Vol. 146, No. 1, 1987 BIOCHEMICAL AND BIOPHYSICAL RESEARCH COMMUNICATIONS
July 15, 1987 Pages 368-377

Cloning and sequencing of Plasmodium falciparum DNA fragments containing repetitive regions potentially coding for histidine-rich proteins: identification of two overlapping reading frames

Reijer Lenstra[1], Luc d'Auriol[1], Beatrice Andrieu[2], Jacques Le Bras[2], and Francis Galibert[1]

[1]Laboratoire d'Hématologie Expérimentale, Hôpital St-Louis, F-75010 Paris, and [2]Departement de Biologie Parasitaire, Institut de Médecine et d'Epidemiologie Tropicales, Hôpital Claude Bernard, F-75019 Paris

Received June 3, 1987

SUMMARY: DNA sequences, potentially coding for histidine-rich proteins, were isolated from a P.falciparum genomic library using an oligonucleotide probe consisting of histidine codon repeats. Sequencing revealed that the different DNA fragments contain long repetitive regions very homologous to the probe. One clone was fully sequenced and contains two open reading frames that overlap in the repetitive region but are located on opposite strands. Analysis suggests that both are coding. One frame could code for a small histidine-rich protein, the other for a protein containing many aspartic acid residues. Southern blotting revealed that these sequences are conserved in all three P.falciparum strains studied.

A family of proteins with the peculiarity of being very rich in histidine was reported to be conserved in many different species of malaria parasites (1). A well known member of this family is the Histidine Rich Protein (HRP) of P.Lophurae (2), a plasmodium species which infects birds. The HRP (calculated MW=45 kD) is largely composed of repeats of a decapeptide containing eight histidine residues (3,4) and it may induce protective immunity in ducks (5). Another interesting member of the family is the knob-protein of P.falciparum (6,7). P.falciparum infects humans and provokes the most severe malarial infection in man. The pathogenicity of the parasite is thought to be largely dependent on the small protrusions, or knobs, on the cell membranes of infected red blood cells. These knobs mediate the attachment of infected red blood cells to the endothelium of postcapillary venules thereby provoking peripheral vascular obstruction (8) which may lead among others to the often lethal cerebral oedema, but also hides the parasites from the immune system and prevents their elimination by the spleen (9). Genes coding for two other histidine-rich proteins of P.falciparum have been sequenced (10,11) and

0006-291X/87 $1.50

both possess a characteristic repetitive structure. Identification of the histidine-rich proteins of P.falciparum may lead to a deeper understanding of the disease and to possibilities of immunisation against the parasite. We report here the cloning of many different genomic DNA fragments of P.falciparum likely to contain sequences coding for different histidine-rich proteins. The full sequence of one clone is shown. This clone contains an open reading frame (ORF) potentially coding for a small histidine-rich protein (calculated MW=10.7 kD) containing a histidine rich repeat. This reading frame is partially overlapped by another ORF, located on the opposite DNA strand, that may code for a protein (calculated MW=30.2 kD) containing a repeat with a high content of aspartic acid residues.

MATERIAL AND METHODS

Preparation of DNA of P.falciparum.
Malaria parasites, from three different strains (FCM-17 from Senegal, FCM-22 from Madagascar, and FCM-29 from Cameroun), were cultured in vitro (12), using deleucositized red cells. Parasite DNA was prepared from infected red blood cells by lysis in 2xSSC (SSC=0,15 M NaCl 0,015 M NaCitrate), 1% Sarcosyl,10 mg/ml proteinase K (Merck). followed by carefull extraction with phenol, chloroform and dialysis against 10 mM Tris HCl pH 7.5, 1mM EDTA.

Digestion, electrophoresis, and Southern blotting of DNA.
Digestion, electrophoresis, and Southern blotting of DNA was done as suggested by Maniatis et al.(13) .

Labelling of the oligonucleotides and M13 recombinants.
The poly-His-codon oligonucleotide was labelled with polynucleotide kinase (Boehringer) [-] and M13 recombinant phage DNA was labelled by primer extension (14).

Construction of the lambda phage library.
A recombinant lambda phage library was constructed essentially according to Maniatis et al.(13). In short, DNA of the P.falciparum strain FCM-17 was partially digested with Mbo I and fragments with a length between 15 and 20 kb. were inserted in lambda EMBL 4 arms followed by in vitro packaging (Packagene, Promega Biotec) and 10^5 recombinants were obtained.

Methods of sequencing.
Templates prepared from recombinant M13 phages were used for sequencing by the dideoxy method (15). M13 recombinant subclones carriyng inserts were used to create templates with deletions in the insert as described by Henikoff(16) allowing sequencing of both strands.

RESULTS AND DISCUSSION

To investigate whether a poly-His-codon oligonucleotide probe [CA(T/C)CA(T/C)CA(T/C)CA(T/C)CA(T/C)CA] could reveal specific DNA frements, total genomic DNA of three P.falciparum strains was digested with EcoR I or BamH I, southern blotted and probed with labelled oligonucleotide. The nitrocellulose membranes were subsequently washed with 6xSSC at 44°C and exposed overnight. The poly-His-codon oligonucleotide revealed a complex restriction pattern displaying DNA fragments differing in size and in intensity of hybridization signal (not shown). All three strains tested showed the same restriction pattern

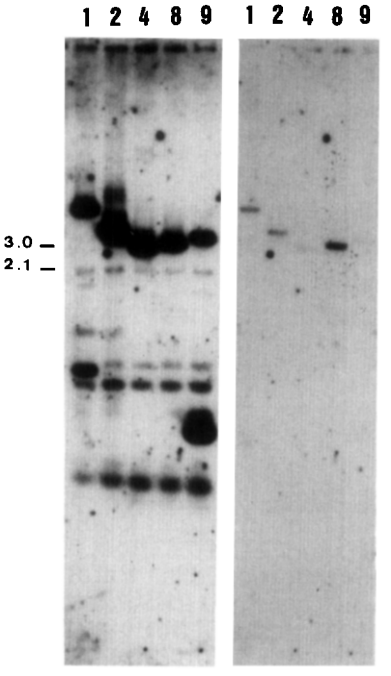

Fig.1. Southern blot of Hae III digested clones 1, 2 ,4, 8, and 9 using the poly-His-codon oligonucleotide probe. Washing was done with 6xSSC at 35°C (on the left) or at 60°C (on the right).

suggesting a specific hybridization of the probe. Since the restriction pattern was also found in two clonal derivatives of the P.falciparum strain FCM-29, the complex restriction pattern is not due to genetic variation in the population of each P.falciparum strain but reveals the existence of many conserved sequences. These findings are suggestive that the genome carries the information for many different histidine-rich proteins. Studies using metabolic labelling of P.falciparum parasites indeed revealed many proteins that are preferentially labelled with [^{3}H] Histidine (results not shown).

To characterize these sequences in more detail a genomic library was screened with the poly-His-codon oligonucleotide. After washing of the nitrocellulose replicas with 6xSSC at 44°C, 25 plaques out of $2x10^4$ gave a positive hybridization signal. The different positive clones became undetectable at various higher washing temperatures suggesting a heterogenous population of positive clones. DNA of ten positive clones was digested with various restriction enzymes, southern blotted and probed with the poly-His-codon oligonucleotide. These analyses revealed that at least 5 out of the 10 recombinant phages contained a different insert. Digestion of the 5 different recombinant DNAs with Hae III gave only one fragment per clone that can be revealed after washing with 6xSSC at 60°C. Both the length of the restriction fragment and the intensity of the hybridization signal were different for each clone (fig.1.). These Hae III

fragments were subcloned in M13 phages and partially or completely sequenced.

The complete sequence of the insert of M13 subclone M4 (derived from EMBL clone 4) is shown in fig.2.. A repetitive region stretching from base 833 to 940 shows high sequence homology with the poly-His-codon oligonucleotide, maximal alignment, with only one mismatch, is possible everywhere between position 887 to 939. The sequence possesses a very high overall AT content (AT/GC =3), a characteristic common to P.falciparum genomic DNA (17). The sequence organisation as depicted in fig.3. shows that the AT content of the sequence is not homogeneous: ORFs, which are less AT-rich, are embedded in regions that consist almost entirely of As and Ts.

One ORF (from base 824 to 1069 included), starting with an initiator codon, has the coding capacity for a small histidine-rich protein of 82 amino acids (calculated MW=10.7 kD). The reading frame is highly repetitive, and is situated after a very AT-rich upstream region but has itself a higher GC content (AT/GC=2). These are the characteristics of most of the genes of the malaria parasites. At base 1058 and 1072 the sequence shows potential donor splice sites (respectively GTAAT and GTAAA) which could bypass the stop codon at position 1073. Obvious potential splice acceptor sites, at the beginning of the downstream ORF, are not assignable. The histidine-rich gene product has a very hydrophilic and basic amino terminal part consisting of a histidine-rich repeat (see fig.4.) attached to a more hydrophobic carboxy terminal part which could function as an anchor sequence for insertion in a membrane.

The longest ORF (from base 1653 to 831 included, see fig.2.) is situated on the opposite strand. An initiator codon is situated at position 1599, so the frame may code for an aspartic acid-rich protein (calculated MW=30.2 kD). The 3' region of this ORF contains a GAT-rich repetition which overlaps the ORF for the histidine-rich protein. The amino terminal part of the aspartic acid-rich protein consists mainly of phenylanaline and is very hydrophobic. The carboxy terminal part is formed of acidic repeats due to the almost exclusive presence of aspartic and glutamic acid residues. The region between both extremities is not very repetitive but contains many lysine residues (13%), and a potential phosphorylation site (Asn-Arg-Ser) (fig.2. position 1124-1116). The lipophilic amino terminal portion could indicate transmembranal transport of the protein or anchorage of this part of the protein in a membrane.

The two ORFs use the same repetitive region but are located on opposite strands: the second base of one reading frame faces the third base in the other reading frame : ($^{1}_{1}$ $^{2}_{3}$ $^{3}_{2}$ $^{1}_{1}$ $^{2}_{3}$ $^{3}_{2}$),see fig.4.. This configuration imposes more restrictions on eventually occuring mutations than an overlap where the third base of a codon in one frame faces the third base of the codon in the opposite reading frame ($^{1}_{2}$ $^{2}_{1}$ $^{3}_{3}$ $^{1}_{2}$ $^{2}_{1}$ $^{3}_{3}$), allowing silent mutations of the third base of the codons in

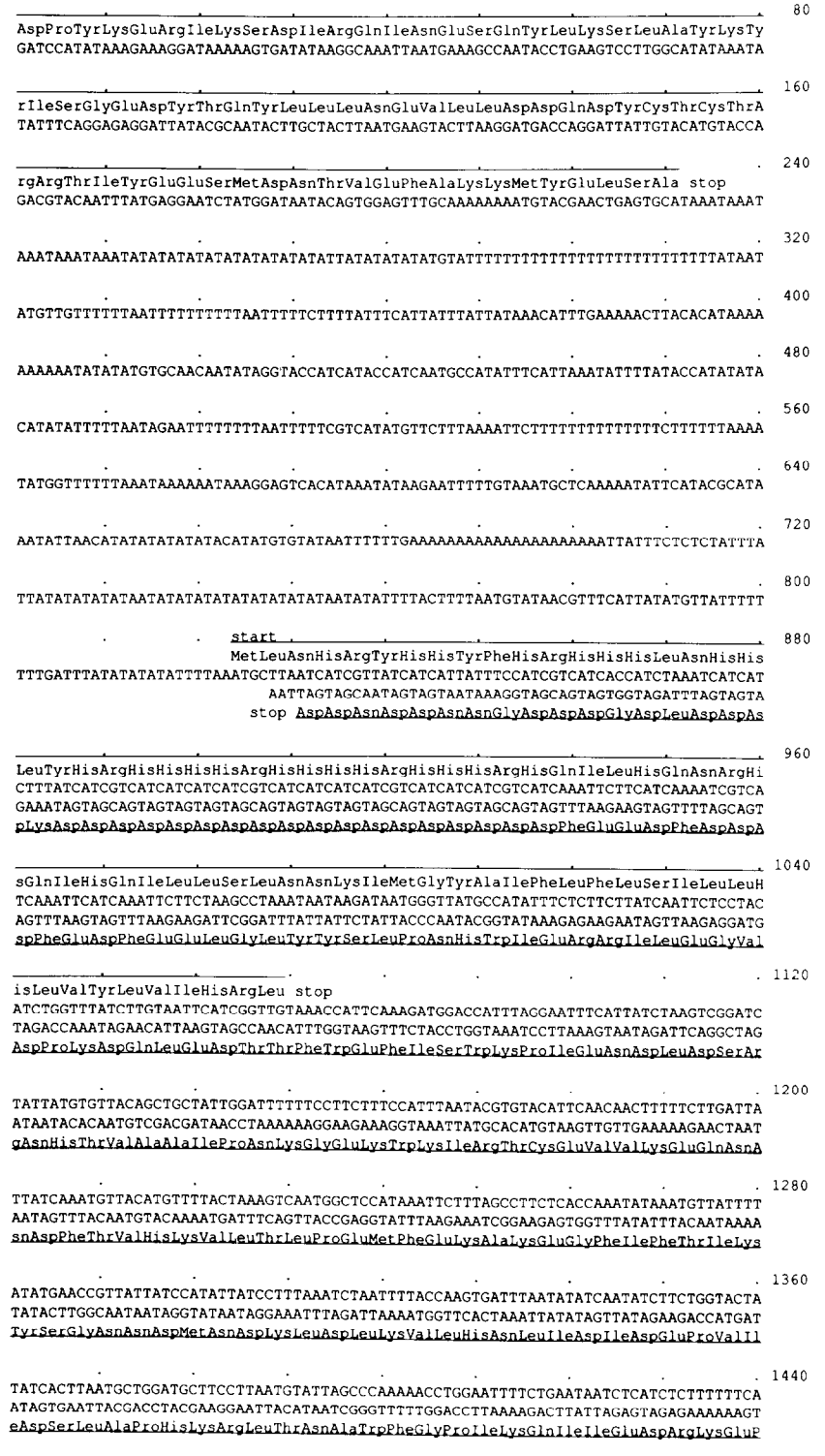

Fig. 2. The nucleic acid sequence of the Hae III fragment of the lambda recombinant clone 4. (subcloned in clone M4), and the deduced amino acid sequences of the potential coding open reading frames.

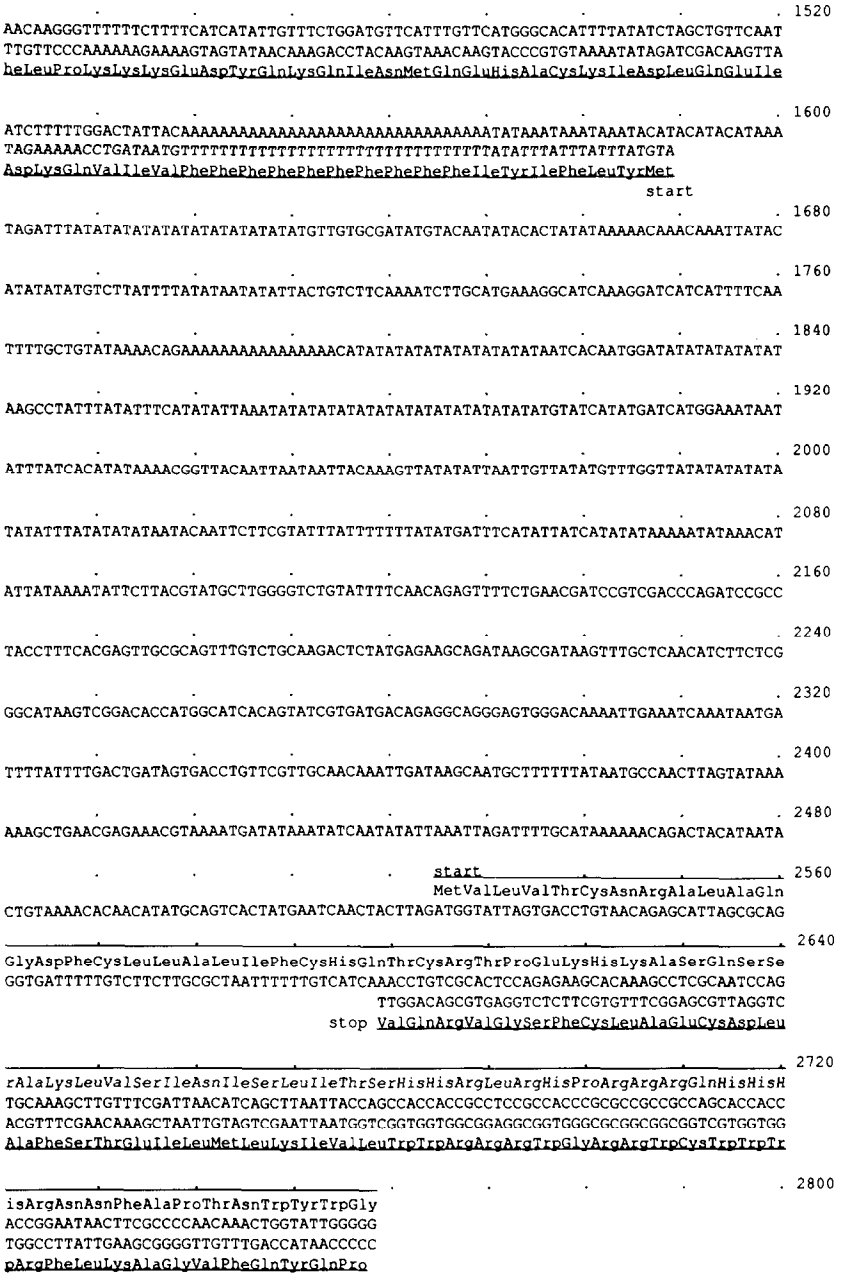

Figure 2—Continued.

both reading frames. In the actual overlap configuration all mutations result in amino acid changes in at least one reading frame. As shown further below an evolutionary pressure on both proteins seems to have restricted the mutational change in the repetitive region. Since both ORFs contain the repetitive region this region was used to probe a Northern blot and screen a cDNA library of 10^5 recombinants but no transcripts of the reading frames were detected (not shown). It is not impossible that these

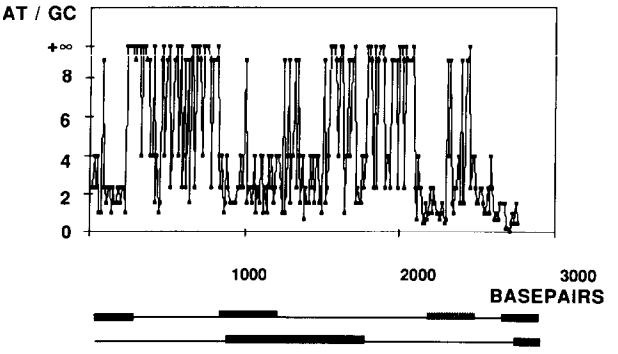

Fig. 3. Correlation between GC content of the nucleic acid sequence and the occurence of open reading frames. The AT to GC ratio from blocks of ten consecutive bases is calculated and plotted as a function of the position. The two DNA strands are represented by black lines beneath the graph, the upper line corresponds to the upper strand shown in fig. 2. The black bars indicate the potentially coding open reading frames, the gray bar indicates an open reading frame without initiator codon.

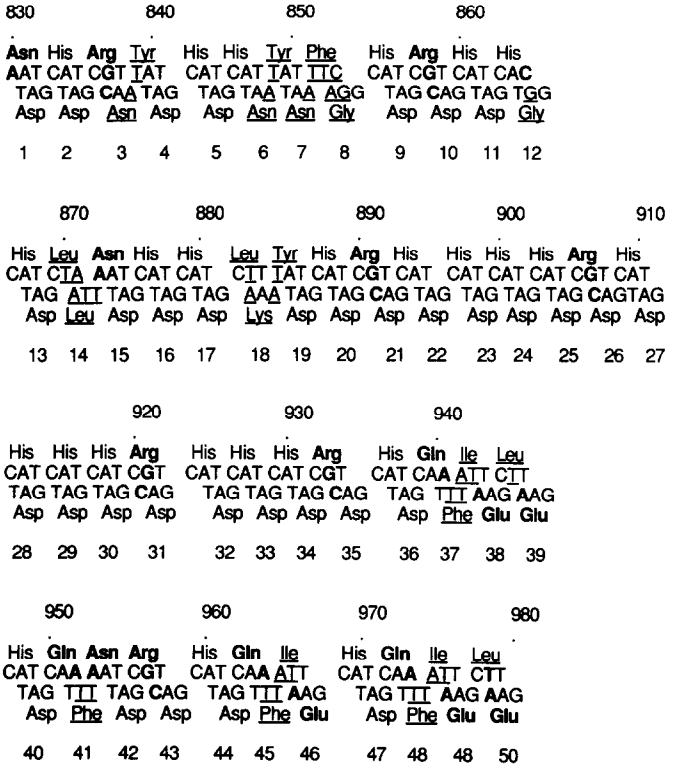

Fig. 4. Structure of the repetitive region (base 830-980) contained in the two open reading frames. The repeats are separated by some blank space. A repeat coding only for histidine (or aspartic acid on the other strand) is considered as an ancestral sequence, and conservative mutations are shown bold while all other mutations are underlined.

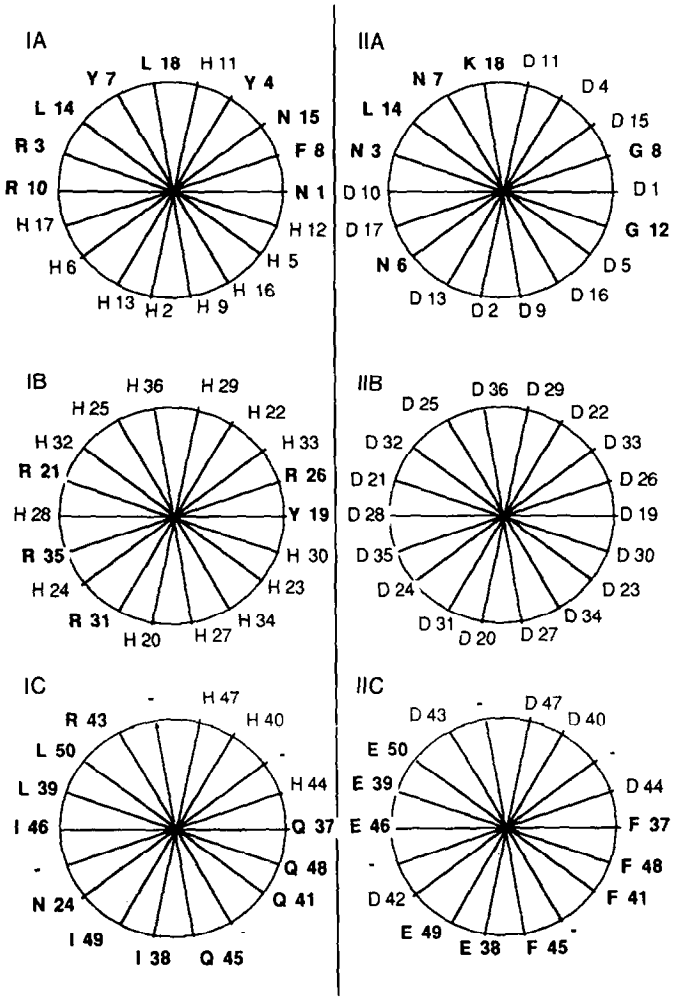

Fig. 5. Alpha helix structures of the amino acid repeats of the histidine-rich and aspartic acid-rich proteins (viewed end on). The changed amino acids (see fig.4.) are printed bold. The 'wheels' IA, IB, and 1C represent the alpha helix model of the histidine-rich repeat, the 'weels' IIA, IIB, and IIC show the aspartic acid-rich repeat. Each 'wheel' represents 5 complete helix turns, and the 'wheels' A, B and C are supperposable (position 1 on position 19 on position 37) to represent the whole repeat as one alphahelix.

genes are not expressed during in vitro culture, or that their level of transcription is to low to allow an easy detection of the transcribed RNAs. Alternatively these genes could be expressed in liver stages, in gametocytes or in mosquito stages.

As shown in fig.4. silent mutations (at positions 837, 858, 891, 907, 921, 933, and 951) in the GAT repeat of the ORF for the aspartic acid rich-protein that changed the GAT codon to GAC (both coding for Asp) changed codons in the opposite ORF from CAT to CGT resulting in a replacement of histidine by arginine in the histidine-rich protein. One silent mutation in the reading frame for the histidine-rich protein (CAT to

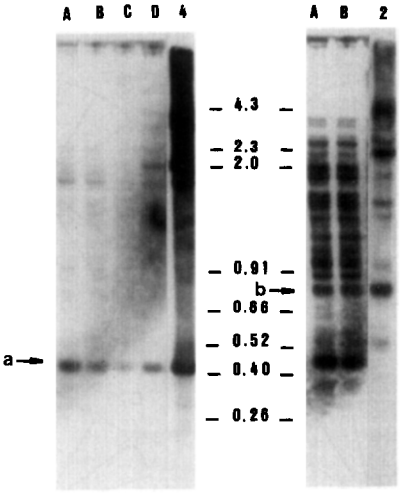

Fig.6. Southern blot of Dra I digested genomic DNA and M13 subclones M4 and M2. Lanes A,B,C,D correspond to P.falciparum strain FCM-17, FCM-29 (clone3), FCM-29 (clone1) and FCM-22 respectively; lanes 4 and 2 correspond to M13 subclones M4 and M2 On the left stringent washing (0.2 xSSc at 65°C) on the right washing with 2xSSC at 60°C.. The arrow **a** indicates the Dra I fragment containing the ORF for the histidine-rich protein gene. The arrow **b** points to a Dra I fragment of subclone M2 that possesses an open reading frame with repeats containing many histidine codons.

CAC), at position 865, resulted in a replacement of GAT by GGT in the opposite ORF, replacing a Asp by Gly in the aspartic acid-rich protein. These amino acid changes are conservative in the sense that a basic amino acid is replaced by a basic amino acid, and an acidic by an acidic amino acid (see fig.3). The amino acid replacements in both repeats result in clustering of the changed amino acids on the same side of an alpha helix structure of the repeats as is shown in fig.5. This may reflect a structural and functional role for these mutations. The clustering of all the phenylalanine residues on the same side of the Asp-rich helix results in a very hydrophobic site in the overall hydrophilic, Asp-rich repeat. The clustering of the non-histidine amino acids in the His-rich helix creates arginine-rich, glutamine-rich or hydrophobic sites which may form various epitopes that differ from the histidine-rich sites. When the histidine-rich repeat is expressed as a beta-galactosidase fusion protein the repeat is not recognized by a rabbitserum raised against a poly-histidine peptide (not shown). This indicates that the immunological properties of the repeats are codetermined by the amino acid substitutions in the repeats.

Since repeats may recombine during DNA cloning, we looked for the presence of the repetitive DNA fragment in genomic DNA of P.falciparum. Southern blotting (fig.6.) shows that the Dra I fragment containing the ORF for the histidine-rich protein, and the greater part of the ORF of the aspartic acid rich protein, is conserved in the three strains, and deletions in the repetitive region that might have occured during cloning are not detected. Subclone M2, which is derived from the EMBL clone 2., contains a

Dra I restriction fragment of the same size as one of the genomic DNA fragments recognized by the probe at lower stringency (see fig.6.) Sequence data (not shown) of this fragment show that it contains several repeats within an ORF that could code for another histidine-rich protein.

ACKNOWLEDGMENTS

We thank Drs Alan Kay and Syed Saeed Hussain for critical reading of the manuscript. R.L. is a recipient fellow of the Commission of the European Communities, division Genetics and Biotechnology. This work was supported by grants from INSERM (SC15, U13) and CNRS (LOI).

REFERENCES

1. Wallach, M. and Sarkar, A. (1984) Prog.Clin.Biol.Res. 155, 109-117.
2. Kilejian, A. (1974) J.Biol.Chem. 249, 4650-4655.
3. Ravetch, J.V., Feder, R., Pavlovec, A. and Blobel, G. (1984) Nature 312, 616-620.
4. Ravetch, J.V., Feder, R., Pavlovec, A. and Blobel, G. (1985) Nature 317, 558.
5. Kilijian, A. (1978) Science 201, 922-924.
6. Kilejian A. (1979) Proc.Natl.Acad.Sci. USA 76, 4650-4653.
7. Kilijian, A. (1980) J.Exp.Med. 151, 1534-1538.
8. Raventos-Suarez, C., Kaul, D.K. and Nagel R.L. (1985) Proc.Natl.Acad. Sci. USA 82, 3829-3833
9. Langreth, S.G. and Peterson E. (1985) Infect. Immun. 47, 760-766.
10. Stahl, H.D., Kemp, D.J., Crewther, P.E., Scanlon, D.B., Woodrow, G., Brown, G.V., Bianco, A.E., Anders, R.F. and Coppel.R.L. (1985) Nucl. Acids Res.13, 7837-7846.
11. Wellems,T.E. and Howard,R.J. (1986) Proc.Natl.Acad.Sci.USA 83, 6065-6069
12. Trager, W. and Jensen, J.B. (1976) Science 193, 673-674.
13. Maniatis, T., Fritsch, E.F., and Sambrook, J. (1982) Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.
14. Burke J.F. (1984) Gene 30, 63-68.
15. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc.Natl.Acad.Sci.USA 74, 5463-5467.
16. Hennikoff, S. (1984) Gene 28, 351-359.
17. Pollack, Y., Katzen, A.L., Spira, D.T. and Golenser J. (1982) Nucl.Acids Res. 10, 539-546.