

The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*

S. Bowman, D. Lawson, D. Basham, D Brown, T. Chillingworth, C. M. Churcher, A. Craig*, R. M. Davies, K. Devlin, T. Feltwell, S. Gentles, R. Gwilliam, N. Hamlin, D. Harris, S. Holroyd, T. Hornsby, P. Horrocks*, K. Jagels, B. Jassal, S. Kyes*, J. McLean, S. Moule, K. Mungall, L. Murphy, K. Oliver, M. A. Quail, M.-A. Rajandream, S. Rutter, J. Skelton, R. Squares, S. Squares, J. E. Sulston, S. Whitehead, J. R. Woodward, C. Newbold* & B. G. Barrell

Pathogen Sequencing Unit, Sanger Centre, Wellcome Trust Genome Campus, Hinxton CB10 1SA, UK

* Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Oxford OX3 9DS, UK

Analysis of *Plasmodium falciparum* chromosome 3, and comparison with chromosome 2, highlights novel features of chromosome organization and gene structure. The sub-telomeric regions of chromosome 3 show a conserved order of features, including repetitive DNA sequences, members of multigene families involved in pathogenesis and antigenic variation, a number of conserved pseudogenes, and several genes of unknown function. A putative centromere has been identified that has a core region of about 2 kilobases with an extremely high (adenine + thymidine) composition and arrays of tandem repeats. We have predicted 215 protein-coding genes and two transfer RNA genes in the 1,060,106-base-pair chromosome sequence. The predicted protein-coding genes can be divided into three main classes: 52.6% are not spliced, 45.1% have a large exon with short additional 5' or 3' exons, and 2.3% have a multiple exon structure more typical of higher eukaryotes.

Malaria remains a major burden to human health in tropical and subtropical areas. In Africa alone, more than one million children under five die from the disease each year¹. Although four members of the *Plasmodium* genus normally infect humans, nearly all deaths are attributable to a single parasite species, *P. falciparum*. The severity of disease caused by this species results primarily from its ability to modify the surface of infected red blood cells by inserting parasite proteins. Parasitized erythrocytes can bind to host endothelial cells, a process called cytoadherence, leading in some cases to their accumulation in specific organs such as the brain and to the development of cerebral malaria². Current approaches to malaria control and treatment rely on measures such as insecticide-impregnated bed nets and chemotherapy. Although considerable resources have been devoted to the development of a vaccine, no effective immunization regime so far exists. Moreover, the rapid spread of resistance to existing and new antimalarial drugs means that in some areas of the world, particularly southeast Asia, reliable prophylaxis is not possible, making treatment difficult³.

In response to these problems, the Malaria Genome Sequencing Consortium⁴ was established to sequence the entire genome of *P. falciparum* as a collaborative venture. In less than a year, the consortium aims to generate almost all of the *P. falciparum* genome sequence in unfinished form, making nearly its entire gene complement accessible for malaria researchers. Focus is already shifting to the development and implementation of whole genome approaches to drug development and vaccine target identification.

Sequencing strategy

P. falciparum has a nuclear genome of around 30 megabases (Mb) divided between 14 chromosomes, which range in size from 0.7 to 3.5 Mb. Sequencing this genome presents significant technical difficulties, mainly because of the biased nucleotide composition of its DNA, with an overall (A + T) content estimated at 82% (ref. 5). In general, fragments of DNA over 5 kilobases (kb) are unstable in *Escherichia coli*, so the large-insert bacterial clones commonly used as templates for sequencing are not available. However, most *P. falciparum* chromosomes can be resolved by pulsed-field gel

electrophoresis (PFGE), and some mapped yeast artificial chromosomes (YACs) exist. Chromosome 3 was sequenced using a whole chromosome 'shotgun' (WCS), an approach that was pioneered during the sequencing of *Saccharomyces cerevisiae* chromosome IX (ref. 6), and which was also used for *P. falciparum* chromosome 2 (ref. 7). Management of the WCS was facilitated by generating several-hundred sequence reads from each of the YACs^{8,9} comprising the minimal tiling path for this chromosome, and using these to place the WCS into a series of discrete bins. A strategy of sequencing YAC clones exclusively was discarded because of the high frequency of chimaerism, deletions and rearrangements inherent to YAC libraries¹⁰. In addition, although the YACs were purified through two pulsed-field gels¹¹, the level of contaminating *S. cerevisiae* sequence remained high owing to preferential cloning of yeast DNA.

Assembly validation

Because of the complexity associated with the assembly of an entire chromosome containing extensive repetitive regions, we confirmed

Figure 1 *P. falciparum* chromosome 3. Order and orientation of genes and predicted genes are shown. Exons are shown as coloured boxes with introns as linking lines. The two tRNA genes are shown as purple boxes with gene names shown in purple. Genes encoding proteins that have been characterized previously in *P. falciparum* have names shown in red. Predicted genes encoding proteins that have similarities to proteins currently unique to apicomplexan species are shown in fuchsia. Predicted genes encoding proteins with similarity to proteins in other eukaryotes for which some functional information is available are shown in yellow. Predicted genes encoding proteins with similarities to proteins of unknown function in other organisms are shown in orange. Predicted genes encoding proteins with similarity solely to bacterial proteins are shown in light blue. Genes encoding predicted proteins having similarity only within a defined protein domain are shown in grey. Predicted genes having no significant similarities are shown in dark green with those that have been confirmed by RT-PCR shown in light green. Pseudogenes are shown in dark blue. N-terminal signal sequences are represented by wavy lines. Repetitive telomeric sequences are shown as hatched boxes, and the location of the predicted centromere (put.CEN) is shown as a red box.

Table 1 Preliminary classification of chromosome-3 proteins

(a)			
Gene name	Description	Gene name	Description
Metabolism		Transport	
PFC0050C	Long chain fatty acid CoA ligase	PFC0125W	ABC transporter
PFC0170C	Lipoamide acyltransferase	PFC0725C	Formate transporter
PFC0275W	Glycerol-3-phosphate dehydrogenase	PFC0840W	E1-E2 P-type ATPase
PFC0395W	Asparagine synthetase	Cell surface	
PFC0710W	Inorganic pyrophosphatase	PFC0005W	<i>var</i> (3D7- <i>var</i> T3-1)
PFC0830W	Triosephosphate isomerase	PFC0010C	<i>rifin</i> (3D7- <i>rif</i> T3-1)
PFC0935C	<i>N</i> -glucosamine-1-phosphate transferase	PFC0025C	<i>stevor</i> (3D7- <i>stevor</i> T3-1)
PFC0950C	Acylaminoacyl peptidase (APH)	PFC0030C	<i>rifin</i> (3D7- <i>rif</i> T3-2)
PFC0995C	Acyl-CoA, cholesterol acyltransferase (ACAT)	PFC0035W	<i>rifin</i> (3D7- <i>rif</i> T3-3)
Cell growth, division, DNA synthesis		PFC0040W	<i>rifin</i> (3D7- <i>rif</i> T3-4)
PFC0305W	Putative microtubule binding protein	PFC0110W	Cytoadherence-linked asexual protein (3D7 clag-3.2)
PFC0340W	DNA polymerase delta, small subunit	PFC0120W	Cytoadherence-linked asexual protein (3D7 clag-3.1)
PFC0385C	Serine/threonine protein kinase	PFC0210C	Circumsporozoite (CS) protein
PFC0525C	Glycogen synthase kinase	PFC0640W	CTRP
PFC0595C	PP2A phosphatase	PFC0800W	Putative band 7 protein (stomatin)
PFC0755C	CDC2-related protein kinase	PFC1095W	<i>rifin</i> (3D7- <i>rif</i> T3-5)
PFC0770C	Kinesin-related protein	PFC1100W	<i>rifin</i> (3D7- <i>rif</i> T3-6)
PFC0860W	Kinesin-related protein	PFC1105C	<i>stevor</i> (3D7- <i>stevor</i> T3-2)
Transcription and post-transcriptional modification		PFC1115C	<i>rifin</i> (3D7- <i>rif</i> T3-7)
PFC00155C	DNA-directed RNA polymerase, 14K subunit	PFC1120C	<i>var</i> (3D7- <i>var</i> T3-2)
PFC0805W	DNA-directed RNA polymerase II, largest subunit	Intracellular trafficking	
PFC0825C	Cleavage and polyadenylation specificity factor	PFC0135C	Chromosome region maintenance protein
Protein synthesis		Cellular organization / biogenesis	
PFC0531C	tRNA-Val	PFC0920W	Histone H2A variant
PFC0532W	tRNA-Ile	Cell rescue, defence, death and ageing	
PFC0200W	60S ribosomal protein L44	PFC0205C	Glutaredoxin
PFC0225C	Elongation factor TS (EF-TS)	PFC0250C	AP endonuclease
PFC0290W	40S ribosomal protein S23	PFC0975C	Cyclophilin
PFC0295C	40S ribosomal protein S12	Unknown function	
PFC0300C	60S ribosomal protein L7	PFC0150W	Homologue of human protein KIAA0249
PFC0400W	60S acidic ribosomal protein P2	PFC0190C	Homologue of <i>D. melanogaster</i> PAST protein
PFC0470W	valyl-tRNA synthetase	PFC0375C	Homologue of <i>C. elegans</i> T08A11.2 protein
PFC0535W	60S ribosomal protein L26	PFC0390W	Putative homologue of <i>C. elegans</i> Y48E1C.2 protein
PFC0635C	Eukaryotic translation initiation factor 4E (EIF-4E)	PFC0410W	Putative homologue of rat brain YT51 protein
PFC0735W	40S ribosomal protein S15A	PFC1025C	Homologue of <i>C. elegans</i> F49C12.11 protein
PFC0775W	40S ribosomal protein S11	PFC1075W	Homologue of PFB0980W
PFC0870W	Elongation factor 1-beta	PFC1080C	Homologue of PFB0985C
PFC1020C	40S ribosomal protein S3A	PFC1085C	Homologue of PFB0990C
Protein destination		PFC1090W	Homologue of PFB0995W
PFC0140C	<i>N</i> -ethylmaleide-sensitive fusion protein NSF		
PFC0255C	Ubiquitin-conjugating enzyme E2		
PFC0285C	T-complex protein, beta subunit		
PFC0310C	ATP-dependent CLP protease		
PFC0350C	T-complex protein, Eta subunit		
PFC0495W	Aspartyl protease		
PFC0520W	26S proteasome regulatory subunit S14		
PFC0745C	Proteasome component C8 (macropain subunit 8)		
PFC0855W	Ubiquitin conjugating enzyme E2-17K		
PFC0900W	T-complex protein I, epsilon subunit		

(b)	
Pfam family	Gene name
Zinc-finger (C3HC4) protein	PFC0510W, PFC0740C
ATP-dependent RNA helicase	PFC0440C, PFC0915W, PFC0955C
PDZ domain protein	PFC0330W, PFC0785C
Serine/threonine protein kinase	PFC0060C, PFC0105W, PFC0485W
Guanine-nucleotide-binding protein	PFC100C, PFC0365W
Ankyrin repeat protein	PFC0160W
Calcium-dependent protein kinase	PFC0420W
Dual-specificity protein phosphatase	PFC0380W
Alpha beta hydrolase	PFC0065C
RNA-binding protein	PFC0865W

a. Gene names and predicted function of the product encoded are listed, with protein classification adapted from that used for *S. cerevisiae*. **b.** Proteins encoded on chromosome 3 that contain a domain identified in existing protein families, but which currently cannot be given a functional classification.

the chromosome 3 sequence by several independent methods. As double-stranded clones were used, with reads produced using both forward and reverse primers, the consistency of the read pairs generated was used as an initial assembly check. Reads derived from mapped YAC clones allowed confirmation of the colinearity of the chromosome assembly and the chromosome 3 YAC map⁸. In addition, the restriction enzyme map generated for this chromosome during the mapping project was used as further confirmation that the sequence had assembled correctly⁸. The order of mapped sequence-tagged site markers and simple sequence-length poly-

morphism microsatellite markers generated from the HB3xDd2 linkage segregation genetic map¹² were also confirmed in the final assembly. Finally, the DNA sequence correlated well with the restriction enzyme pattern for chromosome 3 generated by optical mapping¹³, with an average error of ~5%.

Analysis

The 1,060,106-bp chromosome 3 sequence encodes 215 predicted proteins and two tRNAs (Fig. 1), yielding a mean gene density of one predicted gene every 4.8 kb, which is similar to that in

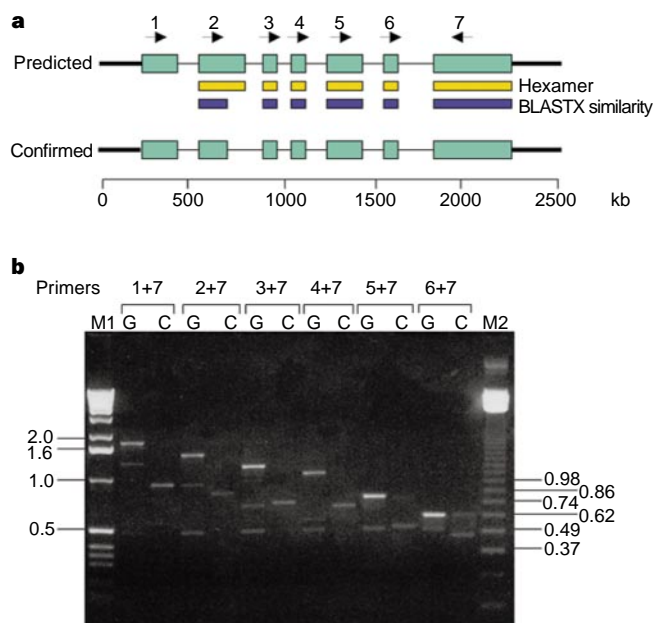


Figure 2 Splicing in the gene PFC0935C. **a**, The original gene prediction is shown, generated using hexamer and GENEFINDER, with exons drawn as coloured boxes and introns as thin lines, together with the supporting hexamer prediction of protein-coding potential and BLASTX similarity data. This is compared with the *in vivo* splicing of this gene, confirmed by RT-PCR and DNA sequencing. Exon 1 is not supported by either protein similarity or hexamer prediction, and exon 2 is incorrectly predicted by the hexamer program. Primers used for exon confirmation are shown as numbered arrows. **b**, A gel showing comparison of PCR products generated from genomic (G) and cDNA (C) template, with primer combinations shown above. Combinations 1 + 7 and 2 + 7 gave slightly smaller products than predicted, an observation confirmed by sequence data. DNA size markers, the 1-kb ladder (lane M1) and the 123-bp ladder (lane M2) are shown in kb.

Caenorhabditis elegans (one gene per 5 kb)¹⁴. The overall (A + T) composition of this chromosome is 80%, with the base composition of exons being 76.8% (A + T) and introns being 84.6% (A + T). Six of the genes on chromosome 3 had been characterized prior to this work. They encode circumsporozoite protein (CSP)^{15,16}, CS protein-TRAP-related protein (CTRP)¹⁷, RNA polymerase II largest subunit¹⁸, elongation factor-TS, CDC2-related protein kinase (EMBL accession no. EM:M86715; TrEMBL accession no. TR:Q25028) and cyclophilin¹⁹, of which only three had been mapped to this chromosome¹⁵⁻¹⁸. Ninety-four of the predicted proteins (43.7%) have significant similarity to existing database entries. Eighty-five (39.5%) have matches to eukaryotic proteins, many providing potential functional information. Nineteen of these (8.8%) are currently unique to *P. falciparum* or other apicomplexan parasites, but few proteins in this group have been functionally characterized. Five predicted proteins (2.3%) have significant similarity solely to bacterial proteins and are likely to be localized to the organelles of the parasite. In total, 27.4% of the predicted proteins on chromosome 3 are members of Pfam protein families²⁰. A further four predicted proteins contain discrete protein domains, but similarity does not extend further than the domain identified (Fig. 1). A preliminary functional classification of the proteins encoded on chromosome 3 is shown in Table 1.

Most of the predicted proteins (202 proteins; 94.0%) contain low-complexity, non-globular regions, as defined by the seg program²¹. However, the percentage of residues defined as low-complexity is 21.6%, indicative of the small size of many of these regions. Such regions have been previously reported⁷ and are often polymorphic between parasite isolates in both housekeeping genes¹⁸ and genes identified by antibody screening^{15,16}. Regions of low

Table 2 Summary of predicted features on *P. falciparum* chromosome 3 and comparison with chromosome 2 (ref. 7)

	Chromosome 3	Chromosome 2
Length (kb)	1,060	945
No. of predicted proteins	215	209
No. of tRNA genes	2	1
Gene density (kb per gene)	4.8	4.5
Predicted genes with introns (%)	47.4	43.1
Percentage (A + T)		
Overall	80.0	80.3
Exons	76.8	75.7
Introns	84.6	86.7
Predicted protein features		
Signal peptides	50 (23%)	47 (22%)
At least one transmembrane domain	86 (40%)	90 (43%)
Multiple transmembrane domains	48 (22%)	27 (13%)
Coiled-coil domains	106 (49%)	111 (53%)
Low-complexity regions	202 (94%)	155 (74%)
Completely low complexity	0 (0%)	17 (8%)
Detectable homologues in other species	71 (33%)	87 (42%)

Genes on chromosome 3 have been predicted using different algorithms from those used for chromosome 2, and analysis of predicted proteins has been performed using different programs and parameters in some cases.

complexity can be divided into two distinct classes: tandem arrays of repeated peptide motifs (such as CSP and CTRP) and homopolymer runs of a single amino-acid residue (such as RNA polymerase II largest subunit and asparagine synthetase). The homopolymer runs represent expansions of amino acids with A/T-rich codons, encoding asparagine, lysine and glutamic acid, and are the predominant type of peptide polymorphism between parasite isolates.

Before the genome project our understanding of gene organization in *P. falciparum* was limited, as few transcriptional units had been characterized^{22,23}. Cloning and sequencing methodologies tended to encourage identification of genes that were highly immunogenic. Most of these genes had a single exon, whereas the remainder had small additional 5' or 3' exons. Analysis of the chromosome 2 and 3 sequences indicates that splicing is a much more common phenomenon in *P. falciparum* than originally thought. Nearly half of the genes on chromosome 3 are predicted to contain at least one intron (102; 47.4%). For 31.9% of these, intron splicing has been confirmed by similarity data to expressed sequence tag (EST) clones, comparative analysis with orthologous genes or by directly using polymerase chain reaction with reverse transcription (RT-PCR). For the majority of spliced genes on chromosome 3 (63 genes), splicing is predicted to be confined to the 2-exon model. Half of the 2-exon gene predictions have a 5' exon length of <100 bp, making the accurate prediction of their initiation ATG codons particularly difficult.

Several predicted genes consist of multiple small exons (up to 15) exhibiting a gene structure more similar to that of higher eukaryotes. We were able to identify five genes of this type on chromosome 3 owing to their similarity to genes in *P. falciparum* or other organisms. PFC0495W is similar to *Eimeria tenella* aspartyl protease; PFC0935C, N-acetylglucosamine-1-phosphate transferase, is most similar to its mouse homologue; and PFC0410W is most similar to the YT51 protein expressed in rat brain. The remaining two genes in this category, PFC0110W and PFC0120W, are members of a *P. falciparum* multigene family that encodes proteins implicated in cytoadherence²⁴. These cytoadherence-linked asexual gene (*clag*) paralogues (*clag* 3.2 and *clag* 3.1, respectively) appear to have arisen from gene duplication; other single copies of *clag* have been localized to chromosomes 2, 4 and 9 (ref. 24). Deletion of the *clag* gene on chromosome 9 abolishes cytoadherence, indicating either that the *clag* genes on chromosome 3 are not functionally equivalent to it, or that their transcription is subject to higher order regulation²⁴. Expression of PFC0410W, PFC0495W, PFC0935C

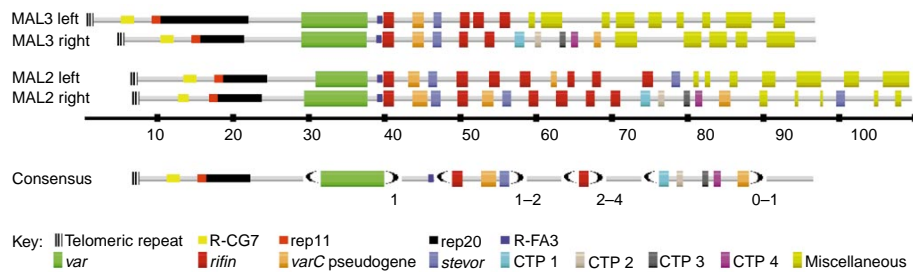


Figure 3 Telomere organization in *P. falciparum*. **a**, The telomeres of chromosomes 2 and 3, showing the higher-order organization of repetitive DNA sequences and telomere-associated multigene families. Four proteins of unknown function encoded on chromosome 3 are similar to proteins encoded on chromosome 2 and have conserved order and orientation (conserved

telomere-encoded proteins: CTPs). These are shown in light blue (PFC1090W/PFB0995W), beige (PFC1085C/PFB990C), grey (PFC1080C/PFB0985C) and fuchsia (PFC1075W/PFB0980W). **b**, A consensus for the arrangement of repetitive DNA sequences and multigene families on *P. falciparum* telomeres based on the telomeres of chromosomes 2 and 3.

and PFC0120W in the asexual blood stages has been confirmed by RT-PCR, and their predicted splice sites corrected by sequencing the cloned RT-PCR products²⁴ (Fig. 2, and data not shown). The complex nature of splicing in this type of gene, and their short exons, makes them difficult to predict in the absence of similarity data; they are predicted inefficiently using the hexamer program (R. Durbin, unpublished software. Documentation, code and data are available from anonymous ftp servers at ftp.sanger.ac.uk/pub/hexamer.; Fig. 2). It is likely that this class of *P. falciparum* genes remains significantly underpredicted from genomic sequence. Sequence data from the RT-PCR products is being used to re-train the current gene-prediction software, but generation of more *P. falciparum* EST sequences or full-length complementary DNA libraries would also facilitate identification of these genes.

We compared the gene-prediction statistics for chromosomes 2 (ref. 7) and 3 (Table 2). As expected, the chromosomes are similar in many respects, including their overall (A + T) composition and coding density. However, the number of genes predicted as having introns is higher for chromosome 3. RT-PCR experiments for some chromosome-3 genes (results not shown) indicate that splicing in *P. falciparum* is actually under-predicted by our current gene-finding methods. Because the initial analysis of chromosome 2 predicts even fewer splicing events⁷, it is likely that splicing has also been underpredicted from the chromosome 2 sequence. It is essential that gene predictions are experimentally confirmed by RT-PCR to generate a larger training set with which to improve gene-finding algorithms.

Direct comparison of the proteins predicted on chromosomes 2 and 3 is hampered because of the different methods used in generating those predictions and the different programs and parameters used in their analysis. For example, the differences observed between these two chromosomes, when comparing predictions of low-complexity sequence, are due at least in part to a different definition of low complexity being used to define the parameters for analysis. Improvement and standardization of gene-finding and protein-analysis methodologies, with subsequent re-analysis of the data from chromosomes 2 and 3, will allow a more accurate comparison of protein features.

Protein targeting

In addition to its nuclear genome, *P. falciparum* contains two organellar genomes, thought to have been acquired as a result of multiple endosymbiotic events. The mitochondrial genome is a 5.9-kb linear molecule present as multiple tandem repeats. A second organellar genome, a 35-kb circular DNA molecule, is located within the apicoplast²⁵. This is an organelle of plastid origin, thought to be unique to apicomplexan parasites, which apparently provides essential metabolic functions²⁶. Gradual loss of genes from the organellar genomes has occurred, such that maintenance of both organelles requires many nuclear-encoded proteins targeted to that

organelle using amino-terminal sequences²⁷. Three predicted proteins on chromosome 3 have N-terminal sequences and implied biological function indicating mitochondrial import (PFC0170C, PFC0225C and PFC0275W). Two predicted proteins have pre-sequences that indicate targeting to the apicoplast (PFC0050C and PFC0310C). An additional three proteins are likely to be localized to an organelle, a conclusion based on their N-terminal sequences and predicted function; however, their precise location cannot be determined from their signal peptides (PFC0470W, PFC495W and PFC725C).

P. falciparum also directs several proteins to the cytoplasm, cytoskeleton and plasma membrane of the infected red blood cell. The proteins implicated as receptors involved in cytoadherence fall into this category; however, the sequences responsible for targeting these molecules remain unidentified.

Telomere structure

Non-coding sequences. The chromosome sequence contains 39 copies of the telomeric repeat sequence (T(G/A)ACCC) at the left telomere and 85 copies at the right telomere, but no attempt has been made to estimate the exact number of copies of this repeat in the intact chromosome. Other repeat sequences (R-CG7, rep11, rep20) occur between the telomere and the gene most proximal to the telomere (*var*). The rep11 sequence is a new repeat family consisting of an 11-bp tandem repeat located immediately telomeric to the rep20 sequences. The R-FA3 repeat sequence²⁸ maps between the *var* gene and the adjacent *rif* gene.

Coding sequences. Of the known multigene families, there are two *var* genes^{29,30}, one at each telomere, four members of the *rif* gene family on the left and three on the right chromosome arm and one copy of the *stevor* family at each telomere. Two members of the *clag* family²⁴ occur in the left subtelomeric region, separated by a region with similarity to *var*, possibly representing the site of a previous recombination event. In addition, several pseudogenes occur at both telomeres that have homology to the *var* 3' exon (the *varC* genes)³¹. The preservation of these pseudogene sequences in the apparently rapidly evolving subtelomeric regions indicates that they may be biologically significant.

The most telomere-proximal sections of all four *P. falciparum* telomeres sequenced to date show a conserved order of repetitive DNA sequences and multigene family members (Fig. 3). In addition, the right telomere of chromosome 3 shows an extended region of similarity with the right telomere of chromosome 2 (ref. 7; Fig. 3). As well as members of known multigene families, there are several predicted genes that share similarity located on each of these chromosomes, which may represent new telomere-associated multigene families (PFC1075W–PFC1090W and PFB0980W–PFB0995W). The cellular location of proteins encoded by the members of these new multigene families is unknown, although we predict that all eight proteins have N-terminal signal peptides.

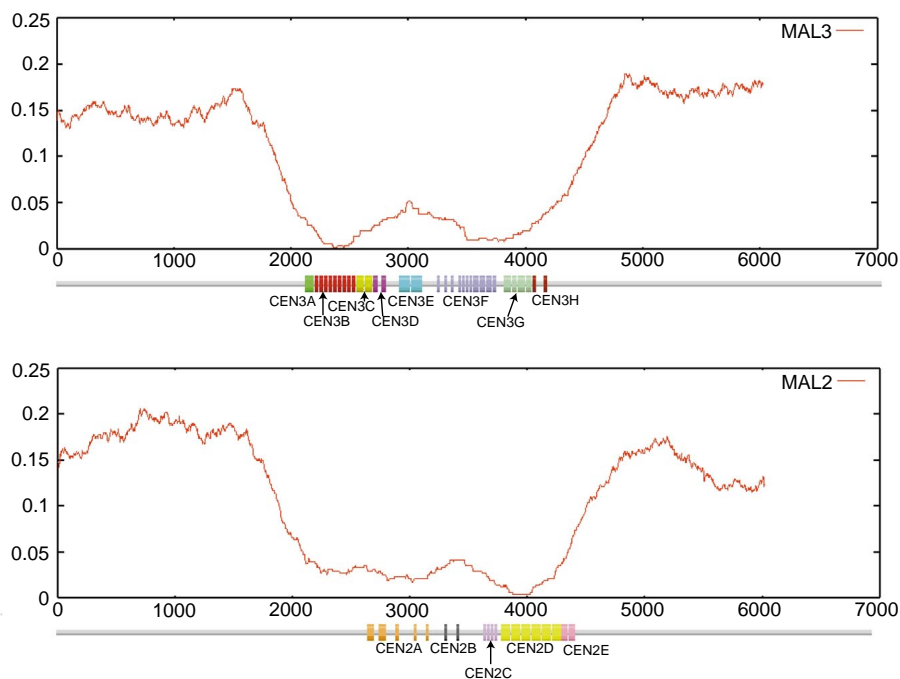


Figure 4 Putative *P. falciparum* centromeres. Plots of (A + T) composition for chromosomes 3 and 2. The Y axes are graduated in units of percentage (G + C) content, ranging from 0% (G + C) to 25% (G + C); the X axis represents DNA sequence in bp. **a**, (A + T) composition of chromosome 3 highlights a region that is extremely (A + T)-rich. Families of repetitive sequences in the predicted

centromere of chromosome 3 are shown as coloured boxes. **b**, A similar region is present on chromosome 2, containing different families of tandem repeat sequences. Alignments of the repeat families can be found at [ftp://www.sanger.ac.uk/Projects/P_falciparum/Centromeres/](http://www.sanger.ac.uk/Projects/P_falciparum/Centromeres/).

Both *rif* and *var* are expressed on the surface of infected red blood cells^{32,33} and undergo antigenic variation, probably as a means of immune evasion³⁴. The products of the *var* locus mediate cytoadherence to a variety of cellular receptors, and thus are important virulence factors. All three of the previously identified multigene families show extreme sequence polymorphism. The clustering of genes sharing similar cellular location and control is likely to be of mechanistic significance and indicates that the newly identified members of the cluster should be studied further. In addition, the regions of shared similarity between the telomeres of chromosomes 2 and 3 indicate that recombination within these regions may be frequent, and could partly explain the extensive polymorphism seen in members of the *stevor*, *rif* and *var* families.

Centromere

No *P. falciparum* centromere has been identified. Centromeres in other eukaryotic species^{35,36} can be divided into two distinct classes: the point centromere seen in budding yeasts, such as *S. cerevisiae*, and the regional centromeres present in *Schizosaccharomyces pombe* and higher eukaryotes. The minimal functional unit of the *S. cerevisiae* centromere consists of an ~125-bp region comprising three parts, the outer CDEI and CDEIII domains that flank a central (A + T)-rich region. *S. cerevisiae* centromeres are located in chromosomal regions of relatively low gene density and high (A + T) composition. The regional class of centromere has been most comprehensively characterized in the fission yeast *S. pombe*. Regional centromeres have characteristic complex arrays of repetitive sequences occurring over an extended region. The three centromeres of *S. pombe* span regions of 40–100 kb, each having a central (A + T)-rich core flanked by chromosome-specific repeat sequences^{37,38}. Sequence in regional centromeres is very variable and this has hampered attempts to find functional centromere sequences in higher eukaryotes³⁵.

We have identified a region on chromosome 3 that is currently the best candidate for the chromosome centromere. This region is

extremely (A + T)-rich, 97.3% over 2.6 kb, with a central region having a slightly higher (G + C) content (Fig. 4a). Analysis of *P. falciparum* chromosome 2 reveals a region with similar structure (Fig. 4b). Both regions have very low coding potential; the nearest predicted genes are 12 kb apart on chromosome 3 (PFC0610c and PFC0615w) and 9 kb apart on chromosome 2 (ref. 7) (PFB0490c and PFB0495w). If these regions are chromosome centromeres, they have a structure more characteristic of regional rather than point centromeres. However, they would represent very compact regional centromeres, with an extremely short core sequence. Analysis of this region on both chromosomes highlights several chromosome-specific tandem repeats (Fig. 4).

Early release of sequence data

Even before this chromosome was completed, several groups had demonstrated the utility of timely release of sequence data in unfinished form. During the course of the project, chromosome 3 sequences were used to confirm chromosome location and sequence data generated independently¹⁹, to identify new members of *P. falciparum* gene families³⁹ and to identify members of families of genes conserved in other genera⁴⁰. □

Methods

Sequencing. *P. falciparum* DNA is denatured at relatively low temperatures because of its extreme (A + T) content⁵, so we had to adapt many standard protocols for this project. DNA used in library preparation was not exposed to either ethidium bromide or ultraviolet (UV) transillumination and we minimized the temperatures to which the DNA was exposed. Gel slices containing chromosome 3 were excised from pulsed-field gels and DNA was extracted from low-melting-point agarose using a modified agarase protocol. After equilibrating with TE buffer, pH 7.4, for several hours, gel slices were loaded into a 2-ml syringe and forced through a 26G needle. Agarase buffer was added to a final concentration of 1×, and 8U β-agarase was added per ml of sample. After incubating at 37 °C for 3 h, the sample was extracted with TE-buffered phenol and the DNA recovered by ethanol precipitation. Libraries were

prepared by fragmenting the DNA by sonication and cloning into pUC18. Sequences were generated using pUC clones with both forward and reverse primers using dye-terminator chemistry.

Assembly. Because of the size and complexity of the project, we modified the standard strategy used for sequence assembly. Initially, sequences generated by the WCS and the YAC skims were assembled using the Phrap assembly program (P. Green, unpublished software), which had been adjusted to handle the large number of reads generated. We divided the chromosome into eight sections based on the YAC map, each section having a separate working database. Contigs that did not contain YAC reads were directed to a repository database, from which they could be recovered once their location had been identified. We expected to have a large number of single reads in the assembly that originated from other *P. falciparum* chromosomes, as the library generated for chromosome 3 was estimated as being 87% pure, based on the hybridization of shotgun reads to chromosome blots. Thus, single reads were not incorporated into any of the databases, but again they could be recovered if necessary. In collaboration with the other laboratories contributing to the Malaria Genome Sequencing Project, the reads generated during the chromosome 3 project that originate from other *P. falciparum* chromosomes will be incorporated into their respective chromosomes as the sequencing project progresses.

Closure. Gaps in the initial assembly were filled by several methods. Initially, oligonucleotide walking from pUC clones bridging contigs was used to fill many of the gaps. A second approach was to perform combinatorial PCR between contigs mapped to the same chromosome region by YAC-derived reads. Products generated by combinatorial PCR were sequenced by oligonucleotide walking. For gaps that could not be filled by PCR, further pUC clones proximal to the gaps were selected by hybridization of a gridded 15x chromosome 3 pUC library to radiolabelled oligonucleotides selected at contig ends⁴¹. Regions of extreme (A + T) composition were resolved by either generating transposon libraries or cloning as very small fragments (50–500 bp) into m13mp18.

Analysis. Completed sections of chromosome 3 were subjected to a series of automatic analyses to reveal possible protein-coding (R. Durbin, unpublished software; P. Green, & L. Hillier, unpublished software) and tRNA⁴² genes, similarities to ESTs⁴³, other proteins^{44,45} and repeat/multigene families. The results were collated in a genome database (ACeDB) that merges overlapping sequences to provide a single contiguous view of the entire chromosome. Documentation, code and data for ACeDB are available from anonymous ftp servers at ftp.sanger.ac.uk/pub/acedb and ncbi.nlm.nih.gov/repository/acedb. Data from the various analyses were viewed interactively through the ACeDB annotator's graphical workbench. GENEFINDER (P. Green & L. Hillier, unpublished software) predictions were confirmed or adjusted to incorporate protein, cDNA and EST matches, hexamer coding potential and repetitive sequences (see http://www.sanger.ac.uk/Projects/P_falciparum/Methods_Analysis for full details of analysis protocols and parameters). We utilized the protein family databases Pfam²⁰ to classify common protein domains in the malaria genome. A number of web-based tools were used to identify transmembrane regions, coiled coil domains, signal peptides and overall suggestions of protein localization (http://www.sanger.ac.uk/Projects/P_falciparum/Toolkit).

Data release. Sequence data generated by the chromosome 3 project were released continuously and were available for searching using the on-site BLAST server and downloading by ftp without restriction. The fully annotated sequence is available for browsing and downloading from http://www.sanger.ac.uk/Projects/P_falciparum. Unfinished sequence data from the remaining eight *P. falciparum* chromosomes in progress at the Sanger Centre can also be accessed through this site.

Received 28 May; accepted 6 July 1999.

1. World Health Organisation *A Global Strategy for Malaria Control* (World Health Organisation, Geneva, 1993).
2. MacPherson, G. G., Warrell, M. J., White, N. J., Looareesuwan, S. & Warrell, D. A. Human cerebral malaria. A quantitative ultrastructural analysis of parasitised erythrocyte sequestration. *Am. J. Pathol.* **119**, 385–401 (1985).
3. White, N. J. The treatment of malaria. *N. Engl. J. Med.* **335**, 800–806 (1996).
4. Hoffman, S. L. *et al.* Funding for malaria genome sequencing. *Nature* **387**, 647 (1997).
5. Pollack, Y., Katzen, A. L., Spira, D. T. & Golenser, J. The genome of *Plasmodium falciparum*. I: DNA base composition. *Nucleic Acids Res.* **10**, 539–546 (1982).
6. Churcher, C. *et al.* The nucleotide sequence of *Saccharomyces cerevisiae* chromosome IX. *Nature (suppl.)* **387**, 84–87 (1997).
7. Gardner, M. J. *et al.* Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science* **282**, 1126–1132 (1998).

8. Thompson, J. K. & Cowman, A. F. AYAC contig and high resolution restriction map of chromosome 3 from *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **90**, 537–542 (1997).
9. Triglia, T. & Kemp, D. J. Large fragments of *Plasmodium falciparum* DNA can be stable when cloned in yeast artificial chromosomes. *Mol. Biochem. Parasitol.* **44**, 207–212 (1991).
10. Gardiner, K. *et al.* YAC analysis and minimal tiling path construction for chromosome 21q. *Somat. Cell. Mol. Genet.* **21**, 399–414 (1995).
11. Vaudin, M. *et al.* The construction and analysis of M13 libraries prepared from YAC DNA. *Nucleic Acids Res.* **23**, 670–674 (1995).
12. Walliker, D. *et al.* Genetic analysis of the human malaria parasite *Plasmodium falciparum*. *Science* **236**, 1661–1666 (1987).
13. Aston, C., Mishra, B. & Schwartz, D. C. Optical mapping and its potential for large scale sequencing projects. *Trends Biotechnol.* **17**, 297–302 (1999).
14. The *C. elegans* sequencing consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012–2018 (1998).
15. Dame, J. B. *et al.* Structure of the gene encoding the immunodominant surface antigen on the sporozoite of the human malaria parasite *Plasmodium falciparum*. *Science* **225**, 593–599 (1984).
16. Enea, V. *et al.* DNA cloning of *Plasmodium falciparum* circumsporozoite gene: amino acid sequence of the repetitive epitope. *Science* **225**, 628–630 (1984).
17. Trottein, F., Triglia, T. & Cowman, A. F. Molecular cloning of a gene from *Plasmodium falciparum* that codes for a protein sharing motifs found in adhesive molecules from mammals and *Plasmodia*. *Mol. Biochem. Parasitol.* **74**, 129–141 (1995).
18. Li, W. B. *et al.* An enlarged largest subunit of *Plasmodium falciparum* RNA polymerase II defines conserved and variable RNA polymerase domains. *Nucleic Acids Res.* **17**, 9621–9636 (1989).
19. Berriman, M. & Fairlamb, A. H. Detailed characterisation of a cyclophilin from the human malaria parasite *Plasmodium falciparum*. *Biochem. J.* **334**, 437–445 (1998).
20. Bateman, A. *et al.* Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic Acids Res.* **27**, 260–262 (1999).
21. Wootton, J. C. & Federhen, S. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* **266**, 554–571 (1996).
22. Lanzer, M., de Bruin, D. & Ravetch, J. V. Transcriptional mapping of a 100 kb locus of *Plasmodium falciparum* identifies a region in which transcription terminates and reinitiates. *EMBO J.* **11**, 1949–1955 (1992).
23. Horrocks, P., Dechering, K. & Lanzer, M. Control of gene expression in *Plasmodium falciparum*. *Mol. Biol. Parasitol.* **95**, 171–181 (1998).
24. Holt, D. C. *et al.* The *clag* gene family of *Plasmodium falciparum*: are there roles other than cytoadherence? *Int. J. Parasitol.* (in the press).
25. Wilson, R. J. M. *et al.* Complete gene map of the plastid-like DNA of the malaria parasite *Plasmodium falciparum*. *J. Mol. Biol.* **261**, 155–172 (1996).
26. Soldati, D. The apicoplast as a potential therapeutic target in *Toxoplasma* and other apicomplexan parasites. *Parasitol. Today* **15**, 5–7 (1999).
27. Waller, R. F. *et al.* Nuclear-encoded proteins target to the plastid in *Toxoplasma gondii* and *Plasmodium falciparum*. *Proc. Natl Acad. Sci. USA* **95**, 12352–12357 (1998).
28. De Bruin, D., Lanzer, M. & Ravetch, J. V. The polymorphic subtelomeric regions of *Plasmodium falciparum* chromosomes contain arrays of repetitive sequence elements. *Proc. Natl Acad. Sci. USA* **91**, 619–623 (1994).
29. Thompson, J. K. *et al.* The chromosomal organisation of the *Plasmodium falciparum* var gene is conserved. *Mol. Biochem. Parasitol.* **87**, 49–60 (1997).
30. Fischer, K. *et al.* Expression of var genes located within polymorphic subtelomeric domains of *Plasmodium falciparum* chromosomes. *Mol. Cell. Biol.* **17**, 3679–3686 (1997).
31. Carcy, B. *et al.* A large multigene family expressed during the erythrocytic schizogony of *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **68**, 221–233 (1997).
32. Leech, J. H., Barnwell, J. W., Miller, L. H. & Howard, R. J. Identification of a strain-specific malarial antigen exposed on the surface of *Plasmodium falciparum*-infected erythrocytes. *J. Exp. Med.* **159**, 1567–1575 (1984).
33. Kyes, S., Rowe, A., Kriek, N. & Newbold, C. I. RiFins: A second family of clonally variant proteins expressed on the surface of red cells infected with *Plasmodium falciparum*. *Proc. Natl Acad. Sci. USA* (in the press).
34. Newbold, C. I. *et al.* PEMP1, polymorphism and pathogenesis. *Ann. Trop. Med. Parasitol.* **91**, 551–557 (1997).
35. Pluta, A. F., Mackay, A. M., Ainsztein, A. M., Goldberg, I. G. & Earnshaw, W. C. The centromere: hub of chromosomal activities. *Science* **270**, 1591–1594 (1995).
36. Clarke, L. Centromeres: proteins, protein complexes, and repeated domains at centromeres of simple eukaryotes. *Curr. Opin. Genet. Dev.* **8**, 212–218 (1998).
37. Baum, M., Ngan, V. K. & Clarke, L. The centromeric K-type repeat and the central core are together sufficient to establish a functional *Schizosaccharomyces pombe* centromere. *Mol. Biol. Cell* **5**, 747–761 (1994).
38. Takahashi, K. *et al.* A low copy number central sequence with strict symmetry and unusual chromatin structure in fission yeast centromere. *Mol. Biol. Cell* **3**, 819–835 (1992).
39. Cheng, Q. *et al.* *stevor* and *rif* are *Plasmodium falciparum* multicopy gene families which potentially encode variant antigens. *Mol. Biochem. Parasitol.* **97**, 161–176 (1998).
40. Garcia, A., Cayla, X., Barik, S. & Langsley, G. A family of PP2 phosphatases in *Plasmodium falciparum* and parasitic protozoa. *Parasitol. Today* **15**, 90–91 (1999).
41. McMurray, A. A., Sulston, J. E. & Quail, M. A. Short-insert libraries as a method for problem solving in genome sequencing. *Genome Res.* **8**, 562–566 (1998).
42. Lowe, T. M. & Eddy, S. R. tRNA scan-SE: a program for improved detection of transfer RNA genes in genome sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
43. Mott, R. EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *Comput. Appl. Biosci.* **13**, 477–478 (1997).
44. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **15**, 403–410 (1990).
45. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).

Acknowledgements. We thank the staff in the subcloning, gel-pouring and media groups for their support; the computer support and software development groups; J. Thompson and A. Cowman for gifts of YAC clones and for advice; D. Schwartz for optical and X. Su for genetic map information; B. Streipen, D. Holt and D. Kemp for communicating results before publication; Z. Christodoulou, R. Pinches and S. Lee for technical assistance; the other members of the Malaria Genome Sequencing consortium for useful discussions; K. Rutherford and R. Summers for help with Fig. 1; and J. Parkhill for critical reading of the manuscript. This work was funded by the Wellcome Trust.

Correspondence and requests for materials should be addressed to S.B. (e-mail: sharen@sanger.ac.uk). Sequence accession numbers and other information can be found on http://www.sanger.ac.uk/Projects/P_falciparum.