sequence GGT GCT by a single nonsynonymous mutation (Fig. 3*B*). The one exception, which differs in two nucleotide positions, is the hexanucleotide sequence GGT AAT, encoding the dipeptide GN. The hexanucleotide repeat sequence is only found within higher order repeats, most commonly 12 or 24 nucleotides in length. The IMR143 and MAD71 genes, however, contain two different 18-nucleotide repeats that differ from the 24-nucleotide repeats in that they lack the consensus GTT GCT or the related sequence GGT GCT.

The 32- and 12-amino acid repeats in the FC27 and K1 alleles are not obviously related to the 4- and 8-amino acid repeats in the IC1/MAD71/NIG32-type alleles. Nevertheless, it has been previously noted that antibodies affinity purified on a synthetic peptide corresponding to three copies of the GGSA repeat present in the MSA-2 of 3D7 and IC1 react with the FC27 MSA-2, possibly due to a cross-reacting epitope within the 32-amino acid repeat (3). The 12-amino acid repeat in K1, which is closely related to a sequence immediately C-terminal to the 32-amino acid repeats in FC27, is also related to sequences immediately C-terminal to the 8- and 4-amino acid repeats in the other alleles. Apart from this short region of sequence with limited homology, the sequences between the repeats and the conserved N- and C-terminal domains are highly conserved within the two MSA-2 allele families but are unrelated between the two families. Point mutations within these flanking variable re-

FIG. 2. Alignment of the translated sequence of the MSA-2 alleles. Dots in the K1 sequence indicate identity with the FC27 sequence, whereas dots in the IC1, MAD71, IMR143, and NIG32 sequences indicate identity with the 3D7 sequence. Dashes represent spaces inserted to maximize alignment. The alignment incorporates spaces to accommodate the various sequence block deletions. The shaded areas indicate the conserved N- and C-terminal regions. The solid bars indicate the signal and anchor sequences. The sequence of FC27 presented here differs from that reported in ref. 2. Two sequencing errors have been corrected—an A has been removed at nucleotide 530 and a C has been inserted at nucleotide 545 correcting a frame shift of five residues.

gions have a nonrandom distribution being clustered close to either end of the repetitive sequences. For example, of the 11 nucleotide differences between the closely related 3D7 and IC1 MSA-2 alleles, 5 of these, all of which result in amino acid changes, are clustered between the 3′ end of the N-terminal conserved domain and the 5′ end of the repetitive sequences (nucleotides 129–165).

DISCUSSION

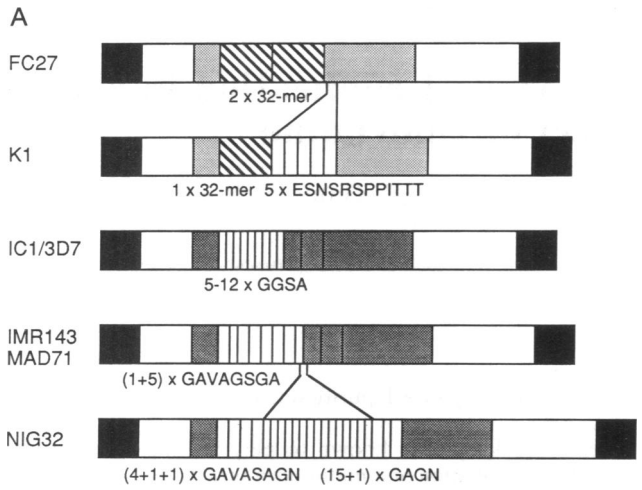
Although some of the diversity in *P. falciparum* results from point mutations leading to gene products of different antigenicity, two other genetic mechanisms contribute to antigenic diversity in *P. falciparum* by causing major primary structural changes in allelic gene products. First, many antigens of *P. falciparum* contain internal segments that consist of short tandemly repeated sequences, which are the major antibody binding sites (10). A variable antigen with this general structure is the S antigen, which is encoded by a single copy gene in the haploid genome of the parasite. S-antigen genes occur as a series of very different but apparently stable alleles coexisting in parasite populations worldwide. In different isolates of *P. falciparum* the central block of repeats found in S antigens varies quite dramatically in sequence, length, and number of repeats and also in reading frame, resulting in marked differences in antigenicity (12, 15–17). Similar repeat variation is seen in the circumsporozoite proteins of different malaria species (18–20).

The second major form of variation is that seen in the MSA-1 gene. Intragenic recombination between two very different alleles can explain the patterns of variation seen in MSA-1 (1, 7, 9). This could only happen during the diploid phase of the sexual cycle. At the same time Mendelian segregation of the chromosomes into the haploid progeny has the potential to continuously reassort the products of diversity generated by any mechanism.

The diversity in MSA-2 described here most closely resembles that of the S antigens as the gene contains an internal variable repetitive segment with relatively conserved regions at either end of the gene. But despite this superficial similarity there are important differences. It is now clear that the S-antigen flanking sequences in different isolates of *P. falciparum* are far more divergent than other *P. falciparum* genes (18). Like these latter genes, the 5′ (43 amino acids) and 3′ (74 amino acids) coding regions of MSA-2 are very conserved with only a single base substitution among the seven alleles we have sequenced. In contrast, the 5′ and 3′ coding regions of the 5 S-antigen genes sequenced fall into three families, which at the nucleotide level vary by ≈20% (16, 17). Immediately flanking the MSA-2 repeats are highly variable nonrepetitive sequences that are of two types. Hence the organization in these regions is reminiscent of the dimorphic alleles of MSA-1. Intragenic recombination generating intermediate forms of MSA-2 does not appear to be a major cause of diversity as appears to be the case for MSA-1.

Another way in which MSA-2 variation resembles S antigen and circumsporozoite protein variation is in the relationship among repeats of different alleles. Although, as with S antigens, the repeats can vary dramatically from isolate to isolate, some underlying patterns can be clearly seen. For example, we have shown that the 11-amino acid repeat of the V1 S antigen can be derived from recombination within the sequence coding for 15-amino acid repeats located toward the C-terminal end of the NF7 S antigen (16). Similarly, the 12-amino acid repeats of the K1 S antigen are closely related to those of NF7 but have undergone a change in reading frame (12). A similar relationship exists between some of the repeats in MSA-2 reported here. For example, the tetrapeptide GAGN in MSA-2 from NIG32 can obviously be derived from the octamer GAVASAGN also present in NIG32. As we pointed out above, these NIG32 repeats are also related to the IMR143 and MAD71 MSA-2 repeats (GAVAGSGA).

Some MSA-2 repeats can be viewed as amplifications of regions of sequence represented once in the 3′ flanking regions of that allelic family. Hence the repeat ESNSRSPPITTT in K1 appears to be an amplification of the sequence ESISPSPPITTT, which occurs only once in FC27, and the repeat GAGN in NIG32 may derive from the single GAGN, which occurs at the C-terminal end of the octapeptide repeats of MAD71 and IMR143 (amino acids 109–112 inclusive in the MAD71 allele), or the closely related GNGA in 3D7 and IC1 (Figs. 1 and 2). Also, the repeat GAVASAGN in NIG32 can be derived by amplification of this sequence, which occurs once at the C-terminal end of the repeat region of MSA-2 in Thai Tn (21). We do not know whether evolution has taken this direction, but a very similar relationship between major



B

GGT GCT	GGT GCT	Consensus	
GGT **G**GT	**A**GT GCT	**3D7**	GGSA
GGT GCT	G**T**T GCT	**MAD71**	GAVA
GGT **T**CT	GGT GCT	"	GSGA
GGT GCT	G**T**T GCT	**NIG32**	GAVA
AGT GCT	GGT **AA**T	"	SAGN
GGT GCT	GGT **AA**T	"	GAGN

C

FC27	E	S	I	S	P	S	P	P	I	T	T	T
K1			N		R							
3D7	A	E	G		S		T		A			
MAD71	A	E	R	G	P		T		A			
IMR143	A	E	R		P		T		A			
IC1	A	E	R		P		T		A			
NIG32	A	E	R		S		T		A			

FIG. 3. (*A*) Schematic diagrams of MSA-2 genes. Black regions at the left and right indicate signal and anchor sequences. Unfilled regions internal to these represent conserved regions, whereas dotted and striped regions represent variable nonrepetitive and repetitive sequences, respectively. The lines connecting FC27/K1 and MAD71/NIG32 indicate the FC27 and MAD71-type 3′ sequences apparently amplified in K1 and NIG32. (*B*) The 4- and 8-amino acid repeats of MSA-2 are related by a hexanucleotide repeat sequence. The sequences of the 4- and 8-amino acid repeats of 3D7 and IMR143 have been aligned as hexanucleotide sequences with the consensus GGT GCT (shown at the top) to highlight the close relationship between them. By use of this alignment there are only four nucleotide mutations between these two repeats, all of which result in amino acid substitutions, indicated by bold lettering. The NIG32 4- and 8-amino acid sequences are also closely related to this consensus. (*C*) The 12-amino acid repeat in the K1 MSA-2 allele is related to a single copy sequence in the FC27 allele. The K1 allele of MSA-2 contains five copies of a 12-amino acid repeat that appears to have been amplified from a sequence represented as a single copy in the FC27 MSA-2 allele. The sequence in K1 differs from the single copy FC27 sequence by two second base mutations, resulting in substitution of an I for an N and P for an R, respectively. Also shown are single copy sequences in the other alleles that are related to the 12-amino acid repeat in K1.

repeat sequences and 3′ flanking sequences has been noted with S antigens. Further, similarities have also been noted with the human minisatellite MS2, an internally repetitive sequence in which alterations were clustered at the 3′ end (22). The mechanism of gene conversion invoked (20) specifically to explain the low numbers of Ts in some malarial repeats would appear not to be relevant here as the repeats have a relatively high T content. However, gene conversion as a mechanism can just as easily explain the MSA-2 sequences as other recombinatorial mechanisms.

The location of MSA-2 on the merozoite surface makes it directly accessible to immune effector mechanisms. It is clear that the MSA-2 repeats encode antigenic epitopes (2) and that antibodies to repeats can inhibit merozoite invasion *in vitro* (5, 23), suggesting that MSA-2 may be a target of protective antibody responses. Therefore, it seems likely that much of the diversity in MSA-2 has evolved under the selection pressure of host immune responses.

Most, if not all, MSA-2 genes have repeat sequences that are related to one or more of the sequences known to date, as none of 44 isolates studied here failed to hybridize with at least one repeat probe. Of the 3 isolates that hybridized with more than one probe, one (311) was a cloned line. We have recently established that the MSA-2 gene of 311 contains another repeat type that is related to the IC1 4-amino acid repeat and the MAD71 8-amino acid repeat. The other double positives may reflect mixed parasite populations.

Our results indicate that variability in MSA-2 repeats is more restricted than variability in S-antigen repeats. Even greater constraints appear to be operating on the nonrepetitive variable regions of MSA-2, as these are of only two types (21). Further studies will be required to establish whether diversity in MSA-2 is restricted enough for this molecule to have potential as a vaccine component.

We thank Etty Bonnici for typing this manuscript. This work was supported by the Australian National Health and Medical Research Council, the John D. and Catherine T. MacArthur Foundation, and the Australian Malaria Vaccine Joint Venture. Support was also provided under the Generic Technology component of the Industry Research and Development Act 1986.

1. Holder, A. A. (1988) *Prog. Allergy* **41**, 72–97.
2. Smythe, J. A., Coppel, R. L., Brown, G. V., Ramasamy, R., Kemp, D. J. & Anders, R. F. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 5195–5199.
3. Smythe, J. A., Peterson, M. G., Coppel, R. L., Saul, A. J., Kemp, D. J. & Anders, R. F. (1990) *Mol. Biochem. Parasitol.* **39**, 227–234.
4. Ramasamy, R. (1987) *Immunol. Cell Biol.* **65**, 419–424.
5. Clark, J. T., Donachie, S., Anand, R., Wilson, C. F., Heidrich, H. G. & McBride, J. S. (1989) *Mol. Biochem. Parasitol.* **32**, 15–24.
6. Fenton, B., Clark, J. T., Wilson, C. F., McBride, J. S. & Walliker, D. (1989) *Mol. Biochem. Parasitol.* **34**, 79–86.
7. Tanabe, K., Mackay, M., Goman, M. & Scaife, J. G. (1987) *J. Mol. Biol.* **195**, 273–287.
8. Certa, U., Rotmann, D., Matile, H. & Reber-Liske, R. (1987) *EMBO J.* **6**, 4137–4142.
9. Peterson, M. G., Coppel, R. L., Moloney, M. B. & Kemp, D. J. (1988) *Mol. Cell Biol.* **8**, 2664–2667.
10. Kemp, D. J., Coppel, R. L. & Anders, R. F. (1987) *Annu. Rev. Microbiol.* **41**, 181–208.
11. Forsyth, K. P., Philip, G., Smith, T., Kum, E., Southwell, B. & Brown, G. V. (1989) *Am. J. Trop. Med. Hyg.* **41**, 259–265.
12. Saint, R. B., Coppel, R. L., Cowman, A. F., Brown, G. V., Shi, P. T., Barzaga, N., Kemp, D. J. & Anders, R. F. (1987) *Mol. Cell Biol.* **7**, 2968–2973.
13. Foote, S. J., Kyle, D. E., Martin, R. K., Oduola, A. M. J., Forsyth, K., Kemp, D. J. & Cowman, A. F. (1990) *Nature (London)* **345**, 255–258.
14. Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B. & Erlich, H. A. (1988) *Science* **239**, 487–491.
15. Cowman, A. F., Saint, R. B., Coppel, R. L., Brown, G. V., Anders, R. F. & Kemp, D. J. (1985) *Cell* **40**, 775–783.
16. Brown, H., Kemp, D. J., Barzaga, N., Brown, G. V., Anders, R. F. & Coppel, R. L. (1987) *Mol. Biol. Med.* **4**, 365–376.
17. Nicholls, S. C., Hillman, Y., Lockyer, M. J., Odink, K. G. & Holder, A. A. (1988) *Mol. Biochem. Parasitol.* **28**, 11–20.
18. McCutchan, T. F., de la Cruz, V. F., Good, M. F. & Wellems, T. E. (1988) *Prog. Allergy* **41**, 173–192.
19. Galinski, M. R., Arnot, D. E., Cochrane, A. H., Barnwell, J. W., Nussenzweig, R. S. & Enea, V. (1987) *Cell* **48**, 311–319.
20. Arnot, D. E., Barnwell, J. W. & Stewart, M. J. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 8102–8106.
21. Thomas, A. W., Carr, D. A., Carter, J. M. & Lyon, J. A. (1990) *Mol. Biochem. Parasitol.* **43**, 211–220.
22. Jeffries, A. J., Neumann, R. & Wilson, V. (1990) *Cell* **60**, 473–485.
23. Epping, R. J., Goldstone, S. D., Ingram, L. T., Upcroft, J. A., Ramasamy, R., Cooper, J. A., Bushell, G. R. & Geysen, H. M. (1988) *Mol. Biochem. Parasitol.* **28**, 1–10.