Variation in the gene encoding a major merozoite surface antigen of the human malaria parasite *Plasmodium falciparum*

James L. Weber, William M. Leininger and Jeffrey A. Lyon

Department of Immunology, Walter Reed Army Institute of Research, Washington, DC 20307-5100, USA

Received 16 December 1985; Accepted 3 March 1986

ABSTRACT

Plasmodium falciparum merozoites have a variable surface protein of about 195,000 molecular weight which may be involved in strain-specific immunity. We have cloned and sequenced a major portion of the gene encoding this antigen from the CAMP strain and have located sites of preferred mung bean nuclease cleavage around the gene. These sites depend on reaction conditions, but at 40% formamide and 2 units of mung bean nuclease per μ g DNA, the intact gene was excised from the chromosome. Comparison of the CAMP strain gene with the same gene from other strains of *P. falciparum* by matching available DNA sequences and by DNA hybridization revealed five regions of homology separated by divergent segments. Two of the variable regions encoded three amino acid repeats, predominantly Ser-Gly-Thr and Thr-Glu-Glu. Implications of these findings on the function of the antigen, and possible mechanisms for generation of variants are discussed.

INTRODUCTION

Immunity to the blood stage forms of the human malaria parasite *Plasmodium falciparum* can be induced by infection, by passive transfer of immune serum or by vaccination, however, this immunity has almost always been found to be strain specific (1-4). Understanding the molecular basis of antigen diversity and the mechanism of generation of variation is therefore critical to vaccine development.

The genes for several malaria antigens have now been cloned and sequenced (5-10). All have been found to contain tandem repeats ranging from 9 to 150 base pairs (bp). Some of the antigen genes like the RESA gene (11) and the *P. falciparum* circumsporozoite protein gene (12) appear to be conserved in sequence between strains, while others like the S antigen gene (13) vary. One *P. falciparum* antigen which has been studied in considerable detail is a merozoite surface glycoprotein of about 195,000 molecular weight (gp195) (14-20). This large antigen is processed about the time of erythrocyte rupture into several smaller species which are also located at the merozoite surface. Tests with monoclonal antibodies against this

protein have shown that the antigen varies among strains.

We have compared our sequence of about 3.6 kilobase pairs (kb) of the 5' portion of the gp195 gene from the CAMP (Malaysia) strain of P. falciparum to sequences from other strains (21-23), and in addition have hybridized the CAMP gene to DNA from other strains. As expected, we found that several segments of the gene, including the repeat regions, varied between strains, but we also found five other regions which exhibited substantial sequence homology, with 76-97% of the bases being identical.

MATERIALS AND METHODS

Parasite DNA and expression library construction

DNA was isolated from cultured P. falciparum of the CAMP strain as described (12) except that as an additional purification step, the DNA was banded on CsCl gradients in the presence of ethidium bromide. The DNA was sheared by mung bean nuclease (Pharmacia P-L Biochemicals) digestion (24) in either 35% or 40% formamide using a range of 1 to 10 units of enzyme per μ g of DNA, and the products of the different reactions were combined. Sizes of the resulting fragments ranged from 10 kb down to very small pieces, with DNA of about 1.5 kb giving maximum ethidium bromide staining. Ends of the DNA fragments were blunted with T4 DNA polymerase before addition of EcoR I linkers. Excess linkers were removed by EcoR I digestion and chromatography over Sephacryl S300 followed by a second round of digestion with EcoR I and chromatography (25). The resulting DNA was ligated into λ gt11 vector treated with alkaline phosphatase (26). When grown in E. coli strain Y1090 (27), 90% of the phage in the library gave white plaques when plated in the presence of 5-bromo-4-chloro-3-indolyl- β -D-galactoside and isopropylthio- β -D-galactopyranoside.

Screening of expression library

The library was screened (19) using owl monkey immune serum. Clones encoding portions of gp195 within the group of monkey serum positive clones were identified by their reaction with monoclonal antibodies recognizing gp195.

Sequencing

The insert from clone a88 was sequenced essentially as described (28). Fragments of the a88 insert generated by sonication were blunt-ended, size selected and ligated into the Sma I site of mp10. Single stranded DNA from the recombinant phage was sequenced using 35 S-dATP and gradient gels (29). Sequence information was compiled using software produced by Intelligenetics Inc. for BIONET.

Figure 1. Mung Bean Nuclease Sites about the gp195 Gene. CAMP strain genomic DNA was digested with mung bean nuclease at the indicated percentages of formamide and units of enzyme per μg of DNA (24). DNA in the outer lanes was additionally digested with *EcoR* I. After electrophoresis and transfer, the DNA fragments were hybridized to the insert from clone a88. Numbers at the left are the sizes in kb of some of the visible bands.

Hybridization

Southern blotting and subsequent hybridization were performed as described (12) except that the DNA fragments used as probes were isolated by preparative agarose gel electrophoresis followed by electroelution into a V-

20 40 60 80 100 120
GAATTCATATATAATTTTTTTCACAAAAACAAAAAAGGGTGTATATACCTTATATATATATATATATACATATGTGAAGGAAATATGTGAATAATATATAA

140 160 180 200 220 240
ATTATAGTTATGATGAATAAATTTTTTATTATAAAAAAAGGCTAATGAAATGCAAAAAAATGTATACATATTTTTCGAAGTCATATTTTAAATATTAACCTATTTTATT

260 280 300 320 340
ATTATTTTTTTTATATATATATTTTAGCTTTAATTCAGTA ATG AAG ATC ATA TTC TTT TTA TGT TCA TTT CTT TTT TTT ATT ATA AAT ACA CAA
Cys Val Thr His Glu Ser Tyr Gln Glu Leu Val Lys Lys Leu Glu Ala Leu Glu Asp Ser Phe Leu Phe Phe Ile Ile Asn Thr Gln

360 380 400 420 440 460 480 500 520
TGT GTA ACA CAT GAA AGT TAT CAA GAA CTT GTC AAA AAA CTA GAA GCT TTA GAA GAT GCA GTA TTG ACA GGT TAT GGT TTA TTT CAT AAG
GAA AAA ATG ATC TTA AAT GAA GAA GAA ATT ACT ACA AAA GGT GCA AGT GCT CAA AGT GGT ACA AGT GGT ACA AGT GGT ACA AGT GGT ACA
Glu Lys Met Ile Leu Asn Glu Glu Glu Ile Thr Thr Lys Gly Ala Ser Ala Gln Ser Gly Thr Ser Gly Thr Ser Gly Thr Ser Gly Thr

540 560 580 600 620 640 660 680 700
AGT GGT ACA AGT GGT ACA AGT GGT ACA AGT GCT CAA AGT GGT ACA AGT GGT ACA AGT GGT ACA AGT GGT ACA AGT GCT CAA
Ser Gly Thr Ser Gly Thr Ser Gly Thr Ser Gly Thr Ser Gly Thr Ser Ala Gln Ser Gly Thr Ser Ala Gln Ser Gly Thr Ser Ala Gln

720 740 760 780 800 820 840 860 880
AGT GGT ACA AGT GGT ACA AGT GGT ACA AGT GGT ACA AGT CCA TCA TCT CGT TCA AAC ACT TTA CCT CGT TCA AAT ACT TCA TCT GGT GCA
Ser Gly Thr Ser Gly Thr Ser Gly Thr Ser Gly Thr Ser Pro Ser Ser Arg Ser Asn Thr Leu Pro Arg Ser Asn Thr Ser Ser Gly Ala

900 920 940 960 980 1000 1020 1040 1060
AGC CCT CCA GCT GAT GCA AGC GAT TCA GAT GCT AAA TCT TAC GCT GAT TTA AAA CCA AGA GTA CGA AAT TAC TTG TTC ACT ATT AAA GAA
Ser Pro Pro Ala Asp Ala Ser Asp Ser Asp Ala Lys Ser Tyr Ala Asp Leu Lys His Arg Val Arg Asn Tyr Leu Phe Thr Ile Lys Glu

1080 1100 1120 1140 1160 1180 1200 1220 1240
CTC AAA TAT CCC GAA CTC TTT GAT TTA ACC AAT CAT ATG TTA ACT TTG TGT GAT AAT ATT CAT GGT TTC AAA TAT TTA ATT GAT GGA TAT
Lys Asp Asn Val Gln Lys Met Leu Thr Lys Leu Cys Asp Asn Ile His Gly Phe Lys Tyr Leu Ile Asp Gly Lys Tyr

1260 1280 1300 1320 1340 1360 1380 1400 1420
GAA GAA ATT AAT GAA TTA TTA TAT AAA TTA AAC TTT TAT TTT GAT TTA TTA AGA GCA AAA TTA AAT GAT GTA TGT GCT AAT GAT TAT TGT
Glu Glu Ile Asn Thr Leu Leu Tyr Phe Asp Lys Asn Phe Tyr Phe Asp Leu Arg Ala Lys Leu Asn Asp Val Cys Ala Asn Asp Tyr Cys

1440 1460 1480 1500 1520 1540 1560 1580 1600
CAA ATA CTT TTC AAT CTT AAT ATT CGT GCA AAT GAA TTA GAC GTA CTT AAA AAA CTT GTG TTC GGA TAT AGA AAA CCA TTA GAC AAT ATT
Gln Ile Pro Phe Asn Leu Lys Ile Arg Ala Asn Glu Leu Asp Val Leu Lys Lys Leu Val Phe Gly Tyr Arg Lys Pro Leu Asp Asn Ile

1620 1640 1660 1680 1700 1720 1740 1760 1780
AAA GAT AAT GTA GGA AAA ATG GAA GAT TAC ATT AAA AAA AAT AAA ACC ATA GCA AAT ATA AAT GAA TTA ATT GAA GGA AGT AAG AAA
Lys Asp Asn Val Gln Asn Glu Lys Met Leu Asp Tyr Ile Lys Lys Asn Lys Thr Thr Ile Ala Asn Ile Asn Glu Leu Ile Glu Gly Ser Lys Gln

1800 1820 1840 1860 1880 1900 1920 1940 1960
ACA ATT GAT CAA AAT AAG AAT GCA GAT AAT GAA GAA GGA AAA AAA TTA TAC CAA GCT CAA TAT GAT CTT CTT ATT TAC AAT AAA CAA
Thr Ile Asp Gln Asn Lys Asn Ala Asp Asn Glu Glu Gly Lys Lys Leu Tyr Gln Ala Gln Tyr Asp Leu Ser Ile Tyr Asn Lys Gln

2000 2020 2040 2060 2080 2100 2120 2140 2160
TTA GAA GAA CCA CAT AAT TTA ATG AGC GTT TTA GAA AAA CGT ATT GAC ACT TTA AAA AAA AAT GAA AAT AAG GAA TTA CTT GAT AAT
Leu Glu Glu Ala His Asn Leu Ile Ser Val Leu Glu Lys Arg Ile Asp Thr Leu Lys Lys Asn Glu Asn Ile Lys Glu Leu Leu Asp Lys

2180 2200 2220 2240 2260 2280 2300 2320 2340
ATA AAT GAA ATT AAA AAT CCC CCA CCG GCC AAT TCT GGA AAT CCA CCA AAT ACT CTC CTT GAT AAG AAC AAA AAA ATC GAG GAA CAC GAA
Ile Asn Glu Ile Lys Asn Pro Pro Pro Ala Asn Ser Gly Asn Thr Pro Asn Thr Leu Leu Asp Lys Lys Lys Lys Lys Lys Lys Lys Lys Lys

2360 2380 2400 2420 2440 2460 2480 2500 2520
GAA AAA ATA AAA GAA ATT GCC AAA ACT ATT AAA TTT AAC ATT GAT AGT TTA TTT ACT GAT CCA CTT GAA TTA GAA TAT TAT AGA GAA
Glu Lys Ile Lys Glu Ile Ala Lys Thr Ile Lys Phe Asn Ile Asp Ser Leu Phe Thr Asp Pro Leu Glu Leu Glu Tyr Tyr Leu Arg Glu

2540 2560 2580 2600 2620 2640 2660 2680 2700
AAA AAT AAA AAA GTT GAT GTA CCA CCT AAA TCA CAA GAT CCT ACG AAA TCT GTT CAA ATA CCA AAA GTT CCT TAT CCA AAT GGT ATT GTA
Lys Asn Lys Lys Val Asp Val Thr Pro Lys Ser Gln Asp Pro Thr Lys Ser Val Gln Ile Pro Lys Val Gln Tyr Pro Asn Gln Tyr Lys

2720 2740 2760 2780 2800 2820 2840 2860 2880
TAT CCT TTA CCA CTC ACT GAT ATT CAT AAT TCA TTA GCT GCA GAT AAT GAT AAA AAT TCA TAT GGT GAT TTA ATG AAT CCT GAT ACT AAA
Tyr Pro Leu Pro Leu Thr Asp Ile His Asn Ser Leu Ala Ala Asp Asn Asp Lys Asn Ser Tyr Gly Asp Leu Met Asn Pro Asp Thr Lys

2900 2920 2940 2960 2980 3000 3020 3040 3060
GAA AAA ATT AAT GAA AAA ATT AAT ACA GAT AAT AAG GAA AGA AAA ATA TTC ATT AAT AAC ATT AAA AAA CAA AAT ATT GAT TTA GAA GAA
Glu Lys Ile Asn Glu Lys Ile Thr Asp Asn Lys Glu Arg Lys Ile Phe Ile Asn Asn Ile Lys Lys Gln Ile Asp Leu Glu Glu Lys

3080 3100 3120 3140 3160 3180 3200 3220 3240
AAA ATT AAT CAC ACA AAA GAA CAA AAT AAA AAA TTA CTT GAA GAT TAT GAA AAG TCA AAA AAG GAT TAT GAA GAA TTA CTT GAA AAA TTT
Lys Ile Asn His Thr Lys Glu Gln Asn Lys Lys Leu Glu Asp Tyr Glu Lys Ser Lys Lys Asp Tyr Glu Glu Leu Leu Lys Phe

3260 3280 3300 3320 3340 3360 3380 3400 3420
TAT GAA ATG AAA TTT AAT AAT AAT TTT GAC AAA GAT GTC GTA GAT AAA ATA TTC AGT GCA AGA TAT ACA TAT AAT GTT GAA AAA CAA AGA
Tyr Glu Met Lys Phe Asn Asn Asn Phe Asp Lys Asp Val Val Asp Lys Ile Phe Ser Ala Arg Tyr Thr Tyr Asn Val Glu Lys Gln Arg

3440 3460 3480 3500 3520 3540 3560 3580 3600
TAT AAT AAT AAA TTT TCA TCC TCT AAT AAT TCT GTA TAT AAT GTT CAA AAA TTA AAA AAG GCT CTT TCA TAT CTT GAA GAT TAT TCT TTA
Tyr Asn Asn Lys Phe Ser Ser Ser Asn Ser Val Tyr Asn Val Gln Lys Leu Lys Lys Ala Leu Ser Tyr Glu Leu Asp Tyr Thr

3620 3640 3660 3680 3700 3720 3740 3760 3780
AGA AAA GGA ATT TCT GAA AAA GAT TTT AAT CAT TAT TAT ACT TTG AAA ACT GGC CTC GAA GCT GAT ATA AAA AAA TTA ACA GAA GAA ATA
Arg Lys Gly Ile Ser Glu Lys Asp Phe Asn His Tyr Tyr Thr Leu Lys Thr Gly Leu Glu Ala Asp Ile Lys Lys Leu Thr Glu Glu Ile

3800 3820 3840 3860 3880 3900 3920 3940 3960
AAG AGT AGT GAA AAC AAA ATT CTA GAA AAA AAT TTT AAA GGA CTA ACA CAT TCA GCA AAT GCT TCC TTA GAA GTA TAT GAT ATT GTA AAA
Lys Ser Ser Glu Asn Lys Ile Leu Glu Lys Asn Phe Lys Gly Leu Thr His Ser Ala Asn Ala Ser Leu Glu Val Tyr Asp Ile Val Lys

3980 4000 4020 4040 4060 4080 4100 4120 4140
TTA CAA GTA CAA AAA GTT TTA TTA ATT AAA AAA ATA GAA GAC TTA AGA AAG ATA GAA TTA TTT TTA AAA AAT GCA CAA CTA AAA GAT AGT
Leu Gln Val Gln Lys Val Leu Leu Ile Lys Lys Ile Glu Asp Leu Arg Lys Ile Glu Leu Phe Leu Lys Asn Ala Gln Leu Lys Asp Ser

4160 4180 4200 4220 4240 4260 4280 4300 4320
ATT CAT GTA CCA AAT ATT TAT AAA CCA CAA AAT AAA CCA GAA CCA TAT TAT TTA ATT GTA TTA AAA GAA GTA GAT AAA TTA AAA GAA
Ile His Val Pro Asn Ile Tyr Lys Pro Gln Asn Lys Pro Glu Pro Tyr Tyr Leu Ile Val Lys Lys Lys Lys Ser Asn Asp Asn Ser Gln

4340 4360 4380 4400 4420 4440 4460 4480 4500
TTT ATA CCA AAA GTA AAA GAC ATG TTA AAG AAA GAA CAA GCT GTC TTA TCA AGT ATT ACA CAA CCT TTA GTC GCA AGC GAA ACA ACT
Phe Ile Pro Lys Val Lys Asp Met Leu Lys Lys Glu Gln Ala Val Leu Ser Ser Ile Thr Gln Pro Leu Val Ala Ala Ser Glu Thr Thr

4520 4540 4560 4580 4600 4620 4640 4660 4680
GAA GAT GGG GGT CAC TCC ACA CCA TTA TCC CCA GGA GAA ACA GAA GTA ACA GAA GAA CAA GAA ACA GAA GAA CTA GGA
Glu Asp Gly Gly His Ser Thr His Thr Leu Ser Gln Ser Gly Glu Thr Glu Val Thr Glu Glu Thr Glu Glu Thr Val Gly

4700 4720 4740 4760 4780 4800 4820 4840 4860
CAC ACA ACA ACG GTA ACA ATA CCA TTA CCA CCA AAA GAA GTA AAA GTT GTT GAA AAT TCA ATA GAA CAT AAG AGT AAT GAC AAT TCA CAA
His Thr Thr Val Thr Ile Thr Leu Pro Pro Lys Glu Val Lys Val Val Glu Asn Ser Ile Glu His Lys Ser Asn Asp Asn Ser Gln

4880 4900 4920 4940 4960 4980 5000 5020 5040
GCC TTG ACA AAA ACA GTT TAT CTA AAG AAA TTA GAT GAA TTT TTA ACT AAA TCA TAT ATA TGT CAT AAA TAT ATT TTA GTA TCA AAC TCT
Ala Leu Thr Lys Thr Val Tyr Leu Lys Lys Leu Asp Glu Phe Leu Thr Lys Ser Tyr Ile Cys His Lys Tyr Ile Leu Val Ser Asn Ser

```

2780      2800      2820      2840      2860
AGT ATG GAC CAA AAA TTA TTA GAG GTA TAT AAT CTT ACT CCA GAA GAA GAA AAT GAA TTA AAA TCA TGT GAT CCA TTA GAT TTA TTA TTT
Ser Met Asp Gln Lys Leu Leu Glu Val Tyr Asn Leu Thr Pro Glu Glu Glu Asn Glu Leu Lys Ser Cys Asp Pro Leu Asp Leu Leu Phe
2880      2900      2920      2940
AAT ATT CAA AAT AAC ATA CCT GCT ATG TAT TCA TTA TAT GAT AGT ATG AAC AAT GAT TTA CAA CAT CTC TTT TTT GAA TTA TAT CAA AAG
Asn Ile Gln Asn Asn Ile Pro Ala Met Tyr Ser Leu Tyr Ser Met Asn Asn Asp Leu Gln His Leu Phe Phe Glu Leu Tyr Gln Lys
2960      2980      3000      3020      3040
GAA ATG ATT TAT TAT TTA CAT AAA CTA AAA GAG GAA AAT CAC ATC AAA AAA TTA TTA GAG GAG CAA AAA ATA ACT GGA ACA TCA TCT
Glu Met Ile Tyr Tyr Leu His Lys Leu Lys Glu Glu Asn His Ile Lys Lys Leu Leu Glu Glu Gln Lys Gln Ile Thr Gly Thr Ser Ser
3060      3080      3100      3120
ACA TCC AGT CCT GGA AAT ACA ACC GTA AAT ACT GCT CAA TCC GCA ACT CAC AGT AAT TCC CAA AAC CAA CAA TCA AAT GCA TCC TCT ACC
Thr Ser Pro Gly Asn Thr Thr Val Asn Thr Thr Ala Gln Ser Ala Thr His Ser Asn Ser Gln Asn Gln Gln Ser Asn Ala Ser Ser Thr
3140      3160      3180      3200      3220
AAT ACC CAA AAT GGT GTA GCT GCA TCA TCT GGT CCT GCT GTA GTT GAA GAA AGT CAT GAT CCC TTA ACA GTA TTG TCT ATT AGT AAC GAT
Asn Thr Gln Asn Gly Val Ala Val Ser Ser Gly Pro Ala Val Val Glu Glu Ser His Asp Pro Leu Thr Val Leu Ser Ile Ser Asn Asp
3240      3260      3280      3300
TTG AAA GGT ATT GTT AGT CTC TTA AAT CTT GGA AAT AAA ACT AAA GTA CCT AAT CCA TTA ACC ATT TCT ACA ACA GAG ATG GAA AAA TTT
Leu Lys Gly Ile Val Ser Leu Leu Asn Leu Gly Asn Lys Thr Lys Val Pro Asn Pro Leu Thr Ile Ser Thr Thr Glu Met Glu Lys Phe
3320      3340      3360      3380
TAT GAG AAT ATT TTA AAA AAT AAT GAT ACC TAT TTT AAT GAT GAT ATC AAA CAA TTC GAT AAA TCT AAT TCA AAA GTA ATT ACA GGT TTG
Tyr Glu Asn Ile Leu Lys Asn Asn Asp Thr Tyr Phe Asn Asp Ile Lys Gln Phe Val Lys Ser Asn Ser Lys Val Ile Thr Gly Leu
3400      3420      3440      3460      3480
ACC GAA ACA CAA AAA AAT GCA TTA AAT GAT GAA ATT AAA AAA TTA AAA GAT ACT TTA CAG TTA TCA TTT GAT TTA TAT AAT AAA TAT AAA
Thr Glu Thr Gln Lys Asn Ala Leu Asn Asp Glu Ile Lys Lys Leu Lys Asp Thr Leu Gln Leu Ser Phe Asp Leu Tyr Asn Lys Tyr Lys
3500      3520      3540      3560      3580
TTA AAA TTA GAT AGA TTA TTT AAT AAG AAA AAA GAA CTT GGC CAA GAC AAA ATG CAA ATT AAA AAA CTT ACT TTA TTA AAA GAA CAA TTA
Leu Lys Leu Asp Arg Leu Phe Asn Lys Lys Lys Glu Leu Gly Gln Asp Lys Met Gln Ile Lys Lys Leu Thr Leu Leu Lys Glu Gln Leu
3598
GAA TCA AAA TTG AAT TC
Glu Ser Lys Leu Asn Ser

```

Figure 2. Sequence of 3.6 kb of the gp195 Gene from the CAMP Strain. Bases are numbered beginning at the 5' *EcoR* I linker. The probable translation start site is base 288. Dots indicate potential glycosylation sites (Asn-X-Thr/Ser), and repeated sequences are underlined.

shaped channel (International Biotechnologies Inc., model UEA) filled with 7.5 M ammonium acetate.

RESULTS

Mung bean nuclease sites about the gp195 gene

Clones encoding gp195 were selected from a genomic DNA expression library prepared using mung bean nuclease to catalyze shearing of the large chromosomal DNA. Preferred sites of mung bean nuclease cleavage about the gp195 gene were determined by hybridizing the genomic DNA cleavage products to the cloned gene as shown in Figure 1. Sites were found to be highly dependent upon reaction conditions. As the formamide or nuclease concentrations were increased, the enzyme cut more frequently. Separate genomic Southern hybridizations showed that cleavage with *EcoR* I alone gave a labeled fragment of 4.9 kilobase pairs (kb), and that the 3' end of this fragment was the same as the common 3' end of most of the clones shown in Figure 4A (data not shown). At 35% formamide and 2 units nuclease per μ g DNA the preferred mung bean sites are outside of the 4.9 kb fragment (Figure 1, first two lanes). At 40%, 2 units/ μ g, the enzyme cuts near the probable translation start site and about 2.0 kb downstream of the common 3' ends of our clones (third and sixth lanes). Therefore under these conditions, the

SGE2	EKMVLNE	GTSGTAVTTSTPG	SGGSVTSGGSGGSA	SVASGG	SGGSA
MAD20	EKMVLNE	GTSGTAVTTSTPG	SSGSVTSGG	SVASVASVG	SGGSA
We'll come	EKMVLNE	GTSGTAVTTSTPG	SKGSVASGGSGGSA	SGGSVASGGSA	
K1	EKMILNE	EEITTKGA	SAQSGTSGTSGTSGPSPGT		
CAMP	EKMILNE	EEITTKGA	SAQSGTSGTSGTSGTSGTSGTSAQSGTSGT		

SGE2	SGGSGN		SRRTNPSDNS	SDSDAKS
MAD20	SGGSGN		SRRTNPSDNS	SDSD
We'll come	SGGSGN		SRRTNPSDNS	SDSDAKS
K1			SPSSRSNTLPRNSTSGASPPADA	SDSDAKS
CAMP	SAQSGTSGTSAQSGTSGTSGTSGT		SPSSRSNTLPRNSTSGASPPADA	SDSDAKS

Figure 3. Comparison of gp195 Repeats from Five Strains. Amino acid sequences are shown for the SGE2 (Zaire), MAD20 (Papua New Guinea), Wellcome (West Africa), K1 (Thailand) and CAMP (Malaysia) strains.

nuclease cuts at sites near the beginning and end of the gene, in agreement with the results of McCutchan et al (24).

At 10 units of mung bean nuclease and 35 or 40% formamide, however, the enzyme cuts within the coding region itself producing prominent fragments of about 3.4 kb and a smear of smaller fragments (Figure 1, fourth and fifth lanes). Several gp195 clones selected from our library (Figure 4A) contained these smaller inserts. We conclude that under the appropriate conditions mung bean nuclease will excise the intact gp195 gene from chromosomal DNA, but that at higher nuclease concentrations, the enzyme will cleave within the gene.

Sequence of CAMP strain gp195 DNA

The complete sequence of our gp195 clone with the largest insert (a88) is shown in Figure 2. A long open reading frame begins at base 288 and continues to the end of our clone. This frame was verified to be correct by sequencing the ends of the insert in clone a50 which is located entirely within the coding region (Figure 4A) and which is expressed as a galactosidase fusion protein (19). Two regions of the sequence encode three amino acid repeats, bases 477-647 encoding primarily Ser-Gly-Thr repeats and bases 2538-2591 encoding primarily Thr-Glu-Glu repeats. Except for the common threonine codon, these two repeat regions show no similarity in sequence. Searches for hydrophobic stretches in the primary sequence identified only the very beginning of the polypeptide as potentially membrane spanning, and this region is probably a signal peptide (21,23). Portions of the gp195 molecule may be anchored to membranes by lipid attachments (30), similar to the trypanosome variable surface glycoprotein (31). Attempts to find inverted repeats or symmetric regions in the CAMP

Figure 4. Maps of the gp195 Gene and Regions of Homology. A. Sizes and positions of six CAMP gp195 clones. B. All 3.6 kb of the cloned and sequenced CAMP gene are shown with the regions of strong homology (at least 75% matching bases) to the Wellcome sequence indicated by the solid lines. Numbers under the homologous regions show the percentages of matching amino acids. The zigzag segments in the CAMP gene are the repeat regions. C. Sau 3A fragments (numbered) and Hinf I fragments (lettered) for the cloned portion of the CAMP gene.

sequence were unsuccessful.

Comparison of CAMP gp195 sequence with gp195 sequences from other strains

The primary repeat region of gp195 has now been sequenced from five different strains (21-23). As shown in Figure 3, the repeats themselves are alike in containing Ser-X-X, but differ in the second and especially third amino acids, in the pattern of repeat variants, and in the number of repeats. Sequences immediately flanking the repeats clearly fall into two distinct families. The CAMP and K1 repeat DNA sequences differ from the repeat sequences in the other family by being offset by one codon. For example, the dominant CAMP repeat of Ser-Gly-Thr is usually encoded by AGT-GGT-ACA whereas the dominant SGE2 repeat of Ser-Gly-Gly is usually encoded by TCA-GGT-GGT. The flanking sequences that distinguish the two gp195 families are in turn flanked by sequences that are virtually identical in all five strains.

The SGE2 and MAD20 sequences end shortly after the repeated region, and except for the repeats code for exactly the same protein sequence as the Wellcome strain. Both the K1 and Wellcome sequences are complete, and again

Figure 5. Hybridization of the CAMP gp195 Gene to *Sau* 3A Digests of DNA from other Strains. The strains (see reference 12 for descriptions) are indicated above each lane. Sizes of standard fragments are indicated in kb to the right of the top panel. Numbers at the left of each panel refer to the fragments shown in Figure 4C. The probe for the top panel was the entire insert from clone a88, while the probes for the bottom panels were the *Hinf* I fragments shown in Figure 4C.

with the exception of the repeat region, code for nearly identical amino acids (99% amino acid homology). In contrast, the CAMP sequence differs considerably from the Wellcome (and K1) sequence. As shown in Figure 4B, five regions of 63% or greater amino acid sequence homology between the CAMP and Wellcome strains are separated by more variable regions. The second variable region shows limited homology between the CAMP and Wellcome sequences with 36% of the amino acids being identical. Also, there was limited sequence homology within the fourth variable region just upstream of the fifth and last region of strong homology. The CAMP and Wellcome sequences were nearly identical in length except for the major repeat region which was 81 bp longer in the CAMP strain, and the fourth variable region which was 75 bp longer in the CAMP strain. At the position of the Thr-Glu-Glu repeats in the CAMP sequence, nonconserved 6 bp repeats encoding predominantly Gln-Ala and Val-Pro were found in the Wellcome and K1 sequences.

Comparison of gp195 genes by hybridization

Every *P. falciparum* strain tested (10 in all so far) has shown an *Eco*R I fragment of about 4.9 kb which hybridizes to the 5' portion of the gp195 gene and which varies slightly in size between strains (22,23,32, data not shown). However, when smaller segments of the genes from several strains were compared by hybridization, a different picture emerged. Hybridization of the insert from clone a88 to *Sau* 3A digests of the DNA from ten different strains of *P. falciparum* is shown in the top panel of Figure 5. Fragments 1,3,5, and 7-9, identified in Figure 4C, are labeled for the CAMP strain in Figure 5. Fragments 2,4 and 6 are too small to be detected. Some of the strains like the Brazilian IMTM 22 strain show strong hybridization with the CAMP probe while others like FCR 3 from Africa show only limited cross hybridization.

To confirm the fragment labeling in the top panel of Figure 5 and to investigate further the segments of the gene that can cross hybridize, *Sau* 3A digests of the CAMP, IMTM 22 and FCR 3 strains were probed with each of the three *Hinf* I fragments of the a88 insert labeled a, b, c in Figure 4C. As shown in the bottom panels of Figure 5, probe a hybridized to all three strains; reaction with the CAMP strain was probably so much stronger because of the presence in this strain of homologous repeats. Probe b also reacted with all three strains, however, probe c did not react with the FCR 3 strain.

DISCUSSION

For malaria parasites, mung bean nuclease has been extremely useful in constructing genomic expression libraries. Our results confirm that under the appropriate conditions, mung bean nuclease can excise intact genes from the chromosomal DNA. However, this phenomenon is strongly dependent on reaction conditions.

Our comparison of gp195 genes from different strains by analysis of DNA sequences and by hybridization are consistent. Strong homology is found only for portions of the molecule, especially the regions flanking the primary repeat region and the segment before the EcoR I site within the gene. All seven cysteine residues within our sequence are also found in the Wellcome and K1 sequences. Potential glycosylation sites (dots in Figure 2), however, are not conserved between strains. 99% of the 281 bp of our sequence upstream of the translation start site are identical to the Wellcome and K1 sequences.

A minimum estimate of rate of change of the gp195 variable regions over evolutionary time can be obtained by noting that in nature, P. falciparum infects only humans, not the great apes or monkeys (33), and that human ancestors diverged from the African great ape ancestors about 3.5 million years ago (34), the maximum time that two strains could have had to diverge. For the second conserved and variable regions, using this time, respective values of 0.9 and 0.05 million years per 1% amino acid sequence change are obtained. The value (unit evolutionary period) for the conserved region is low compared to other proteins, and for the variable region is extremely low (35), especially considering that these are maximum values. These results, combined with the fact that the region upstream of the start codon is 99% conserved between strains, strongly support positive selection for gp195 variants. That conserved regions are found at all may be because these regions are non-immunogenic, because immune responses against these regions do not block parasite growth or because they are functionally or structurally important for the molecule.

The mechanism of variation of the primary repeat region of the gp195 gene, the first variable region, probably differs from the mechanism for the other variable regions. Like the P. falciparum S antigen genes (13) and the P. knowlesi circumsporozoite protein genes (36), the gp195 repeat region shows between strains, strong conservation of flanking regions, yet significant variation in the actual repeat sequences. There are a number of examples of eucaryotic genes with tandem repeats within the coding regions

including collagen genes (37), wheat endosperm protein genes (38), a *Xenopus* transcription factor gene (39) and a proline rich salivary protein gene (40). However malaria parasites seem to have taken special advantage of this phenomenon; every parasite antigen gene which has been sequenced has repeats. Within either of the two families of repeat sequences (Figure 3), steps (especially deletions) for going from one sequence to another can be imagined, but between families, considering the codon shift, steps of change are unclear. Any of various theoretical mechanisms for generation and change of tandem repeats such as unequal homologous recombination, gene conversion or multiple replication initiations (41-44) could at this time explain evolution of the two families. Perhaps malaria parasites change their repeats at a high enough rate to be observed in the lab.

It is interesting that all five published gp195 sequences fall exactly into two families based on the sequences immediately flanking the repeats. Perhaps most *P. falciparum* isolates evolved from only from two ancestral clones. The K1 sequence is clearly in the CAMP family on the basis of the repeat region sequence, but the remainder of the sequence is much closer to the Wellcome strain than the CAMP strain. The simplest explanation for this result is that the K1 gp195 gene is the result of a recombination between a Wellcome-like gene and a CAMP-like gene within the conserved region immediately following the repeats.

Segments of the gp195 gene encoding the different processed products of this protein are beginning to be mapped (19,23). In particular, the 83,000 molecular weight product is encoded at the 5' end of the open reading frame. This coding segment probably extends into the third region of homology, spanning the first two variable regions. We have found that the coding region for a 45,000 molecular weight processed product spans the interior *EcoR* I site at the 3' end of our clones (19) and therefore includes the last region of strong homology.

Since immunity to *P. falciparum* blood infection is strain specific, most likely one or more protective parasite antigens vary among strains. Because monoclonal antibodies against gp195 can block growth of the parasites in culture (16,45) and recent vaccination trials with gp195 in monkeys were successful (32,46), it is a prime candidate for a variable antigen able to induce protection. Also, the likelihood of positive selection of variants during evolution argues that immune responses to gp195 are inhibitory to the parasite. If the gp195 repeat is immunodominant as is the case for the circumsporozoite protein repeat (5,47), then antibodies

against this region could block the gp195 from one strain, yet allow a gp195 from another strain with conserved functional flanks but a different repeat to go unscathed. This hypothesis is now testable for gp195; conserved flanking regions, produced in *E. coli*, can be examined for immunogenicity and for the ability of antibodies against these regions to block growth of the parasites.

ACKNOWLEDGEMENTS

We thank Lisandro Reyes, Kenneth Pinkston and Dale Edwards for technical assistance and Terry Hadley, David Haynes and Jeff Chulay for helpful discussions. This work was supported in part by the National Institutes of Health BIONET Grant (#1 U41 RR-01685-02).

REFERENCES

1. Diggs, C.L. (1980) in *Malaria*, Kreier, J. Ed., Vol 3, pp. 299-313, Academic Press, New York.
2. Miller, L.H., David, P.H. and Hadley, T.J. (1985) *Phil. Trans. R. Soc. Lond. B* 307, 99-115.
3. Boyd, M.F., Stratman-Thomas, W.K. and Kitchen, S.F. (1936) *Amer. J. Trop. Med.* 16, 139-145.
4. Wilson, R.J.M. and Phillips, R.S. (1976) *Nature* 263, 132-134.
5. Dame, J.B., Williams, J.L., McCutchan, T.F., Weber, J.L., Wirtz, R.A., Hockmeyer, W.T., Maloy, W.L., Haynes, J.D., Schneider, I., Roberts, D., Sanders, G.S., Reddy, E.P., Diggs, C.L. and Miller, L.H. (1984) *Science* 225, 593-599.
6. Koenen, M., Scherf, A., Mercereau, O., Langsley, G., Sibilli, L., Dubois, P., Pereira da Silva, L. and Muller-Hill, B. (1984) *Nature* 311, 382-385.
7. Coppel, R.L., Cowman, A.F., Anders, R.F., Bianco, A.E., Saint, R.B., Lingelbach, K.R., Kemp, D.J. and Brown, G.V. (1984) *Nature* 310, 789-792.
8. Hope, I.A., Mackay, M., Hyde, J.E., Goman, M. and Scaife, J. (1985) *Nuc. Acids Res.* 13, 369-379.
9. Stahl, H.D., Crewther, R.F., Anders, R.F., Brown, G.V., Coppel, R.L., Bianco, A.E., Mitchell, G.F. and Kemp, D.J. (1985) *Proc. Natl. Acad. Sci. USA* 82, 543-547.
10. Ravetch, J.V., Kochan, J. and Perkins, M. (1985) *Science* 227, 1593-1597.
11. Cowman, A.F., Coppel, R.L., Saint, R.B., Favaloro, J., Crewther, P.E., Stahl, H.D., Bianco, A.E., Brown, G.V., Anders, R.F. and Kemp, D.J. (1984) *Mol. Biol. Med.* 2, 207-221.
12. Weber, J.L. and Hockmeyer, W.T. (1985) *Mol. Biochem. Parasitol.* 15, 305-316.
13. Cowman, A.F., Saint, R.B., Coppel, R.L., Brown, G.V., Anders, R.F. and Kemp, D.J. (1985) *Cell* 40, 775-783.
14. Hall, R., Osland, A., Hyde, J.E., Simmons, D.L., Hope, I.A. and Scaife, J.G. (1984) *Mol. Biochem. Parasitol.* 11, 61-80.
15. Holder, A.A. and Freeman, R.R. (1984) *J. Exp. Med.* 160, 624-629.
16. Pirson, P.J. and Perkins, M.E. (1985) *J. Imm.* 134, 1946-1951.
17. Lyon, J.A., Haynes, J.D., Diggs, C.L., Chulay, J.D. and Pratt-Rossiter, J.M. (1986) *J. Imm.* in press.

18. McBride, J.S., Newbold, C.I. and Anand, R. (1985) *J. Exp. Med.* 161, 160-180.
19. Lyon, J.A., Geller, R.H., Haynes, J.D., Chulay, J.D. and Weber, J.L. (1986) *Proc. Natl. Acad. Sci. USA* in press.
20. Howard, R.F., Stanley, H.A., Campbell, G.H., Langreth, S.G. and Reese, R.T. (1985) *Mol. Biochem. Parasitol.* 17, 61-77.
21. Mackay, M., Goman, M., Bone, N., Hyde, J.E., Scaife, J., Certa, U., Stunnenberg, H. and Bujard, H. (1985) *EMBO J.* 4, 3823-3829.
22. Cheung, A., Shaw, A.R., Leban, J. and Perrin, L.H. (1985) *EMBO J.* 4, 1007-1012.
23. Holder, A.A., Lockyer, M.J., Odink, K.G., Sandhu, J.S., Riveros-Moreno, V., Nicholls, S.C., Hillman, Y., Davey, L.A., Tizard, M.L.V., Schwarz, R.T. and Freeman, R.R. (1985) *Nature* 317, 270-273.
24. McCutchan, T.F., Hansen, J.L., Dame, J.B. and Mullins, J.A. (1984) *Science* 225, 625-628.
25. Lanar, D.E., Pearce, E.J. and Sher, A. (1985) *Mol. Biochem. Parasitol.* 17, 45-60.
26. Young, R.A., Bloom, B.R., Grosskinsky, C.M., Ivanyi, J., Thomas, D. and Davis, R.W. (1985) *Proc. Natl. Acad. Sci. USA* 82, 2583-2587.
27. Young, R.A. and Davis, R.W. (1983) *Science* 222, 778-782.
28. Deininger, P.L. (1983) *Anal. Biochem.* 129, 216-223.
29. Biggin, M.D., Gibson, T.J. and Hong, G.F. (1983) *Proc. Natl. Acad. Sci. USA* 80, 3963-3965.
30. Haldar, K., Ferguson, M.A.J. and Cross, G.A.M. (1985) *J. Biol. Chem.* 260, 4969-4974.
31. Ferguson, M.A.J., Haldar, K. and Cross, G.A.M. (1985) *J. Biol. Chem.* 260, 4963-4968.
32. Hall, R., Hyde, J.E., Goman, M., Simmons, D.L., Hope, I.A., Mackay, M. and Scaife, J. (1984) *Nature* 311, 379-382.
33. Coatney, G.R., Collins, W.E., Warren, M. and Contacos, P.G. (1971) *The Primate Malarias*, U.S. Government Printing Office, Washington, D.C.
34. Hasegawa, M., Kishino, H. and Yano, T. (1985) *J. Mol. Evol.* 22, 160-174.
35. Wilson, A.C., Carlson, S.S. and White, T.J. (1977) *Ann. Rev. Biochem.* 46, 573-639.
36. Sharma, S., Svec, P., Mitchell, G.H. and Godson, G.N. (1985) *Science* 229, 779-782.
37. Bornstein, P. and Sage, H. (1980) *Ann. Rev. Biochem.* 49, 957-1003.
38. Okita, T.W., Cheesbrough, V. and Reeves, C.D. (1985) *J. Biol. Chem.* 260, 8203-8213.
39. Miller, J., McLachlan, A.D. and Klug, A. (1985) *EMBO J.* 4, 1609-1614.
40. Zeimer, M.A., Swain, W.F., Rutter, W.J., Clements, S., Ann, D.K. and Carlson, D.M. (1984) *J. Biol. Chem.* 259, 10475-10480.
41. Smith, G.P. (1976) *Science* 191, 528-535.
42. Baltimore, D. (1981) *Cell* 24, 592-594.
43. Roberts, J.M., Buck, L.B. and Axel, R. (1983) *Cell* 33, 53-63.
44. Maresca, A., Singer, M.F. and Lee, T.N.H. (1984) *J. Mol. Biol.* 174, 629-649.
45. Schmidt-Ullrich, R., Brown, J., Whittle, H. and Lin, P. (1986) *J. Exp. Med.* 163, 179-188.
46. Perrin, L.H., Merkli, B., Loche, M., Chizzolini, C., Smart, J. and Richle, R. (1984) *J. Exp. Med.* 160, 441-451.
47. Zavala, F., Cochrane, A.H., Nardin, E.H., Nussenzweig, R.S. and Nussenzweig, V. (1983) *J. Exp. Med.* 157, 1947-1957.