Sequence of *Plasmodium falciparum* chromosomes 1, 3–9 and 13

N. Hall*, A. Pain*, M. Berriman*, C. Churcher*, B. Harris*, D. Harris*, K. Mungall*, S. Bowman*†, R. Atkin*, S. Baker*, A. Barron*, K. Brooks*, C. O. Buckee*, C. Burrows*, I. Cherevach*, C. Chillingworth*, T. Chillingworth*, Z. Christodoulou‡, L. Clark*, R. Clark*, C. Corton*, A. Cronin*, R. Davies*, P. Davis*, P. Dear§, F. Dearden*, J. Doggett*, T. Feltwell*, A. Goble*, I. Goodhead*, R. Gwilliam*, N. Hamlin*, Z. Hance*, D. Harper*, H. Hauser*, T. Hornsby*, S. Holroyd*, P. Horrocks‡, S. Humphry*, K. Jagels*, K. D. James*, D. Johnson*, A. Kerhornou*, A. Knights*, B. Konfortov§, S. Kyes‡, N. Larke*, D. Lawson*, N. Lennard*, A. Line*, M. Maddison*, J. McLean*, P. Mooney*, S. Moule*, L. Murphy*, K. Oliver*, D. Ormond*, C. Price*, M. A. Quail*, E. Rabbinowitsch* M.-A. Rajandream*, S. Rutter*, K. M. Rutherford*, M. Sanders*, M. Simmonds*, K. Seeger*, S. Sharp*, R. Smith*, R. Squares*, S. Squares*, K. Stevens*, K. Taylor*, A. Tivey*, L. Unwin*, S. Whitehead*, J. Woodward*, J. E. Sulston*, A. Craig||‡, C. Newbold‡ & B. G. Barrell*

* *The Wellcome Trust Sanger Institute, The Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK*
‡ *The Weatherall Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Headington, Oxford OX3 9DS, UK*
§ *MRC Laboratory of Molecular Biology, Hills Road, Cambridge, CB2 2QH, UK*
|| *Liverpool School of Tropical Medicine, Pembroke Place, Liverpool L3 5QA, UK*

Since the sequencing of the first two chromosomes of the malaria parasite, *Plasmodium falciparum*[1,2], there has been a concerted effort to sequence and assemble the entire genome of this organism. Here we report the sequence of chromosomes 1, 3–9 and 13 of *P. falciparum* clone 3D7—these chromosomes account for approximately 55% of the total genome. We describe the methods used to map, sequence and annotate these chromosomes. By comparing our assemblies with the optical map, we indicate the completeness of the resulting sequence. During annotation, we assign Gene Ontology terms to the predicted gene products, and observe clustering of some malaria-specific terms to specific chromosomes. We identify a highly conserved sequence element found in the intergenic region of internal *var* genes that is not associated with their telomeric counterparts.

Contiguous DNA sequences (contigs) have been obtained for chromosomes 1, 3, 4, 5 and 9, whereas chromosomes 6, 7, 8 and 13 contain a few gaps; most contigs have been ordered and oriented. Table 1 shows the status and content of the chromosomes at the time of writing. As we were unable to produce unbroken sequence from telomere to telomere for all nine chromosomes, contiguous 'pseudo-chromosomes' were constructed by artificially joining all contigs that could be mapped to an individual chromosome. In most cases, the order and orientation of the contigs could be inferred using mapping data[3–5] or read-pair information. Small contigs (of less than 5 kilobases, kb) that could not be mapped onto a chromosome have not been included in the analysis, and thus a small number of genes on the unmapped contigs will be missing from the genome sequence. The construction of pseudo-chromosomes does, however, have the advantage of allowing a global analysis of chromosome structure, and also removes redundancy from the analysis that would otherwise occur owing to contamination between chromosomes during purification and aberrant contigs formed during assembly.

A comparison of the optical maps for the finished chromosomes with virtual restriction digests with two enzymes of the assembled sequences show good agreement (Fig. 1). A misassembly in chromosome 4 is apparent from both comparisons, which we have localized to a region in an internal *var* gene repeat. The depth of coverage in this area suggests that there is a 50-kb perfect repeat. Chromosome 9 has a deletion of 100 kb in comparison with the *Bam*HI optical map, but it compares well with the *Nhe*I map, and with the sequence tagged site (STS) markers and the yeast artificial chromosome (YAC) map. The data strongly suggest that this anomaly is due to an optical mapping error, rather than a problem with the chromosome sequence.

The sizes of the pseudo-chromosomes 6, 7 and 8 also compare well with the predictions from the optical map. Chromosome 13 is 400 kb smaller than the predicted size in the *Nhe*I map, but only 10 kb smaller than the predicted size from the *Bam*HI map. Thus size comparisons between optical maps and digests reveal that very few data are missing from the chromosome assemblies (Fig. 1). When comparing contig order and orientation with the optical map of unfinished chromosomes, many more outliers are visible on the scatter plots (Fig. 1 and Table 1). Only chromosomes 13 and 6 have r^2 values of less than 0.8 in correlation analysis, both against the *Bam*HI maps. Thus for the most part, the contigs are ordered and oriented correctly.

Chromosomes 6, 7 and 8 do not resolve on pulsed field gel electrophoresis, and therefore they were sequenced as a group. Because of this we were unable to group contigs sufficiently to initiate gap closure. In order to overcome this problem, a HAPPY map[6–8] was created, using data from the genome sequence to design primers. HAPPY mapping allows the order and spacing of STS markers to be determined accurately, by following their segregation among roughly *hap*loid samples of randomly fragmented DNA, using the *poly*merase chain reaction. In the first round of mapping, 496 probes were generated which could be arranged on 61 linkage groups with 343 singletons at a lod (log of odds) threshold of 4. A further 30 probes were incorporated to increase the number of linkage groups to 62 at a lod threshold of 5 with 361 singletons. The large number of singletons produced was due to the high level of extra-chromosomal contamination of the purified chromosomes, which we estimated to be around 40%. Despite this, generation of a HAPPY map for chromosomes 6, 7 and 8 has been an invaluable step in grouping contigs to direct the finishing process.

Although gene predictions and annotations were performed by three different groups as part of the sequencing consortium, the predicted overall protein-coding content of each chromosome was very similar (Table 1). Small differences in coding percentage were seen in part due to chromosome size and thus their respective contributions of the telomeric sequences. The gene structures predicted from each group, assessed by comparing gene size, exon size and intron size, were also largely the same (Table 1). As the sequence for some chromosomes is incomplete, it is possible that exons that overlap gaps may be missed. In some cases where frameshifts occur within exons, particular effort has been made to check that these are pseudogenes and not caused by sequencing errors. The consistency of annotations across all chromosomes suggests that the quality of sequence has not seriously affected gene identification. We expect the accuracy of sequence of all chromosomes to be very high owing to the depth of read coverage (Table 1). Chromosome maps showing the location and structure of genes along each chromosome are available (Supplementary Information).

Gene Ontology (GO) was used to classify genes across the entire genome, and as GO had not been previously applied for annotating an intracellular parasite, new parasite-specific GO terms were created[9]. The proportion of genes associated with parasite-specific processes or localized in parasite-specific compartments varies between chromosomes (Fig. 2). Whereas most 'housekeeping' genes appear to be evenly distributed across the chromosomes (Fig. 2a), chromosome 5 appears to have the highest proportion of genes annotated with apicoplast localization (Fig. 2b). Conversely, and unlike chromosome 4, it has a very low proportion of genes associated with host cell invasion or adhesion (Fig. 2b, c). The

† Present address: Syngenta, Jealott's Hill International Research Centre, Bracknell RG42 6EY, UK.

uneven distribution of apicoplast targeted genes on chromosome 5 involves non-orthologous genes, whereas the clustering of genes involved in host cell invasion or adhesion results from duplications of gene families such as *variant antigen* (*var*) and *repetitive interspersed family* (*rif*) genes.

We have identified two previously undescribed clustered gene families; one on chromosome 9 and one on chromosome 13. On chromosome 9, there are 7 copies of a putative protein kinase which show 25–46% amino-acid identity to each other; four of these genes have a predicted signal peptide. Proteomic analysis has shown expression of two of these genes (PFI0105c and PFI0135c)[10]. Chromosome 13 contains a tandem array of 5 parologous genes including *MSP7* (ref. 11) with 15–30% identity to each other. Expression of one of these MSP7-like proteins (MAL13P1.174) has been detected, by proteomic studies, during the asexual stage[12]. The significance of the physical localization and function of these different genes is unknown, so further studies of their expression pattern and cellular localization are required. Protein alignments of these families are available (Supplementary Information).

Bowman *et al.*[2] deduced a consensus pattern of repeats and coding regions for the subtelomeric regions of chromosomes 2 and 3. The overall arrangement of *var*, *rif* and *subtelomeric variable open reading frame* (*stevor*) genes is conserved in nearly all telomeres, but the number and orientation of gene families vary. For example, many subtelomeres contain multiple *var* genes, and some have inverted *var* genes. The right-hand telomere of chromosome 5 has a truncated telomere with a partial inverted *var* gene adjacent to the telomeric repeat, with no rep11 or rep20 repeat units. The telomere-associated repeat elements are involved in co-localization of telomeres within the nucleus[13,14]. This may aid chromosome

Table 1 **Summary statistics**

	Value									
	Whole genome	Chr. 1	Chr. 3	Chr. 4	Chr. 5	Chr. 6	Chr. 7	Chr. 8	Chr. 9	Chr. 13
The genome										
Size (bp)	22,853,764	643,292	1,060,087	1,204,112	1,343,552	1,377,956	1,350,452	1,323,195	1,541,723	2,747,327
No. of gaps	93	0	0	0	0	8	14	24	0	37
Coverage*	14.5	13.3	10.9	16.8	15.1	16.8	15.8	16.2	17.9	17.2
Mapped YACs	–	15	19	18	16	16	17	23	14	29
HAPPY map linkage groups	–	–	–	–	–	17	7	8	–	–
*Bam*HI map length	–	667.9	1,146.6	1,136.8	1,306.8	1,443.8	1,503.7	1,372.8	1,687.9	2,734.9
r^2 *Bam*HI	–	0.994	0.999	0.778	0.998	0.796	0.878	0.986	0.958	0.741
*Nhe*I optical map length (kb)	–	683.8	1,083.5	1,311.1	1,394.8	1,494.7	1,493.5	1,331.4	1,600.0	3,171.8
r^2 *Nhe*I	–	0.999	0.997	0.983	0.998	0.908	0.989	0.878	0.909	0.821
(G + C) content (%)	19.4	20.5	19.9	20.7	19.3	19.7	20.0	19.7	19.0	19.2
No. of genes	5,268	143	239	237	312	312	277	295	365	672
Mean gene length (bp)	2,283.3	1,965.0	2,319.5	2,643.9	2,307.0	2,403.6	2,755.1	2,376.3	2,092.2	2,254.5
Gene density (kb per gene)	4,338.2	4,498.5	4,435.5	5,080.6	4,306.3	4,416.5	4,875.3	4,485.4	4,223.9	4,088.3
Percent coding†	52.6	43.7	52.3	52.0	53.6	54.4	56.5	53.0	49.5	55.1
Genes with introns (%)	53.9	69.9	59.0	58.6	52.6	52.9	56.0	57.3	59.2	52.7
Genes with ESTs (%)	47.4	37.8	51.5	45.1	51.0	52.2	45.5	48.1	52.9	54.6
Gene products detected by proteomics‡ (%)	48.2	50.3	53.1	50.6	54.8	52.8	51.6	55.6	53.4	53.4
Exons										
Number	12,674	373	638	576	736	809	651	784	925	1,656
Mean no. per gene	2.4	2.6	2.7	2.4	2.4	2.6	2.4	2.7	2.5	2.5
(G + C) content (%)	23.7	25.3	23.8	25.2	23.6	23.7	24.1	23.9	23.6	23.1
Mean length (bp)	949.1	753.3	868.9	1,087.9	978.0	927.0	1,172.3	894.2	825.6	914.9
Total length (bp)	12,028,350	280,998	554,355	626,607	719,781	749,937	763,167	701,019	763,644	1,515,033
Introns										
Number	7,406	230	399	339	424	497	374	489	560	984
(G + C) content (%)	13.5	13.5	13.4	13.5	13.6	13.8	13.5	13.6	13.4	13.4
Mean length (bp)	178.7	170.4	163.6	186.3	167.7	169.6	180.9	167.8	172.4	158.1
Total length (bp)	1,323,509	39,183	65,279	63,169	71,122	84,283	67,669	82,031	96,547	155,553
Intergenic regions										
(G + C) content (%)	13.6	14.2	13.6	14.0	13.5	13.9	13.8	13.8	13.2	13.4
Mean length (bp)	1,693.9	1,883.4	1,608.9	1,949.4	1,662.6	1,640.4	1,773.2	1,703.1	1,716.8	1,499.2
RNAs										
No. of tRNA genes	43	0	2	5	5	3	7	0	0	5
No. of 5S rRNA genes	3	0	0	0	0	0	0	0	0	0
No. of 5.8S, 18S, 28S rRNA units	7	1	0	0	1	0	1	2	0	1
The proteome										
Total predicted proteins	5,268	143	239	237	312	312	277	295	365	672
Hypothetical proteins§	3,208	80	140	138	175	168	159	189	219	396
InterPro matches	2,650	64	147	141	151	164	112	147	176	227
Pfam matches	1,746	52	100	96	131	131	91	115	139	ND
Gene Ontology										
Process	1,301	41	58	78	62	77	84	62	83	184
Function	1,244	29	59	60	76	67	66	66	88	189
Component	2,412	88	119	121	140	125	149	145	169	281
Targeted to apicoplast	551	14	29	20	49	17	30	33	43	69
Targeted to mitochondrion	246	3	9	3	20	23	16	17	19	31
Structural features										
Transmembrane domain(s)	1,631	74	79	82	89	92	96	104	117	179
Signal peptide	544	21	33	30	32	31	33	20	46	65
Signal anchor	367	18	9	18	23	23	16	16	34	44

ND, not determined; EST, expressed sequence tag. The optical map lengths were calculated by adding together the lengths of restriction fragments in order to estimate the amount of data missing from each of the unfinished chromosomes. The Pearsons product moment coefficient (r^2) was calculated for each chromosome against each of the optical maps using regression analysis (see Fig. 1). Specialized searches used the following programs and databases: InterPro[29]; Pfam[30]; Gene Ontology[28]. Predictions of apicoplast and mitochondrial targeting were performed using TargetP[31] and MitoProtII[32]; transmembrane domains, TMHMM[33]; and signal peptides and signal anchors, SignalP-2.0 (ref. 27).

* Average number of sequence reads per nucleotide.

† Excluding introns.

‡ Percentage of proteins detected in parasite extracts by two independent proteomic analyses[10,12].

§ Hypothetical proteins are proteins with insufficient similarity to characterized proteins in other organisms to justify provision of functional assignments.

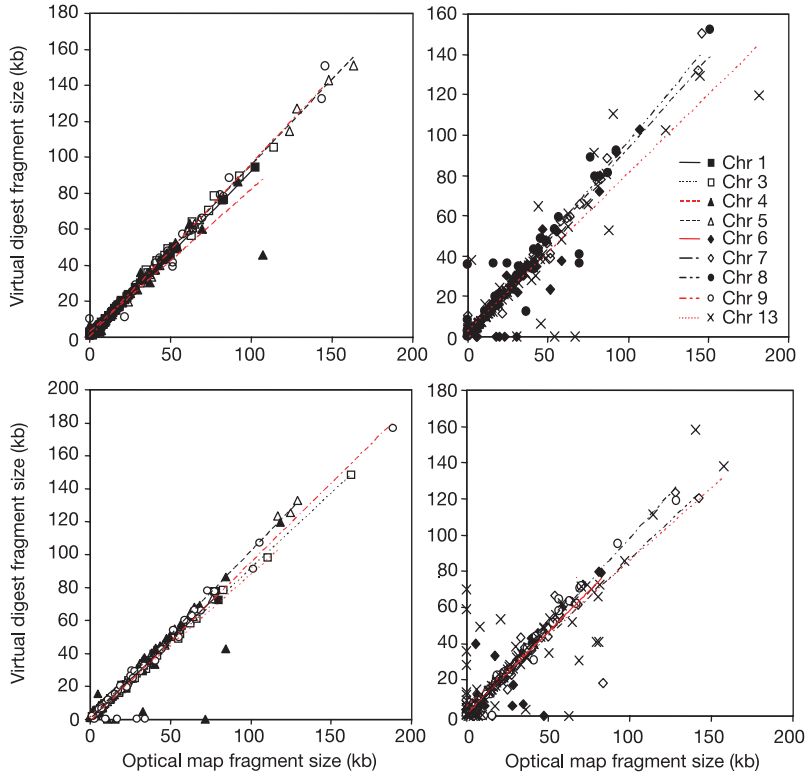

Figure 1 Scatter graphs of virtual restriction digests of completed chromosomes and pseudo-chromosomes against optical map fragment sizes. Top row: completed chromosomes (left) and unfinished chromosomes (right) compared with *Nhe*I optical map. Bottom row; as top row but compared with *Bam*HI optical map. Each point on the graph represents a restriction fragment compared to its corresponding optical map fragment. The lines show the regression for each chromosome.

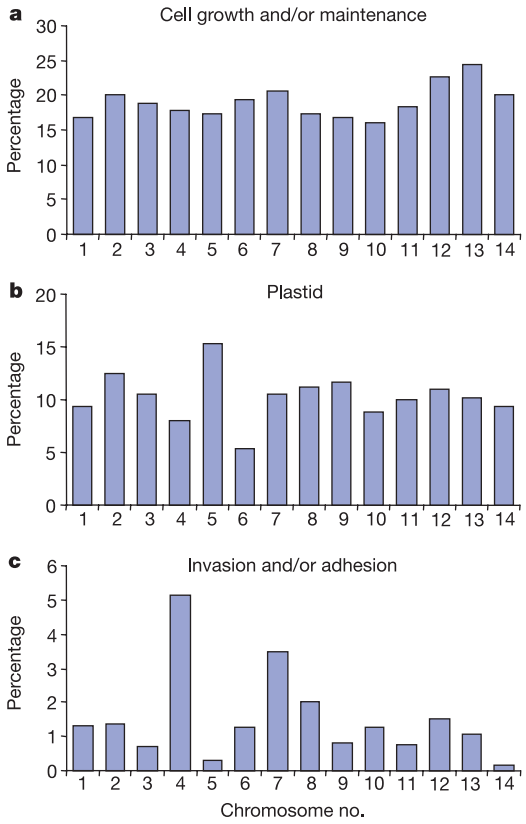

Figure 2 Comparison of the percentage of annotations with specific Gene Ontology terms on each chromosome. **a**, Annotations to 'cell growth and/or maintenance'; **b**, annotations to 'plastid'; **c**, annotations to 'invasion' and/or adhesion.

segregation and increased recombination between subtelomeric genes. Telomere repeats extending from truncated genes are frequently observed in other clones of *P. falciparum*, often leading to transcription of the telomere[13]. This observation suggests that telomere transcription may be involved in telomere maintenance at truncated chromosome ends. As the *var* gene on the right-hand end of chromosome 5 is inverted, there could be transcription of the telomeric repeat.

A putative centromere structure has been predicted in chromosomes 2 and 3 (ref. 2) which is characterized by a 2.6-kb region of 97.3% (A + T) content residing in a gap between coding sequences of at least 9 kb. On inspection of all of the completed chromosomes, we have identified similar structures representing the putative centromeres. There is only ever one per chromosome. All have a region of very high (A + T) content, and a core region of slightly higher (G + C) content, all lying in a gap between coding regions of between 8 and 11 kb. A similar structure has now been identified in the intracellular parasite *Encephalitozoon cuniculi*[15]. The discovery of these elements in all contiguous chromosomes, and now in another organism, suggests they have an important role in chromosome maintenance.

Three of the nine chromosomes that were sequenced by us (namely 4, 7 and 8) contain internal arrays of *var* genes. In the intergenic regions of the internal *var* arrays, we have identified a highly conserved, (G + C)-rich (~40% (G + C) content), sequence element of length ~202 bp (Fig. 3). We have also identified three such (G + C)-rich conserved elements on chromosome 12, sequenced in ref. 16 (not shown in Fig. 3). There are in total 15 of these (G + C)-rich elements in the entire *P. falciparum* genome, with not more than one element present in every internal *var* intergenic region. These (G + C)-rich elements are strictly associated with internal *var* arrays, and were not found in subtelomeric *var* genes, nor near the single internal *var* genes on chromosomes 6

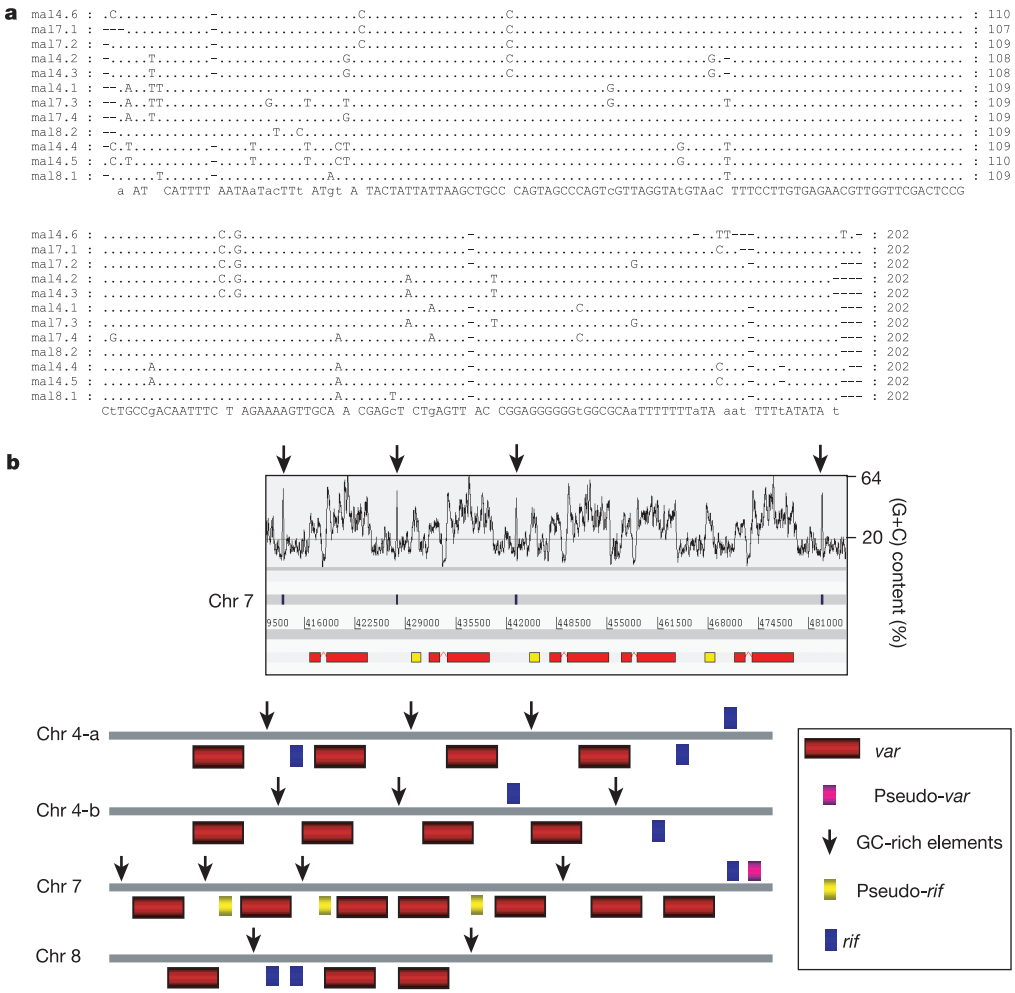

Figure 3 Position and structure of *var*-related (G + C)-rich elements. **a**, Multiple alignment of the (G + C)-rich conserved sequence elements on chromosomes 4, 7 and 8 of *P. falciparum*, using CLUSTAL. Only the non-identical nucleotides across all 12 (G + C)-rich conserved sequence elements are indicated in the alignment, with the consensus sequence indicated at the bottom. The upper-case letters in the consensus sequence denote complete identity across all the (G + C)-rich elements presented in the alignment. Each of these sequence elements is represented with a unique identifier, representing its specific origin. **b**, Location of the (G + C)-rich conserved sequence elements in the intergenic region of internal *var* gene clusters on chromosomes 4, 7 and 8 of *P. falciparum*. Top panel, four (G + C)-rich sequence elements in the intergenic regions of internal *var* gene cluster on chromosome 7. The arrowheads indicate the peaks in the (G + C) plot, corresponding to the location of the (G + C)-rich conserved sequence elements. The exact location of the neighbouring *var* and pseudo-*rif* genes are marked with red and yellow boxes, respectively. Bottom panel, a schematic diagram representing the relative positions of the internal *var* and *rif* genes and the conserved (G + C)-rich sequence elements on chromosomes 4, 7 and 8 (not to scale). The *var* or *rif* genes are placed either on top or bottom of the grey bars, depending on the direction of transcription.

and 12. There is no obvious systematic order of the location of these (G + C)-rich sequence elements with respect to adjacent *var* genes in terms of proximity or direction of transcription of the *var* genes. The specific positioning of these conserved sequence elements between internal *var* genes suggests a possible regulatory function, although a standard BLASTN query in public databases showed no significant similarity to previously identified RNA genes or gene regulatory elements. The (G + C)-rich element does have the potential to form secondary structures when analysed using the MFOLD program (http://bioweb.pasteur.fr/seqanal/interfaces/mfold-simple.html) (data not shown). This could indicate that the (G + C)-rich element is a hitherto unknown transcribed RNA species. *Cis*-acting (G + C)-rich gene regulatory elements have been shown to function as important transcriptional regulators present in the promoter, enhancer and locus control regions of many eukaryotic genes from several species (see ref. 17 for a review). The interaction between specific sites along a DNA molecule has been shown to have a crucial role in the regulation of genetic processes such as DNA replication, site-specific recombination and transposition in other organisms[18]. Control of gene expression through DNA loop formation has also been shown in other organisms[18], while in *P. falciparum* regulation of *var* gene expression by cooperative gene silencing elements in *var* gene introns[19], or by a 5′ flanking *var* gene region regulatory element, has also been described[20]. The potential of the (G + C)-rich sequences to form DNA secondary structures supports a possible function as regulatory elements in *var*-related genetic processes in *P. falciparum*. □

Methods

Sequencing

The DNA was cloned and sequenced according to methods described elsewhere[2,21]. Derived contigs were ordered according to previously derived genetic, optical and physical maps[3–5]. For all unfinished chromosomes, assemblies were screened against mapped contigs to remove extra-chromosomal contamination. For chromosomes 6, 7 and 8 a HAPPY map was generated to assist ordering; briefly, agarose-embedded genomic DNA was released by melting at 65 °C, sheared gently into fragments with a mean size of ~50 kb, and 88 samples, each containing ~0.7 genome-equivalents of fragments, were taken (a

further 8 samples were DNA-free controls). These samples (the mapping panel) were preamplified by PEP (primer extension preamplification), diluted and dispensed into 30 replica panels. Each replica was screened for between 50 and 100 markers using a two-phase polymerase chain reaction (multiplexed forward and reverse primers in phase 1, followed by dilution and a second phase for one marker at a time, using an internal forward primer and the reverse primer). Pairwise lod scores between markers were calculated, linkage groups identified, and maps of each group of three or more markers computed, essentially as described previously[7,8]

Annotation

Genome annotation was carried out using Artemis[22]. Genes were identified by manual curation of the output of the software packages Genefinder (P. Green, unpublished work), GlimmerM[23] and phat[24]. Functional assignments were based on assessment of BLAST and FASTA searches against public databases and domain predictions using InterproScan[25], TMHMM[26] and SignalP[27].

Gene Ontology (GO) terms[28] were manually assigned to gene products for all 14 chromosomes. First, candidate GO terms were selected by sequence-similarity searching a database of peptide sequences and their previously assigned GO terms, drawn from the following databases: Flybase, Mouse Genome Informatics, *Saccharomyces* Genome Database, Swissprot and The *Arabidopsis* Information Resource. After visual inspection of sequence alignments, suitable terms were either assigned directly from the candidate list, or alternatively, higher or lower granularity terms were selected directly from the ontology. When previously characterized genes were identified, terms were selected as above, but alternative experimental evidence codes were used to reflect the fact that the inferences were no longer based on sequence similarity. Some GO terms were also assigned automatically. In particular, 'membrane' was assigned using the transmembrane helix prediction tool TMHMM 2.0 (ref. 26).

Received 31 July; accepted 2 September 2002; doi:10.1038/nature01095.

1. Gardner, M. J. *et al.* Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science* **282,** 1126–1132 (1998).
2. Bowman, S. *et al.* The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*. *Nature* **400,** 532–538 (1999).
3. Su, X. *et al.* A genetic map and recombination parameters of the human malaria parasite *Plasmodium falciparum*. *Science* **286,** 1351–1353 (1999).
4. Lai, Z. *et al.* A shotgun optical map of the entire *Plasmodium falciparum* genome. *Nature Genet.* **23,** 309–313 (1999).
5. de Bruin, D., Lanzer, M. & Ravetch, J. V. Characterization of yeast artificial chromosomes from *Plasmodium falciparum*: construction of a stable, representative library and cloning of telomeric DNA fragments. *Genomics* **14,** 332–339 (1992).
6. Glockner, G. *et al.* Sequence and analysis of chromosome 2 of *Dictyostelium discoideum*. *Nature* **418,** 79–85 (2002).
7. Piper, M. B., Bankier, A. T. & Dear, P. H. A HAPPY map of *Cryptosporidium parvum*. *Genome Res.* **8,** 1299–1307 (1998).
8. Konfortov, B. A., Cohen, H. M., Bankier, A. T. & Dear, P. H. A high-resolution HAPPY map of *Dictyostelium discoideum* chromosome 6. *Genome Res.* **10,** 1737–1742 (2000).
9. Berriman, M., Aslett, M. & Ivens, A. Parasites are GO. *Trends Parasitol.* **17,** 463–464 (2001).
10. Florens, L. *et al.* A proteomic view of the *Plasmodium falciparum* life cycle. *Nature* (this issue).
11. Pachebat, J. A. *et al.* The 22 kDa component of the protein complex on the surface of *Plasmodium falciparum* merozoites is derived from a larger precursor, merozoite surface protein 7. *Mol. Biochem. Parasitol.* **117,** 83–89 (2001).
12. Lasonder, E. *et al.* Analysis of the *Plasmodium falciparum* proteome by high accuracy mass spectrometry. *Nature* (this issue).
13. Figueiredo, L. M., Freitas-Junior, L. H., Bottius, E., Olivo-Marin, J. C. & Scherf, A. A central role for *Plasmodium falciparum* subtelomeric regions in spatial positioning and telomere length regulation. *EMBO J.* **21,** 815–824 (2002).
14. O'Donnell, R. A. *et al.* A genetic screen for improved plasmid segregation reveals a role for Rep20 in the interaction of *Plasmodium falciparum* chromosomes. *EMBO J.* **21,** 1231–1239 (2002).
15. Katinka, M. D. *et al.* Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature* **414,** 450–453 (2001).
16. Hyman, R., Fung, E. & Dennis, R. Sequence of *Plasmodium falciparum* chromosome 12. *Nature* (this issue).
17. Hapgood, J. P., Riedemann, J. & Scherer, S. D. Regulation of gene expression by GC-rich DNA cis-elements. *Cell Biol. Int.* **25,** 17–31 (2001).
18. Adhya, S. Multipartite genetic control elements: communication by DNA loop. *Annu. Rev. Genet.* **23,** 227–2250 (1989).
19. Deitsch, K. W., Calderwood, M. S. & Wellems, T. E. Malaria. Cooperative silencing elements in *var* genes. *Nature* **412,** 875–876 (2001).
20. Vazquez-Macias, A. *et al.* A distinct 5′ flanking var gene region regulates *Plasmodium falciparum* variant erythrocyte surface antigen expression in placental malaria. *Mol. Microbiol.* **45,** 155–167 (2002).
21. Quail, M. A. M13 cloning of mung bean nuclease digested PCR fragments as a means of gap closure within A/T-rich, genome sequencing projects. *DNA Seq.* **12,** 355–359 (2001).
22. Rutherford, K. *et al.* Artemis: sequence visualization and annotation. *Bioinformatics* **16,** 944–945 (2000).
23. Salzberg, S. L., Pertea, M., Delcher, A. L., Gardner, M. J. & Tettelin, H. Interpolated Markov models for eukaryotic gene finding. *Genomics* **59,** 24–31 (1999).
24. Cawley, S. E., Wirth, A. I. & Speed, T. P. Phat—a gene finding program for *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **118,** 167–174 (2001).
25. Zdobnov, E. M. & Apweiler, R. InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17,** 847–848 (2001).
26. Sonnhammer, E. L., von Heijne, G. & Krogh, A. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **6,** 175–182 (1998).
27. Nielsen, H., Brunak, S. & von Heijne, G. Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Eng.* **12,** 3–9 (1999).
28. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nature Genet.* **25,** 25–29 (2000).
29. Apweiler, R. *et al.* The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.* **29,** 37–40 (2001).
30. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **30,** 276–280 (2002).
31. Emanuelsson, O., Nielsen, H., Brunak, S. & von Heijne, G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* **300,** 1005–1016 (2000).
32. Claros, M. G. & Vincens, P. Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur. J. Biochem.* **241,** 779–786 (1996).
33. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305,** 567–580 (2001).

Supplementary Information accompanies the paper on *Nature*'s website (http://www.nature.com/nature).

Acknowledgements

We thank the staff in the computer support and software development groups; J. Thompson and A. Cowman for gifts of YAC clones and for advice; D. Schwartz for optical map data; X. Su for genetic map information; Y. Shaw for help with Fig. 1; M. Harris and M. Ashburner for assistance with the parasite specific GO terms; O. White and M. Gardner for Table 1 and supplementary figures; the other members of the Malaria Genome Sequencing Consortium for discussions; and The Wellcome Trust Plasmodium Genome Mapping Consortium. This work was supported by the Wellcome Trust.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to N.H. (e-mail: nh1@sanger.ac.uk). Sequences have been deposited in the EMBL database with accession numbers AL844501 (chromosome 1), AL844502 (chromosome 3), AL844503 (chromosome 4), AL844504 (chromosome 5), AL844505 (chromosome 6), AL844506 (chromosome 7), AL844507 (chromosome 8), AL844508 (chromosome 9) and AL844509 (chromosome 13). Other information is available at http://www.sanger.ac.uk/Projects/P_falciparum.

Sequence of *Plasmodium falciparum* chromosomes 2, 10, 11 and 14

Malcolm J. Gardner*, Shamira Shallom*, Jane M. Carlton*, Steven L. Salzberg*, Vishvanath Nene*, Azadeh Shoaibi*, Anne Ciecko*, Jeffery Lynn*, Michael Rizzo*, Bruce Weaver*, Behnam Jarrahi*, Michael Brenner*, Babak Parvizi*, Luke Tallon*, Azita Moazzez*, David Granger*, Claire Fujii*, Cheryl Hansen*, James Pederson†, Tamara Feldblyum*, Jeremy Peterson*, Bernard Suh*, Sam Angiuoli*, Mihaela Pertea*, Jonathan Allen*, Jeremy Selengut*, Owen White*, Leda M. Cummings*‡, Hamilton O. Smith*‡, Mark D. Adams*‡ J. Craig Venter*‡, Daniel J. Carucci†, Stephen L. Hoffman†‡ & Claire M. Fraser*

* *The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850, USA*
† *Malaria Program, Naval Medical Research Center, 503 Robert Grant Avenue, Silver Spring, Maryland 20910-7500, USA*

The mosquito-borne malaria parasite *Plasmodium falciparum* kills an estimated 0.7–2.7 million people every year, primarily children in sub-Saharan Africa. Without effective interventions, a variety of factors—including the spread of parasites resistant to antimalarial drugs and the increasing insecticide resistance of mosquitoes—may cause the number of malaria cases to double over the next two decades[1]. To stimulate basic research and facilitate the development of new drugs and vaccines, the genome of *Plasmodium falciparum* clone 3D7 has been sequenced using a chromosome-by-chromosome shotgun strategy[2–4]. We report

‡ Present addresses: National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesdsa, Maryland 20894, USA (L.M.C.); Celera Genomics, 45 West Gude Drive, Rockville, Maryland 20850, USA (H.O.S., M.D.A.); The Center for the Advancement of Genomics, 1901 Research Boulevard, 6th Floor, Rockville, Maryland 20850, USA (C.V.); Sanaria, 308 Argosy Drive, Gaithersburg, Maryland 20878, USA (S.L.H.).