Genome sequence of the human malaria parasite *Plasmodium falciparum*

Malcolm J. Gardner[1], Neil Hall[2], Eula Fung[3], Owen White[1], Matthew Berriman[2], Richard W. Hyman[3], Jane M. Carlton[1], Arnab Pain[2], Karen E. Nelson[1], Sharen Bowman[2]*, Ian T. Paulsen[1], Keith James[2], Jonathan A. Eisen[1], Kim Rutherford[2], Steven L. Salzberg[1], Alister Craig[4], Sue Kyes[5], Man-Suen Chan[5], Vishvanath Nene[1], Shamira J. Shallom[1], Bernard Suh[1], Jeremy Peterson[1], Sam Angiuoli[1], Mihaela Pertea[1], Jonathan Allen[1], Jeremy Selengut[1], Daniel Haft[1], Michael W. Mather[6], Akhil B. Vaidya[6], David M. A. Martin[7], Alan H. Fairlamb[7], Martin J. Fraunholz[8], David S. Roos[8], Stuart A. Ralph[9], Geoffrey I. McFadden[9], Leda M. Cummings[1], G. Mani Subramanian[10], Chris Mungall[11], J. Craig Venter[12], Daniel J. Carucci[13], Stephen L. Hoffman[13]*, Chris Newbold[5], Ronald W. Davis[3], Claire M. Fraser[1] & Bart Barrell[2]

The parasite *Plasmodium falciparum* is responsible for hundreds of millions of cases of malaria, and kills more than one million African children annually. Here we report an analysis of the genome sequence of *P. falciparum* clone 3D7. The 23-megabase nuclear genome consists of 14 chromosomes, encodes about 5,300 genes, and is the most (A + T)-rich genome sequenced to date. Genes involved in antigenic variation are concentrated in the subtelomeric regions of the chromosomes. Compared to the genomes of free-living eukaryotic microbes, the genome of this intracellular parasite encodes fewer enzymes and transporters, but a large proportion of genes are devoted to immune evasion and host–parasite interactions. Many nuclear-encoded proteins are targeted to the apicoplast, an organelle involved in fatty-acid and isoprenoid metabolism. The genome sequence provides the foundation for future studies of this organism, and is being exploited in the search for new drugs and vaccines to fight malaria.

Despite more than a century of efforts to eradicate or control malaria, the disease remains a major and growing threat to the public health and economic development of countries in the tropical and subtropical regions of the world. Approximately 40% of the world's population lives in areas where malaria is transmitted. There are an estimated 300–500 million cases and up to 2.7 million deaths from malaria each year. The mortality levels are greatest in sub-Saharan Africa, where children under 5 years of age account for 90% of all deaths due to malaria[1]. Human malaria is caused by infection with intracellular parasites of the genus *Plasmodium* that are transmitted by *Anopheles* mosquitoes. Of the four species of *Plasmodium* that infect humans, *Plasmodium falciparum* is the most lethal. Resistance to anti-malarial drugs and insecticides, the decay of public health infrastructure, population movements, political unrest, and environmental changes are contributing to the spread of malaria[2]. In countries with endemic malaria, the annual economic growth rates over a 25-year period were 1.5% lower than in other countries. This implies that the cumulative effect of the lower annual economic output in a malaria-endemic country was a 50% reduction in the per capita GDP compared to a non-malarious country[3]. Recent studies suggest that the number of malaria cases may double in 20 years if new methods of control are not devised and implemented[1].

An international effort[4] was launched in 1996 to sequence the *P. falciparum* genome with the expectation that the genome sequence would open new avenues for research. The sequences of two of the 14 chromosomes, representing 8% of the nuclear genome, were published previously[5,6] and the accompanying Letters in this issue describe the sequences of chromosomes 1, 3–9 and 13 (ref. 7), 2, 10, 11 and 14 (ref. 8), and 12 (ref. 9). Here we report an analysis of the genome sequence of *P. falciparum* clone 3D7, including descriptions of chromosome structure, gene content, functional classification of proteins, metabolism and transport, and other features of parasite biology.

Sequencing strategy

A whole chromosome shotgun sequencing strategy was used to determine the genome sequence of *P. falciparum* clone 3D7. This approach was taken because a whole genome shotgun strategy was not feasible or cost-effective with the technology that was available at the beginning of the project. Also, high-quality large insert libraries of (A + T)-rich *P. falciparum* DNA have never been constructed in *Escherichia coli*, which ruled out a clone-by-clone sequencing strategy. The chromosomes were separated on pulsed field gels, and chromosomal DNA was extracted and used to construct shotgun libraries of 1–3-kilobase (kb) fragments of sheared DNA. Eleven of the fourteen chromosomes could be resolved on the gels, but chromosomes 6, 7 and 8 could not be resolved and were sequenced as a group. The shotgun sequences were assembled into contiguous DNA sequences (contigs), in some cases with low coverage shotgun sequences of yeast artificial chromosome (YAC) clones to assist in the ordering of contigs for closure. Sequence tagged sites (STSs)[10], microsatellite markers[11,12] and HAPPY mapping[7] were also used to place and orient contigs during the gap closure process. The high (A + T) content of the genome made gap closure extremely difficult[7–9]. The predicted restriction enzyme maps of the chromosome sequences were compared to optical restriction maps to verify that the chromosomes had been assembled correctly[13]. Chromosomes 1–5, 9 and 12 were closed, whereas chromosomes 6–8, 10, 11, 13 and 14 contained 3–37 gaps (most <2.5 kb) per chromosome at the beginning of genome annotation. Efforts to close the remaining gaps are continuing.

[1] The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850, USA; [2] The Wellcome Trust Sanger Institute, The Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK; [3] Stanford Genome Technology Center, 855 California Avenue, Palo Alto, California 94304, USA; [4] Liverpool School of Tropical Medicine, Pembroke Place, Liverpool L3 5QA, UK; [5] University of Oxford, Weatherall Institute of Molecular Medicine, John Radcliffe Hospital, Headington, Oxford OX3 9DU, UK; [6] Department of Microbiology and Immunology, Drexel University College of Medicine, 2900 Queen Lane, Philadelphia, Pennsylvania 19129, USA; [7] School of Life Sciences, The Wellcome Trust Biocentre, The University of Dundee, Dundee DD1 5EH, UK; [8] Department of Biology and Genomics Institute, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6018, USA; [9] Plant Cell Biology Research Centre, School of Botany, University of Melbourne, Melbourne, VIC 3010, Australia; [10] Celera Genomics, 45 West Gude Drive, Rockville, Maryland 20850, USA; [11] Department of Molecular and Cellular Biology, Berkeley Drosophila Genome Project, University of California, Berkeley, California 94720, USA; [12] The Center for the Advancement of Genomics, 1901 Research Boulevard, 6th Floor, Rockville, Maryland 20850, USA; [13] Malaria Program, Naval Medical Research Center, 503 Robert Grant Avenue, Silver Spring, Maryland 20910-7500, USA.

*Present addresses: Syngenta, Jealott's Hill International Research Centre, Bracknell, RG42 6EY, UK (S.B.); Sanaria, 308 Argosy Drive, Gaithersburg, Maryland 20878, USA (S.L.H.).

Genome structure and content

The *P. falciparum* 3D7 nuclear genome is composed of 22.8 megabases (Mb) distributed among 14 chromosomes ranging in size from approximately 0.643 to 3.29 Mb (Fig. 1, and Supplementary Figs A–N). Thus the *P. falciparum* genome is almost twice the size of the genome of the fission yeast *Schizosaccharomyces pombe*. The overall (A + T) composition is 80.6%, and rises to ~90% in introns and intergenic regions. The structures of protein-encoding genes were predicted using several gene-finding programs and manually curated. Approximately 5,300 protein-encoding genes were identified, about the same as in *S. pombe* (Table 1, and Supplementary Table A). This suggests an average gene density in *P. falciparum* of 1 gene per 4,338 base pairs (bp), slightly higher than was found previously with chromosomes 2 and 3 (1 per 4,500 bp and 1 per 4,800 bp, respectively). The higher gene density reported here is probably the result of improved gene-finding software and larger training sets that enabled the detection of genes overlooked previously[8]. Introns were predicted in 54% of *P. falciparum* genes, a proportion roughly similar to that in *S. pombe* and *Dictyostelium discoideum*, but much higher than observed in *Saccharomyces cerevisiae* where only 5% of genes contain introns. Excluding introns, the mean length of *P. falciparum* genes was 2.3 kb, substantially larger than in the other organisms in which the average gene lengths range from 1.3 to 1.6 kb. *Plasmodium falciparum* genes showed a markedly greater proportion of genes (15.5%) longer than 4 kb compared to *S. pombe* and *S. cerevisiae* (3.0% and 3.6%, respectively). The explanation for the increased gene length in *P. falciparum* is not clear. Many of these large genes encode uncharacterized proteins that may be cytosolic proteins, as they do not possess recognizable signal peptides. No transposable elements or retrotransposons were identified.

Fifty-two per cent of the predicted gene products (2,731) were detected in cell lysates prepared from several stages of the parasite life cycle by high-resolution liquid chromatography and tandem mass spectrometry[14,15], including many predicted proteins with no similarity to proteins in other organisms. In addition, 49% of the genes overlapped (97% identity over at least 100 nucleotides) with expressed sequence tags (ESTs) derived from several life-cycle stages. As the proteomics and EST studies performed to date may not represent a complete sampling of all genes expressed during the complex life cycle of the parasite, this suggests that the annotation process identified substantial portions of most genes. However, in the absence of supporting EST or protein evidence, correct prediction of the 5′ ends of genes and genes with multiple small exons is challenging, and the gene models should be regarded as preliminary. Additional ESTs and full-length complementary DNA sequences[16] are required for the development of better training sets for gene-finding programs and the verification of the predicted genes.

The nuclear genome contains a full set of transfer RNA (tRNA) ligase genes, and 43 tRNAs were identified to bind all codons except TGT and TGC, coding for Cys; it is possible that these tRNAs are located within the currently unsequenced regions. All codons ending in C and T appear to be read by single tRNAs with a G in the first position, which is likely to read both codons via G:U wobble. Each anticodon occurs only once except for methionine (CAT), for which there are two copies, one for translation initiation and one for internal methionines, and the glycine (CCT) anticodon, which occurs twice. An unusual tRNA resembling a selenocysteinyl-tRNA was also found. A putative selenocysteine lyase was identified, which may provide selenium for synthesis of selenoproteins. Increased growth has been observed in selenium-supplemented *Plasmodium* culture[17].

In almost all other eukaryotic organisms sequenced to date, the tRNA genes exhibit extensive redundancy, the only exception being the intracellular parasite *Encephalitozoon cuniculi* which contains 44 tRNAs[18]. Often, the abundance of specific anticodons is correlated with the codon usage of the organism[19,20]. This is not the case in *P. falciparum*, which exhibits minimal redundancy of tRNAs. The mitochondrial genome of *Plasmodium* is small (about 6 kb) and encodes no tRNAs, so the mitochondrion must import tRNAs[21,22]. Through their import, cytoplasmic tRNAs may serve mitochondrial protein synthesis in a manner seen with other organisms[23,24]. The apicoplast genome appears to encode sufficient tRNAs for protein synthesis within the organelle[25].

Unlike many other eukaryotes, the malaria parasite genome does not contain long tandemly repeated arrays of ribosomal RNA (rRNA) genes. Instead, *Plasmodium* parasites contain several single 18S-5.8S-28S rRNA units distributed on different chromosomes.

Table 1 ***Plasmodium falciparum*** **nuclear genome summary and comparison to other organisms**

Feature	Value				
	P. falciparum	*S. pombe*	*S. cerevisiae*	*D. discoideum*	*A. thaliana*
Size (bp)	22,853,764	12,462,637	12,495,682	8,100,000	115,409,949
(G + C) content (%)	19.4	36.0	38.3	22.2	34.9
No. of genes	5,268*	4,929	5,770	2,799	25,498
Mean gene length† (bp)	2,283	1,426	1,424	1,626	1,310
Gene density (bp per gene)	4,338	2,528	2,088	2,600	4,526
Per cent coding	52.6	57.5	70.5	56.3	28.8
Genes with introns (%)	53.9	43	5.0	68	79
Exons					
Number	12,674	ND	ND	6,398	132,982
No. per gene	2.39	ND	NA	2.29	5.18
(G + C) content (%)	23.7	39.6	28.0	28.0	ND
Mean length (bp)	949	ND	ND	711	170
Total length (bp)	12,028,350	ND	ND	4,548,978	33,249,250
Introns					
Number	7,406	4,730	272	3,587	107,784
(G + C) content (%)	13.5	ND	NA	13.0	ND
Mean length (bp)	178.7	81	NA	177	170
Total length (bp)	1,323,509	383,130	ND	643,899	18,055,421
Intergenic regions					
(G + C) content (%)	13.6	ND	ND	14.0	ND
Mean length (bp)	1,694	952	515	786	ND
RNAs					
No. of tRNA genes	43	174	ND	73	ND
No. of 5S rRNA genes	3	30	ND	NA	ND
No. of 5.8S, 18S and 28S rRNA units	7	200–400	ND	NA	700–800

ND, not determined; NA, not applicable. 'No. of genes' for *D. discoideum* are for chromosome 2 (ref. 155) and in some cases represent extrapolations to the entire genome. Sources of data for the other organisms: *S. pombe*[65], *S. cerevisiae*[156], *D. discoideum*[155] and *A. thaliana*[157].
*70% of these genes matched expressed sequence tags or encoded proteins detected by proteomics analyses[14,15].
†Excluding introns.

The sequence encoded by a rRNA gene in one unit differs from the sequence of the corresponding rRNA in the other units. Furthermore, the expression of each rRNA unit is developmentally regulated, resulting in the expression of a different set of rRNAs at different stages of the parasite life cycle[26,27]. It is likely that by changing the properties of its ribosomes the parasite is able to alter the rate of translation, either globally or of specific messenger RNAs (mRNAs), thereby changing the rate of cell growth or altering patterns of cell development. The two types of rRNA genes previously described in *P. falciparum* are the S-type, expressed primarily in the mosquito vector, and the A-type, expressed primarily in the human host. Seven loci encoding rRNAs were identified in the genome sequence (Fig. 1). Two copies of the S-type rRNA genes are located on chromosomes 11 and 13, and two copies of the A-type genes are located on chromosomes 5 and 7. In addition, chromosome 1 contains a third, previously uncharacterized, rRNA unit that encodes 18S and 5.8S rRNAs that are almost identical to the S-type genes on chromosomes 11 and 13, but has a significantly divergent 28S rRNA gene (65% identity to the A-type and 75% identity to the S-type). The expression profiles of these genes are unknown. Chromosome 8 also contains two unusual rRNA gene units that contain 5.8S and 28S rRNA genes but do not encode 18S rRNAs; it is not known whether these genes are functional. The sequences of the 18S and 28S rRNA genes on chromosome 7 and the 28S rRNA gene on chromosome 8 are incomplete as they reside at contig ends. The 5S rRNA is encoded by three identical tandemly arrayed genes on chromosome 14.

Chromosome structure

Plasmodium falciparum chromosomes vary considerably in length, with most of the variation occurring in the subtelomeric regions. Field isolates, even those from individuals residing in a single village[28], exhibit extensive size polymorphism that is thought to be due to recombination events between different parasite clones during meiosis in the mosquito[29]. Chromosome size variation is also observed in cultures of erythrocytic parasites, but is due to chromosome breakage and healing events and not to meiotic recombination[30,31]. Subtelomeric deletions often extend well into the chromosome, and in some cases alter the cell adhesion properties of the parasite owing to the loss of the gene(s) encoding adhesion molecules[32,33]. Because many genes involved in antigenic variation are located in the subtelomeric regions, an understanding of subtelomere structure and functional properties is essential for the elucidation of the mechanisms underlying the generation of antigenic diversity.

The subtelomeric regions of the chromosomes display a striking degree of conservation within the genome that is probably due to promiscuous inter-chromosomal exchange of subtelomeric regions. Subtelomeric exchanges occur in other eukaryotes[34–36], but the regions involved are much smaller (2.5–3.0 kb) in *S. cerevisiae* (data not shown). Previous studies of *P. falciparum* telomeres[37,38] suggested that they contained six blocks of repetitive sequences that were designated telomere-associated repetitive elements (TAREs 1–6).

Whole genome analysis reveals a larger (up to 120 kb), more complex, subtelomeric repeat structure than was observed previously. The conserved regions fall into five large subtelomeric blocks (SBs; Fig. 2). The sequences within blocks 2, 4 and 5 include many tandem repeats in addition to those described previously, as well as non-repetitive regions. Subtelomeric block 1 (SB-1, equivalent to TARE-1), contains the 7-bp telomeric repeat in a variable number of near-exact copies[39]. SB-2 contains several sub-blocks of repeats of different sizes, including TAREs 2–5 and other sequences. The beginning of SB-2 consists of about 1,000–1,300 bp of non-repetitive sequence, followed on some chromosomes by 2.5 copies of a 164-bp repeat. This is followed by another 300 bp of non-repetitive sequence, and then 10 copies of a 135-bp repeat, the main element of TARE-2. TARE-2 is followed by 200 bp of non-repetitive sequence, and then two copies of a highly conserved 63-bp repeat. SB-2 extends for another 6 kb that contains non-repetitive sequence as well as other tandem repeats. Only four of the 28 telomeres are missing SB-2, which always occurs immediately adjacent to SB-1. A notable feature of SB-2 is the conserved order and orientation of each repeat variant as well as the sequence homology extending throughout the block. For almost any two chromosomes that were examined, a consistently ordered series of unique, identical sequences of >30 bp that are distributed across SB-2 were identified, suggesting that SB-2 is a repeat with a complex internal structure occurring once per telomere.

SB-3 consists of the Rep20 element[40], a large block of highly variable copies of a 21-bp repeat. The tandem repeats in SB-3 occur in a random order (Fig. 2). SB-4 has not been described previously, although it does contain the previously described R-FA3 sequence[41]. SB-4 also includes a complex mix of short (<28-bp) tandem repeats, and a 105-bp repeat that occurs once in each subtelomere. Many telomeres contain one or more *var* (variant antigen) gene exons within this block, which appear as gaps in the alignment. In five subtelomeres, fragments of 2–4 kb from SB-4 are duplicated and inverted. SB-5 is found in half of the subtelomeres, does not contain tandem repeats, and extends up to 120 kb into some chromosomes. The arrangement and composition of the subtelomeric blocks suggests frequent recombination between the telomeres.

Centromeres have not been identified experimentally in malaria parasites. However, putative centromeres were identified by comparison of the sequences of chromosomes 2 and 3 (ref. 6). Eleven of the 14 chromosomes contained a single region of 2–3 kb with extremely high (A + T) content (>97%) and imperfect short tandem repeats, features resembling the regional *S. pombe* centromeres; the 3 chromosomes lacking such regions were incomplete.

The proteome

Of the 5,268 predicted proteins, about 60% (3,208 hypothetical proteins) did not have sufficient similarity to proteins in other organisms to justify provision of functional assignments (Table 2). This is similar to what was found previously with chromosomes 2 and 3 (refs 5, 6). Thus, almost two-thirds of the proteins appear to be unique to this organism, a proportion much higher than observed in other eukaryotes. This may be a reflection of the greater evolutionary distance between *Plasmodium* and other eukaryotes that have been sequenced, exacerbated by the reduction of sequence similarity due to the (A + T) richness of the genome. Another 257 proteins (5%) had significant similarity to hypothetical proteins in other organisms. Thirty-one per cent (1,631) of the predicted proteins had one or more transmembrane domains, and 17.3% (911) of the proteins possessed putative signal peptides or signal anchors.

The Gene Ontology (GO)[42] database is a controlled vocabulary that describes the roles of genes and gene products in organisms. GO terms were assigned manually to 2,134 gene products (40%)

Figure 1 Schematic representation of the *P. falciparum* 3D7 genome. ►
Protein-encoding genes are indicated by open diamonds. All genes are depicted at the same scale regardless of their size or structure. The labels indicate the name for each gene. The rows of coloured rectangles represent, from top to bottom for each chromosome, the high-level Gene Ontology assignment for each gene in the 'biological process', 'molecular function', and 'cellular component' ontologies[42]; the life-cycle stage(s) at which each predicted gene product has been detected by proteomics techniques[14,15]; and *Plasmodium yoelii yoelii* genes that exhibit conserved sequence and organization with genes in *P. falciparum*, as shown by a position effect analysis. Rectangles surrounding clusters of *P. yoelii* genes indicate genes shown to be linked in the *P. y. yoelii* genome[165]. Boxes containing coloured arrowheads at the ends of each chromosome indicate subtelomeric blocks (SBs; see text and Fig. 2).

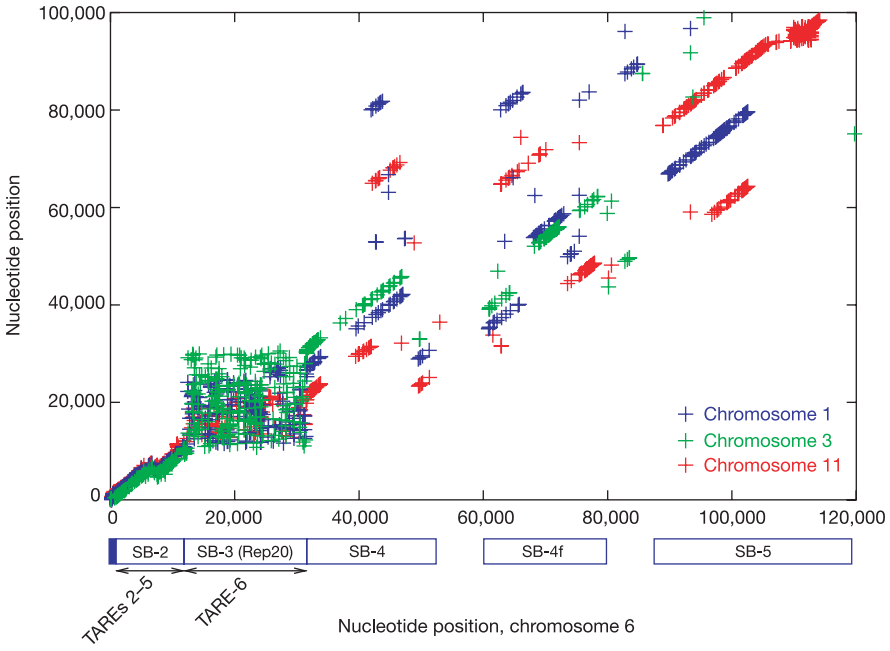

Figure 2 Alignment of subtelomeric regions of chromosomes 1, 3, 6 and 11. MUMmer2[152] alignments showing exact matches between the left subtelomeric regions of chromosome 6 (horizontal axis) and chromosomes 11 (red), 1 (blue) and 3 (green), illustrating the conserved synteny between all telomeres. Each point represents an exact match of 40 bp or longer that is shared by two chromosomes and is not found anywhere else on either chromosome. Each collinear series of points along a diagonal represents an aligned region. SB, subtelomeric block; TARE, telomere-associated repetitive element.

and a comparison of annotation with high-level GO terms for both *S. cerevisiae* and *P. falciparum* is shown in Fig. 3. In almost all categories, higher values can be seen for *S. cerevisiae*, reflecting the greater proportion of the genome that has been characterized compared to *P. falciparum*. There are two exceptions to this pattern that reflect processes specifically connected with the parasite life cycle. At least 1.3% of *P. falciparum* genes are involved in cell-to-cell adhesion or the invasion of host cells. As discussed below (see 'Immune evasion'), *P. falciparum* has 208 genes (3.9%) known to be involved in the evasion of the host immune system. This is reflected in the assignment of many more gene products to the GO term 'physiological processes' in *P. falciparum* than in *S. cerevisiae* (Fig. 3). The comparison with *S. cerevisiae* also reveals that particular categories in *P. falciparum* appear to be under-represented. Sporulation and cell budding are obvious examples (they are included in the category 'other cell growth and/or maintenance'), but very few genes in *P. falciparum* were associated with the 'cell organization and biogenesis', the 'cell cycle', or 'transcription factor' categories compared to *S. cerevisiae* (Fig. 3). These differences do not necessarily imply that fewer malaria genes are involved in these processes, but highlight areas of malaria biology where knowledge is limited.

Table 2 **The *P. falciparum* proteome**

Feature	Number	Per cent
Total predicted proteins	5,268	
Hypothetical proteins	3,208	60.9
InterPro matches	2,650	52.8
Pfam matches	1,746	33.1
Gene Ontology		
Process	1,301	24.7
Function	1,244	23.6
Component	2,412	45.8
Targeted to apicoplast	551	10.4
Targeted to mitochondrion	246	4.7
Structural features		
Transmembrane domain(s)	1,631	31.0
Signal peptide	544	10.3
Signal anchor	367	7.0
Non-secretory protein	4,357	82.7

Of the apicoplast-targeted proteins, 126 were judged on the basis of experimental evidence or the predictions of multiple programs[61,158] to be localized to the apicoplast with high confidence. Predicted apicoplast localization for 425 other proteins is based on an analysis using only one method and is of lower confidence. Predicted mitochondrial localization was based upon BLASTP searches of *S. cerevisiae* mitochondrion-targeted proteins[159] and TargetP[158] and MitoProtII[160] predictions; 148 genes were judged to be targeted to the mitochondrion with a high or medium confidence level, and an additional 98 genes with a lower confidence of mitochondrial targeting. Other specialized searches used the following programs and databases: InterPro[161]; Pfam[162]; Gene Ontology[42]; transmembrane domains, TMHMM[163]; signal peptides and signal anchors, SignalP-2.0[164].

The apicoplast

Malaria parasites and other members of the phylum apicomplexa harbour a relict plastid, homologous to the chloroplasts of plants and algae[25,43,44]. The 'apicoplast' is essential for parasite survival[45,46], but its exact role is unclear. The apicoplast is known to function in the anabolic synthesis of fatty acids[5,47,48], isoprenoids[49] and haeme[50,51], suggesting that one or more of these compounds could be exported from the apicoplast, as is known to occur in plant plastids. The apicoplast arose through a process of secondary endosymbiosis[52–55], in which the ancestor of all apicomplexan parasites engulfed a eukaryotic alga, and retained the algal plastid, itself the product of a prior endosymbiotic event[56]. The 35-kb apicoplast genome encodes only 30 proteins[25], but as in mitochondria and chloroplasts, the apicoplast proteome is supplemented by proteins encoded in the nuclear genome and post-translationally targeted into the organelle by the use of a bipartite targeting signal, consisting of an amino-terminal secretory signal sequence, followed by a plastid transit peptide[55,57–60].

In total, 551 nuclear-encoded proteins (~10% of the predicted nuclear encoded proteins) that may be targeted to the apicoplast were identified using bioinformatic[61] and laboratory-based methods. Apicoplast targeting of a few proteins has been verified by antibody localization and by the targeting of fluorescent fusion proteins to the apicoplast in transgenic *P. falciparum* or *Toxoplasma gondii*[47] parasites. Some proteins may be targeted to both the apicoplast and mitochondrion, as suggested by the observation that the total number of tRNA ligases is inadequate for independent

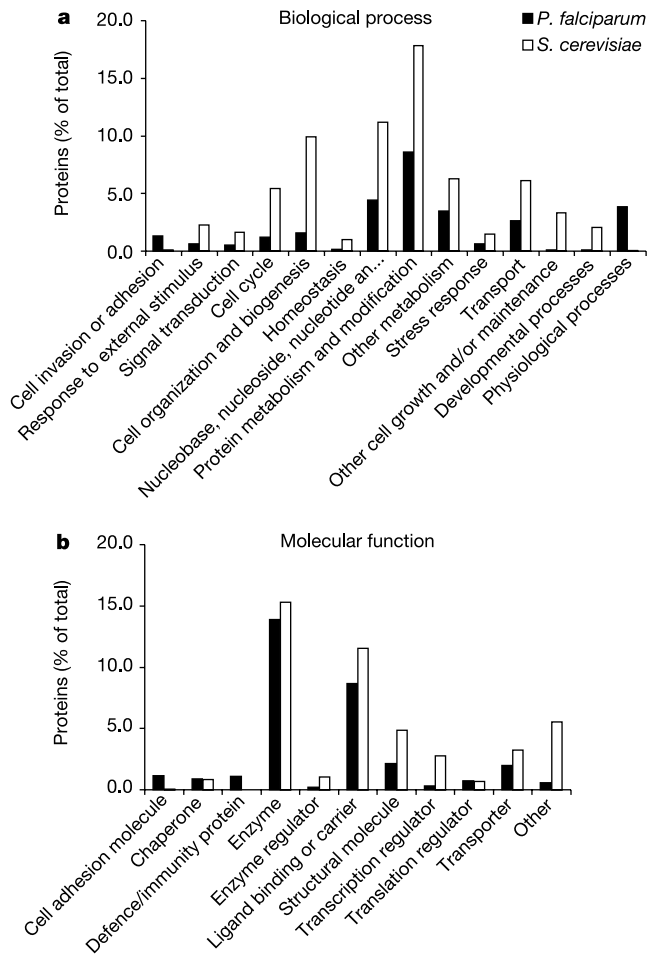

Figure 3 Gene Ontology classifications. Classification of *P. falciparum* proteins according to the 'biological process' (**a**) and 'molecular function' (**b**) ontologies of the Gene Ontology system[42].

protein synthesis in the cytoplasm, mitochondrion and apicoplast. In plants, some proteins lack a transit peptide but are targeted to plastids via an unknown process. Proteins that use an alternative targeting pathway in *P. falciparum* would have escaped detection with the methods used.

Nuclear-encoded apicoplast proteins include housekeeping enzymes involved in DNA replication and repair, transcription, translation and post-translational modifications, cofactor synthesis, protein import, protein turnover, and specific metabolic and transport activities. No genes for photosynthesis or light perception are apparent, although ferredoxin and ferredoxin-NADP reductase are present as vestiges of photosystem I, and probably serve to recycle reducing equivalents[62]. About 60% of the putative apicoplast-targeted proteins are of unknown function. Several metabolic pathways in the organelle are distinct from host pathways and offer potential parasite-specific targets for drug therapy[63] (see 'Metabolism' and 'Transport' sections).

Evolution

Comparative genome analysis with other eukaryotes for which the complete genome is available (excluding the parasite *E. cuniculi*) revealed that, in terms of overall genome content, *P. falciparum* is slightly more similar to *Arabidopsis thaliana* than to other taxa. Although this is consistent with phylogenetic studies[64], it could also be due to the presence in the *P. falciparum* nuclear genome of genes derived from plastids or from the nuclear genome of the secondary endosymbiont. Thus the apparent affinity of *Plasmodium* and *Arabidopsis* might not reflect the true phylogenetic history of the *P. falciparum* lineage. Comparative genomic analysis was also used to identify genes apparently duplicated in the *P. falciparum* lineage since it split from the lineages represented by the other completed genomes (Supplementary Table B).

There are 237 *P. falciparum* proteins with strong matches to proteins in all completed eukaryotic genomes but no matches to proteins, even at low stringency, in any complete prokaryotic proteome (Supplementary Table C). These proteins help to define the differences between eukaryotes and prokaryotes. Proteins in this list include those with roles in cytoskeleton construction and maintenance, chromatin packaging and modification, cell cycle regulation, intracellular signalling, transcription, translation, replication, and many proteins of unknown function. This list overlaps with, but is somewhat larger than, the list generated by an analysis of the *S. pombe* genome[65]. The differences are probably due in part to the different stringencies used to identify the presence or absence of homologues in the two studies.

A large number of nuclear-encoded genes in most eukaryotic species trace their evolutionary origins to genes from organelles that have been transferred to the nucleus during the course of eukaryotic evolution. Similarity searches against other complete genomes were used to identify *P. falciparum* nuclear-encoded genes that may be derived from organellar genomes. Because similarity searches are not an ideal method for inferring evolutionary relatedness[66], phylogenetic analysis was used to gain a more accurate picture of the evolutionary history of these genes. Out of 200 candidates examined, 60 genes were identified as being of probable mitochondrial origin. The proteins encoded by these genes include many with known or expected mitochondrial functions (for example, the tricarboxylic acid (TCA) cycle, protein translation, oxidative damage protection, the synthesis of haem, ubiquinone and pyrimidines), as well as proteins of unknown function. Out of 300 candidates examined, 30 were identified as being of probable plastid origin, including genes with predicted roles in transcription and translation, protein cleavage and degradation, the synthesis of isoprenoids and fatty acids, and those encoding four subunits of the pyruvate dehydrogenase complex. The origin of many candidate organelle-derived genes could not be conclusively determined, in part due to the problems inherent in analysing genes of very high (A + T) content. Nevertheless, it appears likely that the total number of plastid-derived genes in *P. falciparum* will be significantly lower than that in the plant *A. thaliana* (estimated to be over 1,000). Phylogenetic analysis reveals that, as with the *A. thaliana* plastid, many of the genes predicted to be targeted to the apicoplast are apparently not of plastid origin. Of 333 putative apicoplast-targeted genes for which trees were constructed, only 26 could be assigned a probable plastid origin. In contrast, 35 were assigned a probable mitochondrial origin and another 85 might be of mitochondrial origin but are probably not of plastid origin (they group with eukaryotes that have not had plastids in their history, such as humans and fungi, but the relationship to mitochondrial ancestors is not clear). The apparent non-plastid origin of these genes could either be due to inaccuracies in the targeting predictions or to the co-option of genes derived from the mitochondria or the nucleus to function in the plastid, as has been shown to occur in some plant species[67].

Metabolism

Biochemical studies of the malaria parasite have been restricted primarily to the intra-erythrocytic stage of the life cycle, owing to the difficulty of obtaining suitable quantities of material from the other life-cycle stages. Analysis of the genome sequence provides a global view of the metabolic potential of *P. falciparum* irrespective of the life-cycle stage (Fig. 4). Of the 5,268 predicted proteins, 733 (~14%) were identified as enzymes, of which 435 (~8%) were assigned Enzyme Commission (EC) numbers. This is considerably

fewer than the roughly one-quarter to one-third of the genes in bacterial and archaeal genomes that can be mapped to Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway diagrams[68], or the 17% of *S. cerevisiae* open reading frames that can be assigned EC numbers. This suggests either that *P. falciparum* has a smaller proportion of its genome devoted to enzymes, or that enzymes are more difficult to identify in *P. falciparum* by sequence similarity methods. (This difficulty can be attributed either to the great evolutionary distance between *P. falciparum* and other well-studied organisms, or to the high (A + T) content of the genome.) A few genes might have escaped detection because they were located in the small regions of the genome that remain to be sequenced (Table 1). However, many biochemical pathways could be reconstructed in their entirety, suggesting that the similarity-searching approach was for the most part successful, and that the relative paucity of enzymes in *P. falciparum* may be related to its parasitic life-style. A similar picture has emerged in the analysis of transporters (see 'Transport').

In erythrocytic stages, *P. falciparum* relies principally on anaerobic glycolysis for energy production, with regeneration of NAD^+ by conversion of pyruvate to lactate[69]. Genes encoding all of the enzymes necessary for a functional glycolytic pathway were identified, including a phosphofructokinase (PFK) that has sequence similarity to the pyrophosphate-dependent class of enzymes but which is probably ATP-dependent on the basis of the characterization of the homologous enzyme in *Plasmodium berghei*[70,71]. A second putative pyrophosphate-dependent PFK was also identified which possessed N- and carboxy-terminal extensions that could represent targeting sequences.

A gene encoding fructose bisphosphatase could not be detected, suggesting that gluconeogenesis is absent, as are enzymes for synthesis of trehalose, glycogen or other carbohydrate stores. Candidate genes for all but one enzyme of the conventional pentose

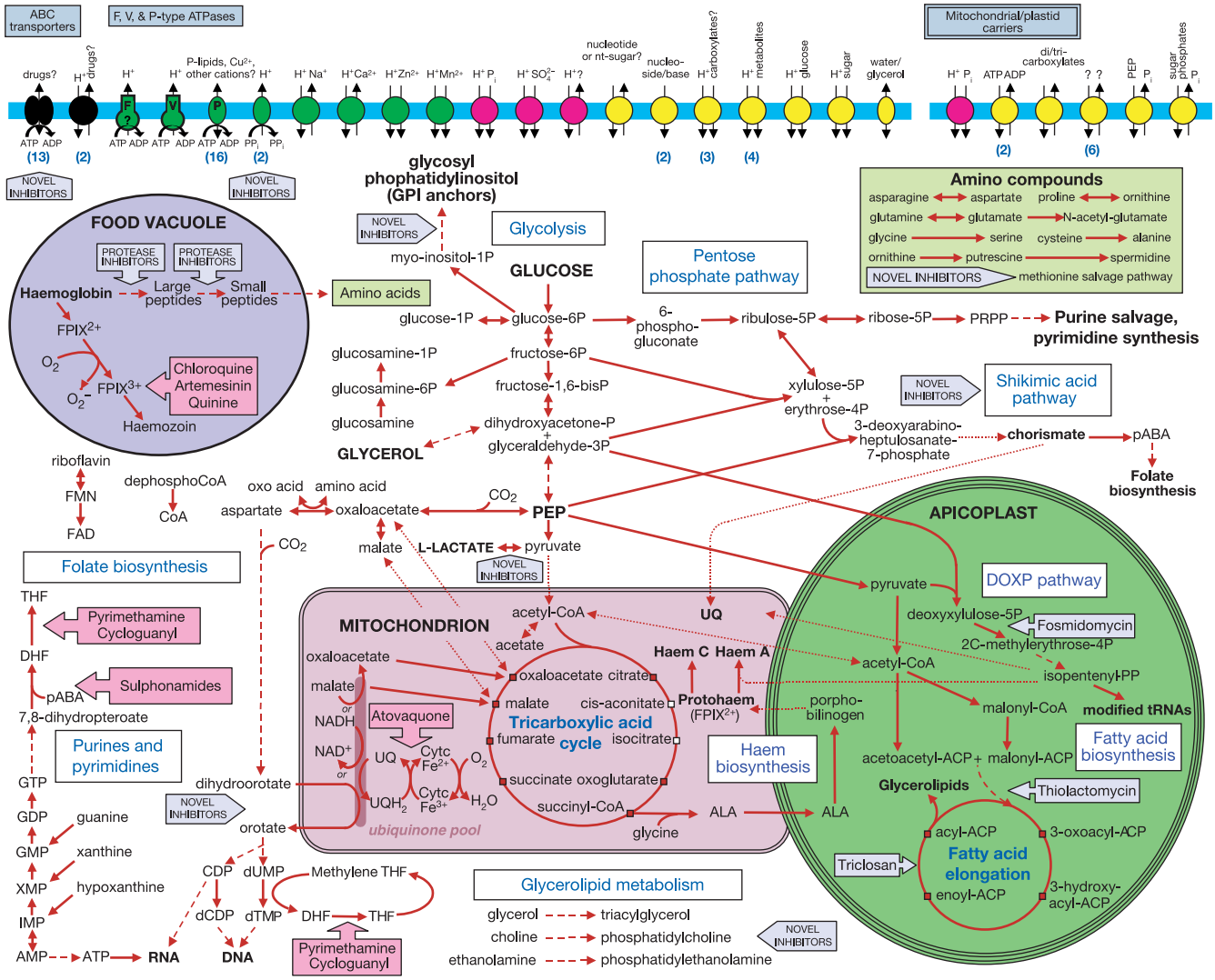

Figure 4 Overview of metabolism and transport in *P. falciparum*. Glucose and glycerol provide the major carbon sources for malaria parasites. Metabolic steps are indicated by arrows, with broken lines indicating multiple intervening steps not shown; dotted arrows indicate incomplete, unknown or questionable pathways. Known or potential organellar localization is shown for pathways associated with the food vacuole, mitochondrion and apicoplast. Small white squares indicate TCA (tricarboxylic acid) cycle metabolites that may be derived from outside the mitochondrion. Fuschia block arrows indicate the steps inhibited by antimalarials; grey block arrows highlight potential drug targets. Transporters are grouped by substrate specificity: inorganic cations (green), inorganic anions (magenta), organic nutrients (yellow), drug efflux and other (black). Arrows indicate direction of transport for substrates (and coupling ions, where appropriate). Numbers in parentheses indicate the presence of multiple transporter genes with similar substrate predictions. Membrane transporters of unknown or putative subcellular localization are shown in a generic membrane (blue bar). Abbreviations: ACP, acyl carrier protein; ALA, aminolevulinic acid; CoA, coenzyme A; DHF, dihydrofolate; DOXP, deoxyxylulose phosphate; $FPIX^{2+}$ and $FPIX^{3+}$, ferro- and ferriprotoporphyrin IX, respectively; pABA, *para*-aminobenzoic acid; PEP, phosphoenolpyruvate; P_i, phosphate; PP_i, pyrophosphate; PRPP, phosphoribosyl pyrophosphate; THF, tetrahydrofolate; UQ, ubiquinone.

phosphate pathway were found. These include a bifunctional glucose-6-phosphate dehydrogenase/6-phosphogluconate dehydrogenase required to generate NADPH and ribose 5-phosphate for other biosynthetic pathways[72,73]. Transaldolase appears to be absent, but erythrose 4-phosphate required for the chorismate pathway could probably be generated from the glycolytic intermediates fructose 6-phosphate and glyceraldehyde 3-phosphate via a putative transketolase (Fig. 4).

The genes necessary for a complete TCA cycle, including a complete pyruvate dehydrogenase complex, were identified. However, it remains unclear whether the TCA cycle is used for the full oxidation of products of glycolysis, or whether it is used to supply intermediates for other biosynthetic pathways. The pyruvate dehydrogenase complex seems to be localized in the apicoplast, and the only protein with significant similarity to aconitases has been reported to be a cytosolic iron-response element binding protein that did not possess aconitase activity[74]. Also, malate dehydrogenase appears to be cytosolic rather than mitochondrial, even though it seems to have originated from the mitochondrial genome[75]. Genes encoding malate-quinone oxidoreductase and type I fumarate dehydratase are present. Malate-quinone oxidoreductase, which is probably targeted to the mitochondrion, may well replace malate dehydrogenase in the TCA cycle, as it does in *Helicobacter pylori*. A gene encoding phosphoenolpyruvate carboxylase (PEPC) was also found. Like bacteria and plants, *P. falciparum* may cope with a drain of TCA cycle intermediates by using phosphoenolpyruvate (PEP) to replenish oxaloacetate (Fig. 4). This would seem to be supported by reports of CO_2-incorporating activity in asexual stage parasite cultures[76]. Thus, the TCA cycle appears to be unconventional in erythrocytic stages, and may serve mainly to synthesize succinyl-CoA, which in turn can be used in the haem biosynthesis pathway.

Genes encoding all subunits of the catalytic F_1 portion of ATP synthase, the protein that confers oligomycin sensitivity, and the gene that encodes the proteolipid subunit *c* for the F_0 portion of ATP synthase, were detected in the parasite genome. The F_0 *a* and *b* subunits could not be detected, raising the question as to whether the ATP synthase is functional. Because parts of the genome sequence are incomplete, the presence of the *a* and *b* subunits could not be ruled out. Erythrocytic parasites derive ATP through glycolysis and the mitochondrial contribution to the ATP pool in these stages appears to be minimal[77,78]. It is possible that the ATP synthase functions in the insect or sexual stages of the parasite. However, in the absence of the F_0 *a* and *b* subunits, an ATP synthase cannot use the proton gradient[79].

A functional mitochondrion requires the generation of an electrochemical gradient across the inner membrane. But the *P. falciparum* genome seems to lack genes encoding components of a conventional NADH dehydrogenase complex I. Instead, a single subunit NADH dehydrogenase gene specifies an enzyme that can accomplish ubiquinone reduction without proton pumping, thus constituting a non-electrogenic step. Other dehydrogenases targeted to the mitochondrion also serve to reduce ubiquinone in *P. falciparum*, including dihydroorotate dehydrogenase, a critical enzyme in the essential pyrimidine biosynthesis pathway[80]. The parasite genome contains some genes specifying ubiquinone synthesis enzymes, in agreement with recent metabolic labelling studies[81]. Re-oxidation of ubiquinol is carried out by the cytochrome *bc1* complex that transfers electrons to cytochrome *c*, and is accompanied by proton translocation[82]. Apocytochrome *b* of this complex is encoded by the mitochondrial genome[21,22], but the rest of the components are encoded by nuclear genes. Ubiquinol cycling is a critical step in mitochondrial physiology, and its selective inhibition by hydroxynaphthoquinones is the basis for their antimalarial action[83]. The final step in electron transport is carried out by the proton-pumping cytochrome *c* oxidase complex, of which only two subunits are encoded in the mitochondrial DNA (mtDNA). In most eukaryotes, subunit II of cytochrome *c* oxidase is encoded by a gene on the mitochondrial genome. In *P. falciparum*, however, the *coxII* gene is divided such that the N-terminal portion is encoded on chromosome 13 and the C-terminal portion on chromosome 14. A similar division of the *coxII* gene is also seen in the unicellular alga, *Chlamydomonas reinhardtii*[84]. An alternative oxidase that transfers electrons directly from ubiquinol to oxygen has been seen in plants as well in many protists, and an earlier biochemical study suggested its presence in *P. falciparum*[85]. The genome sequence, however, fails to reveal such an oxidase gene.

Biochemical, genetic and chemotherapeutic data suggest that malaria and other apicomplexan parasites synthesize chorismate from erythrose 4-phosphate and phosphoenolpyruvate via the shikimate pathway[86–89]. It was initially suggested that the pathway was located in the apicoplast[88], but chorismate synthase is phylogenetically unrelated to plastid isoforms[90] and has subsequently been localized to the cytosol[91]. The genes for the preceding enzymes in the pathway could not be identified with certainty, but a BLASTP search with the *S. cerevisiae* arom polypeptide[92], which catalyses 5 of the preceding steps, identified a protein with a low level of similarity (E value 7.9×10^{-8}).

In many organisms, chorismate is the pivotal precursor to several pathways, including the biosynthesis of aromatic amino acids and ubiquinone. We found no evidence, on the basis of similarity searches, for a role of chorismate in the synthesis of tryptophan, tyrosine or phenylalanine, although *para*-aminobenzoate (pABA) synthase does have a high degree of similarity to anthranilate (2-amino benzoate) synthase, the enzyme catalysing the first step in tryptophan synthesis from chorismate. In accordance with the supposition that the malaria parasite obtains all of its amino acids either by salvage from the host or by globin digestion, we found no enzymes required for the synthesis of other amino acids with the exception of enzymes required for glycine–serine, cysteine–alanine, aspartate–asparagine, proline–ornithine and glutamine–glutamate interconversions. In addition to pABA synthase, all but one of the enzymes (dihydroneopterin aldolase) required for *de novo* synthesis of folate from GTP were identified.

Several studies have shown that the erythrocytic stages of *P. falciparum* are incapable of *de novo* purine synthesis (reviewed in ref. 80). This statement can now be extended to all life-cycle stages, as only adenylsuccinate lyase, one of the 10 enzymes required to make inosine monophosphate (IMP) from phosphoribosyl pyrophosphate, was identified. This enzyme also plays a role in purine salvage by converting IMP to AMP. Purine transporters and enzymes for the interconversion of purine bases and nucleosides are also present. The parasite can synthesize pyrimidines *de novo* from glutamine, bicarbonate and aspartate, and the genes for each step are present. Deoxyribonucleotides are formed via an aerobic ribonucleoside diphosphate reductase[93,94], which is linked via thioredoxin to thioredoxin reductase. Gene knockout experiments have recently shown that thioredoxin reductase is essential for parasite survival[95].

The intraerythrocytic stages of the malaria parasite uses haemoglobin from the erythrocyte cytoplasm as a food source, hydrolysing globin to small peptides, and releasing haem that is detoxified in the form of haemazoin. Although large amounts of haem are toxic to the parasite, *de novo* haem biosynthesis has been reported[96] and presumably provides a mechanism by which the parasite can segregate host-derived haem from haem required for synthesis of its own iron-containing proteins. However, it has been unclear whether *de novo* synthesis occurs using imported host enzymes[97] or parasite-derived enzymes. Genes encoding the first two enzymes in the haem biosynthetic pathway, aminolevulinate synthase[98] and aminolevulinate dehydratase[99], were cloned previously, and genes encoding every other enzyme in the pathway except for uroporphyrinogen-III synthase were found (Fig. 4).

Haem and iron–sulphur clusters form redox prosthetic groups for a wide range of proteins, many of which are localized to the

mitochondrion and apicoplast. The parasite genome appears to encode enzymes required for the synthesis of these molecules. There are two putative cysteine desulphurase genes, one which also has homology to selenocysteine lyase and may be targeted to the mitochondrion, and the second which may be targeted to the apicoplast, suggesting organelle specific generation of elemental sulphur to be used in Fe–S cluster proteins. The subcellular localization of the enzymes involved in haem synthesis is uncertain. Ferrochelatase and two haem lyases are likely to be localized in the mitochondrion.

The role of the apicoplast in type II fatty-acid biosynthesis was described previously[5,47]. The genes encoding all enzymes in the pathway have now been elucidated, except for a thioesterase required for chain termination. No evidence was found for the associative (type I) pathway for fatty-acid biosynthesis common to most eukaryotes. The apicoplast also houses the machinery for mevalonate-independent isoprenoid synthesis. Because it is not present in mammals, the biosynthesis of isopentyl diphosphate from pyruvate and glyceraldehyde-3-phosphate provides several attractive targets for chemotherapy. Three enzymes in the pathway have been identified, including 1-deoxy-D-xylulose-5-phosphate synthase, 1-deoxy-D-xylulose-5-phosphate reductoisomerase[49], and 2C-methyl-D-erythritol 2,4-cyclodiphosphate synthase[100,101]. One predicted protein was similar to the fourth enzyme, 2C-methyl-D-erythritol-4-phosphate cytidyltransferase (BLASTP E value 9.6×10^{-15}).

Transport

On the basis of genome analysis, *P. falciparum* possesses a very limited repertoire of membrane transporters, particularly for uptake of organic nutrients, compared to other sequenced eukaryotes (Fig. 5). For instance, there are only six *P. falciparum* members of the major facilitator superfamily (MFS) and one member of the amino acid/polyamine/choline APC family, less than 10% of the numbers seen in *S. cerevisiae*, S. *pombe* or *Caenorhabditis elegans* (Fig. 5). The apparent lack of solute transporters in *P. falciparum* correlates with the lower percentage of multispanning membrane proteins compared with other eukaryotic organisms (Fig. 5). The predicted transport capabilities of *P. falciparum* resemble those of obligate intracellular prokaryotic parasites, which also possess a limited complement of transporters for organic solutes[102].

A complete catalogue of the identified transporters is presented in Fig. 4. In addition to the glucose/proton symporter[103] and the water/glycerol channel[104], one other probable sugar transporter and three carboxylate transporters were identified; one or more of the latter are probably responsible for the lactate and pyruvate/proton symport activity of *P. falciparum*[105]. Two nucleoside/nucleobase transporters are encoded on the *P. falciparum* genome, one of which has been localized to the parasite plasma membrane[106]. No obvious amino-acid transporters were detected, which emphasizes the importance of haemoglobin digestion within the food vacuole as an important source of amino acids for the erythrocytic stages of the parasite. How the insect stages of the parasite acquire amino acids and other important nutrients is unknown, but four metabolic uptake systems were identified whose substrate specificity could not be predicted with confidence. The parasite may also possess novel proteins that mediate these activities. Nine members of the mitochondrial carrier family are present in *P. falciparum*, including an ATP/ADP exchanger[107] and a di/tri-carboxylate exchanger, probably involved in transport of TCA cycle intermediates across the mitochondrial membrane. Probable phosphoenolpyruvate/phosphate and sugar phosphate/phosphate antiporters most similar to those of plant chloroplasts were identified, suggesting that these transporters are targeted to the apicoplast membrane. The former may enable uptake of phosphoenolpyruvate as a precursor of fatty-acid biosynthesis.

A more extensive set of transporters could be identified for the transport of inorganic ions and for export of drugs and hydrophobic compounds. Sodium/proton and calcium/proton exchangers were identified, as well as other metal cation transporters, including a substantial set of 16 P-type ATPases. An Nramp divalent cation transporter was identified which may be specific for manganese or iron. *Plasmodium falciparum* contains all subunits of V-type ATPases as well as two proton translocating pyrophosphatases[108], which could be used to generate a proton motive force, possibly across the parasite plasma membrane as well as across a vacuolar membrane. The proton pumping pyrophosphatases are not present in mammals, and could form attractive antimalarial targets. Only a single copy of the *P. falciparum* chloroquine-resistance gene *crt* is present, but multiple homologues of the multidrug resistance pump *mdr1* and other predicted multidrug transporters were identified (Fig. 3). Mutations in *crt* seem to have a central role in the development of chloroquine resistance[109].

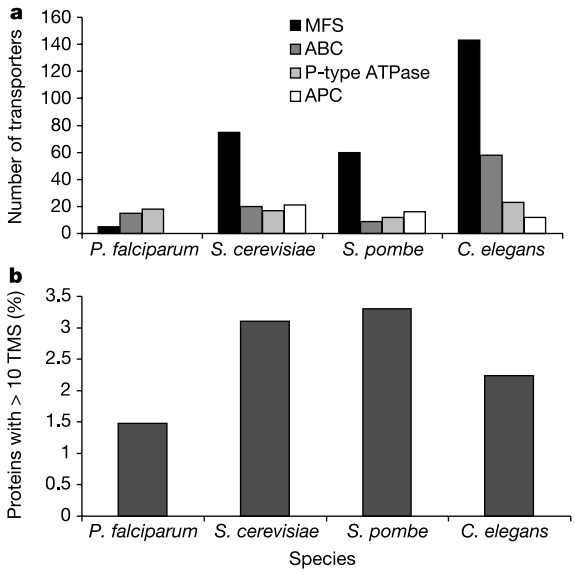

Figure 5 Analysis of transporters in *P. falciparum*. **a**, Comparison of the numbers of transporters belonging to the major facilitator superfamily (MFS), ATP-binding cassette (ABC) family, P-type ATPase family and the amino acid/polyamine/choline (APC) family in *P. falciparum* and other eukaryotes. Analyses were performed as previously described[102]. **b**, Comparison of the numbers of proteins with ten or more predicted transmembrane segments[163] (TMS) in *P. falciparum* and other eukaryotes. Prediction of membrane spanning segments was performed using TMHMM.

Plasmodium falciparum infection of erythrocytes causes a variety of pleiotropic changes in host membrane transport. Patch clamp analysis has described a novel broad-specificity channel activated or inserted in the red blood cell membrane by *P. falciparum* infection that allows uptake of various nutrients[110]. If this channel is encoded by the parasite, it is not obvious from genome analysis, because no clear homologues of eukaryotic sodium, potassium or chloride ion channels could be identified. This suggests that *P. falciparum* may use one or more novel membrane channels for this activity.

DNA replication, repair and recombination

DNA repair processes are involved in maintenance of genomic integrity in response to DNA damaging agents such as irradiation, chemicals and oxygen radicals, as well as errors in DNA metabolism such as misincorporation during DNA replication. The *P. falciparum* genome encodes at least some components of the major DNA repair processes that have been found in other eukaryotes[111,112]. The core of eukaryotic nucleotide excision repair is present (XPB/Rad25, XPG/Rad2, XPF/Rad1, XPD/Rad3, ERCC1) although some highly conserved proteins with more accessory roles

could not be found (for example, XPA/Rad4, XPC). The same is true for homologous recombinational repair with core proteins such as MRE11, DMC1, Rad50 and Rad51 present but accessory proteins such as NBS1 and XRS2 not yet found. These accessory proteins tend to be poorly conserved and have not been found outside of animals or yeast, respectively, and thus may be either absent or difficult to identify in *P. falciparum*. However, it is interesting that Archaea possess many of the core proteins but not the accessory proteins for these repair processes, suggesting that many of the accessory eukaryotic repair proteins evolved after *P. falciparum* diverged from other eukaryotes.

The presence of MutL and MutS homologues including possible orthologues of MSH2, MSH6, MLH1 and PMS1 suggests that *P. falciparum* can perform post-replication mismatch repair. Orthologues of MSH4 and MSH5, which are involved in meiotic crossing over in other eukaryotes, are apparently absent in *P. falciparum*. The repair of at least some damaged bases may be performed by the combined action of the four base excision repair glycosylase homologues and one of the apurinic/apyrimidinic (AP) endonucleases (homologues of Xth and Nfo are present). Experimental evidence suggests that this is done by the long-patch pathway[113].

The presence of a class II photolyase homologue is intriguing, because it is not clear whether *P. falciparum* is exposed to significant amounts of ultraviolet irradiation during its life cycle. It is possible that this protein functions as a blue-light receptor instead of a photolyase, as do members of this gene family in some organisms such as humans. Perhaps most interesting is the apparent absence of homologues of any of the genes encoding enzymes known to be involved in non-homologous end joining (NHEJ) in eukaryotes (for example, Ku70, Ku86, Ligase IV and XRCC1)[112]. NHEJ is involved in the repair of double strand breaks induced by irradiation and chemicals in other eukaryotes (such as yeast and humans), and is also involved in a few cellular processes that create double strand breaks (for example, VDJ recombination in the immune system in humans). The role of NHEJ in repairing radiation-induced double strand breaks varies between species[114]. For example, in humans, cells with defects in NHEJ are highly sensitive to γ-irradiation while yeast mutants are not. Double strand breaks in yeast are repaired primarily by homologous recombination. As NHEJ is involved in regulating telomere stability in other organisms, its apparent absence in *P. falciparum* may explain some of the unusual properties of the telomeres in this species[115].

Secretory pathway

Plasmodium falciparum contains genes encoding proteins that are important in protein transport in other eukaryotic organisms, but the organelles associated with a classical secretory pathway and protein transport are difficult to discern at an ultra-structural level[116]. In order to identify additional proteins that may have a role in protein translocation and secretion, the *P. falciparum* protein database was searched with *S. cerevisiae* proteins with GO assignments for involvement in protein export. We identified potential homologues of important components of the signal recognition particle, the translocon, the signal peptidase complex and many components that allow vesicle assembly, docking and fusion, such as COPI and COPII, clathrin, adaptin, v- and t-SNARE and GTP binding proteins. The presence of Sec62 and Sec63 orthologues raises the possibility of post-translational translocation of proteins, as found in *S. cerevisiae*.

Although *P. falciparum* contains many of the components associated with a classical secretory system and vesicular transport of proteins, the parasite secretory pathway has unusual features. The parasite develops within a parasitophorous vacuole that is formed during the invasion of the host cell, and the parasite modifies the host erythrocyte by the export of parasite-encoded proteins[117]. The mechanism(s) by which these proteins, some of which lack signal peptide sequences, are transported through and targeted beyond the membrane of the parasitophorous vacuole remains unknown. But these mechanisms are of particular importance because many of the proteins that contribute to the development of severe disease are exported to the cytoplasm and plasma membrane of infected erythrocytes.

Attempts to resolve these observations resulted in the proposal of a secondary secretory pathway[118]. More recent studies suggest export of COPII vesicle coat proteins, Sar1 and Sec31, to the erythrocyte cytoplasm as a mechanism of inducing vesicle formation in the host cell, thereby targeting parasite proteins beyond the parasitophorous vacuole, a new model in cell biology[119,120]. A homologue of *N*-ethylmaleimide-sensitive factor (NSF), a component of vesicular transport, has also been located to the erythrocyte cytoplasm[121]. The 41-2 antigen of *P. falciparum*, which is also found in the erythrocyte cytoplasm and plasma membrane[122], is homologous with BET3, a subunit of the *S. cerevisiae* transport protein particle (TRAPP) that mediates endoplasmic reticulum to Golgi vesicle docking and fusion[123]. It is not clear how these proteins are targeted to the cytoplasm, as they lack an obvious signal peptide. Nevertheless, the expanded list of protein-transport-associated genes identified in the *P. falciparum* genome should facilitate the development of specific probes to further elucidate the intra- and extracellular compartments of its protein transport system.

Immune evasion

In common with other organisms, highly variable gene families are clustered towards the telomeres. *Plasmodium falciparum* contains three such families termed *var*, *rif* and *stevor*, which code for proteins known as *P. falciparum* erythrocyte membrane protein 1 (PfEMP1), repetitive interspersed family (rifin) and sub-telomeric variable open reading frame (stevor), respectively[5,124–130]. The 3D7 genome contains 59 *var*, 149 *rif* and 28 *stevor* genes, but for each family there are also a number of pseudogenes and gene truncations present.

The *var* genes code for proteins which are exported to the surface of infected red blood cells where they mediate adherence to host endothelial receptors[131], resulting in the sequestration of infected cells in a variety of organs. These and other adherence properties[132–135] are important virulence factors that contribute to the development of severe disease. Rifins, products of the *rif* genes, are also expressed on the surface of infected red cells and undergo antigenic variation[131]. Proteins encoded by *stevor* genes show sequence similarity to rifins, but they are less polymorphic than the rifins[129]. The function of rifins and stevors is unknown. PfEMP1 proteins are targets of the host protective antibody response[136], but transcriptional switching between *var* genes permits antigenic variation and a means of immune evasion, facilitating chronic infection and transmission. Products of the *var* gene family are thus central to the pathogenesis of malaria and to the induction of protective immunity.

Figure 6 shows the genome-wide arrangement of these multigene families. In the 24 chromosomal ends that have a *var* gene as the first transcriptional unit, there are three basic types of gene arrangement. Eight have the general pattern *var-rif var* + /− $(rif/stevor)_n$, ten can be described as *var*-$(rif/stevor)_n$, three have a *var* gene alone and two have two or more adjacent *var* genes. This telomeric organization is consistent with exchange between chromosome ends, although the extent of this re-assortment may be limited by the varied gene combinations. The *var*, *rif* and *stevor* genes consist of two exons. The first *var* exon is between 3.5 and 9.0 kb in length, polymorphic and encodes an extracellular region of the protein. The second exon is between 1.0 and 1.5 kb, and encodes a conserved cytoplasmic tail that contains acidic amino-acid residues (ATS; 'acidic terminal sequence'). The first *rif* and *stevor* exons are about 50–75 bp in length, and encode a putative signal sequence while the second exon is about 1 kb in length, with the *rif* exon being on average slightly larger than that for *stevor*. The rifin sequences fall into two major

subgroups determined by the presence or absence of a consensus peptide sequence, KEL (X_{15}) IPTCVCR, approximately 100 amino acids from the N terminus. The *var* genes are made up of three recognizable domains known as 'Duffy binding like' (DBL); 'cysteine rich interdomain region' (CIDR) and 'constant2' (C2)[137–139]. Alignment of sequences existing before the *P. falciparum* genome project had placed each of these domains into a number of sub-classes; α to ε for DBL domains, and α to γ for CIDR domains. Despite these recognizable signatures, there is a low level of sequence similarity even between domains of the same sub-type. Alignment and tree construction of the DBL domains identified here showed that a small number did not fit well into existing categories, and have been termed DBL-X. Similar analysis of all 3D7 CIDR sequences showed that with this data they were best described as CIDRα or CIDR non-α, as distinct tree branches for the other domain types were not observed. In terms of domain type and order, 16 types of *var* gene sequences were identified in this study.

Type 1 *var* genes, consisting of DBLα, CIDRα, DBLδ, and CIDR non-α followed by the ATS, are the most common structures, with 38 genes in this category (Fig. 6b). A total of 58 *var* genes commence with a DBLα domain, and in 51 cases this is followed by CIDRα, and in 46 *var* genes the last domain of the first exon is CIDR non-α. Four *var* genes are atypical with the first exon consisting solely of DBL domains (type 3 and type 13). There is non-randomness in the ordering and pairing of DBL and CIDR sub-domains[140], suggesting that some—for example, DBLδ–CIDR non-α and DBLβ–C2 (Table 3)—should either be considered as functional–structural combinations, or that recombination in these areas is not favoured, thereby preserving the arrangement. Eighteen of the 24 telomeric proximal *var* genes are of type 1. With two exceptions, type 4 on chromosome 7 and type 9 on chromosome 11, all of the telomeric *var* genes are transcribed towards the centromere. The inverted position of the two *var* genes may hinder homologous recombination at these loci in telomeric clusters that are formed during asexual multiplication[115]. A further 12 *var* genes are located near to telomeres, with the remaining *var* genes forming internal clusters on chromosomes 4, 7, 8 and 12 and a single internal gene being located on chromosome 6.

Alignment of sequences 1.5 kb upstream of all of the *var* genes revealed three classes of sequences, upsA, upsB and upsC (of which there are 11, 35 and 13 members, respectively) that show preferential association with different *var* genes. Thus, upsB is associated with 22 out of 24 telomeric *var* genes, upsA is found with the two remaining telomeric *var* genes that are transcribed towards the telomere and with most telomere associated *var* genes (9 out of 12) which also point towards the telomere[141]. All 13 upsC sequences are associated with internal *var* clusters. Nearly all the telomeric *var* genes have an (A + T)-rich region approximately 2 kb upstream characterized by a number of poly(A) tracts as well as one or more copies of the consensus GGATCTAG. An analysis of the regions 1.0 kb downstream of *var* genes shows three sequence families, with members of one family being associated primarily with *var* genes

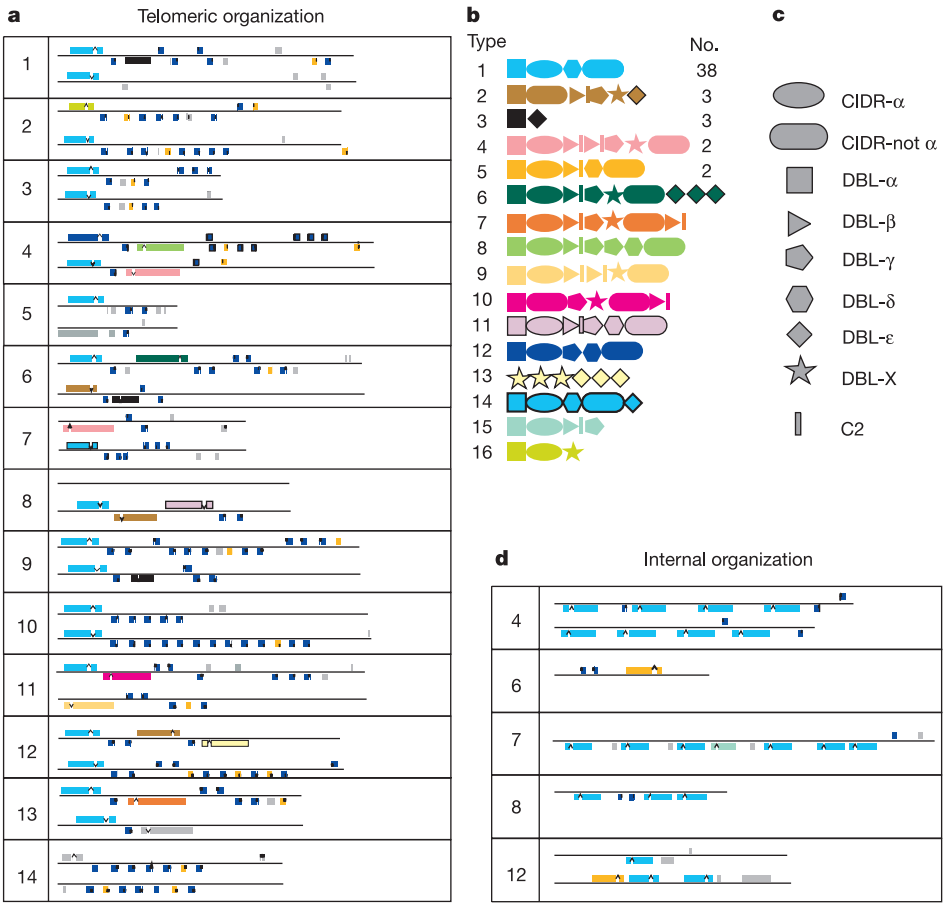

Figure 6 Organization of multi-gene families in *P. falciparum*. **a**, Telomeric regions of all chromosomes showing the relative positions of members of the multi-gene families: *rif* (blue) *stevor* (yellow) and *var* (colour coded as indicated; see **b** and **c**). Grey boxes represent pseudogenes or gene fragments of any of these families. The left telomere is shown above the right. Scale: ~0.6 mm = 1 kb. **b**, **c**, *var* gene domain structure. *var* genes contain three domain types: DBL, of which there are six sequence classes; CIDR, of which there are two sequence classes; and conserved 2 (C2) domains (see text). The relative order of the domains in each gene is indicated (**c**). *var* genes with the same domain types in the same order have been colour coded as an identical class and given an arbitrary number for their type (**b**) and the total number of members of each class in the genome of *P. falciparum* clone 3D7. **d**, Internal multi-gene family clusters. Key as in **a**.

next to the telomeric repeats. The intron sequences within the *var* genes have been associated with locus specific silencing[142]. They vary in length from 170 to ~1,200 bp and are ~89% A/T. On the coding strand, at the 5′ end the non-A/T bases are mainly G residues with 70% of sequences having the consensus TGTTTGGATATATA. The central regions are highly A-rich, and contain a number of semi-conserved motifs. The 3′ region is comparably rich in C, with one or more copies in most genes of the sequence $(TA)_n$ CCCATAAC-TACA. The 3′ end has an extended and atypical splice consensus of ACANATATAGTTA$(T)_n$ TAG. Sequences upstream of *rif* and *stevor* genes also have distinguishable upstream sequences, but a proportion of *rif* genes have the *stevor* type of 5′ sequence. Because the majority of telomeric *var* genes share a similar structure and 5′ and 3′ sequences, they may form a unique group in terms of regulation of gene expression.

The most conserved *var* gene previously identified, which mediates adherence to chondroitin sulphate A in the placenta[143], is incomplete in 3D7 because of deletion of part of exon 1 and all of exon 2. This gene is located on the right telomere of chromosome 5 (Fig 6). The majority of *var* genes sequenced previously had been identified as they mediated adhesion to particular receptors, and most of them had more than four domains in exon 1. The fact that type 1 *var* genes containing only 4 domains predominate in the 3D7 genome suggests that previous analyses had been based on a highly biased sample. The significance of this in terms of the function of type 1 *var* genes remains to be determined.

Immune-evasion mechanisms such as clonal antigenic variation of parasite-derived red cell surface proteins (PfEMP1s, rifins) and modulation of dendritic cell function have been documented in *P. falciparum*[131,132]. A putative homologue of human cytokine macrophage migration inhibitory factor (MIF) was identified in *P. falciparum*. In vertebrates, MIFs have been shown to function as immuno-modulators and as growth factors[144], and in the nematode *Brugia malayi*, recombinant MIF modulated macrophage migration and promoted parasite survival[145]. An MIF-type protein in *P. falciparum* may contribute to the parasite's ability to modulate the immune response by molecular mimicry or participate in other host–parasite interactions.

Implications for vaccine development

An effective malaria vaccine must induce protective immune responses equivalent to, or better than, those provided by naturally acquired immunity or immunization with attenuated sporozoites[146]. To date, about 30 *P. falciparum* antigens that were identified via conventional techniques are being evaluated for use in vaccines, and several have been tested in clinical trials. Partial protection with one vaccine has recently been attained in a field setting[147]. The present genome sequence will stimulate vaccine development by the identification of hundreds of potential antigens that could be scanned for desired properties such as surface expression or limited antigenic diversity. This could be combined with data on stage-specific expression obtained by microarray and proteomics[14,15] analyses to identify potential antigens that are expressed in one or more stages of the life cycle. However, high-throughput immunological assays to identify novel candidate vaccine antigens that are the targets of protective humoral and cellular immune responses in humans need to be developed if the genome sequence is to have an impact on vaccine development. In addition, new methods for maximizing the magnitude, quality and longevity of protective immune responses will be required in order to produce effective malaria vaccines.

Table 3 **Domains of PfEMP1 proteins in *P. falciparum***

Domain type	Number of domains
DBLα	58
DBLβ-C2	18
DBLγ	13
DBLδ	44
DBLε	13
DBL-X	13
CIDRα	51
CIDR non-α	54
Preferred pairings	**Frequency**
DBLα–CIDRα	51/58
DBLβ–C2	18/18
DBLδ–CIDR non-α	44/44
CIDRα–DBLδ	39/51
CIDRα–DBLβ	10/51
DBLβ–C2–DBLγ	10/18
DBLγ–DBL-X	8/13

Top, the total number of each DBL or CIDR domain type in intact *var* genes within the *P. falciparum* 3D7 genome. Bottom, the frequencies of the most common individual domain pairings found within intact *var* genes. The denominator refers to the total number of the first-named domains in intact *var* genes, and the numerator refers to the number of second-named domains found adjacent. See text for discussion of domain types.

Concluding remarks

The *P. falciparum*, *Anopheles gambiae* and *Homo sapiens* genome sequences have been completed in the past two years, and represent new starting points in the centuries-long search for solutions to the malaria problem. For the first time, a wealth of information is available for all three organisms that comprise the life cycle of the malaria parasite, providing abundant opportunities for the study of each species and their complex interactions that result in disease. The rapid pace of improvements in sequencing technology and the declining costs of sequencing have made it possible to begin genome sequencing efforts for *Plasmodium vivax*, the second major human malaria parasite, several malaria parasites of animals, and for many related parasites such as *Theileria* and *Toxoplasma*. These will be extremely useful for comparative purposes. Last, this technology will enable sampling of parasite, vector and host genomes in the field, providing information to support the development, deployment and monitoring of malaria control methods.

In the short term, however, the genome sequences alone provide little relief to those suffering from malaria. The work reported here and elsewhere needs to be accompanied by larger efforts to develop new methods of control, including new drugs and vaccines, improved diagnostics and effective vector control techniques. Much remains to be done. Clearly, research and investments to develop and implement new control measures are needed desperately if the social and economic impacts of malaria are to be relieved. The increased attention given to malaria (and to other infectious diseases affecting tropical countries) at the highest levels of government, and the initiation of programmes such as the Global Fund to Fight AIDS, Tuberculosis and Malaria[148], the Multilateral Initiative on Malaria in Africa[149], the Medicines for Malaria Venture[150], and the Roll Back Malaria campaign[151], provide some hope of progress in this area. It is our hope and expectation that researchers around the globe will use the information and biological insights provided by complete genome sequences to accelerate the search for solutions to diseases affecting the most vulnerable of the world's population. □

Methods

Sequencing, gap closure and annotation

The techniques used at each of the three participating centres for sequencing, closure and annotation are described in the accompanying Letters[7–9]. To ensure that each centres' annotation procedures produced roughly equivalent results, the Wellcome Trust Sanger Institute ('Sanger') and the Institute for Genomic Research ('TIGR') annotated the same 100-kb segment of chromosome 14. The number of genes predicted in this sequence by the two centres was 22 and 23; the discrepancy being due to the merging of two single genes by one centre. Of the 74 exons predicted by the two centres, 50 (68%) were identical, 9 (2%) overlapped, 6 (8%) overlapped and shared one boundary, and the remainder were predicted by one centre but not the other. Thus 88% of the exons predicted by the two centres in the 100-kb fragment were identical or overlapped.

Finished sequence data and annotation were transferred in XML (extensible markup language) format from Sanger and the Stanford Genome Technology Center to TIGR, and

made available to co-authors over the internet. Genes on finished chromosomes were assigned systematic names according the scheme described previously[5]. Genes on unfinished chromosomes were given temporary identifiers.

Analysis of subtelomeric regions

Subtelomeric regions were analysed by the alignment of all of the chromosomes to each other using MUMmer2[152] with a minimum exact match length ranging from 30 to 50 bp. Tandem repeats were identified by extracting a 90-kb region from the ends of all chromosomes and using Tandem Repeat Finder[153] with the following parameter settings: match = 2, mismatch = 7, indel = 7, pm = 75, pi = 10, minscore = 100, maxperiod = 500. Detailed pairwise alignments of internal telomeric blocks were computed with the ssearch program from the Fasta3 package[154].

Evolutionary analyses

Plasmodium falciparum proteins were searched against a database of proteins from all complete genomes as well as from a set of organelle, plasmid and viral genomes. Putative recently duplicated genes were identified as those encoding proteins with better BLASTP matches (based on E value with a 10^{-15} cutoff) to other proteins in *P. falciparum* than to proteins in any other species. Proteins of possible organellar descent were identified as those for which one of the top six prokaryotic matches (based on E value) was to either a protein encoded by an organelle genome or by a species related to the organelle ancestors (members of the *Rickettsia* subgroup of the α-Proteobacteria or cyanobacteria). Because BLAST matches are not an ideal method of inferring evolutionary history, phylogenetic analysis was conducted for all these proteins. For phylogenetic analysis, all homologues of each protein were identified by BLASTP searches of complete genomes and of a non-redundant protein database. Sequences were aligned using CLUSTALW, and phylogenetic trees were inferred using the neighbour-joining algorithms of CLUSTALW and PHYLIP. For comparative analysis of eukaryotes, the proteomes of all eukaryotes for which complete genomes are available (except the highly reduced *E. cuniculi*) were searched against each other. The proportion of proteins in each eukaryotic species that had a BLASTP match in each of the other eukaryotic species was determined, and used to infer a 'whole-genome tree' using the neighbour-joining algorithm. Possible eukaryotic conserved and specific proteins were identified as those with matches to all the complete eukaryotic genomes (10^{-30} E-value cutoff) but without matches to any complete prokaryotic genome (10^{-15} cutoff).

Received 31 July; accepted 2 September 2002; doi:10.1038/nature01097.

1. Breman, J. G. The ears of the hippopotamus: manifestations, determinants, and estimates of the malaria burden. *Am. J. Trop. Med. Hyg.* **64,** 1–11 (2001).
2. Greenwood, B. & Mutabingwa, T. Malaria in 2002. *Nature* **415,** 670–672 (2002).
3. Gallup, J. L. & Sachs, J. D. The economic burden of malaria. *Am. J. Trop. Med. Hyg.* **64,** 85–96 (2001).
4. Hoffman, S. L. *et al.* Funding for malaria genome sequencing. *Nature* **387,** 647 (1997).
5. Gardner, M. J. *et al.* Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science* **282,** 1126–1132 (1998).
6. Bowman, S. *et al.* The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*. *Nature* **400,** 532–538 (1999).
7. Hall, N. *et al.* Sequence of *Plasmodium falciparum* chromosomes 1, 3–9 and 13. *Nature* **419,** 527–531 (2002).
8. Gardner, M. J. *et al.* Sequence of *Plasmodium falciparum* chromosomes 2, 10, 11 and 14. *Nature* **419,** 531–534 (2002).
9. Hyman, R. W. *et al.* Sequence of *Plasmodium falciparum* chromosome 12. *Nature* **419,** 534–537 (2002).
10. Foster, J. & Thompson, J. The *Plasmodium falciparum* genome project: a resource for researchers. *Parasitol. Today* **11,** 1–4 (1995).
11. Su, X. *et al.* A genetic map and recombination parameters of the human malaria parasite *Plasmodium falciparum*. *Science* **286,** 1351–1353 (1999).
12. Su, X. Z. & Wellems, T. E. Toward a high-resolution *Plasmodium falciparum* linkage map: polymorphic markers from hundreds of simple sequence repeats. *Genomics* **33,** 430–444 (1996).
13. Lai, Z. *et al.* A shotgun optical map of the entire *Plasmodium falciparum* genome. *Nature Genet.* **23,** 309–313 (1999).
14. Florens, L. *et al.* A proteomic view of the *Plasmodium falciparum* life cycle. *Nature* **419,** 520–526 (2002).
15. Lasonder, E. *et al.* Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry. *Nature* **419,** 537–542 (2002).
16. Watanabe, J., Sasaki, M., Suzuki, Y. & Sugano, S. FULL-malaria: a database for a full-length enriched cDNA library from human malaria parasite, *Plasmodium falciparum*. *Nucleic Acids Res.* **29,** 70–71 (2001).
17. Gamain, B. *et al.* Increase in glutathione peroxidase activity in malaria parasite after selenium supplementation. *Free Radic. Biol. Med.* **21,** 559–565 (1996).
18. Katinka, M. D. *et al.* Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature* **414,** 450–453 (2001).
19. Moriyama, E. N. & Powell, J. R. Codon usage bias and tRNA abundance in *Drosophila*. *J. Mol. Evol.* **45,** 514–523 (1997).
20. Duret, L. tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes. *Trends Genet.* **16,** 287–289 (2000).
21. Vaidya, A. B., Akella, R. & Suplick, K. Sequences similar to genes for two mitochondrial proteins and portions of ribosomal RNA in tandemly arrayed 6-kilobase-pair DNA of a malaria parasite. *Mol. Biochem. Parasitol.* **35,** 97–107 (1989).
22. Vaidya, A. B., Lashgari, M. S., Pologe, L. G. & Morrisey, J. Structural features of *Plasmodium* cytochrome *b* that may underlie susceptibility to 8-aminoquinolines and hydroxynaphthoquinones. *Mol. Biochem. Parasitol.* **58,** 33–42 (1993).
23. Tan, T. H., Pach, R., Crausaz, A., Ivens, A. & Schneider, A. tRNAs in *Trypanosoma brucei*: genomic organization, expression, and mitochondrial import. *Mol. Cell. Biol.* **22,** 3707–3717 (2002).
24. Tarassov, I. A. & Martin, R. P. Mechanisms of tRNA import into yeast mitochondria: an overview. *Biochimie* **78,** 502–510 (1996).
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. Li, J., Wirtz, R. A., McConkey, G. A., Sattabongkot, J. & McCutchan, T. F. Transition of *Plasmodium vivax* ribosome types corresponds to sporozoite differentiation in the mosquito. *Mol. Biochem. Parasitol.* **65,** 283–289 (1994).
27. Waters, A. P. The ribosomal RNA genes of *Plasmodium*. *Adv. Parasitol.* **34,** 33–79 (1994).
28. Babiker, H. A., Creasey, A. M., Bayoumi, R. A., Walliker, D. & Arnot, D. E. Genetic diversity of *Plasmodium falciparum* in a village in eastern Sudan. 2. Drug resistance, molecular karyotypes and the mdr1 genotype of recent isolates. *Trans. R. Soc. Trop. Med. Hyg.* **85,** 578–583 (1991).
29. Hinterberg, K., Mattei, D., Wellems, T. E. & Scherf, A. Interchromosomal exchange of a large subtelomeric segment in a *Plasmodium falciparum* cross. *EMBO J.* **13,** 4174–4180 (1994).
30. Hernandez, R. R., Hinterberg, K. & Scherf, A. Compartmentalization of genes coding for immunodominant antigens to fragile chromosome ends leads to dispersed subtelomeric gene families and rapid gene evolution in *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **78,** 137–148 (1996).
31. Scherf, A. *et al.* Gene inactivation of Pf11-1 of *Plasmodium falciparum* by chromosome breakage and healing: identification of a gametocyte-specific protein with a potential role in gametocytogenesis. *EMBO J.* **11,** 2293–2301 (1992).
32. Day, K. P. *et al.* Genes necessary for expression of a virulence determinant and for transmission of *Plasmodium falciparum* are located on a 0.3-megabase region of chromosome 9. *Proc. Natl Acad. Sci. USA* **90,** 8292–8296 (1993).
33. Pologe, L. G. & Ravetch, J. V. A chromosomal rearrangement in a *P. falciparum* histidine-rich protein gene is associated with the knobless phenotype. *Nature* **322,** 474–477 (1986).
34. Louis, E. J., Naumova, E. S., Lee, A., Naumov, G. & Haber, J. E. The chromosome end in yeast: its mosaic nature and influence on recombinational dynamics. *Genetics* **136,** 789–802 (1994).
35. van Deutekom, J. C. *et al.* Evidence for subtelomeric exchange of 3.3 kb tandemly repeated units between chromosomes 4q35 and 10q26: implications for genetic counselling and etiology of FSHD1. *Hum. Mol. Genet.* **5,** 1997–2003 (1996).
36. Rudenko, G., McCulloch, R., Dirks-Mulder, A. & Borst, P. Telomere exchange can be an important mechanism of variant surface glycoprotein gene switching in *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* **80,** 65–75 (1996).
37. 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).
38. Scherf, A., Figueiredo, L. M. & Freitas-Junior, L. H. *Plasmodium* telomeres: a pathogen's perspective. *Curr. Opin. Microbiol.* **4,** 409–414 (2001).
39. Vernick, K. D. & McCutchan, T. F. Sequence and structure of a *Plasmodium falciparum* telomere. *Mol. Biochem. Parasitol.* **28,** 85–94 (1988).
40. Oquendo, P. *et al.* Characterisation of a repetitive DNA sequence from the malaria parasite, *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **18,** 89–101 (1986).
41. 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).
42. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nature Genet.* **25,** 25–29 (2000).
43. McFadden, G. I., Reith, M., Munhollan, J. & Lang-Unnasch, N. Plastid in human parasites. *Nature* **381,** 482–483 (1996).
44. Kohler, S. *et al.* A plastid of probable green algal origin in apicomplexan parasites. *Science* **275,** 1485–1489 (1997).
45. Fichera, M. E. & Roos, D. S. A plastid organelle as a drug target in apicomplexan parasites. *Nature* **390,** 407–409 (1997).
46. He, C. Y., Striepen, B., Pletcher, C. H., Murray, J. M. & Roos, D. S. Targeting and processing of nuclear-encoded apicoplast proteins in plastid segregation mutants of *Toxoplasma gondii*. *J. Biol. Chem.* **276,** 28436–28442 (2001).
47. 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).
48. Surolia, N. & Surolia, A. Triclosan offers protection against blood stages of malaria by inhibiting enoyl-ACP reductase of *Plasmodium falciparum*. *Nature Med.* **7,** 167–173 (2001).
49. Jomaa, H. *et al.* Inhibitors of the nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs. *Science* **285,** 1573–1576 (1999).
50. Sato, S. & Wilson, R. J. The genome of *Plasmodium falciparum* encodes an active delta-aminolevulinic acid dehydratase. *Curr. Genet.* **40,** 391–398 (2002).
51. Van Dooren, G. G., Su, V., D'Ombrain, M. C. & McFadden, G. I. Processing of an apicoplast leader sequence in *Plasmodium falciparum* and the identification of a putative leader cleavage enzyme. *J. Biol. Chem.* **277,** 23612–23619 (2002).
52. Wilson, R. J. Progress with parasite plastids. *J. Mol. Biol.* **319,** 257–274 (2002).
53. Stoebe, B. & Kowallik, K. V. Gene-cluster analysis in chloroplast genomics. *Trends Genet.* **15,** 344–347 (1999).
54. Fast, N. M., Kissinger, J. C., Roos, D. S. & Keeling, P. J. Nuclear-encoded, plastid-targeted genes suggest a single common origin for apicomplexan and dinoflagellate plastids. *Mol. Biol. Evol* **18,** 418–426 (2001).
55. Roos, D. S. *et al.* Origin, targeting, and function of the apicomplexan plastid. *Curr. Opin. Microbiol.* **2,** 426–432 (1999).
56. Palmer, J. D. & Delwiche, C. F. Second-hand chloroplasts and the case of the disappearing nucleus. *Proc. Natl Acad. Sci. USA* **93,** 7432–7435 (1996).
57. Waller, R. F., Reed, M. B., Cowman, A. F. & McFadden, G. I. Protein trafficking to the plastid of *Plasmodium falciparum* is via the secretory pathway. *EMBO J.* **19,** 1794–1802 (2000).
58. DeRocher, A., Hagen, C. B., Froehlich, J. E., Feagin, J. E. & Parsons, M. Analysis of targeting sequences demonstrates that trafficking to the *Toxoplasma gondii* plastid branches off the secretory system. *J. Cell Sci.* **113** (Part 22), 3969–3977 (2000).
59. van Dooren, G. G., Schwartzbach, S. D., Osafune, T. & McFadden, G. I. Translocation of proteins across the multiple membranes of complex plastids. *Biochim. Biophys. Acta* **1541,** 34–53 (2001).

60. Yung, S., Unnasch, T. R. & Lang-Unnasch, N. Analysis of apicoplast targeting and transit peptide processing in *Toxoplasma gondii* by deletional and insertional mutagenesis. *Mol. Biochem. Parasitol.* **118,** 11–21 (2001).
61. Zuegge, J., Ralph, S., Schmuker, M., McFadden, G. I. & Schneider, G. Deciphering apicoplast targeting signals—feature extraction from nuclear-encoded precursors of *Plasmodium falciparum* apicoplast proteins. *Gene* **280,** 19–26 (2001).
62. Vollmer, M., Thomsen, N., Wiek, S. & Seeber, F. Apicomplexan parasites possess distinct nuclear-encoded, but apicoplast-localized, plant-type ferredoxin-$NADP^+$ reductase and ferredoxin. *J. Biol. Chem.* **276,** 5483–5490 (2001).
63. Ralph, S. A., D'Ombrain, M. C. & McFadden, G. I. The apicoplast as an antimalarial drug target. *Drug Resist. Updat.* **4,** 145–151 (2001).
64. Baldauf, S. L., Roger, A. J., Wenk-Siefert, I. & Doolittle, W. F. A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science* **290,** 972–977 (2000).
65. Wood, V. *et al.* The genome sequence of *Schizosaccharomyces pombe*. *Nature* **415,** 871–880 (2002).
66. Eisen, J. A. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.* **8,** 163–167 (1998).
67. Adams, K. L., Daley, D. O., Whelan, J. & Palmer, J. D. Genes for two mitochondrial ribosomal proteins in flowering plants are derived from their chloroplast or cytosolic counterparts. *Plant Cell* **14,** 931–943 (2002).
68. Kanehisa, M., Goto, S., Kawashima, S. & Nakaya, A. The KEGG databases at GenomeNet. *Nucleic Acids Res.* **30,** 42–46 (2002).
69. Sherman, I. W. in *Malaria Parasite Biology, Pathogenesis, and Protection* (ed. Sherman, I. W.) 135–143 (ASM, Washington DC, 1998).
70. Buckwitz, D., Jacobasch, G., Gerth, C., Holzhutter, H. G. & Thamm, R. A kinetic model of phosphofructokinase from *Plasmodium berghei*. Influence of ATP and fructose-6-phosphate. *Mol. Biochem. Parasitol.* **27,** 225–232 (1988).
71. Buckwitz, D., Jacobasch, G. & Gerth, C. Phosphofructokinase from *Plasmodium berghei*. Influence of Mg^{2+}, ATP and Mg^{2+}-complexed ATP. *Biochem. J.* **267,** 353–357 (1990).
72. Clarke, J. L., Scopes, D. A., Sodeinde, O. & Mason, P. J. Glucose-6-phosphate dehydrogenase-6-phosphogluconolactonase. A novel bifunctional enzyme in malaria parasites. *Eur. J. Biochem.* **268,** 2013–2019 (2001).
73. Miclet, E. *et al.* NMR spectroscopic analysis of the first two steps of the pentose-phosphate pathway elucidates the role of 6-phosphogluconolactonase. *J. Biol. Chem.* **276,** 34840–34846 (2001).
74. Loyevsky, M. *et al.* An IRP-like protein from *Plasmodium falciparum* binds to a mammalian iron-responsive element. *Blood* **98,** 2555–2562 (2001).
75. Lang-Unnasch, N. Purification and properties of *Plasmodium falciparum* malate dehydrogenase. *Mol. Biochem. Parasitol.* **50,** 17–25 (1992).
76. Blum, J. J. & Ginsburg, H. Absence of α-ketoglutarate dehydrogenase activity and presence of CO_2-fixing activity in *Plasmodium falciparum* grown *in vitro* in human erythrocytes. *J. Protozool.* **31,** 167–169 (1984).
77. Fry, M. & Beesley, J. E. Mitochondria of mammalian *Plasmodium* spp. *Parasitology* **102,** 17–26 (1991).
78. Vaidya, A. B. in *Malaria: Parasite Biology, Pathogenesis, and Protection* (ed. Sherman, I. W.) 355–368 (ASM, Washington DC, 1998).
79. Papa, S., Zanotti, F. & Gaballo, A. The structural and functional connection between the catalytic and proton translocating sectors of the mitochondrial F_1F_0-ATP synthase. *J. Bioenerg. Biomembr.* **32,** 401–411 (2000).
80. Sherman, I. W. in *Malaria: Parasite Biology, Pathogenesis, and Protection* (ed. Sherman, I. W.) 177–184 (ASM, Washington DC, 1998).
81. de Macedo, C. S., Uhrig, M. L., Kimura, E. A. & Katzin, A. M. Characterization of the isoprenoid chain of coenzyme Q in *Plasmodium falciparum*. *FEMS Microbiol. Lett.* **207,** 13–20 (2002).
82. Trumpower, B. L. & Gennis, R. B. Energy transduction by cytochrome complexes in mitochondrial and bacterial respiration: the enzymology of coupling electron transfer reactions to transmembrane proton translocation. *Annu. Rev. Biochem.* **63,** 675–716 (1994).
83. Vaidya, A. B., McIntosh, M. T. & Srivastava, I. K. *Membrane Structure in Disease and Drug Therapy* (ed. Zimmer, G.) (Marcel Dekker, New York, 2000).
84. Perez-Martinez, X. *et al.* Subunit II of cytochrome c oxidase in Chlamydomonad algae is a heterodimer encoded by two independent nuclear genes. *J. Biol. Chem.* **276,** 11302–11309 (2001).
85. Murphy, A. D. & Lang-Unnasch, N. Alternative oxidase inhibitors potentiate the activity of atovaquone against *Plasmodium falciparum*. *Antimicrob. Agents Chemother.* **43,** 651–654 (1999).
86. Dieckmann, A. & Jung, A. Mechanisms of sulfadoxine resistance in *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **19,** 143–147 (1986).
87. McConkey, G. A. Targeting the shikimate pathway in the malaria parasite *Plasmodium falciparum*. *Antimicrob. Agents Chemother.* **43,** 175–177 (1999).
88. Roberts, F. *et al.* Evidence for the shikimate pathway in apicomplexan parasites. *Nature* **393,** 801–805 (1998).
89. Roberts, C. W. *et al.* The shikimate pathway and its branches in apicomplexan parasites. *J. Infect. Dis.* **185** (Suppl. 1), S25–S36 (2002).
90. Keeling, P. J. *et al.* Shikimate pathway in apicomplexan parasites. *Nature* **397,** 219–220 (1999).
91. Fitzpatrick, T. *et al.* Subcellular localization and characterization of chorismate synthase in the apicomplexan *Plasmodium falciparum*. *Mol. Microbiol.* **40,** 65–75 (2001).
92. Duncan, K., Edwards, R. M. & Coggins, J. R. The pentafunctional arom enzyme of *Saccharomyces cerevisiae* is a mosaic of monofunctional domains. *Biochem. J.* **246,** 375–386 (1987).
93. Rubin, H. *et al.* Cloning, sequence determination, and regulation of the ribonucleotide reductase subunits from *Plasmodium falciparum*: a target for antimalarial therapy. *Proc. Natl Acad. Sci. USA* **90,** 9280–9284 (1993).
94. Chakrabarti, D., Schuster, S. M. & Chakrabarti, R. Cloning and characterization of subunit genes of ribonucleotide reductase, a cell-cycle-regulated enzyme, from *Plasmodium falciparum*. *Proc. Natl Acad. Sci. USA* **90,** 12020–12024 (1993).
95. Krnajski, Z., Gilberger, T. W., Walter, R. D. & Muller, S. The malaria parasite *Plasmodium falciparum* possesses a functional thioredoxin system. *Mol. Biochem. Parasitol.* **112,** 219–228 (2001).
96. Bonday, Z. Q., Dhanasekaran, S., Rangarajan, P. N. & Padmanaban, G. Import of host δ-aminolevulinate dehydratase into the malarial parasite: identification of a new drug target. *Nature Med.* **6,** 898–903 (2000).
97. Bonday, Z. Q., Taketani, S., Gupta, P. D. & Padmanaban, G. Heme biosynthesis by the malarial parasite. Import of δ-aminolevulinate dehydrase from the host red cell. *J. Biol. Chem.* **272,** 21839–21846 (1997).
98. Wilson, C. M., Smith, A. B. & Baylon, R. V. Characterization of the δ-aminolevulinate synthase gene homologue in *P. falciparum*. *Mol. Biochem. Parasitol.* **75,** 271–276 (1996).
99. Sato, S., Tews, I. & Wilson, R. J. Impact of a plastid-bearing endocytobiont on apicomplexan genomes. *Int. J. Parasitol.* **30,** 427–439 (2000).
100. Rohdich, F. *et al.* Biosynthesis of terpenoids. 2C-Methyl-D-erythritol 2,4-cyclodiphosphate synthase (IspF) from *Plasmodium falciparum*. *Eur. J. Biochem.* **268,** 3190–3197 (2001).
101. Kemp, L. E., Bond, C. S. & Hunter, W. N. Structure of 2C-methyl-D-erythritol 2,4-cyclodiphosphate synthase: an essential enzyme for isoprenoid biosynthesis and target for antimicrobial drug development. *Proc. Natl Acad. Sci. USA* **99,** 6591–6596 (2002).
102. Paulsen, I. T., Nguyen, L., Sliwinski, M. K., Rabus, R. & Saier, M. H. Jr Microbial genome analyses: comparative transport capabilities in eighteen prokaryotes. *J. Mol. Biol.* **301,** 75–100 (2000).
103. Woodrow, C. J., Burchmore, R. J. & Krishna, S. Hexose permeation pathways in *Plasmodium falciparum*-infected erythrocytes. *Proc. Natl Acad. Sci. USA* **97,** 9931–9936 (2000).
104. Hansen, M., Kun, J. F., Schultz, J. E. & Beitz, E. A single, bi-functional aquaglyceroporin in blood-stage *Plasmodium falciparum* malaria parasites. *J. Biol. Chem.* **277,** 4874–4882 (2002).
105. Elliott, J. L., Saliba, K. J. & Kirk, K. Transport of lactate and pyruvate in the intraerythrocytic malaria parasite, *Plasmodium falciparum*. *Biochem. J.* **355,** 733–739 (2001).
106. Rager, N., Mamoun, C. B., Carter, N. S., Goldberg, D. E. & Ullman, B. Localization of the *Plasmodium falciparum* PfNT1 nucleoside transporter to the parasite plasma membrane. *J. Biol. Chem.* **276,** 41095–41099 (2001).
107. Dyer, M., Wong, I. H., Jackson, M., Huynh, P. & Mikkelsen, R. Isolation and sequence analysis of a cDNA encoding an adenine nucleotide translocator from *Plasmodium falciparum*. *Biochim. Biophys. Acta* **1186,** 133–136 (1994).
108. McIntosh, M. T., Drozdowicz, Y. M., Laroiya, K., Rea, P. A. & Vaidya, A. B. Two classes of plant-like vacuolar-type H^+-pyrophosphatases in malaria parasites. *Mol. Biochem. Parasitol.* **114,** 183–195 (2001).
109. Fidock, A. D. *et al.* Mutations in the *P. falciparum* digestive vacuole transmembrane protein PfCRT and evidence for their role in chloroquine resistance. *Mol. Cell* **6,** 861–871 (2000).
110. Desai, S. A., Bezrukov, S. M. & Zimmerberg, J. A voltage-dependent channel involved in nutrient uptake by red blood cells infected with the malaria parasite. *Nature* **406,** 1001–1005 (2000).
111. Eisen, J. A. & Hanawalt, P. C. A phylogenomic study of DNA repair genes, proteins, and processes. *Mutat. Res.* **435,** 171–213 (1999).
112. Wood, R. D., Mitchell, M., Sgouros, J. & Lindahl, T. Human DNA repair genes. *Science* **291,** 1284–1289 (2001).
113. Haltiwanger, B. M. *et al.* DNA base excision repair in human malaria parasites is predominantly by a long-patch pathway. *Biochemistry* **39,** 763–772 (2000).
114. Critchlow, S. E. & Jackson, S. P. DNA end-joining: from yeast to man. *Trends Biochem. Sci.* **23,** 394–398 (1998).
115. Freitas-Junior, L. H. *et al.* Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of *P. falciparum*. *Nature* **407,** 1018–1022 (2000).
116. Bannister, L. H., Hopkins, J. M., Fowler, R. E., Krishna, S. & Mitchell, G. H. A brief illustrated guide to the ultrastructure of *Plasmodium falciparum* asexual blood stages. *Parasitol. Today* **16,** 427–433 (2000).
117. van Dooren, G. G., Waller, R. F., Joiner, K. A., Roos, D. S. & McFadden, G. I. Traffic jams: protein transport in *Plasmodium falciparum*. *Parasitol. Today* **16,** 421–427 (2000).
118. Wiser, M. F., Lanners, H. N., Bafford, R. A. & Favaloro, J. M. A novel alternate secretory pathway for the export of *Plasmodium* proteins into the host erythrocyte. *Proc. Natl Acad. Sci. USA* **94,** 9108–9113 (1997).
119. Albano, F. R. *et al.* A homologue of Sar1p localises to a novel trafficking pathway in malaria-infected erythrocytes. *Eur. J. Cell Biol.* **78,** 453–462 (1999).
120. Adisa, A., Albano, F. R., Reeder, J., Foley, M. & Tilley, L. Evidence for a role for a *Plasmodium falciparum* homologue of Sec31p in the export of proteins to the surface of malaria parasite-infected erythrocytes. *J. Cell Sci.* **114,** 3377–3386 (2001).
121. Hayashi, M. *et al.* A homologue of N-ethylmaleimide-sensitive factor in the malaria parasite *Plasmodium falciparum* is exported and localized in vesicular structures in the cytoplasm of infected erythrocytes in the brefeldin A-sensitive pathway. *J. Biol. Chem.* **276,** 15249–15255 (2001).
122. Knapp, B., Hundt, E. & Kupper, H. A. A new blood stage antigen of *Plasmodium falciparum* transported to the erythrocyte surface. *Mol. Biochem. Parasitol.* **37,** 47–56 (1989).
123. Sacher, M. *et al.* TRAPP, a highly conserved novel complex on the cis-Golgi that mediates vesicle docking and fusion. *EMBO J.* **17,** 2494–2503 (1998).
124. 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).
125. Weber, J. L. Interspersed repetitive DNA from *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **29,** 117–124 (1988).
126. Su, Z. *et al.* The large diverse gene family *var* encodes proteins involved in cytoadherence and antigenic variation of *Plasmodium falciparum*-infected erythrocytes. *Cell* **82,** 89–100 (1995).
127. Baruch, D. I. *et al.* Cloning the *P. falciparum* gene encoding PfEMP1, a malarial variant antigen and adherence receptor on the surface of parasitized human erythrocytes. *Cell* **82,** 77–87 (1995).
128. Smith, J. D. *et al.* Switches in expression of *Plasmodium falciparum var* genes correlate with changes in antigenic and cytoadherent phenotypes of infected erythrocytes. *Cell* **82,** 101–110 (1995).
129. 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).
130. Kyes, S. A., Rowe, J. 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* **96,** 9333–9338 (1999).
131. Kyes, S., Horrocks, P. & Newbold, C. Antigenic variation at the infected red cell surface in malaria. *Annu. Rev. Microbiol.* **55,** 673–707 (2001).
132. Urban, B. C. *et al. Plasmodium falciparum*-infected erythrocytes modulate the maturation of dendritic cells. *Nature* **400,** 73–77 (1999).
133. Pain, A. *et al.* Platelet-mediated clumping of *Plasmodium falciparum*-infected erythrocytes is a common adhesive phenotype and is associated with severe malaria. *Proc. Natl Acad. Sci. USA* **98,** 1805–1810 (2001).

134. Fried, M. & Duffy, P. E. Adherence of *Plasmodium falciparum* to chondroitin sulfate A in the human placenta. *Science* **272,** 1502–1504 (1996).
135. Udomsangpetch, R. *et al. Plasmodium falciparum*-infected erythrocytes form spontaneous erythrocyte rosettes. *J. Exp. Med.* **169,** 1835–1840 (1989).
136. Bull, P. C. *et al.* Parasite antigens on the infected red cell surface are targets for naturally acquired immunity to malaria. *Nature Med.* **4,** 358–360 (1998).
137. Peterson, D. S., Miller, L. H. & Wellems, T. E. Isolation of multiple sequences from the *Plasmodium falciparum* genome that encode conserved domains homologous to those in erythrocyte binding proteins. *Proc. Natl Acad. Sci. USA* **92,** 7100–7104 (1995).
138. Baruch, D. I. *et al.* Identification of a region of PfEMP1 that mediates adherence of *Plasmodium falciparum* infected erythrocytes to CD36: conserved function with variant sequence. *Blood* **90,** 3766–3775 (1997).
139. Smith, J. D., Gamain, B., Baruch, D. I. & Kyes, S. Decoding the language of *var* genes and *Plasmodium falciparum* sequestration. *Trends Parasitol.* **17,** 538–545 (2001).
140. Smith, J. D. *et al.* Identification of a *Plasmodium falciparum* intercellular adhesion molecule-1 binding domain: a parasite adhesion trait implicated in cerebral malaria. *Proc. Natl Acad. Sci. USA* **97,** 1766–1771 (2000).
141. Voss, T. S. *et al.* Genomic distribution and functional characterisation of two distinct and conserved *Plasmodium falciparum var* gene 5′ flanking sequences. *Mol. Biochem. Parasitol.* **107,** 103–115 (2000).
142. Deitsch, K. W., Calderwood, M. S. & Wellems, T. E. Malaria. Cooperative silencing elements in *var* genes. *Nature* **412,** 875–876 (2001).
143. Rowe, J. A., Kyes, S. A., Rogerson, S. J., Babiker, H. A. & Raza, A. Identification of a conserved *Plasmodium falciparum var* gene implicated in malaria in pregnancy. *J. Infect. Dis.* **185,** 1207–1211 (2002).
144. Lue, H., Kleemann, R., Calandra, T., Roger, T. & Bernhagen, J. Macrophage migration inhibitory factor (MIF): mechanisms of action and role in disease. *Microbes Infect.* **4,** 449–460 (2002).
145. Pastrana, D. V. *et al.* Filarial nematode parasites secrete a homologue of the human cytokine macrophage migration inhibitory factor. *Infect. Immun.* **66,** 5955–5963 (1998).
146. Richie, T. L. & Saul, A. Progress and challenges for malaria vaccines. *Nature* **415,** 694–701 (2002).
147. Bojang, K. A. *et al.* Efficacy of RTS,S/AS02 malaria vaccine against *Plasmodium falciparum* infection in semi-immune adult men in The Gambia: a randomised trial. *Lancet* **358,** 1927–1934 (2001).
148. Kapp, C. Global fund on AIDS, tuberculosis, and malaria holds first board meeting. *Lancet* **359,** 414 (2002).
149. Nchinda, T. C. Malaria: a reemerging disease in Africa. *Emerg. Infect. Dis.* **4,** 398–403 (1998).
150. Ridley, R. G. Medical need, scientific opportunity and the drive for antimalarial drugs. *Nature* **415,** 686–693 (2002).
151. Nabarro, D. N. & Tayler, E. M. The "roll back malaria" campaign. *Science* **280,** 2067–2068 (1998).
152. Delcher, A. L., Phillippy, A., Carlton, J. & Salzberg, S. L. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.* **30,** 2478–2483 (2002).
153. Benson, G. Tandem repeats finder: a program to analyse DNA sequences. *Nucleic Acids Res.* **27,** 573–580 (1999).
154. Pearson, W. R. Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.* **132,** 185–219 (2000).
155. Glockner, G. *et al.* Sequence and analysis of chromosome 2 of *Dictyostelium discoideum*. *Nature* **418,** 79–85 (2002).
156. Wood, V., Rutherford, K. M., Ivens, A., Rajandream, M. A. & Barrell, B. A re-annotation of the *Saccharomyces cerevisiae* genome. *Comp. Funct. Genom.* **2,** 143–154 (2001).
157. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408,** 796–815 (2000).
158. 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).
159. Scharfe, C. *et al.* MITOP, the mitochondrial proteome database: 2000 update. *Nucleic Acids Res.* **28,** 155–158 (2000).
160. Claros, M. G. & Vincens, P. Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur. J. Biochem.* **241,** 779–786 (1996).
161. 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).
162. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **30,** 276–280 (2002).
163. 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).
164. Nielsen, H., Engelbrecht, J., Brunak, S. & von Heijne, G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **10,** 1–6 (1997).
165. Carlton, J. M. *et al.* Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium Yoelii yoelii*. *Nature* **419,** 512–519 (2002).

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

Acknowledgements

We thank our colleagues at The Wellcome Trust Sanger Institute, The Institute for Genomic Research, the Stanford Genome Technology Center, and the Naval Medical Research Center for their support. We thank J. Foster for providing markers for chromosome 14; R. Huestis and K. Fischer for providing RT–PCR data for chromosomes 2 and 3 before publication; A. Waters for assistance with ribosomal RNAs; S. Cawley for assistance with phat; and M. Crawford and R. Wang for discussions. This work was supported by the Wellcome Trust, the Burroughs Wellcome Fund, the National Institute for Allergy and Infectious Diseases, the Naval Medical Research Center, and the US Army Medical Research and Materiel Command.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.J.G. (e-mail: gardner@tigr.org).

Sequences and annotation are available at the following websites: PlasmoDB (http://plasmodb.org), The Institute for Genomic Research (http://www.tigr.org), the Wellcome Trust Sanger Institute (http://www.sanger.ac.uk/Projects/Protozoa/), and the Stanford Genome Technology Center (http://www-sequence.stanford.edu/group/malaria). Chromosome sequences were submitted to EMBL or GenBank with accession numbers AL844501–AL844509 (chromosomes 1, 3–9 and 13), AE001362.2 (chromosome 2), AE014185–AE014187 (chromosomes 10, 11 and 14) and AE014188 (chromosome 12).