AD_____

Award Number: DAMD17-98-2-8005

TITLE: Malaria Genome Sequencing Project

PRINCIPAL INVESTIGATOR: Malcolm J. Gardner, Ph.D.

CONTRACTING ORGANIZATION: The Institute for Genomic Research
Rockville, MD 20850

REPORT DATE: January 2004

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	January 2004	Final(17 Dec 1997 - 16 Dec 2003)

4. TITLE AND SUBTITLE

Malaria Genome Sequencing Project

5. FUNDING NUMBERS

DAMD17-98-2-8005

6. AUTHOR(S)

Malcolm J. Gardner, Ph.D.

20040421 055

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

The Institute for Genomic Research
Rockville, MD 20850

E-Mail: gardner@tigr.org

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

Original contains color plates: ALL DTIC reproductions will be in black and white

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 Words)

The objectives of this Cooperative Agreement were: **Specific Aim 1,** sequence 3.5 Mb of *P. falciparum* genomic DNA; **Specific Aim 2,** annotate the sequence; **Specific Aim 3,** release the information to the scientific community. Two **Specific Aims** were added to the Cooperative Agreement: **Specific Aim 4,** sequencing of *P. yoelii* to 5X coverage; **Specific Aim 5,** sequencing of *P. vivax* to 5X coverage. In 2002 the complete genome sequence of *P. falciparum* was published in *Nature* in collaboration with the Sanger Institute and Stanford University. A comparative analysis of the genome of the rodent malaria parasite *P. yoelii* with that of *P. falciparum* was also published. The *P. vivax* genome has been sequenced to 10X coverage and genome closure is underway. Preliminary *P. vivax* sequence data has been released to the scientific community. Proteomic analyses of several stages of the *P. falciparum* life cycle by colleagues at the Scripps Research Institute were also supported by this Agreement. In addition, colleagues at the Malaria Program, Naval Medical Research Center have identified 700 genes expressed in *P. falciparum* sporozoites and 1,200 genes expressed in *P. yoelii* liver stages. Bioinformatic approaches were used to identify 300 genes encoding potential liver-stage antigens. The *P. falciparum*, *P. yoelii*, and *P. vivax* genome sequences generated by this effort have supported malaria research worldwide.

14. SUBJECT TERMS

Plasmodium falciparum, Plasmodium vivax, Plasmodium yoelii, malaria, Genome, genomics, chromosome, functional genomics, proteomics

15. NUMBER OF PAGES

178

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Table of Contents

Introduction

Malaria is caused by apicomplexan parasites of the genus *Plasmodium*. It is a major public health problem in many tropical areas of the world, and also affects many individuals and military forces that visit these areas. In 1994 the World Health Organization estimated that there were 300-500 million cases and up to 2.7 million deaths caused by malaria each year. Because of increased parasite resistance to chloroquine and other antimalarials the situation is expected to worsen considerably. These dire facts have stimulated efforts to develop an international, coordinated strategy for malaria research and control [1]. Development of new drugs and vaccines against malaria will undoubtedly be an important factor in control of the disease. However, despite recent progress, drug and vaccine development has been a slow and difficult process, hampered by the complex life cycle of the parasite, a limited number of drug and vaccine targets, and our incomplete understanding of parasite biology and host-parasite interactions.

The advent of microbial genomics, i.e. the ability to sequence and study the entire genomes of microbes, should accelerate the process of drug and vaccine development for microbial pathogens. As pointed out by Bloom, the complete genome sequence provides the "sequence of every virulence determinant, every protein antigen, and every drug target" in an organism [2], and establishes an excellent starting point for this process. In 1995, an international consortium including the National Institutes of Health, the Wellcome Trust, the Burroughs Wellcome Fund, and the US Department of Defense was formed (Malaria Genome Sequencing Project) to finance and coordinate genome sequencing of the human malaria parasite *Plasmodium falciparum*[3]. Later, because of the improvement in sequencing technologies that led to a dramatic reduction in sequencing costs, the consortium expanded its efforts to include other species of *Plasmodium*. Another major goal of the consortium was to foster close collaboration between members of the consortium and other agencies such as the World Health Organization, so that the knowledge generated by the Project could be rapidly applied to basic research and antimalarial drug and vaccine development programs worldwide. Participating centers included the Naval Medical Research Center, the Wellcome Trust Sanger Institute, and the Stanford University Genome Technology Center.

Body

This report describes progress in the Malaria Genome Sequencing Project achieved by The Institute for Genomic Research and the Malaria Program, Naval Medical Research Center, under Cooperative Research Agreement DAMD17-98-2-8005, over the period from Dec. '97 to Dec '03. The Specific Aims of the work supported by this agreement are listed below. Specific Aims 1-3 were contained

in the original Cooperative Agreement. Specific Aims 4-5 were added to the Cooperative Agreement through modifications.

The Cooperative Agreement was initially scheduled to expire in December 2002. However, we were granted a 12-month no-cost extension to allow us to complete a newly-expanded Specific Aim 5 (Sequencing of *P. vivax* to 5X coverage). The project concluded on December 16, 2003.

1. Determine the sequence of 3.5 megabases of the *P. falciparum* genome (clone 3D7):

a) Construct small-insert shotgun libraries (1-2 kb inserts) of chromosomal DNA isolated from preparative pulsed-field gels.

b) Sequence a sufficiently large number of randomly selected clones from a shotgun library to provide 10-fold coverage of the selected chromosome.

c) Construct P1 artificial chromosome (PAC) libraries (inserts up to 20 kb) of chromosomal DNA isolated from preparative pulsed-field gels.

d) If necessary, generate additional STS markers for the chromosome by i) mapping unique-sequence contigs derived from assembly of the random sequences to chromosome, ii) mapping end-sequences from chromosome-specific PAC clones to YACs.

e) Use TIGR Assembler to assemble random sequence fragments, and order contigs by comparison to the STS markers on each chromosome.

f) Close any remaining gaps in the chromosome sequence by PCR and primer-walking using *P. falciparum* genomic DNA or the YAC, BAC, or PAC clones from each chromosome as templates.

2. Analyze and annotate the genome sequence:

a) employ a variety of computer techniques to predict gene structures and relate them to known proteins by similarity searches against databases; identify untranslated features such as tRNA genes, rRNA genes, insertion sequences and repetitive elements; determine potential regulatory sequences and ribosome binding sites; use these data to identify metabolic pathways in *P. falciparum.*

3. Establish a publicly-accessible *P. falciparum* genome database and submit sequences to GenBank.

4. Perform whole genome shotgun sequencing of the rodent malaria parasite *Plasmodium yoelii* to 3X coverage, assemble into contigs,

annotate the contigs, make the data available on the TIGR web site, and submit the data to GenBank.

5. Perform whole genome shotgun sequencing of the human malaria parasite *Plasmodium vivax* to 5X coverage, assemble the contigs, annotate the contigs, make the data available on the TIGR web site, and submit the data to GenBank.

We are pleased to report that Specific Aims 1-4 have been completed. In previous annual reports we announced the publication in *Science* of the first complete sequence of a malarial chromosome (chromosome 2) [4]; development of a *Plasmodium* gene finding program, GlimmerM [5]; introduction of optical restriction mapping technology for rapid mapping of whole *Plasmodium* chromosomes [6,7]; the publication in *Nature* of the complete *Plasmodium falciparum* genome in collaboration with the Sanger Institute, Stanford University, the NMRC, and others[8,9]; a comparative analysis of the *P. falciparum* and the *P. yoelii* genome sequence at 5X coverage [10]; and an analysis of the *P. falciparum* proteome [11].

Specific Aim 5, to determine the genome sequence of *P. vivax* up to 5X coverage, is still underway. Due to rapidly declining sequencing costs we were able to obtain 9X genome coverage (almost twice what was anticipated), assemble the genome, and begin the gap closure process. Preliminary data has been released on the TIGR web site. We have applied for funding to complete the genome sequence from the Microbial Sequencing Centers program supported by the National Institute for Allergy and Infectious Diseases (NIAID). TIGR has been awarded a 5-year $65 million contract under this program, and if our proposal to complete the *P. vivax* genome is accepted, it is highly likely that the work will be done at TIGR.

Sequencing of *P. falciparum* chromosomes 2, 10, 11, and 14 (Specific Aims 1, 2, 3)

Sequencing of chromosomes 2, 10, 11, and 14 was funded primarily by grants from the NIAID (chromosomes 2, 10 and 11) and the Burroughs Wellcome Fund (chromosome 14). Funds from this collaborative agreement were used to accelerate the sequencing, assist in closure and annotation, and facilitate rapid utilization of the sequence data by the DoD vaccine and drug development groups. In previous years we described the isolation of chromosomal DNA, preparation of shotgun libraries, random sequencing, assembly, gap closure, production and public release of preliminary annotation (Annual Reports 1999-2002).

Annotation and publication of the *P. falciparum* genome sequence (Specific Aims 2 and 3)

Last year focused primarily on the closure of the last few gaps in the chromosomes and the final annotation and publication of the *P. falciparum* genome sequence in collaboration with the other members of the *P. falciparum* genome consortium (Annual Report 2003). After extensive discussions with counterparts at the Sanger Institute and Stanford University, an agreement to collaborate on the joint analysis and publication of entire *P. falciparum* genome sequence was reached. This whole genome overview was to be accompanied by a series of papers by each sequencing center on the chromosomes sequenced by each group. The whole genome overview and chromosome papers were to be published in a single issue of a journal. In addition, a comparative analysis of the *P. falciparum* and *P. yoelii* genomes based upon the 5X coverage of the *P. yoelii* sequence was to be be published along with the *P. falciparum* papers.

The principal investigator of this agreement was selected to be the coordinator of the annotation effort and the lead author on the final publication. Furthermore, TIGR was chosen to be the central repository of all the *P. falciparum* genome data. From January through June of 2002, TIGR collected the chromosome sequences and associated annotation from the other sequencing centers and coordinated the analysis of the genome sequence and the preparation of whole genome and a series of chromosome manuscripts for publication. The manuscripts were submitted for publication in July 2002 and published in *Nature* on Oct. 3, 2002 [8,9] (Annual Report 2003).

Sequencing of *P. yoelii* to 5X coverage (Specific Aim 4)

A secondary goal established at the initiation of the malaria genome project was to sequence the genome of a another species of *Plasmodium* so as to be able to perform a series of comparative analyses.

After discussions with NMRC, we elected to proceed with sequencing of *P. yoelii*. Reductions in the costs of sequencing allowed us to perform this work without requesting additional funds (Annual Report 2001). The genome was sequenced to 5X coverage and a comparative analysis with the *P. falciparum* genome was performed. This work was published in *Nature* on Oct. 3, 2002 [12] (Annual Report 2003).

Sequencing of *P. vivax* to 5X coverage (Specific Aim 5)

P. vivax is the second most important human malaria parasite. It cause 70-80 million cases of malaria each year and is responsible for over 50% of malaria cases in Central and South America, Asia and the Indian sub-continent

[13]. In the 2002 Annual Report, we described the addition of this Specific Aim to the Cooperative Agreement. We later obtained permission from NIAID to use surplus funds that remained from a cooperative agreement that supported the sequencing of *P. falciparum* chromosomes 10 and 11 (U01 AI42243; PI Malcolm Gardner) to obtain an additional 3X sequence coverage of the *P. vivax* genome (Annual Report 2003). This work was managed by Dr. Jane Carlton, an Associate Investigator in the Parasite Genomics Group at TIGR.

The Salvador I strain of *P. vivax*, isolated from a naturally acquired infection of a patient from El Salvador was chosen for sequencing [14]. This strain has been passaged through human volunteers and *Aotus* (owl) and *Saimiri* (squirrel) monkeys by mosquito and blood infection, it has been the subject of drug susceptibility and relapse activity studies [15], and it has been used to test the immunogenicity and protective efficacy of recombinant antigen constructs [16,17]. Salvador I chromosomes can be separated by pulsed-field gel electrophoresis for karyotype and physical mapping studies[18], and more than 7,000 genome survey sequences (GSSs) have been generated for this strain [19]. Thus, like the 3D7 clone of *P. falciparum*, it is often regarded as the standard reference strain for *P. vivax*. Genomic DNA was provided by John Barnwell at the Centers for Disease Control, from parasites grown in splenectomized *Saimiri* monkeys.

A whole genome shotgun strategy was used for the random sequencing part of the project. Mass sequencing of *P. vivax* genomic libraries started in the Spring of 2002. All shotgun sequence data have been released periodically during the project, and the final 10X coverage sequence data can be downloaded and searched via TIGR's *P. vivax* web pages (http://www.tigr.org/tdb/e2k1/pva1/ and Table 1).

At 10X coverage of the genome, random sequencing was stopped and closure of the gaps between the contigs commenced until all funds for the project were depleted in December 2003. Table 1 shows the current status of the genome sequence. Significant head-way was made into closing the genome, and currently almost 23 Mb of the predicted 25-26 Mb genome is contained in just 38 scaffolds (a scaffold is a group of ordered and orientated contigs known to be physically linked to each other by paired read information). Approximately 100 gaps between the contigs in scaffolds remain to be closed, in addition to 15 gaps between scaffolds that were identified through synteny studies with the *P. falciparum* genome. Five scaffolds that range in size from 328 kb to 1.9 Mb contain telomeric sequences at one end, indicating that at least five chromosome ends have been assembled. More than half of the genome has been pinned to a physical chromosome map. However, the low-resolution of the map, which contains only 18 genome markers, prevents any further mapping of the scaffolds. The complete 6 kb mitochondrial genome has been closed.

Table 1. Current status of the *P. vivax* genome sequence.

	At completion of random sequencing (4.2.03)	After 8 months of closure (12.16.03)
Total no. reads	378,878	383,267
Total no. contigs in scaffolds	1,837	1,094
Largest contig	694,650 bp	1,996,761 bp
Total no. scaffolds > 10 kb	71	38*
Largest scaffold	1,417,402 bp	2,119,814 bp
No. intra-scaffold gaps	825	108
No. telomeres identified	-	5

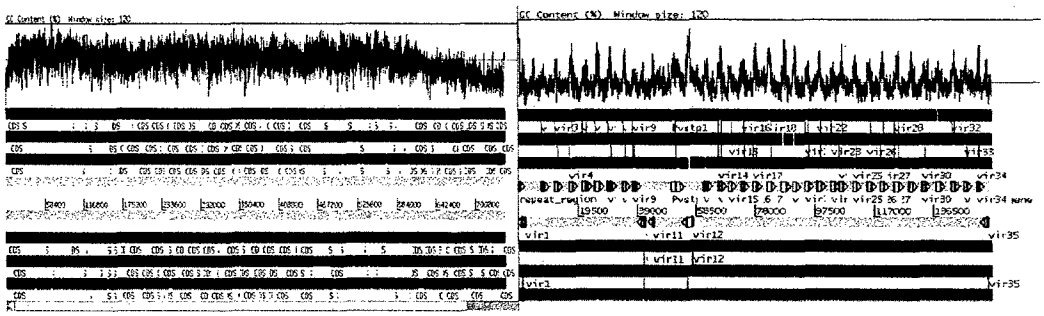
* These 38 scaffolds contain 22.89 Mb of the ~25 Mb genome

It has become apparent during the gap-closure process and since publication of two large contigs of the *P. vivax* genome, that the subtelomeric regions of *P. vivax* may be signifcantly more AT-rich than the internal conserved chromosome core. For example, the AT content of a 200 kb sequence from one chromosome of a field isolate was found to increase from 52% in the conserved core to 75% at the telomere proximal end [20], whereas a 150 kb contig containing a telomere from one chromosome was found to have a mainly uniform AT composition of ~79% [21]. The *vir* genes, important antigen genes implicated in antigenic variation and found in subtelomeric regions as mentioned above, are also known to be highly AT-rich (average AT ~ 79%). Figure 1 shows side-by-side comparisons of a telomeric contig, 3448, from the current genome data, compared with the published 150 kb telomeric contig [22]. It is clear that the GC content of contig 3448 is decreasing towards the telomeric region, but the length of the subtelomeric region is much less than in the 150 kb contig, indicating either that the complete subtelomeric region has not been sequenced, or the subtelomeric regions of the laboratory strain Salvador I are truncated compared to field isolates of *P. vivax*. Additional evidence for under-representation of the subtelomeric regions in the sequence data include the presence of many *vir* genes on short (average length 2.7 kb) contigs which cannot be linked through any paired read information to other contigs (approximately 65 contigs total with a combined length of 173 kb). This indicates that these non-coding regions are highly AT-rich and were not cloned efficiently in the shotgun libraries used for random sequencing, most likely due to their instability in plasmid vectors. Exactly how much subtelomeric sequence may be missing is not clear.

A "white paper" to request the funds required to complete the *P. vivax* genome sequence has been submitted to the NIAID under the Microbial Sequencing Centers program (http://www.niaid.nih.gov/dmid/genomes/mscs/default.htm).

Figure 1. Grahical representation of genome contig 3448 (left) and published 150 kb telomeric contig (right). The top graph is GC content plotted over the length of each contig; the horizontal line through each plot represents the mean GC, 46% for contig 3448 and 21% for the 150 kb contig. The six black horizontal lines beneath the plot represent the six reading frames; open reading frames are identified in blue. *vir* genes are annotated on the 150 kb contig.

Contig 3448 150 kb telomeric contig



Proteomics studies

A major goal of the malaria genome project is to identify antigens for vaccine development. Analysis of the genome sequence data can be used to identify potential antigens but does not by itself provide all of the information required for selection and prioritization of vaccine candidates. For example, the genome sequence itself does not specify at which point in the life cycle a gene is transcribed, or whether the protein product of a gene is actually present in the parasite. To identify proteins present in various stages of the parasite life cycle, we have begun to use proteomics techniques to directly identify parasite proteins in cell lysates.

In the 2002 and 2003 Annual Reports, we described studies that were performed by Dr. John Yates and colleagues at the Scripps Research Institute, partly funded by a subcontract from TIGR under this cooperative agreement. Briefly, proteins in parasite lysates were digested with proteases and the resulting peptides were separated by high-resolution liquid chromatography. The peptides were then injected into a tandem mass spectrometer. Spectra of each peptide were matched against predicted spectra of the peptides predicted from the genome sequence. In this way peptides generated from cell lysates were used to identify the proteins present in the cell lysate. Our role has been mainly to provide Dr. Yates's group with genomic sequence data from *P. falciparum* and *P. yoelii*, which they used to identify peptides derived from parasite lysates. Over 2,400 *P. falciparum* proteins, about 45% of the total proteins predicted from the genome sequence, were identified, including approx 500 proteins from

sporozoite stages[11]. The NMRC is using this data to select antigens for vaccine development (Annual Reports 2002 and 2003).

Functional Genomics

Funds from this Cooperative Agreement were transferred to the Malaria Program, NMCR under a CRADA amendment (NCRADA-NMRI-96NMR505). The funds were used to conduct functional genomics studies of *Plasmodium* to further identify candidate molecules for malaria vaccine development.

The NMRC, under the supervision of CAPT Daniel J. Carucci, performed the following activities:

1. Characterization of sporozoite ESTs. A cDNA library was constructed from *P. falciparum* sporozoites isolated from mosquitos salivary glands. The DNA sequencing of clones identified 700 expressed sequence tags (ESTs). Further analysis using reverse transcriptase PCR (RT-PCR) has verified the expression patterns of some of these genes. Also, a comparison to other databases of ESTs from different stages and species of parasite provided many insights into gene expression in sporozoites. A manuscript describing these findings is in preparation.

2. Identification of liver-stage ESTs. The transcriptional repertoire of the liver stage of *Plasmodium* has remained unknown due to the inaccessibility of these parasite stages. We overcame this hindrance by utilizing laser capture microdissection (LCM) to provide a high quality source of parasite mRNA for the construction of a liver stage cDNA library. Sequencing and annotation of this library demonstrated expression of over 1,200 *P. yoelii* genes during development in the hepatocyte. This is the first comprehensive analysis of gene expression undertaken for the liver stage of any malaria parasite, and provides insights into the differential expression of *P. yoelii* genes during this critical stage for vaccine development. Using comparative genomics we have identified hundreds of *P. falciparum* orthologs. A manuscript has been prepared and submitted for publication.

3. Bioinformatics. Computational approaches and bioinformatic analyses were used in combination with the various malaria large-scale databases such as *P. yoelli* liver stage EST, *P. falciparum* sporozoite EST, *P. falciparum* proteome [11], transcriptome [23] and genome [24] in order to identify and prioritize putative pre-erythrocytic genes. This selection process identified approximately 300 genes as potential novel vaccine targets.

4. Recombinational Cloning and Functional Analysis. We examined the feasibility of a high-throughput cloning approach using the Gateway

system (Invitrogen, Inc.) to create a large set of expression clones encoding *P. falciparum* single-exon genes. We have successfully optimized this cloning strategy to generate master DNA clones representing the 300 pre-erythrocytic genes identified above. These master clones were subsequently used to generate multiple "destination" Gateway constructs resulting in a complete set of recombinant clones in multiple types of expression vectors. Examples of such expression vectors include; DNA vaccine, recombinant protein expression, cell transfection, yeast-2-hybrid (Y2H). DNA vaccine constructs were used to immunize mice and raise antibodies for parasite protein localization studies. Recombinant protein expression constructs are being used in both *E. coli* and cell-free systems for the production of small-scale proteins. We have also generated a set of Y2H constructs that are currently being used in a comprehensive *P. falciparum* interactome project that will assist in elucidating the function of these malaria proteins. We have initiated a process of depositing these master clones to the MR4 repository to be available to the malaria research community. A manuscript has been prepared and submitted for publication.

Key Research Accomplishments

1. The sequences of chromosomes 2, 10, 11, and 14 were completed.
2. Chromosomes 2, 10, 11, and 14 were annotated at TIGR.
3. In collaboration with the Wellcome Trust Sanger Institute and Stanford University, the entire *P. falciparum* genome was annotated.
4. The *P. yoelii* genome sequence obtained at 5X coverage was annotated.
5. The *P. falciparum* and *P. yoelii* genome sequences were published in the Oct 3rd, 2002 issue of *Nature.*
6. Proteomic analyses of *P. falciparum* sporozoites, merozoites, trophozoites, and gametocytes were performed by John Yate's group at the Scripps Research Insitute under a subcontract to this award. The results were published in the Oct 3rd, 2002 issue of *Nature.*
7. Sequencing of the *P. vivax* genome reached 10X coverage. The genome was assembled and to date 87% of intrascaffold gaps have been closed. Preliminary data were released on the TIGR web site.
8. A *P. falciparum* sporozoite cDNA library was prepared and used to generate 700 ESTs.

9. Laser capture microdissection was used to isolate *P. yoelii* liver stage parasites and identify 1200 genes expressed in liver stages.

10. Bioinformatics approaches and genome and gene transcription data were used to identify 300 genes encoding potential pre-erythrocytic stage antigens.

11. The Gateway (Invitrogen, Inc.) recombinational cloning system was used to clone the 300 genes encoding potential pre-erythrocytic stage antigens. DNA vaccines encoding these antigens and recombinant proteins are being prepared in order to generate antibodies for protein localization studies.

Reportable Outcomes

Journal articles and book chapters

1. Gardner, M. J. in *Microbial Genomes* (eds. Fraser, C. M., Read, T. D. & Nelson, K. E.) (Humana Press, Totowa, In press).
2. Hall, N. & Gardner, M. J. in *Genomes and the Molecular Biology of Malaria Parasites* (ed. Waters, A. P.) (Horizon Scientific Press, Wymondham, In press).
3. Gardner, M. J. Sequencing the genome of the malaria parasite: the impact of genomics on health in developing countries. *Sust Dev Int* **Summer 2003**, 125-127 (2003).
4. Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R. W., Carlton, J. M., Pain, A., Nelson, K. E., Bowman, S., Paulsen, I. T., James, K., Eisen, J. A., Rutherford, K., Salzberg, S. L., Craig, A., Kyes, S., Chan, M. S., Nene, V., Shallom, S. J., Suh, B., Peterson, J., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., Haft, D., Mather, M. W., Vaidya, A. B., Martin, D. M., Fairlamb, A. H., Fraunholz, M. J., Roos, D. S., Ralph, S. A., McFadden, G. I., Cummings, L. M., Subramanian, G. M., Mungall, C., Venter, J. C., Carucci, D. J., Hoffman, S. L., Newbold, C., Davis, R. W., Fraser, C. M. & Barrell, B. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**, 498-511 (2002). *(Appendices B and C)*
5. Gardner, M. J., Shallom, S., Carlton, J. M., Salzberg, S. L., Nene, V., Shoaibi, A., Ciecko, A., Lynn, J., Rizzo, M., Weaver, B., Jarrahi, B., Brenner, M., Parvizi, B., Tallon, L., Moazzez, A., Granger, D., Fujii, C. Y., Hansen, C., Pederson, J., Feldblyum, T., Peterson, J., Suh, B., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., White, O., Cummings, L. M., Smith, H. O., Adams, M. D., Venter, J. C., Carucci, D. J., Hoffman, S. L. & Fraser, C. M. Sequence of *Plasmodium falciparum* chromosomes 2, 10, 11, and 14. *Nature* **419**, 531-534 (2002). *(Appendices B and C)*

6. Carlton, J. M., Angiuoli, S. V., Suh, B. B., Kooij, T., Pertea, M., Ermolaeva, M. D., Allen, D. R., Silva, J. C., Selengut, J. D., Koo, H. L., Peterson, J. D., Pop, M., Kosack, D. S., Shumway, M. F., Bidwell, S. L., Shallom, S. J., Shoaibi, A., Cummings, L. M., Cho, J. K., Quackenbush, J., van Aken, S. E., Riedmuller, S. B., Feldbylum, T. V., Florens, L., Yates III, J. R., Raine, D. J., Sinden, R. E., Harris, M. A., Cunningham, D. A., Preiser, P. R., Bergman, L. W., Vaidya, A. B., van Lin, L. H., Janse, C. J., Waters, A. P., Smith, H. O., White, O. R., Salzberg, S. L., Venter, J. C., Fraser, C. M., Hoffman, S. L., Gardner, M. J. & Carucci, D. J. Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii. Nature* **419**, 512-519 (2002). *(Appendices B and C)*
7. Florens, L., Washburn, M. P., Raine, J. D., Anthony, R. M., Grainger, M., Haynes, J. D., Moch, J. K., Muster, N., Sacci, J. B., Tabb, D. L., Witney, A. A., Wolters, D., Wu, Y., Gardner, M. J., Holder, A. A., Sinden, R. E., Yates, J. R., 3rd & Carucci, D. J. A proteomic view of the *Plasmodium falciparum* life cycle. *Nature* **419**, 520-526 (2002). *(Appendices B and C)*
8. Carucci, D. J., Horrocks, P. & Gardner, M. J. in *Methods in Molecular Medicine, Malaria Methods and Protocols* (ed. Doolan, D. L.) (Humana Press, Totowa, NJ, 2002).
9. Kappe, S. H., Gardner, M. J., Brown, S. M., Ross, J., Matuschewski, K., Ribeiro, J. M., Adams, J. H., Quackenbush, J., Cho, J., Carucci, D. J., Hoffman, S. L. & Nussenzweig, V. Exploring the transcriptome of the malaria sporozoite stage. *Proc Natl Acad Sci U S A* **98**, 9895-900 (2001). *(Appendix F)*
10. Gardner, M. J. A status report on the sequencing and annotation of the P. falciparum genome. *Mol Biochem Parasitol* **118**, 133-8 (2001). (Appendix G)
11. Pertea, M., Salzberg, S. L. & Gardner, M. J. Finding genes in Plasmodium falciparum. *Nature* **404**, 34; discussion 34-5 (2000).
12. Salzberg, S. L., Pertea, M., Delcher, A., Gardner, M. J. & Tettelin, H. Interpolated Markov models for eukaryotic gene finding. *Genomics* **59**, 24-31 (1999). *(Appendix H)*
13. Lai, Z., Jing, J., Aston, C., Clarke, V., Apodaca, J., Dimlanta, E. T., Carucci, D. J., Gardner, M. J., Mishra, B., Anantharaman, T., Paxia, S., Hoffman, S. L., Venter, J. C., Huff, E. J. & Schwartz, D. C. A shotgun optical map of the entire *Plasmodium falciparum* genome. *Nat Genet* **23**, 309-313 (1999). (Appendix I)
14. Jing, J., Aston, C., Zhongwu, L., Carucci, D. J., Gardner, M. J., Venter, J. C. & Schwartz, D. C. Optical mapping of *Plasmodium falciparum* chromosome 2. *Genome Res* **9**, 175-181 (1999). *(Appendix J)*
15. Gardner , M. J., Tettelin, H., Carucci, D. J., Cummings, L. M., Smith, H. O., Fraser, C. M., Venter, J. C. & Hoffman, S. L. The malaria genome sequencing project: complete sequence of *P. falciparum* chromosome 2. *Parassitologia* **41**, 69-75 (1999). *(Appendix K)*
16. Gardner, M. J. The genome of the malaria parasite. *Curr Opin Genet Devel* **9**, 704-708 (1999). *(Appendix L)*

17. Gardner, M. J., Tettelin, H., Carucci, D. J., Cummings, L. M., Aravind, L., Koonin, E. V., Shallom, S., Mason, T., Yu, K., Fujii, C., Pedersen, J., Shen, K., Jing, J., Schwartz, D. C., Pertea, M., Salzberg, S., Zhou, L., Sutton, G. G., Clayton, R. L., White, O., Smith, H. O., Fraser, C. M., Adams, M. D., Venter, J. C. & Hoffman, S. L. Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science* **282**, 1126-1132 (1998). *(Appendix M)*
18. Gardner, M. J., Tettelin, H., Carruci, D. J., Cummings, L. M., Adams, M. D., Smith, H. O., Venter, J. C. & Hoffman, S. L. The Malaria Genome Sequencing Project. *Protist* **149**, 109-112 (1998). *(Appendix N)*
19. Carucci, D. J., Gardner, M. J., Tettelin, H., Cummings, L. M., Smith, H. O., Adams, M. D., Venter, J. C. & Hoffman, S. L. Sequencing the genome of *Plasmodium falciparum*. *Curr Opin Infect Dis* **11**, 531-534 (1998).
20. Carucci, D. J., Gardner, M. J., Tettelin, H., Cummings, L. M., Smith, H. O., Adams, M. D., Hoffman, S. L. & Venter, J. C. The malaria genome sequencing project. *Expert Reviews in Molecular Medicine* **http://www-ermm.cbcu.cam.ac.uk/dcn/txt001dcn.htm** (1998).

Manuscripts

1. João Carlos Aguiar, Jose M. Ribeiro, Peter L. Blair, Fengying Huang, Joshua A. Russell, Patricia de la Vega, Adam Witney and Daniel J. Carucci. "The Identification of *Plasmodium falciparum* genes expressed on sporozoite stages."

2. João Carlos Aguiar, Jose M. Ribeiro, Peter L. Blair, Fengying Huang, Joshua A. Russell, Patricia de la Vega, Adam Witney and Daniel J. Carucci. "Transcriptional Analysis of *Plasmodium yoelii* Liver Stage Gene Expression".

3. João Carlos Aguiar, Joshua LaBaer, Peter L. Blair, Victoria Y. Shamailova, Malvika Koundinya, Joshua A. Russell, Fengying Huang, Kristi Strang, Wenhong Mar, Robert M. Anthony, Adam Witney, Sonia R. Caruana, Leonardo Brizuela, John B. Sacci Jr., Stephen L. Hoffman and Daniel J. Carucci. "High throughput Generation of *P. falciparum* Functional Molecules by Recombinational Cloning".

4. Denise L. Doolan, Joao C. Aguiar, Walter R. Weiss, Alex Sette, Phil L. Felgner, David P. Regis, Paula Quinones-Casas, John R. Yates, III, Peter L. Blair, Tom L. Richie, Stephen L. Hoffman and Daniel J. Carucci. "Utilization of genomic sequence information to develop malaria vaccines".

Oral and poster presentations

1. Joao C. Aguiar, Peter L. Blair Fengying Huang, Joshua A. Russell, Patricia de la Vega, Adam A. Witney, Jose M. Ribiero, John B. Sacci, and

Daniel J. Carucci. "Discovering Malaria Pre-erythrocytic Genes." Molecular Parasitology Meeting. Woods Hole, MA. September, 2003. *(Oral presentation)*

2. Denise Doolan. "Utilization of genomic sequence information to develop malaria vaccines." American Societyof Tropical Medicine and Hygiene, 52nd Annual Meeting and Centennial Celebration. Philadelphia, PA. December 3-7, 2003 *(Oral presentation)*
3. Gardner, M. J. "Insights into parasite biology derived from the sequencing of parasite genomes, functional genomics, and proteomics." Functional genomics - genome communication. Chalmers University of Technology. Gothenburg, Sweden. August 28-29, 2003. (*Oral presentation*)
4. Gardner, M. J. "Genome sequence of the human malaria parasite *Plasmodium falciparum*." Utilisation of genome data for the development of new diagnostics, therapeutics, and vaccines. National Science Foundation. Colombo, Sri Lanka. Jan. 7, 2003. (*Oral presentation*)
5. Gardner, M. J. "Genome sequence of the human malaria parasite *Plasmodium falciparum*." 14th Annual Genome Sequencing and Analysis Conference. Boston, MA. Oct. 2-5, 2002. (*Organizer and chair of the "Host-Pathogen Genomics" session, oral presentation*)
6. Gardner, M. J. "Genome sequence of the human malaria parasite *Plasmodium falciparum*." Press Conference on the Publication of the *Anopheles gambiae* and *Plasmodium falciparum* Genomes. Headquarters, American Association for the Advancement of Science. Washington, D.C. Oct. 2, 2002. (*Oral presentation*)
7. "Genome sequences of *Plasmodium falciparum* and *Theileria parva*." International Centre of Insect Physiology and Ecology. Nairobi, Kenya. Nov. 26, 2002. *(Seminar)*
8. "Genome sequences of *Plasmodium falciparum* and *Theileria parva*." International Livestock Research Institute. Nairobi, Kenya. Nov. 27, 2002. *(Seminar)*
9. Gardner, M. J. "Genome sequence of the human malaria parasite *Plasmodium falciparum*." Burroughs Wellcome Fund Symposium: The Complete *Plasmodium falciparum* Genome, Insights and Surpises. American Society of Tropical Medicine and Hygiene 51st Annual Meeting. Denver, CO. Nov. 10-14, 2002. (*Oral presentation*)
10. Gardner, M. J. "Progress towards completion of the *Plasmodium falciparum* genome." Malaria: Progress, Problems and Plans in the Genomic Era. Johns Hopkins University School of Public Health. Baltimore, MD. Jan. 27-29, 2002. (*Oral presentation*)
11. Gardner, M. J. "Genome sequence of the human malaria parasite *Plasmodium falciparum*." Genomes Around Us: What Are We Learning? AAAS Annual Meeting. Boston, MA. Feb. 14-19, 2002. (*Oral presentation*)
12. Gardner, M. J. "Completion of the *Plasmodium falciparum* genome." 2nd ASM & TIGR Conference on Microbial Genomes. Las Vegas, NV. Feb. 10-13, 2002. (*Oral presentation*)
13. Gardner, M. J. "Genome sequence of the human malaria parasite

Plasmodium falciparum." The Third Pan-African Malaria Conference. Arusha International Conference Center. Arusha, Tanzania. Nov. 17-22, 2002. (*Organizer of "Malaria Genomics" side meeting, oral presentation*)

14. Gardner, M. J. "*Theileria* sequencing project at ILRI and TIGR." 11th Meeting of the Malaria Genome Sequencing Consortium. Wellcome Trust Genome Campus. Hinxton, U.K. June 5-6, 2001. (*Oral presentation*)
15. Gardner, M. J. "The Malaria Genome Sequencing Project." Symposium on "Human, Microbial and Vector Genome Projects: The Road to New Interventions for Tropical Diseases." American Society of Tropical Medicine and Hygiene 50th Annual Meeting. Atlanta, GA. Nov. 11-15, 2001. (*Oral presentation*)
16. Gardner, M. J. *Plasmodium falciparum* Genome Annotation Meeting. The Institute for Genomic Research. Rockville, MD. Dec. 7-8, 2001. (*Organizer and Chair, oral presentation*)
17. Gardner, M. J. "Update on sequencing of *P. falciparum* chromosomes 10, 11, and 14." 10th Meeting of the Malaria Genome Sequencing Consortium. Philadelphia, PA. Feb. 2-4, 2001. (*Oral presentation*)
18. Gardner, M. J. "Update on plans for *P. falciparum* publication." 11th Meeting of the Malaria Genome Sequencing Consortium. The Wellcome Trust Sanger Institute. Hinxton, England. June 5-6, 2001. (*Oral presentation*)
19. Gardner, M. J. "Sequencing the genome of *Plasmodium falciparum*." 36th Joint Conference on Parasitic Diseases. National Institutes of Health. Bethesda, MD. July 23, 2001. (*Oral presentation*)
20. "Complete sequence of *P. falciparum* chromosome 2." International Livestock Research Institute. Nairobi, Kenya. Feb. 12, 2000. *(Seminar)*
21. "The *Plasmodium falciparum* genome project." Harvard School of Public Health. Cambridge, MA. Apr. 18, 2000. *(Seminar)*
22. Gardner, M. J. "The malaria genome project." Harvard Malaria Institute Initiative Workshop: Genomes to Drugs. Harvard Faculty Club. Cambridge, MA. July 24-25, 2000. (*Oral presentation*)
23. Gardner, M. J. "Progress report on sequencing of *P. falciparum* chromosome 14." Ninth Malaria Genome Consortium Meeting. The Wellcome Trust Genome Campus. Hinxton, U.K. June 4-6, 2000. (*Oral presentation*)
24. Gardner, M. J. "Malaria Genome Sequencing Project." Third Annual Conference on Microbial Genomes. Chantilly, VA. Jan. 29 - Feb. 1, 1999. (*Oral presentation*)
25. Gardner, M. J. "Update on the sequencing of *P. falciparum chromosome 14*." Seventh Malaria Genome Sequencing Meeting. Wellcome Trust Genome Campus. Hinxton, U.K. July 21-23, 1999. (*Oral presentation*)
26. Gardner, M. J. & Cummings, L. M. "Sequencing of *P. falciparum* chromosomes 10, 11, and 14." Malaria Genome Symposium, American Society of Tropical Medicine and Hygiene 48th Annual Meeting. Washington, D.C. Nov. 28 - Dec. 2, 1999. (*Oral presentation*)
27. Gardner, M. J. "Malaria research after the genome project." British Society

of Parasitology 11th Malaria Meeting. Imperial College. London, U.K. Sept. 20-22, 1999. (*Oral presentation*)

28. Gardner, M. J. "Microbial genome sequencing and vaccine development." Society for Industrial Microbiology. Arlington, VA. August 2, 1999. (*Oral presentation*)
29. Gardner, M. J. "Sequencing of microbial genomes and the implications for vaccine development." 6th Annual IBC Conference of Vaccine Technologies. Arlington, VA. March 1999. (*Oral presentation*)
30. "The malaria genome project: sequencing of *P. falciparum* chromosome 2." Universidad de Puerto Rico, Recinto de Ciencias Medicas. San Juan, Puerto Rico. Apr. 12, 1999. *(Seminar)*
31. Shallom, S., Tettelin, H., Cummings, L. M., Gardner, M. J., Carucci, D. J., Adams, M. D., Hoffman, S. L. & Venter, J. C. in *10th International Genome Sequencing and Analysis Conference* (Miami Beach, FL, 1998).
32. Gardner, M. J. "Chromosome 2 sequence of *Plasmodium falciparum*." Workshop on the Functional Analysis of the Malaria Genome. The Institute for Genomic Research. Rockville, MD. Nov. 9-10, 1998. (*Oral presentation*)
33. Gardner, M. J. "Complete sequence of *Plasmodium falciparum* chromosome 2." The Malaria Challenge After One Hundred Years of Malariology. Accademia Nazionale dei Lincei. Rome, Italy. Nov. 16-18, 1998. (*Oral presentation*)
34. Gardner, M. J., Tettelin, H., Carucci, D. J., Cummings, L. M., Aravind, L., Koonin, E. V., Smith, H. O., Adams, M. D., Venter, J. C. & Hoffman, S. L. "Complete nucleotide sequence of chromosome 2 of the human malaria parasite *Plasmodium falciparum*." 10th International Genome Sequencing and Analysis Conference. Miami Beach, FL. Sept. 1998. (*Oral presentation*)
35. Gardner, M. J. "The malaria genome: chromosome 2." Gordon Conference on Malaria. Somerville College, Oxford University. Oxford, U.K. July 27-31, 1998. (*Oral presentation*)
36. "Application of genomics to parasitology: sequencing of chromosome 2 of *P. falciparum*." Johns Hopkins University. Baltimore, MD. *(Lecture)*
37. Gardner, M. J. "Sequencing of *P. falciparum* chromosome 2." Scientific Working Group on Utilization of Genomic Information for Tropical Diseases Drug and Vaccine Discovery. World Health Organization. Geneva, Switzerland. Feb. 18-20, 1998. (*Oral presentation*)
38. Gardner, M. J., Cummings, L., Tettelin, H., Carucci, D. C., Smith, H. O., Hoffman, S. L. & Venter, J. C. "Sequencing of chromosome 2 of the human malaria parasite *Plasmodium falciparum*." Second Annual Conference on Microbial Genomes. Hilton Head, SC. Jan. 31 - Feb 4., 1998. (*Oral presentation*)
39. Gardner, M. J., Carucci, D. J., Cummings, L., Tettelin, H., Adams, M., Smith, H. O., Hoffman, S. L. & Venter, J. C. "Progress report on sequencing of *Plasmodium falciparum* chromosome 2." Malaria Genome Sequencing Meeting. Holiday Inn. Cambridge, U.K. (*Oral presentation*)

40. Gardner, M. J. "The malaria genome project: strategies and current status." Malaria Genome Symposium, American Society of Tropical Medicine and Hygiene Annual Meeting. Orlando, FL. (*Oral presentation*)
41. Cummings, L. M., Tettelin, H., Carucci, D. J., Gardner, M. J., Shen, K., Pedersen, J., Shallom, S., Smith, H. O., Hoffman, S. L., Adams, M. D. & Venter, J. C. "Malaria genome project: sequencing of *Plasmodium falciparum* chromosome 2." 9th International Genome Sequencing and Analysis Conference. Hilton Head, SC. (*Oral presentation.*)

Web sites, CD-ROM, and artwork

1. *Plasmodium* genome: scientific achievement and medical opportunity (CD-ROM). Nature Publishing Group (2002).
2. *Plasmodium falciparum:* Malaria enters the genomic era (Poster distributed with Oct. 2, 2002 issue of *Nature*). Nature Publishing Group.
3. The *Plasmodium falciparum* Genome Project (web site). http://www.tigr.org/tdb/e2k1/pfa1/.
4. The *Plasmodium yoelii yoelii* Genome Sequencing Program (web site). http://www.tigr.org/tdb/e2k1/pya1/.
5. The *Plasmodium vivax* Genome Sequencing Program (web site). http://www.tigr.org/tdb/e2k1/pva1/.

Patent application

Provisional patent application. Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum* and proteins of said chromosome useful in antimalaria vaccines and diagnostic reagents. Filed by the Naval Medical Research Center. Docket number 82017.

Funding applied for

A "white paper" to request the funds required to complete the *P. vivax* genome sequence has been submitted to the NIAID under the Microbial Sequencing Centers program (http://www.niaid.nih.gov/dmid/genomes/mscs/default.htm).

Conclusions

The objectives of this 5-year Cooperative Agreement between TIGR and the Malaria Program, NMRC, were to: **Specific Aim 1**, sequence 3.5 Mb of *P. falciparum* genomic DNA; **Specific Aim 2**, annotate the sequence; **Specific Aim 3**, release the information to the scientific community. Two additional Specific Aims were added to the Cooperative Agreement: **Specific Aim 4**, sequencing of *P. yoelii* to 3X coverage; **Specific Aim 5**, sequencing of *P. vivax* to 3X coverage.

By publishing the complete genome sequence of *P. falciparum*, and chromosomes 2, 10, 11 and 14, we have **completed Specific Aims 1-3.** By sequencing *P. yoelii* to 5X coverage and publishing an analysis of this genome we have **completed Specific Aim 4**. **Specific Aim 5**, sequencing of *P. vivax* to 5X coverage has also been completed, but by using funds from other sources we were able to increase the sequence coverage to 10X and close X % of the intrascaffold gaps. We have requested funds from the NIAID Microbial Sequencing Centers program to complete the genome sequence.

This project, in which the overall goal was to sequence the genome of the human malaria parasite *P. falciparum*, has been extremely successful. In collaboration with colleagues at the Sanger Institute and the Stanford Genome Technology Center, the genome sequence was determined over a six-year period and published in a special issue of *Nature*. Preliminary data was released to the malaria research community throughout the project, prior to publication. These resources have been of tremendous value to malaria researchers worldwide and have facilitated hundreds of studies of parasite biochemistry, genetics, evolution, immunology, molecular and cellular biology, and pathogenesis [25 26 27 28 29]. In practical terms, the genome data has led directly to the identification of new vaccine candidate antigens [27] and many new drug targets [30 31]. The genome sequences also provided the foundation for further studies in functional genomics [23 32,33] and proteomics [11,34]. To date, the published

articles on the *P. falciparum* genome and chromosomes have been cited in published journal articles over 500 times (Institute for Scientific Information web site). Just one year after its publication, the *P. falciparum* genome paper [24] is one of the most highly cited papers in the malaria field.

Rapid improvements in sequencing technology and the concommitant reductions in costs over the life of this project allowed us to expand the original goals of this project to include the sequencing of the rodent malaria parasite *Plasmodium yoelii yoelii* [10], which is used as model system for studies of malaria vaccines. In addition, we have completed a draft genome sequence of the second major human malaria parasite *Plasmodium vivax.* The genome sequences of these organisms provide many opportunities for further studies of the biology, evolution, and pathogenesis of malaria parasites.

References

1. Butler, D., Maurice, J. & O'Brien, C. Briefing malaria. *Nature* **386**, 535-540 (1997).
2. Bloom, B. R. A microbial minimalist. *Nature* **378**, 236 (1995).
3. Hoffman, S. L., Bancroft, W. H., Gottlieb, M., James, S. L., Bond, E. C., Stephenson, J. R. & Morgan, M. J. Funding for malaria genome sequencing. *Nature* **387**, 647 (1997).
4. Gardner, M. J., Tettelin, H., Carucci, D. J., Cummings, L. M., Aravind, L., Koonin, E. V., Shallom, S., Mason, T., Yu, K., Fujii, C., Pedersen, J., Shen, K., Jing, J., Schwartz, D. C., Pertea, M., Salzberg, S., Zhou, L., Sutton, G. G., Clayton, R. L., White, O., Smith, H. O., Fraser, C. M., Adams, M. D., Venter, J. C. & Hoffman, S. L. Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science* **282**, 1126-1132 (1998).
5. Salzberg, S. L., Pertea, M., Delcher, A., Gardner, M. J. & Tettelin, H. Interpolated Markov models for eukaryotic gene finding. *Genomics* **59**, 24-31 (1999).
6. Jing, J., Aston, C., Zhongwu, L., Carucci, D. J., Gardner, M. J., Venter, J. C. & Schwartz, D. C. Optical mapping of *Plasmodium falciparum* chromosome 2. *Genome Research* **9**, 175-181 (1999).
7. Lai, Z., Jing, J., Aston, C., Clarke, V., Apodaca, J., Dimlanta, E. T., Carucci, D. J., Gardner, M. J., Mishra, B., Anantharaman, T., Paxia, S., Hoffman, S. L., Venter, J. C., Huff, E. J. & Schwartz, D. C. A shotgun optical map of the entire *Plasmodium falciparum* genome. *Nature Genetics* **23**, 309-313 (1999).
8. Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R. W., Carlton, J. M., Pain, A., Nelson, K. E., Bowman, S., Paulsen, I. T., James, K., Eisen, J. A., Rutherford, K., Salzberg, S. L., Craig, A., Kyes, S., Chan, M.-S., Nene, V., Shallom, S. J., Suh, B., Peterson, J., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., Haft, D., Mather, M. W., Vaidya, A. B., Martin,

D., Fairlamb, A. H., Fraunholz, M. J., Roos, D. S., Ralph, S., McFadden, G. I., Cummings, L. M., Subramanian, G. M., Mungall, C., Venter, J. C., Carucci, D. J., Hoffman, S. L., Newbold, C., Davis, R. W., Fraser, C. M. & Barrell, B. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**, 498-511 (2002).

9. Gardner, M. J., Shallom, S., Carlton, J. M., Salzberg, S. L., Nene, V., Shoaibi, A., Ciecko, A., Lynn, J., Rizzo, M., Weaver, B., Jarrahi, B., Brenner, M., Parvizi, B., Tallon, L., Moazzez, A., Granger, D., Fujii, C. Y., Hansen, C., Pederson, J., Feldblyum, T., Peterson, J., Suh, B., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., White, O., Cummings, L. M., Smith, H. O., Adams, M. D., Venter, J. C., Carucci, D. J., Hoffman, S. L. & Fraser, C. M. Sequence of *Plasmodium falciparum* chromosomes 2, 10, 11, and 14. *Nature* **419**, 531-534 (2002).
10. Carlton, J. M., Angiuoli, S. V., Suh, B. B., Kooij, T. W., Pertea, M., Silva, J. C., Ermolaeva, M. D., Allen, J. E., Selengut, J. D., Koo, H. L., Peterson, J. D., Pop, M., Kosack, D. S., Shumway, M. F., Bidwell, S. L., Shallom, S. J., van Aken, S. E., Riedmuller, S. B., Feldblyum, T. V., Cho, J. K., Quackenbush, J., Sedegah, M., Shoaibi, A., Cummings, L. M., Florens, L., Yates, J. R., Raine, J. D., Sinden, R. E., Harris, M. A., Cunningham, D. A., Preiser, P. R., Bergman, L. W., Vaidya, A. B., van Lin, L. H., Janse, C. J., Waters, A. P., Smith, H. O., White, O. R., Salzberg, S. L., Venter, J. C., Fraser, C. M., Hoffman, S. L., Gardner, M. J. & Carucci, D. J. Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. *Nature* **419**, 512-9 (2002).
11. Florens, L., Washburn, M. P., Raine, J. D., Anthony, R. M., Grainger, M., Haynes, J. D., Moch, J. K., Muster, N., Sacci, J. B., Tabb, D. L., Witney, A. A., Wolters, D., Wu, Y., Gardner, M. J., Holder, A. A., Sinden, R. E., Yates, J. R., 3rd & Carucci, D. J. A proteomic view of the *Plasmodium falciparum* life cycle. *Nature* **419**, 520-526 (2002).
12. Carlton, J. M., Angiuoli, S. V., Suh, B. B., Kooij, T., Pertea, M., Ermolaeva, M. D., Allen, D. R., Silva, J. C., Selengut, J. D., Koo, H. L., Peterson, J. D., Pop, M., Kosack, D. S., Shumway, M. F., Bidwell, S. L., Shallom, S. J., Shoaibi, A., Cummings, L. M., Cho, J. K., Quackenbush, J., van Aken, S. E., Riedmuller, S. B., Feldbylum, T. V., Florens, L., Yates III, J. R., Raine, D. J., Sinden, R. E., Harris, M. A., Cunningham, D. A., Preiser, P. R., Bergman, L. W., Vaidya, A. B., van Lin, L. H., Janse, C. J., Waters, A. P., Smith, H. O., White, O. R., Salzberg, S. L., Venter, J. C., Fraser, C. M., Hoffman, S. L., Gardner, M. J. & Carucci, D. J. Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii*. *Nature* **419**, 512-519 (2002).
13. Mendis, K., Sina, B. J., Marchesini, P. & Carter, R. The neglected burden of *Plasmodium vivax* malaria. *Am J Trop Med Hyg* **64**, 97-106 (2001).
14. Collins, W. E., Contacos, P. G., Krotoski, W. A. & Howard, W. A. Transmission of four Central American strains of *Plasmodium vivax* from monkey to man. *J Parasitol* **58**, 332-335 (1972).

15. Contacos, P. G., Collins, W. E., Jeffery, G. M., Krotoski, W. A. & Howard, W. A. Studies on the characterization of plasmodium vivax strains from Central America. *Am J Trop Med Hyg* **21**, 707-12 (1972).
16. Collins, W. E., Sullivan, J. S., Morris, C. L., Galland, G. G., Jue, D. L., Fang, S., Wohlhueter, R., Reed, R. C., Yang, C., Hunter, R. L. & Lal, A. A. Protective immunity induced in squirrel monkeys with a multiple antigen construct against the circumsporozoite protein of Plasmodium vivax. *Am J Trop Med Hyg* **56**, 200-10 (1997).
17. Yang, C., Collins, W. E., Xiao, L., Saekhou, A. M., Reed, R. C., Nelson, C. O., Hunter, R. L., Jue, D. L., Fang, S., Wohlhueter, R. M., Udhayakumar, V. & Lal, A. A. Induction of protective antibodies in Saimiri monkeys by immunization with a multiple antigen construct (MAC) containing the Plasmodium vivax circumsporozoite protein repeat region and a universal T helper epitope of tetanus toxin. *Vaccine* **15**, 377-86 (1997).
18. Carlton, J. M.-R., Galinski, M. R., Barnwell, J. W. & Dame, J. B. Karyotype and synteny among the chromosomes of all four species of human malaria parasite. *Molecular and Biochemical Parasitology* **101**, 23-32 (1999).
19. Carlton, J. M., Muller, R., Yowell, C. A., Fluegge, M. R., Sturrock, K. A., Pritt, J. R., Vargas-Serrato, E., Galinski, M. R., Barnwell, J. W., Mulder, N., Kanapin, A., Cawley, S. E., Hide, W. A. & Dame, J. B. Profiling the malaria genome: a gene survey of three species of malaria parasite with comparison to other apicomplexan species. *Mol Biochem Parasitol* **118**, 201-10 (2001).
20. Tchavtchitch, M., Fischer, K., Huestis, R. & Saul, A. The sequence of a 200 kb portion of a *Plasmodium vivax* chromosome reveals a high degree of conservation with *Plasmodium falciparum* chromosome 3. *Mol Biochem Parasitol* **118**, 211-22 (2001).
21. del Portillo, H. A., Fernandez-Becerra, C., Bowman, S., Oliver, K., Preuss, M., Sanchez, C. P., Schneider, N. K., Villalobos, J. M., Rajandream, M. A., Harris, D., Pereira da Silva, L. H., Barrell, B. & Lanzer, M. A superfamily of variant genes encoded in the subtelomeric region of Plasmodium vivax. *Nature* **410**, 839-42 (2001).
22. del Portillo, H. A., Fernandez-Becerra, C., Bowman, S., Oliver, K., Preuss, M., Sanchez, C. P., Schneider, N. K., Villalobos, J. M., Rajandream, M. A., Harris, D., da Silva, L. H., Barrell, B. & Lanzer, M. A superfamily of variant genes encoded in the subtelomeric region of *Plasmodium vivax*. *Nature* **410**, 839-42. (2001).
23. Le Roch, K. G., Zhou, Y., Blair, P. L., Grainger, M., Moch, J. K., Haynes, J. D., De La Vega, P., Holder, A. A., Batalov, S., Carucci, D. J. & Winzeler, E. A. Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science* **301**, 1503-8 (2003).
24. Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R. W., Carlton, J. M., Pain, A., Nelson, K. E., Bowman, S., Paulsen, I. T., James, K., Eisen, J. A., Rutherford, K., Salzberg, S. L., Craig, A., Kyes, S., Chan, M. S., Nene, V., Shallom, S. J., Suh, B., Peterson, J., Angiuoli, S., Pertea,

M., Allen, J., Selengut, J., Haft, D., Mather, M. W., Vaidya, A. B., Martin, D. M., Fairlamb, A. H., Fraunholz, M. J., Roos, D. S., Ralph, S. A., McFadden, G. I., Cummings, L. M., Subramanian, G. M., Mungall, C., Venter, J. C., Carucci, D. J., Hoffman, S. L., Newbold, C., Davis, R. W., Fraser, C. M. & Barrell, B. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**, 498-511 (2002).

25. Horrocks, P., Bowman, S., Kyes, S., Waters, A. P. & Craig, A. Entering the post-genomic era of malaria research. *Bull World Health Organ* **78**, 1424-37 (2000).
26. Craig, A., Kyes, S., Ranson, H. & Hemingway, J. Malaria parasite and vector genomes: partners in crime. *Trends Parasitol* **19**, 356-62 (2003).
27. Hoffman, S. L., Subramanian, G. M., Collins, F. H. & Venter, J. C. *Plasmodium*, human and *Anopheles* genomics and malaria. *Nature* **415**, 702-9 (2002).
28. Greenwood, B. & Mutabingwa, T. Malaria in 2002. *Nature* **415**, 670-2 (2002).
29. Wellems, T. E., Su, X., Ferdig, M. & Fidock, D. A. Genome projects, genetic analysis, and the changing landscape of malaria research. *Current Opinions in Microbiology* **2**, 415-9 (1999).
30. Macreadie, I., Ginsburg, H., Sirawaraporn, W. & Tilley, L. Antimalarial drug development and new targets. *Parasitol Today* **16**, 438-44 (2000).
31. Wilson, R. J. Progress with parasite plastids. *J Mol Biol* **319**, 257-74 (2002).
32. Bozdech, Z., Zhu, J., Joachimiak, M. P., Cohen, F. E., Pulliam, B. & DeRisi, J. L. Expression profiling of the schizont and trophozoite stages of *Plasmodium falciparum* with a long-oligonucleotide microarray. *Genome Biol* **4**, R9 (2003).
33. Bozdech, Z., Llinas, M., Pulliam, B. L., Wong, E. D., Zhu, J. & DeRisi, J. L. The Transcriptome of the Intraerythrocytic Developmental Cycle of Plasmodium falciparum. *PLoS Biol* **1**, 85-100 (2003).
34. Lasonder, E., Ishihama, Y., Andersen, J. S., Vermunt, A. M., Pain, A., Sauerwein, R. W., Eling, W. M., Hall, N., Waters, A. P., Stunnenberg, H. G. & Mann, M. Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry. *Nature* **419**, 537-42 (2002).

Appendices

Appendix A. List of personnel receiving pay from the research effort.

Appendix B. October 3, 2002 issue of special issue of *Nature* on "*Plasmodium* genomics."

Appendix C. Special Reprint of October 3, 2002 issue of *Nature* containing the following reprints:

Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R. W., Carlton, J. M., Pain, A., Nelson, K. E., Bowman, S., Paulsen, I. T., James, K., Eisen, J. A., Rutherford, K., Salzberg, S. L., Craig, A., Kyes, S., Chan, M.-S., Nene, V., Shallom, S. J., Suh, B., Peterson, J., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., Haft, D., Mather, M. W., Vaidya, A. B., Martin, D., Fairlamb, A. H., Fraunholz, M. J., Roos, D. S., Ralph, S., McFadden, G. I., Cummings, L. M., Subramanian, G. M., Mungall, C., Venter, J. C., Carucci, D. J., Hoffman, S. L., Newbold, C., Davis, R. W., Fraser, C. M. & Barrell, B. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**, 498-511 (2002).

Gardner, M. J., Shallom, S., Carlton, J. M., Salzberg, S. L., Nene, V., Shoaibi, A., Ciecko, A., Lynn, J., Rizzo, M., Weaver, B., Jarrahi, B., Brenner, M., Parvizi, B., Tallon, L., Moazzez, A., Granger, D., Fujii, C. Y., Hansen, C., Pederson, J., Feldblyum, T., Peterson, J., Suh, B., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., White, O., Cummings, L. M., Smith, H. O., Adams, M. D., Venter, J. C., Carucci, D. J., Hoffman, S. L. & Fraser, C. M. Sequence of *Plasmodium falciparum* chromosomes 2, 10, 11, and 14. *Nature* **419**, 531-534 (2002).

Carlton, J. M., Angiuoli, S. V., Suh, B. B., Kooij, T., Pertea, M., Ermolaeva, M. D., Allen, D. R., Silva, J. C., Selengut, J. D., Koo, H. L., Peterson, J. D., Pop, M., Kosack, D. S., Shumway, M. F., Bidwell, S. L., Shallom, S. J., Shoaibi, A., Cummings, L. M., Cho, J. K., Quackenbush, J., van Aken, S. E., Riedmuller, S. B., Feldbylum, T. V., Florens, L., Yates III, J. R., Raine, D. J., Sinden, R. E., Harris, M. A., Cunningham, D. A., Preiser, P. R., Bergman, L. W., Vaidya, A. B., van Lin, L. H., Janse, C. J., Waters, A. P., Smith, H. O., White, O. R., Salzberg, S. L., Venter, J. C., Fraser, C. M., Hoffman, S. L., Gardner, M. J. & Carucci, D. J. Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii*. *Nature* **419**, 512-519 (2002).

Florens, L., Washburn, M. P., Raine, J. D., Anthony, R. M., Grainger, M., Haynes, J. D., Moch, J. K., Muster, N., Sacci, J. B., Tabb, D. L., Witney, A. A., Wolters, D., Wu, Y., Gardner, M. J., Holder, A. A., Sinden, R. E., Yates, J. R., 3rd & Carucci, D. J. A proteomic view of the *Plasmodium falciparum* life cycle. *Nature* **419**, 520-526 (2002).

Appendix D. *Plasmodium* genome: scientific achievement and medical opportunity (CD-ROM). Nature Publishing Group (2002).

Appendix E. Kappe, S. H., Gardner, M. J., Brown, S. M., Ross, J., Matuschewski, K., Ribeiro, J. M., Adams, J. H., Quackenbush, J., Cho, J., Carucci, D. J., Hoffman, S. L. & Nussenzweig, V. Exploring the transcriptome of the malaria sporozoite stage. *Proc Natl Acad Sci U S A* **98**, 9895-900 (2001).

Appendix F. Gardner, M. J. A status report on the sequencing and annotation of the P. falciparum genome. Mol Biochem Parasitol 118, 133-8 (2001).

Appendix G. Salzberg, S. L., Pertea, M., Delcher, A., Gardner, M. J. & Tettelin, H. Interpolated Markov models for eukaryotic gene finding. *Genomics* **59**, 24-31 (1999).

Appendix H. Lai, Z., Jing, J., Aston, C., Clarke, V., Apodaca, J., Dimlanta, E. T., Carucci, D. J., Gardner, M. J., Mishra, B., Anantharaman, T., Paxia, S., Hoffman, S. L., Venter, J. C., Huff, E. J. & Schwartz, D. C. A shotgun optical map of the entire *Plasmodium falciparum* genome. *Nat Genet* **23**, 309-313 (1999).

Appendix I. Jing, J., Aston, C., Zhongwu, L., Carucci, D. J., Gardner, M. J., Venter, J. C. & Schwartz, D. C. Optical mapping of *Plasmodium falciparum* chromosome 2. *Genome Res* **9**, 175-181 (1999).

Appendix J. Gardner , M. J., Tettelin, H., Carucci, D. J., Cummings, L. M., Smith, H. O., Fraser, C. M., Venter, J. C. & Hoffman, S. L. The malaria genome sequencing project: complete sequence of *P. falciparum* chromosome 2. *Parassitologia* **41**, 69-75 (1999).

Appendix K. Gardner, M. J. The genome of the malaria parasite. *Curr Opin Genet Devel* **9**, 704-708 (1999).

Appendix L. Gardner, M. J., Tettelin, H., Carucci, D. J., Cummings, L. M., Aravind, L., Koonin, E. V., Shallom, S., Mason, T., Yu, K., Fujii, C., Pedersen, J., Shen, K., Jing, J., Schwartz, D. C., Pertea, M., Salzberg, S., Zhou, L., Sutton, G. G., Clayton, R. L., White, O., Smith, H. O., Fraser, C. M., Adams, M. D., Venter, J. C. & Hoffman, S. L. Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science* **282**, 1126-1132 (1998).

Appendix M. Gardner, M. J., Tettelin, H., Carruci, D. J., Cummings, L. M., Adams, M. D., Smith, H. O., Venter, J. C. & Hoffman, S. L. The Malaria Genome Sequencing Project. *Protist* **149**, 109-112 (1998).

Appendix A: Personal Receiving Salary fromCooperative Agreement

ABRAMZON, SOFIYA
AGUIAR, MARIA
AHN, SUSIE
ANGIUOLI, SAMUEL
ANTHONY, ROBERT
ATHEARN, RYAN
BAILEY, MARSHA
BENTON, JONATHAN
BERA, JAYATI
BERRY, KRISTI
BIDWELL, SHELBY
BORKOWSKI, NICOLE T.
BORMAN, JON
BOWMAN, CHERYL
BRENNER, MICHAEL I.
BUCHOFF, JEFFREY
BURGESS, SHANDRECA
BURR, PATRICK
CALDWELL, LAUREN
CARLTON, JANE
CARNEIRO DA SILVA, JOANA
CARTY, HEATHER
CHAUDHARY, ABHILASHA
CHEN, DAN
CHEN, MINGHUA
CHEN, YIXIN
CIECKO, ANNE
COVARRUBIAS, MIGUEL
CRONIN, LISA
CUMMINGS, LEDA
CURTISS, RAHIM
DELCHER, ARTHUR
DINTERMAN, SHELLY
DOMINGO, ALEXANDER
DRAGOI, IOANA
DUGUE, NATALIE REDDIX
ELLIOTT, DAVID
ERMOLAEVA, MARIA
FARRELL, ANGELA
FUJII, CLAIRE
GANGESTAD, DONNA
GANSBERGER, KRISTEN
GARCIA_HERNANDEZ, ENRIQUE
GARDNER, MALCOLM
GARRETT, MINA
GEBREGEORGIS, ELIZABETH
GEER, KEITA
GERMAN, OKSANA
GILL, JOHN
GLADNEY, EMILY
GRIMES, ELIZABETH
HALPIN, REBECCA
HANCE, MARK E.
HANSEN, CHERYL
HEREFORD, NATALIE
HILL, JESSICA
HOLLEY, TARA
HOLMES, MICHAEL
IMPRAIM, MARJORIE
JACKSON, JACQUELINE
JARRAHI, BEHNAM
JENKINS, CHELTON
JENKINS, JENNIFER
JIANG, LINGXIA
JONES, KRISTINE
KALB, ERICA
KANG, KATHERINE
KHOURI, HODA
KOO, HEAN
KOSACK, DANIEL
KRIGA, YULIYA
KURUSHKO, ALENA
KWON, ERIKA
LARKIN, CHRISTOPHER
LEE, KATHERINE
LEE, PERVIS C.
LEE, YUANDAN
LEVITSKAIA, IRINA
LEWIS, MATTHEW
LU, CHARLES
LYNN, JEFFREY
MAHURKAR, ANUP
MAJOROS, WILLIAM

MASON, TANYA
MCDANIEL, JOE
MILITSCHER, JENNIFER
MOAZZEZ, AZITA
MOFFAT, KELLY
NELSON, KEITH
NENE, VISHVANATH
NORCUTT, KARA
OIGUENBLICK, EMILIA
PAI, GRACE H.
PARKSEY, DEBBIE
PARVIZI, BABAK
PERTEA, MIHAELA
PETROGRADSKAYA, INNA
RADUNE (BUSHMAN), DIANA
REDDIX-DUGUE,NATALIE
RIEDMULLER, STEVEN
RIGGS, FLORENCE
RIZZO, MICHAEL
ROMERO, CALUDIA
ROONEY, TIMOTHY
RUCH, KAREN
RYABTSEVA, TAMARA
SAVINOVA, LYUDMILA
SCANLAN, DAVID
SCANLAN, DAVID
SCHATZ, MICHAEL
SELLERS, PATRICK
SENTER, JACQUELINE
SHALLOM, SHAMIRA
SHETTY, JYOTI
SHUMWAY, MARTIN
SHVARTSBEYN, ALLA
SILVA, JOANA
SITZ, JEFF
SKOVORODNEV, ALEXANDER
SMIRNOVA, TATYANA
SMITH, SHANNON
SOMMER, DANIEL
SOSA, JULIA
STEWART, AMY
SUH, BERNARD
TALLON, LUKE J.
TOMS, BRADLEY
TRAN, BAO
TRAN, KEVIN
TSITRIN, TAMARA
UTTERBACK, TERESA
VAN AKEN, SUSAN
VISWANATHAN, LAKSHMI DEVI
VON ARX, ANNA
WANLESS, DAVID
WEAVER, BRUCE
WILLIAMS, MARY
WILLIAMS, MONICA
ZADORY, DANIEL
ZHAO, YONGMEI
ZHAO, YONGMEI
ZHOU, LIWEI
ZHURKIN, MIKHAIL
ZSCHOCHE, CHRISTINA

3 October 2002

International weekly journal of science

nature

$10.00

www.nature.com/nature

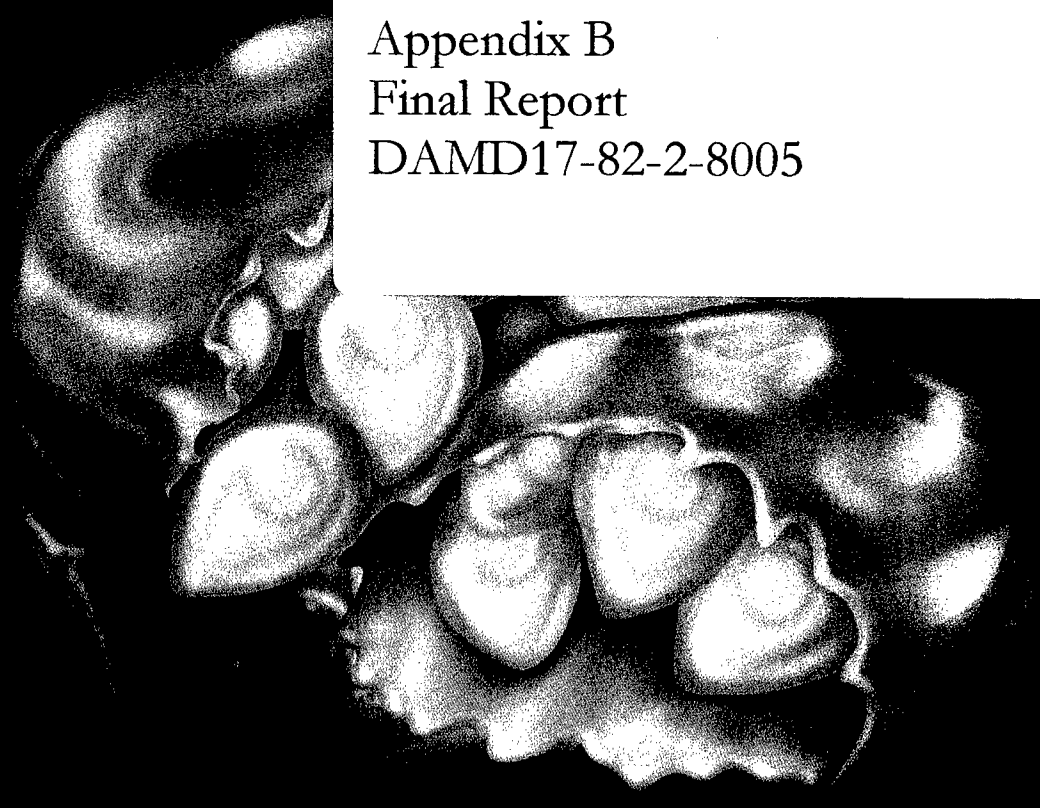
Plasmodium genomics

Genomics and proteomics pave the way for controlling malaria

Cold antihydrogen
CERN delivers

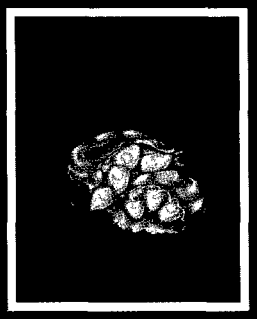
Antarctic ice
Flow reversals

Antigen presentation
A 'customizing' protease



nature

Plasmodium genomes



Cover illustration
Infected red blood cells liberating *P. falciparum* merozoite parasites. (Image credit: Debbie Marsels.)

Good 'omics' for the poor?

This week marks a milestone in malaria research, with the publication of complete genome sequences for the human parasite, the apicomplexan *Plasmodium falciparum* (this issue of *Nature*), and its vector, the mosquito *Anopheles gambiae* (*Science*, 4 October). Some may warn against excessive optimism, because the global problems caused by malaria are daunting (see Malaria Insight, *Nature* **415**, 669–715, 2002; News and Views, this issue, pp. 493–497; News Features, pp. 426–430). As an antidote to pessimism, let us celebrate the science you will find in this section, which will without doubt aid researchers in the fight against malaria.

Plasmodium falciparum is the first eukaryotic parasite for which we have a complete genome (pp. 498, 527, 531 & 534). The large proportion (70%) of predicted genes already validated experimentally allow firm conclusions to be drawn about the evolution of metabolic pathways. Comparison with the human genome also reveals some pathways that are specific to the pathogen or its peculiar organelles; these will usher in the development of specific drugs with lesser side effects. Completion of a second full genome, that of the model rodent malaria parasite *P. yoelii yoelii*, allows, for the first time, the comparison of two eukaryotic species within a single genus: *Plasmodium* (page 512). Despite their evolutionary similarity, these pathogens exhibit striking differences in their immune evasion strategies. The immediate availability of two state-of-the-art proteomics studies provides stimulating new insights for the development of both drugs and vaccines (pp. 520, 537). Newly discovered patterns of gene expression during the *Plasmodium* life cycle will lead to strategies for targeting several parasitic stages at once.

The fruits of applying this knowledge may take years to materialize, so could this be just another end of a beginning? We believe not. This major achievement will maintain the momentum in the scientific community worldwide. Researchers have already used the freely available PlasmoDB database (p. 490) to identify new potential antimalarial drugs (*Nature Medicine* **7**, 167; 2001). In the same spirit, all of this section's contents, along with seminal malaria research, news and features articles previously published in *Nature*, are available free online (**www.nature.com/nature/malaria**). A CD-ROM containing similar items, plus an interactive GenePlot from PlasmoDB, will be distributed with a future issue of *Nature*, bringing this wealth of information to researchers in countries with limited Internet access. That high-tech genomics and proteomics are being mobilized against the emblem disease of poverty is a good omen indeed.

Tanguy Chouard **Senior Editor**
Ursula Wales **Senior Editor**
Rita Shaw **Chief Biology Editor**

The *Plasmodium* genome database

Designing and mining a eukaryotic genomics resource.

The Plasmodium Genome Database Collaborative

As reported elsewhere in this issue (M. J. Gardner *et al.* *Nature* **419**, 498–511; 2002), a reference genome sequence for the human malaria parasite *Plasmodium falciparum* is now complete. But how are researchers to access *P. falciparum* genome sequence data, integrate this resource with other relevant data sets, and exploit the resulting information for functional studies, including identification of novel drug targets and candidate vaccine antigens?

The *Plasmodium* genome database (PlasmoDB, see http://PlasmoDB.org) contains information from multiple sources, including DNA sequence data and curated annotations, automated gene model predictions, predicted proteins and protein motifs, cross-species comparisons, optical and genetic mapping data, information on population polymorphisms, expression data generated by a variety of complementary strategies, and proteomics data. Integrating this information at a single site provides 'one-stop shopping' for genomics-scale data sets related to malaria parasites.

The use of a relational database architecture enables users to ask complex questions. For example, immunologists trying to develop an antimalaria vaccine might wish to identify potential immunodominant surface antigens. Drug developers might wish to identify enzymes expressed in bloodstream parasites that differ significantly from their human counterparts. Researchers interested in antigenic variation and how the parasite adheres to cells (a cause of malaria pathogenesis) might wish to identify all gene families in the parasite genome; those interested in genome organization might be interested in the chromosomal location of these proteins; evolutionary biologists might wish to examine all genes for which clear orthologues are known from a range of species; and so on.

Universal access

It has taken six years to complete the *P. falciparum* genome sequence. In the meantime, interim data were periodically released by the three sequencing centres involved in this project, to advance research on basic malaria biology, and drug and vaccine development. PlasmoDB was developed to make this information available to the research community, notwithstanding the challenges posed by unfinished sequence data. This web-accessible database provides access to the entire genome sequence of the 3D7 reference strain of *P. falciparum*, together with computationally predicted and manually curated genes and gene models, protein feature predictions and functional annotation.

PlasmoDB went live in June 2000 — more than two years before today's formal completion of the *P. falciparum* reference sequence. The website receives several thousand hits each day from more than 100 countries, numbers that are certain to rise significantly with the release of the complete genome sequence. The result can be measured in the scores, possibly hundreds, of publications that have resulted, and in new targets now being assessed for drug and vaccine development.

Malaria biologists are a more diverse and dispersed community than those who study fruitfly or yeast genomes. They encompass field scientists in Cameroon, epidemiologists in Papua New Guinea, pharmaceutical developers in India, molecular geneticists in Brazil, and so on. Because many malaria researchers lack reliable high-speed Internet access, a platform-independent CD-ROM (to be distributed with *Nature* in a few weeks' time) has been developed to provide universal free access to the complete genome sequence and annotations currently available for this malaria parasite. More than a series of 'flat-file' images, *P. falciparum* GenePlot is a true database, providing a graphical user interface for browsing, querying, downloading and manipulating the genome and annotations on a desktop computer without web access.

It has been a stimulating challenge to see how many commonly asked questions can be accommodated in the CD-ROM format. For example, while local implementation of BLAST searches requires substantial memory and computational speed (and GenBank is too large to include on a single CD), GenePlot can be asked to find and retrieve all predicted proteins with similarity to proteases, based on text indices derived from precomputed BLAST comparisons of the entire *P. falciparum* genome against all of GenBank.

The initial motivation behind the GenePlot CD was to make the genome accessible to malaria biologists with limited Internet connectivity, but this format has also proved enormously popular with well-connected users. Having the data literally 'in hand' provides scientists everywhere with a sense of ownership and involvement in the *Plasmodium* genome project, expediting the pace of research and discovery related to malaria parasites and the devastating diseases they cause.

Unfinished business

In most genomics projects, initial mapping studies (desirable even with the advent of whole-genome shotgun sequencing) are followed by a random sequencing phase, then by a phase focusing on closure of remaining gaps to produce a 'finished' sequence (which may still contain numerous gaps, depending on complexity and size of the genome, time, patience and funding). Annotation is conducted to various levels of depth. Database development makes the information accessible to the user community. Finally, functional studies (transcript profiling, proteomic studies, genome-scale knockouts, and so on) become possible once the complete, annotated sequence is available to end-users.

There are good reasons for this sequential strategy. Gap closure is expensive, and so makes little sense while random sequencing may still yield useful information. Manual annotation of assembled sequences is also laborious, and is best deferred until the genome sequence is complete. For large, complex eukaryotic genomes, years may pass between the initial sequencing and the availability of this information in practical form for researchers in the lab. Such delays cause considerable frustration, as individual genes could be identified long before assembly of a finished genome.

Problems associated with unfinished data, and the accompanying need for user education regarding the interpretation of these results, provided the first challenge for PlasmoDB. Specific information missing in incomplete data

The CD-ROM containing *P. falciparum* GenePlot and other malaria-related resources, including *Nature*'s malaria Insight of 7 February 2002 and the papers reported elsewhere in this issue, will be provided to all *Nature* subscribers in a few weeks' time. It can also be obtained from helpcd@plasmodb.org or the Malaria Research and Reference Reagent Resource Center (MR4) by an e-mail request to malaria@atcc.org, with "Nature malaria CD-ROM" in the subject line. A full postal address must be included in the body of the message.

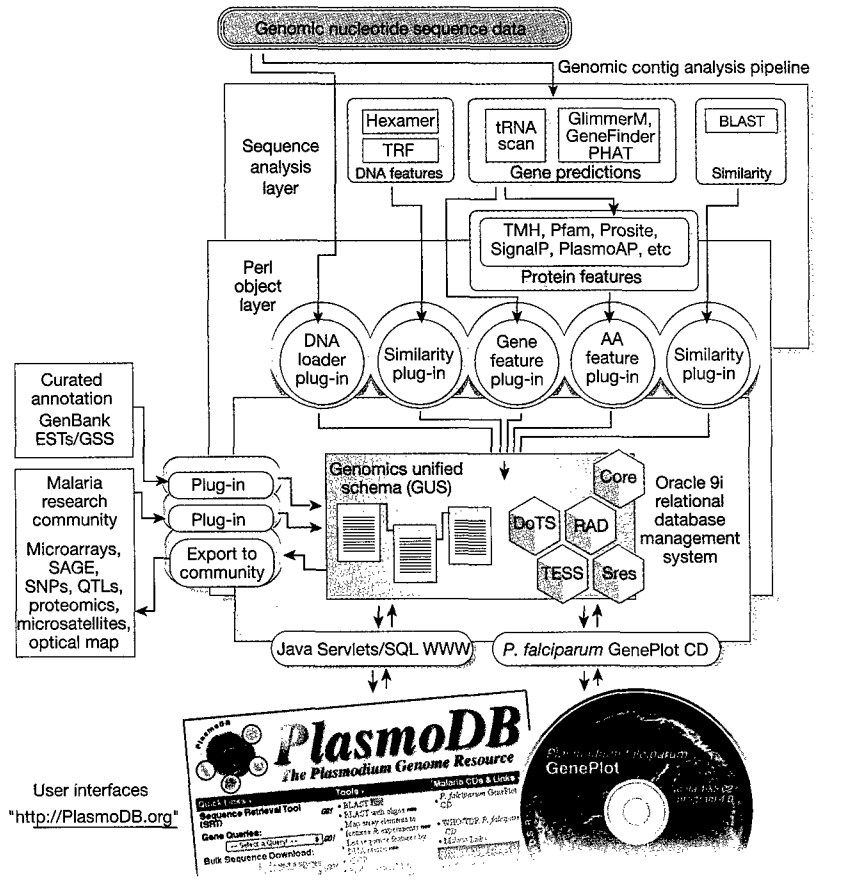
sets limits confidence that a particular gene is absent from the organism. Contaminating sequences from cloning vectors and host cells may be present. Redundancy in the data set attributable to incomplete or inaccurate assemblies poses a further problem, particularly for the A/T-rich *P. falciparum* genome. In PlasmoDB, possible redundancy or inaccurate assembly was identified by high-stringency comparisons of each sequence with the entire genome; and comparison of DNA sequences with optical and genetic maps. The importance of these tools for *P. falciparum* declines as the genome project approaches completion, but they remain valuable for new projects, such as the other *Plasmodium* species now being sequenced.

Unfinished sequence data also pose challenges for gene identification and analysis, as the constantly changing nature of this information makes time-consuming manual annotation impossible. Comparisons with GenBank, computational gene-finding algorithms and protein feature analyses are feasible (Box 1), but generate a bewildering range of predictions: which of four competing gene predictions is most likely to be correct? Which of 60 sequences exhibiting similarity to cathepsin is really a protease? Automated analysis can help to provide provisional assignments early, before manual curation of the finished sequence. Even after first-pass annotation, these analyses can help to suggest alternative possibilities whenever new experimental information suggests inaccuracies in the curated annotation.

Many disciplines

Many disciplines accommodate large data sets (MRI imaging, weather forecasting, ecological and econometric modelling, and so on), but this is a relatively new problem for molecular and cell biologists. How to collect the deluge of data engulfing us from genomics, transcriptomics, proteomics, glycomics, pharmacogenomics, vaccinomics, and even more hideously named approaches? What kind of tools will be required to analyse — and to integrate — these massive 'omics'-scale data sets? How can we use all this information to treat malaria?

PlasmoDB is based on a relational database architecture (GUS; Box 1), built around biologically relevant relationships following the central dogma of biology: 'gene to messenger RNA to protein'. Parallel views for other organisms (including other *Plasmodium* species) allow phylogenetic comparisons. Because all this information is in a single database, queries can combine searches for particular genes of interest with RNA and protein expression analysis, studies on population genetic polymorphisms, and cross-species comparison. One can envisage the incorporation of other data types, such as publication records, clinical outcome data, genomic information from the mosquito vector *Anopheles gambiae*, protein structural information



The architecture of PlasmoDB

PlasmoDB is not itself a database, but a web interface that uses an underlying relational database (GUS, for genomics unified schema), which stores and integrates nucleotide sequences, annotation, information on gene expression and regulation, controlled vocabularies/ontologies, and evidence for these annotations. GUS is organism-independent and also contains the human and mouse genomes (www.allgenes.org). The schema, associated code and project-independent data are at www.gusdb.org.

Primary *P. falciparum* sequence data are subjected to automated analyses (sequence analysis layer), including the identification of motifs and simple repeats; comparison against the entire genome to identify gene families, repetitive elements and redundancy; searching for intron/exon structure, using several algorithms trained on experimentally validated *P. falciparum* sequences; conceptual gene translation and identification of potential protein motifs; and comparisons with the non-redundant GenBank/EMBL database (results retained in a text-queryable index). Genomic contig sequences are aligned to optical restriction maps and microsatellite linkage groups using hidden Markov models for fragment length and ePCR.

The GUS schema employs views that are used in an object layer for parent–child relationships. To facilitate data loading, Perl was used to create a 'thin' object layer in which each relational table is treated as an object. GUS is partitioned into distinct name spaces. Core contains workflow tables, tracking how each row in the database is populated (data provenance). Sres (shared resources) contains controlled vocabularies and ontologies, such as taxonomy, anatomy and disease tables. TESS captures descriptions (grammar representations) for genetic regulatory regions (not currently implemented for PlasmoDB). DoTS houses sequence and sequence annotation. Any sequence span can have multiple features mapped to it, and gene predictions can be associated with multiple transcripts and proteins. Each predicted or experimentally determined transcript may itself have multiple features and similarities, as can each protein entry. RAD handles data from high-throughput technologies for studying gene expression. RAD currently accommodates expression data from SAGE (serial analysis of gene expression), cDNA and oligonucleotide glass slide microarrays, and Affymetrix chips, and is extensible to accommodate information from other platforms. Sample information, together with other experimental descriptions, can be entered directly into the database via web-based forms.

The RAD schema is compliant with MIAME guidelines (www.mged.org/). A microarray gene expression (MAGE) object model and XML-based language have been developed for data exchange, and importers and exporters are being built for RAD to MAGE-ML.

from high-throughput crystallography studies, and chemical compound libraries.

PlasmoDB provides graphic and text-based views of all available *Plasmodium* genomic sequences, curated annotations, and tools for retrieval of these data. But the sheer wealth of information can make browsing difficult, so the database allows the user to define custom views. For all their visual appeal, however, static, precomputed views are inherently restricted, and so fail to answer many genomic-scale questions that arise in the laboratory.

The relational database underlying PlasmoDB permits queries that integrate diverse data types, as illustrated by questions relating to drug and vaccine development (Table 1). For example, a medicinal chemist might be interested in *P. falciparum* dihydrofolate reductase (DHFR), the target of the drug pyrimethamine used in common antimalarial agents. The gene encoding this enzyme can be identified by EC number or GO function, text searches of curated annotation using the enzyme name as a key word, text searches against BLAST results, motif searches for protein sequence signatures, BLAST similarity to DHFR sequences from other species, or searches based on protein structural predictions. Degenerate searches are also possible, such as searching for all proteases. The results returned would undoubtedly contain false positives, but these can be weeded out by scientists familiar with protease characteristics. Candidate cytoskeletal proteins can be identified by similar strategies, or searches based on protein structural predictions. Such searches can then be refined, for example by identifying sequences conserved in multiple malaria parasites, or those that are sufficiently distinct from human orthologues to provide a basis for selective inhibition.

Information on metabolic pathways and/or subcellular localization can also be used to inform database queries. For example, PlasmoDB enables the identification of proteins likely to be associated with the apicoplast — a distinctive organelle that has received considerable attention as a candidate drug target — on the basis of curated annotation, exploiting the structured gene ontology (GO) vocabulary. Alternatively, the origins of this organelle by horizontal transfer of an algal chloroplast can be exploited as the basis for a text search for genes exhibiting sequence similarity to plastid, chloroplast or plant genes. Phylogenetic comparison with plant species is not currently supported in PlasmoDB, but all nucleotide and predicted protein sequences can be downloaded by users for local analysis.

Combining gene and protein predictions with the results from RNA and/or protein expression analysis enables enzymes being considered for antimalarial drug development to be filtered, removing any proteins not expressed in blood-stage parasites. Integrating these data with functional studies, polymorphism data, publications, or small-molecule databases, would allow further refinement.

User query	*Computational strategy/approach*
Drug development	
Dihydrofolate reductase	GO function*, EC*, text*, motif* and BLAST* searches
Proteases	Text* and motif* searches; GO function or process*
Cytoskeletal genes conserved in multiple *Plasmodium* species	GO cellular component*; protein structural predictions; phylogenetic cross-comparison with other *Plasmodium* species*
Differ significantly from probable human orthologues	Comparison with human sequences*
Apicoplast pathway enzymes	GO function, process or component*; text search for 'chloroplast or plastid'*; phylogenetic comparison with plants/algae
Expressed in blood-stage parasites	Expression profiling studies*; proteomics data*
Essential for parasite survival	Curated annotation from pharmacological/genetic studies; literature searches
Validated drug targets (in other systems)	Drug databases; literature databases
Availability of candidate inhibitors	Small-molecule databases; DOCKing algorithms
Vaccine candidates	
Known antigens (AMA1, MSP1, SERA)	Text*, motif* and BLAST* searches
Multigene families	Self-BLAST analysis*
Associated with the parasite on infected cell surface	Protein features: signal sequences*, transmembrane domains*, potential acylation sites or GPI anchors. Similarity to known membrane and surface proteins in other systems
Immunodominant	B- and/or T-cell epitope predictions*
Unlikely to be deleted	Non-telomeric*; conserved in multiple *P. falciparum* isolates
Likely to be under positive (immune) selection	DNA/protein features: repetitive/low-complexitysequence*. High ratio of non-silent/silent polymorphisms (from phylogenetic cross-comparisons and population genetics studies)*

*Searches currently supported by PlasmoDB

For immunologists, computationally accessible queries allow identification of particular genes of interest as vaccine antigens (see Table 1). Additional gene-family members can be recognized on the basis of sequence similarity. Probable surface antigens can be identified from the presence of signal sequences, transmembrane domains, acylation signals or glycophosphatidylinositol (GPI) anchor motifs. Additional queries of immunological relevance might include the presence of predicted immunodominant epitopes, expression in life-cycle stage(s) of interest, conservation in multiple *P. falciparum* isolates, and evidence of immune selection based on highly repetitive elements, low-complexity sequence or polymorphisms identified in population genetic studies.

PlasmoDB can be used to build complex queries using boolean operators. For example, searching PlasmoDB release 3.3 for all genes predicted to contain a secretory signal sequence yields 1,952 hits. Because this search used curated annotations plus the predictions from any one of several distinct gene-finding algorithms, the results are several-fold redundant, yielding about 800 distinct genes, or more than 15% of the parasite genome. More than twice as many proteins (5,003) are predicted to contain transmembrane domains, but the intersection of these results yields only 1,083 hits (about 400 distinct proteins) exhibiting both features. Next, the database can be searched for all messenger RNAs known from expressed sequence tag (EST) evidence, yielding 3,057 hits (searches based on microarray or proteomics evidence are also possible). The intersection between these secretory pathway and expression searches identifies a grand total of 190 candidates, probably corresponding to fewer than 100 distinct genes.

Two key points emerge from these queries. First, the power of a database devoted to mining genomics-scale data sets comes from its ability to form relational (integrated) queries, allowing researchers to frame their own questions. No encyclopaedic version of precomputed analyses and 'canned' queries will ever provide all possible answers in advance. For example, neither computational analysis nor manual curation would have been likely to identify enzymes associated with the apicoplast before this organelle was discovered and its targeting signals mapped.

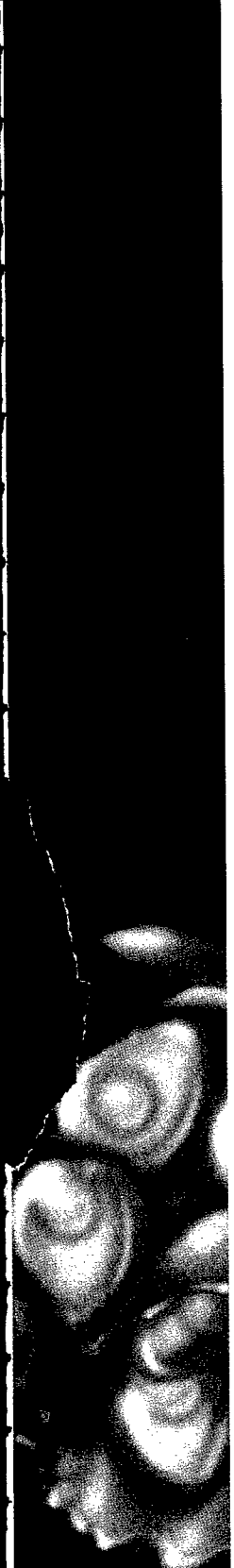
Second, the goal of these queries is not to get the 'right' answer (a provably correct list of valid drug targets or vaccine antigens), but to reduce the options, filtering the overwhelming number of sequences in the genome down to a few genes amenable to experimental analysis — in short, to let computers do what computers do well, and to let people do what people do well. Integrating the results of such studies into the database completes the loop, with computational and experimental analysis in the lab building on each other to accelerate the pace of biological research.

Jessica C. Kissinger, Brian P. Brunk, Jonathan Crabtree, Martin J. Fraunholz, Bindu Gajria, Arthur J. Milgram, David S. Pearson, Jonathan Schug, Amit Bahl, Sharon J. Diskin, Hagai Ginsburg, Gregory R. Grant, Dinesh Gupta, Philip Labo, Li Li, Matthew D. Mailman, Shannon K. McWeeney, Patricia Whetzel, Christian J. Stoeckert Jr and David S. Roos are associated with the Departments of Biology and Genetics, Center for Bioinformatics and Genomics Institute, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6018, USA. Address for correspondence: droos@sas.upenn.edu

The grand assault

Russell F. Doolittle

The complete genome sequence of the parasite responsible for most of the world's human malaria has been determined. The nature of the genome meant that this was a difficult project, requiring considerable ingenuity.



The parasite *Plasmodium falciparum*, responsible for most human malaria, is among the most studied pathogens of all time, probably surpassed only by the human immunodeficiency virus and the tuberculosis bacterium *Mycobacterium tuberculosis*. The extent of human suffering caused by malaria and its devastating costs have long been recognized by international bodies, and many initiatives have been taken over the years to try to defeat this insidious microbe[1]. In 1996, an international consortium of scientists from more than a dozen institutions set out to determine the 23 million base pairs of DNA that make up the organism's genome sequence. Their massive effort — which ended up going well beyond simple sequencing — is reported on pages 498–542 of this issue[2–8]. The avowed goal of the project was to search for chinks in the parasite's armour, so that new and effective drugs and vaccines might be developed.

Sequencing strategy

The strategy for determining the *P. falciparum* genome sequence depended on first physically separating its 14 chromosomes by the technique of pulsed gel electrophoresis. In fact, three of the chromosomes (numbers 6, 7 and 8) could not be separated from each other and were simply taken as a combined unit. Three different teams then attacked different chromosomes: a team led by the Sanger Centre, Cambridge, UK, sequenced nine[3]; The Institute for Genomic Research (TIGR), Maryland, and others took on four[4]; and a group centred on Stanford University, California, did the other[5].

In broadest outline, the DNA was mechanically sheared into random fragments, the fragments were inserted into bacteria (where they were copied every time the bacteria multiplied), and individual bacterial colonies were collected. The DNA inserts from these clones were sequenced automatically, and their order determined by assembling overlapping sequences together using a computer. Around half a million individual fragments were sequenced. As in most genome projects on this scale, the sequence determination is not really complete, and several gaps and ambiguities remain. Nonetheless, even at 95% 'finished', the published sequence must be regarded as a milestone that will be a major asset to biomedical researchers.

But why did this project take so long? After all, 23 million base pairs is not a particularly large genome by current standards (Fig. 1, overleaf), and the much larger genome of the fruitfly *Drosophila melanogaster* was apparently completed in less than a year[9]. The single biggest hurdle was the extremely biased base composition of the *P. falciparum* genome. More than 80% of the bases are either As or Ts (as opposed to Cs or Gs). In fact, regions of the genome that do not code for genes average more than 85% As or Ts, and runs of 50 As or Ts are common. Most of the genomes already sequenced have been much less skewed in their base composition.

The extreme bias made the assembly process — by which individual clones are put in their correct order by an iterative overlap process — particularly challenging. Usually, if one clone has a distinctive sequence at one end (say, its 3′ end) and another the exact same sequence at the other (5′) end, it is assumed that these sequences overlap in the genome. But for *P. falciparum*, so many clone ends were AT-rich that it was difficult to assign overlaps. As a result, new stratagems had to be devised for ordering many of the chromosomal pieces, including a heavy reliance on genetic and physical 'maps' of genomic landmarks.

For example, one type of physical map used was an 'optical' map. Here, a purified chromosome is cut into segments with an enzyme known to cut DNA at particular sequences, and the segments are separated according to size by gel electrophoresis, producing the optical map. Meanwhile, the postulated sequence is 'virtually' fragmented in a computer by breaking it at the theoretical sites at which the chosen enzyme cuts. The hypothetical fragments are then sorted by length, generating a virtual map. The agreement between the optical and virtual maps for most chromosomes was reassuringly good[3].

Identifying the genes

Extreme AT-richness aside, finding the genes in any eukaryotic genome can be problematic, because the protein-coding parts of genes (exons) are interrupted by non-coding regions (introns). (Eukaryotes can be loosely defined as organisms whose cells have nuclei and cytoskeletons, distinguishing them from the Bacteria and the Archaea — neither of which has introns in their coding sequences.) Although computer programs can identify the ends of exons that need to be joined together to form mature gene products, these are seldom 100% accurate. So there is often a need for validation that is not required in bacterial gene analysis. Such confirmation can be provided by studies of complementary DNA sequences, which directly reflect the mature gene products, or by identification of the encoded proteins themselves[6,7]. To achieve the latter, the consortium used ultrasensitive mass spectrometry — the application of which is almost sure to become a standard component of future genome projects of this sort.

Frustratingly, possible functions for fully 60% of the postulated 5,279 genes remain unknown, because these

This week's News Features section has two further articles describing reaction to publication of the *Plasmodium* genomes, and discussion of the prospects for malaria control. See pages 426 and 429.

genes match no other sequences in existing data banks. Another 5% of the genes are also classified as 'hypothetical' in this sense, although they do have counterparts — themselves with unknown functions — in other organisms. This is both surprising and disappointing. But we can be sure that many of these genes really do exist. For instance, mass spectrometry identified authentic peptides corresponding to proteins encoded by 2,391 of the genes, including many of those for which functions have not yet been found[6,7].

Remarkably, the consortium also sequenced a second plasmodial genome — that of *P. yoelii yoelii*[8], the cause of a malaria-like disease in wild African rats. More than 60% of the *P. falciparum* genes, including most general housekeeping genes, had close relatives in the *P. yoelii yoelii* genome. But the comparison also revealed a treasure-trove of differences and rearrangements. Many of these are near the ends of chromosomes, in regions that somehow control the impressive ability of plasmodial parasites to change and thereby evade recognition by the host immune system. There is evidence in both species for the kinds of genetic rearrangements and chromosomal exchanges that might be allied to this ability. But none of the genes known to be involved in immune evasion by *P. falciparum* can be recognized in *P. yoelii yoelii*. The hope had been that comparison would reveal host-specific adaptations that could be exploited in some way, but the extreme differences have confounded that strategy.

Having the entire inventory of genes for *P. falciparum* provides a complete map of its metabolic pathways, the genes encoding metabolic enzymes all being recognized by comparison with other organisms. Not only can the pathways active at different stages of the parasite life cycle be delineated, but key points at which the organism's metabolism is known to be vulnerable to attack can be seen in an overall context. For example, quinine — the first and most successful antimalarial drug — acts within a subcellular compartment of the parasite, the food vacuole, in which host haemoglobin is degraded as a foodstuff. Sulphonamides are effective because *P. falciparum* makes its own folic acid vitamins by a scheme involving *p*-aminobenzoic acid — a structure mimicked by sulphanilamide. At least four other known targets of antimalarial drugs reside in the apicoplast, an exotic quadruple-membraned compartment.

Most of these sites of drug action were known well before the genome-inspired metabolic map, although it is reassuring to see the whole landscape. It is the potential for choosing new drug targets that is exciting, however, and the consortium has now pinpointed five. For example, within the food vacuole there are several protein-degrading enzymes that might conceivably be blocked by specific inhibitors. How long it will be before these hopes are realized is unknown, but there are almost too many options to pursue. The question also arises of whether drugs that might be forthcoming would be affordable by those most in need.

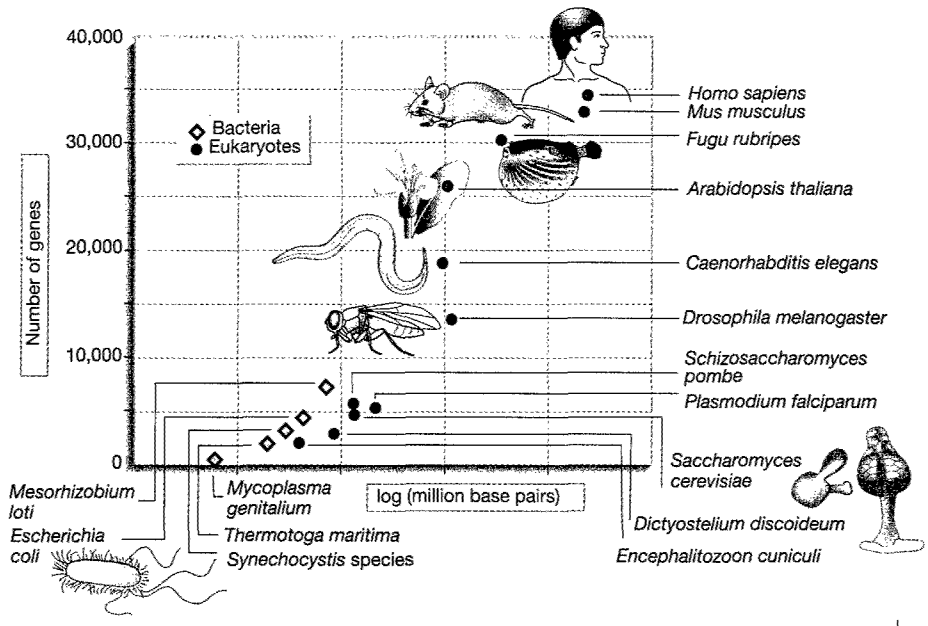


Figure 1 **Some of the genomes sequenced so far. The figure shows the number of genes plotted against genome size for the 12 fully sequenced genomes of eukaryotes and a representative set of bacteria. Note the log scale for genome size, expressed as millions of base pairs.**

During the course of this project there has been a spirited debate as to whether malaria is better attacked by large-scale genome projects[10] or by more traditional public-health measures[11]. The politically correct chorus of response has been that of course both approaches must be undertaken, especially as the true benefits of the genome studies are "down the line"[12,13].

Initially, the cost of the current project was put in the neighbourhood of US$15 million[14]. These are not easy estimates to make because funding usually comes from multiple sources and sometimes involves third parties. To the initial figure should now be added the cost of many parallel projects, including the now published sequence[15] of *Anopheles gambiae*, the mosquito that transmits malaria. The Sanger Centre and TIGR websites also list half a dozen other protozoan parasite genomes under study, including two more plasmodial strains.

So is it worth it from a medical point of view? That really remains to be seen. If one assesses what has been learned so far from the whole-genome projects (see Fig. 1) of the past six or seven years, it is clear that the 'bio' part of the biomedical enterprise has been the clear winner. Whether one studies molecular evolution or gene transcription, population genetics or developmental biology, cellular mechanics or signal transduction, whole-genome information is what defines the playing field. But for the most part, the promised medical benefits have been slow to materialize. Translating all of this information into new treatments and cures is not a trivial process.

That malaria was near eradication decades ago in some areas as a result of DDT spraying is more than a cruel irony[11,16]. Indeed, since the use of that insecticide was sharply curtailed in the 1970s, 30 million people may have died from *Plasmodium*-infected mosquito bites, and ten times that number have suffered this debilitating disease. As the participants in the current study acknowledge[2], "genome sequences alone provide little relief to those suffering from malaria".

Russell F. Doolittle is at the Center for Molecular Genetics, University of California, San Diego, La Jolla, California 92093-0634, USA.
e-mail: rdoolittle@ucsd.edu

1. Malaria Insight *Nature* **415,** 669–715 (2002).
2. Gardner, M. J. *et al. Nature* **419,** 498–511 (2002).
3. Hall, N. *et al. Nature* **419,** 527–531 (2002).
4. Gardner, M. J. *et al. Nature* **419,** 531–534 (2002).
5. Hyman, R. W. *et al. Nature* **419,** 534–537 (2002).
6. Florens, L. *et al. Nature* **419,** 520–526 (2002).
7. Lasonder, E. *et al. Nature* **419,** 537–542 (2002).
8. Carlton, J. M. *et al. Nature* **419,** 512–519 (2002).
9. Adams, M. D. *et al. Science* **287,** 2185–2195 (2000).
10. Hoffmann, S. L. *Science* **290,** 1509 (2000).
11. Curtis, C. F. *Science* **290,** 1508 (2000).
12. James, A. A. *Science* **291,** 435–436 (2001).
13. Morel, C. M. *Science* **291,** 435–436 (2001).
14. Hoffmann, S. L. *et al. Nature* **387,** 647 (1997).
15. Holt, R. A. *et al. Science* **298,** 129–149 (2002).
16. Jukes, T. H. *Am. Sci.* **51,** 355–361 (1965).

Biological revelations

Dyann F. Wirth

The genome of the malaria parasite was sequenced with the aim of learning more about how the parasite works, and with the hope that this would reveal potential drug targets. Has that hope been realized?

Malaria has confounded some of the best minds of the past century. A hundred years after the discovery that mosquitoes transmit *Plasmodium falciparum*, the major parasite that causes human malaria, we still do not know enough about the disease to defeat it permanently. But the papers on pages 498–542 of this issue[1–7], describing the complete genome sequence of *P. falciparum*, may eventually lead to new drugs and vaccines, and will certainly be an invaluable guide to future research. These papers are a testament to the success of a six-year project undertaken by an international consortium of labs and funding agencies.

First, a bit of background. The malaria parasite leads a complicated life (Fig. 1), existing mainly inside liver cells and red blood cells in its human host and, when residing in mosquitoes (notably *Anopheles gambiae*), being associated with the insect's gut and salivary glands. It undergoes several transformations along the way. The stages of its life cycle were originally described more than 100 years ago and were given names based on morphology, such as merozoite, trophozoite and gametocyte (in humans), and zygote, ookinete and sporozoite (in mosquitoes). One of the most curious features of the human stages is the human immune response — there is much immune activity, but this does not control the infection effectively, nor afford protection against future infections.

Despite massive efforts to eradicate the disease in the 1950s and early 1960s, more people are infected with malaria in Africa today than at any other time in history. Over 500 million people are infected with the disease worldwide, and one-quarter of the population is at risk of infection. More than a million children die of malaria each year, mostly in Africa. And those individuals who survive suffer a combination of anaemia and immune suppression that leaves them vulnerable to other fatal illnesses. Alarmingly, drug resistance in the parasite is now widespread.

These stark facts emphasize the need to find new treatments for the disease and new ways of preventing it. The genome project described in this issue[1–7] was conceived with these goals in mind. With the wealth of information now available at the click of a mouse, malaria researchers have an unprecedented opportunity to find genes that are potentially unique to, or at least substantially different in, *P. falciparum* compared with other species; such genes may make good drug targets, with less risk of side effects.

Even before the whole genome had been sequenced, new drug targets were being identified from searches of the partially assembled sequence data for unique genes[8]. But the total sequence will provide a more complete picture of the parasite's inner workings and the chance to identify vulnerable aspects. So just what have we learnt about the parasite's biology from this package of papers, which comprises its genome sequence[1,4–6]; a comparison of its genome with that of a rodent malaria parasite, *P. yoelii yoelii*[2]; and two proteomics studies of the proteins expressed at different stages in the parasite's life cycle[3,7]? Where are the potential weaknesses? And what have we discovered about the parasite's means of evading the human immune response?

One notable feature of the parasite's genome[1] is the apparent absence of genes for proteins that, in other species, are key to metabolism and the energetics of mitochondria — cellular powerhouses, which produce the energy-storing molecule ATP. For example, the consortium found no predicted genes for two protein components of ATP synthase, a mitochondrial ATP-producing enzyme. (At present, many of the genes are only 'predicted': they have been identified by gene-searching algorithms, but have not yet been confirmed as bona fide genes.) Similarly, there are apparently no genes for components of a conventional NADH dehydrogenase complex, another key mitochondrial enzyme. Perhaps *P. falciparum* generates and stores energy by using novel proteins or mechanisms — potential drug targets. That the mitochondria are active, at least in sporozoites and gametocytes, seems likely, given that the proteomics analyses[3,7] detected fragments of enzymes involved in some typical mitochondrial processes, including the tricarboxylic-acid cycle and oxidative phosphorylation.

Also interesting is the number of predicted genes — some 10% — that encode proteins associated with the apicoplast[1]. This essential cellular compartment is known to be important for the biosynthesis of fatty acids and isoprenoids, components of many membrane proteins, and for iron metabolism. But analysis of these genes should reveal other possible functions, and so new drug targets. The genome sequence also identifies the molecules within the apicoplast

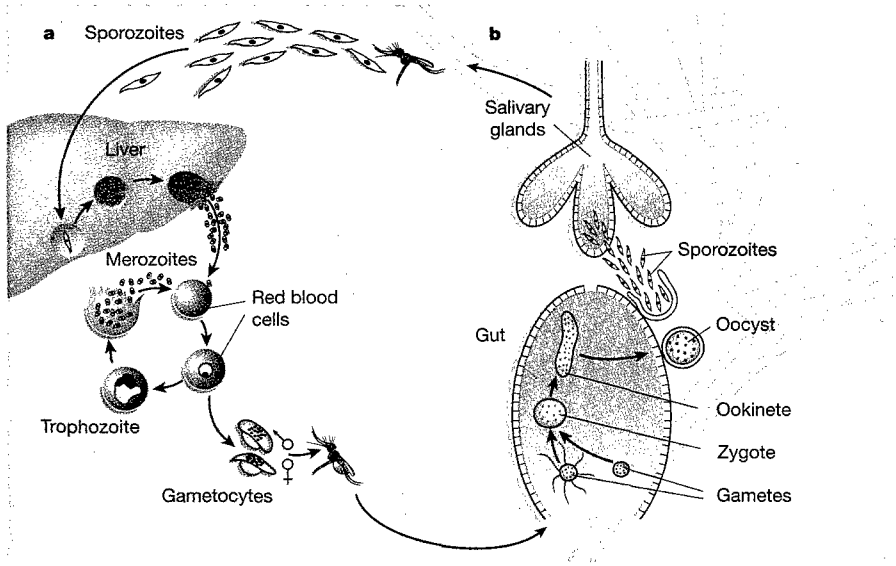


Figure 1 **Life cycle of the parasite *Plasmodium falciparum*.** a, **When a parasite-infected mosquito feeds on a human, it injects the parasites in their sporozoite form. These travel to the liver, where they develop through several stages, finally producing merozoites which invade and multiply, via the trophozoite stage, in red blood cells. Eventually, up to 10% of all red cells become infected. (Clinical features of malaria, including fever and chills, anaemia and cerebral malaria, are all associated with infected red blood cells, and most current drugs target this stage of the life cycle.) The merozoites in a subset of infected red blood cells then develop into gametocytes.** b, **When another mosquito bites the infected human, it takes up blood containing gametocytes, which develop into male and female reproductive cells (gametes). These fuse in the insect's gut to form a zygote. The zygote in turn develops into the ookinete, which crosses the wall of the gut and forms a sporozoite-filled oocyst. When the oocyst bursts, the sporozoites move to the mosquito's salivary glands, and the process begins again.**

that are the targets of several existing drugs[9].

The complex life cycle of *P. falciparum* means that the parasite has had to adapt to several different environments. So it is also intriguing that, compared with the genome of the free-living budding yeast, the parasite genome[1] encodes a limited number of predicted transporter proteins for the active uptake of nutrients from the environment. In fact, entire classes of transporters seem to be missing. It may be that several genes in this class have been overlooked because they are made up of many small coding regions, which can be missed by gene-prediction algorithms. But, taken at face value, this surprising finding implies that adequate amounts of nutrients recognized by the transporters must be present at all stages of the parasite life cycle, so that there is no selective advantage in having many transporters with differing substrate specificities. Alternatively, the parasite may use previously identified pores or channels to acquire nutrients[10,11].

Regulating protein levels

During its life cycle, *P. falciparum* undergoes several developmental changes. One of the most dramatic is sexual differentiation and the formation of gametes, male and female reproductive cells. The proteomics studies[3,7] of these stages have coincidentally shed light on a fundamental question: how does the parasite regulate the levels of its proteins? The genome[1] encodes relatively few predicted proteins that control the transcription of genes into messenger RNAs (the first step in making a protein). Moreover, there seem to be few transcriptional regulatory elements in the genome — or at least, there are few elements that are known from other organisms. Yet the proteomics analyses and previous studies show that protein abundance is tightly regulated.

The proteomics studies also show that proteins involved in processing mRNAs and in protein synthesis (translation) are expressed at higher levels in gametocytes, particularly female gametocytes, than in other stages. Interestingly, proteins that are present in early zygotes — which are produced from gametocytes — seem to be absent in gametocytes, although the mRNAs encoding these proteins are abundantly present. All of this is consistent with the proposal[12] that the regulation of protein levels is controlled through mRNA processing and translation, rather than by gene transcription. Perhaps this is a general feature of the parasite — another potential drug target.

In addition, one of the proteomics studies[3] reveals groups of genes whose regulation appears to be coordinated. Some simultaneously expressed genes are clustered in the genome; comparison of these genes and their flanking sequences may provide further insight into how they are regulated.

Immune evasion

Arguably the most striking features of the *P. falciparum* genome are the regions near the ends of each chromosome[1]. This is where families of genes that encode surface proteins, such as the *var* genes, are found. These proteins, or antigens, can sometimes be recognized by and thus stimulate the human immune system. But they have a great capacity for change, which occurs partly through the exchange of material between chromosome ends. As the genome sequence shows, the very ends of the chromosomes — the telomeres — have a complex arrangement of sequences that may facilitate such exchange (as described in ref. 13) and thereby lead to immune evasion.

The general structure of the chromosome ends is similar to that in the rodent parasite *P. yoelii yoelii*[2]. But, surprisingly, the genes that encode the variant surface antigens in *P. falciparum* are not found in *P. yoelii yoelii*, which has a different family of variant genes, originally described in a less virulent human parasite, *P. vivax*[14]. This is interesting, because it suggests that *P. yoelii yoelii*, which is often used as a model of *P. falciparum*, is in some respects more similar to *P. vivax*. It is tempting to speculate that, despite their dissimilar sequences, the genes at the ends of the *P. falciparum* and *P. yoelii yoelii* chromosomes have similar functions. But that remains to be seen.

Finally, research on the *P. falciparum var* genes has focused on their role in enabling infected red blood cells to stick to small blood vessels in the brain. This feature is associated with the fatal form of the disease, cerebral malaria. So it is interesting that one of the proteomics analyses[3] reveals that the peptides derived from many of the *var* genes occur in sporozoites, which are produced in mosquitoes and invade the human liver during the initial infection. These results point to possible alternative functions for *var* gene products.

The complete picture

One of the most exciting aspects of this huge undertaking is that it can be related to other work. We now have the genome of the mosquito *A. gambiae*[15], together with draft sequences of the human genome[16,17], and so can get a better handle on the interactions among three species that have long been evolving together. It is well known that certain variations in human genes are associated with a reduced susceptibility to malaria, and analysis of different human populations will no doubt reveal more on this. A close look at the mosquito genome should provide similar insights. Study of the parasite genome will reveal much about how *P. falciparum* interacts with its host and carrier, and more about the genes involved in parasite recognition by the human immune system. Decoding the information in these genomes, and translating it into effective remedies, is both a challenge and an opportunity for the scientific community.

Dyann F. Wirth is in the Department of Immunology and Infectious Disease, Harvard School of Public Health, 665 Huntington Avenue, Boston, Massachusetts 02115-6021, USA.
e-mail: dfwirth@hsph.harvard.edu

1. Gardner, M. J. *et al. Nature* **419,** 498–511 (2002).
2. Carlton, J. M. *et al. Nature* **419,** 512–519 (2002).
3. Florens, L. *et al. Nature* **419,** 520–526 (2002).
4. Hall, N. *et al. Nature* **419,** 527–531 (2002).
5. Gardner, M. J. *et al. Nature* **419,** 531–534 (2002).
6. Hyman, R. W. *et al. Nature* **419,** 534–537 (2002).
7. Lasonder, E. *et al. Nature* **419,** 537–542 (2002).
8. Jomaa, H. *et al. Science* **285,** 1573–1576 (1999).
9. Waller, R. F. *et al. Proc. Natl Acad. Sci. USA* **95,** 12352–12357 (1998).
10. Desai, S. A., Bezrukov, S. M. & Zimmerberg, J. *Nature* **406,** 1001–1005 (2000).
11. Kirk, K. *Nature* **406,** 949–951 (2000).
12. Dechering, K. J. *et al. Mol. Cell. Biol.* **19,** 967–978 (1999).
13. Freitas-Junior, L. H. *et al. Nature* **407,** 1018–1022 (2000).
14. del Portillo, H. A. *et al. Nature* **410,** 839–842 (2001).
15. Holt, R. A. *et al. Science* **298,** 129–149 (2002).
16. International Human Genome Sequencing Consortium *Nature* **409,** 860–921 (2001).
17. Venter, J. C. *et al. Science* **291,** 1304–1351 (2001).

The mosquito genome

The post-genomic era opens

Ennio De Gregorio and Bruno Lemaitre

The mosquito *Anopheles gambiae* is the main agent in the transmission of human malaria. Its genome sequence will in time help to devise control strategies, but will be a more immediate boon for insect biologists.

The papers that appear in this issue, describing the genome of the human malaria parasite *Plasmodium falciparum*, are published simultaneously with others in *Science* tackling the genome of the mosquito *Anopheles gambiae*. The connection is obvious: the parasite requires a mosquito to complete its complex life cycle and for transmission from one host to another. These two species are respectively the major parasite causing malaria and the major vector.

Plasmodium is taken up by mosquitoes in blood meals drawn from infected humans (see the life-cycle diagram on page 495). The parasite then undergoes several developmental stages, and crosses two mosquito cell layers that enclose the insect's midgut and salivary glands. Ultimately, *Plasmodium* is passed on when the mosquito bites a new human host, about two weeks after ingesting the first infected blood meal. For more than a century, an objective of malaria control programmes

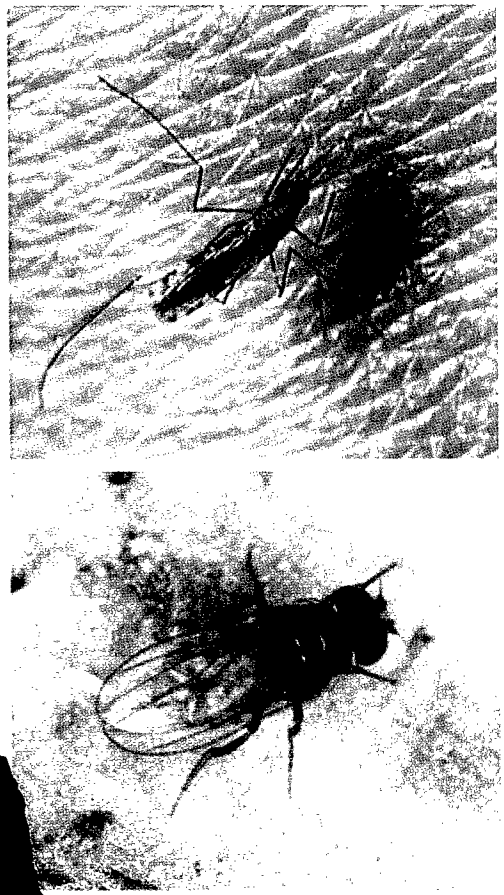


Figure 1 **The mosquito and the fruitfly in typical pose — *Anopheles* (top) on human skin, *Drosophila* on a banana.**

INSTITUT PASTEUR/P. BREY

E. DE GREGORIO

has been to block parasite transmission by mosquitoes. These approaches will clearly benefit from the improved understanding of mosquito biology and mosquito interactions with *P. falciparum* that the genome sequences will make possible.

The *A. gambiae* genome[1] was sequenced by a collaboration between Celera Genomics, the French National Sequencing Centre (Genoscope) and The Institute for Genomics Research (TIGR), in association with several university laboratories. These groups used the same 'shotgun' strategy as that applied for sequencing the human, mouse and fruitfly (*Drosophila melanogaster*) genomes. Random fragments of genomic DNA were first cloned in bacteria, and sequenced, and the overlapping clones were then assembled into contiguous sequences. Unexpectedly, the high levels of genetic variation (polymorphisms) in the reference strain of *A. gambiae* used for sequencing — the PEST strain — made the genomic assembly step difficult. The genetic variation might be explained by the fact that two distinct populations of *A. gambiae* have contributed to the PEST strain, thereby creating a mosaic genome structure. This unprecedented situation required the development of new sequence-assembly strategies, and these will be a considerable asset for future genome projects — as with mosquitoes, not all organisms are available as inbred laboratory strains.

Comparison with the fruitfly

Much of the interest in the *A. gambiae* genome will centre on comparisons with that of *D. melanogaster*, which was published two years ago[2]. These two insects belong to the same taxonomic order, the Diptera, but inhabit distinct environments and have different lifestyles (Fig. 1). *Drosophila melanogaster* feeds on decaying organic matter, such as damaged or rotting fruit, where it also completes its life cycle, whereas *A. gambiae* feeds on sugar nectar and on the blood of vertebrate hosts. Blood meals are required for female mosquitoes to produce eggs; these are laid in water, where larvae develop and hatch. Blood feeding exposes the insect to viruses and parasites — like *Plasmodium*, these other pathogens exploit *Anopheles* as a vector for transmission.

One of the main differences between the two species is that, at 278 million base pairs, the *A. gambiae* genome is much bigger than that of *D. melanogaster* (estimated to be 180 million base pairs). But this difference is not reflected in the total number of genes, which, with 13,000–14,000 genes so far identified in both insects, is surprisingly similar. It seems that, in the course of evolution, *Drosophila* has experienced a progressive reduction both in the regions between genes and in the introns, the non-protein-coding stretches of DNA within genes.

Comparison of the coding sequences reveals that the genomes of *Anopheles* and *Drosophila* are less similar than would be expected for two species that diverged 'only' 250 million years ago. Only half of the genes in the two genomes can be interpreted as orthologues — genes in different species that have common ancestry, although their functions may differ. *Anopheles* and *Drosophila* orthologues show an average of about 56% identity in DNA sequence. As Zdobnov *et al.* point out in another of the papers in *Science*[3], from the sequence standpoint, the two species differ more than do humans and pufferfish — species that diverged 450 million years ago. Some of the protein families present in both mosquito and fruitfly appear to have evolved from a common ancestral gene through independent gene-duplication in each species. The *Anopheles* genome shows several cases of such expansion which might reflect adaptation to its lifestyle. An example is the family of fibrinogen-like proteins (of which there are 58 in *Anopheles* and 13 in *Drosophila*), which in the mosquito are probably used as anticoagulant for the ingested blood meals.

Defence specialization

Insects have efficient immune systems for combating the various pathogens they encounter, and most of our knowledge in this area comes from genetic and molecular studies in *Drosophila*. Finding out how *Anopheles* responds to *Plasmodium* infection is essential for obtaining clues to controlling malaria. Christophides *et al.*[4] analysed the gene families in *A. gambiae* that are linked to insect immunity, and show that they diverge widely from those in *Drosophila*. Good examples are the prophenoloxidase enzymes (nine in the mosquito, three in the fruitfly); these enzymes catalyse the synthesis of melanin, which is associated with several defence reactions in insects.

The study by Christophides *et al.* suggests that *Anopheles* employs the same general defence mechanisms as *Drosophila*, and uses similar pathogen-activated signal-transduction pathways, but that it has adapted recognition and effector immune genes to different types of aggressors. The best characterized effector system in insects consists of antimicrobial peptides, which display a wide spectrum of antibiotic activities. Interestingly, out of seven families of these peptides found in *Drosophila*, only two are also evident in *Anopheles*: five, then, are specific to *Drosophila*. Conversely, at least one mosquito-specific antimicrobial peptide has already been identified and others might be discovered by functional studies in the future. The expression profiles of some *A. gambiae* immune genes also suggest that, like the fruitfly, the mosquito mounts specific immune responses adapted to different types of pathogen[4,5].

The availability of the entire DNA sequence, together with tools such as DNA microarrays and targeted gene disruption[6–8], will make *Anopheles* a powerful model system for studying insect biology. The genomic data will also help in developing strategies to combat malaria and other mosquito-borne human diseases, for example yellow fever, dengue, filariasis and encephalitis. Such strategies will include reducing the number and lifespan of infectious mosquitoes, analysing what attracts them to their human targets, and limiting the capacity of parasites to develop within the insect vector. Malaria is characterized by a highly complex set of interactions between the parasite, the vector and the host. Now that the genomes of all three players have been fully sequenced, the post-genomic era in combating this dreadful disease can really begin.

Ennio De Gregorio and Bruno Lemaitre are at the Centre de Génétique Moléculaire, CNRS, 91198 Gif-sur-Yvette, France.
e-mails: gregorio@cgm.cnrs-gif.fr
lemaitre@cgm.cnrs-gif.fr

1. Holt, R. A. *et al. Science* **298,** 129–149 (2002).
2. Special section on the *Drosophila* genome *Science* **287,** 2181–2224 (2000).
3. Zdobnov, A. M. *et al. Science* **298,** 149–159 (2002).
4. Christophides, G. K. *et al. Science* **298,** 159–165 (2002).
5. Dimopoulos, G. *et al. Proc. Natl Acad. Sci. USA* **99,** 8814–8819 (2002).
6. Catteruccia, F. *et al. Nature* **405,** 959 (2000).
7. Grossman, G. L. *et al. Insect Mol. Biol.* **10,** 597–604 (2001).
8. Blandin, S. *et al. EMBO Rep.* **3,** 852–856 (2002).

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).

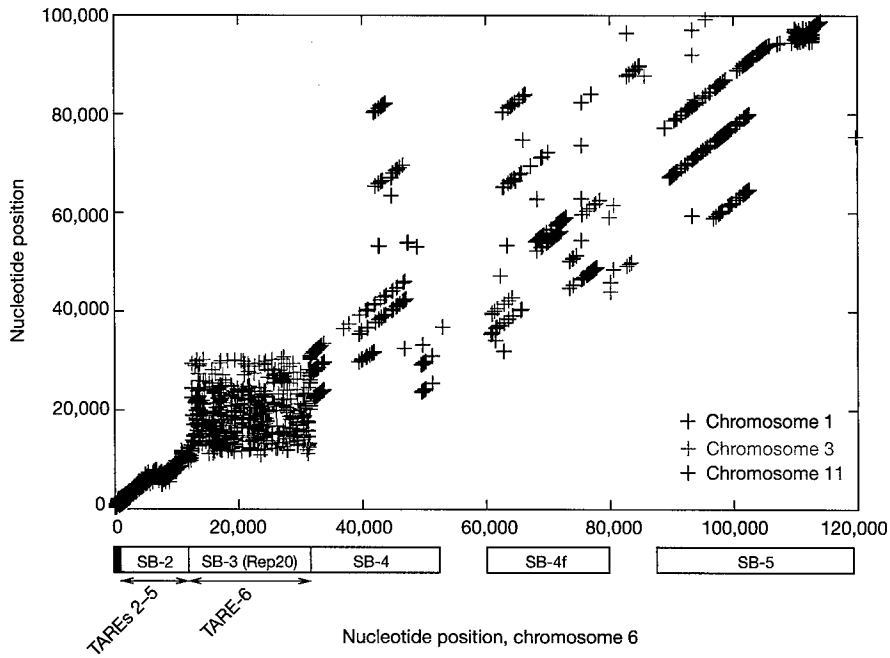


Figure 2 Alignment of subtelomeric regions of chromosomes 1, 3, 6 and 11. MUMmer2[152] alignments showing exact matches between the left subtelomeric regions of chromosome 6 (horizontal axis) and chromosomes 11 (red), 1 (blue) and 3 (green), illustrating the conserved synteny between all telomeres. Each 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

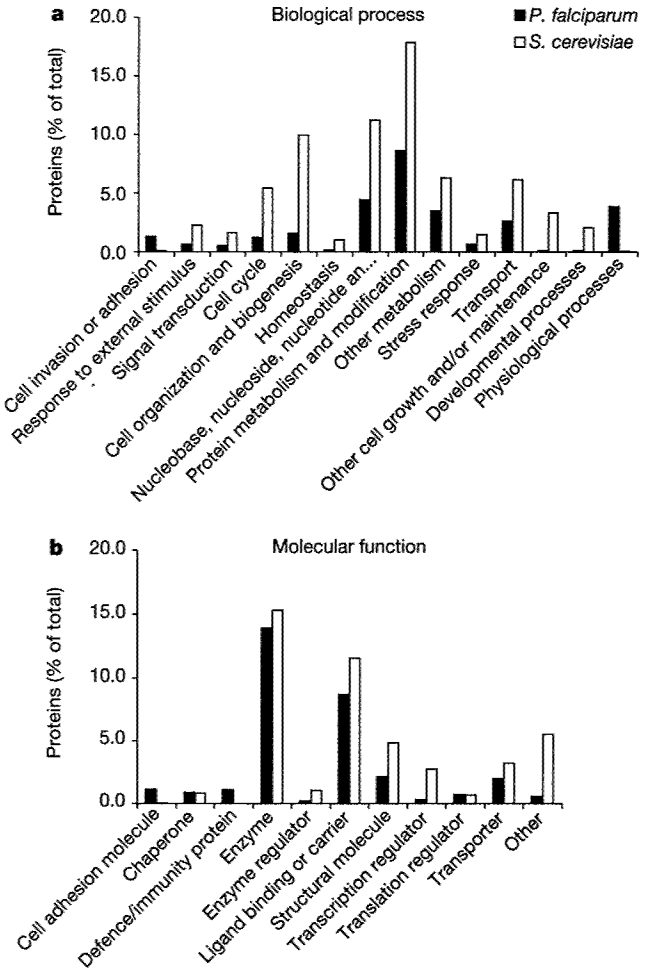


Figure 3 Gene Ontology classifications. Classification of *P. falciparum* proteins according to the 'biological process' (**a**) and 'molecular function' (**b**) ontologies of the Gene Ontology system[42].

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

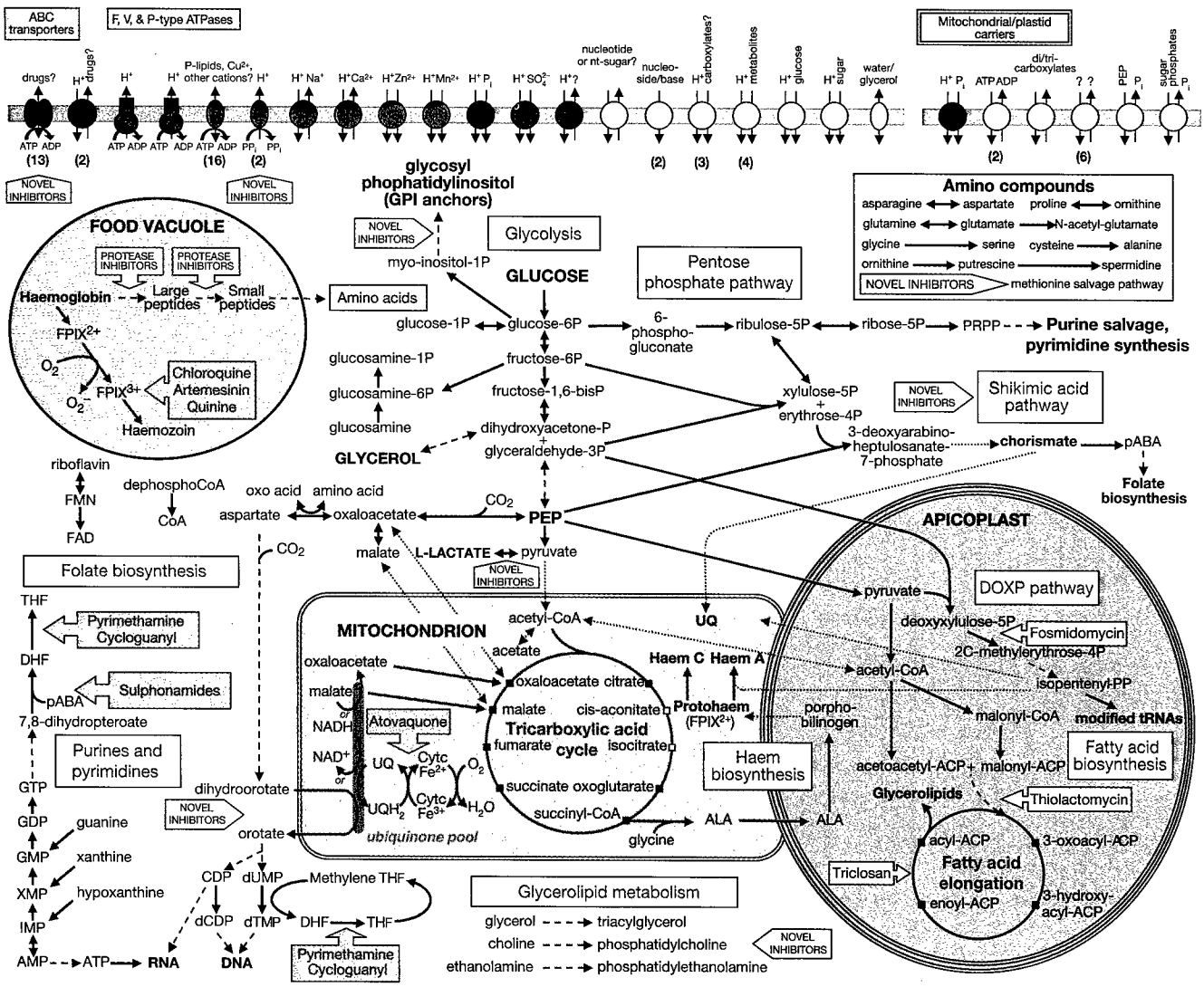


Figure 4 Overview of metabolism and transport in *P. falciparum*. Glucose and glycerol provide the major carbon sources for malaria parasites. Metabolic steps are indicated by arrows, with broken lines indicating multiple intervening steps not shown; dotted arrows indicate incomplete, unknown or questionable 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].

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.

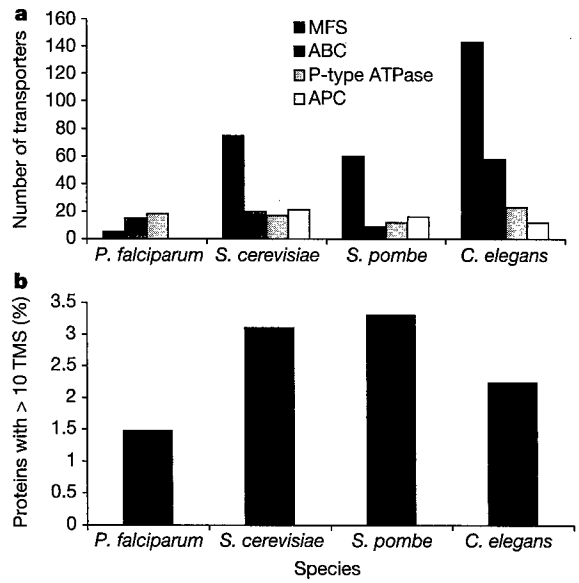


Figure 5 Analysis of transporters in *P. falciparum*. **a**, Comparison of the numbers of transporters belonging to the major facilitator superfamily (MFS), ATP-binding cassette (ABC) family, P-type ATPase family and the amino acid/polyamine/choline (APC) family in *P. falciparum* and other eukaryotes. 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.

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

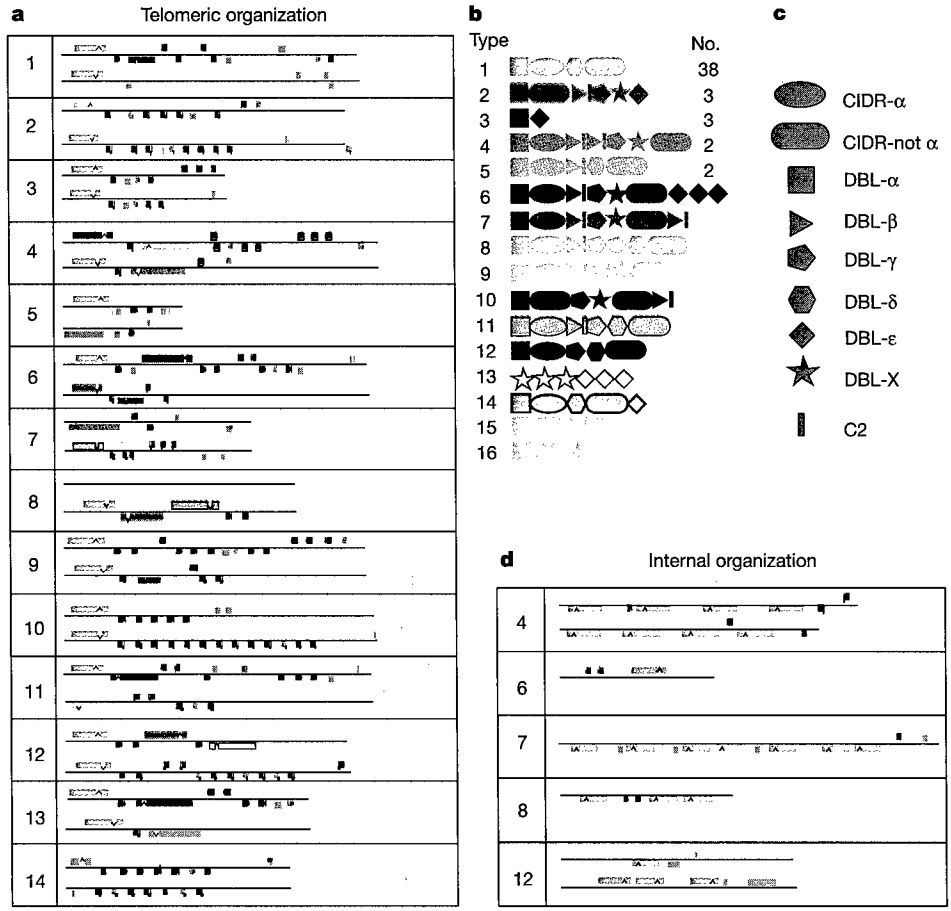


Figure 6 Organization of multi-gene families in *P. falciparum*. **a**, Telomeric regions of all chromosomes showing the relative positions of members of the multi-gene families: *rif* (blue) *stevor* (yellow) and *var* (colour coded as indicated; see **b** and **c**). Grey boxes represent pseudogenes 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., Munholland, 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).

Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*

Jane M. Carlton*, Samuel V. Angiuoli*, Bernard B. Suh*, Taco W. Kooij†, Mihaela Pertea*, Joana C. Silva*, Maria D. Ermolaeva*, Jonathan E. Allen*, Jeremy D. Selengut*, Hean L. Koo*, Jeremy D. Peterson*, Mihai Pop*, Daniel S. Kosack*, Martin F. Shumway*, Shelby L. Bidwell*, Shamira J. Shallom*, Susan E. van Aken*, Steven B. Riedmuller*, Tamara V. Feldblyum*, Jennifer K. Cho*‡, John Quackenbush*, Martha Sedegah§, Azadeh Shoaibi*, Leda M. Cummings*‡, Laurence Florens||, John R. Yates||, J. Dale Raine¶, Robert E. Sinden¶, Michael A. Harris#, Deirdre A. Cunningham☆, Peter R. Preiser☆, Lawrence W. Bergman, Akhil B. Vaidya**, Leo H. van Lin†, Chris J. Janse†, Andrew P. Waters†, Hamilton O. Smith#, Owen R. White*, Steven L. Salzberg*, J. Craig Venter††, Claire M. Fraser*, Stephen L. Hoffman‡§, Malcolm J. Gardner* & Daniel J. Carucci§**

** The Institute for Genomic Research, 9712 Medical Center Drive; and †† The Center for the Advancement of Genomics, 1901 Research Boulevard, Rockville, Maryland 20850, USA*
† Department of Parasitology, Leiden University Medical Centre, PO Box 9600, 2300 RC Leiden, The Netherlands
§ Naval Medical Research Center, Malaria Program (IDD), Silver Spring, Maryland 20910, USA
|| Department of Cell Biology, The Scripps Research Institute, La Jolla, California, 92037, USA
¶ Infection & Immunity Section, Department of Biological Sciences, Imperial College of Science, Technology & Medicine, London, SW7 2AZ, UK
 # Celera Genomics, 45 West Gude Drive, Rockville, Maryland 20850, USA
☆ Division of Parasitology, National Institute for Medical Research, London, UK
*** Division of Molecular Parasitology, Department of Microbiology & Immunology, Drexel University College of Medicine, Philadelphia, Pennsylvania 191929, USA*

Species of malaria parasite that infect rodents have long been used as models for malaria disease research. Here we report the whole-genome shotgun sequence of one species, *Plasmodium yoelii yoelii*, and comparative studies with the genome of the human malaria parasite *Plasmodium falciparum* clone 3D7. A synteny map of 2,212 *P. y. yoelii* contiguous DNA sequences (contigs) aligned to 14 *P. falciparum* chromosomes reveals marked conservation of gene synteny within the body of each chromosome. Of about 5,300 *P. falciparum* genes, more than 3,300 *P. y. yoelii* orthologues of predominantly metabolic function were identified. Over 800 copies of a variant antigen gene located in subtelomeric regions were found. This is the first genome sequence of a model eukaryotic parasite, and it provides insight into the use of such systems in the modelling of *Plasmodium* biology and disease.

For decades, the laboratory mouse has provided an alternative platform for infectious disease research where the pathogen under study is intractable to routine laboratory manipulation. Experimental study of the human malaria parasite *Plasmodium falciparum* is particularly problematic as the complete life cycle cannot be maintained *in vitro*. Four species of rodent malaria (*Plasmodium yoelii*, *Plasmodium berghei*, *Plasmodium chabaudi* and *Plasmodium vinckei*) isolated from wild thicket rats in Africa have been adapted to grow in laboratory rodents[1]. These species reproduce many of the biological characteristics of the human malaria parasite. Many of the experimental procedures refined for use with *P. falciparum* were initially developed for rodent malaria species, a prime example being stable genetic transformation[2]. Thus rodent models of malaria have been used widely and successfully to complement research on *P. falciparum*.

With the advent of the *P. falciparum* Genome Sequencing Project, undertaken by an international consortium of genome sequencing centres and malaria researchers, a series of initiatives has begun to generate substantial genome information from additional *Plasmodium* species[3]. We describe here the genome sequence of the rodent malaria parasite *P. y. yoelii* to fivefold genome coverage. We show that this partial genome sequencing approach, although limited in its application to the study of genome structure, has proved to be an effective means of gene discovery and of jump-starting experimental studies in a model *Plasmodium* species. Furthermore, we show that despite the considerable divergence between the *P. y. yoelii* and *P. falciparum* genomes, sequencing and annotation of the former can substantially improve the accuracy and efficiency of annotation of the latter.

Plasmodium yoelii yoelii genome sequencing and annotation

We applied the whole-genome shotgun (WGS) sequencing approach, used successfully to sequence and assemble the first large eukaryotic genome[4], to achieve fivefold sequence coverage of the genome of a clone of the 17XNL line of *P. y. yoelii* (Table 1). This level of coverage is expected to comprise 99% of the genome[5] assuming random library representation. As with *P. falciparum*, the genomes of rodent malaria parasites are highly (A + T)-rich[6], which adversely affects DNA stability in plasmid libraries. Consequently, all ~220,000 reads were produced from clones originating

Table 1 ***Plasmodium yoelii yoelii* genome coverage statistics**

Data	Component	Value
Genome	No. of contigs	5,687
	Mean contig size (kb)	3.6
	Max. contig size (kb)	51.5
	Cumulative contig length (Mb)	23.1
	No. of singletons	11,732
	No. of groups	2,906
	Max. group size (kb)	69.8
	Cumulative group size (Mb)	21.6
Transcriptome	No. of ESTs	13,080
	Average length (nucleotides)	497
Proteome	No. of gametocyte peptides	1,413
	No. of sporozoite peptides	677

‡ Present addresses: National Center for Biotechnology Information, National Institutes of Health, Bethesda, Maryland 20894, USA (L.M.C.); Genentech, San Francisco, California 94080, USA (J.K.C.); and Sanaria, 308 Argosy Drive, Gaithersburg, Maryland 20878, USA (S.L.H.).

from small (2–3 kilobases (kb)) insert libraries. Contigs were assembled using TIGR Assembler[7]. Contaminating mouse sequences, identified through similarity searches and found to comprise 10% of the total sequence data, were excluded from the analyses. Approximately three-quarters of the contigs could be placed into 2,906 'groups', each group consisting of two or more contigs known to be linked through paired reads as determined by Grouper software[7]. This produced an average group size of 7.4 kb, approximately 4 kb more than the average contig size. This group size is small compared with the group data produced by other partial eukaryotic genome projects, where extensive use of large insert (linking) libraries has enabled the construction of ordered and orientated 'scaffolds'[8], and emphasizes the use of such linking libraries in partial genome projects. The genome size of *P. y. yoelii* is estimated to be 23 megabases (Mb), in agreement with karyotype data[9].

Expression data from the *P. y. yoelii* transcriptome and proteome were generated to aid in gene identification and annotation of the contigs (Table 1). A total of 13,080 expressed sequence tag (EST) sequences generated from clones of an asexual blood-stage *P. y. yoelii* complementary DNA library[10], in combination with other *P. yoelii* ESTs and transcript sequences available from public databases, were assembled and used to compile a gene index[11] of expressed *P. y. yoelii* sequences (http://www.tigr.org/tdb/tgi/pygi/). For protein expression data, multidimensional protein identification technology (MudPIT), which combines high-resolution liquid chromatography with tandem mass spectrometry and database searching, was applied to the gametocyte and salivary gland sporozoite proteomes of *P. y. yoelii*. A total of 1,413 gametocyte and 677 sporozoite peptides were recorded and used for the purposes of gene annotation.

We used two gene-finding programs, GlimmerMExon and Phat[12], to predict coding regions in *P. y. yoelii*. GlimmerMExon is based on the eukaryotic gene finder GlimmerM[13], with modifications developed for analysing the short fragments of DNA that result from partial shotgun sequencing. Gene models based on GlimmerMExon and Phat predictions were refined using Combiner. Annotation of predicted gene models used TIGR's fully automated Eukaryotic Genome Control suite of programs. Gene finding and subsequent annotation were limited to 2,960 contigs (each of which is over 2 kb in size), a subset of sequences that contains more than 20 Mb of the genome. A total of 5,878 complete genes and 1,952 partial genes (defined as genes lacking either an annotated start or stop codon) can be predicted from the nuclear genome data.

Table 2 **Comparison of genome features of *P. falciparum* and *P. y. yoelii***

Feature	*P. y. yoelii*	*P. falciparum*
Size (Mb)	23.1	22.9
No. of chromosomes	14	14
No. of gaps	5,812	93
Coverage*	5	14.5
(G + C) content (%)	22.6	19.4
No. of genes†	5,878	5,268
Mean gene length (bp)	1,298	2,283
Gene density (bp per gene)	2,566	4,338
Per cent coding	50.6	52.6
Genes with introns (%)	54.2	53.9
Genes with ESTs (%)	48.9	49.1
Gene products detected by proteomics (%)	18.2	51.8
Exons		
Mean no. per gene	2.0	2.4
(G + C) content (%)	24.8	23.7
Mean length (bp)	641	949
Introns		
(G + C) content (%)	21.1	13.5
Mean length (bp)	209	179
Total length (bp)	1,687,689	1,323,509
Intergenic regions		
(G + C) content (%)	20.7	13.6
Mean length (bp)	859	1,694
RNAs		
No. of tRNA genes‡	39	43
No. of 5S rRNA genes	3	3
No. of 5.8S, 18S and 28S rRNA units	4	7
Mitochondrial genome		
(G + C) content (%)	31	31
Apicoplast genome		
(G + C) content (%)	15	14

*Average number of sequence reads per nucleotide.
†Total number of full-length genes.
‡The smaller number reflect the partial nature of the *P. y. yoelii* genome data.

Comparative genome analysis

A comparison of several genome features of *P. falciparum* and *P. y. yoelii* is shown in Table 2, demonstrating that many similarities exist between the genomes. Besides the similarly extreme (G + C) compositions, both genomes contain a comparable number of predicted full-length genes, with the higher figure in *P. y. yoelii* due to an extremely high copy number of variant antigen genes (see below). Where differences between the genomes do exist, such as the (G + C) content of the coding portion of the genomes, incompleteness of the *P. y. yoelii* genome data, with the associated problems of accurate gene finding in both species, is likely to be a confounding factor. As an indication of this problem, analysis of *P. y. yoelii* proteomic data identified 83 regions of the genome apparently expressed during sporozoite and/or gametocyte stages but not assigned to a *P. y. yoelii* gene model (data not shown). Many of these peptide hits appear sufficiently close to a model as to indicate a fault with gene boundary prediction rather than a lack of gene prediction *per se*. However, as with the gene model prediction in *P. falciparum*, the gene models of *P. y. yoelii* should be considered preliminary and under revision.

Identifying orthologues of *P. falciparum* vaccine candidate proteins and proteins that are either targets of antimalarial drugs or involved in antimalarial drug resistance mechanisms is a primary goal of model malaria parasite genomics. Using BLASTP[14] with a cutoff E value of 10^{-15} and no low-complexity filtering, 3,310 bidirectional orthologues (defined as genes related to each other through vertical evolutionary descent) can be identified in the full protein complement of *P. falciparum* (5,268 proteins) and the protein complement of *P. y. yoelii* translated from complete gene models (5,878 proteins). A list of vaccine candidate orthologues and orthologues of genes involved in antimalarial drug interactions identified from among the 3,310 orthologues and from additional BLAST analyses is shown in Table 3. Those genes that are not identifiable may either be absent from the partial genome data, or represent genes that have been lost or diverged sufficiently that they are undetectable through similarity searching.

Many of the candidate vaccine antigens under study in *P. falciparum* can be identified in *P. y. yoelii*, including orthologues of several asexual blood-stage antigens known to elicit immune responses in individuals exposed to natural infection (MSP1, AMA1, RAP1, RAP2). As immunity to *P. falciparum* blood-stage infection can be transferred by immune sera, identification of the targets of potentially protective antibody responses after natural infection can provide information beneficial to the selection of candidate antigens for malaria vaccines. We found several orthologues of known *P. falciparum* transmission-blocking candidates; in particular, members of the P48/45 gene family identified previously[15] were confirmed.

We identified several *P. y. yoelii* orthologues of *P. falciparum* biochemical pathway components under study as targets for drug design (Table 3), most notably: (1) the 1-deoxy-D-xylulose 5-phosphate reductoisomerase (DOXPR) gene whose product is inhibited by fosmidomycin in *P. falciparum in vitro* cultures and mice infected with *P. vinckei*[16]; (2) enoyl-acyl carrier protein (ACP) reductase (FABI) whose product is inhibited by triclosan in *P. falciparum in vitro* cultures and mice infected with *P. berghei*[17]; and (3) a gene encoding farnesyl transferase (FTASE), which is inhibited in cultures of *P. falciparum* treated with custom-designed peptidomimetics[18]. The rodent models of malaria have proved

Table 3 ***P. y. yoelii* orthologues of *P. falciparum* candidate vaccine and drug interaction genes**

P. falciparum gene	*Pf* chromosome	ST location*	*Pf* locus	*Py* locus
Candidate vaccine antigens				
Ring-infected erythrocytic surface antigen 1, *resa1*	1	Yes	PFA0110w	Not identified
Merozoite surface protein 4, *msp4*	2	No	PFB0310c	PY07543†
Merozoite surface protein 5, *msp5*	2	No	PFB0305c	PY07543†
Liver stage antigen 3, *lsa3*	2	No	PFB0915w	Not identified
Merozoite surface protein 2, *lsa3*	2	No	PFB0300c	Not identified
Transmission-blocking target antigen 230, *Pfs230*	2	No	PFB0405w	PY03856
Circumsporozoite protein, *csp*	3	No	MAL3P2.11	PY03168
Rhoptry-associated protein 2, *rap2*	5	Yes	PFE0080c	PY03918
Sporozoite surface antigen, *starp*	7	Yes	PF07_0006	Not identified
Morozoite surface protein 1, *msp1*	9	No	PFI475w	PY05748
Liver stage antigen 1, *lsa1*	10	No	PF10_0356	Not identified
Merozoite surface protein 3, *msp3*	10	No	PF10_0345	Not identified
Glutamate-rich protein, *glurp*	10	No	PF10_0344	Not identified
Ookinete surface protein 25, *Pfs25*	10	No	PF10_0303	PY00523
Ookinete surface protein 28, *Pfs28*	10	No	PF10_0302	PY00522
Erythrocyte membrane-associated 332 antigen, *Pf332*	11	No	PF11_0507	PY06496
Apical membrane antigen 1, *ama1*	11	No	PF11_0344	PY01581
Exported protein 1, *exp1*	11	No	PF11_0224	Not identified
Surface sporozoite protein 2, *ssp2*	13	No	PF13_0201	PY03052
Sexual-stage-specific surface antigen 48/45, *Pfs48/45*	13	No	PF13_0247	PY04207
Rhoptry-associated protein 1, *rap1*	14	Yes	PF14_0637	PY00622
Candidate drug interaction genes				
Dihydrofolate reductase, *dhfr*	4	No	PFD0830w	PY04370
Multidrug resistance protein 1, *pfmdr1*	5	No	PFE1150w	PY00245
Translationally controlled tumour protein, *tctp*	5	No	PFE0545c	PY04896
Farnesyl transferase, *ftase*	5	No	PFE0970w	PY06214
Enoyl-acyl carrier reductase, *fabi*	6	No	MAL6P1.275	PY03846
Dihydro-protate dehydrogenase, *dhod*	6	No	MAL6P1.36	PY02580
Chloroquine-resistance transporter, *pfcrt*	7	No	MAL7P1.27	PY05061
Dihydropteroate synthase, *dhps*	8	No	PF08_0095	PY02226
Lactate dehydrogenase, *ldh*	13	No	PF13_0141	PY03885
DOXP reductoisomerase, *doxpr*	14	No	PF14_0641	PY05578

A full listing of all orthologues can be found as Table A in the Supplementary Information. *Pf*, *P. falciparum*; *Py*, *P. y. yoelii*.
*ST, subtelomeric. Defined as >75% of the distance from the centre to the end of the *P. falciparum* chromosome.
†Homologue of *P. falciparum msp4* and *msp5* genes found as a single gene *msp4/5* in *P. y. yoelii* and other rodent malaria species[62].

invaluable both for the study of potency of new antimalarial compounds *in vivo*, and for the elucidation of mechanisms of antimalarial drug resistance.

We applied the Gene Ontology (GO) gene classification system[19], which uses a controlled vocabulary to describe genes and their function, to indicate which classes of gene among the 3,310 orthologues might differ in number between *P. falciparum* and *P. y. yoelii* (Fig. 1). A similar proportion of proteins were identified for most of the GO classes between the two species, with the caveat that fewer total numbers of proteins were identified in *P. y. yoelii* owing to the partial nature of the genome data for this species. However, proteins allocated to the physiological processes, cell invasion and adhesion, and cell communication categories were significantly reduced in *P. y. yoelii*. These classes contain members of three multigene families whose genes are found predominantly in the subtelomeric regions of *P. falciparum* chromosomes: PfEMP1, the protein product of the *var* gene family known to be involved in antigenic variation, cyto-adherence and rosetting, and rifins and stevors, which are clonally variant proteins possibly involved in antigenic variation and evasion of immune responses (reviewed in ref. 20). Apparently, *P. falciparum* has generated species-specific, subtelomeric genes involved in host cell invasion, adhesion and antigenic variation, homologues of which are not found in the *P. y. yoelii* genome.

Gene families of unique interest in the *P. y. yoelii* genome

The largest family of genes identified in the *P. y. yoelii* genome is the *yir* gene family, homologues of the *vir* multigene family recently described in the human malaria parasite *Plasmodium vivax*[21] and in other species of rodent malaria[22]. In *P. vivax*, an estimated 600–1,000 copies of the subtelomerically located *vir* gene encode proteins that are immunovariant in natural infections, indicating a possible functional role in antigenic variation and immune evasion. Within the *P. y. yoelii* genome data, 838 *yir* genes (693 full genes and 145 partial genes) are present (Table 4; see also Supplementary Figs A and B). Almost 75% of the annotated contigs identified as containing subtelomeric sequences (see below) contain *yir* genes, many arranged in a head-to-tail fashion. Expression data indicate that *yir* genes are expressed during sporozoite, gametocyte and erythrocytic stages of the parasite, similar to the expression pattern seen with *P. falciparum var* and *rif* genes[23]. Preliminary

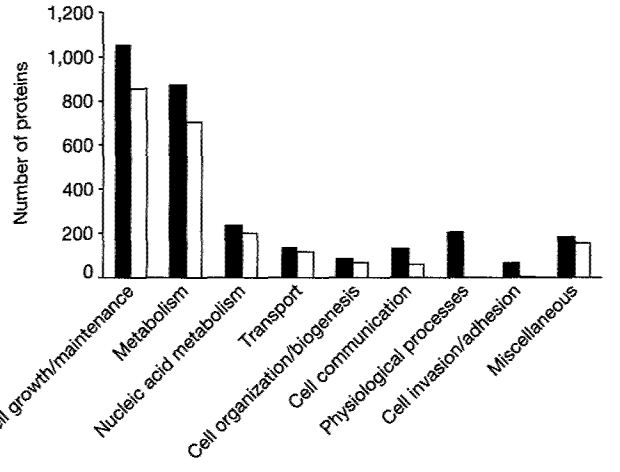


Figure 1 Functional classification comparison between *P. falciparum* and *P. y. yoelii* proteins. We compared the GO terms of proteins assigned to 'biological process' for the orthologous genes identified between the two species. The process group contains 3,041 *P. falciparum* annotations (filled bars), and 2,161 reciprocal annotations are shown for *P. y. yoelii* (open bars). Ten GO classes with similar numbers of *P. falciparum* and *P. y. yoelii* proteins in each are assigned as 'miscellaneous'; that is, cell cycle, external stimulus response, stress response, signal transduction, homeostasis, developmental processes, cell proliferation, membrane fusion, death, cell motility.

Table 4 **Paralogous gene families in *P. y. yoelii***

Gene family	No.	Name	HMM ID	Location in *Py*	*Py* expression*	*Pf* locus	TM/SP†
yir/bir/cir	838	Variant antigen family	TIGR01590	Subtelomeric	Gmt, spz, bs	None	P/A
235 kDa	14	Reticulocyte binding family	TIGR01612	Subtelomeric	Gmt, spz, bs	PFD0110w, MAL13P1.176, PF13_0198, PFL2520w, PFD0110w	P/A
pyst-a	168	Hypothetical	TIGR01599	Subtelomeric	Gmt, spz	PF14_0604	A/A
pyst-b	57	Hypothetical	TIGR01597	Subtelomeric	Bs	None	P/A
pyst-c	21	Hypothetical	TIGR01601, TIGR01604	Subtelomeric	Bs	None	P/P
pyst-d	17	Hypothetical	TIGR01605	Subtelomeric	Gmt	None	P/P
etramp	11	Early transcribed membrane protein family	TIGR01495	Subtelomeric	Gmt, spz, bs	PF13_0012, PF14_0016, PF11_0040, PFB0120w, PF10_0323, MAL12P1.387, PF11_0039, PFL1095c, PF10_0019, PF1745c, PFE1590w, PF10_0164, MAL8P1.6, PFA0195w, PFL0065w, PF14_0729	P/P
pst-a	12	Hydrolase family	TIGR01607	Subtelomeric	Gmt, spz	PFL2530w, PF10_0379, PF14_0738, PF14_0017, PF14_0737, PFI800w, PFI1775w, PF07_0040, PF07_0005, PFA0120c	A/A
rhoph1/clag	2	Rhoptry H1/ cyto-adherence-linked asexual gene family	PF03805	Subtelomeric	Gmt, bs	PFC0110w, PFC0120w, PFI1730w, PFI1710w, PFB0935w	A/P

*Found in, but not limited to: gmt, gametocyte life stage; spz, sporozoite life stage; bs, asexual blood stage.
†TM, transmembrane domain; SP, signal peptide; P, predicted; A, absent. TM and SP predictions were identical for *P. falciparum* and *P. y. yoelii* members of the same gene family. (See ref. 30 for details regarding TM and SP prediction algorithms.)

results using antibodies developed against the conserved regions of the protein have confirmed protein localization at the surface of the infected red blood cell (D.A.C. *et al.*, manuscript in preparation). The number of gene copies in the *P. y. yoelii* genome, the localization and stage-specific expression of gene members, as well as the existence of homologues in other *Plasmodium* species, make this gene family a prime target for the study of mechanisms of immune evasion.

A maximum of 14 members of the *Py235* multigene family can be identified among the *P. y. yoelii* protein data (Table 4). This family expresses proteins that localize to rhoptries (organelles that contain proteins involved in parasite recognition and invasion of host red blood cells). *Py235* genes exhibit a newly discovered form of clonal antigenic variation, whereby each individual merozoite derived from a single parent schizont has the propensity to express a different Py235 protein[24]. Closely related homologues of the *Py235* gene family have been found in other rodent malaria species, and more distantly related homologues have been found in *P. vivax*[25] and *P. falciparum*[26]. The gene copy number identified in the current data set is less than has been predicted in other *P. y. yoelii* lines (30–50 per genome). This could reflect real differences in copy number between lines, but more probably suggests an error in the original estimate or misassembly of extremely closely related sequences. Almost all of the *Py235* genes are found on contigs identified as subtelomeric in the *P. y. yoelii* genome (see Supplementary Fig. C).

Four further paralogous gene families, *pyst-a* to *-d*, are specific to *P. y. yoelii* (Table 4). The *pyst-a* family deserves mention, as it is homologous to a *P. chabaudi* glutamate-rich protein[27] and to a single hypothetical gene on *P. falciparum* chromosome 14, suggesting expansion of this family in the rodent malaria species from a common ancestral *Plasmodium* gene. Two paralogous gene families containing multiple members are homologous to multigene families identified in *P. falciparum*. Gene members of one family, *etramp* (early transcribed membrane protein), have previously been identified in *P. falciparum*[28] and in *P. chabaudi* where a single member has been identified and localized to the parasitophorous vacuole membrane[29].

Telomeres and chromosomal exchange in subtelomeric regions

The telomeric repeat in *P. y. yoelii* is AACCCTG, which differs from the *P. falciparum* telomeric repeat AACCCTA by one nucleotide. A total of 71 contigs were found to contain telomeric repeat sequences arranged in tandem, with the largest array consisting of 186 copies. The *P. y. yoelii* subtelomeric chromosomal regions show little repeat structure compared with those of *P. falciparum*. A survey of tandem repeats in the entire genome found only a few in the telomeric or subtelomeric regions, specifically a 15 base pair (bp) (45 copies) and a 31-bp (up to 10 copies), both of which were found on multiple contigs, and a 36-bp repeat that occurred on one contig. No repeat element that corresponds to Rep20, a highly variable 21-bp unit that spans up to 22 kb in *P. falciparum* telomeres, was found.

The telomeric and subtelomeric regions of *P. y. yoelii* contigs show extensive large-scale similarity, indicating that these regions undergo chromosomal exchange similar to that reported for *P. falciparum* (see ref. 30). The longest subtelomeric contig is approximately 27 kb (see Supplementary Fig. C) and is homologous to other subtelomeric contigs across its entire length, indicating that the region of chromosomal exchange extends at least this distance into the subtelomeres. Recent data have shown that clustering of telomeres at the nuclear periphery in asexual and sexual stage *P. falciparum* parasites may promote sequence exchange between members of subtelomeric virulence genes on heterologous chromosomes, resulting in diversification of antigenic and adhesive phenotypes (see ref. 31 for review). The suggestion of extensive chromosome exchange in *P. y. yoelii* indicates that a similar system for generating antigenic diversity of the *yir*, *Py235* and other gene families located within subtelomeric regions may exist.

A genome-wide synteny map

The *Plasmodium* lineage is estimated to have arisen some 100–180 million years ago[32], and species of the parasite are known to infect birds, mammals and reptiles[33]. On the basis of the analysis of small subunit (SSU) ribosomal RNA sequences, the closest relative to *P. falciparum* is *Plasmodium reichenowi*, a parasite of chimpanzees, with the rodent malaria species forming a distinct clade[34,35]. Early gene mapping studies have shown that regions of gene synteny exist between species of rodent malaria[9] and between human malaria species[36,37], despite extensive chromosome size polymorphisms between homologous chromosomes[38]. This level of gene synteny seems to decrease as the phylogenetic distance between *Plasmodium* species increases[39]. Before the *Plasmodium* genome sequencing

projects, the degree to which conservation of synteny extended across *Plasmodium* genomes was not fully apparent.

Using the *P. falciparum* and *P. y. yoelii* genome data, we have constructed a genome-wide syntenic map between the species. To avoid confounding factors inherent in DNA-based analyses of (A + T)-rich genomes, we first calculated the protein similarity between all possible protein-coding regions in both data sets using MUMmer[40]. Sensitivity was ensured through the use of a minimum word match length of five amino acids chosen to identify seed maximal unique matches (MUMs). By comparison, the recent human–mouse synteny analysis used a match length of 11 (ref. 8). Using this method, which is independent of gene prediction data, 2,212 sequences could be aligned (tiled) to *P. falciparum* chromosomes, representing a cumulative length of 16.4 Mb of sequence, or over 70% of the *P. y. yoelii* genome (see Supplementary Table C). The per cent of each *P. falciparum* chromosome covered with *P. y. yoelii* matches varies from 12% (chromosome 4) to 22% (chromosomes 1 and 14), with an average of about 18%. The spatial arrangement of the tiling paths (see Fig. 1 in ref. 30) confirms previous suggestions[9] that most of the conserved matches are found within the body of *Plasmodium* chromosomes, and confirms the absence of *var*, *rif* and *stevor* homologues in the *P. y. yoelii* genome.

Although the tiling paths indicate the degree of conservation of gene order between *P. falciparum* and *P. y. yoelii*, longer stretches of contiguous *P. y. yoelii* sequence are necessary to examine this feature in depth. Accordingly, we carried out linkage of many *P. y. yoelii* assemblies adjacent to each other along the tiling paths. First, 1,050 adjacent contigs were linked on the basis of paired reads as determined by Grouper software. Second, *P. y. yoelii* ESTs were aligned to the tiling paths, and those found to overlap sequences adjacent in the tiling path were used as evidence to link a further 236 *P. y. yoelii* sequences. Third, amplification of the sequence between adjacent contigs in the tiling paths linked a further 817 assemblies. Linkage of *P. y. yoelii* sequences by these methods resulted in the formation of 457 syntenic groups from 2,212 original contigs, ranging in length from a few kilobases to more than 800 kb. Syntenic groups were assigned to a *P. y. yoelii* chromosome where possible through the use of a partial physical map[9]. Thus, long contiguous sections of the *P. y. yoelii* genome with accompanying *P. y. yoelii* chromosomal location can be assigned to each *P. falciparum* chromosome (see Fig. 1 in ref. 30). The degree of conservation of gene order between the species was examined using ordered and orientated syntenic groups and Position Effect software. Of 4,300 *P. y. yoelii* genes within the syntenic groups, 3,145 (73%) were found to match a region of *P. falciparum* in conserved order.

One section of the syntenic map between *P. falciparum* and *P. y. yoelii* in particular—associated with *P. falciparum* chromosomes 4 and 10 and *P. y. yoelii* chromosome 5—provides a detailed snapshot of synteny between the species. Chromosome 5 of *P. y. yoelii* has received particular attention owing to the localization of a number of sexual-stage-specific genes to it[41], and because truncated versions of the chromosome are found in lines of the rodent malaria parasite *P. berghei*, which is defective in gametocytogenesis[42]. Genomic resources available for *P. berghei* chromosome 5 include chromosome markers and long-range restriction maps[41]. Exploiting the high level of synteny of rodent malaria parasite chromosomes[9], these tools were applied in combination with further mapping studies to close the syntenic map of chromosome 5 of *P. y. yoelii* (Fig. 2).

Approximately 0.8 Mb of *P. y. yoelii* chromosome 5 (estimated total length of 1.5 Mb) could be linked into one group that is syntenic to *P. falciparum* chromosome 10 and *P. falciparum* chromosome 4. From a total of 243 genes predicted in the syntenic region of *P. falciparum* chromosome 10, and 34 genes predicted in the syntenic region of chromosome 4, 171 (70%) and 22 (65%) of these, respectively, have homologues along *P. y. yoelii* chromosome 5 that appear in the same order. Pairs of homologous genes that map to regions of conserved synteny between *P. y. yoelii* and *P. falciparum* are probably orthologues, confirmed by the finding that most of these homologous pairs are also reciprocal best matches between the *P. falciparum* and *P. y. yoelii* proteins. Genes in the synteny gap on chromosome 10 (Fig. 2) include a glutamate-rich protein, S antigen, MSP3, MSP6 and liver stage antigen 1, several of which are prime vaccine antigen candidates in *P. falciparum*. Genes in the synteny gap on chromosome 4 include four *var* and two *rif* genes, which make up one of the four internal clusters of *var/rif* genes found in *P. falciparum* (see ref. 30). A series of uncharacterized hypothetical genes occur on the contigs that overlap these regions in *P. y. yoelii*.

An intriguing finding from the study of chromosome 5 has been the analysis of the syntenic break point between *P. falciparum* chromosomes 4 and 10. The final *P. y. yoelii* contig in the tiling path with significant synteny to *P. falciparum* chromosome 10 also contains the external transcribed sequence (ETS) of the SSU rRNA C unit. The synteny resumes on *P. falciparum* chromosome 4 in a *P. y. yoelii* contig that also contains the ETS of the large subunit (LSU) of the same rRNA unit. (No rRNA unit sequences are located on *P. falciparum* chromosomes 4 and 10; matches to contigs containing these genes occur in coding regions of other genes.) Both *P. y. yoelii* contigs are linked to each other through a third contig that contains the remaining elements (SSU, 5.8S, LSU, and internal transcribed sequences 1 and 2) of the complete rRNA unit (Fig. 2). Thus it seems that the break in synteny between *Plasmo-*

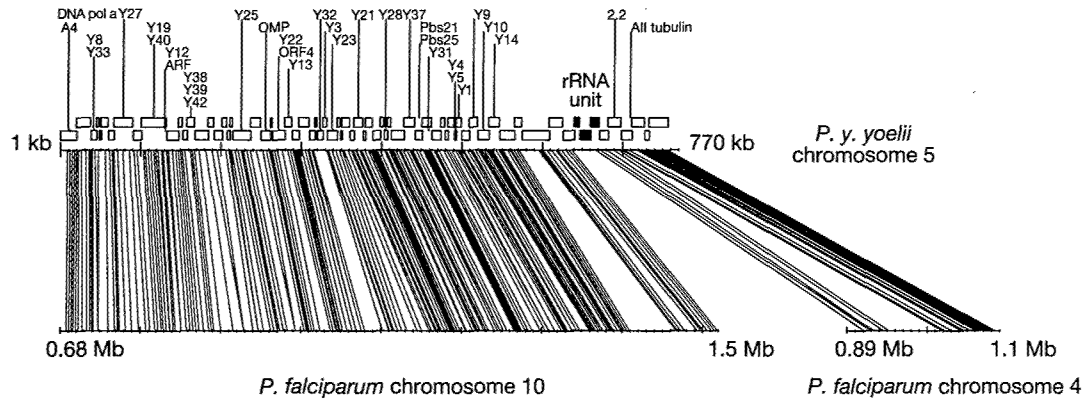


Figure 2 Conservation of gene synteny between *P. y. yoelii* chromosome 5 and *P. falciparum* chromosomes 4 and 10. Physical marker data used to confirm contig order in the tiling path of *P. y. yoelii* chromosome 5 are shown above the contigs (open boxes). Each coloured line represents a pair of orthologous genes present in the two species shown anchored to its respective location in the two genomes. Contigs containing the *P. y. yoelii* rRNA unit are shown as filled boxes.

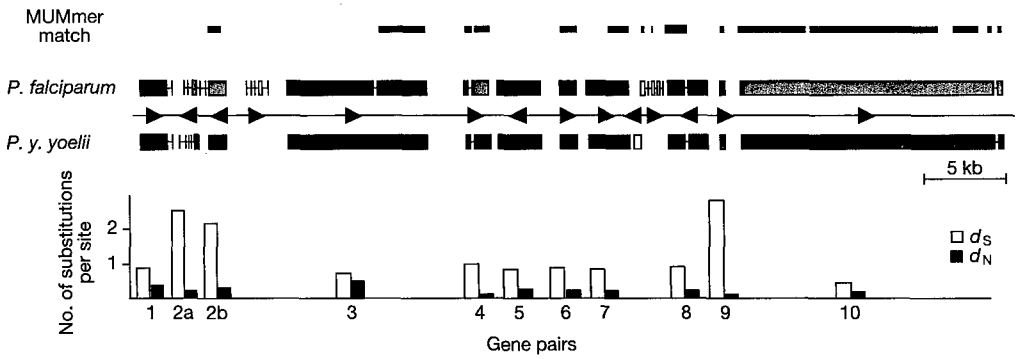


Figure 3 Global alignment scheme of a syntenic region between *P. falciparum* and *P. y. yoelii* encompassing ten orthologous gene pairs and nine intergenic regions. White boxes represent genes that have no orthologue and were excluded from analysis; green boxes represent gene models that were refined; red boxes represent unaltered gene models; arrowheads represent gene orientation on the DNA molecule. Clusters of MUMmer matches between the two species are represented as thick blue lines. For the ten orthologous gene pairs, synonymous mutations per synonymous site (d_S, open bars) and non-synonymous mutations per non-synonymous site (d_N, filled bars) were estimated and plotted.

dium chromosomes has occurred within a single rRNA unit, a phenomenon first reported in prokaryotes[43]. Six rRNA units reside as individual operons on *P. falciparum* chromosomes 1, 5, 7, 8, 11 and 13 respectively (ref. 30), in contrast to rodent malaria species that have four[44]. Intriguingly, breaks in the synteny between *P. y. yoelii* and *P. falciparum* can be mapped to almost all rRNA unit loci on the *P. falciparum* chromosomes (see Fig. 1 of ref. 30). A full analysis of this potential phenomenon is outside the scope of this study, but these results provide preliminary evidence for one possible mechanism underlying synteny breakage that may have occurred during evolution of the *Plasmodium* genus—that of chromosome breakage and recombination at sites of rRNA units.

Comparative alignment of syntenic regions

Recent comparative studies have revealed that the fine detail of short stretches of the rodent and human malaria parasite genomes is remarkably conserved[45], and that such comparisons are useful for gene prediction and evolutionary studies. Accordingly, we used a comparison of the longest assembly of *P. y. yoelii* (MALPY00395, 51.3 kb) and its syntenic region in *P. falciparum* (chromosome 7, at coordinates 1,131–1,183 kb) as a case study for a preliminary evolutionary analysis of the two genomes. Gene prediction programs run against these two regions identified 11 genes in the syntenic region of both species (Fig. 3), eight of which are orthologous gene pairs (genes 1, 3–8 and 10). The structures of two additional gene pairs (genes 2a/b and 9) were refined through manual curation of erroneous gene boundaries. Three hypothetical genes, two in *P. falciparum* and one in *P. y. yoelii*, had no discernible orthologue in the other species; the presence of multiple stop codons in these areas suggests that the genes may have become pseudogenes. A global alignment at the DNA level of the syntenic region (Fig. 3) reveals the similarity between species in intergenic regions to be almost negligible, as mirrored in similar syntenic comparisons of mouse and human[46,47]. Moreover, the mutation saturation observed in intergenic regions suggests that 'phylogenetic footprinting' can be used to identify conserved motifs between species that may be involved in gene regulation.

In contrast to intergenic regions, the similarity between species in coding regions is relatively high. The average number of non-synonymous substitutions per non-synonymous site, d_N, between the two species is 26% (±12%). Synonymous sites, d_S, are saturated (average $d_S > 1$), which supports the lack of similarity observed within intergenic regions. These values are considerably higher than those reported for human–rodent comparisons, which are approximately 7.5% and 45% for non-synonymous and synonymous substitutions, respectively[48]. The cause of such apparent disparities remains unknown, but may be a consequence of extreme genome composition or the short generation time of the parasite.

Rodent malaria species as models for *P. falciparum* biology

The usefulness of rodent malaria species as models for the study of *P. falciparum* is controversial. It is apparent that rodent models are the first port of call when preliminary *in vivo* evidence of antimalarial drug efficacy, immune response to vaccine candidates, and life-cycle adaptations in the face of drug or vaccine challenge are required. Different species of malaria parasite have developed different mechanisms of resistance to the antimalarial drug chloroquine, despite a similar mode of action of the drug (reviewed in ref. 49). It seems that mechanisms developed by the parasite to evade an inhospitable environment, whether caused by antimalarial drugs or the host immune system, may differ widely from species to species. A model involving evolution of different genes in *Plasmodium* species as a response to different host environments is consistent with the comparison of the *P. falciparum* and *P. y. yoelii* genomes presented here; conservation of synteny between the two species is high in regions of housekeeping genes, but not in regions where genes involved in antigenic variation and evasion of the host immune system are located. On the one hand, this can be interpreted as a blow to the systematic identification of all orthologues of antigen genes between *P. falciparum* and *P. y. yoelii* that could be used in the design of a malaria vaccine. On the other hand, a picture is emerging of selecting a model malaria species based on the complement of genes that best fit the phenotypic trait under study. Thus the presence of homologues of the *yir* family may make *P. y. yoelii* an attractive model for studying antigenic variation in *P. vivax*. Furthermore, identification of orthologues in the genomes of relatively distant rodent and human malaria parasites will facilitate finding orthologues in other model malaria species, for example monkey models of malaria such as *Plasmodium knowlesi*. □

Methods

Genome and EST sequencing

Plasmodium yoelii yoelii 17XNL line[50], selected from an isolate taken from the blood of a wild-caught thicket rat in the Central African Republic[51], is a non-lethal strain with a preference for development in reticulocytes. Clone 1.1 was obtained through serial dilution of sporozoites. Parasites were grown in laboratory mice no more than three blood passages from mosquito passage to limit chromosome instability, collected by exsanguination into heparin, and host mouse leukocytes were removed by filtration. Small insert libraries (average insert size 1.6 kb) were constructed in pUC-derived vectors after nebulization of genomic DNA. DNA sequencing of plasmid ends used ABI Big Dye terminator chemistry on ABI3700 sequencing machines. A total of 222,716 sequences (82% success rate), averaging 662 nucleotides in length, were assembled using TIGR Assembler[7]. BLASTN of the *P. y. yoelii* contigs and singletons against the complete set of

Celera mouse contigs[8], using a cutoff of 90% identity over 100 nucleotides, identified contaminating mouse sequences that were subsequently removed. Contigs were assigned to groups using Grouper[52]. Each contig was assigned an identifier in the format 'MALPY00001'.

Proteomic analysis

MudPIT technology and methods were as described in ref. 23. Sporozoites of *P. y. yoelii* were dissected from infected *Anopheles stephensi* mosquito salivary glands, and *P. y. yoelii* gametocytes were prepared as described[53]. Cellular debris from uninfected mosquitoes and mouse erythrocytes were analysed as controls. Tandem mass spectrometry (MS/MS) data sets were searched against several databases: the complete set of *P. y. yoelii* full and partial proteins (7,860 total); 791,324 *P. y. yoelii* open reading frames (stop-to-stop ORFs over 15 amino acids and start-to-stop ORFs over 100 amino acids); 57,885 ORFs from NCBI's RefSeq for human, mouse and rat; 15,570 *Anopheles*, *Aedes* and *Drosophila melanogaster* proteins from GenBank; and 165 common protein contaminants (for example, trypsin, bovine serum albumin).

Gene finding and annotation

The splice site recognition module of GlimmerMExon was trained specifically for *P. yoelii* genome data, using DNA sequences extracted from a set of 1,166 donor and 1,166 acceptor sites confirmed by *P. y. yoelii* ESTs. Phat and the exon recognition module of GlimmerMExon were trained on *P. falciparum* data as described (see ref. 54). Combiner was used to generate a final ranked list of *P. y. yoelii* gene models, and TIGR's Eukaryotic Genome Control suite of programs was used for automated annotation of these (both described in ref. 54). Automated gene names were assigned to proteins by taking the 'equivalogue' name of the hidden Markov model (HMM) associated with the protein where possible, or where no HMM was assigned, on the basis of the best-paired alignment. Each protein was assigned an identifier in the format 'PY00001'.

Paralogous gene families

Proteins encoded by multigene families were identified by a domain-based clustering algorithm developed at TIGR. Families were regarded as potentially *Plasmodium*- or *yoelii*-specific if they were not described by any Pfam[55] or TIGRFAM[56] domains and if the automatic annotation process had not ascribed names corresponding to widely distributed proteins. HMMs for these families were built using the HMMER package version 2.1.1 (ref. 57). Newly constructed models were then used to search the *P. yoelii*, *P. falciparum* and GenBank databases to define the scope of the families.

Telomeric/subtelomeric repeat analysis

Subtelomeric contigs were identified through alignment using MUMmer2 (ref. 40) with a minimum exact match ranging from 30–40 bases. Tandem Repeat Finder[58] used the following settings: match = 2, mismatch = 7, PM (match probability) = 75, PI (indel probability) = 10, minscore = 400, max period = 700.

Comparative analyses

Gene model predictions in the syntenic region of *P. falciparum* chromosome 7 were inspected manually, and bi-directional best hits between gene models that respected conserved syntenies were selected. A global alignment of the two sequences was calculated using Owen[59], and nucleotide sequences of predicted gene models were aligned using CLUSTALW[60] with default parameters, and refined manually. The number of substitutions per synonymous (d_S) and nonsynonymous (d_N) sites were estimated using the Nei and Gojobori method[61]. Conservation of gene order was established using Position Effect (http://www.tigr.org/software), where matches between *P. falciparum* and *P. y. yoelii* genes were calculated using BLASTP with a cutoff E value of 10^{-15}. The query and hit gene from each match were defined as anchor points in gene sets composed of adjacent genes. Up to ten genes upstream and downstream from each anchor gene were used in creating the gene set. An optimal alignment was calculated between the ordered gene sets using BLASTP per cent similarity scores and a linear gap penalty. Low-scoring alignments with a cumulative per cent similarity less than 100 were not used. Each optimal alignment provided a list of matching genes in conserved order between *P. falciparum* and *P. y. yoelii*.

Received 31 July; accepted 30 August 2002; doi:10.1038/nature01099.

1. Carter, R. & Diggs, C. L. *Parasitic Protozoa* 359–465 (Academic, New York/San Francisco/London, 1977).
2. van Dijk, M. R., Waters, A. P. & Janse, C. J. Stable transfection of malaria parasite blood stages. *Science* **268,** 1358–1362 (1995).
3. Carlton, J. M. & Carucci, D. J. Rodent models of malaria in the genomics era. *Trends Parasitol.* **18,** 100–102 (2002).
4. Myers, E. W. *et al.* A whole-genome assembly of *Drosophila*. *Science* **287,** 2196–2204 (2000).
5. Lander, E. S. & Waterman, M. S. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2,** 231–239 (1988).
6. McCutchan, T. F., Dame, J. B., Miller, L. H. & Barnwell, J. Evolutionary relatedness of *Plasmodium* species as determined by the structure of DNA. *Science* **225,** 808–811 (1984).
7. Sutton, G. G., White, O., Adams, M. D. & Kervalage, A. R. TIGR Assembler: a new tool for assembling large shotgun sequencing projects. *Genome Sci. Technol.* **1,** 9–19 (1995).
8. Mural, R. J. *et al.* A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science* **296,** 1661–1671 (2002).
9. Janse, C. J., Carlton, J. M.-R., Walliker, D. & Waters, A. P. Conserved location of genes on polymorphic chromosomes of four species of malaria parasites. *Mol. Biochem. Parasitol.* **68,** 285–296 (1994).
10. Daly, T. M., Long, C. A. & Bergman, L. W. Interaction between two domains of the *P. yoelii* MSP-1 protein detected using the yeast two-hybrid system. *Mol. Biochem. Parasitol.* **117,** 27–35 (2001).
11. Quackenbush, J. *et al.* The TIGR gene indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Res.* **29,** 159–164 (2001).
12. Cawley, S. E., Wirth, A. I. & Speed, T. P. Phat—a gene finding program for *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **118,** 167–174 (2001).
13. Salzberg, S. L., Pertea, M., Delcher, A. L., Gardner, M. J. & Tettelin, H. Interpolated Markov models for eukaryotic gene finding. *Genomics* **59,** 24–31 (1999).
14. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
15. Thompson, J., Janse, C. J. & Waters, A. P. Comparative genomics in *Plasmodium*: a tool for the identification of genes and functional analysis. *Mol. Biochem. Parasitol.* **118,** 147–154 (2001).
16. Jomaa, H. *et al.* Inhibitors of the nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs. *Science* **285,** 1573–1576 (1999).
17. 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).
18. Ohkanda, J. *et al.* Peptidomimetic inhibitors of protein farnesyltransferase show potent antimalarial activity. *Bioorg. Med. Chem. Lett.* **11,** 761–764 (2001).
19. The Gene Ontology Consortium. Creating the gene ontology resource: design and implementation. *Genome Res.* **11,** 1425–1433 (2001).
20. Cooke, B. M., Mohandas, N. & Coppel, R. L. The malaria-infected red blood cell: structural and functional changes. *Adv. Parasitol.* **50,** 1–86 (2001).
21. del Portillo, H. A. *et al.* A superfamily of variant genes encoded in the subtelomeric region of *Plasmodium vivax*. *Nature* **410,** 839–842 (2001).
22. Janssen, C. S., Barrett, M. P., Turner, C. M. & Phillips, R. S. A large gene family for putative variant antigens shared by human and rodent malaria parasites. *Proc. R. Soc. Lond. B* **269,** 431–436 (2002).
23. Florens, L. *et al.* A proteomic view of the *Plasmodium falciparum* life cycle. *Nature* **419,** 520–526 (2002).
24. Preiser, P. R., Jarra, W., Capiod, T. & Snounou, G. A rhoptry-protein-associated mechanism of clonal phenotypic variation in rodent malaria. *Nature* **398,** 618–622 (1999).
25. Galinski, M. R., Xu, M. & Barnwell, J. W. *Plasmodium vivax* reticulocyte binding protein-2 (PvRBP-2) shares structural features with PvRBP-1 and the *Plasmodium yoelii* 235 kDa rhoptry protein family. *Mol. Biochem. Parasitol.* **108,** 257–262 (2000).
26. Rayner, J. C., Galinski, M. R., Ingravallo, P. & Barnwell, J. W. Two *Plasmodium falciparum* genes express merozoite proteins that are related to *Plasmodium vivax* and *Plasmodium yoelii* adhesive proteins involved in host cell selection and invasion. *Proc. Natl Acad. Sci. USA* **97,** 9648–9653 (2000).
27. Wiser, M. F., Giraldo, L. E., Schmitt-Wrede, H. P. & Wunderlich, F. *Plasmodium chabaudi*: immunogenicity of a highly antigenic glutamate-rich protein. *Exp. Parasitol.* **85,** 43–54 (1997).
28. Spielmann, T. & Beck, H. P. Analysis of stage-specific transcription in *Plasmodium falciparum* reveals a set of genes exclusively transcribed in ring stage parasites. *Mol. Biochem. Parasitol.* **111,** 453–458 (2000).
29. Favaloro, J. M., Culvenor, J. G., Anders, R. F. & Kemp, D. J. A *Plasmodium chabaudi* antigen located in the parasitophorous vacuole membrane. *Mol. Biochem. Parasitol.* **62,** 263–270 (1993).
30. Gardner, M. J. *et al.* Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419,** 498–511 (2002).
31. Scherf, A., Figueiredo, L. M. & Freitas-Junior, L. H. *Plasmodium* telomeres: a pathogen's perspective. *Curr. Opin. Microbiol.* **4,** 409–414 (2001).
32. Mu, J. *et al.* Chromosome-wide SNPs reveal an ancient origin for *Plasmodium falciparum*. *Nature* **418,** 323–326 (2002).
33. Garnham, P. C. C. *Malaria Parasites and Other Haemosporidia* (Blackwell Scientific, Oxford, 1966).
34. Escalante, A. A. & Ayala, F. J. Phylogeny of the malarial genus *Plasmodium*, derived from rRNA gene sequences. *Proc. Natl Acad. Sci. USA* **91,** 11373–11377 (1994).
35. Waters, A. P., Higgins, D. G. & McCutchan, T. F. *Plasmodium falciparum* appears to have arisen as a result of lateral transfer between avian and human hosts. *Proc. Natl Acad. Sci. USA* **88,** 3140–3144 (1991).
36. Tchavtchitch, M., Fischer, K., Huestis, R. & Saul, A. The sequence of a 200 kb portion of a *Plasmodium vivax* chromosome reveals a high degree of conservation with *Plasmodium falciparum* chromosome 3. *Mol. Biochem. Parasitol.* **118,** 211–222 (2001).
37. Carlton, J. M.-R., Galinski, M. R., Barnwell, J. W. & Dame, J. B. Karyotype and synteny among the chromosomes of all four species of human malaria parasite. *Mol. Biochem. Parasitol.* **101,** 23–32 (1999).
38. Janse, C. J. Chromosome size polymorphism and DNA rearrangements in *Plasmodium*. *Parasitol. Today* **9,** 19–22 (1993).
39. Carlton, J. M. R., Vinkenoog, R., Waters, A. P. & Walliker, D. Gene synteny in species of *Plasmodium*. *Mol. Biochem. Parasitol.* **93,** 285–294 (1998).
40. 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).
41. van Lin, L. H. M., Pace, T., Janse, C. J., Scotti, R. & Ponzi, R. A long range restriction map of chromosomes 5 of *Plasmodium berghei* demonstates a chromosomes specific symmetrical subtelomeric organisation. *Mol. Biochem. Parasitol.* **86,** 111–115 (1997).
42. Janse, C. J., Ramesar, J., van den Berg, F. M. & Mons, B. *Plasmodium berghei*: *in vivo* generation and selection of karyotype mutants and non-gametocyte producer mutants. *Exp. Parasitol.* **74,** 1–10 (1992).
43. Liu, S. L. & Sanderson, K. E. Rearrangements in the genome of the bacterium *Salmonella typhi*. *Proc. Natl Acad. Sci. USA* **92,** 1018–1022 (1995).
44. Dame, J. B. & McCutchan, T. F. The four ribosomal DNA units of the malaria parasite *Plasmodium berghei*. Identification, restriction map and copy number analysis. *J. Biol. Chem.* **258,** 6984–6990 (1983).
45. van Lin, L. H. *et al.* Interspecies conservation of gene order and intron-exon structure in a genomic locus of high gene density and complexity in *Plasmodium*. *Nucleic Acids Res.* **29,** 2059–2068 (2001).
46. Jareborg, N., Birney, E. & Durbin, R. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Res.* **9,** 815–824 (1999).
47. Shabalina, S. A., Ogurtsov, A. Y., Kondrashov, V. A. & Kondrashov, A. S. Selective constraint in intergenic regions of human and mouse genomes. *Trends Genet.* **17,** 373–376 (2001).
48. Makalowski, W. & Boguski, M. S. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. *Proc. Natl Acad. Sci. USA* **95,** 9407–9412 (1998).
49. Carlton, J. M., Fidock, D. A., Djimde, A., Plowe, C. V. & Wellems, T. E. Conservation of a novel

vacuolar transporter in *Plasmodium* species and its central role in chloroquine resistance of *P. falciparum*. *Curr. Opin. Microbiol.* **4,** 415–420 (2001).
50. Weinbaum, F. I., Evans, C. B. & Tigelaar, R. E. An *in vitro* assay for T cell immunity to malaria in mice. *J. Immunol.* **116,** 1280–1283 (1976).
51. Landau, I. & Chabaud, A. G. Natural infection by 2 plasmodia of the rodent *Thamnomys rutilans* in the Central African Republic. *C.R. Acad. Sci. Hebd. Seances Acad. Sci. D* **261,** 230–232 (1965).
52. Gardner, M. J. *et al.* Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science* **282,** 1126–1132 (1998).
53. Beetsma, A. L., van de Wiel, T. J., Sauerwein, R. W. & Eling, W. M. *Plasmodium berghei* ANKA: purification of large numbers of infectious gametocytes. *Exp. Parasitol.* **88,** 69–72 (1998).
54. Gardner, M. J. *et al.* Sequence of *Plasmodium falciparum* chromosomes 2, 10, **11** and 14. *Nature* **419,** 531–534 (2002).
55. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **30,** 276–280 (2002).
56. Haft, D. H. *et al.* TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Res.* **29,** 41–43 (2001).
57. Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14,** 755–763 (1998).
58. Benson, G. Tandem repeats finder: a program to analyse DNA sequences. *Nucleic Acids Res.* **27,** 573–580 (1999).
59. Ogurtsov, A. Y., Roytberg, M. A., Shabalina, S. A. & Kondrashov, A. S. OWEN: aligning long collinear regions of genomes. *Bioinformatics* (in the press).
60. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22,** 4673–4680 (1994).
61. Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3,** 418–426 (1986).
62. Black, C. G., Wang, L., Hibbs, A. R., Werner, E. & Coppel, R. L. Identification of the *Plasmodium chabaudi* homologue of merozoite surface proteins 4 and 5 of *Plasmodium falciparum*. *Infect. Immun.* **67,** 2075–2081 (1999).

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

Acknowledgements

We thank S. Cawley and T. Pace for collaborative work; J. Mendoza and J. Ramesar for technical support; C. Long for the gift of a *P. y. yoelii* cDNA library; R. Arcilla and W. Weiss for parasite material; and J. Eisen and S. Sullivan for critical reading of the manuscript. L.H.v.L. was supported by an INCO-DEV programme grant from the European Community; T.W.K. was supported by a Rijks Universiteit te Leiden studentship; J.D.R. was supported with funds from the Wellcome Trust. This project was funded by the US Department of Defense through cooperative agreement with the US Army Medical Research and Material Command and by the Naval Medical Research Center. The opinions expressed are those of the authors and do not reflect the official policy of the Department of the Navy, Department of Defense, or the US government.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.M.C. (e-mail: carlton@tigr.org). Access to genome annotation data is available through the TIGR Eukaryotic Projects website (http://www.tigr.org) and PlasmoDB (http://www.plasmodb.org). This whole-genome shotgun project has been deposited at DDBJ/EMBL/GenBank under the project accession number AABL00000000. The version described in this paper is the first version, AABL01000000.

A proteomic view of the *Plasmodium falciparum* life cycle

Laurence Florens*, Michael P. Washburn†, J. Dale Raine‡, Robert M. Anthony§, Munira Grainger‖, J. David Haynes§¶, J. Kathleen Moch§, Nemone Muster*, John B. Sacci§#, David L. Tabb*☆, Adam A. Witney§#, Dirk Wolters†#, Yimin Wu, Malcolm J. Gardner††, Anthony A. Holder‖, Robert E. Sinden‡, John R. Yates*† & Daniel J. Carucci§**

** Department of Cell Biology, The Scripps Research Institute, SR-11, 10550 North Torrey Pines Road, La Jolla, California 92037, USA*
† Department of Proteomics and Metabolomics, Torrey Mesa Research Institute, Syngenta Research & Technology, 3115 Merryfield Row, San Diego, California 92121-1125, USA
‡ Infection and Immunity Section, Department of Biological Sciences, Imperial College of Science, Technology & Medicine, Sir Alexander Fleming Building, South Kensington, London SW7 2AZ, UK
§ Naval Medical Research Center, Malaria Program (IDD), 503 Robert Grant Avenue, Room 3A40; and ¶ Department of Immunology, Walter Reed Army Institute of Research, Silver Spring, Maryland 20910-7500, USA
‖ The Division of Parasitology, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK
☆ Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA
*** Malaria Research and Reference Reagent Resource Center, American Type Culture Collection, 10801 University Boulevard, Manassas, Virginia 20110-2209, USA*
†† The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850, USA

The completion of the *Plasmodium falciparum* clone 3D7 genome provides a basis on which to conduct comparative proteomics studies of this human pathogen. Here, we applied a high-throughput proteomics approach to identify new potential drug and vaccine targets and to better understand the biology of this complex protozoan parasite. We characterized four stages of the parasite life cycle (sporozoites, merozoites, trophozoites and gametocytes) by multidimensional protein identification technology. Functional profiling of over 2,400 proteins agreed with the physiology of each stage. Unexpectedly, the antigenically variant proteins of *var* and *rif* genes, defined as molecules on the surface of infected erythrocytes, were also largely expressed in sporozoites. The detection of chromosomal clusters encoding co-expressed proteins suggested a potential mechanism for controlling gene expression.

The life cycle of *Plasmodium* is extraordinarily complex, requiring specialized protein expression for life in both invertebrate and vertebrate host environments, for intracellular and extracellular survival, for invasion of multiple cell types, and for evasion of host immune responses. Interventional strategies including anti-malarial vaccines and drugs will be most effective if targeted at specific parasite life stages and/or specific proteins expressed at these stages. The genomes of *P. falciparum*[1] and *P. yoelii yoelii*[2] are now completed and offer the promise of identifying new and effective drug and vaccine targets.

Functional genomics has fundamentally changed the traditional gene-by-gene approach of the pre-genomic era by capitalizing on the success of genome sequencing efforts. DNA microarrays have been successfully used to study differential gene expression in the abundant blood stages of the *Plasmodium* parasite[3,4]. However, transcriptional analysis by DNA microarrays generally requires microgram quantities of RNA and has been restricted to stages that can be cultivated *in vitro*, limiting current large-scale gene expression analyses to the blood stages of *P. falciparum*. As several key stages of the parasite life cycle, in particular the pre-erythrocytic stages, are not readily accessible to study, and as differential gene expression is in fact a surrogate for protein expression, global proteomic analyses offer a unique means of determining not only protein expression, but also subcellular localization and post-translational modifications.

We report here a comprehensive view of the protein complements isolated from sporozoites (the infectious form injected by the mosquito), merozoites (the invasive stage of the erythrocytes), trophozoites (the form multiplying in erythrocytes), and gametocytes (sexual stages) of the human malaria parasite *P. falciparum*. These proteomes were analysed by multidimensional protein identification technology (MudPIT), which combines in-line, high-resolution liquid chromatography and tandem mass spectrometry[5]. Two levels of control were implemented to differentiate parasite from host proteins. By using combined host–parasite sequence databases and noninfected controls, 2,415 parasite proteins were confidently identified out of thousands of host proteins; that is, 46% of all gene products were detected in four stages of the *Plasmodium* life cycle (Supplementary Table 1).

Comparative proteomics throughout the life cycle

The sporozoite proteome appeared markedly different from the other stages (Table 1). Almost half (49%) of the sporozoite proteins

Table 1 **Comparative summary of the protein lists for each stage**

Protein count	Sporozoites	Merozoites	Trophozoites	Gametocytes
152	X	X	X	X
197	–	X	X	X
53	X	–	X	X
28	X	X	–	X
36	X	X	X	–
148	–	–	X	X
73	–	X	–	X
120	X	–	–	X
84	–	X	X	–
80	X	–	X	–
65	X	X	–	–
376	–	–	–	X
286	–	–	X	–
204	–	X	–	–
513	X	–	–	–
2,415	1,049	839	1,036	1,147

Whole-cell protein lysates were obtained from, on average, 17×10^6 sporozoites, 4.5×10^9 trophozoites, 2.75×10^9 merozoites, and 6.5×10^9 gametocytes.

Present addresses: BRB 13-009, Department of Microbiology and Immunology, University of Maryland School of Medicine, 655 W. Baltimore St., Baltimore, Maryland 21201, USA (J.B.S.); Department of Medical Microbiology, St George's Hospital Medical School, Cranmer Terrace, London SW17 0RE, UK (A.A.W.); and Ruhr-University Bochum, Institute of Analytical Chemistry, 44780 Bochum, Germany (D.W.).

were unique to this stage, which shared an average of 25% of its proteins with any other stage. On the other hand, trophozoites, merozoites and gametocytes had between 20% and 33% unique proteins, and they shared between 39% and 56% of their proteins. Consequently, only 152 proteins (6%) were common to all four stages. Those common proteins were mostly housekeeping proteins such as ribosomal proteins, transcription factors, histones and cytoskeletal proteins (Supplementary Table 1). Proteins were sorted into main functional classes based on the Munich Information Centre for Protein Sequences (MIPS) catalogue[6], with some adaptations for classes specific to the parasite, such as cell surface and apical organelle proteins (Fig. 1). When considering the annotated proteins in the database, some marked differences appeared between sporozoites and blood stages (Fig. 1). Although great care was taken to ensure that the results reflect the state of the parasite in the host, a portion of the data set may reflect the parasite's response to different purification treatments. However, the stage-specific detection of known protein markers at each stage established the relevance of our data set.

The merozoite proteome

Merozoites are released from an infected erythrocyte, and after a short period in the plasma, bind to and invade new erythrocytes. Proteins on the surface and in the apical organelles of the merozoite mediate cell recognition and invasion in an active process involving an actin-myosin motor. Four putative components of the invasion motor[7], merozoite cap protein-1 (MCP1), actin, myosin A, and myosin A tail domain interacting protein (MTIP), were abundant merozoite proteins (Supplementary Table 2). Abundant merozoite surface proteins (MSPs) such as MSP1 and MSP2 are linked by a glycosylphosphatidyl (GPI) anchor to the membrane, and both have been implicated in immune evasion (reviewed in ref. 8). A second family of peripheral membrane proteins, represented by MSP3 and MSP6, was also detected (Fig. 2a), although these proteins are largely soluble proteins of the parasitophorous vacuole, which are released on schizont rupture. Other vacuolar proteins, such as the acidic basic repeat antigen (ABRA) and serine repeat antigen (SERA), were detected in the merozoite fraction, but some such as S-antigen[9] were not (Supplementary Table 2). Notably, MSP8 and a related MSP8-like protein were only identified in sporozoites (Fig. 2a). Some MSPs are diverse in sequence and may be extensively modified by proteolysis; these features, together with the association of a variety of peripheral and soluble proteins, provide for a complex surface architecture.

Many apical organellar proteins, in the micronemes and rhoptries, have a single transmembrane domain. Among these proteins, apical membrane antigen 1 (AMA1) and MAEBL were found in both sporozoite and merozoite preparations (Fig. 2a). Erythrocyte-binding antigens (EBA), such as EBA 175 and EBA 140/BAEBL, were found only in the merozoite and trophozoite fractions. Of note, the reticulocyte-binding protein (PfRH) family (PFD0110w, MAL13P1.176, PF13_01998, PFL2520w and PFD1150c), which has similarity with the Py235 family of *P. y. yoelii* rhoptry proteins and the *Plasmodium vivax* reticulocyte-binding proteins, was not detected in the merozoite fraction. Some PfRH proteins were, however, detected in sporozoites (Fig. 2a), including RH3, which is a transcribed pseudogene in blood stages[10]. Components of the low molecular mass rhoptry complex, the rhoptry-associated proteins (RAP) 1, 2 and 3, were all found in merozoites. RAP1 was also detected in sporozoites. The high molecular mass rhoptry protein complex (RhopH), together with ring-infected erythrocyte surface antigen (RESA), which is a component of dense granules, is transferred intact to new erythrocytes at or after invasion and may contribute to the host cell remodelling process. RhopH1, RhopH2 (PFI1445w; Ling, I. T., *et al.*, unpublished data) and RhopH3 were found in the merozoite proteome. RhopH1 (PFC0120w/PFC0110w) has been shown to be a member of the cyto-adherence linked asexual gene family (CLAG)[11]; however, the presence of CLAG9 in the merozoite fraction (Fig. 2a) suggests that CLAG9 may also be a RhopH protein, casting some doubt on the proposed role for this protein in cyto-adherence[12].

The trophozoite proteome

After erythrocyte invasion the parasite modifies the host cell. The principal modifications during the initial trophozoite phase (lasting about 30 h) allow the parasite to transport molecules in and out of the cell, to prepare the surface of the red blood cell to mediate cyto-adherence, and to digest the cytoplasmic contents, particularly haemoglobin, in its food vacuole. In the next phase of schizogony (the final ~18 h of the asexual development in the blood cell), nuclear division is followed by merozoite formation and release.

Knob-associated histidine-rich protein (KAHRP) and erythrocyte membrane proteins 2 and 3 (EMP2 and -3) bind to the erythrocyte cytoskeleton (Fig. 2a). Of the proteins of the parasitophorous vacuole and the tubovesicular membrane structure extending into the cytoplasm of the red blood cell, three (the skeleton-binding protein 1, and exported proteins EXP1 and EXP2) were represented by peptides (Fig. 2a); although a fourth (Sar1 homologue, small GTP-binding protein; PFD0810w) was not. It is likely that one or more of the hypothetical proteins detected only in the trophozoite sample are involved in these unusual structures.

Digestion of haemoglobin is a major parasite catabolic process[13]. Members of the plasmepsin family (aspartic proteinases; PF14_0075 to PF14_0078)[14], falcipain family (cysteine proteinases; PF11_0161,

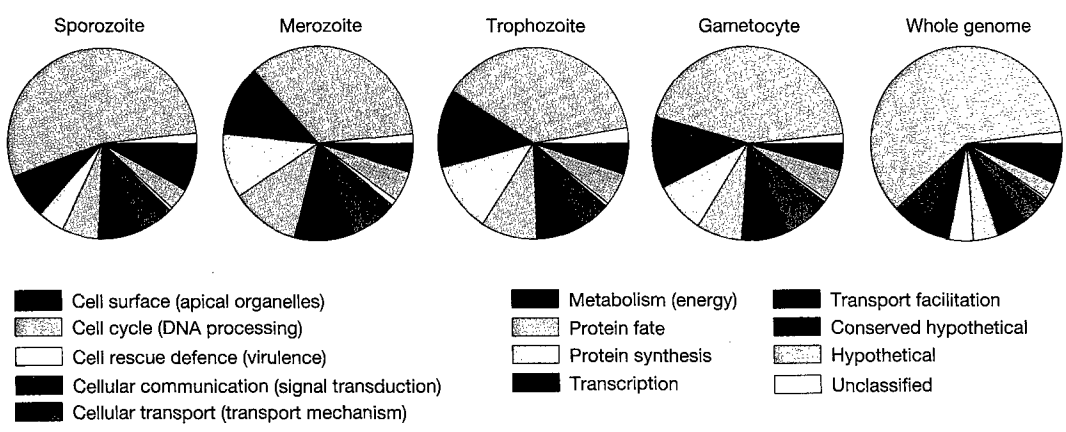


Figure 1 Functional profiles of expressed proteins. Proteins identified in each stage are plotted as a function of their broad functional classification as defined by the MIPS catalogue[6]. To avoid redundancy, only one class was assigned per protein. The complete protein list is given in Supplementary Table 1.

PF11_0162 and PF11_0165)[15], and falcilysin (a metallopeptidase; PF13_0322)[16] implicated in this process were all clearly identified (Supplementary Table 1). Several proteases expressed in the merozoite and trophozoite fractions, and not involved in haemoglobin digestion, may be important in parasite release at the end of schizogony, invasion of the new cell, or merozoite protein processing. Possible candidates for this mechanism include cysteine proteinases of the falcipain and SERA families, or subtilisins such as SUB1 and SUB2, both located in apical organelles (Fig. 2a).

The gametocyte proteome

Stage V gametocytes are dimorphic, with a male:female ratio of 1:4. They are arrested in the cell cycle until they enter the mosquito where development is induced within minutes to form the male and female gametes. Gametocyte structure reflects these ensuing fates; that is, the female has abundant ribosomes and endoplasmic reticulum/vesicular network to re-initiate translation, whereas the male is largely devoid of ribosomes and is terminally differentiated[17].

Gametocyte-specific transcription factors, RNA-binding proteins, and gametocyte-specific proteins involved in the regulation of messenger RNA processing (particularly splicing factors, RNA helicases, RNA-binding proteins, ribonucleoproteins (RNPs) and small nuclear ribonucleoprotein particles (snRNPS)) were highly represented in the gametocyte proteome (Supplementary Table 1). Transcription in the terminally differentiated gametocytes is 'suppressed', but the female gametocytes contain mRNAs encoding gamete/zygote/ookinete surface antigens (for example, P25/28)

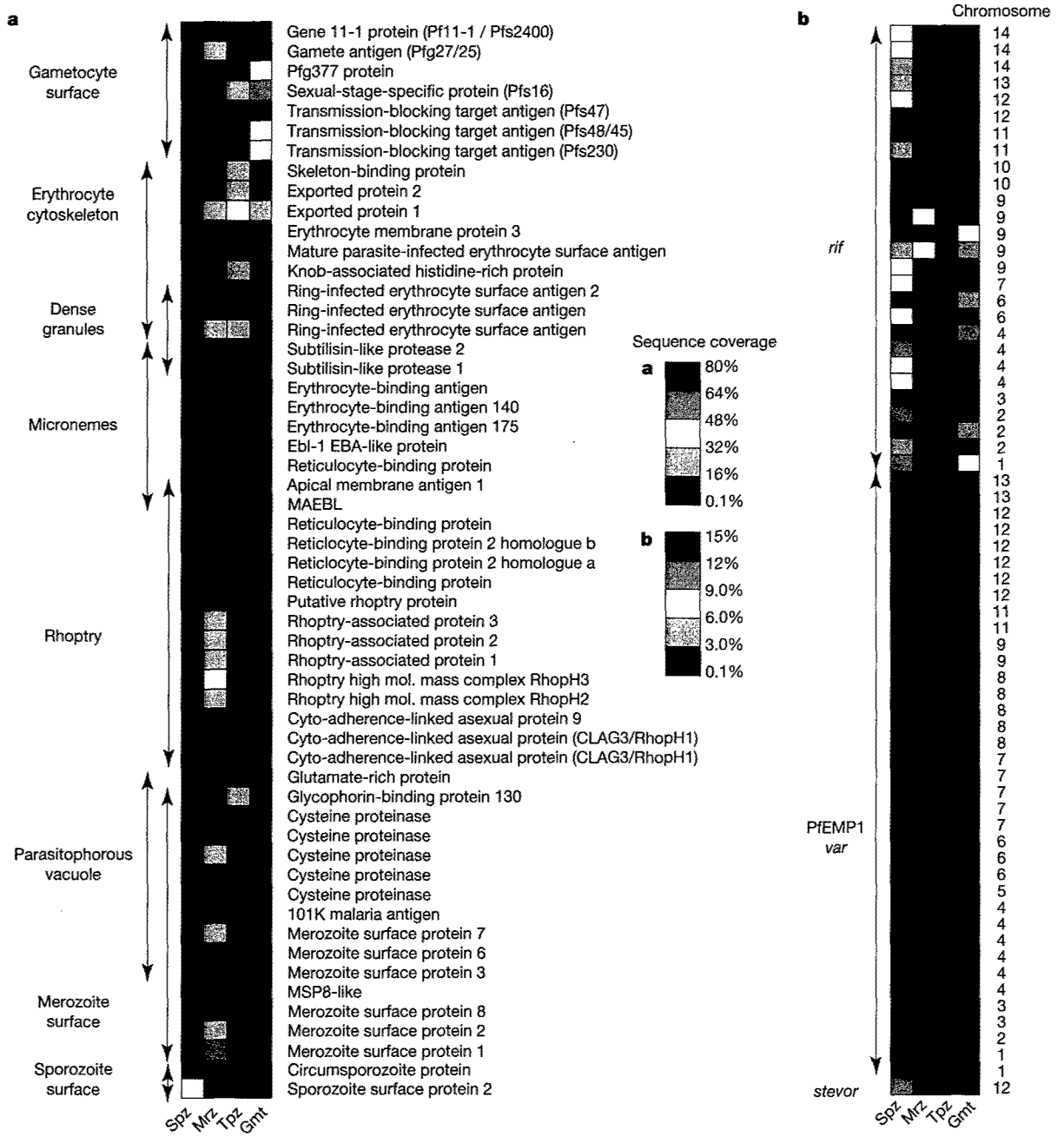


Figure 2 Expression patterns of known stage-specific proteins. **a**, Cell surface, organelle, and secreted proteins are plotted as a function of their known subcellular localization. **b**, *stevor*, *var* and *rif* polymorphic surface variants are plotted as a function of the chromosome encoding their genes. The matrices are colour-coded by sequence coverage measured in each stage (proteins not detected in a stage are represented by black squares). Locus names associated with these proteins are listed in Supplementary Table 2. Spz, sporozoite; mrz, merozoite; tpz, trophozoite; gmt, gametocyte.

that are subject to post-transcriptional control; this control is released rapidly during gamete development[17]. Ribosomal proteins were largely represented: 82% of known small subunit (SSU) proteins and 69% of known large subunit (LSU) proteins were detected in gametocytes compared to 94% and 82%, respectively, from all stages examined (Supplementary Table 1). We suggest that this reflects the accumulation of ribosomes in the female gametocyte to accommodate for the sudden increase in protein synthesis required during gametogenesis and early zygote development.

Other protein groupings highly represented in the gametocyte were in the cell cycle/DNA processing and energy classes (Fig. 1). The former is consistent with the biological observation that the mature gametocyte is arrested in G0 of the cell cycle and will require a full complement of pre-existing cell cycle regulatory cascades to respond, within seconds, to the gametogenesis stimuli (that is, xanthurenic acid and a drop in temperature)[18]. Metabolic pathways of the malaria parasite may be stage-specific, with asexual blood stage parasites dependent on glycolysis and conversion of pyruvate to lactate (L-lactate dehydrogenase) for energy. In the gametocyte and sporozoite preparations, peptides from enzymes involved in the mitochondrial tricarboxylic acid (TCA) cycle and oxidative phosphorylation were identified (Table 2). This observation suggests that gametocytes have fully functional mitochondria as a pre-adaptation to life in the mosquito, as suggested by morphological and biochemical studies[19] and their sensitivity to anti-malarials attacking respiration (primaquine and artimesinin-based products)[17]. It will be interesting to observe whether other mosquito and liver stages, which show similar drug sensitivities, express the same metabolic proteome.

Cell surface proteins (Fig. 1) included most of the known surface antigens (Fig. 2a and Supplementary Table 2). However, Pfs35 and a sexual stage-specific kinase (PF13_0258) were not detected. Nevertheless the cultured gametocytes analysed in this study expressed a specific repertoire of rifin and PfEMP1 proteins (Fig. 2b and Supplementary Table 2). Together these observations suggest that the gametocyte, which is very long-lived in the red blood cell (that is, 9–12 days compared with 2 days for the pathogenic asexual parasites), expresses a limited repertoire of the highly polymorphic families of surface antigens so widely represented in the asexual parasites.

The sporozoite proteome

Sporozoites are injected by the mosquito during ingestion of a blood meal. Although, they are in the blood stream for only minutes, sporozoites probably require mechanisms to evade the host humoral immune system in order for at least a fraction of the thousands of sporozoites injected by the mosquito to survive the hostile environment in the blood and successfully invade hepatocytes.

The main class of annotated sporozoite proteins identified was cell surface and organelle proteins (Fig. 1). Sporozoites are an invasive stage and possess the apical complex machinery involved in host cell invasion. As observed in the analysis of the *P. y. yoelii* sporozoite transcriptome[20], actin and myosin were found in the motile sporozoites (Supplementary Table 2). Many proteins associated with rhoptry, micronemes and dense granules were detected (Fig. 2a). Among the proteins found were known markers of the sporozoite stage, such as the circumsporozoite protein (CSP) and sporozoite surface protein 2 (SSP2; also known as TRAP), both present in large quantities at the sporozoite surface (Fig. 2a). Peptides derived from CTRP (circumsporozoite protein and thrombospondin-related adhesive protein (TRAP)-related protein), an ookinete cell surface protein involved in recognition and/or motility[21], were detected in the sporozoite fractions (Supplementary Table 1).

Most surprisingly, peptides derived from multiple *var* (coding for PfEMP1) and *rif* genes were identified in the sporozoite samples. PfEMP1 and rifins are coded for by large multigene families (*var* and *rif*)[22,23] and are present on the surface of the infected red blood cell. No peptides derived from *rif* genes were identified in the trophozoite sample, whereas sporozoites expressed 21 different rifins and 25 PfEMP1 isoforms (Fig. 2b); that is, a total of 14% of the *rif* genes and 33% of the *var* genes encoded by the genome. Furthermore, very little overlap was observed between stages: only ten PfEMP1 and two rifin isoforms expressed in sporozoites were found in other stages. Whereas in the blood stream the asexual stage parasites undergo asexual multiplication and therefore have an opportunity to undergo antigenic 'switching' of the variant antigen genes, the non-replicative sporozoites may not have this opportunity. Expressing such a polymorphic array of *var* (PfEMP1) and *rif* genes could be part of a sporozoite survival mechanism.

Chromosomal clusters encoding co-expressed proteins

The distinct proteomes of each stage of the *Plasmodium* life cycle suggested that there is a highly coordinated expression of *Plasmodium* genes involved in common processes. Co-expression groups are a widespread phenomenon in eukaryotes, where mRNA array analyses have been used to establish gene expression profiles. Analysis of co-regulated gene groups facilitates both searching for regulatory motifs common to co-regulated genes, and predicting protein function on the basis of the 'guilt by association' model. Furthermore, mRNA analyses in *Saccharomyces cerevisiae*[24] and *Homo sapiens*[25,26] have demonstrated that co-regulated genes do not map to random locations in the genome but are in fact

Table 2 **Examples on enzymes in stage-specific metabolic pathways**

Locus	Stage				Enzyme	EC number†	Reaction catalysed
	Spz*	Mrz*	Tpz*	Gmt*			
End of glycolysis							
PF10_0363	1.2	–	2.4	–	Pyruvate kinase	2.7.1.40	*P*-enolpyruvate to pyruvate
MAL6P1.160	8.6	66.9	18.8	14.7	Pyruvate kinase		
PF13_0141	46.2	83.9	70.9	78.8	L-lactate dehydrogenase	1.1.1.27	Pyruvate to lactate
TCA cycle and oxidative phosphorylation							
PF10_0218	12.3	–	–	–	Citrate synthase	4.1.3.7	Acetyl coA + oxaloacetate to citrate
PF13_0242	3.2	–	16.9	8.8	Isocitrate dehydrogenase (NADP)	1.1.1.41	Isocitrate to 2-oxoglutarate + CO_2
PF08_0045	2.9	–	2.2	23.1	2-Oxoglutarate dehydrogenase e1 component	1.2.4.2	2-Oxoglutarate to succinyl CoA
PF10_0334	–	–	3.5	27.7	Flavoprotein subunit of succinate dehydrogenase	1.3.5.1	Succinate to fumarate
PFL0630w	3.7	–	–	12.1	Iron-sulphur subunit of succinate dehydrogenase		
PF14_0373	–	–	–	12.7	Ubiquinol cytochrome oxidoreductase	1.10.2.2.	Ubiquinol to cytochrome c reductase in electron transport
PFB0795w	–	–	–	14.2	ATP synthase F1, α-subunit		
PFI1365w	–	–	–	8.8	Cytochrome *c* oxidase subunit	1.9.3.1	
PFI1340w	–	–	–	8.8	Fumarate hydratase	4.2.1.2	Fumarate to malate
MAL6P1.242	30.4	–	–	40.9	Malate dehydrogenase	1.1.1.37	Malate to oxaloacetate

Plasmodium metabolic pathways can be found at http://www.sites.huji.ac.il/malaria/. Spz, sporozoite; mrz, merozoite; tpz, trophozoite; gmt, gametocyte.
*The sequence coverage (that is, the percentage of the protein sequence covered by identified peptides) measured in each stage is reported.
†Enzyme Commission (EC) numbers are reported for each protein.

frequently organized into gene clusters on a chromosome. Gene clustering in *Plasmodium* species has been demonstrated. Ordered arrays of genes involved in virulence and antigenic variation (for example, *var*, *vir* and *rif* genes) are located in the subtelomeric regions of the chromosomes[27,28].

To determine whether gene clustering exists along the entire *P. falciparum* genome, genes whose protein products were detected in our analysis were mapped onto all 14 chromosomes in a stage-dependent manner (Fig. 3a). The 2,415 proteins identified represented an average of 45% of the open reading frames (ORFs) predicted per chromosome. The number of protein hits by chromosome was similar for all stages: sporozoite, merozoite, trophozoite and gametocyte protein lists constituting 19.7%, 15.8%, 19.5% and 21.6% of the predicted ORFs per chromosomes, respectively. Groups of three or more consecutive loci whose protein products were detected in a particular stage were defined as chromosomal clusters encoding co-expressed proteins (Fig. 3b). On the basis of this definition a total of 98 clusters containing 3 loci, 32 clusters containing 4 loci, 5 clusters containing 5 loci, and 3 clusters containing 6 loci were identified (Supplementary Table 3). For each chromosome, the frequency of finding clusters encoding co-expressed proteins containing 3–6 adjacent loci markedly exceeded the probability of finding such clusters by chance (see the footnote of Supplementary Table 3 for details on the probability calculation). Therefore, chromosomal clusters encoding co-expressed proteins were prevalent in the *P. falciparum* genome.

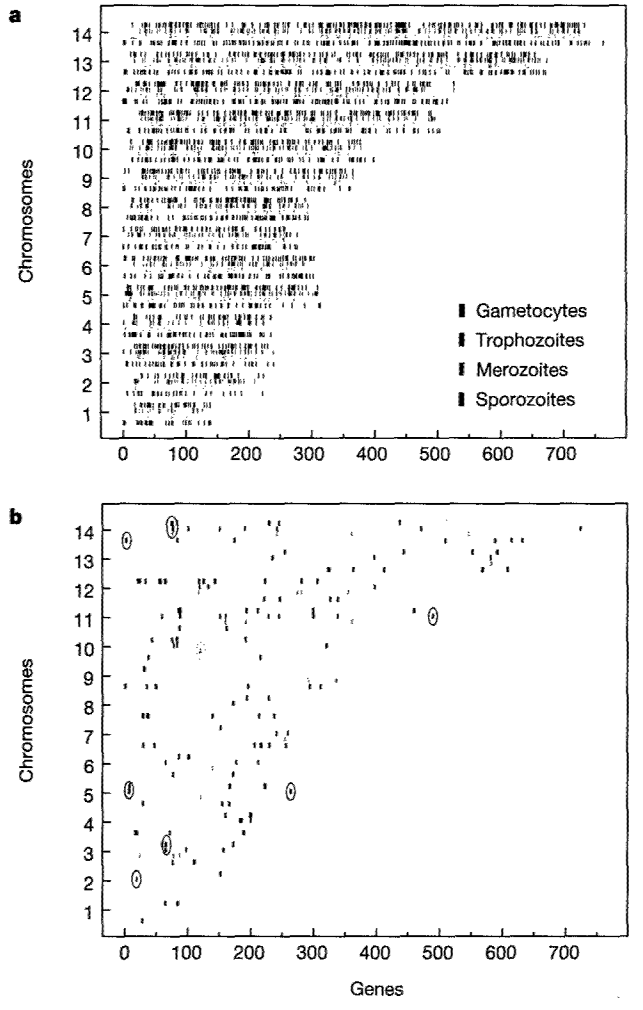


Figure 3 Distribution of expressed proteins by chromosome. **a**, For each stage, genes whose products were detected (coloured vertical bars) are plotted in the order they appear on their chromosome (grey boxes). **b**, Groups of at least three consecutive expressed genes are defined as chromosomal clusters of co-expressed proteins. Examples of such clusters, circled in **b**, are specified in Table 3 and the complete description of the 138 clusters can be found in Supplementary Table 3.

Functionally related genes have been shown to cluster in the *S. cerevisiae*[24] and human genomes[26]. This phenomenon also occurs in *P. falciparum*. A total of 138 clusters encoding co-expressed proteins were identified and 67 of them (49%) contained at least two loci that have been functionally annotated. Of these 67 clusters, 30 contained at least two loci whose annotation clearly indicates that the proteins are functionally related. For example, clusters on chromosomes 3, 5 and 10 contained ribosomal proteins, proteins involved in protein modification, and proteins involved in nucleotide metabolism, respectively (Table 3). Chromosome 14 contained a cluster of four aspartic proteases co-expressed in all of the blood stages (Table 3). This cluster was not detected in sporozoites, where no haemoglobin degradation is expected to occur. Interestingly, whereas the falcipain gene cluster on chromosome 11 appeared in our analysis as a cluster of co-expressed proteins (Supplementary Table 3), the SERA gene cluster on chromosome 2, coding for proteins that share a papain-like sequence motif[29], did not. Of the ten sporozoite-specific clusters, five involved *var* and *rif* genes, such as the *rif* cluster located in the subtelomeric domain of chromosome 14 (Table 3). On the basis of their presence in clusters encoding co-expressed proteins, we were able to suggest functional roles for 24 proteins annotated as hypothetical in the *P. falciparum* genome (Supplementary Table 3). For example, a gametocyte-specific cluster on chromosome 13 encoded two transmission-blocking antigens (Pfs48/45 and Pfs47) and a hypothetical protein, PF13_0246, which might be a gametocyte surface protein. Two clusters on chromosomes 2 and 11 were highly specific to the trophozoite stage (Table 3). Each of these clusters contained well-known secreted and surface proteins, namely KAHRP, PfEMP3, antigen 332, and RESA, all of which have been implicated in knob formation. The highly coordinated expression of these genes makes the three hypothetical proteins listed in these trophozoite-specific gene clusters possible candidates for involvement in cyto-adherence.

Discussion

Although sample handling is a principal consideration when studying pathogens, the expression of large numbers of previously identified proteins was consistent with their published expression profiles, validating our data set as a meaningful sampling of each stage's proteome. This is a particularly important aspect of our analysis as 65% of the 5,276 genes encoded by the *P. falciparum* genome are annotated as hypothetical[1], and of the 2,415 expressed proteins we identified, 51% are hypothetical proteins (Supplementary Table 1). Our results confirmed that these hypothetical ORFs predicted by gene modelling algorithms were indeed coding regions. Furthermore, from all four stages analysed, we identified 439 proteins predicted to have at least one transmembrane segment or a GPI addition signal (18% of the data set) and 304 soluble proteins with a signal sequence; that is, potentially secreted or located to organelles. Well over half of the secreted proteins and integral membrane proteins detected were annotated as hypothetical (Supplementary Table 4). The obvious interest in this class of proteins is that, with no homology to known proteins, they represent potential *Plasmodium*-specific proteins and may provide targets for new drug and vaccine development.

Our comprehensive large-scale analysis of protein expression showed that most surface proteins are more widely expressed than initially thought. In particular, the *var* and *rif* genes, which were thought to be involved in immune evasion only in the blood stage, have now been shown to be expressed in apparently large and varied numbers at the sporozoite stage. These surface proteins might be involved in general interaction processes with host cells and/or immune evasion. An alternative hypothesis is that stage-specific

Table 3 **Examples of chromosomal gene clusters encoding co-expressed proteins**

Chromosome	ID	Locus	Stage				Description	Class	SP	TM
			Spz	Mrz	Tpz	Gmt				
3	64	PFC0285c	2.1	12.7	33.2	18.7	T-complex protein β-subunit	Protein fate	0	0
3	65	PFC0290w	8.3	–	33.8	18.6	40S ribosomal protein S23	Protein synthesis	0	0
3	66	PFC0295c	–	14.9	52.5	21.3	40S ribosomal protein S12	Protein synthesis	0	0
3	67	PFC0300c	–	12.1	30.4	17.9	60S ribosomal protein L7	Protein synthesis	0	0
5	263	PFE1345c	–	–	1.9	1.6	Minichromosome maintenance protein 3	Cell transport	0	0
5	264	PFE1350c	–	–	22.4	–	Ubiquitin-conjugating enzyme	Protein fate	0	0
5	265	PFE1355	–	4.8	2.6	2.6	Ubiquitin carboxy-terminal hydrolase	Protein fate	0	0
5	266	PFE1360c	–	–	7.7	–	Methionine aminopeptidase	Protein fate	0	0
10	119	PF10_0121	10.8	74.5	29	–	Hypoxanthine phosphoribosyltransferase	Metabolism	0	0
10	120	PF10_0122	5.4	6.1	–	6.1	Phosphoglucomutase	Metabolism	0	0
10	121	PF10_0123	–	11.7	–	–	GMP synthetase	Metabolism	0	0
10	122	PF10_0124	0.9	1.8	–	–	Hypothetical protein		0	0
14	74	PF14_0074	26.6	–	–	4.9	Hypothetical protein		0	0
14	75	PF14_0075	–	26.5	43.2	47.4	Plasmepsin	Protein fate	1	0
14	76	PF14_0076	–	6.6	35.2	10	Plasmepsin 1	Protein fate	1	0
14	77	PF14_0077	–	21.2	43	11.5	Plasmepsin 2	Protein fate	1	0
14	78	PF14_0078	–	14.2	52.8	29.9	HAP protein	Protein fate	1	0
14	2	PF14_0002	3.5	–	–	–	Rifin	Surface or organelles	0	1
14	3	PF14_0003	7.9	–	–	–	Rifin	Surface or organelles	1	2
14	4	PF14_0004	6.5	–	–	–	Rifin	Surface or organelles	1	2
2	18	PFB0090c	–	–	3	–	Hypothetical protein, conserved		0	0
2	19	PFB0095c	–	–	3.4	–	Erythrocyte membrane protein 3	Surface or organelles	1	0
2	20	PFB0100c	–	1.5	24.8	–	Knob-associated histidine-rich protein	Surface or organelles	1	0
11	489	PF11_0506	–	–	6.3	4.4	Hypothetical protein		0	1
11	490	PF11_0507	–	–	0.8	–	Antigen 332	Surface or organelles	0	0
11	491	PF11_0508	–	–	3.3	–	Hypothetical protein		0	0
11	492	PF11_0509	–	6.4	3	–	RESA	Surface or organelles	0	0
13	443	PF13_0246	4.5	–	–	8.6	Hypothetical protein		0	0
13	444	PF13_0247	–	–	–	32.4	Transmission-blocking target antigen precursor (Pfs48/45)	Surface or organelles	1	1
13	445	PF13_0248	–	–	–	7.1	Transmission-blocking target antigen precursor (Pfs47)	Surface or organelles	1	1

Clusters of at least three consecutive genes encoding co-expressed proteins are reported with their position (ID) on the chromosome, the sequence coverage measured for these proteins in each stage (%), their current annotation and functional class, and the predicted presence of signal peptide (SP) or transmembrane domains (TM) (based on the TMHMM[43], a transmembrane (TM) helices prediction method based on a hidden Markov model (HMM), big-PI Predictor[44] and SignalP[45] algorithms).

regulation is not as exact as previously thought.

One mechanism of protein expression control that contributes to stage specificity in *P. falciparum* arises from the chromosomal clustering of genes encoding co-expressed proteins. The clusters described in this study demonstrate a widespread high order of chromosomal organization in *P. falciparum* and probably correspond to regions of open chromatin allowing for co-regulated gene expression. The high (A + T) content of the *P. falciparum* genome makes the identification of regulatory sequences such as promoters and enhancers challenging[31,32]. Focusing analyses on stage-specific and multi-stage clusters will facilitate finding stage-specific and general *cis*-acting sequences in the *Plasmodium* genome and will help decipher gene expression regulation during the parasite life cycle.

The malaria parasite is a complex multi-stage organism, which has co-evolved in mosquitoes and vertebrates for millions of years. Designing drugs or vaccines that substantially and persistently interrupt the life cycle of this complex parasite will require a comprehensive understanding of its biology. The *P. falciparum* genome sequence and comparative proteomics approaches may initiate new strategies for controlling the devastating disease caused by this parasite. □

Methods

Parasite material

Plasmodium falciparum clone 3D7 (Oxford) was used throughout. Sporozoites were initially isolated from the salivary glands of *Anopheles stephansi* mosquitoes, 14 days after infection, by centrifugation in a Renograffin 60 gradient, as described[33]. Four sporozoite samples were used as is. A fifth sample underwent an additional purification step on Dynabeads M-450 Epoxy coupled to NFS1 (an anti-*P. falciparum* CS protein monoclonal antibody)[34] according to the manufacturer's instructions (Dynal). Trophozoite-infected erythrocytes from synchronized cultures were purified on 70% Percoll-alanine[30], and the trophozoites released from the erythrocytes[35]. Of the of 260 parasitized erythrocytes counted by Giemsa-stained thin-blood film, 100% were identified as trophozoites. Merozoites were prepared essentially as described in ref. 36, using highly synchronized schizonts and purifying the merozoites by passage through membrane filters. Starting with synchronized asexual parasites grown in suspension culture as described[37,38], gametocytes were prepared by daily media changes of static cultures at 37 °C. When there were very few mature asexual stages present, gametocyte-infected erythrocytes were collected from the 52.5%/45% and 45%/30% interfaces of a Percoll gradient[39]. The gametocytes consisted mostly of stage IV and V parasites with minor contamination (<3%) from mixed asexual stage parasites. Finally, cellular debris from the upper bodies of parasite-free *A. stephansi* and non-infected human erythrocytes were used as controls for sporozoites and blood-stage parasites, respectively. Every effort was made to minimize enzymatic activity and protein degradation during sampling, and the subsequent isolation of the parasites; however, we cannot exclude that some of the differences in protein profiles that we observe between the different life-cycle stages may be a consequence of the sample-handling procedures.

Cell lysis

Five sporozoite, four merozoite, four trophozoite and three gametocyte preparations were lysed, digested and analysed independently. Cell pellets were first diluted ten times in 100 mM Tris-HCl pH 8.5, and incubated in ice for 1 h. After centrifugation at 18,000 *g* for 30 min, supernatants were set aside and microsomal membrane pellets were washed in 0.1 M sodium carbonate, pH 11.6. Soluble and insoluble protein fractions were separated by centrifugation at 18,000 *g* for 30 min. Supernatants obtained from both centrifugation steps were either combined (sporozoites, trophozoites and merozoites) or digested and analysed independently (gametocytes).

Peptide generation and analysis

The method follows that of Washburn *et al.*[5], with the exception that Tris(2-carboxyethyl)phosphine hydrochloride (TCEP-HCl; Pierce) was used to reduce urea-denatured proteins. Peptide mixtures were analysed through MudPIT as described[5].

Protein sequence databases

The *P. falciparum* database contained 5,283 protein sequences. Spectra resulting from contaminant mosquito and erythrocyte peptides had to be taken into account in the sporozoite and blood-stage samples, respectively. Tandem mass spectrometry (MS/MS) data sets from blood stages were therefore searched against a database containing both *P. falciparum* protein sequences and 24,006 ORFs from the human, mouse and rat RefSeq NCBI databases. At the date of the searches, the *Anopheles gambiae* genome was not available. The NCBI database contained 922 *Anopheles* and 313 *Aedes* proteins, which were combined to the 14,335 ORFs of the NCBI *Drosophila melanogaster*[40] database to create a control diptera database. Finally, these databases were complemented with a set of 172 known protein contaminants, such as proteases, bovine serum albumin and human keratins.

MS/MS data set analysis

The SEQUEST algorithm was used to match MS/MS spectra to peptides in the sequence databases[41]. To account for carboxyamidomethylation, MS/MS data sets were searched with a relative molecular mass of 57,000 (M_r 57K) added to the average molecular mass of cysteines. Peptide hits were filtered and sorted with DTASelect[42]. Spectra/peptide matches were only retained if they were at least half-tryptic (Lys or Arg at either end of the identified peptide) and with minimum cross-correlation scores (XCorr) of 1.8 for +1, 2.5 for +2, and 3.5 for +3 spectra and DeltaCn (top match's XCorr minus the second-best match's XCorr divided by the top match's XCorr) of 0.08. Peptide hits were deemed unambiguous only if they were not found in non-infected controls and were uniquely assigned to parasite proteins by searching against combined parasite–host databases. Finally, for low coverage loci, peptide/spectrum matches were visually assessed on two main criteria: any given MS/MS spectrum had to be clearly above the baseline noise, and both *b* and *y* ion series had to show continuity. The Contrast tool[42] was used to compare and merge protein lists from replicate sample runs and to compare the proteomes established for the four stages.

Received 31 July; accepted 9 September 2002; doi:10.1038/nature01107.

1. Gardner, M. J. *et al.* Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419,** 498–511 (2002).
2. Carlton, J. M. *et al.* Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii*. *Nature* **419,** 512–519 (2002).
3. Ben Mamoun, C. *et al.* Co-ordinated programme of gene expression during asexual intraerythrocytic development of the human malaria parasite *Plasmodium falciparum* revealed by microarray analysis. *Mol. Microbiol.* **39,** 26–36 (2001).
4. Hayward, R. E. *et al.* Shotgun DNA microarrays and stage-specific gene expression in *Plasmodium falciparum* malaria. *Mol. Microbiol.* **35,** 6–14 (2000).
5. Washburn, M. P., Wolters, D. & Yates, J. R. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnol.* **19,** 242–247 (2001).
6. Mewes, H. W. *et al.* MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* **30,** 31–34 (2002).
7. Pinder, J. C. *et al.* Actomyosin motor in the merozoite of the malaria parasite, *Plasmodium falciparum*: implications for red cell invasion. *J. Cell Sci.* **111,** 1831–1839 (1998).
8. Holder, A. A. *Malaria Vaccine Development: a Multi-immune Response and Multi-stage Perspective* (ed. Hoffman, S. L.) 77–104 (ASM Press, Washington, 1996).
9. Coppel, R. L. *et al.* Isolate-specific S-antigen of *Plasmodium falciparum* contains a repeated sequence of eleven amino acids. *Nature* **306,** 751–756 (1983).
10. Taylor, H. M. *et al. Plasmodium falciparum* homologue of the genes for *Plasmodium vivax* and *Plasmodium yoelii* adhesive proteins, which is transcribed but not translated. *Infect. Immun.* **69,** 3635–3645 (2001).
11. Kaneko, O. *et al.* The high molecular mass rhoptry protein, RhopH1, is encoded by members of the clag multigene family in *Plasmodium falciparum* and *Plasmodium yoelii*. *Mol. Biochem. Parasitol.* **118,** 223–231 (2001).
12. Trenholme, K. R. *et al.* clag9: A cytoadherence gene in *Plasmodium falciparum* essential for binding of parasitized erythrocytes to CD36. *Proc. Natl Acad. Sci. USA* **97,** 4029–4033 (2000).
13. Klemba, M. & Goldberg, D. E. Biological roles of proteases in parasitic protozoa. *Annu. Rev. Biochem.* **71,** 275–305 (2002).
14. Banerjee, R. *et al.* Four plasmepsins are active in the *Plasmodium falciparum* food vacuole, including a protease with an active-site histidine. *Proc. Natl Acad. Sci. USA* **99,** 990–995 (2002).
15. Rosenthal, P. J., Sijwali, P. S., Singh, A. & Shenai, B. R. Cysteine proteases of malaria parasites: targets for chemotherapy. *Curr. Pharm. Des.* **8,** 1659–1672 (2002).
16. Eggleson, K. K., Duffin, K. L. & Goldberg, D. E. Identification and characterization of falcilysin, a metallopeptidase involved in hemoglobin catabolism within the malaria parasite *Plasmodium falciparum*. *J. Biol. Chem.* **274,** 32411–32417 (1999).
17. Sinden, R. E., Butcher, G. A., Billker, O. & Fleck, S. L. Regulation of infectivity of *Plasmodium* to the mosquito vector. *Adv. Parasitol.* **38,** 53–117 (1996).
18. Billker, O., Shaw, M. K., Margo, G. & Sinden, R. E. Identification of xanthurenic acid as the putative inducer of malaria development in the mosquito. *Nature* **392,** 289–292 (1998).
19. Krungkrai, J., Prapunwattana, P. & Krungkrai, S. R. Ultrastructure and function of mitochondria in gametocytic stage of *Plasmodium falciparum*. *Parasite* **7,** 19–26 (2000).
20. Kappe, S. H. *et al.* Exploring the transcriptome of the malaria sporozoite stage. *Proc. Natl Acad. Sci. USA* **98,** 9895–9900 (2001).
21. Dessens, J. T. *et al.* CTRP is essential for mosquito infection by malaria ookinetes. *EMBO J.* **18,** 6221–6227 (1999).
22. Deitsch, K. W. & Wellems, T. E. Membrane modifications in erythrocytes parasitized by *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **76,** 1–10 (1996).
23. 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).
24. Cohen, B. A., Mitra, R. D., Hughes, J. D. & Church, G. M. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nature Genet.* **26,** 183–186 (2000).
25. Caron, H. *et al.* The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* **291,** 1289–1292 (2001).
26. Lercher, M. J., Urrutia, A. O. & Hurst, L. D. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nature Genet.* **31,** 180–183 (2002).
27. Hernandez-Rivas, R. *et al.* Expressed var genes are found in *Plasmodium falciparum* subtelomeric regions. *Mol. Cell Biol.* **17,** 604–611 (1997).
28. del Portillo, H. A. *et al.* A superfamily of variant genes encoded in the subtelomeric region of *Plasmodium vivax*. *Nature* **410,** 839–842 (2001).
29. Gardner, M. J. *et al.* Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science* **282,** 1126–1132 (1998).
30. Kanaani, J. & Ginsburg, H. Metabolic interconnection between the human malarial parasite *Plasmodium falciparum* and its host erythrocyte. *J. Biol. Chem.* **264,** 3194–3199 (1989).
31. Dechering, K. J. *et al.* Isolation and functional characterization of two distinct sexual-stage-specific promoters of the human malaria parasite *Plasmodium falciparum*. *Mol. Cell Biol.* **19,** 967–978 (1999).
32. Lockhart, D. J. & Winzeler, E. A. Genomics, gene expression and DNA arrays. *Nature* **405,** 827–836 (2000).
33. Pacheco, N. D., Strome, C. P., Mitchell, F., Bawden, M. P. & Beaudoin, R. L. Rapid, large-scale isolation of *Plasmodium berghei* sporozoites from infected mosquitoes. *J. Parasitol.* **65,** 414–417 (1979).
34. Mellouk, S. *et al.* Evaluation of an *in vitro* assay aimed at measuring protective antibodies against sporozoites. *Bull. World Health Organ.* **68** Suppl., 52–59 (1990).
35. Rabilloud, T. *et al.* Analysis of membrane proteins by two-dimensional electrophoresis: comparison of the proteins extracted from normal or *Plasmodium falciparum*-infected erythrocyte ghosts. *Electrophoresis* **20,** 3603–3610 (1999).
36. Blackman, M. J. Purification of *Plasmodium falciparum* merozoites for analysis of the processing of merozoite surface protein-1. *Methods Cell Biol.* **45,** 213–220 (1994).
37. Haynes, J. D. & Moch, J. K. Automated synchronization of *Plasmodium falciparum* parasites by culture in a temperature-cycling incubator. *Methods Mol. Med.* **72,** 489–497 (2002).
38. Haynes, J. D., Moch, J. K. & Smoot, D. S. Erythrocytic malaria growth or invasion inhibition assays with emphasis on suspension culture GIA. *Methods Mol. Med.* **72,** 535–554 (2002).
39. Carter, R., Ranford-Cartwright, L. & Alano, P. The culture and preparation of gametocytes of *Plasmodium falciparum* for immunochemical, molecular, and mosquito infectivity studies. *Methods Mol. Biol.* **21,** 67–88 (1993).
40. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287,** 2185–2195 (2000).
41. Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5,** 976–989 (1994).
42. Tabb, D. L., McDonald, W. H. & Yates, J. R. DTASelect and contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.* **1,** 21–26 (2002).
43. 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).
44. Eisenhaber, B., Bork, P. & Eisenhaber, F. Sequence properties of GPI-anchored proteins near the omega-site: constraints for the polypeptide binding site of the putative transamidase. *Protein Eng.* **11,** 1155–1161 (1998).
45. 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).

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

Acknowledgements

We are grateful to J. Graumann, R. Sadygov, G. Chukkapalli, A. Majumdar and R. Sinkovits for computer programming; C. Deciu for the probability calculations; and C. Delahunty and C. Vieille for critical reading of the manuscript. The authors acknowledge the support of the Office of Naval Research, the US Army Medical Research and Material Command, and the National Institutes of Health (to J.R.Y.). J.D.R. is funded by a Wellcome Trust Prize Studentship. We thank the scientists and funding agencies comprising the international Malaria Genome Project for making sequence data from the genome of *P. falciparum* clone 3D7 public before publication of the completed sequence. The opinions expressed are those of the authors and do not reflect the official policy of the Department of the Navy, Department of Defense, or the US government.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.R.Y. (e-mail: jyates@scripps.edu).

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

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

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

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

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

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

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

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

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

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

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

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

We have identified two previously undescribed clustered gene families; one on chromosome 9 and one on chromosome 13. On chromosome 9, there are 7 copies of a putative protein kinase which show 25–46% amino-acid identity to each other; four of these genes have a predicted signal peptide. Proteomic analysis has shown expression of two of these genes (PFI0105c and PFI0135c)[10]. Chromosome 13 contains a tandem array of 5 parologous genes including *msp7* (ref. 11) with 15–30% identity to each other. Expression of one of these MSP7-like proteins (MAL13P1.174) has been detected, by proteomic studies, during the asexual stage[12]. The significance of the physical localization and function of these different genes is unknown, so further studies of their expression pattern and cellular localization are required. Protein alignments of these families are available (Supplementary Information).

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

Table 1 **Summary statistics**

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

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

*Average number of sequence reads per nucleotide.

†Excluding introns.

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

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

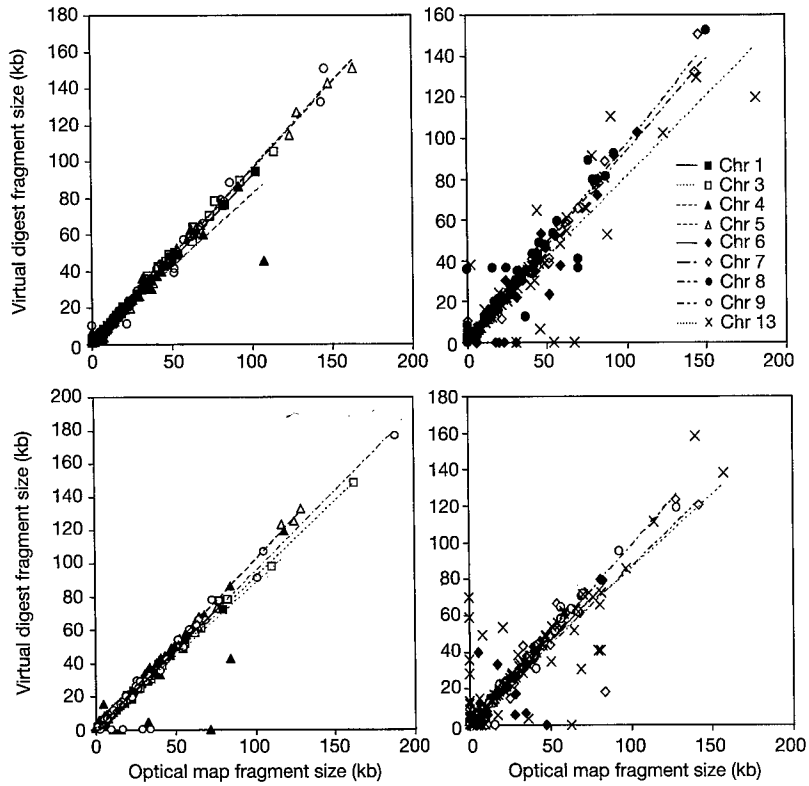


Figure 1 Scatter graphs of virtual restriction digests of completed chromosomes and pseudo-chromosomes against optical map fragment sizes. Top row: completed chromosomes (left) and unfinished chromosomes (right) compared with *Nhe*I optical map. Bottom row; as top row but compared with *Bam*HI optical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

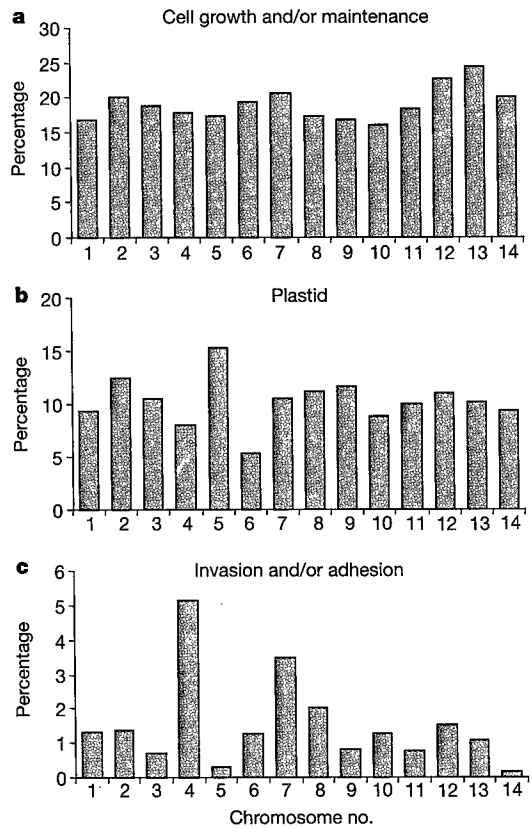


Figure 2 Comparison of the percentage of annotations with specific Gene Ontology terms on each chromosome. **a**, Annotations to 'cell growth and/or maintenance'; **b**, annotations to 'plastid'; **c**, annotations to 'invasion' and/or adhesion.

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

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

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

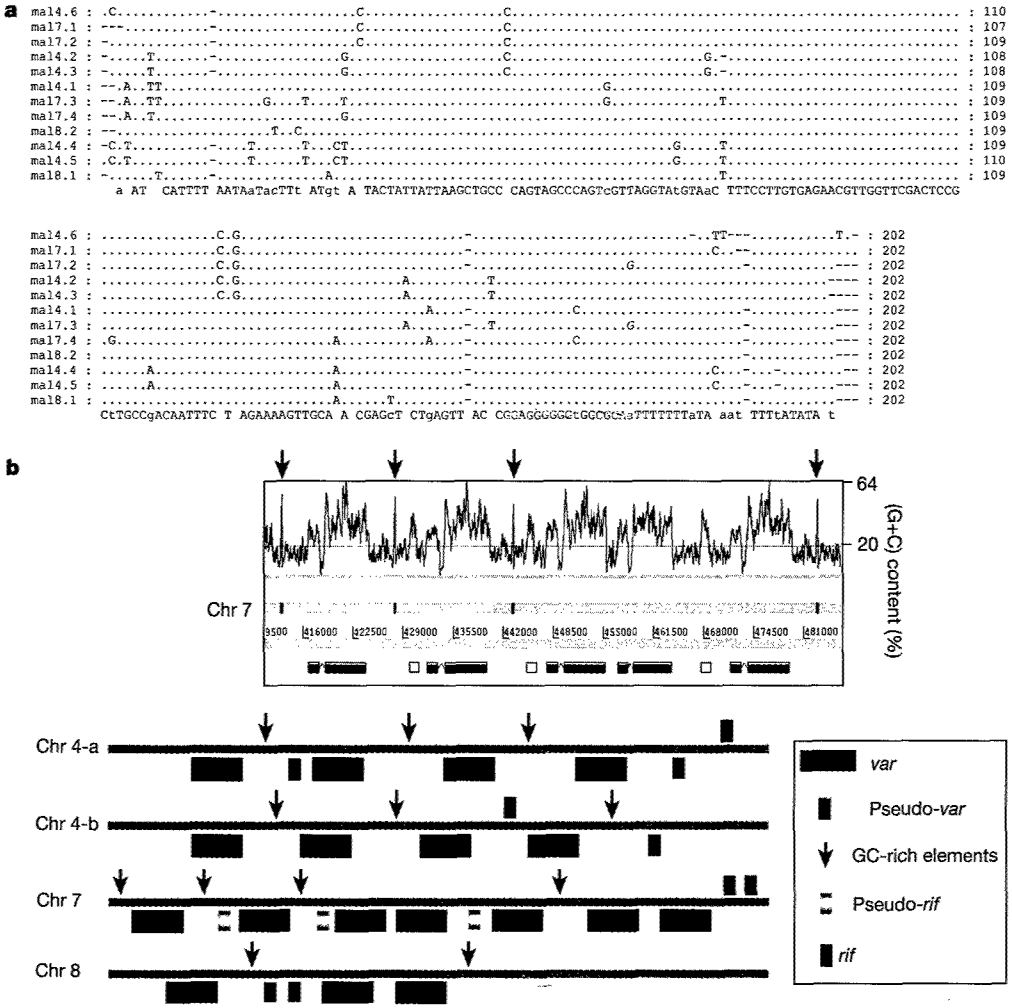


Figure 3 Position and structure of *var*-related (G + C)-rich elements. **a**, Multiple alignment of the (G + C)-rich conserved sequence elements on chromosomes 4, 7 and 8 of *P. falciparum*, using CLUSTAL. Only the non-identical nucleotides across all 12 (G + C)-rich conserved sequence elements are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

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

Methods

Sequencing

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

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

Annotation

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

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

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

1. Gardner, M. J. *et al.* Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science* **282,** 1126–1132 (1998).
2. Bowman, S. *et al.* The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*. *Nature* **400,** 532–538 (1999).
3. Su, X. *et al.* A genetic map and recombination parameters of the human malaria parasite *Plasmodium falciparum*. *Science* **286,** 1351–1353 (1999).
4. Lai, Z. *et al.* A shotgun optical map of the entire *Plasmodium falciparum* genome. *Nature Genet.* **23,** 309–313 (1999).
5. de Bruin, D., Lanzer, M. & Ravetch, J. V. Characterization of yeast artificial chromosomes from *Plasmodium falciparum*: construction of a stable, representative library and cloning of telomeric DNA fragments. *Genomics* **14,** 332–339 (1992).
6. Glockner, G. *et al.* Sequence and analysis of chromosome 2 of *Dictyostelium discoideum*. *Nature* **418,** 79–85 (2002).
7. Piper, M. B., Bankier, A. T. & Dear, P. H. A HAPPY map of *Cryptosporidium parvum*. *Genome Res.* **8,** 1299–1307 (1998).
8. Konfortov, B. A., Cohen, H. M., Bankier, A. T. & Dear, P. H. A high-resolution HAPPY map of *Dictyostelium discoideum* chromosome 6. *Genome Res.* **10,** 1737–1742 (2000).
9. Berriman, M., Aslett, M. & Ivens, A. Parasites are GO. *Trends Parasitol.* **17,** 463–464 (2001).
10. Florens, L. *et al.* A proteomic view of the *Plasmodium falciparum* life cycle. *Nature* **419,** 520–526 (2002).
11. Pachebat, J. A. *et al.* The 22 kDa component of the protein complex on the surface of *Plasmodium falciparum* merozoites is derived from a larger precursor, merozoite surface protein 7. *Mol. Biochem. Parasitol.* **117,** 83–89 (2001).
12. Lasonder, E. *et al.* Analysis of the *Plasmodium falciparum* proteome by high accuracy mass spectrometry. *Nature* **419,** 531–542 (2002).
13. Figueiredo, L. M., Freitas-Junior, L. H., Bottius, E., Olivo-Marin, J. C. & Scherf, A. A central role for *Plasmodium falciparum* subtelomeric regions in spatial positioning and telomere length regulation. *EMBO J.* **21,** 815–824 (2002).
14. O'Donnell, R. A. *et al.* A genetic screen for improved plasmid segregation reveals a role for Rep20 in the interaction of *Plasmodium falciparum* chromosomes. *EMBO J.* **21,** 1231–1239 (2002).
15. Katinka, M. D. *et al.* Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature* **414,** 450–453 (2001).
16. Hyman, R., Fung, E. & Dennis, R. W. *et al.* Sequence of *Plasmodium falciparum* chromosome 12. *Nature* **419,** 534–536 (2002).
17. Hapgood, J. P., Riedemann, J. & Scherer, S. D. Regulation of gene expression by GC-rich DNA cis-elements. *Cell Biol. Int.* **25,** 17–31 (2001).
18. Adhya, S. Multipartite genetic control elements: communication by DNA loop. *Annu. Rev. Genet.* **23,** 227–2250 (1989).
19. Deitsch, K. W., Calderwood, M. S. & Wellems, T. E. Malaria. Cooperative silencing elements in *var* genes. *Nature* **412,** 875–876 (2001).
20. Vazquez-Macias, A. *et al.* A distinct 5′ flanking var gene region regulates *Plasmodium falciparum* variant erythrocyte surface antigen expression in placental malaria. *Mol. Microbiol.* **45,** 155–167 (2002).
21. Quail, M. A. M13 cloning of mung bean nuclease digested PCR fragments as a means of gap closure within A/T-rich, genome sequencing projects. *DNA Seq.* **12,** 355–359 (2001).
22. Rutherford, K. *et al.* Artemis: sequence visualization and annotation. *Bioinformatics* **16,** 944–945 (2000).
23. Salzberg, S. L., Pertea, M., Delcher, A. L., Gardner, M. J. & Tettelin, H. Interpolated Markov models for eukaryotic gene finding. *Genomics* **59,** 24–31 (1999).
24. Cawley, S. E., Wirth, A. I. & Speed, T. P. Phat—a gene finding program for *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **118,** 167–174 (2001).
25. Zdobnov, E. M. & Apweiler, R. InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17,** 847–848 (2001).
26. Sonnhammer, E. L., von Heijne, G. & Krogh, A. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **6,** 175–182 (1998).
27. Nielsen, H., Brunak, S. & von Heijne, G. Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Eng.* **12,** 3–9 (1999).
28. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nature Genet.* **25,** 25–29 (2000).
29. Apweiler, R. *et al.* The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.* **29,** 37–40 (2001).
30. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **30,** 276–280 (2002).
31. Emanuelsson, O., Nielsen, H., Brunak, S. & von Heijne, G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* **300,** 1005–1016 (2000).
32. Claros, M. G. & Vincens, P. Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur. J. Biochem.* **241,** 779–786 (1996).
33. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305,** 567–580 (2001).

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

Acknowledgements

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

Competing interests statement

The authors declare that they have no competing financial interests.

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

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

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

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

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

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

here the nucleotide sequences of chromosomes 10, 11 and 14, and a re-analysis of the chromosome 2 sequence[5]. These chromosomes represent about 35% of the 23-megabase *P. falciparum* genome.

P. falciparum chromosomes were resolved on preparative pulsed field gels, and used to prepare shotgun libraries of 1–2-kilobase (kb) DNA fragments in plasmid vectors. Sequences of randomly selected clones were assembled, and gaps were closed using primer walking on plasmid templates or polymerase chain reaction (PCR) products. The cross-contamination of the chromosomal libraries with sequences from other chromosomes (up to 25%) and the high (A + T) content (80.6%) of *P. falciparum* DNA caused extreme difficulties in the gap closure process. Intergenic regions and introns frequently contained long runs of up to 50 consecutive A or T residues that were difficult to clone and sequence. The high (A + T) content of the chromosomes also prevented the construction of large insert libraries that could be used to construct scaffolds of ordered and oriented contiguous DNA sequences (contigs) during assembly. Similar but more severe problems were reported in the sequencing of the (A + T)-rich chromosome 2 of the slime mould *Dictyostelium discoideum*[6], illustrating the need to develop better methods for the cloning and sequencing of very (A + T)-rich genomes. The reported sequences contain three or four short gaps (<2 kb) in each chromosome. Contigs comprising these chromosomes were joined end-to-end before annotation. Efforts to close the remaining gaps will continue.

Examination of the sequences of chromosomes 2, 10, 11 and 14 revealed that the structure of these chromosomes was similar to that of the other chromosomes. All contained the 97–99% (A + T) putative centromeric sequences reported previously[7]. Conserved subtelomeric sequences[2] were observed in chromosomes 2, 10 and 11, but most of these elements had been deleted from both ends of chromosome 14. The termini of chromosome 14 consisted of telomeric hexamer repeats fused directly to truncated *var* (variant antigen) genes. Deletions of this type are thought to be due to chromosome breakage and healing events that occur during *in vitro* cultivation of the parasite.

Annotation procedures have improved since the publication of the *P. falciparum* chromosome 2 sequence[5]. A gene finding program, phat (pretty handy annotation tool[8]), was developed, supplementing the GlimmerM program[9] used previously. In this work, GlimmerM and phat were retrained on a larger training set of well-

Table 1 **Summary statistics**

Feature	Value				
	Whole genome	Chromosome 2	Chromosome 10	Chromosome 11	Chromosome 14
The genome					
Size (bp)	22,853,764	947,102	1,694,445	2,035,250	3,291,006
No. of gaps	93	0	4	3	3
Coverage*	14.5	11.1	15.6	11.3	9.2
(G + C) content (%)	19.4	19.7	19.7	19.0	18.4
No. of genes	5,268	223 (209)	403	492	769
Mean gene length (bp)†	2,283.3	2,079.1 (2,105.1)	2,085.8	2,127.7	2,315.1
Gene density (bp per gene)	4,338.2	4,247.1 (4,531.6)	4,204.6	4,136.7	4,279.6
Percent coding	52.6	49.0 (46.5)	49.6	51.4	54.1
Genes with introns (%)	53.9	57.0 (43.1)	51.4	50.4	49.9
Genes with ESTs (%)	49.1	46.2	48.1	48.4	46.9
Gene products detected by proteomics‡ (%)	51.8	43.5	49.1	51.0	52.1
Exons					
Number	12,674	510 (353)	892	1,094	1,757
Mean no. per gene	2.4	2.3 (1.7)	2.2	2.2	2.3
(G + C) content (%)	23.7	24.4 (24.3)	24.5	23.5	22.8
Mean length (bp)	949.1	909.1 (1,246.3)	942.3	956.9	1,013.3
Total length (bp)	12,028,350	463,647 (439,944)	840,576	1,046,814	1,780,305
Introns					
Number	7,406	287 (144)	489	602	988
(G + C) content (%)	13.5	13.4 (13.4)	13.6	13.7	13.5
Mean length (bp)	178.7	202.4 (208.4)	234.5	189.4	185.5
Total length (bp)	1,323,509	58,080 (30,006)	114,676	114,012	183,240
Intergenic regions					
(G + C) content (%)	13.6	13.5 (14.1)	13.6	14.1	13.2
Mean length (bp)	1,693.9	1,702.3 (2,063.2)	1,678.5	1,768.5	1,717.2
RNAs					
No. of tRNA genes	43	1	0	2	2
No. of 5S rRNA genes	3	0	0	0	3
No. of 5.8S, 18S and 28S rRNA units	7	0	0	1	0
The proteome					
Total predicted proteins	5,268	223	403	492	769
Hypothetical proteins§	3,208	121	265	339	485
InterPro matches	2,650	116	210	283	455
Pfam matches	1,746	77	133	184	275
Gene Ontology					
Process	1,301	63	89	110	168
Function	1,244	54	74	95	174
Component	2,412	120	181	220	308
Targeted to apicoplast	551	28	36	52	73
Targeted to mitochondrion	246	10	13	17	33
Structural features					
Transmembrane domain(s)	1,631	87	133	141	202
Signal peptide	544	28	41	52	63
Signal anchor	367	19	32	31	51

Numbers in parentheses under chromosome 2 indicate values obtained in the previous annotation[5]. Specialized searches used the following programs and databases: InterPro[21], Pfam[19] and Gene Ontology[22]. Predictions of apicoplast and mitochondrial targeting were performed using TargetP[26] and MitoProtII[25]; transmembrane domains, TMHMM[24]; and signal peptides and signal anchors, SignalP-2.0 (ref. 23).

*Average number of sequence reads per nucleotide. EST, expressed sequence tag.

†Excluding introns.

‡Percent of proteins detected in parasite extracts by two independent proteomic analyses[29,30].

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

characterized genes, complementary DNAs (cDNAs) and products of PCR with reverse transcription (RT–PCR) (total length 540 kb) than was used in the earlier work. A program called Combiner was used to evaluate the GlimmerM and phat predictions, as well as the results of searches against nucleotide and protein databases, to construct consensus gene models. To assess the effect of these modifications, chromosome 2 was re-annotated and the results were compared with the previous annotation.

Application of these automated annotation procedures and manual curation of the resulting gene models for chromosome 2 produced 223 gene models. The revised procedures detected 21 genes not predicted previously, and 13 of the existing chromosome 2 models collapsed into six models in the new annotation. Of the 21 new gene models, all but one had no significant similarity to proteins in a non-redundant amino-acid database. However, at least a portion of each of the 21 gene models had been predicted independently by both GlimmerM and phat, suggesting that many of these models were likely to represent coding sequences. On the other hand, five of the new gene models encoded proteins less than 100 amino acids in length, and may be less likely to encode proteins.

Another major difference was the detection of additional small exons. In the earlier annotation of chromosome 2, the 209 predicted genes contained 353 exons, or an average of 1.7 exons per gene. The revised procedures reported here revealed 510 exons, or 2.3 exons per gene; 60% of the new exons were predicted to be additions to the gene models reported previously. Most cases involved the addition of one or two exons per gene. In three notable cases, however, 7 to 12 small exons were added to the earlier gene models, and almost all of the new exons had been predicted by both of the gene finding programs. Overall, use of the revised annotation procedures resulted in the detection of additional genes and many small exons, which is reflected in the higher gene density and shorter mean exon length in the newly annotated chromosome 2 sequence compared with the previous annotation (Table 1). Despite these improvements in software and training sets, gene finding in *P. falciparum* remains challenging, and the gene structures presented here should be regarded as preliminary until confirmed by sequence information obtained from cDNAs or RT–PCR experiments[10]. Accurate prediction of the 5′ ends of genes is particularly difficult. Generation of larger training sets, including additional expressed sequence tags (ESTs) and full-length cDNAs, would greatly improve the sensitivity and accuracy of gene predictions.

These annotation procedures were also applied to the analysis of chromosomes 10, 11 and 14 (Table 1; maps of these chromosomes are available as Supplementary Information). The 10 short gaps in the chromosomes should not have interfered with the gene predictions; only the genes adjacent to the gaps might have been affected. All three chromosomes were similar in terms of gene density, coding percentage and other parameters. A complete description of the parasite genome is contained in the accompanying Article[2].

Annotation of chromosomes 10, 11 and 14 revealed four proteins with sequence similarity to SR proteins, a family of conserved splicing factors that contain RNA-binding domains and a protein interaction domain rich in Ser and Arg residues (SR domain; PF10_0047, PF10_0217, PF11_0200, PF14_0656). Three additional putative SR proteins were identified on chromosomes 5 and 13 (PFE0160c, PFE0865c, MAL13P1.120). SR proteins are thought to bind to exonic splicing enhancers (ESEs), short (6–9 bp) sequences within exons that assist in the recognition of nearby splice sites, and to interact with components of the spliceosome[11]. ESEs have previously been characterized only in multicellular organisms. To determine whether *P. falciparum* may use ESEs as part of its splicing machinery, a Gibbs sampling algorithm for motif detection[12] was applied to a set of *P. falciparum* exons to detect any exonic splicing enhancers (ESEs). The exons were extracted from the set of well-characterized genes used to train the GlimmerM gene finder. Regions of 50 bp regions were selected from both ends of the internal exons and divided into two different data sets, representing the exon regions adjacent to both 5′ and 3′ splice sites. At least 10 runs of the Gibbs sampler were performed for each data set in order to identify the most probable motif with a length of 5–9 nucleotides. The motif with the highest maximum *a posteriori* probability was retained. This analysis identified a motif with the consensus GAAGAA, which is identical to ESEs found in human exons[13,14]. The identification of several putative SR proteins, and sequences identical to the ESEs in humans, suggests that some features of exon recognition and splicing observed in higher eukaryotes may be conserved in *P. falciparum*. □

Methods

Sequencing and closure

P. falciparum clone 3D7 was selected for sequencing because it can complete all phases of the life cycle, and had been used in a genetic cross[15] and the Wellcome Trust Malaria Genome Mapping Project[16]. High-molecular-mass genomic DNA was subjected to electrophoresis on preparative pulsed field gels, and chromosomes were excised. DNA was extracted from the gel, sheared, and cloned into the pUC18 vector as described[5] (chromosomes 2, 14) or into a modified pUC18 vector via *Bst*XI linkers (chromosomes 10, 11). Sequences were assembled and gaps were closed by primer walking on plasmid DNAs or genomic PCR products, or by transposon insertion[5]. Ordering of contigs was facilitated by the use of sequence tagged sites[16] and microsatellite markers[17]. The final assembly of each chromosome was verified by comparison with *Bam*HI and *Nhe*I optical restriction maps[18]. The average difference in size between the experimentally determined restriction fragments and the fragments predicted from the sequence was approximately 5–6% for chromosomes 11 and 14 for both enzymes. For chromosome 10, the average difference in fragment sizes was 6.1% for the *Nhe*I map, but the *Bam*HI optical and prediction restriction maps could not be aligned. Because the *Nhe*I optical restriction map agreed with that predicted from the sequence, the chromosome 10 assembly was judged to be correct.

Annotation

GlimmerM[9] and phat[8] were trained on 117 *P. falciparum* genes and 39 cDNAs taken from GenBank, plus 32 genes from chromosomes 2 and 3 that had been verified by RT–PCR (provided by R. Huestis and K. Fischer; the training set is available at http://www.tigr.org/software/glimmerm/data). The GlimmerM and phat predictions, and sequence alignments of the chromosomes to protein and cDNA databases, were evaluated by the Combiner program. The program used a linear weighting method and dynamic programming to construct consensus gene models that were curated manually using AnnotationStation (AffyMetrix Inc.). Predicted proteins were searched against a non-redundant amino-acid database using BLASTP; other features were identified by searches against the Pfam[19], PROSITE[20] and InterPro[21] databases. The results of all analyses were reviewed using Manatee, a tool that interfaces with a relational database of the information produced by the annotation software. Predicted gene products were manually assigned Gene Ontology [22] terms. Signal peptides and signal anchors were predicted with SignalP-2.0 (ref. 23). Transmembrane helices were predicted with TMHMM[24]. Mitochondrial- and apicoplast-targeted proteins were predicted by MitoProtII[25], TargetP[26] and PATS[27]. tRNA-ScanSE[28] was used to identify transfer RNAs.

Received 6 August; accepted 2 September 2002; doi:10.1038/nature01094.

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. Gardner, M. J. *et al.* Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419,** 498–511 (2002).
3. Hall, N. *et al.* Sequence of *Plasmodium falciparum* chromosomes 1, 3–9 and 13. *Nature* **419,** 527–531 (2002).
4. Hyman, R. W. *et al.* Sequence of *Plasmodium falciparum* chromosome 12. *Nature* **419,** 534–537 (2002).
5. Gardner, M. J. *et al.* Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science* **282,** 1126–1132 (1998).
6. Glockner, G. *et al.* Sequence and analysis of chromosome 2 of *Dictyostelium discoideum*. *Nature* **418,** 79–85 (2002).
7. Bowman, S. *et al.* The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*. *Nature* **400,** 532–538 (1999).
8. Cawley, S. E., Wirth, A. I. & Speed, T. P. Phat—a gene finding program for *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **118,** 167–174 (2001).
9. Salzberg, S. L., Pertea, M., Delcher, A., Gardner, M. J. & Tettelin, H. Interpolated Markov models for eukaryotic gene finding. *Genomics* **59,** 24–31 (1999).
10. Huestis, R. & Fischer, K. Prediction of many new exons and introns in *Plasmodium falciparum* chromosome 2. *Mol. Biochem. Parasitol.* **118,** 187–199 (2001).
11. Maniatis, T. & Tasic, B. Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* **418,** 236–243 (2002).
12. Lawrence, C. E. *et al.* Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262,** 208–214 (1993).
13. Ramchatesingh, J., Zahler, A. M., Neugebauer, K. M., Roth, M. B. & Cooper, T. A. A subset of SR proteins activates splicing of the cardiac troponin T alternative exon by direct interactions with an exonic enhancer. *Mol. Cell Biol.* **15,** 4898–4907 (1995).
14. Fairbrother, W. G., Yeh, R. F., Sharp, P. A. & Burge, C. B. Predictive identification of exonic splicing enhancers in human genes. *Science* **297,** 1007–1013 (2002).

15. Walliker, D., Quayki, I., Wellems, T. E. & McCutchan, T. F. Genetic analysis of the human malaria parasite *Plasmodium falciparum*. *Science* **236,** 1661–1666 (1987).
16. Foster, J. & Thompson, J. The *Plasmodium falciparum* genome project: a resource for researchers. *Parasitol. Today* **11,** 1–4 (1995).
17. Su, X. *et al.* A genetic map and recombination parameters of the human malaria parasite *Plasmodium falciparum*. *Science* **286,** 1351–1353 (1999).
18. Lai, Z. *et al.* A shotgun optical map of the entire *Plasmodium falciparum* genome. *Nature Genet.* **23,** 309–313 (1999).
19. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **30,** 276–280 (2002).
20. Falquet, L. *et al.* The PROSITE database, its status in 2002. *Nucleic Acids Res.* **30,** 235–238 (2002).
21. 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).
22. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nature Genet.* **25,** 25–29 (2000).
23. 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).
24. 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).
25. Claros, M. G. & Vincens, P. Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur. J. Biochem.* **241,** 779–786 (1996).
26. 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).
27. 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).
28. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25,** 955–964 (1997).
29. Florens, L. *et al.* A proteomic view of the *Plasmodium falciparum* life cycle. *Nature* **419,** 520–526 (2002).
30. Lasonder, E. *et al.* Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry. *Nature* **419,** 537–542 (2002).

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

Acknowledgements

We thank our colleagues at The Institute for Genomic Research and the Naval Medical Research Center for support; 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; and S. Cawley for assistance with phat. This work was supported by 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). Chromosome sequences have been deposited in GenBank with accession numbers AE001362.2 (chromosome 2), AE014185 (chromosome 10), AE01486 (chromosome 11) and AE01487 (chromosome 14), and in PlasmoDB (http://plasmodb.org).

Sequence of *Plasmodium falciparum* chromosome 12

Richard W. Hyman, Eula Fung, Aaron Conway, Omar Kurdi, Jennifer Mao, Molly Miranda, Brian Nakao, Don Rowley, Tomoaki Tamaki, Fawn Wang & Ronald W. Davis

Stanford Genome Technology Center, 855 California Avenue, Palo Alto, California 94304 USA, and Departments of Biochemistry and Genetics, Stanford University Medical College, Stanford University, Stanford, California 94305, USA

The human malaria parasite *Plasmodium falciparum* is responsible for the death of more than a million people every year[1]. To stimulate basic research on the disease, and to promote the development of effective drugs and vaccines against the parasite, the complete genome of *P. falciparum* clone 3D7 has been sequenced, using a chromosome-by-chromosome shotgun strategy[2–4]. Here we report the nucleotide sequence of the third largest of the parasite's 14 chromosomes, chromosome 12, which comprises about 10% of the 23-megabase genome. As the most (A + T)-rich (80.6%) genome sequenced to date, the *P. falciparum* genome presented severe problems during the assembly of primary sequence reads. We discuss the methodology that yielded a finished and fully contiguous sequence for chromosome 12. The biological implications of the sequence data are more thoroughly discussed in an accompanying Article (ref. 3).

At the inception of the Malaria Genome Project, our colleagues at the Institute for Genomic Research (TIGR) and the Wellcome Trust Sanger Institute (WTSI) sequenced *P. falciparum* chromosomes 2 and 3 (refs 5, 6). We chose to sequence the third-largest *P. falciparum* chromosome, chromosome 12, which comprises about 10% of the genome. We made this choice because a 'tiling path' had just been published[7]. (A tiling path is an ordered set of recombinant DNAs covering a large DNA sequence, such as chromosome 12. In this case, the tiling path is composed of yeast artificial chromosomes (YACs) with sequence-tagged sites (STSs, mapped sequence markers).) We predicted that the YACs and the STSs would be helpful in positioning sequence contigs (stretches of contiguous sequence) along *P. falciparum* chromosome 12.

From the published data[7], we defined a 21 YAC tiling path across *P. falciparum* chromosome 12 (Supplementary Fig. 1). However, we did not want to rely exclusively on sequencing YACs because of three important concerns, which turned out to be warranted. (1) Base changes in the sequence can occur during the construction of any recombinant DNA/YAC, and mutations can occur during passage of any YAC in yeast. (2) One or more YACs in the tiling path might not overlap a neighbouring YAC, creating a physical gap in the sequence. (3) Three of the YACs in the tiling path were derived from *P. falciparum* clone B8 rather than clone 3D7. Polymorphisms between the DNAs of the two strains could hinder the assembly process. Therefore, we devised the following overall strategy. We would sequence random pieces of (that is, use 'shotgun sequencing' on) each of the YACs in the minimum tiling path to low coverage—just enough to establish a 'bin' (a group of related sequences). The bins would give us physical position information across *P. falciparum* chromosome 12. The STSs would give us physical position information within each bin. In addition, we would shotgun-sequence *P. falciparum* chromosome 12 itself. The sequence of each chromosome 12 shotgun sequence 'read' (a sequence of length 100–600 bases derived from a piece of DNA) would be compared to the sequences in each bin. When there was a good match, the read would stay in that bin. This process is highly iterative.

The 21 YACs comprising the minimum tiling path varied considerably in size, with a range of 40–220 kilobases (kb; ref. 7). Our shotgun sequence coverage of the YACs also varied considerably, with a range of 0.5–9.7 YAC coverage (Supplementary Table 1). However, with the exception of four YACs with which we experimented with high coverage early in this project, the shotgun sequence coverage of the remaining YACs was low, as originally planned. In total, there are 14,159 YAC reads (2.6-fold chromosome 12 coverage) supporting the final chromosome 12 sequence. In addition, we produced 69,532 *P. falciparum* chromosome 12 shotgun reads (11.3-fold chromosome 12 coverage) that support the chromosome 12 consensus sequence (Supplementary Table 2). After assembling all of the shotgun sequence data, nearly all of the contigs could be placed unambiguously relative to each other, based on the YAC bins and the STSs. The few remaining contigs were positioned unambiguously by using the genetic map of *P. falciparum* chromosome 12 constructed through the use of microsatellite markers derived from our chromosome 12 sequence[8,9]. The very few remaining contigs were placed unambiguously by use of the data that accrued during the process of 'finishing' (identifying and replacing all problems in the assembled sequence).

Every part of the assembled sequence of *P. falciparum* chromosome 12 was carefully examined to identify problems in the sequence. These problems were of many types, including (but not limited to) gaps in the sequence, weakly supported sequence,

ambiguities in the sequence, and sequence in only one direction. The problems in the assembled sequence were resolved during the process of finishing. As a result of finishing, we added an additional 7,500 reads (1.09-fold chromosome 12 coverage) in support of the consensus sequence. The final phase of the finishing process was sequence validation. We manually scanned through the *P. falciparum* chromosome 12 consensus sequence, and noted regions (sometimes as large as several hundred base pairs) where we were dissatisfied with the supporting reads. The substantial majority of these regions were composed (almost entirely) of sets of various tandem repeats. As such, there was little reason to try to re-sequence across these regions. However, we were concerned that we may have missed some unique sequence buried in among the repeats.

We therefore undertook a validation process whereby we compared the lengths of PCR (polymerase chain reaction) products with the lengths predicted by the consensus sequence. Manually, we designed three pairs of nested custom primers and performed PCR reactions with those pairs of primers, using total *P. falciparum* genomic DNA as the template. The lengths of the PCR products were determined experimentally by gel electrophoresis. We could predict the lengths of these PCR products by counting bases in the consensus sequence. Overall, we successfully completed 219 validation reactions. Because we attempted three PCR reactions for every weakly supported place in the consensus sequence, we achieved, at least, one PCR product for virtually all positions. For 201 reactions (92%), the PCR product's measured length was within experimental error of the predicted length. For 18 reactions, representing eight positions on the consensus sequence, the predicted and experimental lengths disagreed by just beyond experimental error. In these cases, we prepared, and sequenced, the appropriate PCR products. For validation using completely independent data, we made use of the two optical restriction enzyme cleavage maps of *P. falciparum* chromosome 12 (ref. 10). (We note that we did not use these optical restriction enzyme cleavage patterns previously: for example, to place contigs relative to each other along chromosome 12.) We normalized the two published cleavage maps to 2.27 megabases, the length of chromosome 12 as determined by our sequence. A comparison of those normalized values to the virtual fragment sizes predicted from our sequence revealed an average discrepancy of less than 6%, which represents excellent agreement.

Table 1 **Summary of relevant features**

Feature	Value	
	Whole genome	Chr. 12
The genome		
Size (bp)	22,853,764	2,271,477
No. of gaps*	93	0
Coverage†	14.5	14.9
(G + C) content (%)	19.4	19.3
No. of genes‡	5,268	526
Mean gene length (bp)	2,283.3	2,303.1
Gene density (bp per gene)	4,338.2	4,318.4
Per cent coding§	52.6	53.3
Genes with introns (%)	53.9	51.1
Genes with ESTs (%)	49.1	48.7
Gene products detected by proteomics‖ (%)	51.8	51.0
Exons		
Number	12,674	1,270
Mean no. per gene	2.4	2.4
(G + C) content (%)	23.7	23.8
Mean length (bp)	949.1	953.9
Total length (bp)	12,028,350	1,211,430
Introns		
Number	7,406	744
(G + C) content (%)	13.5	13.4
Mean length (bp)	178.7	172.9
Total length (bp)	1,323,509	128,665
Intergenic regions		
(G + C) content (%)	13.6	13.6
Mean length (bp)	1,693.9	1,703.6
RNAs		
No. of tRNA genes	43	3
No. of 5S rRNA genes	3	0
No. of 5.8S, 18S and 28S rRNA units	7	0
The proteome		
Total predicted proteins	5,268	526
Hypothetical proteins¶	3,208	332
InterPro matches	2,650	256
Pfam matches	1,746	222
Gene Ontology		
Process	1,301	142
Function	1,244	125
Component	2,412	246
Targeted to apicoplast	551	58
Targeted to mitochondrion	246	32
Structural features		
Transmembrane domain(s)	1,631	155
Signal peptide	544	49
Signal anchor	367	33

EST, expressed sequence tag. Specialized searches used the following programs and databases: InterPro[18]; Pfam[19]; Gene Ontology[20]. Predictions of apicoplast and mitochondrial targeting were performed using TargetP[21] and MitoProtII[22]; transmembrane domains, TMHMM[23]; and signal peptides and signal anchors, SignalP-2.0[24].
*Most gaps are probably <2.5 kb.
†Average number of sequence reads per nucleotide.
‡70% of these genes had similarity to expressed sequence tags or encoded proteins detected by proteomics analyses[11,12].
§Excluding introns.
‖Per cent of proteins detected in parasite extracts by two independent proteomic analyses[11,12].
¶Hypothetical proteins are proteins with insufficient similarity to characterized proteins in other organisms to justify provision of functional assignments.

Our final consensus sequence for *P. falciparum* chromosome 12 is composed of 2,271,477 base pairs (bp) (Table 1). The sequence is completely contiguous; there are no gaps. This sequence is supported by a total of 91,191 reads (14.9-fold chromosome 12 coverage). Overall, the guanine-plus-cytosine (G + C) content of chromosome 12 is 19.3%. As expected from this very low (G + C) content, the *P. falciparum* chromosome 12 sequence contains many long runs of consecutive adenine and thymine residues. Runs of, at least, 20 such bases cover 18% of the chromosome 12 sequence. Bowman *et al.*[6] were able to identify a region of extremely low (G + C) content as the best candidate location for the centromere of *P. falciparum* chromosome 3. Our chromosome 12 sequence contains an analogous region between base positions 1,282,701 and 1,284,791 (2,090 bp; 0.092% of chromosome 12). That region has a (G + C) content of 1.9%, is composed of the short tandem repeats characteristic of centromeres, and is, therefore, the putatative centromere of *P. falciparum* chromosome 12. To predict the genes encoded by *P. falciparum* chromosome 12, we used 'gene-calling' software in parallel with our colleagues at the WTSI[2]. *Plasmodium falciparum* chromosome 12 is predicted to encode 529 genes (Table 1, and Supplementary Fig. 2), including 23 genes from known *Plasmodium*-specific protein-encoding gene families (eight *vars*, twelve *rifs*, and three *stevors*[3]) and three transfer RNA genes. The segmental (G + C) content affects the speed and accuracy of sequencing. The predicted exons are, on average, 23.8% (G + C), which is significantly higher than the overall average (19.3%). The predicted introns are, on average, 13.4% (G + C), which is significantly lower than the overall average. All of the chromosome 12 numbers in Table 1 are in accord with the equivalent numbers for the other 13 *P. falciparum* chromosomes[2,4].

Independent data support some of the 526 *P. falciparum* chromosome 12 predicted protein-encoding genes/exons, although support for an exon does not necessarily validate the entire gene predicted to contain that exon. Of the 526 predicted genes, 174 (33%) have good matches to sequences in GenBank, while 256 (48.7%) have excellent matches to expressed sequence tags (ESTs, which are short sequences derived from messenger RNAs). In two accompanying publications, Florens *et al.*[11] and Lasonder *et al.*[12] report the *P. falciparum* proteome (the sum of all proteins encoded by the *P. falciparum* genome) at the main stages of the complex *P. falciparum* life cycle. Peptides were identified for 268 (51.0%) of the predicted protein-coding sequences of chromosome 12. Our colleagues at the WTSI assigned Gene Ontology (GO) categories to the predicted *P. falciparum* genes[2] (Table 1).

When sequencing recombinant DNAs/YACs, the possibility always exists that the recombinant DNAs/YACs do not represent accurately the sequence of the original DNA. Bases may have been added, subtracted and/or changed during the biochemical construction of the YACs, and mutations may have occurred during passage of the YACs in yeast. We can address this issue for three of the *P. falciparum* strain 3D7 YACs (341, 293 and 25; Supplementary Table 1), because we had shotgun sequenced these three YACs to high coverage (5.7-, 6.3- and 8.4-fold YAC coverage, respectively; Supplementary Table 1). We separately assembled these three YACs using only the YAC-derived reads, and identified regions of high-quality, well-supported assembled sequence. Then, using the software cross_match[13], we compared the YAC-derived consensus sequence with the overall consensus sequence. From a comparison of a total of 94,151 bp from the three YACs, we found two separate single-base differences. Thus, the resulting frequency of difference between YAC sequence and chromosome sequence is 2 bp/94,151 bp, or 0.000021. Of the three strain B8 YACs that are part of the chromosome 12 tiling path, we sequenced only YAC B8-420 to high YAC coverage (12.9-fold YAC coverage; Supplementary Table 1). We assembled solely YAC B8-420 reads, and identified regions of high quality, well-supported assembled sequence. These regions encompass a total of 43,375 bp. Again, using the software cross_match, we compared the high-quality strain B8 sequence to our chromosome 12 consensus sequence over the same 43,375 bp. We found 56 differences of several types, including single-base differences and small deletions/insertions. The resulting DNA polymorphism frequency between the *P. falciparum* strain 3D7 sequence and the strain B8 sequence is 56 bp/43,375 bp, or 0.0013. This frequency is 61 times greater than the mutation frequency (0.0013/0.000021 = 61). □

Methods

DNA sequencing

Plasmodium falciparum chromosome 12 DNA twice purified by contour-clamped homogeneous electric field (CHEF) gel electrophoresis, *P. falciparum* genomic DNA for use as template in PCR reactions, and the appropriate PCR reaction conditions using HotStar kits (Stratagene) were supplied by D. Carucci and his team at the Navy Medical Research Center (NMRC). Yeast/YAC stocks and relevant information were supplied by J. Thompson of the Walter and Eliza Hall Institute (WEHI). The yeast/YACs were grown as described[7]. Agarose plugs containing the YACs were prepared. YACs were twice purified by CHEF gel electrophoresis. The first CHEF gel was composed of standard agarose. The second CHEF gel was composed of low-melting-point (LMP) agarose. YACs were freed from the LMP agarose by agarase digestion at 37 °C. For the construction of shotgun sequencing libraries, the *P. falciparum* chromosome 12 and YAC DNAs were first point-sink sheared (a random shearing process) to an average size of 1 kb for the M13-based vector and 2 kb for the pUC-based vector, as previously described[14]. Both M13-based and pUC-based sequencing libraries were constructed from the *P. falciparum* chromosome 12 DNA. Only M13-based sequencing libraries were constructed from the YAC DNAs.

Software

The public software phred was used to call the bases and to assign a quality score to each base[13,15]; a 'phred score' of 20 or higher is considered good quality. All of the sequence data presented here refer solely to good-quality sequence. The public software phrap was used to assemble the shotgun reads[15]. Consed was used to edit the assembled sequence[16]. The final gene set was chosen through a manual review of the data. Each base in the open reading frames (ORFs) of the *P. falciparum* chromosome 12 consensus sequence is supported by, at least, three good-quality reads with, at least, one read in each direction. However, because of resource limitations, there are still a few regions (mostly in stretches of repeated sequences) supported by reads in only one direction.

Finishing

There were two types of gaps in the assembled sequence. (1) Plasmid bridges (also known as 'sequence gaps'). A plasmid bridge connects two contigs, and is composed of paired, opposing reads wherein one read is in one contig while the other read is in a second contig. To sequence across plasmid bridges, we designed custom primers in both directions. Using those primers and the particular plasmid as template, we performed primer-extension sequencing. When necessary, we designed custom primers for an additional round of primer-extension sequencing. The addition of primer-extension reads often attracted previously unassembled shotgun reads to that position. (2) Physical gaps. We used two strategies to close physical gaps. The first was to use existing templates that pointed into a physical gap. We designed custom primers and coupled these with their respective templates for primer-extension sequencing. This procedure extended good-quality sequence into a gap. In those cases where there was still more length to the templates pointing into a gap than covered by good-quality sequence, we again designed custom primers and undertook additional rounds of primer-extension sequencing. When the primer-extension procedure failed to close a physical gap or could not be used because no templates pointed into a gap, we turned to the second, PCR-based, strategy. We designed three nested pairs of custom primers across each physical gap. We used the primer pairs, along with total *P. falciparum* genomic DNA as the template, for PCR reactions. The PCR products were gel-purified and sequenced. Using these two strategies separately or in combination, we were successful in closing every sequence and physical gap in our *P. falciparum* chromosome 12 sequence.

In addition to the gaps, some regions in the assembled sequence of chromosome 12 had good-quality reads in only one direction. Both directions are required, because the sequence in one direction is a check on the sequence in the complementary direction. Therefore, achieving good-quality sequence reads in both directions was a high priority. Where templates existed in the opposing direction, we designed custom primers and undertook primer-extension sequencing on those templates. Where templates did not exist in the opposing direction, we used two different strategies to achieve sequence in the missing direction. One strategy was to undertake an M13 template-based procedure with the existing templates. For this procedure, we started with an M13-based template and used PCR to synthesize the complementary strand in the opposing direction. Then, we sequenced that new DNA strand using primer-extension chemistry. This procedure is often called 'M13-reverses'. The second strategy to achieve sequence in the missing, opposing direction was to construct one or more PCR products across the region. The once-missing, opposing strand of the PCR product was sequenced.

There were many other places in the assembled sequence of chromosome 12 where the sequence was thin (supported by only a few shotgun reads), or ambiguous, or of low quality, and so on. For example, the sequences on both sides of homopolymers of adenine, which occur frequently on this very (adenine + thymine)-rich DNA, were often of low quality. Replacing those thin, weak or ambiguous sequences with good-quality sequence was part of the finishing process. We manually scanned along the entire sequence of chromosome 12, examining both the quality and number of the individual reads and the quality of the consensus sequence. Wherever that quality was low, thin or ambiguous, we designed custom primers for the existing templates. The primers were paired with their appropriate templates for primer-extension sequencing. When this procedure failed, or when there were regions of poor-quality sequence on both strands, we constructed PCR products across the regions and sequenced these PCR products.

PCR products

Because of the very high (A + T) content of *P. falciparum* DNA, the annealing and extension temperatures for PCR reactions are significantly lower (and the extension time significantly longer) than the usual PCR reactions. These lower temperatures might allow slightly mismatched primer/template combinations to be stable and, therefore, amplified. In addition, because of the cost, finishing primers were not purified, so that oligonucleotides of related sequences might be present as contamination in the primer preparations. These related primers might have reasonable matches in the very complex *P. falciparum* genomic DNA template and, therefore, could contribute unwanted primer/template combinations that could be amplified. Therefore, we often found that the products of our PCR reactions were one major DNA product and several minor DNA products, as seen on agarose gels after electrophoresis. As such combinations of DNA do not sequence cleanly, all PCR products to be sequenced were LMP gel-purified.

Annotation

As part of the automated annotation process, the sequences of apparent ORFs were compared to the sequences in GenBank, using the BLAST program[17]. Positive quantitative results were posted. Then, we undertook an experiment in community annotation by inviting the world-wide scientific community to enter our website and annotate any particular ORF, or gene, or gene family, of their choice. At the time of writing, 18 scientists have annotated 52 genes. The participating annotators are: A. Danchin, C. Doerig, A. H. Fairlamb, P. Horrocks, J. E. Hyde, G. Plunkett, S. Rahlfs, P. Rathod, P. A. Rea, M. Seaman, C. Slomianny, J. Tyler, J. Kadonaga, C. Vaquero, C. Boschet, J. Vinetz, L. Wilming and M. F. Wiser. This pilot experiment in community annotation has been a modest, but real, success.

Received 14 June; accepted 9 September 2002; doi:10.1038/nature01102.

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. Hall, N. *et al.* Sequence of *Plasmodium falciparum* chromosomes 1, 3–9 and 13. *Nature* **419,** 527–531 (2002).
3. Gardner, M. J. *et al.* Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419,** 498–511 (2002).
4. Gardner, M. J. *et al.* Sequence of *Plasmodium falciparum* chromosomes 2, 10, 11 and 14. *Nature* **419,** 531–534 (2002).
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. Rubio, J. P., Thompson, J. K. & Cowman, A. F. The var genes of *Plasmodium falciparum* are located in the subtelomeric region of most chromosomes. *EMBO J.* **15,** 4069–4077 (1996).
8. Su, X. *et al.* A genetic map and recombination parameters of the human malaria parasite *Plasmodium falciparum*. *Science* **286,** 1351–1353 (1999).
9. Su, X. Z. & Wellems, T. E. *Plasmodium falciparum*: assignment of microsatellite markers to chromosomes by PFG-PCR. *Exp. Parasitol.* **91,** 367–369 (1999).
10. Jing, J. *et al.* Optical mapping of *Plasmodium falciparum* chromosome 2. *Genome Res.* **9,** 175–181 (1999).
11. Florens, L. *et al.* A proteomic view of the *Plasmodium falciparum* life cycle. *Nature* **419,** 520–526 (2002).

12. Lasonder, E. *et al.* Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry. *Nature* **419,** 537–542 (2002).
13. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8,** 186–194 (1998).
14. Oefner, P. J. *et al.* Efficient random subcloning of DNA sheared in a recirculating point-sink flow system. *Nucleic Acids Res.* **24,** 3879–3886 (1996).
15. Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8,** 175–185 (1998).
16. Gordon, D., Abajian, C. & Green, P. Consed: a graphical tool for sequence finishing. *Genome Res.* **8,** 195–202 (1998).
17. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
18. 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).
19. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **30,** 276–280 (2002).
20. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nature Genet.* **25,** 25–29 (2000).
21. 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).
22. Claros, M. G. & Vincens, P. Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur. J. Biochem.* **241,** 779–786 (1996).
23. 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).
24. 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).

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

Acknowledgements

We acknowledge the generosity of the participating scientists at Stanford University, TIGR, the WTSI, the NMRC and Oxford University. We also thank N. Hall, M. Berriman, A. Pain and B. Barrell for their time and expertise during the gene-calling annotation process, and are grateful to the members of our Stanford Genome Technology Center for their assistance throughout this project. We thank the Burroughs Wellcome Fund for support that allowed us to participate in the international Malaria Genome Project.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.W.H. (e-mail: hyman@sequence.stanford.edu). The GenBank accession number of the sequence of *P. falciparum* (clone 3D7) chromosome 12 is AEO14188.

Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry

Edwin Lasonder*†, Yasushi Ishihama*, Jens S. Andersen*, Adriaan M. W. Vermunt†, Arnab Pain‡, Robert W. Sauerwein§, Wijnand M. C. Eling§, Neil Hall‡, Andrew P. Waters‖, Hendrik G. Stunnenberg† & Matthias Mann*

* *Center for Experimental BioInformatics, Department of Biochemistry and Molecular Biology, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark*
† *Department of Molecular Biology, NCMLS, University of Nijmegen, Geert Grooteplein 26, 6525 GA Nijmegen, The Netherlands*
‡ *The Wellcome Trust Sanger Institute, The Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK*
§ *Department of Medical Microbiology, NCMLS, University Medical Centre, PO Box 9101, 6500 HB Nijmegen, The Netherlands*
‖ *Leiden Malaria Research Group, Department of Parasitology, Centre for Infectious Disease, Leiden University Medical Center, Albinusdreef 2, 2333 ZA Leiden, The Netherlands*

The annotated genomes of organisms define a 'blueprint' of their possible gene products. Post-genome analyses attempt to confirm and modify the annotation and impose a sense of the spatial, temporal and developmental usage of genetic information by the organism. Here we describe a large-scale, high-accuracy (average deviation less than 0.02 Da at 1,000 Da) mass spectrometric proteome analysis[1–3] of selected stages of the human malaria parasite *Plasmodium falciparum*. The analysis revealed 1,289 proteins of which 714 proteins were identified in asexual blood stages, 931 in gametocytes and 645 in gametes. The last two groups provide insights into the biology of the sexual stages of the parasite, and include conserved, stage-specific, secreted and membrane-associated proteins. A subset of these proteins contain domains that indicate a role in cell–cell interactions, and therefore can be evaluated as potential components of a malaria vaccine formulation. We also report a set of peptides with significant matches in the parasite genome but not in the protein set predicted by computational methods.

The *Plasmodium falciparum* parasite is pre-committed to one of three different developmental pathways on re-invasion of a host erythrocyte[4]. Either it develops asexually, resulting in proliferation, or it develops into a male or a female gametocyte—sexual precursor forms that maintain a stable G1 cell cycle arrest and circulate in the peripheral blood stream when mature. Gametocytes are activated in the mid-gut of the mosquito when ingested by the vector through the consumption of a blood meal, and rapidly develop into mature gametes that fertilize to form zygotes. The zygotes mature into a motile invasive form, the ookinete, which is adapted for colonization of the mosquito. Clearly, sexual development and fertilization are essential processes within the parasite life cycle and one strategy of vaccination (transmission blocking) seeks their interruption through antibody-based blockades. Although good candidates for such vaccines exist, it is an accepted view that an effective vaccine will need to target several stages of the parasite and several components of the different forms of the parasite[5]. High-throughput proteome studies on pure parasite forms are a rapid and sensitive means to discover such vaccine candidates.

To define the proteome of the asexual and sexual blood stages of the malaria parasite *P. falciparum* (NF54 isolate), purified asexual (trophozoites and schizonts, Fig. 1a, left panel) and sexual stage parasites (gametocytes, right panel) or gametes (not shown) were extracted by freeze–thawing and centrifugation, yielding soluble and insoluble (pellet) fractions (Fig. 1b). The result of a typical gametocyte extraction is shown in Fig. 1c, revealing that most of the *P. falciparum* and red blood cell (RBC) proteins were present in the soluble fraction, whereas membrane proteins such as the gametocyte-specific cell surface protein Pfs48/45 (ref. 6) were found exclusively in the pellet fraction, as revealed by western blotting (Fig. 1c). These complex protein mixtures were then analysed 'gel free' (differentially extracted membrane fractions) or separated into ten molecular mass fractions by one-dimensional gel electrophoresis followed by excision of equally spaced bands after precisely removing haemoglobin and globin (Fig. 1c, right panel), and tryptic digestion. The tryptic peptides were separated by reversed phase liquid chromatography coupled to quadrupole time-of-flight mass spectrometry for peptide sequencing (nanoLC-MS/MS). Iterative calibration algorithms were used to achieve a final, average absolute mass accuracy of better than 20 parts per million (p.p.m.) in both the precursor and fragment ions, or a mass deviation of 0.03 Da for a typical tryptic peptide of mass 1,300 Da.

These high-accuracy spectra were searched against a combined human and draft *P. falciparum* database, using probability-based scoring in which the fragment ions are matched against the calculated fragments of all tryptic peptides from the human and parasite sequences[7]. A total of 7,548 distinct peptides from the putative set of malaria proteins were matched with significant probability scores (Supplementary Table A). These peptides mapped to 1,709 malaria proteins. Additional constraints were applied to the peptides, including peptide size, discrimination to the next best match, and features of the tandem mass spectra such as

the dominant amino-terminal and inhibited carboxy-terminal fragmentation of proline. These criteria resulted in a list of 1,289 unique proteins that were identified with high confidence, comprising about 23% of the current proteome prediction[8] (Supplementary Table B). Roughly one-third of the malaria proteins identified appeared in both the soluble and insoluble fractions; the remaining two-thirds were present either in the insoluble or in the soluble fraction (Fig. 2a). Taken together, the analysis revealed 714 malaria proteins in asexual stage parasite preparations, 931 in gametocytes and 645 in gametes. The analyses also resulted in identification of more than 1,000 human proteins.

At present, 315 malaria proteins have been found solely in gametocytes, 226 in trophozoites and schizonts, and 97 in gametes (Fig. 2b). A total of 575 proteins were found only in sexual stages (gametocytes plus gametes) and 488 proteins were present in both asexual and sexual stages. The definitive classification of a protein as stage-specific is tentative however, and awaits full analysis of all life cycle stages. Moreover, biological samples are rarely completely pure or synchronous (see Methods). The well-characterized, sexual stage-specific surface antigen, Pfs48/45, highlights this point. Three Pfs48/45-derived peptides were detected in the asexual preparations as opposed to 57 in gametocytes. Integration of the peptide ion currents, a quantity roughly proportional to molar amount of peptide species (see Supplementary Fig. A), yielded a protein abundance ratio of more than eight between the two stages. The unexpected presence of Pfs48/45 in asexual stages is probably due to a small percentage of committed gametocytes in the asexual preparations from previous cycles of growth that cannot be physically separated from asexual-stage parasites. This interpretation is supported by western and northern blot analyses[9] and was further corroborated by quantitative polymerase chain reaction with reverse transcription (RT–PCR) analysis (Fig. 1c and Table 1). Notwithstanding the restrictions, the data are of sufficiently high quality and coverage to justify the classification of a large number of known and new proteins as 'stage-specific'. In fact, the small gametocyte contaminations in the asexual preparation strengthens the stage-specific classification of those proteins that are found exclusively in sexual stage parasites such as Pfg377, gene 11-1 and the hypothetical protein PF14_0039. A total of 517 peptides from the sexual stage antigen Pfg377 (ref. 10) were detected in the gametocytes and gamete preparations whereas not a single Pfg377 peptide was detected in the asexual stage, illustrating the exquisite sensitivity of the analytical procedure as well as the quality of the biological sample. Collectively, mass spectrometric and quantitative RT–PCR analysis (see below) indicates that expression of Pfg377, gene 11-1 and the hypothetical protein PF14_0039 is turned on later

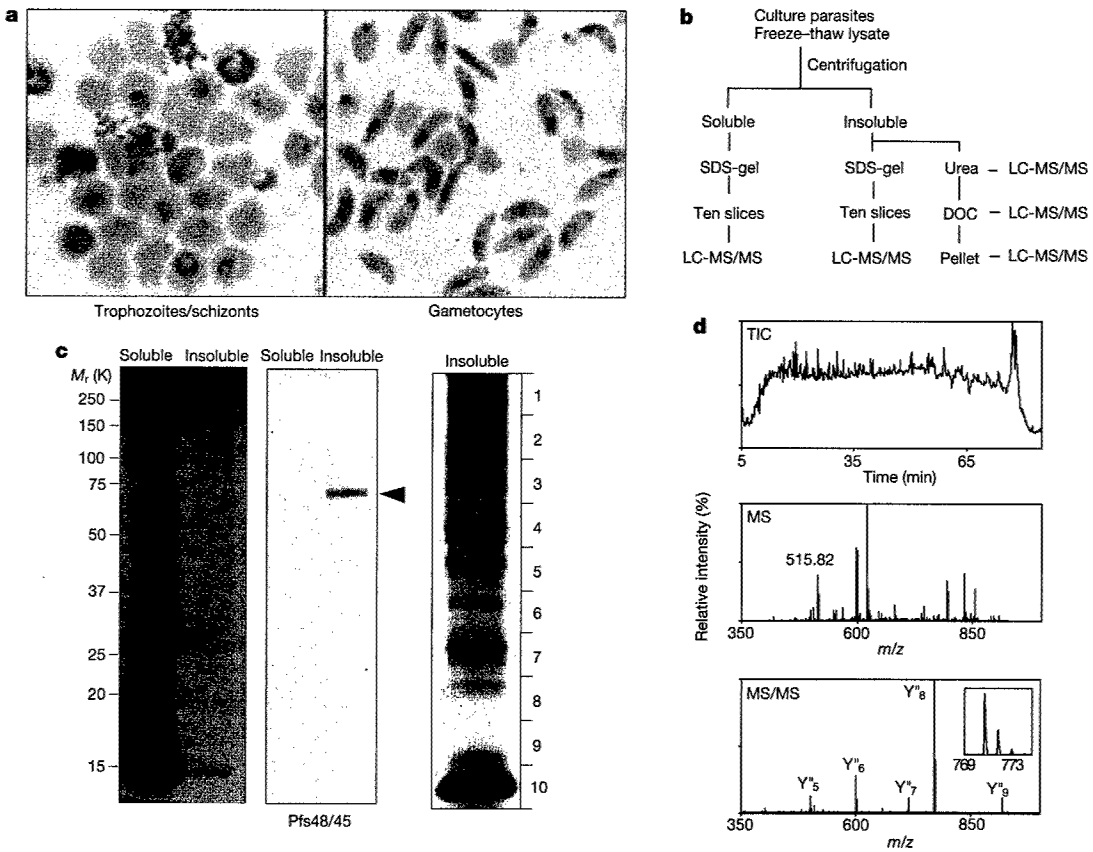


Figure 1 Differential extraction of parasite-infected red blood cells (RBCs) and flow chart of MS analysis. **a**, Giemsa staining of purified trophozoites and schizonts (left panel; asexual stage parasites), and gametocytes (right panel; sexual stage parasites). **b**, Extraction procedure for infected RBCs using freeze–thaw lyses and centrifugation, yielding a soluble and insoluble (pellet) fraction. **c**, The soluble and insoluble fraction from 5×10^5 gametocytes were separated on a 10% SDS gel and stained with Coomassie blue (left panel) or processed for western blotting and developed with a rabbit polyclonal antibody to Pfs48/45. The arrowhead indicates Pfs48/45. The right hand panel shows the insoluble fraction of 2×10^7 gametocytes. **d**, Mass spectrometric analysis of one of the gel slices shown in **c**. The top panel shows the summed ion current of all peptides eluting at a particular time. The middle panel shows a typical mass spectrum obtained from the eluting peptides. The bottom panel shows a fragmentation spectrum of the peptide with a mass of 515.82 indicated in the middle panel and magnification (inset) of one of the fragments. Database searching with this fragmentation information results in identification of peptide LFGIVGSIPK from Pfs48/45.

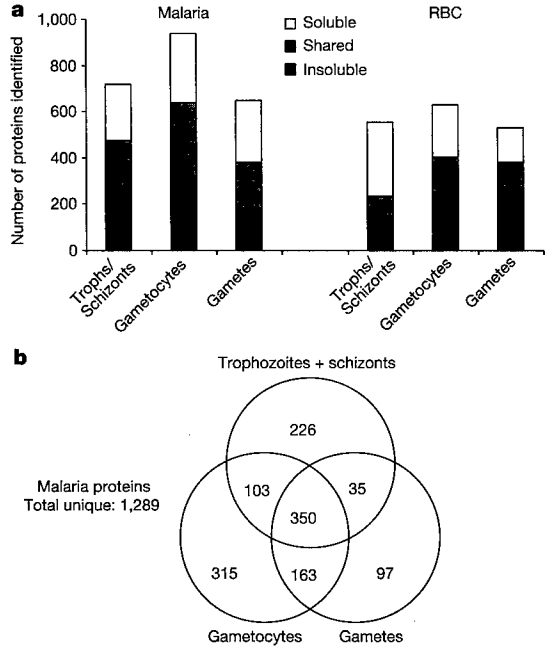


Figure 2 Schematic representation of proteomic data. **a**, Compilation of unique and common proteins in the soluble and insoluble fractions. Bars represent the number of proteins in the different preparations (trophozoites/schizonts, gametocytes and gametes) that were assigned to malaria or human RBCs. Proteins present solely in soluble (open bars) or insoluble (black bars) fractions as well as protein common to both fractions (shaded bars) are indicated. **b**, Venn diagram of the distribution of identified malaria proteins over the three blood stages. The distribution between trophozoites plus schizonts, gametocytes and gametes is shown.

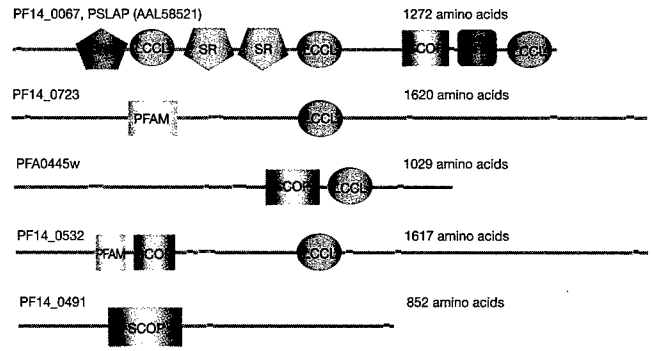


Figure 3 LCCL/lectin domain proteins expressed in sexual stages of *P. falciparum*. The SMART[22]-generated graphic descriptions of the five proteins expressed in sexual stages that contain LCCL/lectin domains are shown. The *P. falciparum* homologue of PSLAP[16] (GenBank accession number AAL58521), PF14_0067, contains a signal sequence (cleavage: amino acids, aa, 22 and 23); 3 LCCL domains (position: aa, 275–362, E value 1.28×10^{-3}; 670–676, E value 8.11×10^{-5}; and 1169–1255, E value 6.89×10^{-16}); 2 scavenger receptor (SR) domains (aa 401–515, E value 3.41×10^{-6} and 528–642, E value 3.10×10^{-7}); 1 lipoxygenase homology 2 (LH2) domain (aa 157–265, E value 8.20×10^{-3}); and one SCOP ConA lectin-like domain (aa 911–1013, E value 1×10^{-4}). PF14_0723 contains a signal peptide (cleavage: aa 19 and 20); 1 LCCL domain (aa 752–843, E value 1.4×10^{-21}); and a PFAM predicted discoidin domain (aa 296–420, E value 8.7×10^{-2}). PFA0445w contains a predicted signal sequence (cleavage: aa 24 and 25); 1 LCCL domain (aa 740–827, E value 4.49×10^{-6}); 1 SCOP kringle family domain (aa 43–95, E value 1.61×10^{-2}; not shown); and 1 SCOP dld7pm galactose-binding motif (aa 593–714, E value 4.01×10^{-3}). PF14_0532, also recognized as a gametocyte-specific transcript in *P. berghei* (AF491294), contains signal sequence (cleavage: aa 23 and 24); 1 LCCL domain (aa 724–815, E value 8.87×10^{-4}); and 1 PFAM ricin domain (aa 222–269, E value 4.0×10^{-2}). PF14_0491 is not predicted to contain a signal sequence but contains a SCOP dlacc anthrax protective antigen family with an immunoglobulin fold (aa 204–372, E value 3×10^{-9}) and SCOP/FN2 kringle-like (aa 30–80, E value 6.7×10^{-2}; not shown).

during gametogenesis than Pfs48/45.

Among the identified asexual proteins described previously are stage-specific antigens KAHRP, PfEMP3, PfSAR1 and PfSBP1, which are involved in the assembly of structures specific to the blood stage, such as knobs[11]. Schizont-specific proteins associated with the merozoite surface and invasive organelles such as MSP1, -2, -3, -6, -7 and -7a, or AMA1 and rhoptry-associated proteins RAP1, RAP2 and Rhop1 (for review of the proteins involved, see ref. 12) were also readily detected. Additionally, numerous hypothetical proteins containing a predicted signal sequence (61) and/or a transmembrane domain (189) were discovered, thus extending the repertoire of confirmed gene products and of potential new vaccine candidates (Supplementary Table B). Our analysis did not reveal any high-scoring peptides in the asexual blood stage preparations that could be assigned unambiguously to a member of the highly variable surface molecules of the PfEMP1 family, which are

Table 1 **Quantification of stage-specific proteins**

Protein name	Accession no.	RT–PCR*	MS-XIC†
Asexual stage-specific proteins (known)			
Apical membrane antigen 1 (AMA1)	PF11_0344	3 (1)	>2
Cyto-adherence-linked asexual protein (CLAG9)	PFI1730w	13 (3)	>5
Mature parasite-infected erythrocyte surface antigen (MESA)	PFE0040c	7 (4)	>12‡
Merozoite surface protein 1 (MSP1)	PFI1475w	66 (37)	>60
Cysteine protease	PFB0340c	221 (61)	>8
Sexual stage-specific proteins (known)			
Actin II	PF14_0124	26 (4)	>10
Pfs16 surface antigen	PF11_0318	8 (2)	>6
Sexual stage-specific surface antigen (Pfs48/45)	PF13_0247	43 (19)	8.2 (4.0)‡
Transmission-blocking target antigen (Pfs230)	PFB0405w	61 (20)	>6
Plasmodium falciparum gametocyte antigen (Pfg377)	PFL2405c	289 (79)	>20
Gene 11-1	PF10_0374	160 (14)	>8
Sexual stage antigen Pfs25	PF10_0303	858 (290)	>4
Sexual stage-specific proteins (novel)			
Hypothetical protein	PF14_0039	597 (208)	>17
Hypothetical protein	PF11_0310	9 (5)	>2
Hypothetical protein	PF11_0413	5 (1)	2.4 (0.5)

*Ratio between stages. RT–PCR normalized to heat shock protein 86, HSP86. Standard deviations are indicated in parentheses.
†Ratio between stages. Values were calculated from ion currents for the precise peptide molecular mass using maximum peak heights of peptides with more than 30 ion counts.
‡Extracted ion chromatogram (XIC) of the peptides of these proteins are presented in Supplementary Fig. A.

knob-associated and primarily responsible for cyto-adherence of parasites to the peripheral vascular endothelium. PfEMP1 may be expressed at very low abundance in *P. falciparum* strain NF54, or poorly extracted by our methods. Our mass spectrometry analyses revealed almost all of the known proteins specific to *P. falciparum* gametocytes and gametes (Supplementary Table B). For example, four of the six members of the P48/45 family that are transcribed in sexual stages were readily identified. In addition we detected the sexual stage-specific Pfs48/45 paralogue, Pf47, for the first time.

To corroborate and extend the stage specificity, we compared not only the absence or presence of particular peptides in the respective fractions, but also integrated the ion currents of peptides (see Supplementary Fig. A). This is important because it is possible that a peptide is not selected for sequencing owing to the complexity of the sample, but is nevertheless present at significant amounts. For example, MESA was found with 18 peptides in the asexual stage. The ion currents at the corresponding elution times in the preparation of a sexual stage were inspected but did not show any signal at the respective peptide masses. Similarly, analysis of peptide ion currents corresponding to hypothetical proteins such as PF14_0039 revealed a protein abundance ratio of at least 17:1 between the sexual and asexual stages. We next performed quantitative RT–PCR using messenger RNA from parallel asexual, gametocyte and gamete parasite preparations and gene-specific primer sets for a group of hypothetical and known proteins and an arbitrary selection of putative sexual and asexual stage-specific proteins. RT–PCR ratios of signals obtained in asexual versus sexual parasites ranged from 3-fold to over 850-fold (Table 1). In accord with the mass spectrometric findings, mRNA from Pfg377, gene 11-1 and the protein PF14_0039, but not Pfs48/45, were detected exclusively in gametocyte and gamete preparations. Comparison between the results obtained with mass spectrometry and quantitative RT–PCR shows that the changes observed in protein abundance between the asexual and sexual stages were also reflected in their mRNA levels.

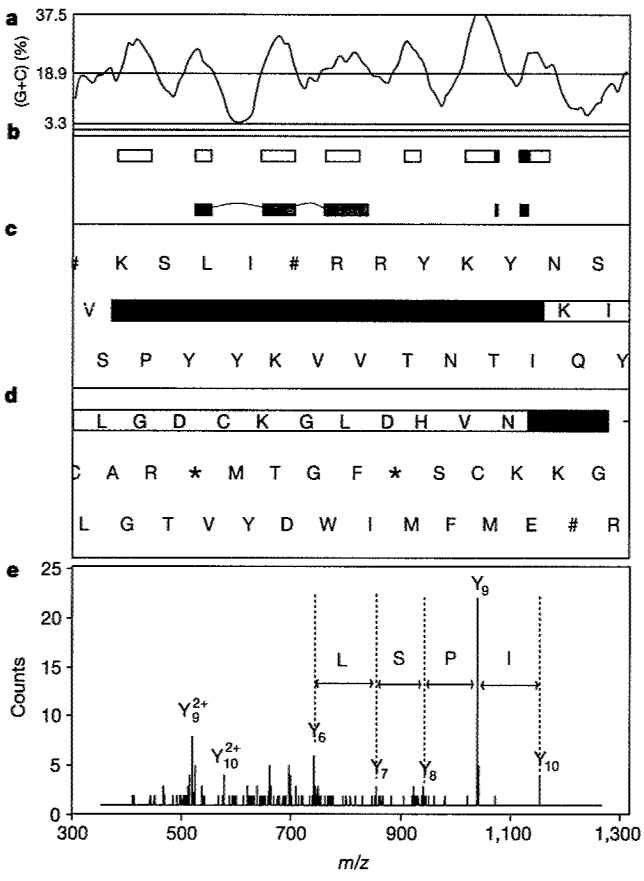


Figure 4 Refinement of gene structure and re-annotation using proteomic data. **a**, (G + C) plot in an ARTEMIS window corresponding to the genomic region of the gene PF11_0169 on chromosome 11 of *P. falciparum*. **b**, Alternative gene models for PF11_0169. The green and yellow gene models represent the original and refined gene structure of PF11_0169, respectively. The red blocks represent the peptide LQIPSLNIIQVR and its occurrence in respect to the green and the yellow gene models. **c**, **d**, Sections from an ARTEMIS window showing the splice-donor site of exon I and the splice-acceptor site of exon II of the yellow gene model. The three-frame translation is shown with hash or asterisk, denoting stop codons. The red blocks together represent the full-length peptide LQIPSLNIIQVR. **e**, Tandem mass spectrum of the orphan peptide LQIPSLNIIQVR. C-terminal fragment ions (Y ions) are indicated as well as a partial sequence.

Erythrocytic stages of the parasites are under enhanced oxidative stress and are particularly vulnerable to exogenous challenges by reactive oxygen species. Therefore it is interesting to note that proteins involved in protection against oxidative stress—the glyoxalase enzymatic system associated with glutathione—seem to be upregulated in gametes and gametocytes, such as glyoxalases I (PF11_0145) and II (PFL0285w), and two different glutaredoxins (MAL6P1.72 and PFC0271c)[13]. Some of the enzymes involved in the synthesis and metabolism of glutathione such as glutathione *S*-transferase (PF14_0187) and γ-glutamylcysteine synthetase (PFI0925w) are exclusively present in sexual stages. Others were found in sexual as well as asexual stages; these include glutathione peroxidase (PFL0595c) and glutathione reductase (PF14_0192). Furthermore, thioredoxin systems are also readily found in sexual stages (thioredoxin, PF14_0545 and PF13_0272 (secreted); thioredoxin peroxidase, PFL0725w; and thioredoxin reductase, PFL0725w).

Another stage-specific feature is the motility of male gametes. This involves the axoneme, a specialized structure consisting of two parallel microtubules thought to be connected by dynein complexes, which so far have only been demonstrated in mature merozoites[14]. The full *P. falciparum* genome reveals 12 genes that are expected to encode different dynein forms. They are conserved throughout *Plasmodium* species and have maximum homology to flagellar dyneins (for example, PF11_0240). Our analysis indicates that at least six of these (three light and three heavy chains) are expressed exclusively in sexual stages. Further specialized motility-associated proteins that may assist in the formation of the axoneme, such as actin II, α-tubulin II, β-tubulin, putative kinesins, as well as predicted tubulin-specific chaperones, are also evident.

The male gamete has been shown to bind to sialic acid on erythrocytes through a lectin-like activity[15]. Therefore, one anticipated finding within the proteome of the sexual stage was the presence of proteins (protein families) that exhibit appropriate adhesion properties. Our analysis revealed five proteins expressing predicted lectin domains of which four contain an additional LCCL motif thought to be involved in protein–protein interactions—all five are expressed exclusively in sexual stages (Fig. 3). One of these proteins, PSLAP (PF14_0067; AAL58521), has been characterized previously as gametocyte-specific and has been shown to contain multiple different domains suggestive of a role in cell–cell interactions[16]. A second (PF14_0723) has been characterized as a gametocyte-specific transcript in the rodent parasite *Plasmodium berghei* (GenBank accession number AF491294). Four of the five are predicted to be secreted proteins (PF14_0723, PFA0445w, PF14_0532, PSLAP). A fifth protein (PF14_0491) contains only a lectin domain and is predicted to be non-secreted; however, we believe that this is due to annotation based on a short exon-splice model, because alternative models would create a signal peptide. Interrogation of the *Plasmodium* genomes (*P. falciparum* and *P. y. yoelii*) suggests that these proteins are conserved in both models

and in other *Plasmodium* species that infect humans, and that expression of LCCL-domain-containing proteins is restricted to the parasite stages found in the mid-gut of the mosquito. Future investigations will determine the precise role of these proteins in the complexity of gamete fertilization, erythrocyte binding by male gametes, interactions with the host and vector immune systems, and the surface carbohydrate structures of the mosquito mid-gut.

The gel slice approach chosen for proteome analysis allowed us to correlate apparent molecular masses with gene annotation derived from theoretical mass predictions, revealing outliers that may be due to protein processing or to incomplete gene annotation. Protein Pfg377, for example, probably represents a case of protein processing (Supplementary Fig. B). The C-terminal portion of Pfg377 was covered with 69 sequenced peptides solely in the protein mixture from gel slice 5 (relative molecular mass (M_r) range of 100,000 to 140,000 (100K–140K)), whereas 15 peptides matching the N-terminal portion were found exclusively in gel 7 (65K–84K). Northern blot analysis[10] and quantitative RT–PCR using 5′ and 3′ specific primer sets (data not shown) indicate that the proteins are encoded by a single gene that is processed after translation.

The parasite genome is highly (A + T)-rich and has proven difficult to assemble and annotate[8]. For this reason, we also performed direct genome searches[17] where the malaria genome rather than the predicted set of proteins was searched by the mass spectrometric data. A large number of additional hits emerged. After analysing these 'orphan' peptide hits in the genome of *P. falciparum*, we were able to identify additional exons or assign different exon–intron boundaries of previously annotated genes using the ARTEMIS annotation tool[18]. One such orphan peptide, LQIPSLNIIQVR, identified two additional exons of the gene PF11_0169 on chromosome 11, annotated as a hypothetical protein. With this information we were able to identify two further exons and modify the gene structure appropriately. This led to re-annotation of the previously annotated hypothetical protein PF11_0169 as a putative homologue of sno-type pyridoxin biosynthesis protein (Fig. 4). The orphan peptide not only identified the two exons of the gene PF11_0169 but also verified the boundaries of the intron between the exons I and II. With the refined gene model we were able to re-annotate the gene with appropriate Gene Ontology (GO) terms. Initial analysis similar to the one outlined above led to eight definite and six probable gene annotation changes. Supplementary Table C lists more than 100 high-quality orphan peptides that can be used in gene annotation and the finding of new genes.

We have applied state-of-the-art proteomics techniques to obtain valuable information about stage-specific expression and localization of malaria proteins as well as information useful in confirming and extending the bioinformatics analysis of the proteome. Further work could encompass additional stages of the parasite life cycle and make use of the rapidly evolving mass spectrometric technology, associated bioinformatics, as well as direct comparison of stages using stable isotope techniques. □

Methods

Preparation of parasites

Plasmodium falciparum parasites were cultured using a semi-automated culture system. Asexual stages (trophozoites and schizonts) were purified using the Variomacs system essentially as described[14]. For the production of gametocytes, parasites were cultured for 14 days without addition of red blood cells. The gametocyte cultures were treated with 50 mM *N*-acetyl-glucosamine from day 8 to 12 after the start of the culture to remove asexual stage parasites. After a total of 14 days culture the gametocytes were collected and purified using the Variomacs system as described[14]. For the production of gametes, purified gametocytes were pelleted and incubated in G_m buffer containing 100 μM xanthurenic acid for a period of 3 h at 21 °C (ref. 19). (G_m buffer contains 1.67 mg ml^{-1} glucose, 8 mg ml^{-1} NaCl, 1 mg ml^{-1} Tris, pH 8.1.) Every effort was made to minimize enzymatic activity and protein degradation during sampling and the subsequent isolation of the parasites. However, we cannot exclude that some of the differences in protein profiles that we observe between the different life-cycle stages may be a consequence of the sample-handling procedures.

Sample preparation

The infected red blood cells (10^7 cells) of each stage were divided into a soluble and insoluble fraction by freeze–thawing several times and pelleted by centrifugation at 13,000 r.p.m. (16,100 *g*). Proteins were extracted in SDS–polyacrylamide gel electrophoresis (PAGE) loading buffer and separated into ten fractions on a 10% protein gel. The separated proteins were treated with dithiothreitol (DTT) and iodoacetamide, and in-gel digested with trypsin[20]. Proteins were also extracted by 8 M urea, 100 mM Tris-HCl, pH 8.0, treated with DTT and iodoacetamide, and digested in-solution with endoprotease Lys-C and trypsin after dilution to 2 M urea. Proteins from the insoluble fractions were further extracted with 1% deoxycholic acid (DOC), 100 mM Tris-HCl, pH 8.0, and digested together with the remaining pellet as above in the presence of 0.05% DOC.

NanoLC-MS/MS analysis of malaria proteins

Peptide mixtures were loaded onto 75-μm ID columns packed with 3-μm C18 particles (Vydac) and eluted into a quadruple time-of-flight mass spectrometer (QSTAR, Sciex-Applied Biosystems). Fragment ion spectra were recorded using information-dependent acquisition and duty-cycle enhancement (see ref. 21 and references therein for a more detailed description of the method). The three malaria stages studied were analysed at least in duplicate. In total, more than 100 nanoLC-MS/MS runs were analysed, yielding more than 200,000 peptide-sequencing events.

Data analysis

Peak lists containing the precursor masses and the corresponding MS/MS fragment masses were generated from the original data file and searched in the annotated *P. falciparum* database (Sanger/TIGR) combined with the human IPI database (European Bioinformatics Institute) using the Mascot program (Matrix Science). Identified peptides that were not unique or had a score less than 20 were removed. Proteins identified with a combined peptide score of higher than 60 were considered significant, and lower scoring proteins were manually verified or rejected. Iterative calibration algorithms on the basis of identified peptides were used to achieve a final average absolute mass accuracy of better than 20 p.p.m. in both the precursor and fragment ions. Relative protein abundance between the malaria stages was based on the total number of unique peptides identified for each protein and was further supported by extracted ion chromatograms (XIC) for individual peptides within an elution time window of 3 min (see also Supplementary Fig. A).

Quantitative real-time RT–PCR

Relative mRNA abundance was quantified by real-time PCR with the GeneAmp 5700 Sequence Detection System (Applied Biosystems) using the SYBR Green PCR Master Mix kit (Applied Biosystems). Total RNA was isolated from two stages: asexual blood stage (trophozoites and schizonts) and the sexual blood stage (gametocytes), both prepared as described above. RNA was isolated from 4×10^7 parasites using TRIzol reagent (Gibco BRL). A total of 2 μg of RNA was used for complementary DNA synthesis. RNA was treated with DNaseI (Pharmacia) after which cDNA was synthesized with random hexamers and Superscript II enzyme (Gibco BRL), according to standard protocols. cDNA was dissolved in 50 μl diethyl pyrocarbonate (DEPC)-treated H_2O. A total of 3 μl of × 10 and × 100 diluted cDNA was used for real-time PCR. Primers were designed by Primer Express software (Applied Biosystems). The mRNA levels for each gene were normalized against heat shock protein 86. For each gene the relative mRNA abundance was determined by calculating the ratio of asexual:gametocyte stage and gametocyte:asexual stage.

Received 31 July; accepted 9 September 2002; doi:10.1038/nature01111.

1. Pandey, A. & Mann, M. Proteomics to study genes and genomes. *Nature* **405,** 837–846 (2000).
2. Link, A. J. *et al.* Direct analysis of protein complexes using mass spectrometry. *Nature Biotechnol.* **17,** 676–682 (1999).
3. Griffin, T. J. & Aebersold, R. Advances in proteome analysis by mass spectrometry. *J. Biol. Chem.* **276,** 45497–45500 (2001).
4. Bruce, M. C., Alano, P., Duthie, S. & Carter, R. Commitment of the malaria parasite *Plasmodium falciparum* to sexual and asexual development. *Parasitology* **100,** 191–200 (1990).
5. Richie, T. L. & Saul, A. Progress and challenges for malaria vaccines. *Nature* **415,** 694–701 (2002).
6. Kocken, C. H. *et al.* Cloning and expression of the gene coding for the transmission blocking target antigen Pfs48/45 of *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **61,** 59–68 (1993).
7. Perkins, D. N., Pappin, D. J., Creasy, D. M. & Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20,** 3551–3567 (1999).
8. Gardner, M. J. *et al.* Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419,** 498–511 (2002).
9. van Dijk, M. R. *et al.* A central role for P48/45 in malaria parasite male gamete fertility. *Cell* **104,** 153–164 (2001).
10. Alano, P. *et al.* COS cell expression cloning of Pfg377, a *Plasmodium falciparum* gametocyte antigen associated with osmiophilic bodies. *Mol. Biochem. Parasitol.* **74,** 143–156 (1995).
11. Wickham, M. E. *et al.* Trafficking and assembly of the cytoadherence complex in *Plasmodium falciparum*-infected human erythrocytes. *EMBO J.* **20,** 5636–5649 (2001).

12. Cowman, A. F. *et al.* Functional analysis of proteins involved in *Plasmodium falciparum* merozoite invasion of red blood cells. *FEBS Lett.* **476,** 84–88 (2000).
13. Rahlfs, S., Fischer, M. & Becker, K. *Plasmodium falciparum* possesses a classical glutaredoxin and a second, glutaredoxin-like protein with a PICOT homology domain. *J. Biol. Chem.* **276,** 37133–37140 (2001).
14. Fowler, R. E. *et al.* Microtubule associated motor proteins of *Plasmodium falciparum* merozoites. *Mol. Biochem. Parasitol.* **117,** 187–200 (2001).
15. Templeton, T. J., Keister, D. B., Muratova, O., Procter, J. L. & Kaslow, D. C. Adherence of erythrocytes during exflagellation of *Plasmodium falciparum* microgametes is dependent on erythrocyte surface sialic acid and glycophorins. *J. Exp. Med.* **187,** 1599–1609 (1998).
16. Delrieu, I. *et al.* PSLAP, a protein with multiple adhesive motifs, is expressed in *Plasmodium falciparum* gametocytes. *Mol. Biochem. Parasitol.* **121,** 11–20 (2002).
17. Kuster, B., Mortensen, P., Andersen, J. S. & Mann, M. Mass spectrometry allows direct identification of proteins in large genomes. *Proteomics* **1,** 641–650 (2001).
18. Rutherford, K. *et al.* Artemis: sequence visualization and annotation. *Bioinformatics* **16,** 944–945 (2000).
19. Brooks, S. R. & Williamson, K. C. Proteolysis of *Plasmodium falciparum* surface antigen, Pfs230, during gametogenesis. *Mol. Biochem. Parasitol.* **106,** 77–82 (2000).
20. Shevchenko, A., Wilm, M., Vorm, O. & Mann, M. Mass spectrometric sequencing of proteins from silver stained polyacrylamide gels. *Anal. Chem.* **68,** 850–858 (1996).
21. Rappsilber, J., Ryder, U., Lamond, A. I. & Mann, M. Large-scale proteomic analysis of the human spliceosome. *Genome Res.* **12,** 1231–1245 (2002).
22. Schultz, J., Milpetz, F., Bork, P. & Ponting, C. P. SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl Acad. Sci. USA* **95,** 5857–5864 (1998).

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

Acknowledgements

We thank our colleagues for help and discussions. This work was supported by the Danish National Research Foundation, the Dutch Science Foundation (NWO), the European Union and the World Health Organization (WHO) Special Program for Research and Training in Tropical Diseases.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to H.G.S. (e-mail: h.stunnenberg@ncmls.kun.nl) or M.M. (e-mail: mann@bmb.sdu.dk). Sequence data for the genes newly annotated according to the present study can be found at http://www.sanger.ac.uk/Projects/P_falciparum.

nature

Reprinted from Vol. 419, no. 6906, 3 October 2002

Plasmodium genome

Appendix C
Final Report
DAMD17-82-2-8005

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].

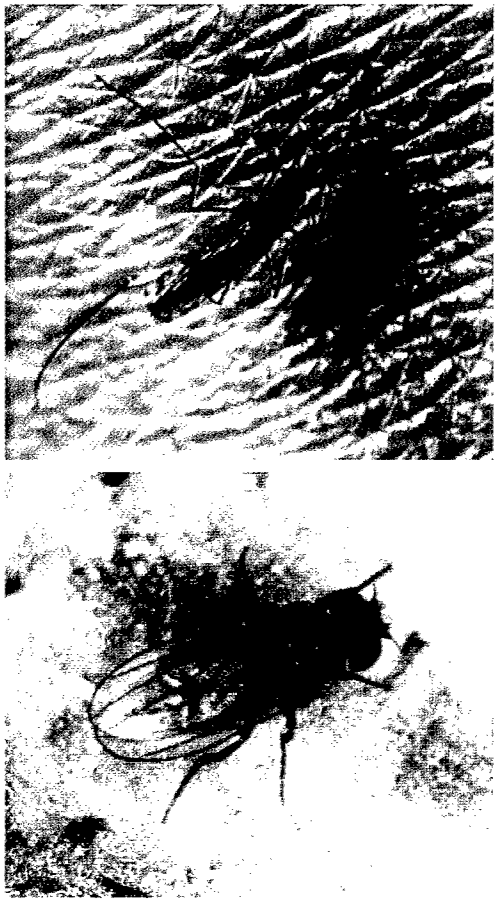
An international effort[4] was launched in 1996 to sequence the *P. falciparum* genome with the expectation that the genome sequence would open new avenues for research. The sequences of two of the 14 chromosomes, representing 8% of the nuclear genome, were published previously[5,6] and the accompanying 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.).



INSTITUT PASTEUR/P. BREY

E. DE GREGORIO

Figure 1 **The mosquito and the fruitfly in typical pose — *Anopheles* (top) on human skin, *Drosophila* on a banana.**

has been to block parasite transmission by mosquitoes. These approaches will clearly benefit from the improved understanding of mosquito biology and mosquito interactions with *P. falciparum* that the genome sequences will make possible.

The *A. gambiae* genome[1] was sequenced by a collaboration between Celera Genomics, the French National Sequencing Centre (Genoscope) and The Institute for Genomics Research (TIGR), in association with several university laboratories. These groups used the same 'shotgun' strategy as that applied for sequencing the human, mouse and fruitfly (*Drosophila melanogaster*) genomes. Random fragments of genomic DNA were first cloned in bacteria, and sequenced, and the overlapping clones were then assembled into contiguous sequences. Unexpectedly, the high levels of genetic variation (polymorphisms) in the reference strain of *A. gambiae* used for sequencing — the PEST strain — made the genomic assembly step difficult. The genetic variation might be explained by the fact that two distinct populations of *A. gambiae* have contributed to the PEST strain, thereby creating a mosaic genome structure. This unprecedented situation required the development of new sequence-assembly strategies, and these will be a considerable asset for future genome projects — as with mosquitoes, not all organisms are available as inbred laboratory strains.

Comparisons with the fruitfly

Much of the interest in the *A. gambiae* genome will centre on comparisons with that of *D. melanogaster*, which was published two years ago[2]. These two insects belong to the same taxonomic order, the Diptera, but inhabit distinct environments and have different lifestyles (Fig. 1). *Drosophila melanogaster* feeds on decaying organic matter, such as damaged or rotting fruit, where it also completes its life cycle, whereas *A. gambiae* feeds on sugar nectar and on the blood of vertebrate hosts. Blood meals are required for female mosquitoes to produce eggs; these are laid in water, where larvae develop and hatch. Blood feeding exposes the insect to viruses and parasites — like *Plasmodium*, these other pathogens exploit *Anopheles* as a vector for transmission.

One of the main differences between the two species is that, at 278 million base pairs, the *A. gambiae* genome is much bigger than that of *D. melanogaster* (estimated to be 180 million base pairs). But this difference is not reflected in the total number of genes, which, with 13,000–14,000 genes so far identified in both insects, is surprisingly similar. It seems that, in the course of evolution, *Drosophila* has experienced a progressive reduction both in the regions between genes and in the introns, the non-protein-coding stretches of DNA within genes.

Comparison of the coding sequences reveals that the genomes of *Anopheles* and *Drosophila* are less similar than would be expected for two species that diverged 'only' 250 million years ago. Only half of the genes in the two genomes can be interpreted as orthologues — genes in different species that have common ancestry, although their functions may differ. *Anopheles* and *Drosophila* orthologues show an average of about 56% identity in DNA sequence. As Zdobnov *et al.* point out in another of the papers in *Science*[3], from the sequence standpoint, the two species differ more than do humans and pufferfish — species that diverged 450 million years ago. Some of the protein families present in both mosquito and fruitfly appear to have evolved from a common ancestral gene through independent gene-duplication in each species. The *Anopheles* genome shows several cases of such expansion which might reflect adaptation to its lifestyle. An example is the family of fibrinogen-like proteins (of which there are 58 in *Anopheles* and 13 in *Drosophila*), which in the mosquito are probably used as anticoagulant for the ingested blood meals.

Insect immunity

Insects have efficient immune systems for combating the various pathogens they encounter, and most of our knowledge in this area comes from genetic and molecular studies in *Drosophila*. Finding out how *Anopheles* responds to *Plasmodium* infection is essential for obtaining clues to controlling malaria. Christophides *et al.*[4] analysed the gene families in *A. gambiae* that are linked to insect immunity, and show that they diverge widely from those in *Drosophila*. Good examples are the prophenoloxidase enzymes (nine in the mosquito, three in the fruitfly); these enzymes catalyse the synthesis of melanin, which is associated with several defence reactions in insects.

The study by Christophides *et al.* suggests that *Anopheles* employs the same general defence mechanisms as *Drosophila*, and uses similar pathogen-activated signal-transduction pathways, but that it has adapted recognition and effector immune genes to different types of aggressors. The best characterized effector system in insects consists of antimicrobial peptides, which display a wide spectrum of antibiotic activities. Interestingly, out of seven families of these peptides found in *Drosophila*, only two are also evident in *Anopheles*: five, then, are specific to *Drosophila*. Conversely, at least one mosquito-specific antimicrobial peptide has already been identified and others might be discovered by functional studies in the future. The expression profiles of some *A. gambiae* immune genes also suggest that, like the fruitfly, the mosquito mounts specific immune responses adapted to different types of pathogen[4,5].

The availability of the entire DNA sequence, together with tools such as DNA microarrays and targeted gene disruption[6–8], will make *Anopheles* a powerful model system for studying insect biology. The genomic data will also help in developing strategies to combat malaria and other mosquito-borne human diseases, for example yellow fever, dengue, filariasis and encephalitis. Such strategies will include reducing the number and lifespan of infectious mosquitoes, analysing what attracts them to their human targets, and limiting the capacity of parasites to develop within the insect vector. Malaria is characterized by a highly complex set of interactions between the parasite, the vector and the host. Now that the genomes of all three players have been fully sequenced, the post-genomic era in combating this dreadful disease can really begin.

Ennio De Gregorio and Bruno Lemaitre are at the Centre de Génétique Moléculaire, CNRS, 91198 Gif-sur-Yvette, France.
e-mails: gregorio@cgm.cnrs-gif.fr
lemaitre@cgm.cnrs-gif.fr

1. Holt, R. A. *et al. Science* **298**, 129–149 (2002).
2. Special section on the *Drosophila* genome *Science* **287**, 2181–2224 (2000).
3. Zdobnov, A. M. *et al. Science* **298**, 149–159 (2002).
4. Christophides, G. K. *et al. Science* **298**, 159–165 (2002).
5. Dimopoulos, G. *et al. Proc. Natl Acad. Sci. USA* **99**, 8814–8819 (2002).
6. Catteruccia, F. *et al. Nature* **405**, 959 (2000).
7. Grossman, G. L. *et al. Insect Mol. Biol.* **10**, 597–604 (2001).
8. Blandin, S. *et al. EMBO Rep.* **3**, 852–856 (2002).

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).

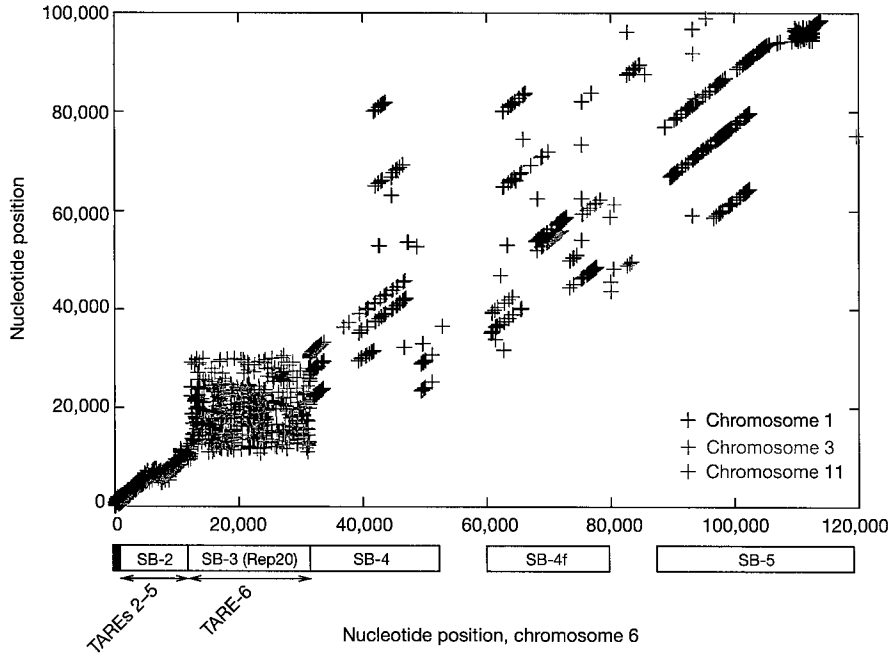


Figure 2 Alignment of subtelomeric regions of chromosomes 1, 3, 6 and 11. MUMmer2[152] alignments showing exact matches between the left subtelomeric regions of chromosome 6 (horizontal axis) and chromosomes 11 (red), 1 (blue) and 3 (green), illustrating the conserved synteny between all telomeres. Each 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

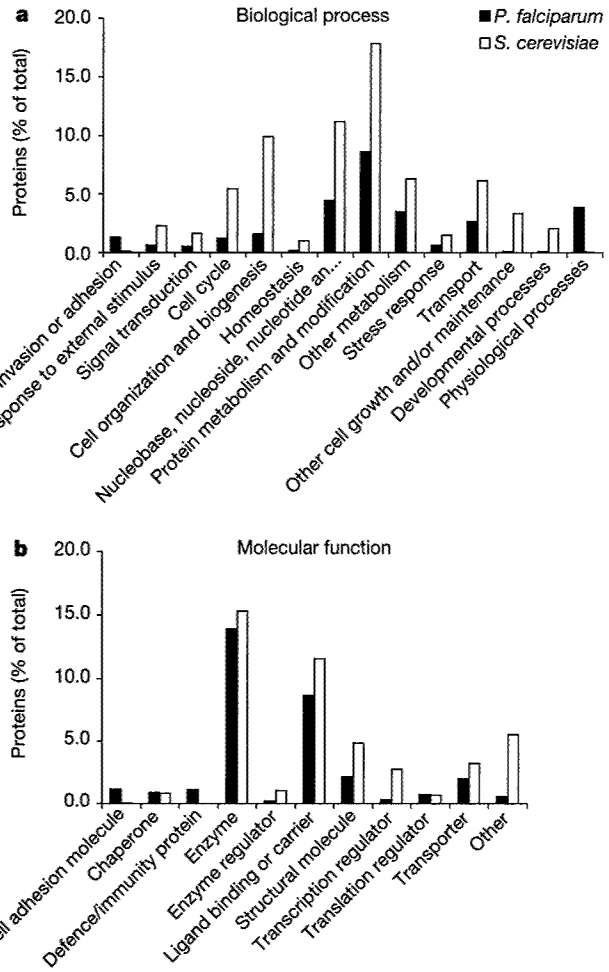


Figure 3 Gene Ontology classifications. Classification of *P. falciparum* proteins according to the 'biological process' (**a**) and 'molecular function' (**b**) ontologies of the Gene Ontology system[42].

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

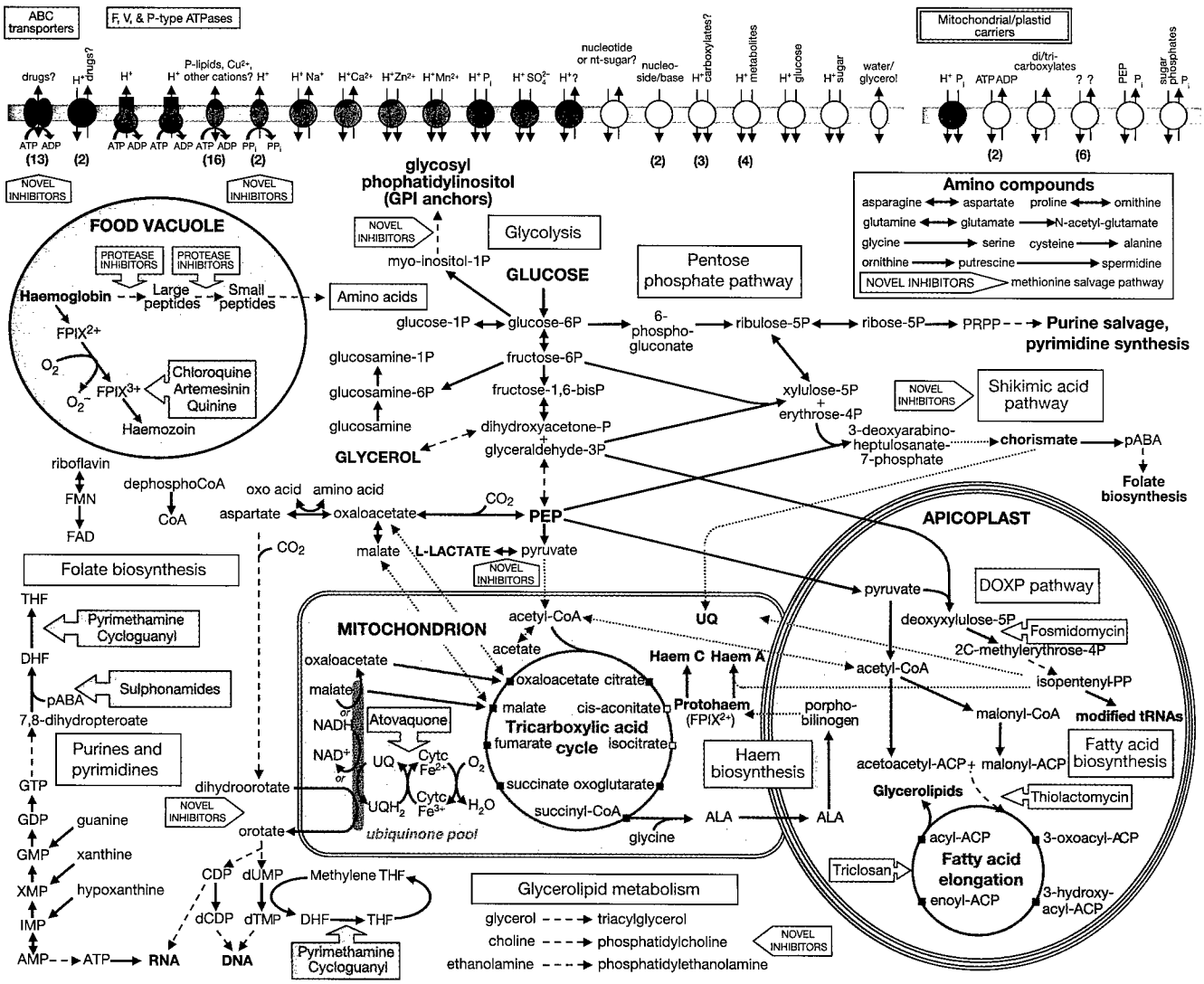


Figure 4 Overview of metabolism and transport in *P. falciparum*. Glucose and glycerol provide the major carbon sources for malaria parasites. Metabolic steps are indicated by arrows, with broken lines indicating multiple intervening steps not shown; dotted arrows indicate incomplete, unknown or questionable 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].

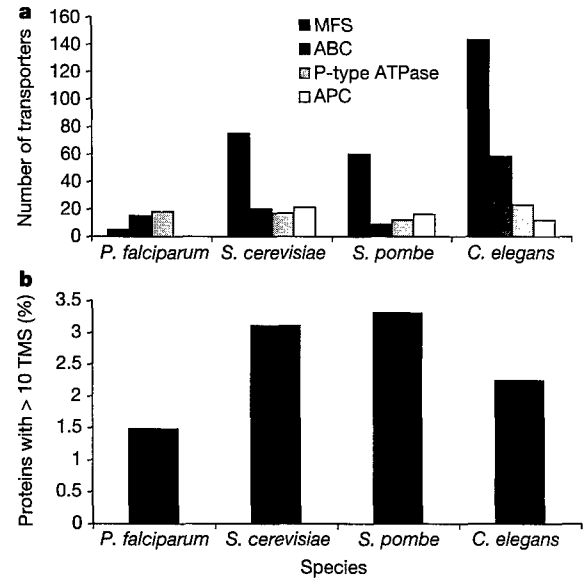


Figure 5 Analysis of transporters in *P. falciparum*. **a**, Comparison of the numbers of transporters belonging to the major facilitator superfamily (MFS), ATP-binding cassette (ABC) family, P-type ATPase family and the amino acid/polyamine/choline (APC) family in *P. falciparum* and other eukaryotes. 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

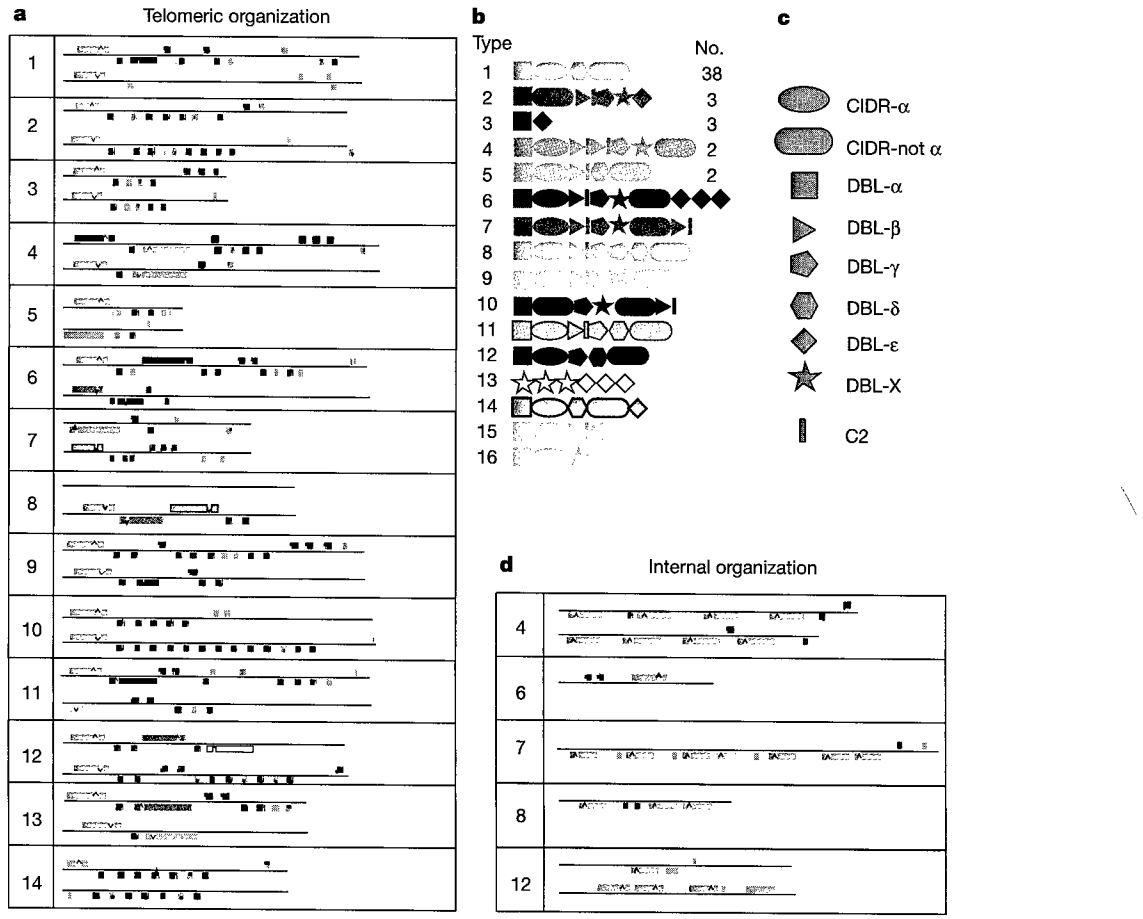


Figure 6 Organization of multi-gene families in *P. falciparum*. **a**, Telomeric regions of all chromosomes showing the relative positions of members of the multi-gene families: *rif* (blue) *stevor* (yellow) and *var* (colour coded as indicated; see **b** and **c**). Grey boxes represent pseudogenes 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).

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

Annotation

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

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

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

1. Gardner, M. J. *et al.* Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science* **282,** 1126–1132 (1998).
2. Bowman, S. *et al.* The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*. *Nature* **400,** 532–538 (1999).
3. Su, X. *et al.* A genetic map and recombination parameters of the human malaria parasite *Plasmodium falciparum*. *Science* **286,** 1351–1353 (1999).
4. Lai, Z. *et al.* A shotgun optical map of the entire *Plasmodium falciparum* genome. *Nature Genet.* **23,** 309–313 (1999).
5. de Bruin, D., Lanzer, M. & Ravetch, J. V. Characterization of yeast artificial chromosomes from *Plasmodium falciparum*: construction of a stable, representative library and cloning of telomeric DNA fragments. *Genomics* **14,** 332–339 (1992).
6. Glockner, G. *et al.* Sequence and analysis of chromosome 2 of *Dictyostelium discoideum*. *Nature* **418,** 79–85 (2002).
7. Piper, M. B., Bankier, A. T. & Dear, P. H. A HAPPY map of *Cryptosporidium parvum*. *Genome Res.* **8,** 1299–1307 (1998).
8. Konfortov, B. A., Cohen, H. M., Bankier, A. T. & Dear, P. H. A high-resolution HAPPY map of *Dictyostelium discoideum* chromosome 6. *Genome Res.* **10,** 1737–1742 (2000).
9. Berriman, M., Aslett, M. & Ivens, A. Parasites are GO. *Trends Parasitol.* **17,** 463–464 (2001).
10. Florens, L. *et al.* A proteomic view of the *Plasmodium falciparum* life cycle. *Nature* **419,** 520–526 (2002).
11. Pachebat, J. A. *et al.* The 22 kDa component of the protein complex on the surface of *Plasmodium falciparum* merozoites is derived from a larger precursor, merozoite surface protein 7. *Mol. Biochem. Parasitol.* **117,** 83–89 (2001).
12. Lasonder, E. *et al.* Analysis of the *Plasmodium falciparum* proteome by high accuracy mass spectrometry. *Nature* **419,** 531–542 (2002).
13. Figueiredo, L. M., Freitas-Junior, L. H., Bottius, E., Olivo-Marin, J. C. & Scherf, A. A central role for *Plasmodium falciparum* subtelomeric regions in spatial positioning and telomere length regulation. *EMBO J.* **21,** 815–824 (2002).
14. O'Donnell, R. A. *et al.* A genetic screen for improved plasmid segregation reveals a role for Rep20 in the interaction of *Plasmodium falciparum* chromosomes. *EMBO J.* **21,** 1231–1239 (2002).
15. Katinka, M. D. *et al.* Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature* **414,** 450–453 (2001).
16. Hyman, R., Fung, E. & Dennis, R. W. *et al.* Sequence of *Plasmodium falciparum* chromosome 12. *Nature* **419,** 534–536 (2002).
17. Hapgood, J. P., Riedemann, J. & Scherer, S. D. Regulation of gene expression by GC-rich DNA cis-elements. *Cell Biol. Int.* **25,** 17–31 (2001).
18. Adhya, S. Multipartite genetic control elements: communication by DNA loop. *Annu. Rev. Genet.* **23,** 227–2250 (1989).
19. Deitsch, K. W., Calderwood, M. S. & Wellems, T. E. Malaria. Cooperative silencing elements in *var* genes. *Nature* **412,** 875–876 (2001).
20. Vazquez-Macias, A. *et al.* A distinct 5′ flanking var gene region regulates *Plasmodium falciparum* variant erythrocyte surface antigen expression in placental malaria. *Mol. Microbiol.* **45,** 155–167 (2002).
21. Quail, M. A. M13 cloning of mung bean nuclease digested PCR fragments as a means of gap closure within A/T-rich, genome sequencing projects. *DNA Seq.* **12,** 355–359 (2001).
22. Rutherford, K. *et al.* Artemis: sequence visualization and annotation. *Bioinformatics* **16,** 944–945 (2000).
23. Salzberg, S. L., Pertea, M., Delcher, A. L., Gardner, M. J. & Tettelin, H. Interpolated Markov models for eukaryotic gene finding. *Genomics* **59,** 24–31 (1999).
24. Cawley, S. E., Wirth, A. I. & Speed, T. P. Phat—a gene finding program for *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **118,** 167–174 (2001).
25. Zdobnov, E. M. & Apweiler, R. InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17,** 847–848 (2001).
26. Sonnhammer, E. L., von Heijne, G. & Krogh, A. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **6,** 175–182 (1998).
27. Nielsen, H., Brunak, S. & von Heijne, G. Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Eng.* **12,** 3–9 (1999).
28. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nature Genet.* **25,** 25–29 (2000).
29. Apweiler, R. *et al.* The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.* **29,** 37–40 (2001).
30. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **30,** 276–280 (2002).
31. Emanuelsson, O., Nielsen, H., Brunak, S. & von Heijne, G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* **300,** 1005–1016 (2000).
32. Claros, M. G. & Vincens, P. Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur. J. Biochem.* **241,** 779–786 (1996).
33. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305,** 567–580 (2001).

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

Acknowledgements

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

Competing interests statement

The authors declare that they have no competing financial interests.

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

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

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

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

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

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

here the nucleotide sequences of chromosomes 10, 11 and 14, and a re-analysis of the chromosome 2 sequence[5]. These chromosomes represent about 35% of the 23-megabase *P. falciparum* genome.

P. falciparum chromosomes were resolved on preparative pulsed field gels, and used to prepare shotgun libraries of 1–2-kilobase (kb) DNA fragments in plasmid vectors. Sequences of randomly selected clones were assembled, and gaps were closed using primer walking on plasmid templates or polymerase chain reaction (PCR) products. The cross-contamination of the chromosomal libraries with sequences from other chromosomes (up to 25%) and the high (A + T) content (80.6%) of *P. falciparum* DNA caused extreme difficulties in the gap closure process. Intergenic regions and introns frequently contained long runs of up to 50 consecutive A or T residues that were difficult to clone and sequence. The high (A + T) content of the chromosomes also prevented the construction of large insert libraries that could be used to construct scaffolds of ordered and oriented contiguous DNA sequences (contigs) during assembly. Similar but more severe problems were reported in the sequencing of the (A + T)-rich chromosome 2 of the slime mould *Dictyostelium discoideum*[6], illustrating the need to develop better methods for the cloning and sequencing of very (A + T)-rich genomes. The reported sequences contain three or four short gaps (<2 kb) in each chromosome. Contigs comprising these chromosomes were joined end-to-end before annotation. Efforts to close the remaining gaps will continue.

Examination of the sequences of chromosomes 2, 10, 11 and 14 revealed that the structure of these chromosomes was similar to that of the other chromosomes. All contained the 97–99% (A + T) putative centromeric sequences reported previously[7]. Conserved subtelomeric sequences[2] were observed in chromosomes 2, 10 and 11, but most of these elements had been deleted from both ends of chromosome 14. The termini of chromosome 14 consisted of telomeric hexamer repeats fused directly to truncated *var* (variant antigen) genes. Deletions of this type are thought to be due to chromosome breakage and healing events that occur during *in vitro* cultivation of the parasite.

Annotation procedures have improved since the publication of the *P. falciparum* chromosome 2 sequence[5]. A gene finding program, phat (pretty handy annotation tool[8]), was developed, supplementing the GlimmerM program[9] used previously. In this work, GlimmerM and phat were retrained on a larger training set of well-

Table 1 **Summary statistics**

Feature	Value				
	Whole genome	Chromosome 2	Chromosome 10	Chromosome 11	Chromosome 14
The genome					
Size (bp)	22,853,764	947,102	1,694,445	2,035,250	3,291,006
No. of gaps	93	0	4	3	3
Coverage*	14.5	11.1	15.6	11.3	9.2
(G + C) content (%)	19.4	19.7	19.7	19.0	18.4
No. of genes	5,268	223 (209)	403	492	769
Mean gene length (bp)†	2,283.3	2,079.1 (2,105.1)	2,085.8	2,127.7	2,315.1
Gene density (bp per gene)	4,338.2	4,247.1 (4,531.6)	4,204.6	4,136.7	4,279.6
Percent coding	52.6	49.0 (46.5)	49.6	51.4	54.1
Genes with introns (%)	53.9	57.0 (43.1)	51.4	50.4	49.9
Genes with ESTs (%)	49.1	46.2	48.1	48.4	46.9
Gene products detected by proteomics‡ (%)	51.8	43.5	49.1	51.0	52.1
Exons					
Number	12,674	510 (353)	892	1,094	1,757
Mean no. per gene	2.4	2.3 (1.7)	2.2	2.2	2.3
(G + C) content (%)	23.7	24.4 (24.3)	24.5	23.5	22.8
Mean length (bp)	949.1	909.1 (1,246.3)	942.3	956.9	1,013.3
Total length (bp)	12,028,350	463,647 (439,944)	840,576	1,046,814	1,780,305
Introns					
Number	7,406	287 (144)	489	602	988
(G + C) content (%)	13.5	13.4 (13.4)	13.6	13.7	13.5
Mean length (bp)	178.7	202.4 (208.4)	234.5	189.4	185.5
Total length (bp)	1,323,509	58,080 (30,006)	114,676	114,012	183,240
Intergenic regions					
(G + C) content (%)	13.6	13.5 (14.1)	13.6	14.1	13.2
Mean length (bp)	1,693.9	1,702.3 (2,063.2)	1,678.5	1,768.5	1,717.2
RNAs					
No. of tRNA genes	43	1	0	2	2
No. of 5S rRNA genes	3	0	0	0	3
No. of 5.8S, 18S and 28S rRNA units	7	0	0	1	0
The proteome					
Total predicted proteins	5,268	223	403	492	769
Hypothetical proteins§	3,208	121	265	339	485
InterPro matches	2,650	116	210	283	455
Pfam matches	1,746	77	133	184	275
Gene Ontology					
Process	1,301	63	89	110	168
Function	1,244	54	74	95	174
Component	2,412	120	181	220	308
Targeted to apicoplast	551	28	36	52	73
Targeted to mitochondrion	246	10	13	17	33
Structural features					
Transmembrane domain(s)	1,631	87	133	141	202
Signal peptide	544	28	41	52	63
Signal anchor	367	19	32	31	51

Numbers in parentheses under chromosome 2 indicate values obtained in the previous annotation[5]. Specialized searches used the following programs and databases: InterPro[21], Pfam[19] and Gene Ontology[22]. Predictions of apicoplast and mitochondrial targeting were performed using TargetP[26] and MitoProtII[25]; transmembrane domains, TMHMM[24]; and signal peptides and signal anchors, SignalP-2.0 (ref. 23).

*Average number of sequence reads per nucleotide. EST, expressed sequence tag.

†Excluding introns.

‡Percent of proteins detected in parasite extracts by two independent proteomic analyses[29,30].

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

characterized genes, complementary DNAs (cDNAs) and products of PCR with reverse transcription (RT–PCR) (total length 540 kb) than was used in the earlier work. A program called Combiner was used to evaluate the GlimmerM and phat predictions, as well as the results of searches against nucleotide and protein databases, to construct consensus gene models. To assess the effect of these modifications, chromosome 2 was re-annotated and the results were compared with the previous annotation.

Application of these automated annotation procedures and manual curation of the resulting gene models for chromosome 2 produced 223 gene models. The revised procedures detected 21 genes not predicted previously, and 13 of the existing chromosome 2 models collapsed into six models in the new annotation. Of the 21 new gene models, all but one had no significant similarity to proteins in a non-redundant amino-acid database. However, at least a portion of each of the 21 gene models had been predicted independently by both GlimmerM and phat, suggesting that many of these models were likely to represent coding sequences. On the other hand, five of the new gene models encoded proteins less than 100 amino acids in length, and may be less likely to encode proteins.

Another major difference was the detection of additional small exons. In the earlier annotation of chromosome 2, the 209 predicted genes contained 353 exons, or an average of 1.7 exons per gene. The revised procedures reported here revealed 510 exons, or 2.3 exons per gene; 60% of the new exons were predicted to be additions to the gene models reported previously. Most cases involved the addition of one or two exons per gene. In three notable cases, however, 7 to 12 small exons were added to the earlier gene models, and almost all of the new exons had been predicted by both of the gene finding programs. Overall, use of the revised annotation procedures resulted in the detection of additional genes and many small exons, which is reflected in the higher gene density and shorter mean exon length in the newly annotated chromosome 2 sequence compared with the previous annotation (Table 1). Despite these improvements in software and training sets, gene finding in *P. falciparum* remains challenging, and the gene structures presented here should be regarded as preliminary until confirmed by sequence information obtained from cDNAs or RT–PCR experiments[10]. Accurate prediction of the 5′ ends of genes is particularly difficult. Generation of larger training sets, including additional expressed sequence tags (ESTs) and full-length cDNAs, would greatly improve the sensitivity and accuracy of gene predictions.

These annotation procedures were also applied to the analysis of chromosomes 10, 11 and 14 (Table 1; maps of these chromosomes are available as Supplementary Information). The 10 short gaps in the chromosomes should not have interfered with the gene predictions; only the genes adjacent to the gaps might have been affected. All three chromosomes were similar in terms of gene density, coding percentage and other parameters. A complete description of the parasite genome is contained in the accompanying Article[2].

Annotation of chromosomes 10, 11 and 14 revealed four proteins with sequence similarity to SR proteins, a family of conserved splicing factors that contain RNA-binding domains and a protein interaction domain rich in Ser and Arg residues (SR domain; PF10_0047, PF10_0217, PF11_0200, PF14_0656). Three additional putative SR proteins were identified on chromosomes 5 and 13 (PFE0160c, PFE0865c, MAL13P1.120). SR proteins are thought to bind to exonic splicing enhancers (ESEs), short (6–9 bp) sequences within exons that assist in the recognition of nearby splice sites, and to interact with components of the spliceosome[11]. ESEs have previously been characterized only in multicellular organisms. To determine whether *P. falciparum* may use ESEs as part of its splicing machinery, a Gibbs sampling algorithm for motif detection[12] was applied to a set of *P. falciparum* exons to detect any exonic splicing enhancers (ESEs). The exons were extracted from the set of well-characterized genes used to train the GlimmerM gene finder. Regions of 50 bp regions were selected from both ends of the internal exons and divided into two different data sets, representing the exon regions adjacent to both 5′ and 3′ splice sites. At least 10 runs of the Gibbs sampler were performed for each data set in order to identify the most probable motif with a length of 5–9 nucleotides. The motif with the highest maximum *a posteriori* probability was retained. This analysis identified a motif with the consensus GAAGAA, which is identical to ESEs found in human exons[13,14]. The identification of several putative SR proteins, and sequences identical to the ESEs in humans, suggests that some features of exon recognition and splicing observed in higher eukaryotes may be conserved in *P. falciparum*. □

Methods

Sequencing and closure

P. falciparum clone 3D7 was selected for sequencing because it can complete all phases of the life cycle, and had been used in a genetic cross[15] and the Wellcome Trust Malaria Genome Mapping Project[16]. High-molecular-mass genomic DNA was subjected to electrophoresis on preparative pulsed field gels, and chromosomes were excised. DNA was extracted from the gel, sheared, and cloned into the pUC18 vector as described[5] (chromosomes 2, 14) or into a modified pUC18 vector via *Bst*XI linkers (chromosomes 10, 11). Sequences were assembled and gaps were closed by primer walking on plasmid DNAs or genomic PCR products, or by transposon insertion[5]. Ordering of contigs was facilitated by the use of sequence tagged sites[16] and microsatellite markers[17]. The final assembly of each chromosome was verified by comparison with *Bam*HI and *Nhe*I optical restriction maps[18]. The average difference in size between the experimentally determined restriction fragments and the fragments predicted from the sequence was approximately 5–6% for chromosomes 11 and 14 for both enzymes. For chromosome 10, the average difference in fragment sizes was 6.1% for the *Nhe*I map, but the *Bam*HI optical and prediction restriction maps could not be aligned. Because the *Nhe*I optical restriction map agreed with that predicted from the sequence, the chromosome 10 assembly was judged to be correct.

Annotation

GlimmerM[9] and phat[8] were trained on 117 *P. falciparum* genes and 39 cDNAs taken from GenBank, plus 32 genes from chromosomes 2 and 3 that had been verified by RT–PCR (provided by R. Huestis and K. Fischer; the training set is available at http://www.tigr.org/software/glimmerm/data). The GlimmerM and phat predictions, and sequence alignments of the chromosomes to protein and cDNA databases, were evaluated by the Combiner program. The program used a linear weighting method and dynamic programming to construct consensus gene models that were curated manually using AnnotationStation (AffyMetrix Inc.). Predicted proteins were searched against a non-redundant amino-acid database using BLASTP; other features were identified by searches against the Pfam[19], PROSITE[20] and InterPro[21] databases. The results of all analyses were reviewed using Manatee, a tool that interfaces with a relational database of the information produced by the annotation software. Predicted gene products were manually assigned Gene Ontology [22] terms. Signal peptides and signal anchors were predicted with SignalP-2.0 (ref. 23). Transmembrane helices were predicted with TMHMM[24]. Mitochondrial- and apicoplast-targeted proteins were predicted by MitoProtII[25], TargetP[26] and PATS[27]. tRNA-ScanSE[28] was used to identify transfer RNAs.

Received 6 August; accepted 2 September 2002; doi:10.1038/nature01094.

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. Gardner, M. J. *et al.* Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419,** 498–511 (2002).
3. Hall, N. *et al.* Sequence of *Plasmodium falciparum* chromosomes 1, 3–9 and 13. *Nature* **419,** 527–531 (2002).
4. Hyman, R. W. *et al.* Sequence of *Plasmodium falciparum* chromosome 12. *Nature* **419,** 534–537 (2002).
5. Gardner, M. J. *et al.* Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science* **282,** 1126–1132 (1998).
6. Glockner, G. *et al.* Sequence and analysis of chromosome 2 of *Dictyostelium discoideum*. *Nature* **418,** 79–85 (2002).
7. Bowman, S. *et al.* The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*. *Nature* **400,** 532–538 (1999).
8. Cawley, S. E., Wirth, A. I. & Speed, T. P. Phat—a gene finding program for *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **118,** 167–174 (2001).
9. Salzberg, S. L., Pertea, M., Delcher, A., Gardner, M. J. & Tettelin, H. Interpolated Markov models for eukaryotic gene finding. *Genomics* **59,** 24–31 (1999).
10. Huestis, R. & Fischer, K. Prediction of many new exons and introns in *Plasmodium falciparum* chromosome 2. *Mol. Biochem. Parasitol.* **118,** 187–199 (2001).
11. Maniatis, T. & Tasic, B. Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* **418,** 236–243 (2002).
12. Lawrence, C. E. *et al.* Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262,** 208–214 (1993).
13. Ramchatesingh, J., Zahler, A. M., Neugebauer, K. M., Roth, M. B. & Cooper, T. A. A subset of SR proteins activates splicing of the cardiac troponin T alternative exon by direct interactions with an exonic enhancer. *Mol. Cell Biol.* **15,** 4898–4907 (1995).
14. Fairbrother, W. G., Yeh, R. F., Sharp, P. A. & Burge, C. B. Predictive identification of exonic splicing enhancers in human genes. *Science* **297,** 1007–1013 (2002).

15. Walliker, D., Quayki, I., Wellems, T. E. & McCutchan, T. F. Genetic analysis of the human malaria parasite *Plasmodium falciparum*. *Science* **236,** 1661–1666 (1987).
16. Foster, J. & Thompson, J. The *Plasmodium falciparum* genome project: a resource for researchers. *Parasitol. Today* **11,** 1–4 (1995).
17. Su, X. *et al.* A genetic map and recombination parameters of the human malaria parasite *Plasmodium falciparum*. *Science* **286,** 1351–1353 (1999).
18. Lai, Z. *et al.* A shotgun optical map of the entire *Plasmodium falciparum* genome. *Nature Genet.* **23,** 309–313 (1999).
19. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **30,** 276–280 (2002).
20. Falquet, L. *et al.* The PROSITE database, its status in 2002. *Nucleic Acids Res.* **30,** 235–238 (2002).
21. 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).
22. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nature Genet.* **25,** 25–29 (2000).
23. 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).
24. 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).
25. Claros, M. G. & Vincens, P. Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur. J. Biochem.* **241,** 779–786 (1996).
26. 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).
27. 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).
28. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25,** 955–964 (1997).
29. Florens, L. *et al.* A proteomic view of the *Plasmodium falciparum* life cycle. *Nature* **419,** 520–526 (2002).
30. Lasonder, E. *et al.* Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry. *Nature* **419,** 537–542 (2002).

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

Acknowledgements

We thank our colleagues at The Institute for Genomic Research and the Naval Medical Research Center for support; 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; and S. Cawley for assistance with phat. This work was supported by 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). Chromosome sequences have been deposited in GenBank with accession numbers AE001362.2 (chromosome 2), AE014185 (chromosome 10), AE01486 (chromosome 11) and AE01487 (chromosome 14), and in PlasmoDB (http://plasmodb.org).

Sequence of *Plasmodium falciparum* chromosome 12

Richard W. Hyman, Eula Fung, Aaron Conway, Omar Kurdi, Jennifer Mao, Molly Miranda, Brian Nakao, Don Rowley, Tomoaki Tamaki, Fawn Wang & Ronald W. Davis

Stanford Genome Technology Center, 855 California Avenue, Palo Alto, California 94304 USA, and Departments of Biochemistry and Genetics, Stanford University Medical College, Stanford University, Stanford, California 94305, USA

The human malaria parasite *Plasmodium falciparum* is responsible for the death of more than a million people every year[1]. To stimulate basic research on the disease, and to promote the development of effective drugs and vaccines against the parasite, the complete genome of *P. falciparum* clone 3D7 has been sequenced, using a chromosome-by-chromosome shotgun strategy[2–4]. Here we report the nucleotide sequence of the third largest of the parasite's 14 chromosomes, chromosome 12, which comprises about 10% of the 23-megabase genome. As the most (A + T)-rich (80.6%) genome sequenced to date, the *P. falciparum* genome presented severe problems during the assembly of primary sequence reads. We discuss the methodology that yielded a finished and fully contiguous sequence for chromosome 12. The biological implications of the sequence data are more thoroughly discussed in an accompanying Article (ref. 3).

At the inception of the Malaria Genome Project, our colleagues at the Institute for Genomic Research (TIGR) and the Wellcome Trust Sanger Institute (WTSI) sequenced *P. falciparum* chromosomes 2 and 3 (refs 5, 6). We chose to sequence the third-largest *P. falciparum* chromosome, chromosome 12, which comprises about 10% of the genome. We made this choice because a 'tiling path' had just been published[7]. (A tiling path is an ordered set of recombinant DNAs covering a large DNA sequence, such as chromosome 12. In this case, the tiling path is composed of yeast artificial chromosomes (YACs) with sequence-tagged sites (STSs, mapped sequence markers).) We predicted that the YACs and the STSs would be helpful in positioning sequence contigs (stretches of contiguous sequence) along *P. falciparum* chromosome 12.

From the published data[7], we defined a 21 YAC tiling path across *P. falciparum* chromosome 12 (Supplementary Fig. 1). However, we did not want to rely exclusively on sequencing YACs because of three important concerns, which turned out to be warranted. (1) Base changes in the sequence can occur during the construction of any recombinant DNA/YAC, and mutations can occur during passage of any YAC in yeast. (2) One or more YACs in the tiling path might not overlap a neighbouring YAC, creating a physical gap in the sequence. (3) Three of the YACs in the tiling path were derived from *P. falciparum* clone B8 rather than clone 3D7. Polymorphisms between the DNAs of the two strains could hinder the assembly process. Therefore, we devised the following overall strategy. We would sequence random pieces of (that is, use 'shotgun sequencing' on) each of the YACs in the minimum tiling path to low coverage—just enough to establish a 'bin' (a group of related sequences). The bins would give us physical position information across *P. falciparum* chromosome 12. The STSs would give us physical position information within each bin. In addition, we would shotgun-sequence *P. falciparum* chromosome 12 itself. The sequence of each chromosome 12 shotgun sequence 'read' (a sequence of length 100–600 bases derived from a piece of DNA) would be compared to the sequences in each bin. When there was a good match, the read would stay in that bin. This process is highly iterative.

The 21 YACs comprising the minimum tiling path varied considerably in size, with a range of 40–220 kilobases (kb; ref. 7). Our shotgun sequence coverage of the YACs also varied considerably, with a range of 0.5–9.7 YAC coverage (Supplementary Table 1). However, with the exception of four YACs with which we experimented with high coverage early in this project, the shotgun sequence coverage of the remaining YACs was low, as originally planned. In total, there are 14,159 YAC reads (2.6-fold chromosome 12 coverage) supporting the final chromosome 12 sequence. In addition, we produced 69,532 *P. falciparum* chromosome 12 shotgun reads (11.3-fold chromosome 12 coverage) that support the chromosome 12 consensus sequence (Supplementary Table 2). After assembling all of the shotgun sequence data, nearly all of the contigs could be placed unambiguously relative to each other, based on the YAC bins and the STSs. The few remaining contigs were positioned unambiguously by using the genetic map of *P. falciparum* chromosome 12 constructed through the use of microsatellite markers derived from our chromosome 12 sequence[8,9]. The very few remaining contigs were placed unambiguously by use of the data that accrued during the process of 'finishing' (identifying and replacing all problems in the assembled sequence).

Every part of the assembled sequence of *P. falciparum* chromosome 12 was carefully examined to identify problems in the sequence. These problems were of many types, including (but not limited to) gaps in the sequence, weakly supported sequence,

Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*

Jane M. Carlton*, Samuel V. Angiuoli*, Bernard B. Suh*, Taco W. Kooij†, Mihaela Pertea*, Joana C. Silva*, Maria D. Ermolaeva*, Jonathan E. Allen*, Jeremy D. Selengut*, Hean L. Koo*, Jeremy D. Peterson*, Mihai Pop*, Daniel S. Kosack*, Martin F. Shumway*, Shelby L. Bidwell*, Shamira J. Shallom*, Susan E. van Aken*, Steven B. Riedmuller*, Tamara V. Feldblyum*, Jennifer K. Cho*‡, John Quackenbush*, Martha Sedegah§, Azadeh Shoaibi*, Leda M. Cummings*‡, Laurence Florens||, John R. Yates||, J. Dale Raine¶, Robert E. Sinden¶, Michael A. Harris#, Deirdre A. Cunningham☆, Peter R. Preiser☆, Lawrence W. Bergman, Akhil B. Vaidya**, Leo H. van Lin†, Chris J. Janse†, Andrew P. Waters†, Hamilton O. Smith#, Owen R. White*, Steven L. Salzberg*, J. Craig Venter††, Claire M. Fraser*, Stephen L. Hoffman‡§, Malcolm J. Gardner* & Daniel J. Carucci§**

* *The Institute for Genomic Research, 9712 Medical Center Drive; and* †† *The Center for the Advancement of Genomics, 1901 Research Boulevard, Rockville, Maryland 20850, USA*
† *Department of Parasitology, Leiden University Medical Centre, PO Box 9600, 2300 RC Leiden, The Netherlands*
§ *Naval Medical Research Center, Malaria Program (IDD), Silver Spring, Maryland 20910, USA*
|| *Department of Cell Biology, The Scripps Research Institute, La Jolla, California, 92037, USA*
¶ *Infection & Immunity Section, Department of Biological Sciences, Imperial College of Science, Technology & Medicine, London, SW7 2AZ, UK*
# *Celera Genomics, 45 West Gude Drive, Rockville, Maryland 20850, USA*
☆ *Division of Parasitology, National Institute for Medical Research, London, UK*
** *Division of Molecular Parasitology, Department of Microbiology & Immunology, Drexel University College of Medicine, Philadelphia, Pennsylvania 191929, USA*

Species of malaria parasite that infect rodents have long been used as models for malaria disease research. Here we report the whole-genome shotgun sequence of one species, *Plasmodium yoelii yoelii*, and comparative studies with the genome of the human malaria parasite *Plasmodium falciparum* clone 3D7. A synteny map of 2,212 *P. y. yoelii* contiguous DNA sequences (contigs) aligned to 14 *P. falciparum* chromosomes reveals marked conservation of gene synteny within the body of each chromosome. Of about 5,300 *P. falciparum* genes, more than 3,300 *P. y. yoelii* orthologues of predominantly metabolic function were identified. Over 800 copies of a variant antigen gene located in subtelomeric regions were found. This is the first genome sequence of a model eukaryotic parasite, and it provides insight into the use of such systems in the modelling of *Plasmodium* biology and disease.

For decades, the laboratory mouse has provided an alternative platform for infectious disease research where the pathogen under study is intractable to routine laboratory manipulation. Experimental study of the human malaria parasite *Plasmodium falciparum* is particularly problematic as the complete life cycle cannot be maintained *in vitro*. Four species of rodent malaria (*Plasmodium yoelii*, *Plasmodium berghei*, *Plasmodium chabaudi* and *Plasmodium vinckei*) isolated from wild thicket rats in Africa have been adapted to grow in laboratory rodents[1]. These species reproduce many of the biological characteristics of the human malaria parasite. Many of the experimental procedures refined for use with *P. falciparum* were initially developed for rodent malaria species, a prime example being stable genetic transformation[2]. Thus rodent models of malaria have been used widely and successfully to complement research on *P. falciparum*.

With the advent of the *P. falciparum* Genome Sequencing Project, undertaken by an international consortium of genome sequencing centres and malaria researchers, a series of initiatives has begun to generate substantial genome information from additional *Plasmodium* species[3]. We describe here the genome sequence of the rodent malaria parasite *P. y. yoelii* to fivefold genome coverage. We show that this partial genome sequencing approach, although limited in its application to the study of genome structure, has proved to be an effective means of gene discovery and of jump-starting experimental studies in a model *Plasmodium* species. Furthermore, we show that despite the considerable divergence between the *P. y. yoelii* and *P. falciparum* genomes, sequencing and annotation of the former can substantially improve the accuracy and efficiency of annotation of the latter.

Plasmodium yoelii yoelii genome sequencing and annotation

We applied the whole-genome shotgun (WGS) sequencing approach, used successfully to sequence and assemble the first large eukaryotic genome[4], to achieve fivefold sequence coverage of the genome of a clone of the 17XNL line of *P. y. yoelii* (Table 1). This level of coverage is expected to comprise 99% of the genome[5] assuming random library representation. As with *P. falciparum*, the genomes of rodent malaria parasites are highly (A + T)-rich[6], which adversely affects DNA stability in plasmid libraries. Consequently, all ~220,000 reads were produced from clones originating

Table 1 ***Plasmodium yoelii yoelii* genome coverage statistics**

Data	Component	Value
Genome	No. of contigs	5,687
	Mean contig size (kb)	3.6
	Max. contig size (kb)	51.5
	Cumulative contig length (Mb)	23.1
	No. of singletons	11,732
	No. of groups	2,906
	Max. group size (kb)	69.8
	Cumulative group size (Mb)	21.6
Transcriptome	No. of ESTs	13,080
	Average length (nucleotides)	497
Proteome	No. of gametocyte peptides	1,413
	No. of sporozoite peptides	677

‡ Present addresses: National Center for Biotechnology Information, National Institutes of Health, Bethesda, Maryland 20894, USA (L.M.C.); Genentech, San Francisco, California 94080, USA (J.K.C.); and Sanaria, 308 Argosy Drive, Gaithersburg, Maryland 20878, USA (S.L.H.).

from small (2–3 kilobases (kb)) insert libraries. Contigs were assembled using TIGR Assembler[7]. Contaminating mouse sequences, identified through similarity searches and found to comprise 10% of the total sequence data, were excluded from the analyses. Approximately three-quarters of the contigs could be placed into 2,906 'groups', each group consisting of two or more contigs known to be linked through paired reads as determined by Grouper software[7]. This produced an average group size of 7.4 kb, approximately 4 kb more than the average contig size. This group size is small compared with the group data produced by other partial eukaryotic genome projects, where extensive use of large insert (linking) libraries has enabled the construction of ordered and orientated 'scaffolds'[8], and emphasizes the use of such linking libraries in partial genome projects. The genome size of *P. y. yoelii* is estimated to be 23 megabases (Mb), in agreement with karyotype data[9].

Expression data from the *P. y. yoelii* transcriptome and proteome were generated to aid in gene identification and annotation of the contigs (Table 1). A total of 13,080 expressed sequence tag (EST) sequences generated from clones of an asexual blood-stage *P. y. yoelii* complementary DNA library[10], in combination with other *P. yoelii* ESTs and transcript sequences available from public databases, were assembled and used to compile a gene index[11] of expressed *P. y. yoelii* sequences (http://www.tigr.org/tdb/tgi/pygi/). For protein expression data, multidimensional protein identification technology (MudPIT), which combines high-resolution liquid chromatography with tandem mass spectrometry and database searching, was applied to the gametocyte and salivary gland sporozoite proteomes of *P. y. yoelii*. A total of 1,413 gametocyte and 677 sporozoite peptides were recorded and used for the purposes of gene annotation.

We used two gene-finding programs, GlimmerMExon and Phat[12], to predict coding regions in *P. y. yoelii*. GlimmerMExon is based on the eukaryotic gene finder GlimmerM[13], with modifications developed for analysing the short fragments of DNA that result from partial shotgun sequencing. Gene models based on GlimmerMExon and Phat predictions were refined using Combiner. Annotation of predicted gene models used TIGR's fully automated Eukaryotic Genome Control suite of programs. Gene finding and subsequent annotation were limited to 2,960 contigs (each of which is over 2 kb in size), a subset of sequences that contains more than 20 Mb of the genome. A total of 5,878 complete genes and 1,952 partial genes (defined as genes lacking either an annotated start or stop codon) can be predicted from the nuclear genome data.

Table 2 **Comparison of genome features of *P. falciparum* and *P. y. yoelii***

Feature	*P. y. yoelii*	*P. falciparum*
Size (Mb)	23.1	22.9
No. of chromosomes	14	14
No. of gaps	5,812	93
Coverage*	5	14.5
(G + C) content (%)	22.6	19.4
No. of genes†	5,878	5,268
Mean gene length (bp)	1,298	2,283
Gene density (bp per gene)	2,566	4,338
Per cent coding	50.6	52.6
Genes with introns (%)	54.2	53.9
Genes with ESTs (%)	48.9	49.1
Gene products detected by proteomics (%)	18.2	51.8
Exons		
Mean no. per gene	2.0	2.4
(G + C) content (%)	24.8	23.7
Mean length (bp)	641	949
Introns		
(G + C) content (%)	21.1	13.5
Mean length (bp)	209	179
Total length (bp)	1,687,689	1,323,509
Intergenic regions		
(G + C) content (%)	20.7	13.6
Mean length (bp)	859	1,694
RNAs		
No. of tRNA genes‡	39	43
No. of 5S rRNA genes	3	3
No. of 5.8S, 18S and 28S rRNA units	4	7
Mitochondrial genome		
(G + C) content (%)	31	31
Apicoplast genome		
(G + C) content (%)	15	14

*Average number of sequence reads per nucleotide.
†Total number of full-length genes.
‡The smaller number reflect the partial nature of the *P. y. yoelii* genome data.

Comparative genome analysis

A comparison of several genome features of *P. falciparum* and *P. y. yoelii* is shown in Table 2, demonstrating that many similarities exist between the genomes. Besides the similarly extreme (G + C) compositions, both genomes contain a comparable number of predicted full-length genes, with the higher figure in *P. y. yoelii* due to an extremely high copy number of variant antigen genes (see below). Where differences between the genomes do exist, such as the (G + C) content of the coding portion of the genomes, incompleteness of the *P. y. yoelii* genome data, with the associated problems of accurate gene finding in both species, is likely to be a confounding factor. As an indication of this problem, analysis of *P. y. yoelii* proteomic data identified 83 regions of the genome apparently expressed during sporozoite and/or gametocyte stages but not assigned to a *P. y. yoelii* gene model (data not shown). Many of these peptide hits appear sufficiently close to a model as to indicate a fault with gene boundary prediction rather than a lack of gene prediction *per se*. However, as with the gene model prediction in *P. falciparum*, the gene models of *P. y. yoelii* should be considered preliminary and under revision.

Identifying orthologues of *P. falciparum* vaccine candidate proteins and proteins that are either targets of antimalarial drugs or involved in antimalarial drug resistance mechanisms is a primary goal of model malaria parasite genomics. Using BLASTP[14] with a cutoff E value of 10^{-15} and no low-complexity filtering, 3,310 bi-directional orthologues (defined as genes related to each other through vertical evolutionary descent) can be identified in the full protein complement of *P. falciparum* (5,268 proteins) and the protein complement of *P. y. yoelii* translated from complete gene models (5,878 proteins). A list of vaccine candidate orthologues and orthologues of genes involved in antimalarial drug interactions identified from among the 3,310 orthologues and from additional BLAST analyses is shown in Table 3. Those genes that are not identifiable may either be absent from the partial genome data, or represent genes that have been lost or diverged sufficiently that they are undetectable through similarity searching.

Many of the candidate vaccine antigens under study in *P. falciparum* can be identified in *P. y. yoelii*, including orthologues of several asexual blood-stage antigens known to elicit immune responses in individuals exposed to natural infection (MSP1, AMA1, RAP1, RAP2). As immunity to *P. falciparum* blood-stage infection can be transferred by immune sera, identification of the targets of potentially protective antibody responses after natural infection can provide information beneficial to the selection of candidate antigens for malaria vaccines. We found several orthologues of known *P. falciparum* transmission-blocking candidates; in particular, members of the P48/45 gene family identified previously[15] were confirmed.

We identified several *P. y. yoelii* orthologues of *P. falciparum* biochemical pathway components under study as targets for drug design (Table 3), most notably: (1) the 1-deoxy-D-xylulose 5-phosphate reductoisomerase (DOXPR) gene whose product is inhibited by fosmidomycin in *P. falciparum in vitro* cultures and mice infected with *P. vinckei*[16]; (2) enoyl-acyl carrier protein (ACP) reductase (FABI) whose product is inhibited by triclosan in *P. falciparum in vitro* cultures and mice infected with *P. berghei*[17]; and (3) a gene encoding farnesyl transferase (FTASE), which is inhibited in cultures of *P. falciparum* treated with custom-designed peptidomimetics[18]. The rodent models of malaria have proved

Table 3 ***P. y. yoelii* orthologues of *P. falciparum* candidate vaccine and drug interaction genes**

P. falciparum gene	*Pf* chromosome	ST location*	*Pf* locus	*Py* locus
Candidate vaccine antigens				
Ring-infected erythrocytic surface antigen 1, *resa1*	1	Yes	PFA0110w	Not identified
Merozoite surface protein 4, *msp4*	2	No	PFB0310c	PY07543†
Merozoite surface protein 5, *msp5*	2	No	PFB0305c	PY07543†
Liver stage antigen 3, *lsa3*	2	No	PFB0915w	Not identified
Merozoite surface protein 2, *lsa3*	2	No	PFB0300c	Not identified
Transmission-blocking target antigen 230, *Pfs230*	2	No	PFB0405w	PY03856
Circumsporozoite protein, *csp*	3	No	MAL3P2.11	PY03168
Rhoptry-associated protein 2, *rap2*	5	Yes	PFE0080c	PY03918
Sporozoite surface antigen, *starp*	7	Yes	PF07_0006	Not identified
Morozoite surface protein 1, *msp1*	9	No	PFI475w	PY05748
Liver stage antigen 1, *lsa1*	10	No	PF10_0356	Not identified
Merozoite surface protein 3, *msp3*	10	No	PF10_0345	Not identified
Glutamate-rich protein, *glurp*	10	No	PF10_0344	Not identified
Ookinete surface protein 25, *Pfs25*	10	No	PF10_0303	PY00523
Ookinete surface protein 28, *Pfs28*	10	No	PF10_0302	PY00522
Erythrocyte membrane-associated 332 antigen, *Pf332*	11	No	PF11_0507	PY06496
Apical membrane antigen 1, *ama1*	11	No	PF11_0344	PY01581
Exported protein 1, *exp1*	11	No	PF11_0224	Not identified
Surface sporozoite protein 2, *ssp2*	13	No	PF13_0201	PY03052
Sexual-stage-specific surface antigen 48/45, *Pfs48/45*	13	No	PF13_0247	PY04207
Rhoptry-associated protein 1, *rap1*	14	Yes	PF14_0637	PY00622
Candidate drug interaction genes				
Dihydrofolate reductase, *dhfr*	4	No	PFD0830w	PY04370
Multidrug resistance protein 1, *pfmdr1*	5	No	PFE1150w	PY00245
Translationally controlled tumour protein, *tctp*	5	No	PFE0545c	PY04896
Farnesyl transferase, *ftase*	5	No	PFE0970w	PY06214
Enoyl-acyl carrier reductase, *fabi*	6	No	MAL6P1.275	PY03846
Dihydro-protate dehydrogenase, *dhod*	6	No	MAL6P1.36	PY02580
Chloroquine-resistance transporter, *pfcrt*	7	No	MAL7P1.27	PY05061
Dihydropteroate synthase, *dhps*	8	No	PF08_0095	PY02226
Lactate dehydrogenase, *ldh*	13	No	PF13_0141	PY03885
DOXP reductoisomerase, *doxpr*	14	No	PF14_0641	PY05578

A full listing of all orthologues can be found as Table A in the Supplementary Information. *Pf*, *P. falciparum*; *Py*, *P. y. yoelii*.
*ST, subtelomeric. Defined as >75% of the distance from the centre to the end of the *P. falciparum* chromosome.
†Homologue of *P. falciparum msp4* and *msp5* genes found as a single gene *msp4/5* in *P. y. yoelii* and other rodent malaria species[62].

invaluable both for the study of potency of new antimalarial compounds *in vivo*, and for the elucidation of mechanisms of antimalarial drug resistance.

We applied the Gene Ontology (GO) gene classification system[19], which uses a controlled vocabulary to describe genes and their function, to indicate which classes of gene among the 3,310 orthologues might differ in number between *P. falciparum* and *P. y. yoelii* (Fig. 1). A similar proportion of proteins were identified for most of the GO classes between the two species, with the caveat that fewer total numbers of proteins were identified in *P. y. yoelii* owing to the partial nature of the genome data for this species. However, proteins allocated to the physiological processes, cell invasion and adhesion, and cell communication categories were significantly reduced in *P. y. yoelii*. These classes contain members of three multigene families whose genes are found predominantly in the subtelomeric regions of *P. falciparum* chromosomes: PfEMP1, the protein product of the *var* gene family known to be involved in antigenic variation, cyto-adherence and rosetting, and rifins and stevors, which are clonally variant proteins possibly involved in antigenic variation and evasion of immune responses (reviewed in ref. 20). Apparently, *P. falciparum* has generated species-specific, subtelomeric genes involved in host cell invasion, adhesion and antigenic variation, homologues of which are not found in the *P. y. yoelii* genome.

Gene families of unique interest in the *P. y. yoelii* genome

The largest family of genes identified in the *P. y. yoelii* genome is the *yir* gene family, homologues of the *vir* multigene family recently described in the human malaria parasite *Plasmodium vivax*[21] and in other species of rodent malaria[22]. In *P. vivax*, an estimated 600–1,000 copies of the subtelomerically located *vir* gene encode proteins that are immunovariant in natural infections, indicating a possible functional role in antigenic variation and immune evasion. Within the *P. y. yoelii* genome data, 838 *yir* genes (693 full genes and 145 partial genes) are present (Table 4; see also Supplementary Figs A and B). Almost 75% of the annotated contigs identified as containing subtelomeric sequences (see below) contain *yir* genes, many arranged in a head-to-tail fashion. Expression data indicate that *yir* genes are expressed during sporozoite, gametocyte and erythrocytic stages of the parasite, similar to the expression pattern seen with *P. falciparum var* and *rif* genes[23]. Preliminary

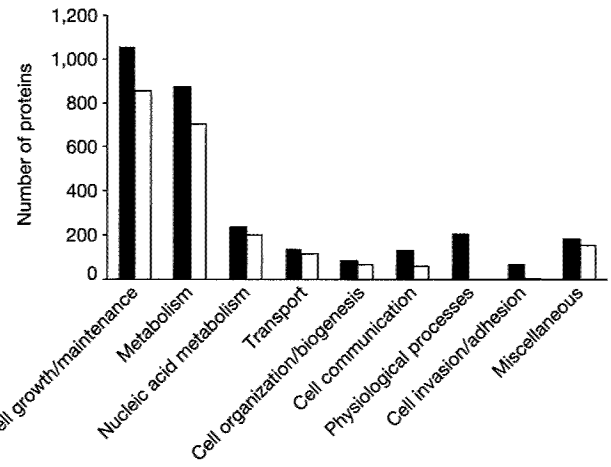


Figure 1 Functional classification comparison between *P. falciparum* and *P. y. yoelii* proteins. We compared the GO terms of proteins assigned to 'biological process' for the orthologous genes identified between the two species. The process group contains 3,041 *P. falciparum* annotations (filled bars), and 2,161 reciprocal annotations are shown for *P. y. yoelii* (open bars). Ten GO classes with similar numbers of *P. falciparum* and *P. y. yoelii* proteins in each are assigned as 'miscellaneous'; that is, cell cycle, external stimulus response, stress response, signal transduction, homeostasis, developmental processes, cell proliferation, membrane fusion, death, cell motility.

Table 4 **Paralogous gene families in *P. y. yoelii***

Gene family	No.	Name	HMM ID	Location in *Py*	*Py* expression*	*Pf* locus	TM/SP†
yir/bir/cir	838	Variant antigen family	TIGR01590	Subtelomeric	Gmt, spz, bs	None	P/A
235 kDa	14	Reticulocyte binding family	TIGR01612	Subtelomeric	Gmt, spz, bs	PFD0110w, MAL13P1.176, PF13_0198, PFL2520w, PFD0110w	P/A
pyst-a	168	Hypothetical	TIGR01599	Subtelomeric	Gmt, spz	PF14_0604	A/A
pyst-b	57	Hypothetical	TIGR01597	Subtelomeric	Bs	None	P/A
pyst-c	21	Hypothetical	TIGR01601, TIGR01604	Subtelomeric	Bs	None	P/P
pyst-d	17	Hypothetical	TIGR01605	Subtelomeric	Gmt	None	P/P
etramp	11	Early transcribed membrane protein family	TIGR01495	Subtelomeric	Gmt, spz, bs	PF13_0012, PF14_0016, PF11_0040, PFB0120w, PF10_0323, MAL12P1.387, PF11_0039, PFL1095c, PF10_0019, PF1745c, PFE1590w, PF10_0164, MAL8P1.6, PFA0195w, PFL0065w, PF14_0729	P/P
pst-a	12	Hydrolase family	TIGR01607	Subtelomeric	Gmt, spz	PFL2530w, PF10_0379, PF14_0738, PF14_0017, PF14_0737, PFI800w, PFI1775w, PF07_0040, PF07_0005, PFA0120c	A/A
rhoph1/clag	2	Rhoptry H1/ cyto-adherence-linked asexual gene family	PF03805	Subtelomeric	Gmt, bs	PFC0110w, PFC0120w, PFI1730w, PFI1710w, PFB0935w	A/P

*Found in, but not limited to: gmt, gametocyte life stage; spz, sporozoite life stage; bs, asexual blood stage.
†TM, transmembrane domain; SP, signal peptide; P, predicted; A, absent. TM and SP predictions were identical for *P. falciparum* and *P. y. yoelii* members of the same gene family. (See ref. 30 for details regarding TM and SP prediction algorithms.)

results using antibodies developed against the conserved regions of the protein have confirmed protein localization at the surface of the infected red blood cell (D.A.C. *et al.*, manuscript in preparation). The number of gene copies in the *P. y. yoelii* genome, the localization and stage-specific expression of gene members, as well as the existence of homologues in other *Plasmodium* species, make this gene family a prime target for the study of mechanisms of immune evasion.

A maximum of 14 members of the *Py235* multigene family can be identified among the *P. y. yoelii* protein data (Table 4). This family expresses proteins that localize to rhoptries (organelles that contain proteins involved in parasite recognition and invasion of host red blood cells). *Py235* genes exhibit a newly discovered form of clonal antigenic variation, whereby each individual merozoite derived from a single parent schizont has the propensity to express a different Py235 protein[24]. Closely related homologues of the *Py235* gene family have been found in other rodent malaria species, and more distantly related homologues have been found in *P. vivax*[25] and *P. falciparum*[26]. The gene copy number identified in the current data set is less than has been predicted in other *P. y. yoelii* lines (30–50 per genome). This could reflect real differences in copy number between lines, but more probably suggests an error in the original estimate or misassembly of extremely closely related sequences. Almost all of the *Py235* genes are found on contigs identified as subtelomeric in the *P. y. yoelii* genome (see Supplementary Fig. C).

Four further paralogous gene families, *pyst-a* to *-d*, are specific to *P. y. yoelii* (Table 4). The *pyst-a* family deserves mention, as it is homologous to a *P. chabaudi* glutamate-rich protein[27] and to a single hypothetical gene on *P. falciparum* chromosome 14, suggesting expansion of this family in the rodent malaria species from a common ancestral *Plasmodium* gene. Two paralogous gene families containing multiple members are homologous to multigene families identified in *P. falciparum*. Gene members of one family, *etramp* (early transcribed membrane protein), have previously been identified in *P. falciparum*[28] and in *P. chabaudi* where a single member has been identified and localized to the parasitophorous vacuole membrane[29].

Telomeres and chromosomal exchange in subtelomeric regions

The telomeric repeat in *P. y. yoelii* is AACCCTG, which differs from the *P. falciparum* telomeric repeat AACCCTA by one nucleotide. A total of 71 contigs were found to contain telomeric repeat sequences arranged in tandem, with the largest array consisting of 186 copies. The *P. y. yoelii* subtelomeric chromosomal regions show little repeat structure compared with those of *P. falciparum*. A survey of tandem repeats in the entire genome found only a few in the telomeric or subtelomeric regions, specifically a 15 base pair (bp) (45 copies) and a 31-bp (up to 10 copies), both of which were found on multiple contigs, and a 36-bp repeat that occurred on one contig. No repeat element that corresponds to Rep20, a highly variable 21-bp unit that spans up to 22 kb in *P. falciparum* telomeres, was found.

The telomeric and subtelomeric regions of *P. y. yoelii* contigs show extensive large-scale similarity, indicating that these regions undergo chromosomal exchange similar to that reported for *P. falciparum* (see ref. 30). The longest subtelomeric contig is approximately 27 kb (see Supplementary Fig. C) and is homologous to other subtelomeric contigs across its entire length, indicating that the region of chromosomal exchange extends at least this distance into the subtelomeres. Recent data have shown that clustering of telomeres at the nuclear periphery in asexual and sexual stage *P. falciparum* parasites may promote sequence exchange between members of subtelomeric virulence genes on heterologous chromosomes, resulting in diversification of antigenic and adhesive phenotypes (see ref. 31 for review). The suggestion of extensive chromosome exchange in *P. y. yoelii* indicates that a similar system for generating antigenic diversity of the *yir*, *Py235* and other gene families located within subtelomeric regions may exist.

A genome-wide synteny map

The *Plasmodium* lineage is estimated to have arisen some 100–180 million years ago[32], and species of the parasite are known to infect birds, mammals and reptiles[33]. On the basis of the analysis of small subunit (SSU) ribosomal RNA sequences, the closest relative to *P. falciparum* is *Plasmodium reichenowi*, a parasite of chimpanzees, with the rodent malaria species forming a distinct clade[34,35]. Early gene mapping studies have shown that regions of gene synteny exist between species of rodent malaria[9] and between human malaria species[36,37], despite extensive chromosome size polymorphisms between homologous chromosomes[38]. This level of gene synteny seems to decrease as the phylogenetic distance between *Plasmodium* species increases[39]. Before the *Plasmodium* genome sequencing

projects, the degree to which conservation of synteny extended across *Plasmodium* genomes was not fully apparent.

Using the *P. falciparum* and *P. y. yoelii* genome data, we have constructed a genome-wide syntenic map between the species. To avoid confounding factors inherent in DNA-based analyses of (A + T)-rich genomes, we first calculated the protein similarity between all possible protein-coding regions in both data sets using MUMmer[40]. Sensitivity was ensured through the use of a minimum word match length of five amino acids chosen to identify seed maximal unique matches (MUMs). By comparison, the recent human–mouse synteny analysis used a match length of 11 (ref. 8). Using this method, which is independent of gene prediction data, 2,212 sequences could be aligned (tiled) to *P. falciparum* chromosomes, representing a cumulative length of 16.4 Mb of sequence, or over 70% of the *P. y. yoelii* genome (see Supplementary Table C). The per cent of each *P. falciparum* chromosome covered with *P. y. yoelii* matches varies from 12% (chromosome 4) to 22% (chromosomes 1 and 14), with an average of about 18%. The spatial arrangement of the tiling paths (see Fig. 1 in ref. 30) confirms previous suggestions[9] that most of the conserved matches are found within the body of *Plasmodium* chromosomes, and confirms the absence of *var*, *rif* and *stevor* homologues in the *P. y. yoelii* genome.

Although the tiling paths indicate the degree of conservation of gene order between *P. falciparum* and *P. y. yoelii*, longer stretches of contiguous *P. y. yoelii* sequence are necessary to examine this feature in depth. Accordingly, we carried out linkage of many *P. y. yoelii* assemblies adjacent to each other along the tiling paths. First, 1,050 adjacent contigs were linked on the basis of paired reads as determined by Grouper software. Second, *P. y. yoelii* ESTs were aligned to the tiling paths, and those found to overlap sequences adjacent in the tiling path were used as evidence to link a further 236 *P. y. yoelii* sequences. Third, amplification of the sequence between adjacent contigs in the tiling paths linked a further 817 assemblies. Linkage of *P. y. yoelii* sequences by these methods resulted in the formation of 457 syntenic groups from 2,212 original contigs, ranging in length from a few kilobases to more than 800 kb. Syntenic groups were assigned to a *P. y. yoelii* chromosome where possible through the use of a partial physical map[9]. Thus, long contiguous sections of the *P. y. yoelii* genome with accompanying *P. y. yoelii* chromosomal location can be assigned to each *P. falciparum* chromosome (see Fig. 1 in ref. 30). The degree of conservation of gene order between the species was examined using ordered and orientated syntenic groups and Position Effect software. Of 4,300 *P. y. yoelii* genes within the syntenic groups, 3,145 (73%) were found to match a region of *P. falciparum* in conserved order.

One section of the syntenic map between *P. falciparum* and *P. y. yoelii* in particular—associated with *P. falciparum* chromosomes 4 and 10 and *P. y. yoelii* chromosome 5—provides a detailed snapshot of synteny between the species. Chromosome 5 of *P. y. yoelii* has received particular attention owing to the localization of a number of sexual-stage-specific genes to it[41], and because truncated versions of the chromosome are found in lines of the rodent malaria parasite *P. berghei*, which is defective in gametocytogenesis[42]. Genomic resources available for *P. berghei* chromosome 5 include chromosome markers and long-range restriction maps[41]. Exploiting the high level of synteny of rodent malaria parasite chromosomes[9], these tools were applied in combination with further mapping studies to close the syntenic map of chromosome 5 of *P. y. yoelii* (Fig. 2).

Approximately 0.8 Mb of *P. y. yoelii* chromosome 5 (estimated total length of 1.5 Mb) could be linked into one group that is syntenic to *P. falciparum* chromosome 10 and *P. falciparum* chromosome 4. From a total of 243 genes predicted in the syntenic region of *P. falciparum* chromosome 10, and 34 genes predicted in the syntenic region of chromosome 4, 171 (70%) and 22 (65%) of these, respectively, have homologues along *P. y. yoelii* chromosome 5 that appear in the same order. Pairs of homologous genes that map to regions of conserved synteny between *P. y. yoelii* and *P. falciparum* are probably orthologues, confirmed by the finding that most of these homologous pairs are also reciprocal best matches between the *P. falciparum* and *P. y. yoelii* proteins. Genes in the synteny gap on chromosome 10 (Fig. 2) include a glutamate-rich protein, S antigen, MSP3, MSP6 and liver stage antigen 1, several of which are prime vaccine antigen candidates in *P. falciparum*. Genes in the synteny gap on chromosome 4 include four *var* and two *rif* genes, which make up one of the four internal clusters of *var/rif* genes found in *P. falciparum* (see ref. 30). A series of uncharacterized hypothetical genes occur on the contigs that overlap these regions in *P. y. yoelii*.

An intriguing finding from the study of chromosome 5 has been the analysis of the syntenic break point between *P. falciparum* chromosomes 4 and 10. The final *P. y. yoelii* contig in the tiling path with significant synteny to *P. falciparum* chromosome 10 also contains the external transcribed sequence (ETS) of the SSU rRNA C unit. The synteny resumes on *P. falciparum* chromosome 4 in a *P. y. yoelii* contig that also contains the ETS of the large subunit (LSU) of the same rRNA unit. (No rRNA unit sequences are located on *P. falciparum* chromosomes 4 and 10; matches to contigs containing these genes occur in coding regions of other genes.) Both *P. y. yoelii* contigs are linked to each other through a third contig that contains the remaining elements (SSU, 5.8S, LSU, and internal transcribed sequences 1 and 2) of the complete rRNA unit (Fig. 2). Thus it seems that the break in synteny between *Plasmo-*

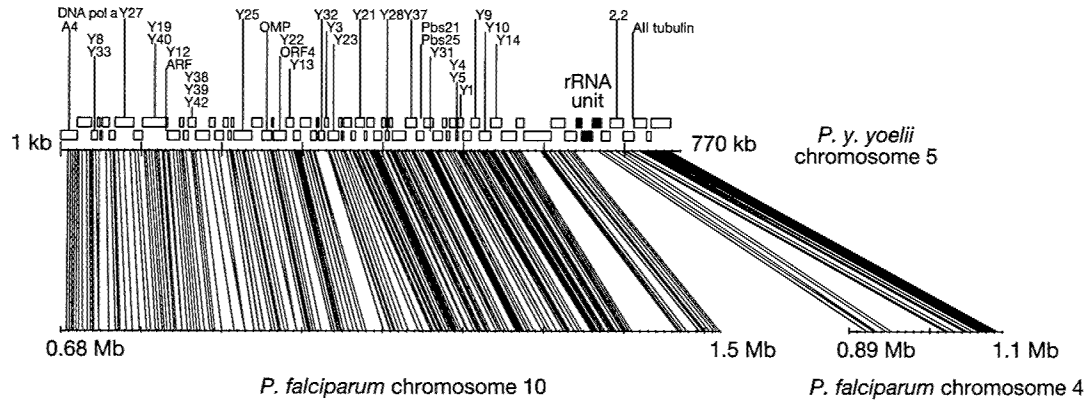


Figure 2 Conservation of gene synteny between *P. y. yoelii* chromosome 5 and *P. falciparum* chromosomes 4 and 10. Physical marker data used to confirm contig order in the tiling path of *P. y. yoelii* chromosome 5 are shown above the contigs (open boxes). Each coloured line represents a pair of orthologous genes present in the two species shown anchored to its respective location in the two genomes. Contigs containing the *P. y. yoelii* rRNA unit are shown as filled boxes.

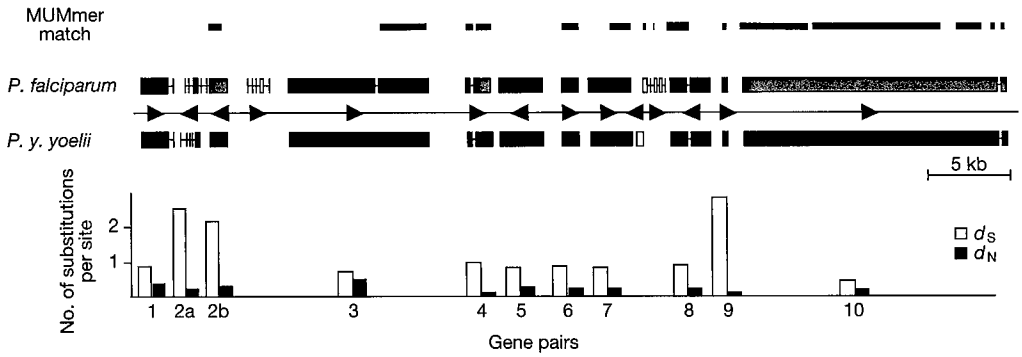


Figure 3 Global alignment scheme of a syntenic region between *P. falciparum* and *P. y. yoelii* encompassing ten orthologous gene pairs and nine intergenic regions. White boxes represent genes that have no orthologue and were excluded from analysis; green boxes represent gene models that were refined; red boxes represent unaltered gene models; arrowheads represent gene orientation on the DNA molecule. Clusters of MUMmer matches between the two species are represented as thick blue lines. For the ten orthologous gene pairs, synonymous mutations per synonymous site (d_S, open bars) and non-synonymous mutations per non-synonymous site (d_N, filled bars) were estimated and plotted.

dium chromosomes has occurred within a single rRNA unit, a phenomenon first reported in prokaryotes[43]. Six rRNA units reside as individual operons on *P. falciparum* chromosomes 1, 5, 7, 8, 11 and 13 respectively (ref. 30), in contrast to rodent malaria species that have four[44]. Intriguingly, breaks in the synteny between *P. y. yoelii* and *P. falciparum* can be mapped to almost all rRNA unit loci on the *P. falciparum* chromosomes (see Fig. 1 of ref. 30). A full analysis of this potential phenomenon is outside the scope of this study, but these results provide preliminary evidence for one possible mechanism underlying synteny breakage that may have occurred during evolution of the *Plasmodium* genus—that of chromosome breakage and recombination at sites of rRNA units.

Comparative alignment of syntenic regions

Recent comparative studies have revealed that the fine detail of short stretches of the rodent and human malaria parasite genomes is remarkably conserved[45], and that such comparisons are useful for gene prediction and evolutionary studies. Accordingly, we used a comparison of the longest assembly of *P. y. yoelii* (MALPY00395, 51.3 kb) and its syntenic region in *P. falciparum* (chromosome 7, at coordinates 1,131–1,183 kb) as a case study for a preliminary evolutionary analysis of the two genomes. Gene prediction programs run against these two regions identified 11 genes in the syntenic region of both species (Fig. 3), eight of which are orthologous gene pairs (genes 1, 3–8 and 10). The structures of two additional gene pairs (genes 2a/b and 9) were refined through manual curation of erroneous gene boundaries. Three hypothetical genes, two in *P. falciparum* and one in *P. y. yoelii*, had no discernible orthologue in the other species; the presence of multiple stop codons in these areas suggests that the genes may have become pseudogenes. A global alignment at the DNA level of the syntenic region (Fig. 3) reveals the similarity between species in intergenic regions to be almost negligible, as mirrored in similar syntenic comparisons of mouse and human[46,47]. Moreover, the mutation saturation observed in intergenic regions suggests that 'phylogenetic footprinting' can be used to identify conserved motifs between species that may be involved in gene regulation.

In contrast to intergenic regions, the similarity between species in coding regions is relatively high. The average number of non-synonymous substitutions per non-synonymous site, d_N, between the two species is 26% (±12%). Synonymous sites, d_S, are saturated (average $d_S > 1$), which supports the lack of similarity observed within intergenic regions. These values are considerably higher than those reported for human–rodent comparisons, which are approximately 7.5% and 45% for non-synonymous and synonymous substitutions, respectively[48]. The cause of such apparent disparities remains unknown, but may be a consequence of extreme genome composition or the short generation time of the parasite.

Rodent malaria species as models for *P. falciparum* biology

The usefulness of rodent malaria species as models for the study of *P. falciparum* is controversial. It is apparent that rodent models are the first port of call when preliminary *in vivo* evidence of antimalarial drug efficacy, immune response to vaccine candidates, and life-cycle adaptations in the face of drug or vaccine challenge are required. Different species of malaria parasite have developed different mechanisms of resistance to the antimalarial drug chloroquine, despite a similar mode of action of the drug (reviewed in ref. 49). It seems that mechanisms developed by the parasite to evade an inhospitable environment, whether caused by antimalarial drugs or the host immune system, may differ widely from species to species. A model involving evolution of different genes in *Plasmodium* species as a response to different host environments is consistent with the comparison of the *P. falciparum* and *P. y. yoelii* genomes presented here; conservation of synteny between the two species is high in regions of housekeeping genes, but not in regions where genes involved in antigenic variation and evasion of the host immune system are located. On the one hand, this can be interpreted as a blow to the systematic identification of all orthologues of antigen genes between *P. falciparum* and *P. y. yoelii* that could be used in the design of a malaria vaccine. On the other hand, a picture is emerging of selecting a model malaria species based on the complement of genes that best fit the phenotypic trait under study. Thus the presence of homologues of the *yir* family may make *P. y. yoelii* an attractive model for studying antigenic variation in *P. vivax*. Furthermore, identification of orthologues in the genomes of relatively distant rodent and human malaria parasites will facilitate finding orthologues in other model malaria species, for example monkey models of malaria such as *Plasmodium knowlesi*. □

Methods

Genome and EST sequencing

Plasmodium yoelii yoelii 17XNL line[50], selected from an isolate taken from the blood of a wild-caught thicket rat in the Central African Republic[51], is a non-lethal strain with a preference for development in reticulocytes. Clone 1.1 was obtained through serial dilution of sporozoites. Parasites were grown in laboratory mice no more than three blood passages from mosquito passage to limit chromosome instability, collected by exsanguination into heparin, and host mouse leukocytes were removed by filtration. Small insert libraries (average insert size 1.6 kb) were constructed in pUC-derived vectors after nebulization of genomic DNA. DNA sequencing of plasmid ends used ABI Big Dye terminator chemistry on ABI3700 sequencing machines. A total of 222,716 sequences (82% success rate), averaging 662 nucleotides in length, were assembled using TIGR Assembler[7]. BLASTN of the *P. y. yoelii* contigs and singletons against the complete set of

Celera mouse contigs[8], using a cutoff of 90% identity over 100 nucleotides, identified contaminating mouse sequences that were subsequently removed. Contigs were assigned to groups using Grouper[52]. Each contig was assigned an identifier in the format 'MALPY00001'.

Proteomic analysis

MudPIT technology and methods were as described in ref. 23. Sporozoites of *P. y. yoelii* were dissected from infected *Anopheles stephensi* mosquito salivary glands, and *P. y. yoelii* gametocytes were prepared as described[53]. Cellular debris from uninfected mosquitoes and mouse erythrocytes were analysed as controls. Tandem mass spectrometry (MS/MS) data sets were searched against several databases: the complete set of *P. y. yoelii* full and partial proteins (7,860 total); 791,324 *P. y. yoelii* open reading frames (stop-to-stop ORFs over 15 amino acids and start-to-stop ORFs over 100 amino acids); 57,885 ORFs from NCBI's RefSeq for human, mouse and rat; 15,570 *Anopheles*, *Aedes* and *Drosophila melanogaster* proteins from GenBank; and 165 common protein contaminants (for example, trypsin, bovine serum albumin).

Gene finding and annotation

The splice site recognition module of GlimmerMExon was trained specifically for *P. yoelii* genome data, using DNA sequences extracted from a set of 1,166 donor and 1,166 acceptor sites confirmed by *P. y. yoelii* ESTs. Phat and the exon recognition module of GlimmerMExon were trained on *P. falciparum* data as described (see ref. 54). Combiner was used to generate a final ranked list of *P. y. yoelii* gene models, and TIGR's Eukaryotic Genome Control suite of programs was used for automated annotation of these (both described in ref. 54). Automated gene names were assigned to proteins by taking the 'equivalogue' name of the hidden Markov model (HMM) associated with the protein where possible, or where no HMM was assigned, on the basis of the best-paired alignment. Each protein was assigned an identifier in the format 'PY00001'.

Paralogous gene families

Proteins encoded by multigene families were identified by a domain-based clustering algorithm developed at TIGR. Families were regarded as potentially *Plasmodium*- or *yoelii*-specific if they were not described by any Pfam[55] or TIGRFAM[56] domains and if the automatic annotation process had not ascribed names corresponding to widely distributed proteins. HMMs for these families were built using the HMMER package version 2.1.1 (ref. 57). Newly constructed models were then used to search the *P. yoelii*, *P. falciparum* and GenBank databases to define the scope of the families.

Telomeric/subtelomeric repeat analysis

Subtelomeric contigs were identified through alignment using MUMmer2 (ref. 40) with a minimum exact match ranging from 30–40 bases. Tandem Repeat Finder[58] used the following settings: match = 2, mismatch = 7, PM (match probability) = 75, PI (indel probability) = 10, minscore = 400, max period = 700.

Comparative analyses

Gene model predictions in the syntenic region of *P. falciparum* chromosome 7 were inspected manually, and bi-directional best hits between gene models that respected conserved syntenies were selected. A global alignment of the two sequences was calculated using Owen[59], and nucleotide sequences of predicted gene models were aligned using CLUSTALW[60] with default parameters, and refined manually. The number of substitutions per synonymous (d_S) and nonsynonymous (d_N) sites were estimated using the Nei and Gojobori method[61]. Conservation of gene order was established using Position Effect (http://www.tigr.org/software), where matches between *P. falciparum* and *P. y. yoelii* genes were calculated using BLASTP with a cutoff E value of 10^{-15}. The query and hit gene from each match were defined as anchor points in gene sets composed of adjacent genes. Up to ten genes upstream and downstream from each anchor gene were used in creating the gene set. An optimal alignment was calculated between the ordered gene sets using BLASTP per cent similarity scores and a linear gap penalty. Low-scoring alignments with a cumulative per cent similarity less than 100 were not used. Each optimal alignment provided a list of matching genes in conserved order between *P. falciparum* and *P. y. yoelii*.

Received 31 July; accepted 30 August 2002; doi:10.1038/nature01099.

1. Carter, R. & Diggs, C. L. *Parasitic Protozoa* 359–465 (Academic, New York/San Francisco/London, 1977).
2. van Dijk, M. R., Waters, A. P. & Janse, C. J. Stable transfection of malaria parasite blood stages. *Science* **268,** 1358–1362 (1995).
3. Carlton, J. M. & Carucci, D. J. Rodent models of malaria in the genomics era. *Trends Parasitol.* **18,** 100–102 (2002).
4. Myers, E. W. *et al.* A whole-genome assembly of *Drosophila*. *Science* **287,** 2196–2204 (2000).
5. Lander, E. S. & Waterman, M. S. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2,** 231–239 (1988).
6. McCutchan, T. F., Dame, J. B., Miller, L. H. & Barnwell, J. Evolutionary relatedness of *Plasmodium* species as determined by the structure of DNA. *Science* **225,** 808–811 (1984).
7. Sutton, G. G., White, O., Adams, M. D. & Kervalage, A. R. TIGR Assembler: a new tool for assembling large shotgun sequencing projects. *Genome Sci. Technol.* **1,** 9–19 (1995).
8. Mural, R. J. *et al.* A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science* **296,** 1661–1671 (2002).
9. Janse, C. J., Carlton, J. M.-R., Walliker, D. & Waters, A. P. Conserved location of genes on polymorphic chromosomes of four species of malaria parasites. *Mol. Biochem. Parasitol.* **68,** 285–296 (1994).
10. Daly, T. M., Long, C. A. & Bergman, L. W. Interaction between two domains of the *P. yoelii* MSP-1 protein detected using the yeast two-hybrid system. *Mol. Biochem. Parasitol.* **117,** 27–35 (2001).
11. Quackenbush, J. *et al.* The TIGR gene indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Res.* **29,** 159–164 (2001).
12. Cawley, S. E., Wirth, A. I. & Speed, T. P. Phat—a gene finding program for *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **118,** 167–174 (2001).
13. Salzberg, S. L., Pertea, M., Delcher, A. L., Gardner, M. J. & Tettelin, H. Interpolated Markov models for eukaryotic gene finding. *Genomics* **59,** 24–31 (1999).
14. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
15. Thompson, J., Janse, C. J. & Waters, A. P. Comparative genomics in *Plasmodium*: a tool for the identification of genes and functional analysis. *Mol. Biochem. Parasitol.* **118,** 147–154 (2001).
16. Jomaa, H. *et al.* Inhibitors of the nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs. *Science* **285,** 1573–1576 (1999).
17. 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).
18. Ohkanda, J. *et al.* Peptidomimetic inhibitors of protein farnesyltransferase show potent antimalarial activity. *Bioorg. Med. Chem. Lett.* **11,** 761–764 (2001).
19. The Gene Ontology Consortium. Creating the gene ontology resource: design and implementation. *Genome Res.* **11,** 1425–1433 (2001).
20. Cooke, B. M., Mohandas, N. & Coppel, R. L. The malaria-infected red blood cell: structural and functional changes. *Adv. Parasitol.* **50,** 1–86 (2001).
21. del Portillo, H. A. *et al.* A superfamily of variant genes encoded in the subtelomeric region of *Plasmodium vivax*. *Nature* **410,** 839–842 (2001).
22. Janssen, C. S., Barrett, M. P., Turner, C. M. & Phillips, R. S. A large gene family for putative variant antigens shared by human and rodent malaria parasites. *Proc. R. Soc. Lond. B* **269,** 431–436 (2002).
23. Florens, L. *et al.* A proteomic view of the *Plasmodium falciparum* life cycle. *Nature* **419,** 520–526 (2002).
24. Preiser, P. R., Jarra, W., Capiod, T. & Snounou, G. A rhoptry-protein-associated mechanism of clonal phenotypic variation in rodent malaria. *Nature* **398,** 618–622 (1999).
25. Galinski, M. R., Xu, M. & Barnwell, J. W. *Plasmodium vivax* reticulocyte binding protein-2 (PvRBP-2) shares structural features with PvRBP-1 and the *Plasmodium yoelii* 235 kDa rhoptry protein family. *Mol. Biochem. Parasitol.* **108,** 257–262 (2000).
26. Rayner, J. C., Galinski, M. R., Ingravallo, P. & Barnwell, J. W. Two *Plasmodium falciparum* genes express merozoite proteins that are related to *Plasmodium vivax* and *Plasmodium yoelii* adhesive proteins involved in host cell selection and invasion. *Proc. Natl Acad. Sci. USA* **97,** 9648–9653 (2000).
27. Wiser, M. F., Giraldo, L. E., Schmitt-Wrede, H. P. & Wunderlich, F. *Plasmodium chabaudi*: immunogenicity of a highly antigenic glutamate-rich protein. *Exp. Parasitol.* **85,** 43–54 (1997).
28. Spielmann, T. & Beck, H. P. Analysis of stage-specific transcription in *Plasmodium falciparum* reveals a set of genes exclusively transcribed in ring stage parasites. *Mol. Biochem. Parasitol.* **111,** 453–458 (2000).
29. Favaloro, J. M., Culvenor, J. G., Anders, R. F. & Kemp, D. J. A *Plasmodium chabaudi* antigen located in the parasitophorous vacuole membrane. *Mol. Biochem. Parasitol.* **62,** 263–270 (1993).
30. Gardner, M. J. *et al.* Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419,** 498–511 (2002).
31. Scherf, A., Figueiredo, L. M. & Freitas-Junior, L. H. *Plasmodium* telomeres: a pathogen's perspective. *Curr. Opin. Microbiol.* **4,** 409–414 (2001).
32. Mu, J. *et al.* Chromosome-wide SNPs reveal an ancient origin for *Plasmodium falciparum*. *Nature* **418,** 323–326 (2002).
33. Garnham, P. C. C. *Malaria Parasites and Other Haemosporidia* (Blackwell Scientific, Oxford, 1966).
34. Escalante, A. A. & Ayala, F. J. Phylogeny of the malarial genus *Plasmodium*, derived from rRNA gene sequences. *Proc. Natl Acad. Sci. USA* **91,** 11373–11377 (1994).
35. Waters, A. P., Higgins, D. G. & McCutchan, T. F. *Plasmodium falciparum* appears to have arisen as a result of lateral transfer between avian and human hosts. *Proc. Natl Acad. Sci. USA* **88,** 3140–3144 (1991).
36. Tchavtchitch, M., Fischer, K., Huestis, R. & Saul, A. The sequence of a 200 kb portion of a *Plasmodium vivax* chromosome reveals a high degree of conservation with *Plasmodium falciparum* chromosome 3. *Mol. Biochem. Parasitol.* **118,** 211–222 (2001).
37. Carlton, J. M.-R., Galinski, M. R., Barnwell, J. W. & Dame, J. B. Karyotype and synteny among the chromosomes of all four species of human malaria parasite. *Mol. Biochem. Parasitol.* **101,** 23–32 (1999).
38. Janse, C. J. Chromosome size polymorphism and DNA rearrangements in *Plasmodium*. *Parasitol. Today* **9,** 19–22 (1993).
39. Carlton, J. M. R., Vinkenoog, R., Waters, A. P. & Walliker, D. Gene synteny in species of *Plasmodium*. *Mol. Biochem. Parasitol.* **93,** 285–294 (1998).
40. 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).
41. van Lin, L. H. M., Pace, T., Janse, C. J., Scotti, R. & Ponzi, R. A long range restriction map of chromosomes 5 of *Plasmodium berghei* demonstates a chromosomes specific symmetrical subtelomeric organisation. *Mol. Biochem. Parasitol.* **86,** 111–115 (1997).
42. Janse, C. J., Ramesar, J., van den Berg, F. M. & Mons, B. *Plasmodium berghei*: *in vivo* generation and selection of karyotype mutants and non-gametocyte producer mutants. *Exp. Parasitol.* **74,** 1–10 (1992).
43. Liu, S. L. & Sanderson, K. E. Rearrangements in the genome of the bacterium *Salmonella typhi*. *Proc. Natl Acad. Sci. USA* **92,** 1018–1022 (1995).
44. Dame, J. B. & McCutchan, T. F. The four ribosomal DNA units of the malaria parasite *Plasmodium berghei*. Identification, restriction map and copy number analysis. *J. Biol. Chem.* **258,** 6984–6990 (1983).
45. van Lin, L. H. *et al.* Interspecies conservation of gene order and intron-exon structure in a genomic locus of high gene density and complexity in *Plasmodium*. *Nucleic Acids Res.* **29,** 2059–2068 (2001).
46. Jareborg, N., Birney, E. & Durbin, R. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Res.* **9,** 815–824 (1999).
47. Shabalina, S. A., Ogurtsov, A. Y., Kondrashov, V. A. & Kondrashov, A. S. Selective constraint in intergenic regions of human and mouse genomes. *Trends Genet.* **17,** 373–376 (2001).
48. Makalowski, W. & Boguski, M. S. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. *Proc. Natl Acad. Sci. USA* **95,** 9407–9412 (1998).
49. Carlton, J. M., Fidock, D. A., Djimde, A., Plowe, C. V. & Wellems, T. E. Conservation of a novel

vacuolar transporter in *Plasmodium* species and its central role in chloroquine resistance of *P. falciparum*. *Curr. Opin. Microbiol.* **4,** 415–420 (2001).
50. Weinbaum, F. I., Evans, C. B. & Tigelaar, R. E. An *in vitro* assay for T cell immunity to malaria in mice. *J. Immunol.* **116,** 1280–1283 (1976).
51. Landau, I. & Chabaud, A. G. Natural infection by 2 plasmodia of the rodent *Thamnomys rutilans* in the Central African Republic. *C.R. Acad. Sci. Hebd. Seances Acad. Sci. D* **261,** 230–232 (1965).
52. Gardner, M. J. *et al.* Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science* **282,** 1126–1132 (1998).
53. Beetsma, A. L., van de Wiel, T. J., Sauerwein, R. W. & Eling, W. M. *Plasmodium berghei* ANKA: purification of large numbers of infectious gametocytes. *Exp. Parasitol.* **88,** 69–72 (1998).
54. Gardner, M. J. *et al.* Sequence of *Plasmodium falciparum* chromosomes 2, 10, 11 and 14. *Nature* **419,** 531–534 (2002).
55. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **30,** 276–280 (2002).
56. Haft, D. H. *et al.* TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Res.* **29,** 41–43 (2001).
57. Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14,** 755–763 (1998).
58. Benson, G. Tandem repeats finder: a program to analyse DNA sequences. *Nucleic Acids Res.* **27,** 573–580 (1999).
59. Ogurtsov, A. Y., Roytberg, M. A., Shabalina, S. A. & Kondrashov, A. S. OWEN: aligning long collinear regions of genomes. *Bioinformatics* (in the press).
60. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22,** 4673–4680 (1994).
61. Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3,** 418–426 (1986).
62. Black, C. G., Wang, L., Hibbs, A. R., Werner, E. & Coppel, R. L. Identification of the *Plasmodium chabaudi* homologue of merozoite surface proteins 4 and 5 of *Plasmodium falciparum*. *Infect. Immun.* **67,** 2075–2081 (1999).

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

Acknowledgements

We thank S. Cawley and T. Pace for collaborative work; J. Mendoza and J. Ramesar for technical support; C. Long for the gift of a *P. y. yoelii* cDNA library; R. Arcilla and W. Weiss for parasite material; and J. Eisen and S. Sullivan for critical reading of the manuscript. L.H.v.L. was supported by an INCO-DEV programme grant from the European Community; T.W.K. was supported by a Rijks Universiteit te Leiden studentship; J.D.R. was supported with funds from the Wellcome Trust. This project was funded by the US Department of Defense through cooperative agreement with the US Army Medical Research and Material Command and by the Naval Medical Research Center. The opinions expressed are those of the authors and do not reflect the official policy of the Department of the Navy, Department of Defense, or the US government.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.M.C. (e-mail: carlton@tigr.org). Access to genome annotation data is available through the TIGR Eukaryotic Projects website (http://www.tigr.org) and PlasmoDB (http://www.plasmodb.org). This whole-genome shotgun project has been deposited at DDBJ/EMBL/GenBank under the project accession number AABL00000000. The version described in this paper is the first version, AABL01000000.

A proteomic view of the *Plasmodium falciparum* life cycle

Laurence Florens*, Michael P. Washburn†, J. Dale Raine‡, Robert M. Anthony§, Munira Grainger‖, J. David Haynes§¶, J. Kathleen Moch§, Nemone Muster*, John B. Sacci§#, David L. Tabb*☆, Adam A. Witney§#, Dirk Wolters†#, Yimin Wu, Malcolm J. Gardner††, Anthony A. Holder‖, Robert E. Sinden‡, John R. Yates*† & Daniel J. Carucci§**

* *Department of Cell Biology, The Scripps Research Institute, SR-11, 10550 North Torrey Pines Road, La Jolla, California 92037, USA*
† *Department of Proteomics and Metabolomics, Torrey Mesa Research Institute, Syngenta Research & Technology, 3115 Merryfield Row, San Diego, California 92121-1125, USA*
‡ *Infection and Immunity Section, Department of Biological Sciences, Imperial College of Science, Technology & Medicine, Sir Alexander Fleming Building, South Kensington, London SW7 2AZ, UK*
§ *Naval Medical Research Center, Malaria Program (IDD), 503 Robert Grant Avenue, Room 3A40; and* ¶ *Department of Immunology, Walter Reed Army Institute of Research, Silver Spring, Maryland 20910-7500, USA*
‖ *The Division of Parasitology, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK*
☆ *Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA*
** *Malaria Research and Reference Reagent Resource Center, American Type Culture Collection, 10801 University Boulevard, Manassas, Virginia 20110-2209, USA*
†† *The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850, USA*

The completion of the *Plasmodium falciparum* clone 3D7 genome provides a basis on which to conduct comparative proteomics studies of this human pathogen. Here, we applied a high-throughput proteomics approach to identify new potential drug and vaccine targets and to better understand the biology of this complex protozoan parasite. We characterized four stages of the parasite life cycle (sporozoites, merozoites, trophozoites and gametocytes) by multidimensional protein identification technology. Functional profiling of over 2,400 proteins agreed with the physiology of each stage. Unexpectedly, the antigenically variant proteins of *var* and *rif* genes, defined as molecules on the surface of infected erythrocytes, were also largely expressed in sporozoites. The detection of chromosomal clusters encoding co-expressed proteins suggested a potential mechanism for controlling gene expression.

The life cycle of *Plasmodium* is extraordinarily complex, requiring specialized protein expression for life in both invertebrate and vertebrate host environments, for intracellular and extracellular survival, for invasion of multiple cell types, and for evasion of host immune responses. Interventional strategies including anti-malarial vaccines and drugs will be most effective if targeted at specific parasite life stages and/or specific proteins expressed at these stages. The genomes of *P. falciparum*[1] and *P. yoelii yoelii*[2] are now completed and offer the promise of identifying new and effective drug and vaccine targets.

Functional genomics has fundamentally changed the traditional gene-by-gene approach of the pre-genomic era by capitalizing on the success of genome sequencing efforts. DNA microarrays have been successfully used to study differential gene expression in the abundant blood stages of the *Plasmodium* parasite[3,4]. However, transcriptional analysis by DNA microarrays generally requires microgram quantities of RNA and has been restricted to stages that can be cultivated *in vitro*, limiting current large-scale gene expression analyses to the blood stages of *P. falciparum*. As several key stages of the parasite life cycle, in particular the pre-erythrocytic stages, are not readily accessible to study, and as differential gene expression is in fact a surrogate for protein expression, global proteomic analyses offer a unique means of determining not only protein expression, but also subcellular localization and post-translational modifications.

We report here a comprehensive view of the protein complements isolated from sporozoites (the infectious form injected by the mosquito), merozoites (the invasive stage of the erythrocytes), trophozoites (the form multiplying in erythrocytes), and gametocytes (sexual stages) of the human malaria parasite *P. falciparum*. These proteomes were analysed by multidimensional protein identification technology (MudPIT), which combines in-line, high-resolution liquid chromatography and tandem mass spectrometry[5]. Two levels of control were implemented to differentiate parasite from host proteins. By using combined host–parasite sequence databases and noninfected controls, 2,415 parasite proteins were confidently identified out of thousands of host proteins; that is, 46% of all gene products were detected in four stages of the *Plasmodium* life cycle (Supplementary Table 1).

Comparative proteomics throughout the life cycle

The sporozoite proteome appeared markedly different from the other stages (Table 1). Almost half (49%) of the sporozoite proteins

Table 1 **Comparative summary of the protein lists for each stage**

Protein count	Sporozoites	Merozoites	Trophozoites	Gametocytes
152	X	X	X	X
197	–	X	X	X
53	X	–	X	X
28	X	X	–	X
36	X	X	X	–
148	–	–	X	X
73	–	X	–	X
120	X	–	–	X
84	–	X	X	–
80	X	–	X	–
65	X	X	–	–
376	–	–	–	X
286	–	–	X	–
204	–	X	–	–
513	X	–	–	–
2,415	1,049	839	1,036	1,147

Whole-cell protein lysates were obtained from, on average, 17×10^6 sporozoites, 4.5×10^9 trophozoites, 2.75×10^9 merozoites, and 6.5×10^9 gametocytes.

Present addresses: BRB 13-009, Department of Microbiology and Immunology, University of Maryland School of Medicine, 655 W. Baltimore St., Baltimore, Maryland 21201, USA (J.B.S.); Department of Medical Microbiology, St George's Hospital Medical School, Cranmer Terrace, London SW17 0RE, UK (A.A.W.); and Ruhr-University Bochum, Institute of Analytical Chemistry, 44780 Bochum, Germany (D.W.).

were unique to this stage, which shared an average of 25% of its proteins with any other stage. On the other hand, trophozoites, merozoites and gametocytes had between 20% and 33% unique proteins, and they shared between 39% and 56% of their proteins. Consequently, only 152 proteins (6%) were common to all four stages. Those common proteins were mostly housekeeping proteins such as ribosomal proteins, transcription factors, histones and cytoskeletal proteins (Supplementary Table 1). Proteins were sorted into main functional classes based on the Munich Information Centre for Protein Sequences (MIPS) catalogue[6], with some adaptations for classes specific to the parasite, such as cell surface and apical organelle proteins (Fig. 1). When considering the annotated proteins in the database, some marked differences appeared between sporozoites and blood stages (Fig. 1). Although great care was taken to ensure that the results reflect the state of the parasite in the host, a portion of the data set may reflect the parasite's response to different purification treatments. However, the stage-specific detection of known protein markers at each stage established the relevance of our data set.

The merozoite proteome

Merozoites are released from an infected erythrocyte, and after a short period in the plasma, bind to and invade new erythrocytes. Proteins on the surface and in the apical organelles of the merozoite mediate cell recognition and invasion in an active process involving an actin-myosin motor. Four putative components of the invasion motor[7], merozoite cap protein-1 (MCP1), actin, myosin A, and myosin A tail domain interacting protein (MTIP), were abundant merozoite proteins (Supplementary Table 2). Abundant merozoite surface proteins (MSPs) such as MSP1 and MSP2 are linked by a glycosylphosphatidyl (GPI) anchor to the membrane, and both have been implicated in immune evasion (reviewed in ref. 8). A second family of peripheral membrane proteins, represented by MSP3 and MSP6, was also detected (Fig. 2a), although these proteins are largely soluble proteins of the parasitophorous vacuole, which are released on schizont rupture. Other vacuolar proteins, such as the acidic basic repeat antigen (ABRA) and serine repeat antigen (SERA), were detected in the merozoite fraction, but some such as S-antigen[9] were not (Supplementary Table 2). Notably, MSP8 and a related MSP8-like protein were only identified in sporozoites (Fig. 2a). Some MSPs are diverse in sequence and may be extensively modified by proteolysis; these features, together with the association of a variety of peripheral and soluble proteins, provide for a complex surface architecture.

Many apical organellar proteins, in the micronemes and rhoptries, have a single transmembrane domain. Among these proteins, apical membrane antigen 1 (AMA1) and MAEBL were found in both sporozoite and merozoite preparations (Fig. 2a). Erythrocyte-binding antigens (EBA), such as EBA 175 and EBA 140/BAEBL, were found only in the merozoite and trophozoite fractions. Of note, the reticulocyte-binding protein (PfRH) family (PFD0110w, MAL13P1.176, PF13_01998, PFL2520w and PFD1150c), which has similarity with the Py235 family of *P. y. yoelii* rhoptry proteins and the *Plasmodium vivax* reticulocyte-binding proteins, was not detected in the merozoite fraction. Some PfRH proteins were, however, detected in sporozoites (Fig. 2a), including RH3, which is a transcribed pseudogene in blood stages[10]. Components of the low molecular mass rhoptry complex, the rhoptry-associated proteins (RAP) 1, 2 and 3, were all found in merozoites. RAP1 was also detected in sporozoites. The high molecular mass rhoptry protein complex (RhopH), together with ring-infected erythrocyte surface antigen (RESA), which is a component of dense granules, is transferred intact to new erythrocytes at or after invasion and may contribute to the host cell remodelling process. RhopH1, RhopH2 (PFI1445w; Ling, I. T., *et al.*, unpublished data) and RhopH3 were found in the merozoite proteome. RhopH1 (PFC0120w/PFC0110w) has been shown to be a member of the cyto-adherence linked asexual gene family (CLAG)[11]; however, the presence of CLAG9 in the merozoite fraction (Fig. 2a) suggests that CLAG9 may also be a RhopH protein, casting some doubt on the proposed role for this protein in cyto-adherence[12].

The trophozoite proteome

After erythrocyte invasion the parasite modifies the host cell. The principal modifications during the initial trophozoite phase (lasting about 30 h) allow the parasite to transport molecules in and out of the cell, to prepare the surface of the red blood cell to mediate cyto-adherence, and to digest the cytoplasmic contents, particularly haemoglobin, in its food vacuole. In the next phase of schizogony (the final ~18 h of the asexual development in the blood cell), nuclear division is followed by merozoite formation and release.

Knob-associated histidine-rich protein (KAHRP) and erythrocyte membrane proteins 2 and 3 (EMP2 and -3) bind to the erythrocyte cytoskeleton (Fig. 2a). Of the proteins of the parasitophorous vacuole and the tubovesicular membrane structure extending into the cytoplasm of the red blood cell, three (the skeleton-binding protein 1, and exported proteins EXP1 and EXP2) were represented by peptides (Fig. 2a); although a fourth (Sar1 homologue, small GTP-binding protein; PFD0810w) was not. It is likely that one or more of the hypothetical proteins detected only in the trophozoite sample are involved in these unusual structures.

Digestion of haemoglobin is a major parasite catabolic process[13]. Members of the plasmepsin family (aspartic proteinases; PF14_0075 to PF14_0078)[14], falcipain family (cysteine proteinases; PF11_0161,

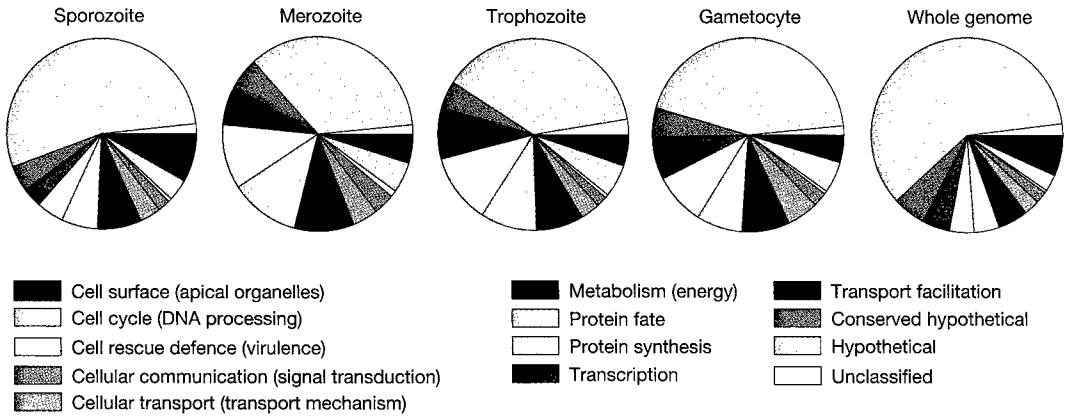


Figure 1 Functional profiles of expressed proteins. Proteins identified in each stage are plotted as a function of their broad functional classification as defined by the MIPS catalogue[6]. To avoid redundancy, only one class was assigned per protein. The complete protein list is given in Supplementary Table 1.

PF11_0162 and PF11_0165)[15], and falcilysin (a metallopeptidase; PF13_0322)[16] implicated in this process were all clearly identified (Supplementary Table 1). Several proteases expressed in the merozoite and trophozoite fractions, and not involved in haemoglobin digestion, may be important in parasite release at the end of schizogony, invasion of the new cell, or merozoite protein processing. Possible candidates for this mechanism include cysteine proteinases of the falcipain and SERA families, or subtilisins such as SUB1 and SUB2, both located in apical organelles (Fig. 2a).

The gametocyte proteome

Stage V gametocytes are dimorphic, with a male:female ratio of 1:4. They are arrested in the cell cycle until they enter the mosquito where development is induced within minutes to form the male and female gametes. Gametocyte structure reflects these ensuing fates; that is, the female has abundant ribosomes and endoplasmic reticulum/vesicular network to re-initiate translation, whereas the male is largely devoid of ribosomes and is terminally differentiated[17].

Gametocyte-specific transcription factors, RNA-binding proteins, and gametocyte-specific proteins involved in the regulation of messenger RNA processing (particularly splicing factors, RNA helicases, RNA-binding proteins, ribonucleoproteins (RNPs) and small nuclear ribonucleoprotein particles (snRNPS)) were highly represented in the gametocyte proteome (Supplementary Table 1). Transcription in the terminally differentiated gametocytes is 'suppressed', but the female gametocytes contain mRNAs encoding gamete/zygote/ookinete surface antigens (for example, P25/28)

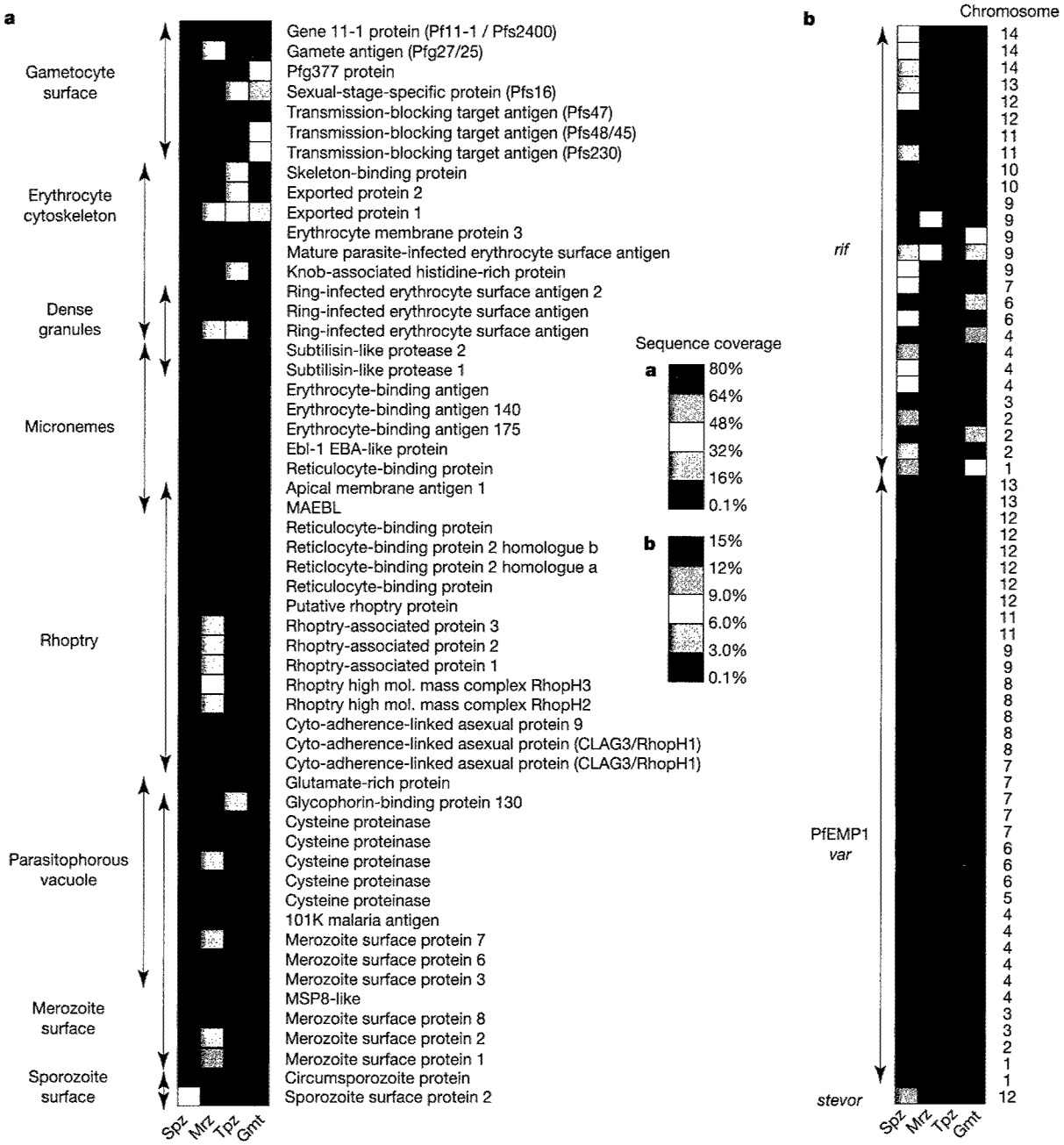


Figure 2 Expression patterns of known stage-specific proteins. **a**, Cell surface, organelle, and secreted proteins are plotted as a function of their known subcellular localization. **b**, *stevor*, *var* and *rif* polymorphic surface variants are plotted as a function of the chromosome encoding their genes. The matrices are colour-coded by sequence coverage measured in each stage (proteins not detected in a stage are represented by black squares). Locus names associated with these proteins are listed in Supplementary Table 2. Spz, sporozoite; mrz, merozoite; tpz, trophozoite; gmt, gametocyte.

that are subject to post-transcriptional control; this control is released rapidly during gamete development[17]. Ribosomal proteins were largely represented: 82% of known small subunit (SSU) proteins and 69% of known large subunit (LSU) proteins were detected in gametocytes compared to 94% and 82%, respectively, from all stages examined (Supplementary Table 1). We suggest that this reflects the accumulation of ribosomes in the female gametocyte to accommodate for the sudden increase in protein synthesis required during gametogenesis and early zygote development.

Other protein groupings highly represented in the gametocyte were in the cell cycle/DNA processing and energy classes (Fig. 1). The former is consistent with the biological observation that the mature gametocyte is arrested in G0 of the cell cycle and will require a full complement of pre-existing cell cycle regulatory cascades to respond, within seconds, to the gametogenesis stimuli (that is, xanthurenic acid and a drop in temperature)[18]. Metabolic pathways of the malaria parasite may be stage-specific, with asexual blood stage parasites dependent on glycolysis and conversion of pyruvate to lactate (L-lactate dehydrogenase) for energy. In the gametocyte and sporozoite preparations, peptides from enzymes involved in the mitochondrial tricarboxylic acid (TCA) cycle and oxidative phosphorylation were identified (Table 2). This observation suggests that gametocytes have fully functional mitochondria as a pre-adaptation to life in the mosquito, as suggested by morphological and biochemical studies[19] and their sensitivity to anti-malarials attacking respiration (primaquine and artimesinin-based products)[17]. It will be interesting to observe whether other mosquito and liver stages, which show similar drug sensitivities, express the same metabolic proteome.

Cell surface proteins (Fig. 1) included most of the known surface antigens (Fig. 2a and Supplementary Table 2). However, Pfs35 and a sexual stage-specific kinase (PF13_0258) were not detected. Nevertheless the cultured gametocytes analysed in this study expressed a specific repertoire of rifin and PfEMP1 proteins (Fig. 2b and Supplementary Table 2). Together these observations suggest that the gametocyte, which is very long-lived in the red blood cell (that is, 9–12 days compared with 2 days for the pathogenic asexual parasites), expresses a limited repertoire of the highly polymorphic families of surface antigens so widely represented in the asexual parasites.

The sporozoite proteome

Sporozoites are injected by the mosquito during ingestion of a blood meal. Although, they are in the blood stream for only minutes, sporozoites probably require mechanisms to evade the host humoral immune system in order for at least a fraction of the thousands of sporozoites injected by the mosquito to survive the hostile environment in the blood and successfully invade hepatocytes.

The main class of annotated sporozoite proteins identified was cell surface and organelle proteins (Fig. 1). Sporozoites are an invasive stage and possess the apical complex machinery involved in host cell invasion. As observed in the analysis of the *P. y. yoelii* sporozoite transcriptome[20], actin and myosin were found in the motile sporozoites (Supplementary Table 2). Many proteins associated with rhoptry, micronemes and dense granules were detected (Fig. 2a). Among the proteins found were known markers of the sporozoite stage, such as the circumsporozoite protein (CSP) and sporozoite surface protein 2 (SSP2; also known as TRAP), both present in large quantities at the sporozoite surface (Fig. 2a). Peptides derived from CTRP (circumsporozoite protein and thrombospondin-related adhesive protein (TRAP)-related protein), an ookinete cell surface protein involved in recognition and/or motility[21], were detected in the sporozoite fractions (Supplementary Table 1).

Most surprisingly, peptides derived from multiple *var* (coding for PfEMP1) and *rif* genes were identified in the sporozoite samples. PfEMP1 and rifins are coded for by large multigene families (*var* and *rif*)[22,23] and are present on the surface of the infected red blood cell. No peptides derived from *rif* genes were identified in the trophozoite sample, whereas sporozoites expressed 21 different rifins and 25 PfEMP1 isoforms (Fig. 2b); that is, a total of 14% of the *rif* genes and 33% of the *var* genes encoded by the genome. Furthermore, very little overlap was observed between stages: only ten PfEMP1 and two rifin isoforms expressed in sporozoites were found in other stages. Whereas in the blood stream the asexual stage parasites undergo asexual multiplication and therefore have an opportunity to undergo antigenic 'switching' of the variant antigen genes, the non-replicative sporozoites may not have this opportunity. Expressing such a polymorphic array of *var* (PfEMP1) and *rif* genes could be part of a sporozoite survival mechanism.

Chromosomal clusters encoding co-expressed proteins

The distinct proteomes of each stage of the *Plasmodium* life cycle suggested that there is a highly coordinated expression of *Plasmodium* genes involved in common processes. Co-expression groups are a widespread phenomenon in eukaryotes, where mRNA array analyses have been used to establish gene expression profiles. Analysis of co-regulated gene groups facilitates both searching for regulatory motifs common to co-regulated genes, and predicting protein function on the basis of the 'guilt by association' model. Furthermore, mRNA analyses in *Saccharomyces cerevisiae*[24] and *Homo sapiens*[25,26] have demonstrated that co-regulated genes do not map to random locations in the genome but are in fact

Table 2 **Examples on enzymes in stage-specific metabolic pathways**

Locus	Stage				Enzyme	EC number†	Reaction catalysed
	Spz*	Mrz*	Tpz*	Gmt*			
End of glycolysis							
PF10_0363	1.2	–	2.4	–	Pyruvate kinase	2.7.1.40	*P*-enolpyruvate to pyruvate
MAL6P1.160	8.6	66.9	18.8	14.7	Pyruvate kinase		
PF13_0141	46.2	83.9	70.9	78.8	L-lactate dehydrogenase	1.1.1.27	Pyruvate to lactate
TCA cycle and oxidative phosphorylation							
PF10_0218	12.3	–	–	–	Citrate synthase	4.1.3.7	Acetyl coA + oxaloacetate to citrate
PF13_0242	3.2	–	16.9	8.8	Isocitrate dehydrogenase (NADP)	1.1.1.41	Isocitrate to 2-oxoglutarate + CO_2
PF08_0045	2.9	–	2.2	23.1	2-Oxoglutarate dehydrogenase e1 component	1.2.4.2	2-Oxoglutarate to succinyl CoA
PF10_0334	–	–	3.5	27.7	Flavoprotein subunit of succinate dehydrogenase	1.3.5.1	Succinate to fumarate
PFL0630w	3.7	–	–	12.1	Iron-sulphur subunit of succinate dehydrogenase		
PF14_0373	–	–	–	12.7	Ubiquinol cytochrome oxidoreductase	1.10.2.2.	Ubiquinol to cytochrome *c* reductase in electron transport
PFB0795w	–	–	–	14.2	ATP synthase F1, α-subunit		
PFI1365w	–	–	–	8.8	Cytochrome *c* oxidase subunit	1.9.3.1	
PFI1340w	–	–	–	8.8	Fumarate hydratase	4.2.1.2	Fumarate to malate
MAL6P1.242	30.4	–	–	40.9	Malate dehydrogenase	1.1.1.37	Malate to oxaloacetate

Plasmodium metabolic pathways can be found at http://www.sites.huji.ac.il/malaria/. Spz, sporozoite; mrz, merozoite; tpz, trophozoite; gmt, gametocyte.
*The sequence coverage (that is, the percentage of the protein sequence covered by identified peptides) measured in each stage is reported.
†Enzyme Commission (EC) numbers are reported for each protein.

frequently organized into gene clusters on a chromosome. Gene clustering in *Plasmodium* species has been demonstrated. Ordered arrays of genes involved in virulence and antigenic variation (for example, *var*, *vir* and *rif* genes) are located in the subtelomeric regions of the chromosomes[27,28].

To determine whether gene clustering exists along the entire *P. falciparum* genome, genes whose protein products were detected in our analysis were mapped onto all 14 chromosomes in a stage-dependent manner (Fig. 3a). The 2,415 proteins identified represented an average of 45% of the open reading frames (ORFs) predicted per chromosome. The number of protein hits by chromosome was similar for all stages: sporozoite, merozoite, trophozoite and gametocyte protein lists constituting 19.7%, 15.8%, 19.5% and 21.6% of the predicted ORFs per chromosomes, respectively. Groups of three or more consecutive loci whose protein products were detected in a particular stage were defined as chromosomal clusters encoding co-expressed proteins (Fig. 3b). On the basis of this definition a total of 98 clusters containing 3 loci, 32 clusters containing 4 loci, 5 clusters containing 5 loci, and 3 clusters containing 6 loci were identified (Supplementary Table 3). For each chromosome, the frequency of finding clusters encoding co-expressed proteins containing 3–6 adjacent loci markedly exceeded the probability of finding such clusters by chance (see the footnote of Supplementary Table 3 for details on the probability calculation). Therefore, chromosomal clusters encoding co-expressed proteins were prevalent in the *P. falciparum* genome.

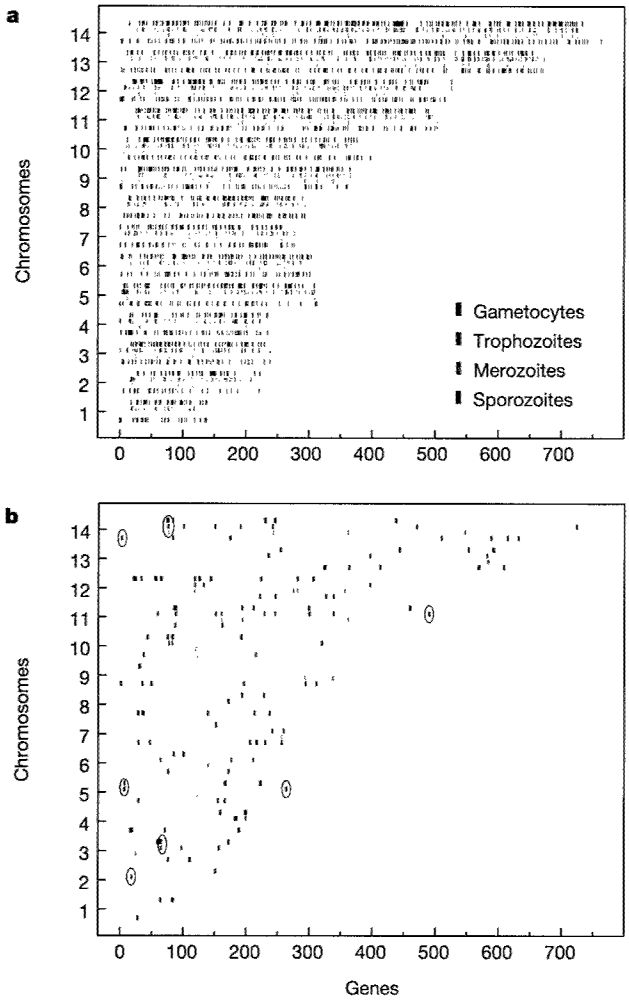


Figure 3 Distribution of expressed proteins by chromosome. **a**, For each stage, genes whose products were detected (coloured vertical bars) are plotted in the order they appear on their chromosome (grey boxes). **b**, Groups of at least three consecutive expressed genes are defined as chromosomal clusters of co-expressed proteins. Examples of such clusters, circled in **b**, are specified in Table 3 and the complete description of the 138 clusters can be found in Supplementary Table 3.

Functionally related genes have been shown to cluster in the *S. cerevisiae*[24] and human genomes[26]. This phenomenon also occurs in *P. falciparum*. A total of 138 clusters encoding co-expressed proteins were identified and 67 of them (49%) contained at least two loci that have been functionally annotated. Of these 67 clusters, 30 contained at least two loci whose annotation clearly indicates that the proteins are functionally related. For example, clusters on chromosomes 3, 5 and 10 contained ribosomal proteins, proteins involved in protein modification, and proteins involved in nucleotide metabolism, respectively (Table 3). Chromosome 14 contained a cluster of four aspartic proteases co-expressed in all of the blood stages (Table 3). This cluster was not detected in sporozoites, where no haemoglobin degradation is expected to occur. Interestingly, whereas the falcipain gene cluster on chromosome 11 appeared in our analysis as a cluster of co-expressed proteins (Supplementary Table 3), the SERA gene cluster on chromosome 2, coding for proteins that share a papain-like sequence motif[29], did not. Of the ten sporozoite-specific clusters, five involved *var* and *rif* genes, such as the *rif* cluster located in the subtelomeric domain of chromosome 14 (Table 3). On the basis of their presence in clusters encoding co-expressed proteins, we were able to suggest functional roles for 24 proteins annotated as hypothetical in the *P. falciparum* genome (Supplementary Table 3). For example, a gametocyte-specific cluster on chromosome 13 encoded two transmission-blocking antigens (Pfs48/45 and Pfs47) and a hypothetical protein, PF13_0246, which might be a gametocyte surface protein. Two clusters on chromosomes 2 and 11 were highly specific to the trophozoite stage (Table 3). Each of these clusters contained well-known secreted and surface proteins, namely KAHRP, PfEMP3, antigen 332, and RESA, all of which have been implicated in knob formation. The highly coordinated expression of these genes makes the three hypothetical proteins listed in these trophozoite-specific gene clusters possible candidates for involvement in cyto-adherence.

Discussion

Although sample handling is a principal consideration when studying pathogens, the expression of large numbers of previously identified proteins was consistent with their published expression profiles, validating our data set as a meaningful sampling of each stage's proteome. This is a particularly important aspect of our analysis as 65% of the 5,276 genes encoded by the *P. falciparum* genome are annotated as hypothetical[1], and of the 2,415 expressed proteins we identified, 51% are hypothetical proteins (Supplementary Table 1). Our results confirmed that these hypothetical ORFs predicted by gene modelling algorithms were indeed coding regions. Furthermore, from all four stages analysed, we identified 439 proteins predicted to have at least one transmembrane segment or a GPI addition signal (18% of the data set) and 304 soluble proteins with a signal sequence; that is, potentially secreted or located to organelles. Well over half of the secreted proteins and integral membrane proteins detected were annotated as hypothetical (Supplementary Table 4). The obvious interest in this class of proteins is that, with no homology to known proteins, they represent potential *Plasmodium*-specific proteins and may provide targets for new drug and vaccine development.

Our comprehensive large-scale analysis of protein expression showed that most surface proteins are more widely expressed than initially thought. In particular, the *var* and *rif* genes, which were thought to be involved in immune evasion only in the blood stage, have now been shown to be expressed in apparently large and varied numbers at the sporozoite stage. These surface proteins might be involved in general interaction processes with host cells and/or immune evasion. An alternative hypothesis is that stage-specific

Table 3 **Examples of chromosomal gene clusters encoding co-expressed proteins**

Chromosome	ID	Locus	Stage				Description	Class	SP	TM
			Spz	Mrz	Tpz	Gmt				
3	64	PFC0285c	2.1	12.7	33.2	18.7	T-complex protein β-subunit	Protein fate	0	0
3	65	PFC0290w	8.3	–	33.8	18.6	40S ribosomal protein S23	Protein synthesis	0	0
3	66	PFC0295c	–	14.9	52.5	21.3	40S ribosomal protein S12	Protein synthesis	0	0
3	67	PFC0300c	–	12.1	30.4	17.9	60S ribosomal protein L7	Protein synthesis	0	0
5	263	PFE1345c	–	–	1.9	1.6	Minichromosome maintenance protein 3	Cell transport	0	0
5	264	PFE1350c	–	–	22.4	–	Ubiquitin-conjugating enzyme	Protein fate	0	0
5	265	PFE1355	–	4.8	2.6	2.6	Ubiquitin carboxy-terminal hydrolase	Protein fate	0	0
5	266	PFE1360c	–	–	7.7	–	Methionine aminopeptidase	Protein fate	0	0
10	119	PF10_0121	10.8	74.5	29	–	Hypoxanthine phosphoribosyltransferase	Metabolism	0	0
10	120	PF10_0122	5.4	6.1	–	6.1	Phosphoglucomutase	Metabolism	0	0
10	121	PF10_0123	–	11.7	–	–	GMP synthetase	Metabolism	0	0
10	122	PF10_0124	0.9	1.8	–	–	Hypothetical protein		0	0
14	74	PF14_0074	26.6	–	–	4.9	Hypothetical protein		0	0
14	75	PF14_0075	–	26.5	43.2	47.4	Plasmepsin	Protein fate	1	0
14	76	PF14_0076	–	6.6	35.2	10	Plasmepsin 1	Protein fate	1	0
14	77	PF14_0077	–	21.2	43	11.5	Plasmepsin 2	Protein fate	1	0
14	78	PF14_0078	–	14.2	52.8	29.9	HAP protein	Protein fate	1	0
14	2	PF14_0002	3.5	–	–	–	Rifin	Surface or organelles	0	1
14	3	PF14_0003	7.9	–	–	–	Rifin	Surface or organelles	1	2
14	4	PF14_0004	6.5	–	–	–	Rifin	Surface or organelles	1	2
2	18	PFB0090c	–	–	3	–	Hypothetical protein, conserved		0	0
2	19	PFB0095c	–	–	3.4	–	Erythrocyte membrane protein 3	Surface or organelles	1	0
2	20	PFB0100c	–	1.5	24.8	–	Knob-associated histidine-rich protein	Surface or organelles	1	0
11	489	PF11_0506	–	–	6.3	4.4	Hypothetical protein		0	1
11	490	PF11_0507	–	–	0.8	–	Antigen 332	Surface or organelles	0	0
11	491	PF11_0508	–	–	3.3	–	Hypothetical protein		0	0
11	492	PF11_0509	–	6.4	3	–	RESA	Surface or organelles	0	0
13	443	PF13_0246	4.5	–	–	8.6	Hypothetical protein		0	0
13	444	PF13_0247	–	–	–	32.4	Transmission-blocking target antigen precursor (Pfs48/45)	Surface or organelles	1	1
13	445	PF13_0248	–	–	–	7.1	Transmission-blocking target antigen precursor (Pfs47)	Surface or organelles	1	1

Clusters of at least three consecutive genes encoding co-expressed proteins are reported with their position (ID) on the chromosome, the sequence coverage measured for these proteins in each stage (%), their current annotation and functional class, and the predicted presence of signal peptide (SP) or transmembrane domains (TM) (based on the TMHMM[43], a transmembrane (TM) helices prediction method based on a hidden Markov model (HMM), big-PI Predictor[44] and SignalP[45] algorithms).

regulation is not as exact as previously thought.

One mechanism of protein expression control that contributes to stage specificity in *P. falciparum* arises from the chromosomal clustering of genes encoding co-expressed proteins. The clusters described in this study demonstrate a widespread high order of chromosomal organization in *P. falciparum* and probably correspond to regions of open chromatin allowing for co-regulated gene expression. The high (A + T) content of the *P. falciparum* genome makes the identification of regulatory sequences such as promoters and enhancers challenging[31,32]. Focusing analyses on stage-specific and multi-stage clusters will facilitate finding stage-specific and general *cis*-acting sequences in the *Plasmodium* genome and will help decipher gene expression regulation during the parasite life cycle.

The malaria parasite is a complex multi-stage organism, which has co-evolved in mosquitoes and vertebrates for millions of years. Designing drugs or vaccines that substantially and persistently interrupt the life cycle of this complex parasite will require a comprehensive understanding of its biology. The *P. falciparum* genome sequence and comparative proteomics approaches may initiate new strategies for controlling the devastating disease caused by this parasite. □

Methods

Parasite material

Plasmodium falciparum clone 3D7 (Oxford) was used throughout. Sporozoites were initially isolated from the salivary glands of *Anopheles stephansi* mosquitoes, 14 days after infection, by centrifugation in a Renograffin 60 gradient, as described[33]. Four sporozoite samples were used as is. A fifth sample underwent an additional purification step on Dynabeads M-450 Epoxy coupled to NFS1 (an anti-*P. falciparum* CS protein monoclonal antibody)[34] according to the manufacturer's instructions (Dynal). Trophozoite-infected erythrocytes from synchronized cultures were purified on 70% Percoll-alanine[30], and the trophozoites released from the erythrocytes[35]. Of the of 260 parasitized erythrocytes counted by Giemsa-stained thin-blood film, 100% were identified as trophozoites. Merozoites were prepared essentially as described in ref. 36, using highly synchronized schizonts and purifying the merozoites by passage through membrane filters. Starting with synchronized asexual parasites grown in suspension culture as described[37,38], gametocytes were prepared by daily media changes of static cultures at 37 °C. When there were very few mature asexual stages present, gametocyte-infected erythrocytes were collected from the 52.5%/45% and 45%/30% interfaces of a Percoll gradient[39]. The gametocytes consisted mostly of stage IV and V parasites with minor contamination (<3%) from mixed asexual stage parasites. Finally, cellular debris from the upper bodies of parasite-free *A. stephansi* and non-infected human erythrocytes were used as controls for sporozoites and blood-stage parasites, respectively. Every effort was made to minimize enzymatic activity and protein degradation during sampling, and the subsequent isolation of the parasites; however, we cannot exclude that some of the differences in protein profiles that we observe between the different life-cycle stages may be a consequence of the sample-handling procedures.

Cell lysis

Five sporozoite, four merozoite, four trophozoite and three gametocyte preparations were lysed, digested and analysed independently. Cell pellets were first diluted ten times in 100 mM Tris-HCl pH 8.5, and incubated in ice for 1 h. After centrifugation at 18,000 *g* for 30 min, supernatants were set aside and microsomal membrane pellets were washed in 0.1 M sodium carbonate, pH 11.6. Soluble and insoluble protein fractions were separated by centrifugation at 18,000 *g* for 30 min. Supernatants obtained from both centrifugation steps were either combined (sporozoites, trophozoites and merozoites) or digested and analysed independently (gametocytes).

Peptide generation and analysis

The method follows that of Washburn *et al.*[5], with the exception that Tris(2-carboxyethyl)phosphine hydrochloride (TCEP-HCl; Pierce) was used to reduce urea-denatured proteins. Peptide mixtures were analysed through MudPIT as described[5].

Protein sequence databases

The *P. falciparum* database contained 5,283 protein sequences. Spectra resulting from contaminant mosquito and erythrocyte peptides had to be taken into account in the sporozoite and blood-stage samples, respectively. Tandem mass spectrometry (MS/MS) data sets from blood stages were therefore searched against a database containing both *P. falciparum* protein sequences and 24,006 ORFs from the human, mouse and rat RefSeq NCBI databases. At the date of the searches, the *Anopheles gambiae* genome was not available. The NCBI database contained 922 *Anopheles* and 313 *Aedes* proteins, which were combined to the 14,335 ORFs of the NCBI *Drosophila melanogaster*[40] database to create a control diptera database. Finally, these databases were complemented with a set of 172 known protein contaminants, such as proteases, bovine serum albumin and human keratins.

MS/MS data set analysis

The SEQUEST algorithm was used to match MS/MS spectra to peptides in the sequence databases[41]. To account for carboxyamidomethylation, MS/MS data sets were searched with a relative molecular mass of 57,000 (M_r 57K) added to the average molecular mass of cysteines. Peptide hits were filtered and sorted with DTASelect[42]. Spectra/peptide matches were only retained if they were at least half-tryptic (Lys or Arg at either end of the identified peptide) and with minimum cross-correlation scores (XCorr) of 1.8 for +1, 2.5 for +2, and 3.5 for +3 spectra and DeltaCn (top match's XCorr minus the second-best match's XCorr divided by the top match's XCorr) of 0.08. Peptide hits were deemed unambiguous only if they were not found in non-infected controls and were uniquely assigned to parasite proteins by searching against combined parasite–host databases. Finally, for low coverage loci, peptide/spectrum matches were visually assessed on two main criteria: any given MS/MS spectrum had to be clearly above the baseline noise, and both *b* and *y* ion series had to show continuity. The Contrast tool[42] was used to compare and merge protein lists from replicate sample runs and to compare the proteomes established for the four stages.

Received 31 July; accepted 9 September 2002; doi:10.1038/nature01107.

1. Gardner, M. J. *et al.* Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419,** 498–511 (2002).
2. Carlton, J. M. *et al.* Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii*. *Nature* **419,** 512–519 (2002).
3. Ben Mamoun, C. *et al.* Co-ordinated programme of gene expression during asexual intraerythrocytic development of the human malaria parasite *Plasmodium falciparum* revealed by microarray analysis. *Mol. Microbiol.* **39,** 26–36 (2001).
4. Hayward, R. E. *et al.* Shotgun DNA microarrays and stage-specific gene expression in *Plasmodium falciparum* malaria. *Mol. Microbiol.* **35,** 6–14 (2000).
5. Washburn, M. P., Wolters, D. & Yates, J. R. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnol.* **19,** 242–247 (2001).
6. Mewes, H. W. *et al.* MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* **30,** 31–34 (2002).
7. Pinder, J. C. *et al.* Actomyosin motor in the merozoite of the malaria parasite, *Plasmodium falciparum*: implications for red cell invasion. *J. Cell Sci.* **111,** 1831–1839 (1998).
8. Holder, A. A. *Malaria Vaccine Development: a Multi-immune Response and Multi-stage Perspective* (ed. Hoffman, S. L.) 77–104 (ASM Press, Washington, 1996).
9. Coppel, R. L. *et al.* Isolate-specific S-antigen of *Plasmodium falciparum* contains a repeated sequence of eleven amino acids. *Nature* **306,** 751–756 (1983).
10. Taylor, H. M. *et al. Plasmodium falciparum* homologue of the genes for *Plasmodium vivax* and *Plasmodium yoelii* adhesive proteins, which is transcribed but not translated. *Infect. Immun.* **69,** 3635–3645 (2001).
11. Kaneko, O. *et al.* The high molecular mass rhoptry protein, RhopH1, is encoded by members of the clag multigene family in *Plasmodium falciparum* and *Plasmodium yoelii*. *Mol. Biochem. Parasitol.* **118,** 223–231 (2001).
12. Trenholme, K. R. *et al.* clag9: A cytoadherence gene in *Plasmodium falciparum* essential for binding of parasitized erythrocytes to CD36. *Proc. Natl Acad. Sci. USA* **97,** 4029–4033 (2000).
13. Klemba, M. & Goldberg, D. E. Biological roles of proteases in parasitic protozoa. *Annu. Rev. Biochem.* **71,** 275–305 (2002).
14. Banerjee, R. *et al.* Four plasmepsins are active in the *Plasmodium falciparum* food vacuole, including a protease with an active-site histidine. *Proc. Natl Acad. Sci. USA* **99,** 990–995 (2002).
15. Rosenthal, P. J., Sijwali, P. S., Singh, A. & Shenai, B. R. Cysteine proteases of malaria parasites: targets for chemotherapy. *Curr. Pharm. Des.* **8,** 1659–1672 (2002).
16. Eggleson, K. K., Duffin, K. L. & Goldberg, D. E. Identification and characterization of falcilysin, a metallopeptidase involved in hemoglobin catabolism within the malaria parasite *Plasmodium falciparum*. *J. Biol. Chem.* **274,** 32411–32417 (1999).
17. Sinden, R. E., Butcher, G. A., Billker, O. & Fleck, S. L. Regulation of infectivity of *Plasmodium* to the mosquito vector. *Adv. Parasitol.* **38,** 53–117 (1996).
18. Billker, O., Shaw, M. K., Margo, G. & Sinden, R. E. Identification of xanthurenic acid as the putative inducer of malaria development in the mosquito. *Nature* **392,** 289–292 (1998).
19. Krungkrai, J., Prapunwattana, P. & Krungkrai, S. R. Ultrastructure and function of mitochondria in gametocytic stage of *Plasmodium falciparum*. *Parasite* **7,** 19–26 (2000).
20. Kappe, S. H. *et al.* Exploring the transcriptome of the malaria sporozoite stage. *Proc. Natl Acad. Sci. USA* **98,** 9895–9900 (2001).
21. Dessens, J. T. *et al.* CTRP is essential for mosquito infection by malaria ookinetes. *EMBO J.* **18,** 6221–6227 (1999).
22. Deitsch, K. W. & Wellems, T. E. Membrane modifications in erythrocytes parasitized by *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **76,** 1–10 (1996).
23. 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).
24. Cohen, B. A., Mitra, R. D., Hughes, J. D. & Church, G. M. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nature Genet.* **26,** 183–186 (2000).
25. Caron, H. *et al.* The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* **291,** 1289–1292 (2001).
26. Lercher, M. J., Urrutia, A. O. & Hurst, L. D. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nature Genet.* **31,** 180–183 (2002).
27. Hernandez-Rivas, R. *et al.* Expressed var genes are found in *Plasmodium falciparum* subtelomeric regions. *Mol. Cell Biol.* **17,** 604–611 (1997).
28. del Portillo, H. A. *et al.* A superfamily of variant genes encoded in the subtelomeric region of *Plasmodium vivax*. *Nature* **410,** 839–842 (2001).
29. Gardner, M. J. *et al.* Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science* **282,** 1126–1132 (1998).
30. Kanaani, J. & Ginsburg, H. Metabolic interconnection between the human malarial parasite *Plasmodium falciparum* and its host erythrocyte. *J. Biol. Chem.* **264,** 3194–3199 (1989).
31. Dechering, K. J. *et al.* Isolation and functional characterization of two distinct sexual-stage-specific promoters of the human malaria parasite *Plasmodium falciparum*. *Mol. Cell Biol.* **19,** 967–978 (1999).
32. Lockhart, D. J. & Winzeler, E. A. Genomics, gene expression and DNA arrays. *Nature* **405,** 827–836 (2000).
33. Pacheco, N. D., Strome, C. P., Mitchell, F., Bawden, M. P. & Beaudoin, R. L. Rapid, large-scale isolation of *Plasmodium berghei* sporozoites from infected mosquitoes. *J. Parasitol.* **65,** 414–417 (1979).
34. Mellouk, S. *et al.* Evaluation of an *in vitro* assay aimed at measuring protective antibodies against sporozoites. *Bull. World Health Organ.* **68** Suppl., 52–59 (1990).
35. Rabilloud, T. *et al.* Analysis of membrane proteins by two-dimensional electrophoresis: comparison of the proteins extracted from normal or *Plasmodium falciparum*-infected erythrocyte ghosts. *Electrophoresis* **20,** 3603–3610 (1999).
36. Blackman, M. J. Purification of *Plasmodium falciparum* merozoites for analysis of the processing of merozoite surface protein-1. *Methods Cell Biol.* **45,** 213–220 (1994).
37. Haynes, J. D. & Moch, J. K. Automated synchronization of *Plasmodium falciparum* parasites by culture in a temperature-cycling incubator. *Methods Mol. Med.* **72,** 489–497 (2002).
38. Haynes, J. D., Moch, J. K. & Smoot, D. S. Erythrocytic malaria growth or invasion inhibition assays with emphasis on suspension culture GIA. *Methods Mol. Med.* **72,** 535–554 (2002).
39. Carter, R., Ranford-Cartwright, L. & Alano, P. The culture and preparation of gametocytes of *Plasmodium falciparum* for immunochemical, molecular, and mosquito infectivity studies. *Methods Mol. Biol.* **21,** 67–88 (1993).
40. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287,** 2185–2195 (2000).
41. Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5,** 976–989 (1994).
42. Tabb, D. L., McDonald, W. H. & Yates, J. R. DTASelect and contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.* **1,** 21–26 (2002).
43. 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).
44. Eisenhaber, B., Bork, P. & Eisenhaber, F. Sequence properties of GPI-anchored proteins near the omega-site: constraints for the polypeptide binding site of the putative transamidase. *Protein Eng.* **11,** 1155–1161 (1998).
45. 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).

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

Acknowledgements

We are grateful to J. Graumann, R. Sadygov, G. Chukkapalli, A. Majumdar and R. Sinkovits for computer programming; C. Deciu for the probability calculations; and C. Delahunty and C. Vieille for critical reading of the manuscript. The authors acknowledge the support of the Office of Naval Research, the US Army Medical Research and Material Command, and the National Institutes of Health (to J.R.Y.). J.D.R. is funded by a Wellcome Trust Prize Studentship. We thank the scientists and funding agencies comprising the international Malaria Genome Project for making sequence data from the genome of *P. falciparum* clone 3D7 public before publication of the completed sequence. The opinions expressed are those of the authors and do not reflect the official policy of the Department of the Navy, Department of Defense, or the US government.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.R.Y. (e-mail: jyates@scripps.edu).

NONPRINT FORM

APPENDIX D

1. Type of Product:	2. Operating System/Version:	3. New Product or Replacement:	4. Type of File:
CD- Rom		New	

5. Language/Utility Program:

6. # of Files/# of Products:	7. Character Set:	8. Disk Capacity:
1		32MB RAM
	9. Compatibility:	**10. Disk Size:**
		4 ½"

11. Title:

Plasmodium genome: Scientific Achievement and Medical Opportunity

12. Performing Organization:	13. Performing Report #:	14. Contract #:
The Institute for Genomic Research		DAMD17-98-2-8005
		15. Program Element #:
16. Sponsor/Monitor: U.S. Army Medical Research & Material Command Fort Detrick, MD 21702-5012	**17. Sponsor/Monitor # Acronym:** MCMR-RMI-S	**19. Project #:**
	18. Sponsor/Monitor #:	**20. Task #:**
		21. Work Unit #:

22. Date:	23. Classification of Product:
	Unclassified
24. Security Classification Authority:	**25. Declassification/Downgrade Schedule:**

26. Distribution/Availability:

Distribution approved for public release; distribution unlimited

NONPRINT FORM

27. Abstract:

28. Classification of Abstract:	**29. Limitation of Abstract:**
Unclassified	Unlimited
30. Subject Terms: human and rodent malaria parasite: Plasmodium falciparum and Plasmodium yoelii	**30a. Classification of Subject Terms:**

31. Required Peripherals:

Adobe Acrobat, Macromedia Flash Player and Apple Quicktime are required to view some parts of this CD-ROM

32. # of Physical Records:	**33. # of Logical Records:**	**34. # of Tracks:**
35. Record Type:	**36. Color:**	**37. Recording System:**
38. Recording Density:	**39. Parity:**	**40. Playtime:**

41. Playback Speed: Standard	**42. Video:**	**43. Text:**	**44. Still Photos:**	**45. Audio:**

46. Other:

47. Documentation/Supplemental Information:

48. Point of Contact and Telephone Number:

Judy Pawlus
301-619-7322

Exploring the transcriptome of the malaria sporozoite stage

Stefan H. I. Kappe*†, Malcolm J. Gardner‡, Stuart M. Brown§, Jessica Ross*, Kai Matuschewski*, Jose M. Ribeiro¶, John H. Adams‖, John Quackenbush‡, Jennifer Cho‡, Daniel J. Carucci**, Stephen L. Hoffman††, and Victor Nussenzweig*

*Michael Heidelberger Division, Department of Pathology, Kaplan Cancer Center, New York University School of Medicine, New York, NY 10016; ‡The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850; §Research Computing Resource, New York University Medical Center, New York, NY 10016; ¶Medical Entomology Section, Laboratory of Parasitic Diseases, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD 20892-0425; ‖Department of Biological Sciences, University of Notre Dame, Notre Dame, IN 46556; **Malaria Program, Naval Medical Research Center, Silver Spring, MD 20910; and ††Celera Genomics, 45 West Gude Drive, Rockville, MD 20850

Edited by Louis H. Miller, National Institutes of Health, Bethesda, MD, and approved June 19, 2001 (received for review April 13, 2001)

Most studies of gene expression in *Plasmodium* have been concerned with asexual and/or sexual erythrocytic stages. Identification and cloning of genes expressed in the preerythrocytic stages lag far behind. We have constructed a high quality cDNA library of the *Plasmodium* sporozoite stage by using the rodent malaria parasite *P. yoelii*, an important model for malaria vaccine development. The technical obstacles associated with limited amounts of RNA material were overcome by PCR-amplifying the transcriptome before cloning. Contamination with mosquito RNA was negligible. Generation of 1,972 expressed sequence tags (EST) resulted in a total of 1,547 unique sequences, allowing insight into sporozoite gene expression. The circumsporozoite protein (CS) and the sporozoite surface protein 2 (SSP2) are well represented in the data set. A BLASTX search with all tags of the nonredundant protein database gave only 161 unique significant matches ($P(N) \leq 10^{-4}$), whereas 1,386 of the unique sequences represented novel sporozoite-expressed genes. We identified ESTs for three proteins that may be involved in host cell invasion and documented their expression in sporozoites. These data should facilitate our understanding of the preerythrocytic *Plasmodium* life cycle stages and the development of preerythrocytic vaccines.

Plasmodium yoelii yoelii | expressed sequence tag

Protozoan parasites of the genus *Plasmodium* are the causative agents of malaria, the most devastating parasitic disease in humans. The parasites occur in distinct morphological and antigenic stages as they progress through a complex life cycle, thwarting decades of efforts to develop an effective malaria vaccine. *Plasmodium* is transmitted via the bite of an infected *Anopheles* mosquito, which releases the sporozoite stage into the skin. Sporozoites enter the bloodstream and, on reaching the liver, invade hepatocytes and develop into exo-erythrocytic forms (EEF). After multiple cycles of DNA replication, the EEF contains thousands of merozoites (liver schizont) that are released into the blood stream and initiate the erythrocytic cycle (asexual blood stage) that causes the disease malaria. Changes in life cycle stages are accompanied by major changes in gene expression and therefore by major changes in antigenic composition. The form of the parasite best studied is the asexual blood stage, mainly because of its comparatively easy experimental accessibility. Therefore, most *Plasmodium* proteins that have been well characterized are expressed during the erythrocytic cycle, among them some major erythrocytic-stage vaccine candidates such as merozoite surface protein-1 (MSP-1) and apical membrane antigen-1 (AMA-1; ref. 1). Erythrocytic-stage vaccines are aimed at inducing an immune response that suppresses or eradicates parasite load in the blood. In contrast, preerythrocytic vaccines are aimed at eliciting an immune response that destroys the sporozoites and the EEF, thereby preventing progression of the parasite to the blood stage. The feasibility of a preerythrocytic vaccine is demonstrated by the fact that immunization with radiation-attenuated sporozoites leads to protective, sterile immunity (2, 3). The effector mechanisms are antibodies (4), cytotoxic T lymphocytes (CTL; ref. 4), and lymphokines (5, 6). Hence, it is desirable to systematically identify proteins synthesized by sporozoites and EEF to select new potential vaccine candidates. Antibodies against surface-exposed sporozoite proteins block hepatocyte entry (7). In addition, sporozoite proteins can be carried over into the invaded hepatocyte and become a target for CTL (8). By using mixtures of these proteins, it might be possible to formulate a vaccine that mimics the sterile immunity achieved by immunization with irradiated sporozoites. Sporozoite proteins could also be the target of transmission-blocking strategies. Past efforts to prepare cDNA libraries of sporozoites and identify new sporozoite antigens were hindered by difficulties in obtaining adequate numbers of purified parasites. Thus far, few sporozoite-expressed proteins have been identified. The best characterized of these proteins are the circumsporozoite protein (CS; ref. 2) and the sporozoite surface protein 2 (SSP2), also called thrombospondin-related anonymous protein (TRAP; refs. 9–11). CS and SSP2/TRAP are involved in the invasion of hepatocytes and are detected in the hepatocyte after sporozoite invasion. Both proteins are found in all *Plasmodium* species examined. A few other sporozoite antigens have been identified in *P. falciparum* (12, 13), but their function is unknown.

To facilitate the identification of genes that are expressed in the sporozoite stage, we have constructed a cDNA library from salivary gland sporozoites of the rodent malaria parasite *Plasmodium yoelii* and generated 1,972 expressed sequence tags (ESTs). We document the quality of the library by the presence of CS and SSP2/TRAP transcripts and the absence of erythrocytic stage-specific transcripts. The sequence data provide insight into sporozoite gene expression. We show sporozoite expression of MAEBL (14), a protein previously thought to be present only in erythrocytic stages. In addition, we identify two putative sporozoite adhesion ligands. Transcripts of a key enzyme of the shikimate pathway (15) are present in the data set, indicating that this pathway is likely to be operational in sporozoites and liver stages.

MICROBIOLOGY

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: CS, circumsporozoite protein; SSP2, sporozoite surface protein 2; TRAP, thrombospondin-related anonymous protein; EST, expressed sequence tag; EEF exo-erythrocytic form; MSP-1, merozoite surface protein-1; MyoA, myosin A; TSR, thrombospondin type 1 repeat; SPATR, secreted protein with altered thrombospondin repeat.

Data deposition: The EST sequences reported in this paper have been deposited in the GenBank dbEST database (accession nos. BG601070–BG603042). Complete gene sequences have been deposited in the GenBank database (accession nos. AF390551–AF390553).

†To whom reprint requests should be addressed. E-mail: kappes01@popmail.med.nyu.edu.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Materials and Methods

Parasite Preparation. Two million *P. yoelii* (17XNL) sporozoites were obtained in a salivary gland homogenate from dissection of 100 infected *Anopheles stephensi* mosquitos. The crude salivary gland homogenate was passed over a DEAE cellulose column to remove contaminating mosquito tissue. Sporozoites (4×10^5) were recovered after purification. The preparation was almost free of mosquito contaminants as judged by microscopic inspection. Sporozoites were immediately subjected to poly(A)$^+$ RNA extraction.

RNA Extraction and cDNA Synthesis. Poly(A)$^+$ RNA was directly isolated from the sporozoites by using the MicroFastTrack procedure (Invitrogen) and was resuspended in a final volume of 10 μl elution buffer (10 mM Tris, pH 7.5). The obtained poly(A)$^+$ RNA was treated with Dnase I (Life Technologies, Rockville, MD) to remove possible genomic DNA contamination. RNA quantification was not possible because of the minute amounts obtained. The RNA was reverse-transcribed by using Superscript II (Life Technologies), a modified oligo(dT) oligonucleotide for first strand priming (5′-AAGCAGTGGTAACAACGCAGAGTACT_{30}VN-3′; V = A/C/G, N = A/C/G/T) and a primer called cap switch oligonucleotide (5′-AAGCAGTGGTAACAACGCAGAGTACGCGGG-3′) that allows extension of the template at the 5′ end (CLONTECH). Second strand synthesis and subsequent PCR amplification was done with an oligonucleotide that anneals to both the modified oligo(dT) oligonucleotide and the cap switch oligonucleotide.

cDNA Cloning and Sequencing. The cDNA was size selected on a CHROMA-SPIN 400 column (CLONTECH) that resulted in a cutoff at ≥300 bp and was ligated into vector pCR4 (Invitrogen). Ligations were transformed into *Escherichia coli* TOP10-competent cells. Template preparation and sequencing were done as described (16). Sequencing was performed in both directions.

Assemblies and Database Searches. All obtained sequences were subjected to vector sequence removal and screened for overlaps, and matching sequences were then assembled by using the TIGR assembler program. The nonredundant (NR) sequence database at the National Center for Biotechnology Information (NCBI) was searched with the complete data set, consisting of the assembled sequences and singletons, by using the Basic Local Alignment Search Tool X (BLASTX) algorithm.

Sources of Sequence Data. Sequence data were obtained from the TIGR *P. yoelii* genome project (www.tigr.org) and the *Plasmodium* genome consortium PlasmoDB (http://PlasmoDB.org).

cDNA Blots. cDNA was separated on agarose gels and transferred to nylon membranes (Roche). Gene-specific probes were prepared by using the digoxigenin (DIG) High Prime Labeling system (Roche). cDNA blots were incubated and washed according to the manufacturer's instructions (Roche).

Reverse Transcription–PCR. Poly(A)$^+$ RNA was reverse-transcribed by using Superscript II. Gene-specific PCR was done by using oligonucleotide primers specific for *P. yoelii MSP-1* (L22551; sense, 5′-GGTAAAAGCTGGCGTCATTGATCC-3′; antisense, 5′-GTCTAATTCAAAATCATCGGCAGG-3′) or *P. yoelii MAEBL* (AF031886; sense, 5′-ATGCTGCTCAATATCAGATTATTGC-3′; antisense, 5′-AACAATTTCATCAAAAGCAACTTCC-3′).

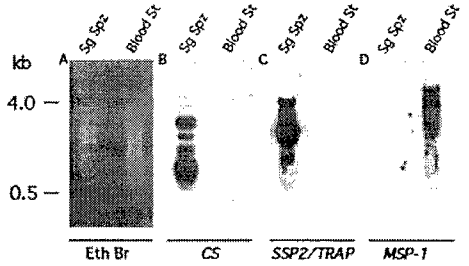


Fig. 1. Quality assessment of the generated cDNA populations. cDNA blot hybridization with stage-specific probes demonstrates that stage-specific transcript representation is not altered by cDNA amplification. (*A*) Ethidium bromide-stained agarose gel of cDNA amplified from salivary gland sporozoites (Sg Spz) or mixed blood stages (Blood St). Note the distinct bands visible in the sporozoite preparation. (*B*) Hybridization to a *CS* probe. (*C*) Hybridization to an *SSP2/TRAP* probe. (*D*) Hybridization to an *MSP-1* probe. Sizes are given in kb.

Indirect Immunofluorescence Assay. Salivary gland sporozoites and midgut sporozoites were incubated in 3% BSA/RPMI medium 1640 on BSA-covered glass-slides for 30 min, fixed, and permeabilized with 0.05% saponin. MAEBL was detected with the polyclonal antisera against the M2 domain or the 3′-carboxyl cysteine-rich region (1:200; ref. 14) and FITC-conjugated goat anti-rabbit IgG (1:100; Kirkegaard & Perry Laboratories).

Results

Quality Assessment of the cDNA Library. The amplified sporozoite cDNAs showed a visible size distribution between 300 and 4,000 bp on ethidium bromide-stained agarose gels, with highest density between 500 and 3,000 bp (Fig. 1*A*). No amplification was detected when the reverse transcription step was omitted (data not shown). To assess the quality of the sporozoite cDNA population, we performed cDNA blot analysis with probes for the sporozoite-expressed *SSP2/TRAP* and *CS*. cDNAs for both proteins were found to be abundant in salivary gland sporozoite preparations but absent in blood stage parasite preparations (Fig. 1 *B* and *C*). Conversely, cDNAs for the blood stage-expressed *MSP-1* were detected in blood stage parasite preparations but absent in sporozoites (Fig. 1*D*). The cDNA blot analysis documented the presence of cDNAs of the approximate full-length size of each transcript. In addition, smaller sized cDNA fragments were present for each transcript, resulting in multiple signals from distinctly sized cDNAs (Fig. 1). To assure that no trace amounts of genomic DNA were amplified, we analyzed the sporozoite cDNA for the presence of introns by using the transcript of myosin A (MyoA), a myosin that is expressed in the sporozoite stage (17). *MyoA* contains two introns, and neither was detected in the sporozoite cDNA preparation (data not shown). Sequencing of 100 clones confirmed the cDNA fragmentation, which was mainly due to internal priming by the modified oligo(dT) oligonucleotide. It annealed to homo-polymeric runs of adenine in the untranslated regions (UTR) and the coding sequences of this AT-rich organism. We took advantage of the AT-richness of the *P. yoelii* genome to differentiate between cDNAs of parasite origin and cDNAs amplified from contaminating mosquito RNA. Based on the total number of cDNA clones of mosquito origin, contamination was estimated to be ≈1%.

Characteristics of the EST Data Set. We obtained a final number of 1,972 sequence reads of sufficient quality to be subjected to further analysis (Table 1). The average length of EST sequence was 377 bp. Six hundred forty-eight of the sequence reads could be assembled into 223 consensus sequences (input files), and

Table 1. General characteristics of the *P. yoelii* sporozoite EST project

ESTs submitted to NCBI	1,972
ESTs in input files	648
Input files	223
Singletons	1,324
Total number of unique sequences	1,547
BLASTX matches	286
Unique BLASTX matches	161
Matches with proteins of unknown function	75
BLASTX matches with *Plasmodium* proteins	70
ESTs for CS	33
ESTs for SSP2/TRAP	13
ESTs for MAEBL	10
ESTs for HSP-70	10

1,324 sequences did not match another sequence in the data set sufficiently to allow assembly (singletons). This analysis gave a total of 1,547 unique sequences. A BLASTN comparison between the 1,547 unique sequences and the incomplete *P. yoelii* genome (2× coverage) database resulted in 1,135 matches. A BLASTX search of the predicted proteins from *the P. falciparum* genome (translated ORFs of >100 bases) resulted in only 356 matches, with a smallest sum probability of $P(N) \leq 10^{-4}$. A BLASTX search of the NR sequence database at NCBI resulted in only 286 matches, with a smallest sum probability of $P(N) \leq 10^{-4}$. Of those, 70 were matches with known *Plasmodium* proteins. The matches were grouped in functional categories shown in Fig. 2 (see Table 2, which is published as supplemental data on the PNAS web site, www.pnas.org, for a complete list of all BLASTX matches). All ESTs have been deposited in the GenBank dbEST database (accession nos. BG601070–BG603042). In addition, data are made available through the *P. yoelii* gene index (http://www.tigr.org/tdb/pygi/).

Functional Groups of ESTs. Ribosomal proteins were not very abundant, with only 7 of the estimated 80 components of the ribosome represented. Only 4 ESTs gave matches with other proteins involved in translation. This low representation of proteins of the translation machinery contrasts with the relative abundance of ribosomal proteins found in EST sequencing projects for *Toxoplasma* tachyzoites (12% of all ESTs; refs. 18 and 19) and *Cryptosporidium* sporozoites (8% of all ESTs; ref. 20). However, a *P. falciparum* blood stage parasite EST project found that proteins involved in translation were also underrepresented (21). There were 18 ESTs in the transcription category, 7 matching a *P. falciparum* RNA recognition motif binding protein and two matching a human zinc finger protein potentially involved in transcription.

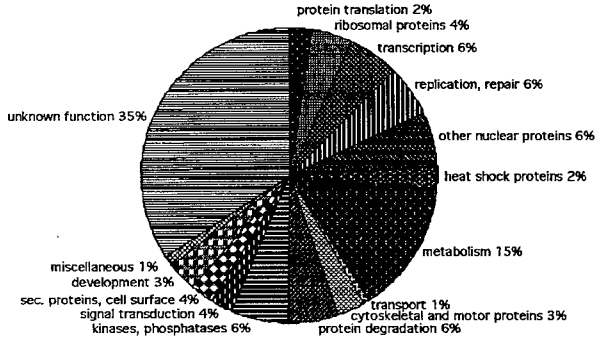


Fig. 2. Functional classification of *P. yoelii* sporozoite ESTs. One hundred sixty-one unique BLASTX matches were classified according to their putative biological function. Refer to Table 2 for a complete list of all BLASTX matches.

Especially significant among the ESTs giving BLASTX matches with proteins involved in metabolic pathways is chorismate synthase, the final enzyme of the shikimate pathway. This pathway generates the aromatic precursor chorismate, which is used for aromatic amino acid biosynthesis. The shikimate pathway is present in plants, fungi, and Apicomplexa (15) but is not found in vertebrates.

The salivary gland sporozoite is highly motile, and its main function is the invasion of the vertebrate hepatocyte. Of relevance to motility and invasion are tags for two apicomplexan unconventional class XIV myosins, MyoA and MyoB. MyoA localized under the plasma membrane within all invasive stages of *Plasmodium* (sporozoite, merozoite, and ookinete; refs. 17, 22, and 23), and a homologous protein was expressed in the *Toxoplasma* tachyzoite (24, 25). This myosin is currently the best candidate for the motor protein that drives Apicomplexan motility and host cell penetration.

Kinases and phosphatases are likely to be involved in the regulation of motility and host cell invasion (26), and we find 10 different input files and singletons in this category. Recently it was shown that a calmodulin-domain kinase, represented with one EST in the data set, played a crucial role in *Toxoplasma* tachyzoite motility and host cell invasion (27). Phospholipase A_2 is represented with one EST. Involvement of secreted phospholipase A_2 in the invasion process was shown in *Toxoplasma* tachyzoites (28). It will be of interest to find out whether this *Plasmodium* homologue has a role in hepatocyte invasion and/or plays a role in the migration of sporozoites through cells before establishing an infection (29).

The group of predicted secreted proteins and proteins that have a membrane anchor are of special interest, because they may be involved in host cell recognition and/or invasion. Within this group is the CS protein, most likely glycosylphasphatidylinositol-anchored, and SSP2/TRAP, a type one transmembrane protein. CS had one of the highest representations in the EST set with 33 matches, and TRAP was represented with 13 matches (Table 1).

Identification of Three Potential Sporozoite Invasion Ligands. Unexpectedly, we found that MAEBL was represented with 10 ESTs (Table 1). It was reported previously that MAEBL is expressed in *P. yoelii* and *P. berghei* merozoites, where it localized to the rhoptry organelles (14, 30). MAEBL is a type one transmembrane protein with a chimeric structure. It shares similarity with apical membrane antigen-1 (AMA-1) in the N-terminal portion, and similarity with the erythrocyte binding protein (EBP) family in the C-terminal portion (31). To ensure that the representation of a merozoite rhoptry protein in our EST library was not an artifact, we hybridized a salivary gland and midgut sporozoite cDNA blot to a *MAEBL*-specific probe, resulting in strong signals for both populations (Fig. 3*A*). In addition, reverse transcription–PCR with gene-specific primers resulted in *MAEBL* amplification from salivary gland sporozoite poly(A)$^+$ RNA and from blood stage poly(A)$^+$ RNA. In contrast, *MSP-1* expression was detected only in blood stages (Fig. 3*B*). A polyclonal antiserum against the carboxyl cysteine-rich region of *P. yoelii* MAEBL strongly reacted with permeabilized *P. yoelii* salivary gland sporozoites and midgut sporozoites in indirect immunofluorescence assay (IFA), indicating that this protein is indeed expressed in the sporozoite stages (Fig. 3 *C* and *D*). MAEBL localization was heterogeneous but was frequently more pronounced in one end of the sporozoites. Similar staining was obtained with a polyclonal antiserum against the M2 domain of MAEBL (data not shown).

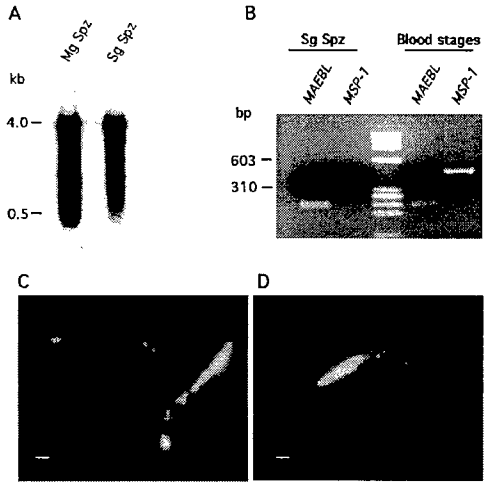


Fig. 3. Sporozoite expression of MAEBL. (*A*) cDNA blot showing *MAEBL* expression in midgut sporozoites (Mg Spz) and salivary gland sporozoites (Sg Spz). (*B*) Reverse transcription–PCR confirming *MAEBL* expression in salivary gland sporozoites. *MAEBL* expression is also detected in blood stages. Amplification with *MSP-1*-specific primers shows *MSP-1* expression in blood stages. *MSP-1* expression is not detected in salivary gland sporozoites. Sizes are given in base pairs (bp). (*C*) Localization of MAEBL by indirect immunofluorescence assay in *P. yoelii* salivary gland sporozoites with antisera against the carboxyl cysteine-rich region. (*D*) Localization of MAEBL by indirect immunofluorescence in *P. yoelii* midgut sporozoites with antisera against the carboxyl cysteine-rich region. Scale bar for *C* and *D* = 1 μm.

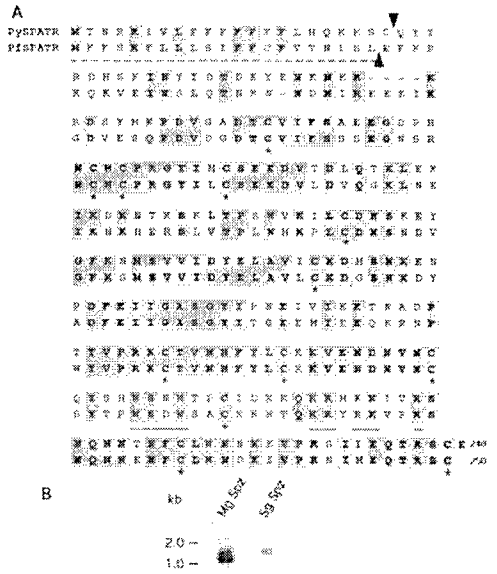


Fig. 4. Alignment of SPATR and expression in sporozoites. (*A*) Comparison of the deduced amino acid sequences of the *P. yoelii* SPATR with the homologue in *P. falciparum* (accession no. C71611). The conserved residues of the altered TSR are underlined with a solid line. The putative signal peptides are underlined with a dashed line. Putative signal peptide cleavage sites are marked with arrowheads (▲, ▼). Conserved cysteine residues are marked with an asterisk (*). Identical residues are shaded dark gray. Conserved amino acid changes are shaded light gray, and radical changes are not shaded. (*B*) cDNA blot demonstrating *SPATR* expression in midgut sporozoites (Mg Spz) and salivary gland sporozoites (Sg Spz). Sizes are given in kb.

One EST in the data set identified another potential sporozoite invasion ligand, matching a hypothetical ORF on chromosome 2 of *P. falciparum* (PFB0570w; ref. 16). We determined the complete ORF for this *P. yoelii* EST. The predicted protein has a putative cleavable signal peptide predicting that it is secreted (Fig. 4*A*). Significantly, the protein carries a motif with similarity to the thrombospondin type 1 repeat (TSR) (32). We therefore named it SPATR (secreted protein with altered thrombospondin repeat). The most conserved motif of the TSR is present (WSXW), followed by a stretch of basic residues. The central CSXTCG that follows the WSXW motif in a number of the TSR superfamily members (33) is not present in SPATR. Interestingly, this motif is present in the TSR of CS but it is not important for CS binding to the hepatocyte surface (34). The *P. yoelii* and *P. falciparum* SPATR proteins share 63% amino acid sequence identity, including 12 conserved cysteine residues (Fig. 4*A*). The N-terminal intron of *SPATR* is conserved in both species (data not shown). This overall similarity suggests that the proteins are homologous. To confirm *SPATR* transcription, we hybridized a salivary gland and midgut sporozoite cDNA blot to a *SPATR*-specific probe. *SPATR* cDNA seemed more abundant in the midgut sporozoite preparations (Fig. 4*B*).

One EST showed weak similarity with Pbs48/45, a member of the six-cysteine (6-cys) superfamily (35). A *P. yoelii* contig from the *P. yoelii* genome project that matched this EST showed a single ORF of 1,440 bp coding for a predicted mature 52-kDa protein. Search of the *P. falciparum* genome database identified a putative homologue that shared 40% amino acid sequence identity with the *P. yoelii* protein (Fig. 5*A*). Both predicted proteins have consensus amino terminal cleavable signal peptides followed by two tandem 6-cys domains. A carboxyl-terminal hydrophobic domain indicated that the proteins could be membrane-anchored by a glycosylphasphatidylinositol linkage. The presence of the 6-cys domain and the overall structure clearly identified the proteins as new members of the 6-cys superfamily. According to the nomenclature of this superfamily by predicted molecular mass of the mature protein, we named the proteins Py52 and Pf52. To confirm *Py52* expression, we hybridized a salivary gland and midgut sporozoite cDNA blot to a *Py52* specific probe. *Py52* cDNA seemed more abundant in the midgut sporozoite preparations (Fig. 5*B*).

Finally, it is noteworthy that none of our ESTs resulted in significant matches with sporozoite-threonine asparagine-rich protein and liver stage antigen-3, proteins that have been described in *P. falciparum* sporozoites (12, 13).

Discussion

The nearly complete genome sequence of *P. falciparum* is now available, and its annotation will be concluded in the near future (36). It has been estimated that the 25–30 megabase genome harbors about 6,000 expressed genes. In addition, a 2× sequence coverage of the *P. yoelii* genome has very recently been completed and made publicly available (www.tigr.org). Malaria parasites occur in a number of different life cycle stages, making it a challenging task to determine which subset of the 6,000 genes is represented in the transcriptome of each stage. Microarrays will be the method of choice for expression analysis in asexual and sexual blood stage parasites where the acquisition of sufficient RNA is not a limitation. Although whole genome microarrays are not yet available, partial arrays from mung bean genomic libraries (37) or blood stage cDNA libraries (38) have been used successfully to study gene expression in blood stages. However, microarray analysis of gene expression in ookinetes, early oocysts, sporozoites, and EEF of mammalian Plasmodia will be difficult because large quantities of these stages are not available.

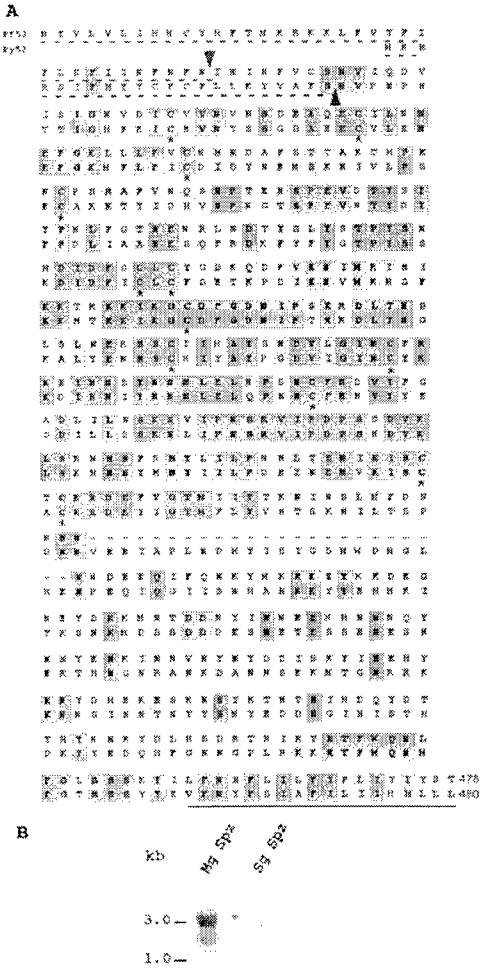


Fig. 5. Alignment of P52 and expression in sporozoites. (*A*) Comparison of the deduced amino acid sequences of the *P. yoelii*, Py52, with the homologue in *P. falciparum*, Pf52. The putative signal peptides are underlined with a dashed line. Putative signal peptide cleavage sites are marked with arrowheads (▲, ▼). Conserved cysteine residues of the tandem 6-cys motifs are marked with an asterisk (*). The carboxyl-terminal hydrophobic putative membrane anchor is underlined with a solid line. Identical residues are shaded dark gray. Conserved amino acid changes are shaded light gray, and radical changes are not shaded. (*B*) cDNA blot demonstrating *Py52* expression in midgut sporozoites (Mg Spz) and salivary gland sporozoites (Sg Spz). Sizes are given in kb.

Herein, we have described a survey of genes expressed in the infectious *Plasmodium* salivary gland sporozoite. We have demonstrated that, with a PCR-based amplification of the transcriptome, it is possible to obtain enough cDNA to construct a library for EST sequence acquisition. CS and SSP2/TRAP are highly expressed in the salivary gland sporozoites. On the basis of Western blot analysis of salivary gland sporozoites, CS is more abundant than SSP2/TRAP (data not shown), and this result is in agreement with the number of ESTs for CS (33 ESTs) and SSP2/TRAP (13 ESTs). We do not know whether the low number of ribosomal protein ESTs in the cDNA data set reflects true abundance of transcripts for those proteins in the sporozoite. PCR amplification of cDNA before cloning and sequencing could have biased the representation. Yet, it is possible that the bulk of proteins of the translation machinery are synthesized in the developing oocyst or in midgut sporozoites. The EST data set gives unprecedented insight into sporozoite gene expression, opening up new avenues of exploration. Expression of chorismate synthase in sporozoites is one example. The shikimate pathway was shown to be functional in blood stage *Plasmodium*, and the herbicide glyphosate had a clear inhibitory effect on parasite growth (15). If the shikimate pathway is also operational in sporozoites and EEF, inhibitory drugs (39) could be used to eliminate the preerythrocytic stages, avoiding progression to the blood stage and therefore disease.

The presence of MAEBL in the sporozoite stage raises interesting questions about its function. Binding of MAEBL to erythrocytes suggested that it had a role in merozoite red blood cell invasion (14). It will be worthwhile to investigate whether MAEBL also has a role in mosquito salivary gland and hepatocyte invasion, and therefore acts as a multifunctional parasite ligand in the merozoite and sporozoite stages. Regardless, its dual expression could make MAEBL the target of an inhibitory immune response against erythrocytic and preerythrocytic stages.

We show here that sporozoites express *SPATR*, coding for a putative secreted protein with a degenerate TSR. The CS protein and SSP2/TRAP each carry a TSR, and both proteins have demonstrated roles in sporozoite motility, host cell attachment, and invasion (34, 40–42). TSRs are also present in CS/TRAP-related protein (43), a protein essential for ookinete motility and host cell invasion (44–46).

The 6-cys motif defines a superfamily of proteins that seems to be restricted to the genus *Plasmodium* (35). Where studied, expression of members of this family was restricted to sexual erythrocytic stages. Recently, targeted gene disruption of *P48/45* identified the protein as a male gamete fertility factor (47). We have identified *Py52* and *Pf52* as genes coding for new members of the 6-cys family. *Py52* is expressed in sporozoites, and, like *SPATR*, *Py52* was expressed at higher level in midgut sporozoites than in salivary gland sporozoites. These expression patterns contrast with expression patterns of *SSP2/TRAP* and *CS*, which appeared equally abundant in both sporozoite stages (data not shown). Although we have not yet analyzed SPATR and Py52 protein expression, it is tempting to speculate, based on transcript level, that both proteins may have a role in sporozoite invasion of the mosquito salivary glands.

We have presented and discussed here only an initial analysis of the EST data set and further characterized a few selected examples with emphasis on putative sporozoite ligands for host cell attachment and invasion. A detailed analysis of all ESTs is beyond the scope of this first description. The amount of redundancy present in the EST data set is relatively low. It is therefore likely that the generation of more sequence data will identify novel sporozoite-expressed genes. However, many ESTs do not have significant database matches, and a number of ESTs produce matches with proteins of unknown function. A comprehensive expression analysis will determine which subset of the identified genes is exclusively expressed in the sporozoite stages. Sporozoite-specific genes are amenable to functional genetic analysis because loss-of-function mutants can be isolated and analyzed (48), a tool not yet available for genes essential in the asexual erythrocytic cycle (49). All told, we can now generate more of the urgently needed information about the sporozoite stage, a stage of the complex malaria life cycle that has so far eluded comprehensive experimental study.

Note Added in Proof. Recently, 1,117 additional ESTs were generated. These ESTs are not included in the analysis presented here. The additional ESTs have been deposited in the GenBank dbEST database (accession nos. BG603043–BG604160) and are also available through the *P. yoelii* gene index (http://www.tigr.org/tdb/pygi/).

We thank Tirza Doniger at the New York University School of Medicine Research Computing Resource for bioinformatics support. This work was supported by National Institutes of Health Grant AI-47102, the United Nations Development Program/World Bank/World Health Organization Special Program for Research and Training in Tropical Diseases (TDR), the Naval Medical Research Center Work Units 61102AA0101BFX and 611102A0101BCX, and a U.S. Army Medical Research and Material Command Contract (DAMD17-98-2-8005). S.H.I.K. is a recipient of the B. Levine fellowship in malaria vaccinology. We thank the scientists and funding agencies comprising the international Malaria Genome Project for making sequence data from the genome of *P. falciparum* (3D7) public prior to publication of the completed sequence. The Sanger Centre (Hinxton, U.K.) provided sequence for chromosomes 1, 3-9, and 13, with financial support from the Wellcome Trust. A consortium composed of the Institute for Genome Research, along with the Naval Medical Research Center (Silver Spring, MD) sequenced chromosomes 2, 10, 11, and 14, with support from the National Institute of Allergy and Infectious Diseases/National Institutes of Health, the Burroughs Wellcome Fund, and the Department of Defense. The Stanford Genome Technology Center sequenced chromosome 12, with support from the Burroughs Wellcome Fund. The Plasmodium Genome Database is a collaborative effort of investigators at the University of Pennsylvania and Monash University (Melbourne, Australia) supported by the Burroughs Wellcome Fund.

1. Holder, A. A. (1996) in *Malaria Vaccine Development: A Multi-Immune Response Approach*, ed. Hoffman, S. L. (Am. Soc. Microbiol., Washington, DC), pp. 35–75.
2. Nussenzweig, V. & Nussenzweig, R. S. (1989) *Adv. Immunol.* **45,** 283–334.
3. Nussenzweig, R. S. & Nussenzweig, V. (1989) *Rev. Infect. Dis.* **11,** S579–S585.
4. Schofield, L., Villaquiran, J., Ferreira, A., Schellekens, H., Nussenzweig, R. S. & Nussenzweig, V. (1987) *Nature (London)* **330,** 664–666.
5. Schofield, L., Ferreira, A., Altszuler, R., Nussenzweig, V. & Nussenzweig, R. S. (1987) *J. Immunol.* **139,** 2020–2025.
6. Ferreira, A., Schofield, L., Enea, V., Schellekens, H., van der Meide, P., Collins, W. E., Nussenzweig, R. S. & Nussenzweig, V. (1986) *Science* **232,** 881–884.
7. Sinnis, P. & Nussenzweig, V. (1996) in *Malaria Vaccine Development: A Multi-Immune Response Approach*, ed. Hoffman, S. L. (Am. Soc. Microbiol., Washington, DC), pp. 15–33.
8. Hoffman, S. L., Franke, E. D., Hollingdale, M. R. & Druilhe, P. (1996) in *Malaria Vaccine Development: A Multi-Immune Response Approach*, ed. Hoffman, S. L. (Am. Soc. Microbiol., Washington, DC), pp. 35–75.
9. Charoenvit, Y., Leef, M. F., Yuan, L. F., Sedegah, M. & Beaudoin, R. L. (1987) *Infect. Immun.* **55,** 604–608.
10. Rogers, W. O., Malik, A., Mellouk, S., Nakamura, K., Rogers, M. D., Szarfman, A., Gordon, D. M., Nussler, A. K., Aikawa, M. & Hoffman, S. L. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 9176–9180.
11. Robson, K. J., Hall, J. R., Jennings, M. W., Harris, T. J., Marsh, K., Newbold, C. I., Tate, V. E. & Weatherall, D. J. (1988) *Nature (London)* **335,** 79–82.
12. Fidock, D. A., Bottius, E., Brahimi, K., Moelans, I. M. D., Aikawa, M., Konings, R. N., Certa, U., Olafsson, P., Kaidoh, T., Asavanich, A., *et al.* (1994) *Mol. Biochem. Parasitol.* **64,** 219–232.
13. Daubersies, P., Thomas, A. W., Millet, P., Brahimi, K., Langermans, J. A. M., Ollomo, B., Mohamed, L. B., Slierendregt, B., Eling, W., Van Belkum, A., *et al.* (2000) *Nat. Med.* **6,** 1258–1263.
14. Kappe, S. H. I., Noe, A. R., Fraser, T. S., Blair, P. L. & Adams, J. H. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 1230–1235.
15. Roberts, F., Roberts, C. W., Johnson, J. J., Kyle, D. E., Krell, T., Coggins, J. R., Coombs, G. H., Milhous, W. K., Tzipori, S., Ferguson, D. J. P., Chakrabarti, D. & McLeod, R. (1998) *Nature (London)* **393,** 801–805.
16. Gardner, M. J., Tettelin, H., Carucci, D. J., Cummings, L. M., Aravind, L., Koonin, E. V., Shallom, S., Mason, T., Yu, K., Fujii, C., *et al.* (1998) *Science* **282,** 1126–1132.
17. Matuschewski, K., Mota, M. M., Pinder, J. C., Nussenzweig, V. & Kappe, S. H. I. (2001) *Mol. Biochem. Parasitol.* **112,** 157–161.
18. Wan, K. L., Blackwell, J. M. & Ajioka, J. W. (1996) *Mol. Biochem. Parasitol.* **75,** 179–186.
19. Ajioka, J. W., Boothroyd, J. C., Brunk, B. P., Hehl, A., Hillier, L., Manger, I. D., Marra, M., Overton, G. C., Roos, D. S., Wan, K. L., *et al.* (1998) *Genome Res.* **8,** 18–28.
20. Strong, W. B. & Nelson, R. G. (2000) *Mol. Biochem. Parasitol.* **107,** 1–32.
21. Chakrabarti, D., Reddy, G. R., Dame, J. B., Almira, E. C., Laipis, P. J., Ferl, R. J., Yang, T. P., Rowe, T. C. & Schuster, S. M. (1994) *Mol. Biochem. Parasitol.* **66,** 97–104.
22. Pinder, J. C., Fowler, R. E., Dluzewski, A. R., Bannister, L. H., Lavin, F. M., Mitchell, G. H., Wilson, R. J. & Gratzer, W. B. (1998) *J. Cell. Sci.* **111,** 1831–1839.
23. Margos, G., Siden-Kiamos, I., Fowler, R. E., Gillman, T. R., Spaccapelo, R., Lycett, G., Vlachou, D., Papagiannakis, G., Eling, W. M., Mitchell, G. H. & Louis, C. (2000) *Mol. Biochem. Parasitol.* **111,** 465–469.
24. Heintzelman, M. B. & Schwartzman, J. D. (1997) *J. Mol. Biol.* **271,** 139–146.
25. Heintzelman, M. B. & Schwartzman, J. D. (1999) *Cell Motil. Cytoskeleton* **44,** 58–67.
26. Bonhomme, A., Bouchot, A., Pezzella, N., Gomez, J., Le Moal, H. & Pinon, J. M. (1999) *FEMS Microbiol. Rev.* **23,** 551–561.
27. Kieschnick, H., Wakefield, T., Narducci, C. A. & Beckers C. (2001) *J. Biol. Chem.* **276,** 12369–12377.
28. Cassaing, S., Fauvel, J., Bessieres, M. H., Guy, S., Seguela, J. P. & Chap, H. (2000) *Int. J. Parasitol.* **30,** 1137–1142.
29. Mota, M. M., Pradel, G., Vanderberg, J. P., Hafalla, J. C. R., Frevert, U., Nussenzweig, R. S., Nussenzweig, V. & Rodriguez, A. (2001) *Science* **291,** 141–144.
30. Kappe, S. H. I., Curley, G. P., Noe, A. R., Dalton, J. P. & Adams, J. H. (1997) *Mol. Biochem. Parasitol.* **89,** 137–148.
31. Adams, J. H., Sim, B. K. L., Dolan, S. A., Fang, X., Kaslow, D. C. & Miller, L. H. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 7085–7089.
32. Lawler, J. & Hynes, R. O. (1986) *J. Cell Biol.* **103,** 1635–1648.
33. Adams, J. C. & Tucker, R. P. (2000) *Dev. Dyn.* **218,** 280–299.
34. Gantt, S. M., Clavijo, P., Bai, X., Esko, J. D. & Sinnis, P. (1997) *J. Biol. Chem.* **272,** 19205–19213.
35. Templeton, T. J. & Kaslow, D. C. (1999) *Mol. Biochem. Parasitol.* **101,** 223–227.
36. Carucci, D. J. & Hoffman, S. L. (2000) *Nat. Med.* **6,** 1–6.
37. Hayward, R. E., Derisi, J. L., Alfadhli, S., Kaslow, D. C., Brown, P. O. & Rathod, P. K. (2000) *Mol. Microbiol.* **35,** 6–14.
38. Mamoun, C. B., Gluzman, I. Y., Hott, C., MacMillan, S. K., Amarakone, A. S., Anderson, D. L., Carlton, J. M.-R., Dame, J. B., Chakrabarti, D., Martin, R. K., *et al.* (2001) *Mol. Microbiol.* **39,** 26–36.
39. McConkey, G. A. (1999) *Antimicrob. Agents Chemother.* **43,** 175–177.
40. Sinnis, P. (1996) *Infect. Agents Dis.* **5,** 182–189.
41. Sultan, A. A., Thathy, V., Frevert, U., Robson, K. J., Crisanti, A., Nussenzweig, V., Nussenzweig, R. S. & Ménard, R. (1997) *Cell* **90,** 511–522.
42. Kappe, S., Bruderer, T., Gantt, S., Fujioka, H., Nussenzweig, V. & Ménard, R. (1999) *J. Cell Biol.* **147,** 937–944.
43. Trottein, F., Triglia, T. & Cowman, A. F. (1995) *Mol. Biochem. Parasitol.* **74,** 129–141.
44. Dessens, J. T., Beetsma, A. L., Dimopoulos, G., Wengelnik, K., Crisanti, A., Kafatos, F. C. & Sinden, R. E. (1999) *EMBO J.* **18,** 6221–6227.
45. Yuda, M., Sakaida, H. & Chinzei, Y. (1999) *J. Exp. Med.* **190,** 1711–1716.
46. Templeton, T. J., Kaslow, D. C. & Fidock, D. A. (2000) *Mol. Microbiol.* **36,** 1–9.
47. van Dijk, M. R., Janse, C. J., Thompson, J., Waters, A. P., Braks, J. A. M., Dodemont, H. J., Stunnenberg, H. G., Van Gemert, G.-J., Sauerwein, R. W. & Eling, W. (2001) *Cell* **104,** 153–164.
48. Ménard, R. & Janse, C. (1997) in *Methods: A Companion to Methods in Enzymology—Analysis of Apicomplexan Parasites* (Academic, Orlando, FL), Vol. 13, pp. 148–157.
49. De Koning-Ward, T. F., Janse, C. J. & Waters, A. P. (2000) *Annu. Rev. Microbiol.* **54,** 157–185.

Appendix F
Final Report
DAMD17-82-2-8005

ELSEVIER

Molecular & Biochemical Parasitology 118 (2001) 133–138

MOLECULAR
& BIOCHEMICAL
PARASITOLOGY

www.parasitology-online.com.

Review

A status report on the sequencing and annotation of the *P. falciparum* genome

Malcolm J. Gardner *

The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA

Abstract

Almost 5 years ago, an international consortium of sequencing centers and funding agencies was formed to sequence the genome of the human malaria parasite *Plasmodium falciparum*. A novel chromosome by chromosome shotgun strategy was devised to sequence this very AT-rich genome. Two of the 14 chromosomes have been completed and the remaining chromosomes are in the final stages of gap closure. The consortium recently developed plans for the annotation and analysis of the complete genome sequence and its publication in 2002. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Genome; *Plasmodium falciparum*; Chromosome; Malaria

1. Introduction

The first complete genome sequence of a free-living organism, *Haemophilus influenzae*, was published in 1995 [1]. Besides proving the speed and cost-effectiveness of the whole genome shotgun (WGS) approach to genome sequencing, this work introduced many scientists to the value of a complete genome sequence in terms of providing insights into the biology, biochemistry, and pathogenicity of microorganisms that cause disease. Several other genome sequences were completed soon after, and today, at least 55 microbial genomes have been sequenced, including both pathogenic and non-pathogenic organisms. As once predicted [2], only a few years after the completion of the first microbial genome sequence, scientists working on many of the most important human pathogens have entered the 'post-genomic era of microbe biology', and are building upon the foundation provided by complete genome sequences to drive research into the development of new drugs and vaccines against these organisms.

Within a few months of the publication of the *H. influenzae* genome, several groups working on the human malaria parasite *P. falciparum* began to investigate the feasibility of determining its genome sequence. At the time, this seemed a daunting and perhaps impossible task. At an estimated 30 megabases (Mb), the *P. falciparum* genome was thought to be approximately 15-fold larger than that of *H. influenzae*, so large that the ~500,000 shotgun sequences required could not have been assembled with the existing assembly software (the genome is now known to be about 25 Mb [3]). In addition, the genome is very AT-rich, and most investigators working on *P. falciparum* were all too familiar with the difficulty of cloning the DNA in *E. coli*, where it was frequently subject to deletions and rearrangements that precluded construction of high-quality, large insert genomic libraries. Were these deletions and rearrangements to occur in the libraries used for sequencing, it would have been impossible to obtain the complete genome sequence. Large fragments (> 100 kb) of *P. falciparum* DNA had been cloned in yeast artificial chromosomes and were relatively stable [4,5], but it was not possible to subclone short fragments of the YAC inserts without a great deal of cross-contamination with yeast DNA, making the YAC libraries unsuitable for large scale sequencing. Another major

* Tel.: +1-301-838-3519; fax: +1-301-838-0208.
E-mail address: gardner@tigr.org (M.J. Gardner).

PII: S0166-6851(01)00390-5

concern was the projected price of the project. With the existing techniques and costs, it was estimated that sequencing of *P. falciparum* would require at least $15 million, a sum not easily obtained from the usual funding sources.

Further discussions amongst members of the malaria research community, the sequencing centers, and representatives from the Wellcome Trust, the National Institute for Allergy and Infectious Diseases, the Burroughs Wellcome Fund, and the U.S. Department of Defense culminated in the formation of an international consortium to sequence the genome of *P. falciparum* [6]. Sequencing was to be conducted by the Pathogen Sequencing Unit at the Sanger Center, the Stanford University Genome Sequencing Center, and The Institute for Genomic Research (TIGR) and the Malaria Program at the Naval Medical Research Center (NMRC). Start-up funds were obtained for projects to investigate various sequencing strategies and to develop reagents prior to initiation of a full-scale effort. Ultimately, a chromosome by chromosome shotgun strategy was devised whereby the 14 chromosomes were purified on pulsed field gels and sequenced individually using a shotgun approach similar to that used for bacterial genomes. This strategy enabled the genome to be divided among the three sequencing centers and partitioned the genome into more manageable segments for assembly and gap closure. The consortium also organized a series of semi-annual meetings beginning in December 1996 [6]. These meetings provided a forum for the sharing of technical information, review and coordination of sequencing and related activities, and development of a data use policy for the use of preliminary sequence data released by the sequencing centers. These meetings have continued to this day, but as the *P. falciparum* sequencing effort gained momentum the meetings evolved to cover such topics as genome databases, functional genomics, and comparative genomics of apicomplexans.

2. Strategy and methodology

The chromosome by chromosome shotgun strategy proved to be fairly effective in the sequencing of *P. falciparum*, although the extreme AT-richness of the genome made the closure process extremely difficult. Briefly, the chromosomes of *P. falciparum* clone 3D7 were resolved on pulsed field gels and chromosomal DNA was extracted by agarase digestion. The DNA was then sheared into 1–2 kb fragments, cloned into plasmid or M13 vectors, and randomly-picked clones were sequenced. Chromosomes 2, 10, 11 and 14 were assigned to TIGR and the NMRC, chromosome 12 to the Stanford group, and the remaining chromosomes were assigned to the Sanger Centre, including the 'blob' of mid-sized chromosomes that could not be resolved on gels. Most of the sequence reactions were performed on ABI 377 slab gel sequencers using dye-terminator chemistry. The sequences were assembled to form contigs using either phrap (www.phrap.org) or TIGR Assembler [7]. The Sanger and Stanford groups also performed low pass sequencing of shotgun libraries prepared from YAC clones previously localized on the chromosomes by the Wellcome Trust Malaria Genome Collaboration [8]. The YAC-derived sequences were used to 'bin' the sequences obtained from the chromosomal libraries into smaller subsets prior to assembly. After assembly, contigs from adjoining regions of the chromosomes were identified by means of forward–reverse links [1], and the groups of linked contigs were mapped to the chromosomes by means of STSs [8,9] or microsatellite markers [10,11], and by comparison to optical restriction maps of the chromosomes [3,12]. Most of the techniques used to close the remaining gaps were basically the same used for other genome projects [1], such as primer walking along plasmid templates that crossed sequence gaps, and closure of physical gaps by PCR amplification and sequencing of genomic DNA fragments that spanned the gaps. Other techniques, however, had to be devised to assist in the closure of very AT-rich regions. Many parts of the genome, such as the putative centromeric sequences identified on chromosomes 2 and 3 by Bowman et al. [13], were over 97% AT, and many regions in the vicinity of long runs of A's and T's proved very difficult to sequence accurately. In these cases, transposon insertion [14] or microlibrary techniques were used to generate a high sequence coverage across the AT-rich area, from which a more accurate sequence could be obtained (potential secondary structures in AT-rich areas that may have interfered with sequencing may also have been disrupted by insertion of a transposon or shotgunning of a fragment during microlibrary construction). These procedures are very labor intensive and time-consuming, however, and dealing with these AT-rich areas is one reason why closure of the *P. falciparum* genome has taken such a long time. Once whole chromosomes or substantial contigs were completed, the sequences were edited to resolve any ambiguities. Optical restriction maps [3,12] proved invaluable for verification that the chromosome sequences had been assembled correctly [14].

3. First glimpses of the *P. falciparum* genome

Completion of the first two chromosome sequences provided detailed pictures of chromosome organization and fascinating previews of the *P. falciparum* genome [14,13]. Some of the major findings included the discovery of two new gene families that were predicted to

encode potentially variant surface antigens (rifins and STEVORS [14–17]), a cluster of four genes of unknown function repeated on one end of chromosomes 2 and 3 [13], genes encoding enzymes of the type II fatty acid biosynthetic pathway previously thought to be restricted to plants and bacteria [14,18], and putative centromeres [13]. Gene density was just under 1 gene per 5 kb, very similar to *C. elegans*, and approximately one-half of genes were predicted to contain introns, although recent studies indicate this may have been an underestimate. Almost two-thirds of the predicted genes had no detectable orthologs in other organisms, suggesting that many aspects of parasite biology have not yet been uncovered despite many years of research.

In addition to the data release associated with the publication of chromosomes 2 and 3, preliminary contigs for all chromosomes have been released periodically, virtually since the beginning of the project, and have been accessible at the sequencing centers' web sites, at NCBI, and more recently at a new community database for malaria genome information, PlasmoDB [19] (www.plasmodb.org). Preliminary annotation is also available at these sites. A data release policy devised by the sequencing centers and the funding agencies, with input from members of the malaria research community, was also established. Although somewhat controversial [20–22], the data release policy allowed many scientists around the world to get early glimpses of the genome sequence data and 'provide ... information that may jump-start biological experimentation' (www.tigr.org/tdb/edb2/pfa1/htmls/), while protecting the right of the sequencing centers to publish whole chromosome or whole genome analyses of the data they had so laboriously produced. Dozens of reports have since been published in which use of the preliminary sequence data was acknowledged. Virtually every area of malaria biology and biochemistry has been positively affected by the release of preliminary genome sequence information. Some of the most outstanding discoveries have been the identification of new drug targets, opening new avenues for the development of novel antimalarials [23–26]. Many other research projects that rely in some fashion on the preliminary sequence data are underway, including the development of full-genome microarrays and proteomics studies.

4. Current status and plans for annotation

The consortium met at the Sanger Centre in June 2001 to review progress in gap closure and make plans for annotation and publication of the *P. falciparum* genome sequence. Chromosomes 2 and 3 have been published. Chromosomes 1, 4, 9–12 and 14 were reported to be in the final stages of closure, with only a handful of gaps per chromosome remaining. Chromosomes in the 'blob' (chromosomes 5–8) and chromosome 13, have full sequence coverage but are lagging behind in gap closure, and still consist of hundreds of contigs per chromosome. Gap closure has been slow due to the paucity of markers for ordering of the contigs. However, the Sanger Centre recently began Happy Mapping [27,28] of the contigs and should complete this task in the fall of 2001. Once these contigs have been grouped and ordered along the chromosomes, gap closure for these chromosomes is expected to accelerate.

The sequencing centers also laid plans for annotation of the genome sequence, leading early in 2002, to submission of joint publication on the analysis of the entire *P. falciparum* genome and a series of papers on the chromosomes by the three sequencing groups. The basic elements of this plan include beginning the annotation on a set of contig sequences representing the best available data for each chromosome. These contigs would be 'frozen' so to permit annotation to proceed on a stable data set, and where possible the contigs will be joined end-to-end in the correct order and orientation to form draft chromosome sequences. As the annotation of these draft chromosomes is underway, closure efforts on the remaining gaps will continue, and the new sequence data generated during the closure process will be merged into the annotated contigs near the end of the process. Each sequencing center will be responsible for annotation of the chromosomes they sequenced, using the software and methods in use at each center. In an attempt to ensure that the annotation done by the participating centers is of equal quality, the same 100 kb sequence will be annotated by the three groups early in the annotation process and the results will be compared to identify any problems. Furthermore, it was agreed that TIGR will maintain a central relational database containing a representation of the sequence data and annotation produced at all three centers, and that the centers will develop procedures for the frequent semi-automated exchanges of data. This will allow all of the annotators to view the same picture of the complete genome and facilitate whole genome analyses. Importantly, this arrangement will also simplify the process of submitting the annotated genome sequence to the PlasmoDB database [19]. This plan has now been put in motion. Many chromosome sequences have been frozen, annotation has begun, and the system for data exchange between centers is being tested.

As with annotation of chromosomes 2 and 3, and other eukaryotic genomes, including the human genome [29–31], annotation of the complete *P. falciparum* genome presents many challenges. A major problem is the difficulty of gene prediction in eukaryotic genomes. Two gene finders specifically designed to predict gene models in the gene-dense *P. falciparum* genome are now available, GlimmerM [32] and phat [33]. Both programs

perform well but predict different gene models in some cases. The human annotator, faced with conflicting models and in many instances with no other evidence such as EST hits or protein matches to confirm either model, has great difficulty in deciding which model, if any, is likely to be the correct one. Subjective criteria (otherwise known as 'the force') must sometimes be employed in selecting one model over another. Another problem is that the gene finders do not detect genes in some regions of the genome where they would be expected to occur, suggesting that some genes may have escaped detection. It was for this reason that the systematic gene nomenclature system devised for *P. falciparum* numbers the genes in increments of five, to allow genes identified later to be neatly inserted into the annotation [14]. The gene finders are also unable to handle the complexities of alternative splicing. In short, gene models are predictions, and investigators using the annotation would be wise to verify the gene models experimentally prior to embarking on detailed studies of these genes. In an attempt to improve gene modeling during the whole genome annotation that is just beginning, the original training set used for GlimmerM [32] was recently updated to include experimentally-verified genes published over the past 3 years, and both GlimmerM and phat have been re-trained on new training set. EST datasets from a variety of organisms have also been updated [34] (www.tigr.org/tdb/tgi.html); these can provide the annotator with experimental evidence to support complete gene models or intron predictions. Genome annotation is an ongoing process, and the upcoming annotation of the complete *P. falciparum* genome should be viewed as the first step in a process that will continue for many years. Continual feedback from the malaria research community, in the form of experimentally verified gene structures, will be essential in order to improve the genome annotation process.

One major improvement over the previous annotation of chromosomes 2 and 3 will be in the assignment of genes into functional role categories. For chromosomes 2 and 3, existing role categories that had been devised for prokaryotes were adapted for use with a eukaryotic organism. Both centers used different schemes, and neither scheme captured the increased complexity of eukaryotic biology. To avoid this situation, the whole genome annotation will use the Gene Ontology (GO) system that is currently used by several organism-specific databases including the *Saccharomyces cerevisiae* database SDB and FlyBase, among others [35]. The GO system consists of three separate ontologies (molecular function, biological process, and cellular component), each with 'a set of structured vocabularies ... that can be used to describe gene products in any organism [36]. A group of parasitologists, coordinated by Matt Berriman at the Sanger Centre, is currently drafting a set of defined terms that describe novel aspects of parasite biology for inclusion into the GO system (e.g. the term 'rosetting' in the biological process ontology). Thus, the *P. falciparum* annotation will use a more powerful and widely utilized gene system of gene product classification, enabling users to gain broader insights into parasite biology from the annotated genome sequence.

5. Sequencing of additional *P. falciparum* clones and *Plasmodium* spp.

The success of the *P. falciparum* sequencing effort led many investigators to call for sequencing of additional *Plasmodium* spp. and a more recent isolate(s) of *P. falciparum*. Of particular interest was generation of sequence data for many of the malaria parasites used as model systems for drug and vaccine development, and for *Plasmodium vivax*, the second most important human malaria parasite. Although the chromosome by chromosome approach to sequencing of *P. falciparum* has been successful, there are a number of reasons why it would be best to avoid this strategy when additional *Plasmodium* spp. are sequenced. One, the introduction of capillary-based sequencers such as the ABI 3700 has dramatically increased the sequencing capacity of genome centers while simultaneously lowering the costs of sequencing. At the time the malaria genome project was started, sequencing of a 30 Mb genome would have been a very large project. Today, the random sequences required for such a project can be generated much more quickly and at lower expense than even 2 years ago, so that dividing the sequencing of a 30 Mb genome between several centers would not be required for purely logistical reasons. In addition, a major problem with the slab gel sequencers was the high frequency of mistracked sequences, which interfered with the assembly and contig grouping procedures. Mistracking does not occur with the capillary based sequencers in which each reaction is contained within a single capillary during electrophoresis. Two, preparations of pulsed field gel purified chromosomal DNA were always cross-contaminated with DNA from other chromosomes. For chromosome 2, we estimated that 20% of the sequences obtained were from other regions of the genome. The cross-contamination resulted in the formation of many short, low coverage contigs that confounded the gap closure process, and since every separate chromosome project generated its own set of 'contaminants,' more sequences had to be generated in the chromosome by chromosome approach to produce the required sequence coverage of a chromosome than might have been required using the whole genome strategy. Third, the chromosome by chromosome strategy was necessitated, in part, by the inability of the existing assembly software to assemble the $\sim$500,000 shotgun sequences

that would have been produced by a whole genome shotgun approach. In fact, the version of the TIGR Assembler that was available early in the chromosome 2 pilot project had difficulty assembling 9000 sequences in a reasonable time frame. More recent versions of TIGR Assembler and new assemblers such as the Celera Assembler [37] can handle much larger data sets. In summary, future efforts to sequence another clone of *P. falciparum*, perhaps from a recent clinical isolate, or another species of *Plasmodium*, could be done using a whole genome approach at any one of several sequencing centers. Several such efforts are already underway or in the planning stages, including the sequencing of *P. yoelii* and *P. vivax* to 5 × coverage (TIGR/NMRC), and five other *Plasmodium* spp. to 3 × coverage (*P. chabaudi*, *P. berghei*, *P. knowlesi*, *P. reichenowi*, and *P. gallinaceum*) by the Sanger Centre. These projects should produce contigs of 2–5 kb representing >90% of the parasites' genomes. Besides providing gene sequences that can be used to facilitate a variety of functional studies, these projects will allow comparative genome analyses of *Plasmodium* spp. that have very different biological characteristics. Several other apicomplexan parasites are also being sequenced, including *Theileria parva* (TIGR and the International Livestock Research Institute), *T. annulata* (The Sanger Centre), and two isolates of *Cryptosporidium parvum* (University of Minnesota and the Medical College of Virginia).

6. Summary

After 5 years of extraordinary effort by the consortium, completion of the *P. falciparum* genome sequence appears imminent. Analysis of the first two chromosomes to be sequenced, which together represented about 8% of the genome, and exciting findings that were made possible by release of preliminary sequence data, have already justified the efforts made to sequence the genome of this deadly parasite. Annotation of the genome sequence has begun following a plan devised be the three sequencing centers and publication of an analysis of the *P. falciparum* genome is expected in 2002. The success of the *P. falciparum* project has spawned similar efforts to determine the genome sequences of additional *Plasmodium* spp. and other apicomplexans. In addition, the human genome sequence [29,30], and the *Anopheles gambiae* genome sequence that is also expected to be completed in 2002 (www.niaid.nih.gov/newsroom/releases/celera.htm), provide opportunities for study of host–vector–parasite relationships. In the years to come, the complete genome sequences of all three members of the *Plasmodium* life cycle will allow investigators to gain a better understanding of parasite biology and will be invaluable resources in the quest to develop new drugs and vaccines to fight malaria.

Acknowledgements

I thank my colleagues at TIGR, the Naval Medical Research Center, and the members of the Malaria Genome Sequencing Consortium at the Sanger Centre and the Stanford Genome Technology Center, for their support. Sequencing of the *P. falciparum* genome at TIGR and the NMRC is supported by the National Institute of Allergy and Infectious Diseases (U01 AI42243), the Burroughs Wellcome Fund (990785), the Department of the Army (DAMD17-98-2-8005), and Naval Medical Research Center Work Unit Nos. 61102A.S13.00101.BFX1431, 612787A.870.00101.EFX.1432, 623002A.810.00101.HFX.1433 and STEP C611102A0101BCX.

References

[1] Fleischmann RD, Adams MD, White O, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 1995;269:496–512.
[2] Bloom BR. A microbial minimalist. Nature 1995;378:236.
[3] Lai Z, Jing J, Aston C, et al. A shotgun optical map of the entire *Plasmodium falciparum* genome. Nat Genet 1999;23:309–13.
[4] Triglia T, Kemp DJ. Large fragments of *Plasmodium falciparum* DNA can be stable when cloned in yeast artificial chromosomes. Mol Biochem Parasitol 1991;44:207–11.
[5] de Bruin D, Lanzer M, Ravetch JV. Characterization of yeast artificial chromosomes from *Plasmodium falciparum*: construction of a stable, representative library and cloning of telomeric DNA fragments. Genomics 1992;14:332–9.
[6] Hoffman SL, Bancroft WH, Gottlieb M, et al. Funding for malaria genome sequencing. Nature 1997;387:647.
[7] Sutton GS, White O, Adams MD, et al. TIGR assembler: a new tool for assembling large shotgun sequencing projects. Genome Sci Technol 1995;1:9–19.
[8] Foster J, Thompson J. The *Plasmodium falciparum* genome project: a resource for researchers. Parasitol Today 1995;11:1–4.
[9] Dame JB, Arnot DE, Bourke PF, et al. Current status of the *Plasmodium falciparum* genome project. Mol Biochem Parasitol 1996;79:1–12.
[10] Su XZ, Wellems TE. Toward a high-resolution *Plasmodium falciparum* linkage map:polymorphic markers from hundreds of simple sequence repeats. Genomics 1996;33:430–44.
[11] Su XZ, Wellems TE. *Plasmodium falciparum*: assignment of microsatellite markers to chromosomes by PFG-PCR. Exp Parasitol 1999;91:367–9.
[12] Jing J, Aston C, Zhongwu L, et al. Optical mapping of *Plasmodium falciparum* chromosome 2. Genome Res 1999;9:175–81.
[13] Bowman S, Lawson D, Basham D, et al. The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*. Nature 1999;400:532–8.
[14] Gardner MJ, Tettelin H, Carucci DJ, et al. Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. Science 1998;282:1126–32.
[15] Cheng Q, Cloonan N, Fischer K, et al. Stevor and rif are *Plasmodium falciparum* multicopy gene families which potentially encode variant antigens. Mol Biochem Parasitol 1998;97:161–76.
[16] Fernandez V, Hommel M, Chen Q, et al. Small, clonally variant antigens expressed on the surface of the *Plasmodium falciparum*-infected erythrocyte are encoded by the rif gene family and are

the target of human immune responses. J Exp Med 1999;190:1393–404.
[17] Kyes SA, Rowe JA, Kriek N, et al. Rifins: a second family of clonally variant proteins expressed on the surface of red cells infected with *Plasmodium falciparum*. Proc Natl Acad Sci USA 1999;96:9333–8.
[18] Waller RF, Keeling PJ, Donald RGK, et al. Nuclear-encoded proteins target to the plastid in *Toxoplasma gondii* and *Plasmodium falciparum*. Proc Natl Acad Sci USA 1998;95:12 352–2 357.
[19] The Plasmodium Genome Database Collaborative. PlasmoDB: an integrative database of the *Plasmodium falciparum* genome. Tools for accessing and analyzing finished and unfinished sequence data. Nucleic Acids Res 2001;29:66–9.
[20] Gottlieb M, McGovern V, Goodwin P, et al. Please don't downgrade the sequencers' role. Nature 2000;406:121–2.
[21] Macilwain C. Biologists challenge sequencers on parasite genome publication. Nature 2000;405:601–2.
[22] Pace T. When public-interest science needs solidarity. Nature 2000;406:122.
[23] Jomaa H, Wiesner J, Sanderbrand S, et al. Inhibitors of the nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs. Science 1999;285:1573–6.
[24] Vollmer M, Thomsen N, Wiek S, et al. Apicomplexan parasites possess distinct nuclear-encoded, but apicoplast-localized, plant-type ferredoxin-NADP + reductase and ferredoxin. J Biol Chem 2001;276:5483–90.
[25] Lee CS, Salcedo E, Wang Q, et al. Characterization of three genes encoding enzymes of the folate biosynthetic pathway in *Plasmodium falciparum*. Parasitology 2001;122(Pt. 1):1–13.
[26] Salcedo E, Cortese JF, Plowe CV, et al. A bifunctional dihydrofolate synthetase—folylpolyglutamate synthetase in *Plasmodium falciparum* identified by functional complementation in yeast and bacteria. Mol Biochem Parasitol 2001;112:239–52.
[27] Dear PH, Cook PR. Happy mapping: linkage mapping using a physical analogue of meiosis. Nucleic Acids Res 1993;21:13–20.
[28] Piper MB, Bankier AT, Dear PH. A HAPPY map of *Cryptosporidium parvum*. Genome Res 1998;8:1299–307.
[29] Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. Science 2001;291:1304–51.
[30] Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. Nature 2001;409:860–921.
[31] Hogenesch JB, Ching KA, Batalov S, et al. A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. Cell 2001;106:413–5.
[32] Salzberg SL, Pertea M, Delcher A, et al. Interpolated Markov models for eukaryotic gene finding. Genomics 1999;59:24–31.
[33] Cawley SE, Wirth AI, Speed TP. Phat—a gene finding program for *Plasmodium falciparum*, Mol Biochem Parasitol, in press.
[34] Carlton JM-R, Muller R, Yowell CA, et al. Profiling the malaria genome: a gene survey of three species of malaria parasite with comparison to other apicomplexan species, Mol Biochem Parasitol 2001:118;201–210.
[35] The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. Nat Genet 2000;25:25–9.
[36] The Gene Ontology Consortium. Creating the gene ontology resource: design and implementation. Genome Res 2001;11:1425–33.
[37] Huson DH, Reinert K, Kravitz SA, et al. Design of a compartmentalized shotgun assembler for the human genome. Bioinformatics 2001;17(Suppl. 1):S132–9.

Interpolated Markov Models for Eukaryotic Gene Finding

Appendix G

Final Report
DAMD17-82-2-8005

Steven L. Salzberg,*,† Mihaela Pertea,† Arthur L. Delcher,‡,§
Malcolm J. Gardner,* and Hervé Tettelin*

*The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850; †Department of Computer Science, Johns Hopkins University, Baltimore, Maryland 21218; ‡Department of Computer Science, Loyola College in Maryland, Baltimore, Maryland 21210; and §Celera Genomics, 45 W. Gude Dr., Rockville, Maryland 20850

Reprinted from GENOMICS, Volume 59, No. 1, July 1, 1999

Genomics **59,** 24–31 (1999)
Article ID geno.1999.5854, available online at http://www.idealibrary.com on IDEAL®

Interpolated Markov Models for Eukaryotic Gene Finding

Steven L. Salzberg,*,†,[1] Mihaela Pertea,† Arthur L. Delcher,‡,§
Malcolm J. Gardner,* and Hervé Tettelin*

*The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850; †Department of Computer Science, Johns Hopkins University, Baltimore, Maryland 21218; ‡Department of Computer Science, Loyola College in Maryland, Baltimore, Maryland 21210; and §Celera Genomics, 45 W. Gude Dr., Rockville, Maryland 20850

Received January 19, 1999; accepted April 13, 1999

Computational gene finding research has emphasized the development of gene finders for bacterial and human DNA. This has left genome projects for some small eukaryotes without a system that addresses their needs. This paper reports on a new system, GLIMMERM, that was developed to find genes in the malaria parasite *Plasmodium falciparum.* Because the gene density in *P. falciparum* is relatively high, the system design was based on a successful bacterial gene finder, GLIMMER. The system was augmented with specially trained modules to find splice sites and was trained on all available data from the *P. falciparum* genome. Although a precise evaluation of its accuracy is impossible at this time, laboratory tests (using RT-PCR) on a small selection of predicted genes confirmed all of those predictions. With the rapid progress in sequencing the genome of *P. falciparum,* the availability of this new gene finder will greatly facilitate the annotation process.

1. INTRODUCTION

The gene finding research community has focused considerable effort on human and bacterial genome sequence analysis. This is not surprising given the attention paid to both areas. The Human Genome Project has produced many millions of nucleotides of sequence, and the importance of rapidly identifying the genes in this sequence cannot be overstated. This task is made difficult by the fact that only 1 to 3% of human genomic sequence is estimated to code for proteins. On the bacterial side, 20 complete bacterial and archaeal genomes have already been published, with dozens more expected in the next 2 years. Gene finders for these prokaryotes have an advantage in that approximately 90% of the DNA of these genomes is coding; thus the task reduces in many cases to choosing between competing reading frames. On the other hand, the demand for accuracy is correspondingly much higher in the prokaryotic world.

[1] To whom correspondence should be addressed. Telephone: (301) 315-2537. Fax: (301) 838-0209. E-mail: salzberg@tigr.org.

In between these two genomic worlds lies a vast array of eukaryotic organisms whose genomes range in size from that of a large prokaryote (on the order of tens of millions of nucleotides) to those that are larger than human (billions of nucleotides). Their gene density tends to be much lower than that of bacteria, but many organisms have a much higher gene density than humans. For example, the genome of the eukaryote *Saccharomyces cerevisiae* has approximately one gene every 5 kb. This corresponds to a gene density of 20%. Recently, chromosome 2 of the malaria parasite *Plasmodium falciparum* was completed (Gardner *et al.,* 1998), and this organism too has a gene density of 20%. The remaining 13 chromosomes from malaria should be completed over the course of the next few years. The much larger (120 million nucleotides) genome of *Arabidopsis thaliana,* which also is expected to have a gene density of approximately 20%, should be completed in the same time frame, and many projects are under way to sequence other small eukaryotes.

Because of their relatively high gene density with respect to human DNA, using a gene finder developed for human sequence (or other organisms with low gene density, including most vertebrates and larger plant genomes) may not be the optimal approach for *P. falciparum* and other small eukaryotes. Prokaryotic gene finders are not well suited to this task because of their inability to handle introns. It is possible to retrain human gene finders using different data (for example, GENSCAN (Burge and Karlin, 1997) has been trained with *Arabidopsis* data), but one still runs the risk that because these systems have been optimized to find genes in DNA that is only 3% coding, they may miss many genes in genomes such as *P. falciparum.*

This paper describes a gene finder developed specifically for small eukaryotes with a gene density of around 20%. This system, GLIMMERM, was built and trained using data from *P. falciparum,* the malaria parasite. It was then used as the principal gene finder for chromosome 2 of *P. falciparum,* which contains 210 genes (209 protein coding genes plus one tRNA) (Gardner *et al.,* 1998). Most of these genes were found by

0888-7543/99 $30.00

GLIMMERM, and as described below, some predictions were confirmed by additional laboratory experiments.

The basis of GLIMMERM is a dynamic programming algorithm that considers all combinations of possible exons for inclusion in a gene model and chooses the best of these combinations. Dynamic programming (DP) has been the basis of many successful eukaryotic gene finders. Hidden Markov model (HMM) systems use a DP algorithm called Viterbi that is a special case of the algorithm here; these HMM methods include VEIL (Henderson *et al.,* 1997); GENSCAN (Burge and Karlin, 1997), which uses semi-Markov HMMs; and Genie (Kulp *et al.,* 1996), which uses generalized HMMs. Very recently, Wirth (1998) described a gene finder for *P. falciparum* based on generalized HMMs, but it is not yet available for comparison. The Morgan system (Salzberg *et al.,* 1996, 1998a) uses a DP algorithm in combination with a decision tree program, and GeneParser (Snyder and Stormo, 1995) uses DP combined with a neural network program. These latter two DP formulations are most similar to the formulation used for GLIMMERM.

2. METHODS AND ALGORITHMS

The phrase "gene model" will be used to denote a particular combination of exons and introns that the system is considering as a possible gene. The decision about what gene model is best is a combination of the strength of the splice sites and the score of the exons produced by an interpolated Markov model (IMM). The methods for producing the IMM and splice site scores are described next, followed by the description of the dynamic programming algorithm that uses these scores.

2.1. Interpolated Markov Models

Markov chains are a family of methods for computing the probability of an event based on a fixed number of previous events. (More formally, a Markov chain is a sequence of random variables X_i, where the probability distribution for each X_i depends only on $X_{i-1}, \ldots, X_{i-k}$ for some constant k.) In the context of DNA sequence analysis, Markov chains predict a base by examining a fixed number of bases just prior to that base in the sequence. The most common type of Markov chain is a fixed-order chain, in which the number of previous bases to examine is a constant. For example, a fifth-order Markov chain will predict a base by looking at the five previous bases. Markov chains, and fifth-order chains in particular, have proven to be effective at gene prediction in bacterial genomes (Borodovsky and McIninch, 1993; Borodovsky *et al.,* 1995).

IMMs are a generalization of fixed-order Markov chains. The main distinction is that rather than deciding in advance how many bases to consider for each prediction, these models will use varying numbers of bases for each prediction. In some contexts they will use 5 bases, while in others they might use 6 or more bases, and in yet other cases they may use 4 or fewer bases. This allows IMMs to be sensitive to how common a particular oligomer is in a given genome. In a given genome, many 5-mers might occur rarely and should not be used for prediction; here the IMM will fall back on a shorter Markov chain. On the other hand, certain 8-mers may occur very frequently, and for those the IMM can use this longer context and make a better prediction. In addition, the IMM can combine the evidence from the eighth-order Markov chain and the fifth-order chain in such cases. Thus it has all the information available to a fifth-order chain plus additional information. It is also worth noting that both IMMs and fifth-order Markov chains should outperform methods based on codon usage statistics. (Cf. Saul and Battistutta (1988), a codon usage method specific to *P. falciparum.* Note that at the time of that work, much less *Plasmodium* data were available, and higher-order statistics might have been inaccurate as a result.)

IMMs form the basis of the GLIMMER system for finding genes in bacteria and archaea (Salzberg *et al.,* 1998b). GLIMMER correctly identifies approximately 98% of the genes in bacteria without any human intervention and with a very limited number of false-positives. It has been used as the gene finder for *Borrelia burgdorferi* (Fraser *et al.,* 1997), *Treponema pallidum* (Fraser *et al.,* 1998), *Chlamydia trachomatis* (Stephens *et al.,* 1998), *Thermotoga maritima* (Nelson *et al.,* submitted for publication), and others. Based on the success of GLIMMER in bacterial sequence annotation, we thought that IMMs should make a good foundation for eukaryotic gene finding. This is particularly true of small eukaryotes like *P. falciparum* in which the gene density is intermediate between that of prokaryotes and higher eukaryotes.

Details of how to construct an IMM for sequence data can be found in the original GLIMMER publication (Salzberg *et al.,* 1998b); GLIMMERM uses the same IMM algorithm as that described there. In brief, GLIMMERM builds IMMs from a set of DNA sequences chosen for training. For coding regions, it builds three separate IMMs, one for each codon position. (This is known as a 3-periodic Markov model (Borodovsky and McIninch, 1993).) These IMMs include zeroth- through eighth-order Markov chains, as well as weights computed for every oligomer of 8 bases or less that appears in the training data. These weights and Markov models are interpolated to produce a score for each base in any potential coding sequence. The logs of these scores are summed to score each coding region.

2.2. Splice Site Identification

The approach used by GLIMMERM to determine the splice sites is similar to that used in the Morgan human gene finding system (Salzberg *et al.,* 1998a). A second-order Markov chain model is used to score a 16-base region around donor sites and a 29-base region around acceptor sites. For both donor and acceptor sites in *P. falciparum,* a wide range of different regions were tested, and these sizes performed best. Two second-order Markov models were built for each type of site. First, a "true" Markov model was created from existing data on known 5′ and 3′ consensus sites. These data were collected by exhaustively combing the literature for every documented exon–intron boundary. A "false" Markov model was built from a large number of randomly chosen false splice sites, i.e., sequences that contained the consensus GT or AG dinucleotide but that were not true splice sites. The score of a site $s_i, s_{i+1}, \ldots, s_j$ was computed by each Markov model according to the formula

$$S(i, j) = \sum_{k=i}^{j} M_{s,k},$$

where

$$M_{s,k} = \ln(f((s_{k-2}, s_{k-1}, s_k), k)/f((s_{k-2}, s_{k-1}), k - 1)),$$

and $f(s, k)$ is the frequency of substring s ending at location k. Note that for the leftmost position in the splice site region, M is taken to be the probability given by the zeroth-order Markov model, and for the second position, M is given by the first-order model. The score for a given splice site is computed by taking the difference of the scores obtained from the true site Markov model and the false site model.

After building the models, we scored all the true splice sites and a large selection of randomly chosen false sites. We then set minimum cut-off scores to identify correctly most (or all) true sites and measured how many false-positives we would expect with various thresholds. The splice sites for training the Markov models were taken from the 119 genes (described under Results and Discussion) used to train the IMMs, all of which had laboratory evidence to support them. These genes contained only 81 introns in total, which did not gener-

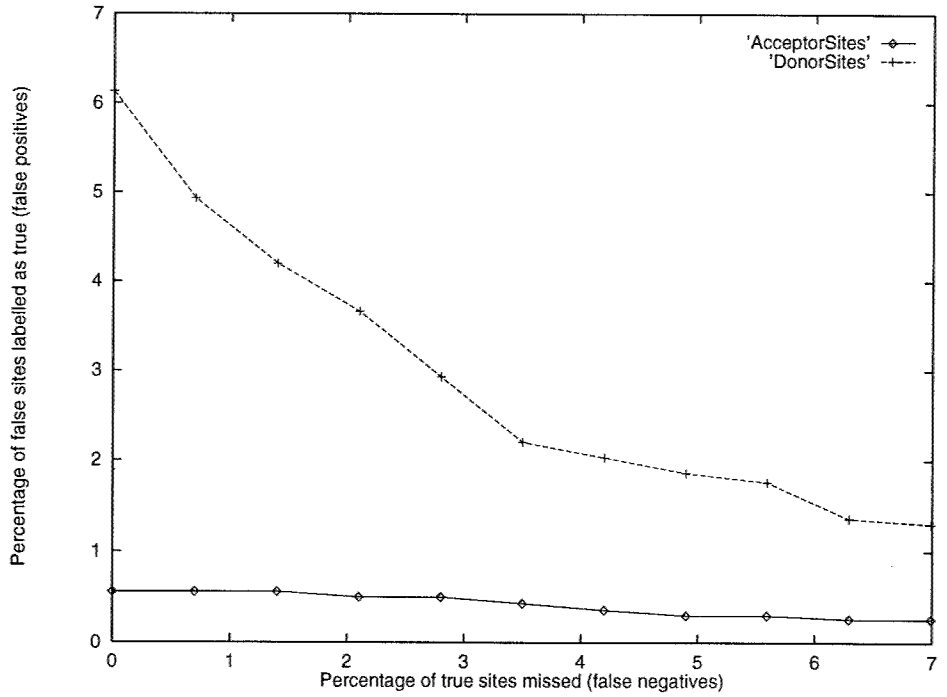


FIG. 1. Trade-off between false-positive rates and false-negative rates for the Markov chain method that recognizes exon–intron splice sites. Data represent the accuracy on sites annotated in chromosome 2 of *P. falciparum*.

ate enough data to produce a very reliable second-order Markov model. Therefore, after an initial training pass using the 81 introns, we used GLIMMERM itself to predict additional introns in chromosome 2, selected the best of these, and added them to the training set. Of course this is a "circular" training protocol, but this represents our attempt to squeeze the best performance we could from limited data. As the sequencing of the remaining chromosomes continues, and as ESTs yield further hard evidence on introns, the available pool of reliable data for training the splice site models should grow dramatically. Alignments with protein sequences from other organisms will provide additional evidence about intron locations. The Markov chain models will consequently improve in accuracy. We intend to continue retraining these models as the genome sequencing progresses.

Figure 1 shows the trade-off in thresholds for the splice site recognition function in *P. falciparum* and shows the trade-off between sensitivity and selectivity for the Markov chain method on the 143 donor and acceptor sites in chromosome 2. Acceptor sites are much easier to recognize: with a false-negative rate of 0% (corresponding to a sensitivity of 100%, meaning that all true sites will be recognized), the false-positive rate—the percentage of AG dinucleotides that will incorrectly be called acceptor sites—is just 0.56%. For donor sites, a 0% false-negative rate corresponds to a rather high 6.1% false-positive rate. Setting the system so that it misses 4 of the 143 (2.8%) donor sites in chromosome 2 would reduce this false-positive rate to 2.9%. The Markov thresholds used here are set so that no true splice sites will be missed.

2.3. Dynamic Programming

GLIMMERM's use of dynamic programming allows it to prune out a large number of possible exon–intron combinations and focus its analysis only on relatively high-scoring combinations (called "parses"). The input to the algorithm is any genomic DNA sequence in FASTA format; small sequences as well as entire chromosomes can be input. The output is a partitioning of the DNA into coding regions interleaved with noncoding regions, on both the main and the complementary strands of the sequence.

As in many other gene finders (Salzberg, 1998), there are a number of assumptions used by GLIMMERM when predicting genes in the DNA sequence. The main assumptions are (1) the coding region of every gene begins with a start codon ATG, (2) a gene has no in-frame stop codons except the very last codon, and (3) each exon is in a consistent reading frame with the previous exon. These constraints significantly enhance the efficiency of computing the optimal gene models, by restricting the search space of the DP algorithm. On the other hand, genuine frameshifts cannot be detected by the system.

The dynamic programming algorithm fills in a structure *Parse,* in which each element *Parse* $[t, n, S]$ denotes the optimal parse of the subsequence that begins at location n and ends at the stop codon at location S. The variable t specifies the type of signal at n, which can be donor, acceptor, start (codon), or stop (codon). More specifically, *Parse* is an ordered list of labeled positions indicating the end-points of a set of exons. For example,

$$Parse[\text{start}, 100, 540] = \langle \text{start}, 100\rangle, \langle \text{donor}, 240\rangle, \langle \text{acceptor}, 380\rangle, \langle \text{stop}, 540\rangle$$

indicates a pair of exons at positions [100 . . . 239] and [380 . . . 539]. A complete gene model is represented as a list *Parse* [start, n, S]. Other elements are partial parses, beginning at a location of type t ($t \neq$ start) and ending at a stop codon S.

The DP algorithm processes the input sequence left to right, looking for stop codons. At each stop codon S, it searches back in the 5′ direction and finds all possible genes ending at that stop. It chooses the highest scoring gene to store in *Parse*. More concisely,

$$Parse[t, n, S] = \langle t, n\rangle, Parse[t_{\text{next}}, i, S],$$

where i is the location that achieves the maximum score

$$\max_{n<i<S}\{Score(\langle t, n\rangle, Parse[t_{\text{next}}, i, S])\},$$

and t_{next} is the type logically following the type t in a parse. For example, if t = acceptor, then t_{next} can be either donor or stop. *Score* (*Parse* [t, n, S]) is the score given by the IMMs to the coding region obtained by concatenating all the exons in the parse delimited by

Parse [t, n, S]. For example, if n is an acceptor site, the algorithm considers all sites i that can follow n and chooses the best one. These would include donor sites, if n is the beginning of an internal exon, and stop codons, if n is the final coding exon. Because the algorithm works backward from each stop codon S, the entry *Parse* [t_{next}, i, S] is computed prior to *Parse* [t, n, S]. The only positions that are considered as possible donor and acceptor sites are those that score above the threshold determined by the Markov chains described previously.

The algorithm incorporates special cases for each of the four types t to prune the search space further. These are as follows:

1. If the interval $(n \ldots i)$ is the coding portion of an exon, its IMM score must exceed a fixed, preset threshold.
2. If two internal exons $(n \ldots i_1)$ and $(n \ldots i_2)$ both result in identical IMM scores, choose the one that maximizes the length of the coding part of the parse. Note that this rule makes GLIMMERM prefer longer gene models.
3. If $(n \ldots i)$ is an intron, then its AT content must be at least 70%. This constraint is based on the observation that all *P. falciparum* introns in the training set had an AT content of above 70%, with only 1% of introns having an AT content under 75%. In contrast, *P. falciparum* exons have an AT content of 70–75%.
4. The length of an intron must be between 50 and 1500 bp; 73 and 1066 bp were the extreme lengths for the introns in the training set.
5. The total length of the coding portions of a gene model represented in *Parse* [start, n, S] must be greater than 200 bp.
6. If n is a stop codon, the algorithm searches backward for all gene models ending at n. Many stop codons can be quickly eliminated because they follow too closely another stop codon in the same reading frame. Thus there is no way to create a gene model ending at these stops—any genes ending at the stop would be too short. The high AT content of *P. falciparum* and the resulting high frequency of stop codons make this step particularly effective.

An attempt was made to use IMMs to score introns as well as exons, but this did not improve the results. Therefore, when t is a donor site and t_{next} is an acceptor, we have

$$Score(\langle \text{donor}, n\rangle, Parse[\text{acceptor}, i, S]) = Score(Parse[\text{acceptor}, i, S]).$$

The algorithm is run separately on both the direct and the complementary strands of the input. GLIMMERM then makes one more pass over the list of putative genes to reject overlapping genes. If genes overlap by less than a fixed amount (30 bp by default), then the overlap is ignored, and both genes are reported in the output. Most overlapping genes are competing gene models that share a stop codon and have different exon locations. Genes that overlap by more than 30 bp are rescored using the IMM, and the gene with the best score is retained. If the scores of two or more overlapping models differ from the maximum score by less than a small preset amount, then GLIMMERM considers the scores equivalent and outputs all the models as possible genes. In these instances, it marks the longest gene as the preferred model.

2.4. Code Availability

The complete GLIMMERM system is available from the authors; it has already been shared with other malaria genome sequencing centers. The code includes routines for retraining the system on data from other organisms. A version of the system trained on *A. thaliana* genes is currently under development. Total processing time to find all genes in malaria chromosome 2 (approximately one million nucleotides) is about 50 min on a Pentium 450 processor running Linux.

2.5. Annotating a Genome

In its current form, GLIMMERM produces multiple gene models for some genes. When no database matches and no other computational evidence were found to support a GLIMMERM prediction, the chromosome 2 annotation reflects the highest scoring model. Although many of these are likely to be correct, it is undoubtedly the case that some are not. Further investigation is required to confirm these predictions (but see below for laboratory evidence confirming a small subset).

The GLIMMERM algorithm was used as one of a suite of tools. Accurate gene identification depends on using every tool available, and the description here should not be taken as implying that GLIMMERM alone can find all genes in *P. falciparum* or any other genome. However, it was a central component in a larger strategy. Other important computational tools used by the malaria chromosome 2 team were as follows: (1) searches of a nonredundant protein sequence database using gapped BLAST and PSI-BLAST (Altschul *et al.,* 1990, 1997); (2) gapped alignments of DNA to protein and EST sequence databases using DDS and DPS (Huang *et al.,* 1997); (3) prediction of putative signal peptides using SignalP (Nielsen *et al.,* 1997); (4) prediction of transmembrane domains with PHThtm (Rost *et al.,* 1995); (5) prediction of nonglobular structures with SEG (Wootton and Federhen, 1996); and (6) a graphical tool to allow annotators to view all the evidence together. In addition, the project used additional aligment tools developed at The Institute for Genomic Research to detect frameshift errors: these tools allow an annotator to detect when a sequence alignment extends beyond the start and stop codons indicated by other tools. In some cases this indicates errors in sequencing, which can be corrected; in other cases it indicates either a genuine frameshift that occurs during translation or a mutation that has changed the length of the translated protein. Any comprehensive annotation effort needs these computational tools and more to produce reasonably accurate gene annotations.

3. RESULTS AND DISCUSSION

GLIMMERM was used as the primary gene finder for chromosome 2 of *P. falciparum.* Chromosome 2 has 209 protein-coding genes spread over approximately one million bases, for a gene density of one gene per 4.5 kb (1/4.5 kb). This contrasts with a density of 1/kb in bacteria, 1/2 kb in yeast, 1/7 kb in *C. elegans,* and 1/50 kb (estimated) in human. Of the 209 protein-coding genes, 43% had at least one intron, and those genes with introns usually had just one or two introns (Gardner *et al.,* 1998). Below we attempt to quantify GLIMMERM's accuracy on these genes.

3.1. Training

To train the IMM, we needed to collect as much coding sequence as possible from *P. falciparum* itself. We exhaustively surveyed the literature to collect every complete sequence that was backed by laboratory evidence. Our survey collected 119 complete coding sequences from 108 GenBank entries representing all 14 chromosomes, of which just 6 genes came from chromosome 2. (This database is available by e-mail upon request from the authors.) Note that by length, chromosome 2 comprises approximately 3% of the genome, so it is unsurprising that just 6/119 genes were from chromosome 2. GenBank contains more than 108 entries from *P. falciparum,* but other entries do not have clear evidence supporting their splice sites. This training set provided the initial data for the splice site models as well.

An important point to emphasize here is that *P.*

TABLE 1

Performance of GLIMMERM on Genes Whose Structure Is Completely Known from Independent Laboratory Evidence

Name	Len	Intr	Comment	Common name
PFB0100c	654	1	Perfect match	Knob-associated His-rich prt
PFB0295w	471	0	Perfect match	Adenylosuccinate lyase (OO)
PFB0300c	272	0	Perfect match	Merozoite surface antigen MSP-2
PFB0305c	272	1	Perfect match	Merozoite surface antigen MSP-5 (EGF domain)
PFB0310c	272	1	Perfect match, highest score from 5 models	Merozoite surface antigen MSP-4 (EGF domain)
PFB0340c	997	3	Perfect match, second highest score from 4 models	SERA antigen/papain-like Protease with active Ser
PFB0405w	3135	0	Perfect match, higher score from 2 models	Transmission blocking Target antigen PfS230

Note. All seven genes had perfect matches to the system's predictions, meaning that the start codon, stop codon, and every splice site were correctly predicted. The column headings give the gene name, its length in amino acids, number of introns (Intr), a comment on GLIMMERM's prediction, and the common name of the protein.

falciparum has an unusually high 82% AT content. As a consequence of this high AT content, stop codons are very frequent (e.g., TAA will occur especially often) in noncoding DNA. This makes it much more likely that long open reading frames (ORFs) represent coding sequence. This fact was used to generate additional training data for GLIMMERM: ORFs greater than 500 bp in the chromosome 2 sequence were assumed to be coding regions and were used in the IMM training. These were added to the list generated by the literature search.

3.2. Accuracy on Known Genes

The 209 genes included in the chromosome 2 annotation were found with GLIMMERM's help. To evaluate the accuracy of the system, it is helpful to consider only those genes from this set for which independent evidence can be found to confirm their existence.

The best way to measure the program's accuracy is to consider its accuracy on those proteins whose exon–intron structure is known precisely from laboratory studies. There are seven genes from chromosome 2 of *P. falciparum* that currently fit into this category; i.e., the sequence from start to stop has been completely characterized. Of these seven, six were included in the training set, and one (PFB0100c) was not.

GLIMMERM's performance on this small set of genes is shown in Table 1. For the two-exon gene PFB0100c, the only independently confirmed gene that was not included in the training set, the system predicted only one model: the correct one. For all seven of the genes, GLIMMERM's output contained a model that matched perfectly. For four of the genes, the correct model was the only one output by the system. For PFB0310c and PFB0405c, GLIMMERM produced five and two competing models, respectively, but in each case the highest scoring one was correct. Only for PFB0340c, a four-exon gene, was GLIMMERM's correct model not the highest scoring one. The system gave a slightly higher score to a model that used a different donor site for the first exon. GLIMMERM's alternate prediction would have a 23-aa insertion in this 997-aa protein.

3.3. Laboratory Tests

An ideal way of measuring the accuracy of GLIMMERM precisely would be to test each of its predictions in the laboratory to see whether they are expressed as predicted. Although a complete test of all predictions would be difficult and time-consuming, one careful set of experiments was conducted as part of the chromosome 2 study.

Because many of the proteins predicted by GLIMMERM had unusual nonglobular domains, the chromosome 2 project team ran a reverse transcriptase (RT-PCR) experiment for 13 of these genes (Gardner *et al.*, 1998) to determine whether or not they were real. These genes are shown in Table 2. The RT-PCR focused its attention on nonglobular domains, not entire proteins, so it could not confirm every detail of the GLIMMERM predictions. In particular, it did not test the exon–intron boundaries for the two genes in this set

TABLE 2

The Set of Genes with Nonglobular Domains for Which RT-PCR Experiments Were Conducted to Confirm Expression

Name	Length	Intr	Common name
PFB0130w	538	0	Prenyl transferase
PFB0145c	1979	0	Hypothetical protein
PFB0180w	560	1	prt with 5′-3′ exonuclease domain
PFB0265c	1516	0	RAD2 endonuclease
PFB0380c	2010	0	Phosphatase (acid phosphatase family)
PFB0435c	1138	7	Predicted amine transporter
PFB0500c	235	0	RAB GTPase
PFB0520w	1233	0	Novel protein kinase
PFB0525w	610	0	Asparaginyl-tRNA synthetase
PFB0685c	885	0	ATP-dependent acyl-CoA synthetase
PFB0720c	899	0	Ori. recognition complex subunit 5 (ATPase)
PFB0755w	1398	0	Hypothetical protein
PFB0880w	426	0	FAD-dependent oxidoreductase

Note. Length is shown in amino acids, and Intr gives the number of introns. In the two genes containing introns, the nonglobular domains are contained within exons.

that contain introns, because the nonglobular domains in those genes do not cross those boundaries. This experiment confirmed that all 13 of the nonglobular domains are expressed; i.e., the predictions for those regions were correct. To our knowledge, this is the first time ever that computational gene predictions provided the impetus for experiments that in turn confirmed the predictions.

Eleven of these 13 genes have sequence homology to known proteins from other organisms. It is worth noting that the nonglobular domains of the *P. falciparum* proteins did *not* occur in the homologs. For example, PFB0180c contains a 176-amino-acid nonglobular insert that is absent from four homologous bacterial exonuclease domains (shown in Fig. 2 of Gardner *et al.,* (1998)). GLIMMERM's prediction for this gene was confirmed by amplifying and then sequencing a region that contained the nonglobular domain. This example points out that the presence of a homologous protein sequence does not always produce an accurate gene prediction.

3.4. Comparison on Genes with Homologs

Of the 209 genes in chromosome 2, 119 have homologous proteins in the public sequence databases. (The training set also contained 119 genes, but the identity of these two numbers is merely coincidence.) The existence of homologs, which come from a wide range of other organisms, provides strong independent evidence that these genes are real. We therefore used these genes to make further measurements of GLIMMERM's accuracy.

Of the 119 genes, 7 were already mentioned: these are the genes from chromosome 2 whose exon–intron structure was known from previously published laboratory studies. Six of those were included in the training set, which leaves 113 genes in chromosome 2 that were *not* included in the training set and for which we have good hints of their exon–intron structure. Because these are homologs, parts of some genes may not align well, making the predicted exon–intron structure less certain.

GLIMMERM finds 98 of these 113 genes (87%) exactly; i.e., the positions of the start codon, the boundaries of each exon and intron, and the stop codon correspond to what is indicated by the alignments to homologous genes. Of these, 22 have competing gene models that score higher, meaning that a human annotator had to examine the output and decide, based on the alignment, to use a model other than the highest-scoring one.

Of the 15 genes that GLIMMERM did not find exactly, 14 were found but had slightly modified coding regions. Seven intronless genes were predicted with incorrect start codons. Three 2-exon genes were broken into two genes each. Four 3-exon genes were predicted with an incorrect first exon but correct second and third exons.

Only one of the genes with homologs, ribosomal protein S30, was missed completely; ribosomal proteins often have a strikingly different composition from other genes and are known to be difficult for content-based gene finders to locate. These will not be missed as long as genomic data are searched against databases of known ribosomal proteins.

In summary, chromosome 2 contains 113 genes that were not included in the set of 119 genes used to train GLIMMERM's IMM. Portions of some of these genes, those with ORFs greater than 500 bp, were extracted automatically and added to the IMM; this portion of the training is fully automatic and requires no human intervention. The splice site training also included some data from chromosome 2, as explained above. A similar procedure can be performed on future chromosomes to extract additional splicing data: first use a sequence alignment program to find homologous genes, extract splice sites from those, and add those splice sites to the Markov chain models. This will allow users of the system to improve the system's performance before making a final run on their chromosomes. Assuming this or a similar protocol is followed, the estimates given here should extrapolate reliably to those chromosomes. Of the 113 genes with homologs, GLIMMERM is able to annotate automatically 76 (66%) if its top-scoring prediction is assumed correct. If a human annotator is available to confirm or reject predictions, then this number grows to 87% (98/113). In most cases the differences between competing models are small, involving one splice site or the start codon. Information from alignments or from other programs—for example, identification of signal peptides—allowed the human annotators to override GLIMMERM's first choice in selected cases.

3.5. Comparison to Chromosome 2 Annotation

Of the 209 genes currently annotated for chromosome 2, GLIMMERM finds 178 exactly. Of these, 40 have competing gene models that score higher; human annotators chose a different model for the final annotation. Of the remaining 31 genes, GLIMMERM finds the stop codons correctly for 14. Different starts appear in the final annotation for several reasons, for example, the existence of a match to a protein sequence that starts at a different start codon. (Note that it is possible that GLIMMERM is still correct in these cases.) The system finds the correct start but the wrong stop codon for 4 genes; this occurs in multiexon genes in which a splice site was missed and one of the exons was incorrectly extended until it hit a stop codon. The 11 remaining partial hits are cases for which GLIMMERM predicts some but not all exons correctly; for example, several multiexon genes are each broken into two separate genes.

Only 2 of the 209 genes are missed completely. One is ribosomal protein S30, which was mentioned above. The second is a predicted integral membrane protein of 192 aa predicted by a preliminary version

of GLIMMERM (before retraining the splice site models). A separate program was used to predict the function of this protein; it did not align to any known sequences.

The improved splice site Markov models resulted in GLIMMERM's generating 41 fewer gene models than before. In addition to the one missed gene just described, it generated 5 new gene models. Of these, one appears to encode a genuine protein, and we are currently investigating this to see if it should be added to the published annotation.

A significant caveat to include with these results is that GLIMMERM often produces multiple competing models that the human annotator must resolve. Most genes with three or more exons result in multiple models. The system indicates which model scores the highest, but as indicated above, 40 of the "correct" gene models had alternative parses that scored higher. These alternative parses share some exons but use different splice sites for others. A human annotator looking at additional evidence, such as alignments to homologous proteins or predictions of signal peptides, was able to overrule the system's top choice in these cases. It is likely that in other cases where no evidence besides GLIMMERM's prediction is available, some of the published annotation may still be in error (all such proteins are annotated as hypotheticals). After each set of multiple gene models was collapsed into one model, the gene list still contains 266 genes. (All of the models can be downloaded on the Web at www.tigr.org/~salzberg/GlimmerMchr2output.html.) These means that, since only 209 genes appeared in the final annotation, the annotators eliminated another 57 gene models entirely from the output. These decisions were somewhat subjective: frequently the putative genes were short or they consisted mostly of low-complexity sequence, and this was not enough to convince the human annotators that the genes were real. In many cases the annotators are probably correct, but it is simply impossible at this point to say with confidence that all of the deleted genes are false-positives. Only further evidence will allow us to decide, but this makes clear the importance of continuing to update and improve genome annotation over time.

ACKNOWLEDGMENTS

S.L.S. is supported by the National Human Genome Research Institute at NIH under Grant K01-HG00022-1. S.L.S., A.L.D., and M.P. are supported in part by the National Science Foundation under Grant IRI-9530462. M.J.G. and H.T. were supported by a supplement to NIAID Grant R01 AI40125-01, which was made possible with funds from NIH's Office for Research on Minority Health and Department of the Army Cooperative Agreement DAMD17-98-2-8005.

REFERENCES

Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410.

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. (1997). Gapped blast and psi-blast: A new generation of protein database search programs. *Nucleic Acids Res.* **25**(17), 3389–3402.

Borodovsky, M., and McIninch, J. (1993). Genemark: Parallel gene recognition for both DNA strands. *Comput. Chem.* **17**(2), 123–133.

Borodovsky, M., McIninch, J., Koonin, E., Rudd, K., Medigue, C., and Danchin, A. (1995). Detection of new genes in the bacterial genome using Markov models for three gene classes. *Nucleic Acids Res.* **23,** 3554–3562.

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268,** 78–94.

Fraser, C., Casjens, S., Huang, W., Sutton, G., Clayton, R., Lathigra, R., White, O., Ketchum, K., Dodson, R., Hickey, E., Gwinn, M., Dougherty, B., Tomb, J.-F., Fleischmann, R., Richardson, D., Peterson, J., Kerlavage, A., Quackenbush, Salzberg, S., Hanson, M., van R., Vugt, Palmer, N., Adams, M., Gocayne, J., Weidman, J., Utterback, T., Watthey, L., McDonald, L., Artiach, P., Bowman, C., Garland, S., Fujii, C., Cotton, M., Horst, K., Roberts, K., Hatch, B., Smith, H., and Venter, J. (1997). Genomic sequence of a lyme disease spirochaete, *Borrelia burgdorferi. Nature* **390**(6660), 580–586.

Fraser, C., Norris, S., Weinstock, G., White, O., Sutton, G., Clayton, R., Dodson, R., Gwinn, M., Hickey, E., Ketchum, K., Sodergren, E., Hardham, J., McLeod, M., Salzberg, S., Khalak, H., Weidman, J., Howell, J., Chidambaram, M., Utterback, T., Watthey, L., McDonald, L., Artiach, P., Bowman, C., Garland, S., Fujii, C., Cotton, M., Horst, K., Roberts, K., Hatch, B., Smith, H., and Venter, J. (1998). Complete genomic sequence of *Treponema pallidum,* the syphilis spirochete. *Science* **281,** 375–388.

Gardner, M., Tettelin, H., Carucci, D., Cummings, L., Aravind, L., Koonin, E., Shallom, S., Mason, T., Yu, K., Fujii, C., Pederson, J., Shen, K., Jing, J., Aston, C., Lai, Z., Schwartz, D., Pertea, M., Salzberg, S., Zhou, L., Sutton, G., Clayton, R., White, O., Smith, H., Fraser, C., Adams, M., Venter, J., and Hoffman, S. (1998). Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum. Science* **282,** 1126–1132.

Henderson, J., Salzberg, S., and Fasman, K. (1997). Finding genes in human DNA with a hidden Markov model. *J. Computat. Biol.* **4**(2), 127–141.

Huang, X., Adams, M., Zhou, H., and Kerlavage, A. (1997). A tool for analyzing and annotating genomic sequences. *Genomics* **46,** 37–45.

Kulp, D., Haussler, D., Reese, M. G., and Eeckman, F. H. (1996). A generalized hidden Markov model for the recognition of human genes in DNA. *In* "ISMB-96: Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology," pp. 134–141, AAAI Press. Menlo Park, CA.

Nelson, K., Clayton, R., Gill, S., Gwinn, M., Dodson, R., Haft, D., Hickey, E., Peterson, J., Nelson, W., Ketchum, K., McDonald, L., Utterback, T., Malek, J., Linher, K., Garrett, M., Stewart, A., Cotton, M., Pratt, M., Phillips, C., Richardson, D., Heidelberg, J., Sutton, G., Fleischmann, R., White, O., Salzberg, S., Smith, H., Venter, J., and Fraser, C. Genome sequence of *Thermotoga maritima:* Evidence for lateral gene transfer between archaea and bacteria. Submitted for publication.

Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **10**(1), 1–6.

Rost, B., Casadio, R., Fariselli, P., and Sander, C. (1995). Transmembrane helices predicted at 95% accuracy. *Protein Sci.* **4**(3), 521–533.

Salzberg, S. (1998). Decision trees and Markov chains for gene finding. *In* "Computational Methods in Molecular Biology" (S. Salzberg, D. Searls, and S. Kasif, Eds.), pp. 187–203, Elsevier, Amsterdam.

Salzberg, S., Chen, X., Henderson, J., and Fasman, K. (1996). Finding genes in DNA using decision trees and dynamic programming. *In* "ISMB-96: Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology" pp. 201–210, AAAI Press, Menlo Park, CA.

Salzberg, S., Delcher, A., Fasman, K., and Henderson, J. (1998a). A decision tree system for finding genes in DNA. *J. Computat. Biol.* **5**(4), 667–680.

Salzberg, S., Delcher, A., Kasif, S., and White, O. (1998b). Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.* **26**(2), 544–548.

Saul, A., and Battistutta, D. (1988). Codon usage in *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **27,** 35–42.

Snyder, E. E., and Stormo, G. D. (1995). Identification of coding regions in genomic DNA. *J. Mol. Biol.* **248,** 1–18.

Stephens, R., Kalman, S., Lammel, C., Fan, J., Marathe, R., Aravind, L., Mitchell, W., Olinger, L., Tatusov, R., Zhao, Q., Koonin, E., and Davis, R. (1998). Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. *Science* **282**(5389), 754–759.

Wirth, A. (1998). "A *Plasmodium falciparum* genefinder," Honours thesis, Department of Mathematics and Statistics, University of Melbourne.

Wootton, J., and Federhen, S. (1996). Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* **266,** 554–71.

Appendix H
Final Report
DAMD17-82-2-8005

A shotgun optical map of the entire *Plasmodium falciparum* genome

Zhongwu Lai[1], Junping Jing[1], Christopher Aston[1], Virginia Clarke[1], Jennifer Apodaca[1], Eileen T. Dimalanta[1], Daniel J. Carucci[3], Malcolm J. Gardner[4], Bud Mishra[2], Thomas S. Anantharaman[2], Salvatore Paxia[2], Stephen L. Hoffman[3], J. Craig Venter[4], Edward J. Huff[1] & David C. Schwartz[1,5]

The unicellular parasite *Plasmodium falciparum* is the cause of human malaria, resulting in 1.7–2.5 million deaths each year[1]. To develop new means to treat or prevent malaria, the Malaria Genome Consortium was formed to sequence and annotate the entire 24.6-Mb genome[2]. The plan, already underway, is to sequence libraries created from chromosomal DNA separated by pulsed-field gel electrophoresis (PFGE). The AT-rich genome of *P. falciparum* presents problems in terms of reliable library construction and the relative paucity of dense physical markers or extensive genetic resources. To deal with these problems, we reasoned that a high-resolution, ordered restriction map covering the entire genome could serve as a scaffold for the alignment and verification of sequence contigs developed by members of the consortium. Thus optical mapping was advanced to use simply extracted, unfractionated genomic DNA as its principal substrate. Ordered restriction maps (*Bam*HI and *Nhe*I) derived from single molecules were assembled into 14 deep contigs corresponding to the molecular karyotype determined by PFGE (ref. 3).

Optical mapping is now a proven means for the construction of accurate, ordered restriction maps from ensembles of individual DNA molecules derived from a variety of clone types, including bacterial artificial chromosomes[4] (BACs), yeast artificial chromosomes[5] (YACs) and small insert clones[6]. We previously developed approaches for mapping clone DNA samples that relied on the analysis of large numbers of identical DNA molecules. Here, the challenge was to develop ways to generate restriction maps of a population of randomly sheared DNA molecules directly extracted from cells that were obviously non-identical. Problems to be solved included the development of techniques for mounting very large DNA molecules onto surfaces and new methods for accurately mapping individual molecules, which were uniquely represented within a population. Finally, new algorithms were necessary to assemble such maps into gap-free contigs covering all 14 chromosomes of the *P. falciparum* genome.

We developed an optical mapping approach, termed shotgun optical mapping, that used large (250–3,000 kb), randomly sheared genomic DNA molecules as the substrate for map construction (Fig. 1*a–e*). Random fragmentation of genomic DNA occurred naturally as a consequence of careful pipetting and other manipulations. Surface-mounted molecules were digested using *Bam*HI and *Nhe*I (refs 6–8). Because genomic DNA molecules frequently extended through multiple digital image fields, we developed an automated image acquisition system (GenCol) to overlap digital images with proper registration (Figs 1*c* and 2). Map construction techniques were altered to take into account local restriction endonuclease efficiencies (the rate of partial

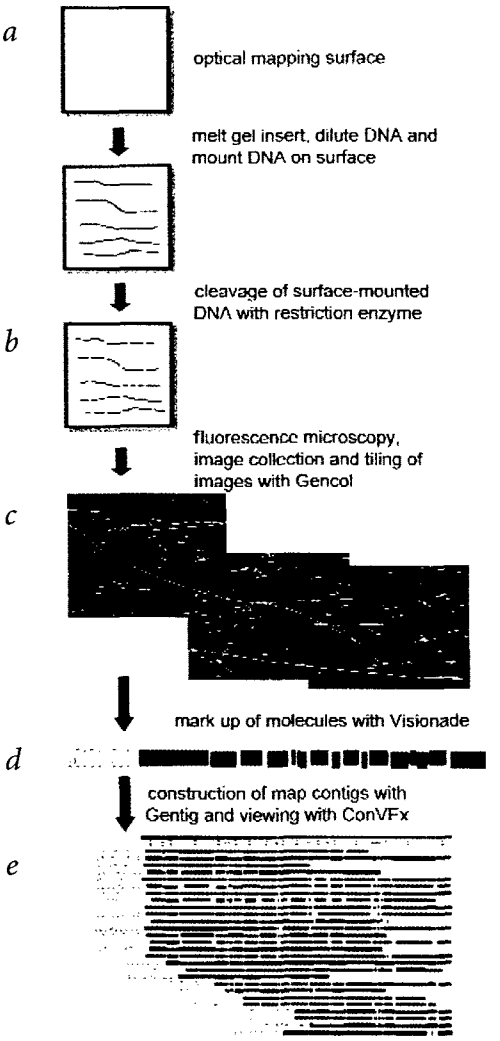


Fig. 1 Schematic of shotgun optical mapping approach. ***a***, Shotgun optical mapping used large (250–3,000 kb), randomly sheared genomic DNA molecules as the substrate for map construction. ***b***, Random fragmentation of genomic DNA occurred naturally as a consequence of careful pipetting and other manipulations. Surface-mounted molecules were digested using *Bam*HI and *Nhe*I (ref. 8). ***c***, Because genomic DNA molecules frequently extended through multiple image fields, an automated image acquisition system was developed (GenCol) and used to overlap images with proper registration. ***d***, Map construction techniques take into account local restriction endonuclease efficiencies (the rate of partial digestion) and the analysis of molecule populations that differed in composition and mass. ***e***, These steps were necessary to enable accurate construction of map contigs.

[1]*W.M. Keck Laboratory for Biomolecular Imaging, Department of Chemistry and* [2]*Courant Institute of Mathematical Sciences, Department of Computer Science, New York University, Department of Chemistry, New York, New York, USA.* [3]*Malaria Program, Naval Medical Research Center and* [4]*The Institute for Genomic Research, Rockville, Maryland, USA.* [5]*Present address: University of Wisconsin-Madison, Departments of Chemistry and Genetics, UW Biotechnology Center, Madison, Wisconsin, USA. Correspondence should be addressed to D.C.S. (e-mail: schwad01@mcrcr.med.nyu.edu).*

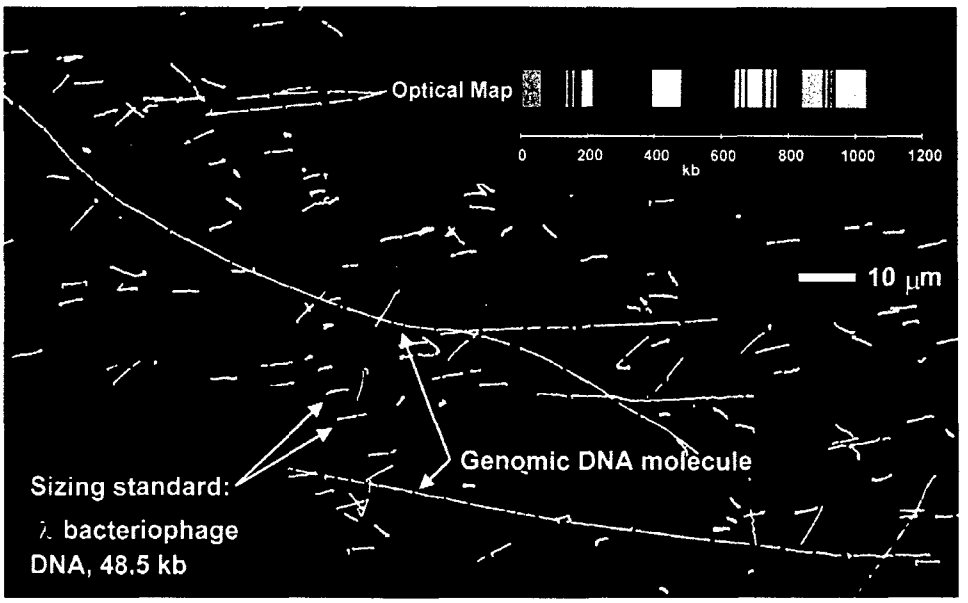


Fig. 2 Digital fluorescence micrograph and map of a typical genomic DNA molecule. A *P. falciparum* molecule digested with *Nhe*I is shown with its corresponding optical map. Comparison with the consensus optical map shows this molecule to be an intact chromosome 3. Image composed by tiling a series of 63× (objective power) images using GenCol. Co-mounted λ bacteriophage DNA is used as a sizing standard and to estimate cutting efficiencies.

digestion) and the analysis of molecule populations that differed in composition and mass. These steps were necessary to enable accurate construction of map contigs.

Previous map construction techniques using cloned DNA molecules[5,6,9] determined restriction-fragment mass on the basis of relative measures of integrated fluorescence intensities or apparent lengths. Thus, fragment masses were reported as a fraction of the total clone size (1.0), and later converted to kilobases by independent measure of clone masses (that is, cloning vector sequence[10]). Additionally, maps derived from ensembles of identical molecules were averaged to construct final maps. In contrast, here, we independently sized restriction fragments in genomic shotgun optical mapping using λ bacteriophage DNA that was co-mounted and digested in parallel (Fig. 2). These molecules were also used to locally monitor the restriction digestion efficiency, and to infer the extent of digestion on a per molecule (genomic) basis. Cutting efficiencies were in excess of 80%. This assessment provided a critical set of parameters for the contig assembly program, 'Gentig'[8,11,12], to reliably overlap maps derived from individual DNA molecules.

Gentig assembled maps into a number of deep contigs, but did not assign every single-molecule map to a contig. The program assembled contigs using 50% of the available molecules, which corresponded to 70% of the total mass of the molecules. In other words, the program was better able to construct contigs from the longer single-molecule maps. Finishing work using spreadsheets assembled the data into 14 contigs corresponding to the PFGE-generated molecular karyotype, with a total genome size of 24.16 Mb (Table 1). *Bam*HI and *Nhe*I maps had an average fragment size of 30.6 kb and 30.1 kb, respectively. We constructed consensus maps (Fig. 3) by simple averaging of aligned restriction-fragment masses (typically 6–26 fragments) derived from overlapping DNA molecules. Overall, chromosome sizes were largely consistent with PFGE results, with the total optical genome size being approximately 7% smaller, indicating that no previously uncharacterized nuclear component was found.

We previously constructed a high-resolution optical map of *P. falciparum* chromosome 2 (ref. 7). The starting material was a PFGE gel slice containing fractionated chromosome 2 DNA. We now constructed a whole-genome optical map using total, unfractionated genomic DNA as the starting material and resolved all 14 chromosomes, including electrophoretically unseparable ones (chromosomes 5–9, termed the 'blob'), at the level of data (optical map contigs) rather than as physical entities (that is, gel bands).

Table 1 • *P. falciparum* whole-genome optical mapping

Chr.	PFGE (Mb)	*Nhe*I (Mb)	*Bam*HI (Mb)	Ave. (Mb)	Diff. (Mb)	Linkage/ confirmation	Orientation
1	0.65/0.65*	0.684	0.668	0.676	0.016	1,3	+
2	1.0/0.947*	0.958	1.037	0.997	0.079	1,3	+
3	1.2/1.060*	1.084	1.096	1.090	0.012	1,2	+
4	1.4	1.311	1.306	1.309	0.005	1	
5	1.6	1.331	1.337	1.334	0.006		
6	1.6	1.395	1.373	1.384	0.022		
7	1.7	1.494	1.444	1.469	0.050		
8	1.7	1.495	1.504	1.499	0.009		
9	1.8	1.600	1.595	1.598	0.005		
10	2.1	1.808	1.688	1.748	0.120	1	
11	2.3	2.097	2.089	2.093	0.008	1	
12	2.4	2.478	2.361	2.419	0.117	1	
13	3.2	3.172	3.022	3.097	0.150	2	+
14	3.4	3.436	3.404	3.420	0.032	1,3	+
Total	26.05	24.341	23.974	24.157	0.367		

*Size from sequencing. Linkage/confirmation was obtained as follows: by mapping PFGE-purified chromosomal material (1); by mapping chromosome-specific YACs (2); or by sequence information (3). +, *Bam*HI and *Nhe*I maps have been oriented. Chr., chromosome; Ave., average size; Diff., difference between *Bam*HI and *Nhe*I maps.

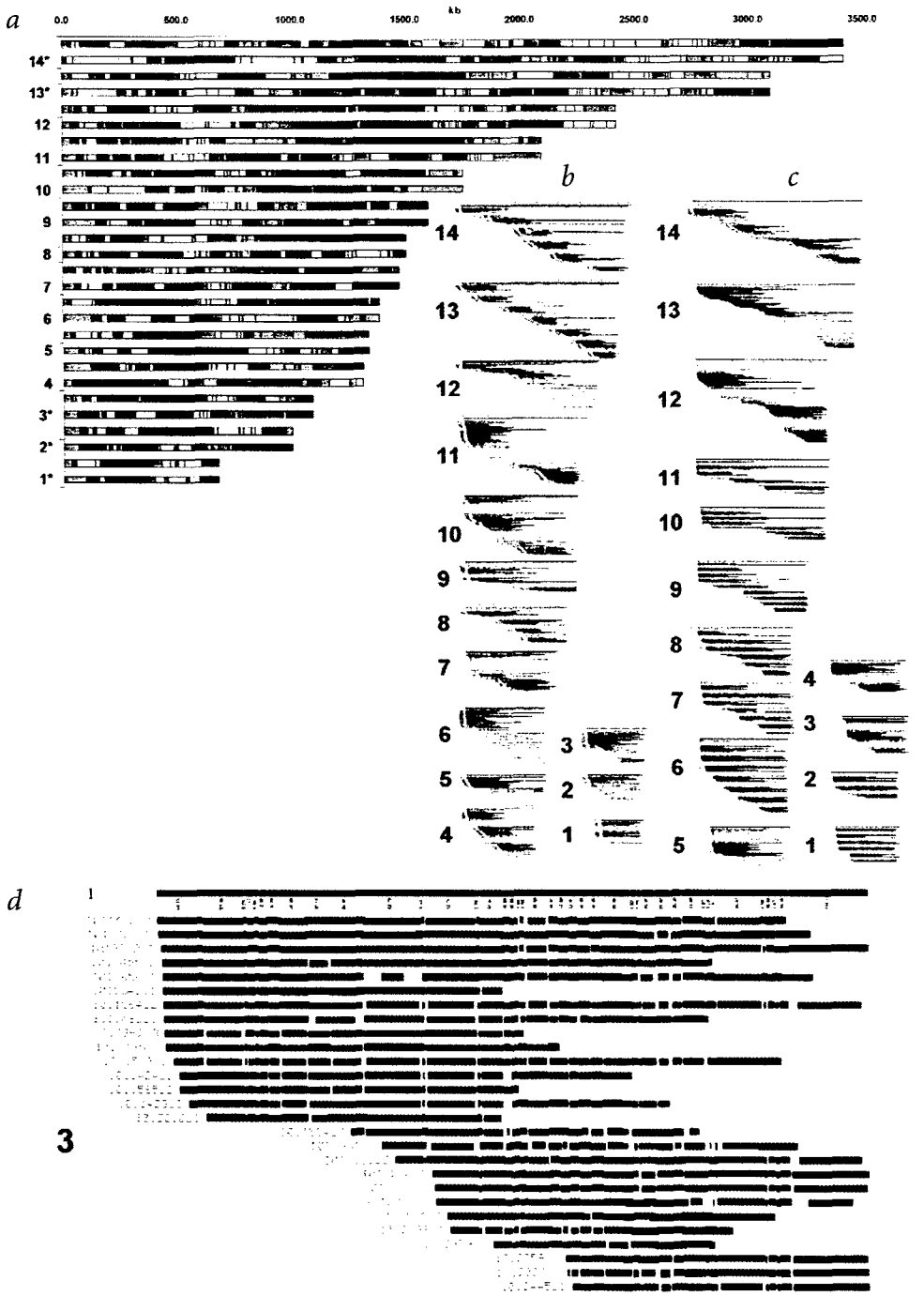


Fig. 3 High-resolution optical mapping of the *P. falciparum* genome using *Nhe*I and *Bam*HI. We mapped 944 molecules with *Nhe*I; the average molecule length was 588 kb, corresponding to 23× coverage. We mapped 1,116 molecules with *Bam*HI; the average molecule length was 666 kb, corresponding to 31× coverage. ***a***, Gap-free, consensus *Nhe*I and *Bam*HI maps were generated across all 14 *P. falciparum* chromosomes using the map contig assembly program Gentig. ***b***,***c***, *Nhe*I and *Bam*HI map alignments determined by Gentig, displayed by ConVEx. Fragment sizes of consensus maps (blue lines) shown in (a) were determined from the alignment and averaging of maps derived from 6–26 underlying individual molecules (green lines), 230–2,716 kb. ***d***, Enlargement of contig for chromosome 3 (*Nhe*I) shown in ConVEx displays maps (green) scaled to the consensus map (blue). These data can be accessed at http://carbon.biotech.wisc.edu/plasmodium. Bar lengths reflect measured fragment sizes. Fragments that overlap are shaded.

To assess errors produced by shotgun optical mapping, we compared optical restriction maps for chromosome 2 with restriction maps generated *in silico* using a previously assembled sequence[13]. We found good correspondence between the two maps. The sequence shows chromosome 2 to be 947 kb versus 958 kb by optical mapping with *Nhe*I and 1,037 kb with *Bam*HI. Only one 600-bp *Bam*HI fragment was missing in the entire genome optical map. The *Nhe*I optical map included all fragments above 400 bp predicted from sequence. The average absolute relative error in sizing fragments was 4.6% for *Nhe*I and 5.0% for *Bam*HI. Likewise, similar errors for chromosome 3 were determined by comparing optical maps with sequence data (*Nhe*I, 4.4%; *Bam*HI, 4.1%; total optical size versus sequence, *Nhe*I, 1,084 kb; *Bam*HI, 1,147 kb; versus 1,060 kb; D. Lawson, pers. comm.). These sizing errors were similar to those associated with PFGE.

Some large *Nhe*I and *Bam*HI fragments were noticeable at the telomeric ends. A telomere of one of the 'blob' chromosomes (chromosome 7) is composed of three consecutive 6-kb *Bam*HI fragments. Optical mapping can estimate numbers of repetitive regions if the repeats contain recognition sites for the endonuclease used. Subtelomeric regions in *P. falciparum*, however, are characterized by 21-bp tandem repeats[14], which are too small to be detected by optical mapping.

We used several approaches to verify and to link our optical maps with the PFGE molecular karyotypes, which number chromosomes according to mobility. Chromosomes that were identified and the orientations of *Bam*HI and *Nhe*I maps are shown

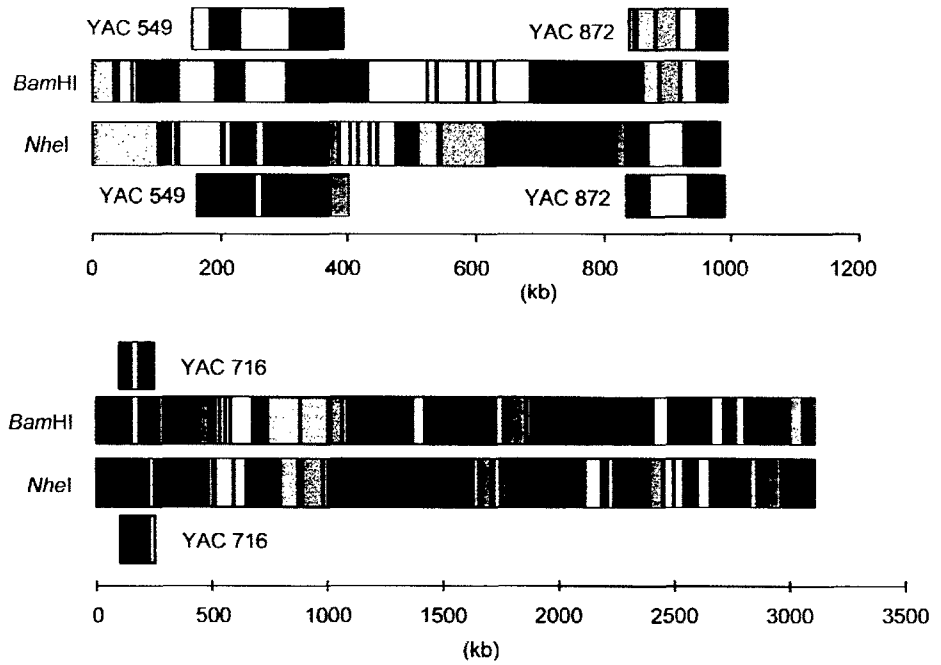


Fig. 4 Identification of chromosomes and alignment of *Nhe*I and *Bam*HI maps by mapping chromosome-specific YAC clones. Chromosome 3 and 13-specific YAC maps were aligned with the optical maps and the two enzyme maps were then oriented and linked. Each YAC is ~150 kb.

(Table 1). We confirmed chromosomal identities of some optical maps by optical mapping of PFGE-purified chromosomal DNA (ref. 7) with *Nhe*I or *Bam*HI. Here, most maps formed a contig, which aligned with a specific consensus map. Despite the fact that the largest and smallest *P. falciparum* chromosomes are resolved by PFGE, the gel slices contained DNA molecules from other chromosomes. There was, however, a sufficiently large population of molecules that formed a contig with a particular chromosome (>50%) to be able to identify it as being the chromosome predicted from the molecular karyotype. When many chromosomes are similar in size, such as chromosomes 5–9, there are many possible orientations of the maps, thus this approach was not viable. Chromosome-specific YAC clones were also optically mapped for further confirmation of chromosomal orientation and linkage. We aligned the resulting maps with a specific contig in the consensus maps (Fig. 4). YAC clones were not available for those chromosomes in the 'blob', so we were unable to identify or link these optical maps. As such, we have assigned numbers to these chromosomes according to their optically determined masses (Table 1). Maps can also be linked together by a series of double digestions, by the use of available sequence information or by Southern blot using chromosome-specific probes.

Because unicellular parasites have relatively small chromosomes that do not visibly condense, PFGE has provided a means by which chromosomal entities can be physically mapped and studied at the molecular level[15–17]. In fact, PFGE separations are currently providing the very material that the international malaria consortium is using to create chromosomal-specific libraries for large-scale sequencing efforts (http://www-ermm.cbcu.cam.ac.uk/dcn/txt001dcn.htm). Unfortunately, parasites such as *P. falciparum* can have karyotypically complex genomes, which confound PFGE analysis by displaying similarly sized chromosomes. Furthermore, very large or circular chromosomes are difficult to physically identify or characterize[18]. Although the shotgun sequencing of entire microorganism genomes[19,20] has obviated physical mapping to some extent, high-quality, finished sequence remains laborious to generate.

Many issues regarding the efficient sequencing of lower eukaryotes remain to be fully resolved, especially when available map resources are minimal. In the case of *Saccharomyces cerevisiae*, the entire genome was sequenced by a large consortium of laboratories on a per chromosome basis[21]. Their tasks were facilitated by the availability of extensive physical and genetic maps, plus an assortment of well-characterized libraries. These substantial genome resources provided ample means for the needed sequence verification efforts, and aids for the sequence-assembly process. In a similar, though much less distributive fashion, the *Caenorhabditis elegans* genome was recently completely sequenced[22]. Given the rapid pace of electrophoretic sequencing technology[23,24] and the accumulation of resources in sequence acquisition and analysis, new ways to efficiently sequence lower eukaryotes, particularly those implicated in human disease, must be developed to optimally leverage map resources created by optical mapping.

The optical maps presented here have been used by members of the consortium[13,25] as scaffolds to verify and facilitate sequence assemblies. In general, the maps were integrated into the sequence assembly process, in much the same way as any other physical maps. In particular, our maps have provided reliable landmarks for sequence assembly where traditional maps are somewhat sparse. Compared with sequence-tagged site (STS) or EST maps, in which landmark order is known but physical distance is approximate, optical restriction maps are constructed from landmarks (restriction sites) that are precisely characterized by physical distance. Another advantage is the speed of map construction: the maps presented here required only 4–6 months to generate. Given these and other advantages, future work will center on the algorithmic integration of high-resolution optical maps with primary sequence reads to more fully automate the sequence assembly and verification process. Finally, we plan to use optical mapping as the basis for developing of new ways to study genomic variations that fall between, or outside of, the capabilities of sequence-based approaches and cytogenetic observation.

Methods

Parasite preparation. We cultivated *P. falciparum* (clone 3D7) in erythrocytes using standard techniques[26]. Possible alterations of the genome that can occur in continuous culture[27] were minimized by keeping parasite aliquots frozen in liquid N_2 until needed. We then cultivated parasites only as long as necessary and prepared agarose-embedded parasites as described[7].

Mounting and digestion of DNA on optical mapping surfaces. We prepared derivatized glass optical mapping surfaces as described[7,28]. We diluted genomic DNA in TE buffer containing a sizing standard (λ bacteriophage DNA, 50 ng/ml), which was co-mounted with the genomic DNA by spreading the sample into the space between the surface and a microscope slide. DNA molecules were digested with *Nhe*I or *Bam*HI (ref. 8). λ bacteriophage DNA (48.5 kb; New England Biolabs) is cut once by *Nhe*I. λ DASH II bacteriophage DNA (41.9 kb; Stratagene) is cut twice by *Bam*HI. Therefore, we also used standards to identify regions on the surface where the digestion efficiency exceeded 70%. We stained DNA with YOYO-1 homodimer (Molecular Probes), before fluorescence microscopy. *P. falciparum* DNA has an AT content of 80–85%, and λ bacteriophage DNA has an AT content of 50%. The YOYO-1 fluorochrome used for DNA staining preferentially intercalates between GC pairs with increased emission quantum yield[29]. We therefore applied a correction factor to each fragment size to correct for this variation in fluorochrome incorporation.

Image acquisition, processing and map construction. We collected digital images of DNA molecules with a cooled charge coupled device (CCD) camera (Princeton Instruments) using Optical Map Maker (OMM) software as described[6]. Because genomic DNA molecules span multiple microscope image fields, we developed 'GenCol', an image acquisition and management software that was used to automatically collect and overlap consecutive CCD images with proper pixel registration. GenCol used a precise fitting routine, and the resulting 'super-images' covered the entire length of single DNA molecules, spanning several microscope fields. Restriction fragments were marked up with 'Visionade'[28], a semi-automatic visualization/editing program, which was run on super-images. Files created from marked-up images of molecules were then sent to map construction software, which automatically determined the restriction fragment masses, characterized internal DNA standard molecules and produced finished maps from single genomic molecules. The integrated fluorescence intensities of λ bacteriophage DNA standards, co-mounted with the genomic molecules, were used to measure the size of the *P. falciparum* restriction fragments on a per image basis. Cutting efficiencies (on a per image basis) were determined from scoring cut sites on sizing standard molecules contained in the same field as the genomic DNA molecules. Knowledge of endonuclease cutting efficiencies was critical for accurate contig construction.

Contig assembly by Gentig. Sophisticated statistical methods are used to overcome errors associated with partial digestion and mass determination[11,12]. Gentig finds overlapped molecules and assembles them into contigs. It computes contigs of genomic maps using a heuristic algorithm for finding the best scoring set of contigs (overlapping maps), because finding the optimal placement is in general computationally too expensive. The entire *P. falciparum* genome data set can be assembled into contigs in ~20 min. Gentig assembled consensus maps for each chromosome by averaging the fragment sizes from the individual maps underlying the contigs.

Contig viewing and editing by 'ConVEx'. We viewed contigs using 'ConVEx' (contig visualization and exploration tool). ConVEx is a multi-scale zoomable interface for visualization and exploration of large, high-resolution contiged restriction maps. Users can examine the consensus maps together with the raw uncorrected data. ConVEx also has a 'lens' mechanism that provides annotation and editing features, allowing communication of features such as STS markers, and even the underlying sequence reads.

Chromosome isolation by PFGE. The genome of *P. falciparum* is ~25 Mb, consisting of 14 chromosomes ranging from 0.6 to 3.5 Mb (ref. 28). PFGE resolves most of the *P. falciparum* chromosomes, except 5–9, which are of similar sizes and co-migrate. PFGE-purified chromosomal DNA was prepared as described[8] and used as a substrate for optical mapping.

YAC isolation and mapping. We cultured yeast cells in AHC media and prepared agarose-embedded cells using standard methods[3]. We purified YAC DNA using PFGE (POE apparatus, 1% gel in 0.5×TBE, pulse time 3 s, 5 s; switch time 32 s; 150 volts for 24 h; ref. 30). Optical maps of YAC clones were prepared with *Nhe*I and *Bam*HI as described above.

Acknowledgements

We thank D. Lawson and T. Wellems for clones and other valuable reagents. This work was supported by the Burroughs Wellcome Fund, NIH, and the Naval Medical Research and Development Command work unit STEP C611102A0101BCX. Additional support came from NCHGR (2 RO1 HG00225-01-09) and NCI1(RO1CA 79063-1).

Received 26 May; accepted 20 September 1999

1. World Health Organization. World malaria situation in 1994. Part I. Population at risk. *Wkly Epidemiol. Rec.* **72**, 269–274 (1997).
2. Wirth, D. Malaria: a 21st century solution for an ancient disease. *Nature Med.* **4**, 1360–1362 (1998).
3. Schwartz, D.C. & Cantor, C.R. Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. *Cell* **37**, 67–75 (1984).
4. Cai, W., Aburatani, H., Housman, D., Wang, Y. & Schwartz, D.C. Ordered restriction endonuclease maps of yeast artificial chromosomes created by optical mapping on surfaces. *Proc. Natl Acad. Sci. USA* **92**, 5164–5168 (1995).
5. Cai, W. *et al.* High resolution restriction maps of bacterial artificial chromosomes constructed by optical mapping. *Proc. Natl Acad. Sci. USA* **95**, 3390–3395 (1998).
6. Jing, J. *et al.* Automated high resolution optical mapping using arrayed, fluid fixed, DNA molecules. *Proc. Natl Acad. Sci. USA* **95**, 8046–8051 (1998).
7. Jing, J. *et al.* Optical mapping of *Plasmodium falciparum* chromosome 2. *Genome Res.* **9**, 175–181 (1999).
8. Lin, J. *et al.* Whole genome shotgun optical mapping of *Deinococcus radiodurans*. *Science* **285**, 1558–1562 (1999).
9. Schwartz, D.C. & Samad, A. Optical mapping approaches to molecular genomics. *Curr. Opin. Biotechnol.* **8**, 70–74 (1997).
10. Meng, X., Benson, K., Chada, K., Huff, E.J. & Schwartz, D.C. Optical mapping of lambda bacteriophage clones using restriction endonucleases. *Nature Genet.* **9**, 432–438 (1995).
11. Anantharaman, T.S., Mishra, B. & Schwartz, D.C. Genomics via Optical Mapping III: contiging genomic DNA and variations. in *Courant Technical Report 760* (Courant Institute, New York University, New York, 1998).
12. Anantharaman, T.S., Mishra, B. & Schwartz, D.C. Genomics via Optical Mapping III: contiging genomic DNA and variations. *The Seventh International Conference on Intelligent Systems for Molecular Biology* **7**, 18–27 (1999).
13. Gardner, M.J. *et al.* Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science* **282**, 1126–1132 (1998).
14. Pace, T., Ponzi, M., Scotti, R. & Frontali, C. Structure and superstructure of *Plasmodium falciparum* subtelomeric regions. *Mol. Biochem. Parasitol.* **69**, 257–268 (1995).
15. Van der Ploeg, L.H.T., Schwartz, D.C., Cantor, C.R. & Borst, P. Antigenic variation in *Trypanosoma brucei* analyzed by electrophoretic separation of chromosome sized DNA molecules. *Cell* **37**, 77–84 (1984).
16. Spithill, T.W. & Samaras, N. The molecular karyotype of *Leishmania major* and mapping of α and β tubulin gene families to multiple unlinked chromosomal loci. *Nucleic Acid Res.* **13**, 4155–4169 (1985).
17. Ahamada, S., Wery, M. & Hamers, R. Rodent malaria parasites: molecular karyotypes characterize species, subspecies and lines. *Parasite* **1**, 31–38 (1994).
18. Moritz, K.B. & Roth, G.E. Complexity of germline and somatic DNA in *Ascaris*. *Nature* **259**, 55–57 (1976).
19. Fleischmann, R.D. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512 (1995).
20. Fraser, C.M. *et al.* The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**, 397–403 (1995).
21. Goffeau, A. *et al.* Life with 6000 genes. *Science* **274**, 546, 563–567 (1996).
22. The *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012–2018 (1998).
23. Mullikin, J.C. & McMurray, A.A. Sequencing the genome, fast. *Science* **283**, 1867–1868 (1999).
24. Venter, J.C. *et al.* Shotgun sequencing of the human genome. *Science* **280**, 1540–1542 (1998).
25. Bowman, S. *et al.* The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*. *Nature* **400**, 532–538 (1999).
26. Trager, W. & Jensen, J.B. Human malaria parasites in continuous culture. *Science* **193**, 673–675 (1976).
27. Corcoran, L.M., Forsyth, K.P., Bianco, A.E., Brown, G.V. & Kemp, D.J. Chromosome size polymorphism in *Plasmodium falciparum* can involve deletions and are frequent in nature parasite populations. *Cell* **44**, 87–95 (1986).
28. Aston, C., Hiort, C. & Schwartz, D.C. Optical mapping: an approach for fine mapping. *Methods Enzymol.* **303**, 55–73 (1999).
29. Netzel, T.L., Nafisi, K., Zhao, M., Lenhard, J.R. & Johnson, I. Base-content dependence of emission enhancements, quantum yields, and lifetimes for cyanine dyes bound to double-strand DNA: photophysical properties of monomeric and bichromophoric DNA stains. *J. Phys. Chem.* **99**, 17936–17947 (1995).
30. Schwartz, D.C. & Koval, M. Conformational dynamics of individual DNA molecules during gel electrophoresis. *Nature* **338**, 520–522 (1989).

Appendix I
Final Report
DAMD17-82-2-8005

Optical Mapping of *Plasmodium falciparum* Chromosome 2

Junping Jing,[1] Zhongwu Lai,[1] Christopher Aston,[1] Jieyi Lin,[1] Daniel J. Carucci,[2] Malcolm J. Gardner,[3] Bud Mishra,[4] Thomas S. Anantharaman,[4] Hervé Tettelin,[3] Leda M. Cummings,[3] Stephen L. Hoffman,[2] J. Craig Venter,[3] and David C. Schwartz[1,5]

[1]*W.M. Keck Laboratory for Biomolecular Imaging, New York University, Department of Chemistry, New York, New York 10003 USA;* [2]*Malaria Program, Naval Medical Research Institute, Rockville, Maryland 20852 USA;* [3]*The Institute for Genomic Research, Rockville, Maryland 20850 USA;* [4]*Courant Institute of Mathematical Sciences, New York University, Department of Computer Science, New York, New York 10012 USA*

Detailed restriction maps of microbial genomes are a valuable resource in genome sequencing studies but are toilsome to construct by contig construction of maps derived from cloned DNA. Analysis of genomic DNA enables large stretches of the genome to be mapped and circumvents library construction and associated cloning artifacts. We used pulsed-field gel electrophoresis purified *Plasmodium falciparum* chromosome 2 DNA as the starting material for optical mapping, a system for making ordered restriction maps from ensembles of individual DNA molecules. DNA molecules were bound to derivatized glass surfaces, cleaved with *Nhe*I or *Bam*HI, and imaged by digital fluorescence microscopy. Large pieces of the chromosome containing ordered DNA restriction fragments were mapped. Maps were assembled from 50 molecules producing an average contig depth of 15 molecules and high-resolution restriction maps covering the entire chromosome. Chromosome 2 was found to be 976 kb by optical mapping with *Nhe*I, and 946 kb with *Bam*HI, which compares closely to the published size of 947 kb from large-scale sequencing. The maps were used to further verify assemblies from the plasmid library used for sequencing. Maps generated in silico from the sequence data were compared to the optical mapping data, and good correspondence was found. Such high-resolution restriction maps may become an indispensable resource for large-scale genome sequencing projects.

Optical mapping is a system for the construction of ordered restriction maps from single molecules (Schwartz et al. 1993; Anantharaman et al. 1997). Individual DNA molecules bound to derivatized glass surfaces and cleaved with restriction enzymes are imaged by digital fluorescence microscopy. Resulting cut sites are visualized as gaps between cleaved DNA fragments, which retain their original order (Cai et al. 1995, 1998). Optical mapping has been used to prepare maps of a number of large insert clone types such as bacterial artificial chromosomes (Cai et al. 1998) and most recently genomic DNA (J. Lin, R. Qi, C. Aston, J. Jing, T.S. Anantharam, B. Mishra, D. White, J.C. Venter, and D.C. Schwartz, in prep). A shotgun mapping strategy was developed in parallel for several microorganisms using large fragments of randomly sheared DNA that were mapped with high cutting efficiencies. The numerous overlapping restriction site landmarks and a measurable cutting efficiency combined together to enable accurate contig assembly without the use of cloned DNA (Anathraman et al. 1998). Because library construction was obviated, it was possible to map large *Plasmodium falciparum* (*P. falciparum*) DNA fragments, which are AT-rich and notoriously difficult to clone because of deletion and rearrangement in *Escherchia coli* (Gardner et al. 1998). Because cloning artifacts were precluded, this enabled accurate maps to be generated. Furthermore, small amounts of starting material were used, facilitating the mapping of this and potentially other parasites that are problematic to culture or clone.

Sequencing of chromosome 2 of *P. falciparum* was completed recently by Gardner and colleagues (Gardner et al. 1998), as part of an international consortium sequencing the whole *P. falciparum* genome (Foster and Thompson 1995; Dame et al. 1996). Existing physical maps of *P. falciparum* chromosomes [chromosome 3; (Thompson and Cowman 1997) and chromosome 4 (Sinnis and Wellems 1988; Watanabe and Inselberg, 1994)], prepared by restriction digestion, gel fingerprinting, and hybridization of probes are of moderate resolution and not ideally suited for systematic sequence verification. To assess the feasibility of optically mapping a whole eukaryotic chromosome, we constructed high-resolution, ordered restriction maps of *P. falciparum* chromosome 2 using genomic DNA and later compared these maps with those generated in

[5]Corresponding author.
E-MAIL schwad01@mcrcr.med.nyu.edu; FAX (212) 995-8487.

9:175–181 ©1999 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/99 $5.00; www.genome.org

silico from the sequence data. Such restriction maps reveal the architecture of large spans of the genome and have value in shotgun sequencing efforts because they provide ideal scaffolds for sequence assembly, finishing, and verification. Gaps that form between contigs can be characterized in terms of location and breadth, thereby facilitating closure techniques.

RESULTS

P. falciparum Chromosome 2 DNA Sample

A chromosome 2 gel slice was used as starting material. Despite the AT-rich nature of the *P. falciparum* genome (80–85%), melting of low-gelling-temperature agarose inserts did not affect the integrity of the DNA and the chromosomal DNA was competent for optical mapping. Previously, we mounted DNA molecules by sandwiching the sample between an optical mapping surface and a microscope slide, followed by peeling the surface from the slide. DNA molecules were stretched and fixed onto the surface. This method works very well with clone types such as bacteriophage, cosmid, and BAC (Cai et al. 1995, 1998); however, larger genomic DNA molecules tend to form crossed molecules. We improved this approach by adding the sample to the edge formed by the placement of a surface onto a slide. The liquid DNA sample spreads into the space between the surface and the slide by capillary action. Consequently, DNA breakage was minimized, molecules tended to elongate in the same direction, and crossed molecules were also minimized (see Fig. 1).

*Nhe*I and *Bam*HI Maps for *P. falciparum* Chromosome 2

The genomic DNA was mapped with either *Nhe*I (Fig. 1A) or *Bam*HI (Fig. 1B). Fragment sizes were calculated by comparison with comounted λ bacteriophage DNA (48.5 kb). *P. falciparum* DNA has an AT content of 80–85% and λ bacteriophage DNA has an AT content of 50%. The YOYO-1 fluorochrome used for DNA staining intercalates preferentially between GC pairs with increased emission quantum yield (Netzel et al. 1995). A correction factor was therefore applied to each fragment size to correct for this massively different fluorochrome incorporation. λ bacteriophage DNA was used also to determine areas on the surface where cutting efficiency was highest. Cutting efficiencies were > 80%. Maps were obtained from individual molecules of ~350 kb. Consensus maps were assembled from 50 molecules generating an average contig depth of 15 molecules. Chromosome 2 was found to be 976 kb by optical mapping with *Nhe*I, and 946 kb by optical mapping with *Bam*HI (average size 961 kb). There were 40 fragments in the *Nhe*I map, ranging from 1.5–115 kb, with average fragment size 24 kb (Fig. 2). There were 30 fragments in the *Bam*HI map ranging from 0.5–80 kb, with average fragment size 32 kb (Fig. 2). Each fragment size in the consensus map was averaged from 10 to 15 fragments. Although *P. falciparum* chromosome 2 migrates as a distinct band by PFGE, we found the gel slice to contain only 60% chromosome 2-specific DNA. The remaining optical mapping data was rejected.

Integration of Optical Maps and Sequence Data

The chromosome 2 sequence assembled by Gardner and colleagues shows chromosome 2 to be 947 kb (Gardner et al. 1998) versus 976 kb by optical mapping with *Nhe*I and 946 kb with *Bam*HI. The optical restriction maps were compared to restriction maps predicted from the sequence, and there was very good correspondence between the two, indicating that there were no major rearrangements or errors in the assembled sequence (Table 1). The optical map included all fragments above 500 bp predicted from sequence. The overall agreement between these maps and the sequence was therefore excellent, with the average fragment size difference below 600 bp (relative error 4.3%) for the *Nhe*I map. The average fragment size difference for the *Bam*HI map was 1.2 kb (relative error 5.8%). However, there were several notable differences. Large differences in size for the fragments at each end of the chromosome were noted (Tables 1 and 2). This is because the sequence for these subtelomeric regions is still under construction. PCR products spanning subtelomeric gaps are being sequenced currently. The optical map sizes were larger than those predicted from sequence for certain other fragments (Tables 1 and 2). These differences were due to large fluorescence inten-

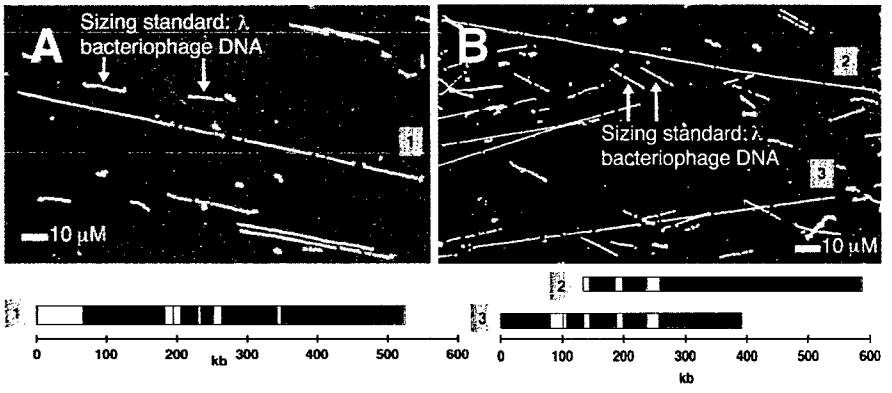


Figure 1 Typical *P. falciparum* chromosome 2 molecules and their corresponding optical maps. (*A*) digested with *Nhe*I (*B*) digested with *Bam*HI. Maps derived from the two *Bam*HI-digested molecules in (*B*) can be aligned.

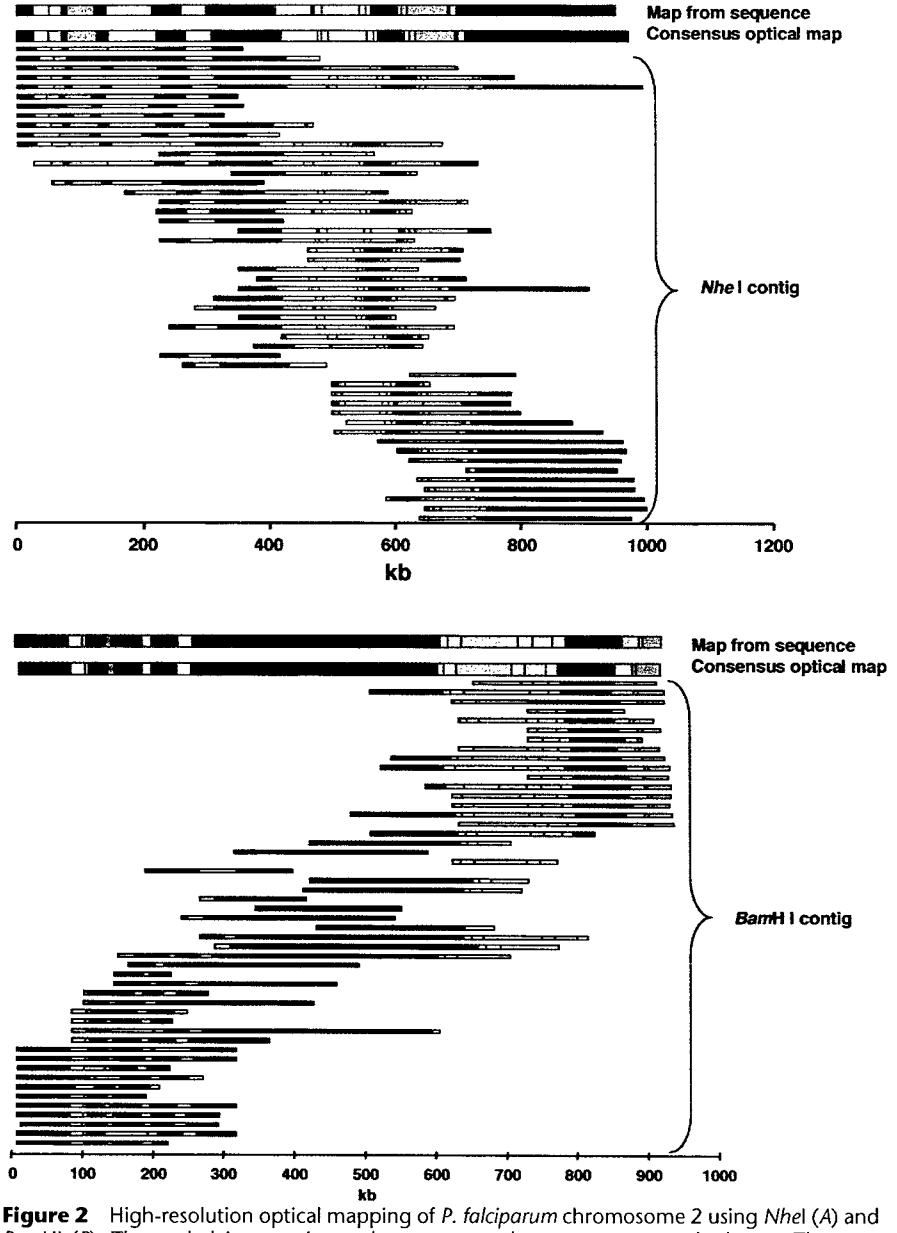


Figure 2 High-resolution optical mapping of *P. falciparum* chromosome 2 using *Nhe*I (*A*) and *Bam*HI (*B*). The underlying contig used to generate the consensus map is shown. The map predicted from sequence information is shown for comparison.

sity measurements falsely caused by crossed molecules. Currently, we combine length measurements with fluorescence intensity measurements to improve on our sizing of these fragments. Chromosome 2 maps using these new measurements show no exceptional errors (not shown; Jing et al., in prep). The map was used to facilitate sequence verification. Optical maps can also be used at the earlier sequence-assembly stage to form a scaffold for assembly of contigs formed from sequencing. Linking of single-enzyme maps produces a much higher resolution multi-enzyme map that is rich in information. Smaller contigs can be placed confidently on a multi-enzyme map. Nowadays, mapping is rarely done in the absence of sequencing. Figure 3 shows a comparison of a multi-enzyme map generated by optical mapping with that predicted from sequence. The maps are in complete agreement across the whole length of the chromosome. Given even small amounts of sequence (~100 kb), maps can be linked and verified readily.

Map Confirmation by Southern Blotting

To confirm the optical maps independently of sequence data, pulsed-field gels of total *P. falciparum* DNA digested with *Nhe*I or *Bam*HI were run and blotted. Plasmid clones used as sequencing templates provided the probes to analyze the Southern blots. Restriction fragment sizes of the blots closely compared in size to the fragments determined by optical mapping and those predicted from the preliminary sequence. Probe PF2CM93 hybridized to a 7.5 kb band generated by *Nhe*I digestion and PFGE. The fragment size predicted from sequence information was 7.6 kb. The corresponding fragment size from the optical map was also 7.6 kb (Table 1). The same probe hybridized to a 41-kb band generated by *Bam*HI digestion and PFGE. The fragment size predicted from sequence information was 41.3 kb. The corresponding fragment size from the optical map was 40.8 kb (Table 2). Probe PF2NA66 also generated data with fragment sizes that were very similar (Tables 1 and 2). By using the same probe on DNA digested with the two different enzymes, the optical maps were oriented and linked with one another.

DISCUSSION

We have generated a high resolution *Nhe*I and the *Bam*HI optical restriction map of *P. falciparum* chromosome 2, which was used in sequence verification. Despite the fact that chromosome 2 is resolved easily by PFGE, we found the chromosome 2 gel slices to contain only 60% chromosome 2-specific DNA. The balance

Table 1. Comparison of *Nhe*I Optical Map with Restriction Map Predicted from Sequence

Optical map (kb)	Map predicted from sequence (kb)	Difference (kb)	Relative difference (%)	Hybridizing probe
71.8	66.597	5.24		
114.5	115.147	0.63	0.6	
10.3	10.226	0.02	0.2	
3.4	3.359	0.07	2.1	
7.9	7.856	0.05	0.6	
24.7	23.684	1.03	4.4	
6.8	4.933	1.88	38.0	
16.5	14.553	1.97	13.6	
3.2	2.875	0.30	10.3	
	0.177			
11.5	11.425	0.10	0.9	
4.1	3.768	0.30	7.9	
63.8	63.252	0.50	0.8	
10.0	10.018	0.01	0.1	
6.7	6.431	0.27	4.2	
8.9	9.248	0.31	3.3	
28.7	27.327	1.34	4.9	
4.3	4.357	0.07	1.6	
7.6	7.581	0.01	0.01	PF2CM93
11.0	10.588	0.44	4.2	
60.5	60.324	0.21	0.4	
12.3	11.935	0.40	3.3	
4.1	3.964	0.12	3.0	
58.2	57.925	0.25	0.4	
5.5	5.381	0.07	1.3	
	0.363			
1.6	1.546	0.02	1.5	
23.4	22.405	0.96	4.3	
35.1	34.171	0.91	2.6	
18.1	17.156	0.93	5.4	
3.1	2.947	0.16	5.4	
24.9	25.138	0.28	1.1	
40.8	40.107	0.73	1.8	
20.8	20.176	0.59	2.9	
25.1	24.476	0.62	2.5	
77.3	75.172	2.15	2.9	PF2NA66
16.6	16.637	0.07	0.4	
48.0	45.683	2.30	5.0	
9.4	8.546	0.88	10.3	
20.1	18.986	1.15	6.0	
23.9	23.192	0.75	3.2	
32.1	14.897	5.65		
976.5	934.513	0.60	4.3	

was contaminated with DNA molecules from other chromosomes. Consequently, a portion of the optical mapping data was rejected. Should we have mapped other chromosomes using the same strategy we could not predict the acquisition of concise data from chromosomes, which are less resolvable by PFGE, such as chromosomes 5–9.

To check the fidelity of the optical maps independently, Southern blotting of chromosome 2 DNA was performed. Sequenced small-insert clones were used as probes, enabling the optical maps to be cross-checked against the sequence. In all, the optical maps were verified against sequence data and Southern blot analysis, and were found to be very accurate. A more directed operation would be to use sequence-templates as probes for hybridizations to generate a series of anchors for sequence assembly. Such templates would be placed precisely onto the optical map, in terms of physical distance (kb) and would be critical for finishing genomic regions of high complexity; namely, tandem or inverted repeats of high homology and short sequence length. This approach would also readily assemble data acquired using different techniques and would allow the placement of very short sequence contigs onto a map. For example, STS markers or ESTs could be assigned to restriction fragments on a whole genome optical map.

Optical maps of entire chromosomes should also find utility at the sequence-assembly stage in which numerous large contigs are formed, but have unknown order along a chromosome. Traditional approaches to establish contig order rely, in part, on combinatorial PCR, or sequence alignment with physical landmarks,

Table 2. Comparison of *Bam*HI Optical Map with Restriction Map Predicted from Sequence

Optical map (kb)	Map predicted from sequence (kb)	Difference (kb)	Relative difference (%)	Hybridizing probe
77.1	76.648	0.42		
19.9	20.955	1.07	5.09	
7.5	6.81	0.65	9.52	
26.1	27.054	0.95	3.52	
9.9	9.831	0.11	1.15	
41.0	43.295	2.28	5.26	
12.4	13.647	1.22	8.92	
3.7	3.754	0.02	0.67	
34.8	35.985	1.18	3.28	
21.1	20.22	0.91	4.51	
63.6	61.785	1.80	2.92	
55.9	55.217	0.73	1.32	
41.3	40.788	0.50	1.22	PF2CM93
67.3	70.318	3.05	4.33	
46.7	46.943	0.23	0.49	
81.2	87.327	6.14	7.03	
2.0	1.786	0.20	11.35	
8.9	11.633	2.68	23.07	
18.6	17.953	0.69	3.85	
80.8	83.96	3.16	3.77	
19.9	20.665	0.78	3.76	
31.1	30.351	0.72	2.39	
17.4	17.959	0.56	3.10	
28.6	30.812	2.22	7.21	PF2NA66
52.2	49.95	2.26	4.52	
2.0	1.813	0.18	9.70	
24.9	24.79	0.07	0.28	
6.0	5.315	0.65	12.28	
0.5	0.621	0.12	19.48	
34.8	16.346	6.93		
937.2	934.531	1.25	5.86	

which are usually well defined in terms of order but not physical distance. This is where optical maps can streamline the final assembly process by reducing the required number of PCR reactions, by providing an easily interpretable physical scaffold with which sequence contigs can be aligned. The alignment process is to simply generate restriction maps in silico from the sequence data and compare this with the optical maps. When multiple enzymes are used independently and resulting maps are aligned properly, the composite map decreases the size of the sequence contig necessary for confident alignment to the final scaffold.

The information content of a multiple restriction enzyme map is greater than the sum of its parts (Lander and Waterman 1988). We used the sequence data to align the *Nhe*I and *Bam*HI restriction maps with respect to each other, creating a composite map. We expected to find a number of restriction site reversals in this composite. That is, given our sizing errors, closely spaced fragments in the composite map may not be represented in the correct order, and would possibly shift relative position. To our initial astonishment, we found only one instance of reversal. Given this result, we decided to evaluate its statistical significance.

One way to evaluate the quality of a composite enzyme map is to examine how well it preserves the order of the restriction sites. For instance, if we create two maps, one with a restriction enzyme A and the other with the restriction enzyme B, and combine the two maps in correct order, it is still possible that the sizing error in the individual fragments may create a situation, in which a restriction site of type B appears before A, whereas the correct order (in the sequence) is A followed by B— restriction sites shift. Assume that both enzymes cut at the same rate E, and the genome (or chromosome) length is L. Then the total number of fragments of each type is $N = LE$. If the sizing error in a fragment is σ (for instance 1 kb), then the maximum sizing error occurs in the middle of the map and is bounded by $(\sqrt{N/2})\sigma$ (a rather conservative estimate). Thus, a fragment of length l, and cuts of type A in one end and of type B in the other end, may appear in the computed map as a fragment whose length is a random variable with mean l and standard deviation $\sigma' = (\sqrt{N})\sigma$. Thus the probability that this fragment will appear in the reversed order is bounded by $\Phi(l/\sigma')$, where

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-u^2/2} du$$

Furthermore, the length of the fragment with cuts A and B is distributed as $2Ee^{-2El}$. Thus, a random fragment of this kind has a length longer than σ' with probability $e^{-2E\sigma'}$ and a simple estimate shows that the probability of reversal is bounded by

$$(1 - e^{-2E\sigma'})\Phi(0) + e^{-2E\sigma'}\Phi(1)$$

Consider the following values of the parameters $L = 980$ kb, $E = 1/30{,}000$, $\sigma = 1$ kb. For these values, $\sigma' = 5.7$ kb and the average fragment length (with two enzymes) is 15 kb. The above estimate indicates that the probability of reversal is bounded by 0.27. A somewhat better estimate can lower this value to 0.17. As the expected number of fragments with cuts A following B (or B following A) is ~30, one would expect to see fewer than five reversals.

However, the composite map created by optical mapping has only one reversal. The probability of this situation (with fewer than 1 reversal) occurring is ~1 in 40. More exactly, this probability is $(1 - p)^{30} + 30\ p\ (1 - p)^{29} = 0.023$. This difference may signal the requirement for more sophisticated analysis, or indicates the presence of a potentially useful physical effect. A closer examination of the data reveals that the error in the fragment sizes in the composite map has a normal distribution with mean, 0.02 kb and standard deviation, 2.01 kb. Surprisingly, the error in the cut locations has a mean, −1.78 kb and a standard deviation, 1.82 kb, indicative of the presence of systematic (e.g., sequence-specific) error and much smaller unsystematic error. A recalculation of the expected number of reversals with the observed values ($\sigma' = 1.82$ kb) results in slightly more than two reversals, making the observed number of reversals of only one much more likely (~1 in 7 as opposed to 1 in 40). Note that as our estimate of σ' is for the worst-case situation, we believe a more realistic analysis may close the gap. On the other hand, this may be caused by another biochemical effect that we

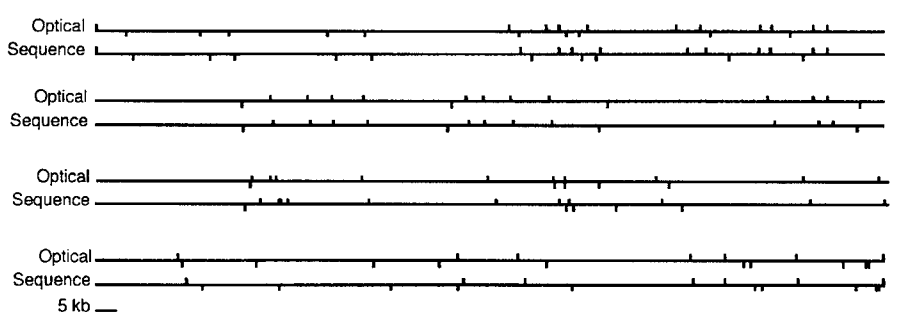


Figure 3 The use of sequence information to link single enzyme maps. The top map was generated by normalizing the single enzyme maps to be the same size (961 kb). The resulting multienzyme map was aligned with the map predicted from sequence. The median relative error is 7%. The average absolute error is 1.4 kb. Upper tick marks are *Nhe*I sites; lower tick marks are *Bam*HI sites.

do not account for in our analysis. More experiments and analyses are required to resolve this situation.

Current optical mapping studies of *P. falciparum* use whole genomic DNA as starting material. The chromosomes are resolved at the level of data rather than as physical entities. The data segregates into 14 deep contigs corresponding to the various chromosomes. Chromosome 2 can be resolved based on size and the near complete correspondence with the data shown in this paper (one 600-bp *Bam*HI fragment is missing on the whole genome map). The success of this project has prompted the Malaria Genome Consortium to recommend support of whole genome mapping to assist in closure of chromosomes, as well as for verification of the final assembly.

In summary, we describe the construction of an ordered restriction map of *P. falciparum* chromosome 2 using optical mapping of genomic DNA. A combined approach using shotgun sequencing and optical mapping will facilitate sequence assembly and finishing of large and complex genomes.

METHODS

Parasite Preparation

P. falciparum (clone 3D7) was cultivated using standard techniques (Trager and Jensen 1976). To minimize possible alterations of the genome that can occur in continuous culture (Corcoran et al. 1986), parasite aliquots were kept frozen in liquid N_2 until needed and then cultivated only as long as necessary. Parasites were cultivated to late trophozoite/early schizont stages and enriched on a Plasmagel gradient. The parasitized red blood cells were washed once with several volumes of 10 mM Tris (pH 8), 0.85% NaCl and the parasites were freed from the erythrocytes by incubation in ice-cold 0.5% acetic acid in dH_2O for 5 min, followed by several washes in cold buffer. The parasites were resuspended to a concentration of 2×10^9/ml in buffer and maintained in a 50°C waterbath. An equal volume of 1% InCert agarose (FMC, Rockland, ME) in buffer, prewarmed to 50°C, was mixed with the prewarmed parasites and the mixture was added to a $1 \times 1 \times 10$-cm gel mold, plugged at one end with solidified agarose, and was allowed to cool to 4°C. The agarose-embedded parasites were pushed out of the mold and incubated with 50 ml of proteinase K solution (2 mg/ml proteinase K in 1% Sarkosyl, 0.5 M EDTA) at 50°C for 48 hr with one change of proteinase K solution and were stored in 50 mM EDTA at 4°C (Schwartz and Cantor 1984).

Chromosome 2 Isolation by PFGE

Uniform parasite slices were taken with a glass coverslip using two offset microscope slides as guides. One half to one quarter of a single slice was sufficient per lane. Parasite slices were arranged end to end on the flat side of the gel comb. The parasites were fixed to the comb by a small bead of molten (60°C) agarose. The comb was then placed into the gel mold and molten agarose [1.2% SeaPlaque (FMC) in 0.5× TBE] poured around the parasite-containing slices. Once cooled, the comb was removed and the space filled with molten agarose. A CHEF DRIII apparatus (Bio-Rad, Hercules, CA) was used for all PFGE (Schwartz and Cantor 1984) chromosome separations. Gels were run with 180–250 sec of ramped pulse time at 3.7 V/cm and 120° field angle, for 90 hr at 14°C with recirculating buffer at ~1 l/min, using *Saccharomyces cerevisiae* and/or *Hansenula wingei* PFGE size markers (Bio-Rad). To minimize UV damage to the DNA, gel slices were removed from the ends of the gel, stained with ethidium bromide (5 μg/ml), and visualized by long wave (320 nm) UV light. Notches corresponding to the individual chromosomes were made in the agarose gel and used as guides to cut the chromosome from the gel. The chromosome-containing gel slices were stored in 50 mM EDTA at 4°C until needed. The gel was stained with ethidium bromide to verify the chromosome excision. The genome of *P. falciparum* is 26–30 Mb in size, consisting of 14 chromosomes ranging in size from 0.6–3.5 Mb (Foote and Kemp 1989). PFGE resolves most of the *P. falciparum* chromosomes, except 5–9 which are similar sizes and comigrate. The gel band containing *Plasmodium falciparum* chromosome 2 was resolved easily, cut from the gel, melted at 72°C for 7 min and incubated with agarose at 40°C for 2 hr. The melted agarose band was diluted in TE to a final DNA concentration suitable for optical mapping (~20 pg/μl.

Mounting and Digestion of DNA on Optical Mapping Surface

Optical mapping surfaces were prepared as described previously (Aston et al. 1999). Briefly, glass coverslips (18×18 mm^2; FISHER Finest, Pittsburgh, PA) were cleaned by boiling in concentrated nitric, then hydrochloric acid. Surfaces were derivatized with 3-aminopropyldiethoxymethyl silane (APDEMS; Aldrich Chemical, Milwaukee, WI). One surface was placed onto a microscope slide. A DNA sample (10 μl) was added to the edge between the surface and the slide and spread into the space between the surface and the slide. The surface was then peeled off from the slide. Digestion was performed by adding 100 μl of digestion solution [50 mM NaCl, 10 mM Tris-HCl (pH 7.9), 10 mM $MgCl_2$, 0.02% Triton X-100, 20 units of restriction endonuclease; New England Biolabs, Beverly, MA] onto the surface and incubating at 37°C from 15 to 30 min. The buffer was aspirated and the surface washed with water before staining of DNA with YOYO-1 homodimer (Molecular Probes, Eugene, OR), prior to fluorescence microscopy. Comounted λ bacteriophage DNA (New England Biolabs) was used as a sizing standard and also to estimate cutting efficiencies.

Image Acquisition, Processing, and Map Construction

DNA molecules were imaged by digital fluorescence microscopy. The optical mapping surface was scanned by the operator for individual digested DNA molecules of adequate length and quality to be collected for image processing and map making. Images were collected with a cooled charge coupled device (CCD) camera (Princeton Instruments, Trenton, NJ) using Optical Map Maker (OMM) software, as described previously (Jing et al. 1998). Images of DNA fragments were processed using a modified version of NIH Image (Huff 1996) which integrates fluorescence intensity for each fragment. These values were used to assemble an ordered restriction map for each molecule. Fluorescence intensity of λ bacteriophage DNA standards was used to measure the size of the *P. falciparum* restriction fragments on a per image basis. Cutting efficiences (on a per image basis) were determined from scoring

cut sites on sizing standard molecules contained in the same field as the genomic DNA molecules. Standard molecules were cut once by *Nhe*I and five times by *Bam*HI. The map for the entire chromosome 2 was manually assembled into contigs by aligning overlapping regions of congruent cut sites. If there were no overlapping regions, the molecules were considered to be from a contaminating *P. falciparum* chromosome and were discarded. Consensus maps for chromosome 2 were assembled by averaging the fragment sizes from the individual maps derived from maps underlying the contigs.

Southern Blotting of *P. falciparum* Genomic DNA

P. falciparum genomic DNA (10 μg) was digested with *Nhe*I or *Bam*HI, resolved by PFGE (POE apparatus, 1% gel in 0.5× TBE, pulse time, 1 sec, 2 sec; switch time, 12 sec, 150 V, for 24 hr) (Schwartz and Koval 1989), blotted, and hybridized with probes derived from small insert clones used for sequencing (PF2CM93 and PF2NA66). Probes were labeled by random priming.

ACKNOWLEDGMENTS

This work was supported by the Burroughs Wellcome Fund and the Naval Medical Research and Development Command work unit STEP C611102A0101BCX. The opinions and assertions herein are those of the authors and are not to be construed as official or as reflecting the views of the U.S. Navy or naval service at large.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Ananthraman, T.S., B. Mishra, and D.C. Schwartz,. 1997. Genomics via optical mapping II: Restriction maps. *J. Comput. Bio.* **4:** 91–118.

Anantharaman, T.S., B. Mishra, and D.C. Schwartz. 1998. Genomics via optical mapping III: Contiging genomic DNA and variations. *Courant Technical Report #760*, Courant Institute, New York.

Aston, C., C. Hiort, and D.C. Schwartz. 1999. Optical mapping: An approach for fine mapping. *Methods Enzymol.* **303:** (in press).

Cai, W., H. Aburatani, D. Housman, Y. Wang, and D.C. Schwartz. 1995. Ordered restriction endonuclease maps of yeast artificial chromosomes created by optical mapping on surfaces. *Proc. Natl. Acad. Sci.* **92:** 5164–5168.

Cai, W., J. Jing, B. Irvin, L. Ohler, E. Rose, U. Kim, M. Simon, and D.C. Schwartz. 1998. High resolution restriction maps of bacterial artificial chromosomes constructed by optical mapping. *Proc. Natl. Acad. Sci.* **95:** 3390–3395.

Corcoran, L.M., K.P. Forsyth, A.E. Bianco, G.V. Brown, and D.J. Kemp. 1986. Chromosome size polymorphism in plasmodium falciparum can involve deletions and are frequent in nature parasite populations. *Cell* **44:** 87–95.

Dame, J.B., D.E. Arnot, P.F. Bourke, D. Chakrabarti, Z. Christodoulou, R.L. Coppel, F. Cowman, A.G. Craig, K. Fischer, J. Foster et al. 1996. Current status of the *Plasmodium falciparum* genome project. *Mol. Biochem. Parasitol.* **79:** 1–12.

Foote, S.J. and D.J. Kemp. 1989. Chromosomes of malaria parasites. *Trends Genet.* **5:** 337–342.

Foster, J. and J. Thompson. 1995. The *Plasmodium falciparum* genome project: A resource for researchers. The Wellcome Trust Malaria Genome Collaboration. *Parasitol. Today* **11:** 1–4.

Gardner, M.J., H. Tettelin, D.J. Carucci, L.M. Cummings, L. Aravind, E.V. Koonin, S. Shallom, T. Mason, K. Yu, C. Fujii et al. 1998. Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science* **282:** 1126–1132.

Huff, E. 1996. Ph.D. thesis. Department of Chemisty, New York University, New York, NY.

Jing, J., J. Reed, J. Huang, X. Hu, V. Clarke, J. Edington, D. Housman, T. Anantharaman, E. Huff, B. Mishra et al. 1998. Automated high resolution optical mapping using arrayed, fluid fixed, DNA molecules. *Proc. Natl. Acad. Sci.* **95:** 8046–8051.

Lander, E.S. and M.S Waterman. 1988. Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics* **2:** 231–239.

Lin, J., R. Qi, C. Aston, J. Jing, T.S. Anantharaman, B. Mishra, O. White, J.C. Venter, and D.C. Schwartz. 1998. Complete shotgun optical mapping of *Deinococcus radiodurans* and *Escherichia coli* K12 using genomic DNA molecules. Submitted.

Netzel, T.L., K. Nafisi, M. Zhao, J.R. Lenhard, and I. Johnson. 1995. Base-content dependence of emission enhancements, quantum yields, and lifetimes for cyanine dyes bound to double-strand DNA: Photophysical properties of monomeric and bichromophoric DNA stains. *J. Phys. Chem.* **99:** 17936–17947.

Schwartz, D.C. and C.R. Cantor. 1984. Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. *Cell* **37:** 67–75.

Schwartz, D.C. and M. Koval. 1989. Conformational dynamics of individual DNA molecules during gel electrophoresis. *Nature* **338:** 520–522.

Schwartz, D.C., X. Li, L. Hernandez, S. Ramnarain, E. Huff, and Y. Wang. 1993. Ordered restriction maps of *Saccharomyces cerevisae* chromosomes constructed by optical mapping. *Science* **262:** 110–114.

Sinnis, P. and T.E. Wellems. 1988. Long-range restriction maps of *Plasmodium falciparum* chromosomes: Crossingover and size variation among geographically distant isolates. *Genomics* **3:** 287–295.

Trager, W. and J.B. Jensen. 1976. Human malaria parasites in continuous culture. *Science* **193:** 673–675.

Thompson, J.K. and A.F. Cowman 1997. A YAC contig and high resolution map of chromosome 3 from *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **90:** 537–542.

Watanabe, J. and J. Inselburg. 1994. Establishing a physical map of chromosome No. 4 of *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **65:** 189–199.

Received October 5, 1998; accepted in revised form December 15, 1998.

Parassitologia 41: 69-75, 1999

Appendix J
Final Report
DAMD17-82-2-8005

The malaria genome sequencing project: complete sequence of *Plasmodium falciparum* chromosome 2

M.J. Gardner[1], H. Tettelin[1], D.J. Carucci[2], L.M. Cummings[1], H.O. Smith[1], C.M. Fraser[1], J.C. Venter[1], S.L. Hoffman[2]

[1] *The Institute for Genomic Research;* [2] *Malaria Program, Naval Medical Research Center, Rockville, MD, USA.*

Abstract. An international consortium has been formed to sequence the entire genome of the human malaria parasite *Plasmodium falciparum*. We sequenced chromosome 2 of clone 3D7 using a shotgun sequencing strategy. Chromosome 2 is 947 kb in length, has a base composition of 80.2% A+T, and contains 210 predicted genes. In comparison to the *Saccharomyces cerevisiae* genome, chromosome 2 has a lower gene density, a greater proportion of genes containing introns, and nearly twice as many proteins containing predicted non-globular domains. A group of putative surface proteins was identified, rifins, which are encoded by a gene family comprising up to 7% of the protein-encoding genes in the genome. The rifins exhibit considerable sequence diversity and may play an important role in antigenic variation. Sixteen genes encoded on chromosome 2 showed signs of a plastid or mitochondrial origin, including several genes involved in fatty acid biosynthesis. Completion of the chromosome 2 sequence demonstrated that the A+T-rich genome of *P. falciparum* can be sequenced by the shotgun approach. Within 2-3 years, the sequence of almost all *P. falciparum* genes will have been determined, paving the way for genetic, biochemical, and immunological research aimed at developing new drugs and vaccines against malaria.

Key words: *Plasmodium falciparum*, malaria, chromosome 2, rifins, genomics, malaria genome sequencing project.

In 1995, the first complete genome sequence of a free-living organism, *Haemophilus influenzae*, was published (Fleischmann *et al.*, 1995). The publication of the *H. influenzae* genome sequence marked a turning point in biology. As noted by Bloom, it heralded a post-genomics era of microbe biology when the complete genomes of most human pathogens would have been sequenced, providing a vast database of sequence information that would enable researchers to focus on studies of the biology and pathogenicity of these organisms (Bloom, 1995). This research in turn would lead to the development of new drugs and vaccines to treat and prevent diseases caused by these pathogens, and would be especially useful for research on organisms difficult to grow. Since then, there has been a flurry of effort to sequence the genomes of other pathogens, and the genomes of organisms that cause diseases such as syphylis (*Treponema pallidum*), ulcers (*Helicobacter pylori*), Lyme disease (*Borrelia burgdorferi*), tuberculosis (*Mycobacterium tuberculosis*), and trachoma (*Chlamydia trachomatis*) have been completed (Fraser *et al.*, 1997, 1998; Tomb *et al.*, 1997; Cole *et al.*, 1998; Stephens *et al.*, 1998). The genomes of several microbes of environmental importance have also been sequenced, as has the genome of the yeast *Saccharomyces cerevisiae* (see <http://www.tigr.org/tdb/mdb/mdb.html> for a complete listing of microbial genomes that have been sequenced or that are in progress). There is no evidence, so far, that the pace of sequencing has slackened, and that more than 60 microbial genomes is currently underway.

The completion of the first few microbial genomes caused several groups to contemplate the sequencing of the *Plasmodium falciparum* genome. It was realized that determination of the complete *P. falciparum* genome sequence would be of great value to malariologists given the difficulty of studying this organism in the laboratory, with large parts of the life cycle being difficult, expensive, or impossible to maintain in the laboratory. Furthermore, techniques such as DNA microarrays and transfection had been developed, providing researchers with new tools to study the expression and function of genes and gene products in malaria parasites. Several groups initiated pilot sequencing projects, and an international consortium including malaria researchers, genome laboratories, bioinformatics centers, and funding agencies was formed to coordinate the project, facilitate collaboration, and ensure that the data would be provided to the scientific community in a timely and useful manner (Hoffman *et al.*, 1997). The consortium met every 6 months during the start-up phase of the project and continues to meet regularly as the work proceeds.

At the time the *P. falciparum* project was started,

Invited contribution to the Malariology Centenary Conference "The malaria challenge after one hundred years of malariology" held in Rome at the Accademia Nazionale dei Lincei, 16-19 November 1998.

Correspondence: Dr Malcolm J. Gardner, The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA, Tel ++1 301 8383519, Fax ++1 301 8380208, e-mail: gardner@tigr.org

several prokaryotic and archaeal genomes had been finished, and sequencing of the genomes of yeast and *Caenorhabditis elegans* were nearing completion. Two strategies had been used in these projects. The clone-by-clone method, used to sequence the *Escherichia coli* and *S. cerevisiae* genomes, for example, involved sequencing of large-insert clones from cosmid, lambda, and YAC libraries (Blattner *et al.*, 1997). The clones sequenced were selected after the construction of a physical map, which provided a tiling path of overlapping clones spanning the genome. The other method, pioneered at TIGR, was the whole genome shotgun method, which used a genomic library of sheared 1-2 kb fragments prepared in a plasmid vector (Fleischmann *et al.*, 1995). Thousands of randomly selected small insert clones were picked and sequenced, and custom fragment assembly software was used to assemble the overlapping fragments into a contiguous sequence. This method proved to be very efficient in that construction of a physical map was not required prior to sequencing. However, very robust software for fragment assembly had to be developed that was able to handle many thousands of individual sequence reads and also deal with the repetitive sequences present in bacterial genomes. In addition, relational databases and software were developed to manage the gap closure, finishing, and annotation processes.

Sequencing of the *P. falciparum* genome raised some formidable technical challenges, however. At ~28 Mb, the *P. falciparum* genome was almost 20-fold larger than the *H. influenzae* genome and seemed too large to tackle by the whole genome shotgun method because of the computational requirements of the assembly process. Closure of the many gaps that would have remained after the initial assembly would also have been difficult with such a large genome and few sequence markers to guide the closure process. On the other hand, the clone-by-clone approach was ruled out because large-insert (>20 kb) genomic libraries of very AT-rich *P. falciparum* DNA in plasmid, lambda, and cosmid vectors that could be used for sequencing were not available. Although large-insert yeast artificial chromosome (YAC) libraries of *P. falciparum* (Foster and Thompson, 1995) had been constructed which appeared to be stable, YACs are not very well suited to high-throughput sequencing projects. Consequently, a new approach was adopted in which individual chromosomes were resolved on pulsed-field gels and used to prepare chromosome-specific shotgun libraries in plasmid and M13 vectors. Randomly-selected clones were then sequenced and assembled in the same way as for a whole-genome shotgun project. Some laboratories also performed low-coverage sequencing of shotgun libraries prepared from YACs previously mapped on the chromosomes (Foster and Thompson, 1995); the YAC shotgun sequences helped to group sequences from the same part of the chromosome and assisted in gap closure. Adoption of the chromosome-by-chromosome shotgun strategy allowed the sequencing effort to be distributed among the different sequencing centers.

Three groups are involved in the sequencing effort: TIGR and the Malaria Program of the US Naval Medical Research Center (NMRC); the Sanger Centre in the UK; and Stanford University. The current status of the project (as of July 1999) is summarized in Table 1. Once the problems that had been encountered in library construction, sequencing, assembly and gap closure were solved, all 3 groups began to make rapid progress. The complete sequence of chromosome 2 (0.95 Mb) was recently published by the TIGR/NMRC group (Gardner *et al.*, 1998), and the Sanger Center has virtually finished chromosome 3 (1.1 Mb). Work on the other chromosomes is well underway. The chromosome 2 sequence was submitted to GenBank and the sequence and annotation is available at TIGR's web site and at the NCBI (Table 1). Preliminary unedited sequence data is also available for downloading, browsing or searching on web sites maintained at each laboratory.

Table 1. Chromosome assignments and current status of the Malaria Genome Sequencing Project. [a] Estimated chromosome sizes for *P. falciparum* clone 3D7 were taken from Dame *et al.* (1996) or from the sequence data. [b] NIAID, National Institute for Allergy and Infectious Diseases; DoD, US Department of Defense; BWF, Burroughs Wellcome Fund. [c] Complete annotation (chromosome 2) or preliminary data can be viewed at web sites maintained by the sequencing centers: TIGR/NMRC <http://www.tigr.org/tdb/mdb/pfdb/pfdb.html>; the Sanger Centre <http://www.sanger.ac.uk/Projects/P_falciparum/>; Stanford University <http://baggage.stanford.edu/group/malaria/start.html>.

Cromosome(s)[a]	Size (Mb)	Laboratory	Funding[b]	Status (as of 7/99)[c]
1	0.8	Sanger Centre	Wellcome Trust	Closure
2	0.95	TIGR/NMRC	NIAID, DoD	Completed (Gardner *et al.*, 1998)
3	1.1	Sanger Centre	Wellcome Trust	Completed (Bowman *et al.*, in press)
4	1.4	Sanger Centre	Wellcome Trust	Closure
5-8	1.6	Sanger Centre	Wellcome Trust	Sequencing
9	1.8	Sanger Centre	Wellcome Trust	Sequencing
10	2.1	TIGR/NMRC	NIAID, DoD	Sequencing
11	2.3	TIGR/NMRC	NIAID, DoD	Closure
12	2.5	Stanford University	BWF	Closure
13	3.2	Sanger Centre	Wellcome Trust	Sequencing
14	3.4	TIGR/NMRC	BWF, DoD	Closure

Sequencing of the first *P. falciparum* chromosome

At the beginning of the Malaria Genome Sequencing Project, *P. falciparum* clone 3D7 was chosen for sequencing because it can complete all stages of the life cycle, was used in a genetic cross (Walliker *et al.*, 1987), and had been used in the Wellcome Trust Malaria Genome Mapping Project (Foster and Thompson, 1995). The TIGR/NMRC group began a pilot project to sequence chromosome 2, which was selected because it could be easily resolved on pulsed-field gels, and being about 1 Mb in size it was not too large to present unsurmountable difficulties in assembly or gap closure. *P. falciparum* chromosomes were resolved on preparative pulsed-field gels and the chromosome 2 bands from several gels were cut out, adjusted to 0.3 M sodium acetate to prevent melting of the AT-rich DNA, and digested with agarose. The DNA was sheared by nebulization and a shotgun library was prepared in pUC18 as described (Fleischmann *et al.*, 1995) except that treatment with *E. coli* DNA polymerase I was performed after the second ligation step to close nicks prior to electroporation. During all steps of the library construction process, the exposure of the DNA to UV light was minimized to avoid damage to the DNA that would reduce the cloning efficiency, particularly of the very AT-rich intergenic sequences. In addition, to prevent generation of non-randomness, the library was not amplified prior to sequencing. Rather, the ligation mixtures were stored at –20°C, and as needed aliquots were electroporated into DH10B cells and spread on ampicillin diffusion plates. The shotgun library contained 1×10^5 recombinants and had an average insert size of 1.6 kb.

Initial sequencing was done with dye-primer chemistry used previously to sequence *H. influenzae* and the other microbial genomes. However, when sequencing the *P. falciparum* clones we observed an apparent artifact with the dye-primer chemistry that resulted in runs of G nucleotide base calls to be incorrectly made following long runs of AT-rich sequence. The artifact did not occur when FS+ dye-terminator chemistry was used on the same template DNAs, and the dye-terminator chemistry also produced significantly longer sequence reads than the dye-primer chemistry. Therefore the rest of the random-phase sequencing was performed using the dye-terminator chemistry. Over 23,000 individual sequences were collected, which was equivalent to about 10x coverage of the chromosome. This is greater coverage than is normally done in a shotgun project, but the excess coverage was thought to be necessary to compensate for the presence of non-chromosome 2 DNA in the library arising from the pulsed-field gel purification of the DNA, and for the expected non-randomness of the shotgun library due to the AT-rich inserts.

The sequence reads were assembled using a version of TIGR Assembler (Sutton *et al.*, 1995) that was extensively modified to assemble the AT-rich and repeat-rich *Plasmodium* sequences. TIGR Assembler identifies and aligns overlapping fragments in two steps. The initial step in assembly is to locate all *n*-mer oligonucleotides shared between fragment pairs. The software views all fragment pairs with a high degree of *n*-mer similarity as potentially overlapping, and in the second step the Smith-Waterman method is used to align the fragments. In the bacterial genome projects the value of *n* used was typically 10-12 nucleotides. However, using n=10 with AT-rich *Plasmodium* DNA resulted in incorrect identification of thousands of potential fragment overlaps, so that the program spent an inordinate amount of time attempting to align the spurious matches. Increasing *n* from 10 to 32 much reduced this problem and significantly lowered the time required for assembly.

After the assembly, 610 contigs were obtained and the largest contigwas 50 kb. Neighboring contigs were identified and ordered by the program GROUPER, which searches for plasmid templates with forward and reverse reads in different contigs (clone links), and for overlapping contigs that failed to merge under the stringent overlap criteria required by TIGR Assembler (grasta links). Contigs within a group are separated by sequence gaps which can be closed by primer walking on the templates identified as clone links, or by editing of the termini of contigs with grasta links. The ends of groups represent physical gaps for which no shotgun clone could be identified. Ten groups of 114 contigs were localized on the chromosome by comparison to STS markers (Lanzer *et al.*, 1993). Closure of physical and sequence gaps used approaches described previously (Fleischmann *et al.*, 1995), with a few modifications to compensate for the AT-richness of the DNA. To close the 9 physical gaps in the central region of the chromosome, PCR reactions using genomic DNA as template were performed with primers from the ends of adjacent groups. PCR products were purified and sequenced using dye-terminator chemistry. This process closed 3 physical gaps immediately, but PCR products from 2 gaps contained very AT-rich sequence which could not be sequenced completely, and remained as sequence gaps. Those physical gaps for which PCR products could not be obtained in the first step were reasoned to be too large for PCR, and to contain one or more of the unlocalized groups. We therefore performed combinatorial PCRs with one primer from the end of a localized group and the second primer from the ends of all free groups larger than 2.5 kb. Two gaps were closed by the combinatoral strategy. Finally, 1 physical gap was closed after editing and reassembly, and another gap was closed by sequencing of a 'missing mate' (i.e., resequencing of a clone for which either the forward or reverse sequencing reaction had failed during the random phase). Five methods were used to close sequence gaps. For contigs which overlapped but had not been merged during assembly, editing and resequencing were per-

formed to close the gaps. Many sequence gaps were caused by artifacts in dye-primer reactions, particularly in extremely AT-rich areas. Long homopolymer stretches of up to 50 consecutive A or T residues also caused the sequence quality to decline downstream of the homopolymer region. These artifacts either prevented the merging of overlapping contigs or produced short sequences that did not extend to the neighboring contig. Some of these problem areas could be solved by trimming of the low quality sequence that prevented merging of the contigs. For other gaps, templates from short or low-quality dye-primer reactions in the vicinity of sequence gaps were identified and resequenced with dye-terminator chemistry; the longer reads of high-quality sequence provided by the dye-terminator reactions was sufficient to close many gaps. For those gaps that remained, primer walking on plasmid templates linking adjacent contigs was used. Finally, there were 5 sequence gaps that could not be closed by the above methods because the sequence was too AT-rich for primer synthesis and walking. To close these gaps, the artificial transposon AT-2 (Devine and Boeke, 1994) was inserted into one of the templates spanning each sequence gap, multiple subclones of each template were sequenced using transposon-specific primers, and the sequences were assembled to close the gap. The chromosome 2 sequence was edited manually using TIGR Editor, and where necessary additional sequencing reactions were performed to improve coverage and resolve sequence ambiguities. One major concern, given the well-known propensity for AT-rich *P. falciparum* sequences to rearrange in *E. coli*, was whether the assembled sequence was an accurate representation of the genomic sequence. To independently confirm the colinearity of the assembled sequence and genomic DNA, *Nhe*I and *Bam*HI optical restriction maps of chromosome 2 DNA were prepared and compared with restriction maps predicted from the sequence (Jing *et al.*, 1999). The relative error of predicted and observed fragment sizes was less than 6%, which proved that there were no major rearrangements in the assembled sequence.

Annotation of *P. falciparum* chromosome 2

Annotation of the chromosome 2 sequence followed the procedures used previously during the annotation of other genomes, including BLAST searching of all open reading frames (ORFs) against a protein sequence database. In addition, to assist in defining the intron/exon boundaries, a new eukaryotic gene finding program was developed specifically for use in this project (Salzberg *et al.*, 1999). This program, GlimmerM, was trained on a set of 117 *P. falciparum* sequences taken from Genbank. Gene models based on the GlimmerM predictions, the similarity of the ORFs to known proteins, and prediction of putative signal peptides and transmembrane domains were constructed.

Chromosome 2 of *P. falciparum* is 947,103 bp in length and 80.2% A+T (Gardner *et al.*, 1998). It possesses typical eukaryotic telomeres and subtelomeric regions containing several kb of rep20 tandem repeats, variant antigen genes (*var*), and a potential new family of variant surface antigens related to the RIF-1 elements (repetitive interspersed family) (Weber, 1988). The large central region encodes many single copy genes and several genes that are tandemly repeated (Fig. 1). Two hundred and nine protein-encoding genes and a gene encoding $tRNA^{Glu}$ were predicted on chromosome 2, giving a gene density of one gene per 4.5 kb, which is significantly lower than in yeast (one gene per 2 kb) but higher than in *C. elegans* (one gene per 7 kb). It was estimated that 43 of the 209 protein-encoding genes contained at least one intron, with most such genes consisting of 2 or 3 exons. Two genes, however, contained 8 exons. Extrapolation of the chromosome 2 data to the entire 28 Mb *P. falciparum* genome suggests that it contains 6,200 genes, 2,600 of which may contain introns. Thus, in terms of intron content and gene density the *P. falciparum* genome appears to be intermediate between the compact yeast genome and the intron-rich genomes of multicellular eukaryotes.

Of the 209 protein encoding genes, only 87 (42%) appeared to have homologs outside *Plasmodium*, suggesting that almost 60% of the genes encoded on this chromosome are so far 'unique' to *Plasmodium*. The proportion of unique genes is almost 2-fold greater than has been observed in other organisms, and confirms that there is much biology to be uncovered in future studies of this parasite. As sequencing of other related parasites proceeds, some of these proteins will undoubtedly be found to have homologs in apicomplexans such as *Toxoplasma* (Ajioka *et al.*, 1998) and *Eimeria*, and hence may be found to be characteristic of apicomplexan parasites. Most of the remaining unidentified proteins on chromosome 2 were predicted to consist primarily of non-globular domains, i.e. domains that are composed of low complexity sequences that do not form compact folded structures (Wootton and Federhen, 1996). The abundance of non-globular domains or proteins in *Plasmodium* was very unusual, and was about half that observed in *S. cerevisiae*, *C. elegans*, and humans. In addition, 13 proteins contained large regions (>30 amino acids) with predicted non-globular structure inserted directly into globular domains, a phenomenon so far unique to *Plasmodium*. These non-globular insertions did not exhibit the AT-bias typical of introns, were not flanked by consensus splice sites, and based on RT-PCR analysis of several genes encoding non-globular domains, were likely to be expressed in the proteins. The abundance of the non-globular domains in *Plasmodium* proteins suggests that they provide as yet unknown selective advantages to the parasite. Study of these proteins containing non-globular inserts may also provide new insight into the general principles of protein folding.

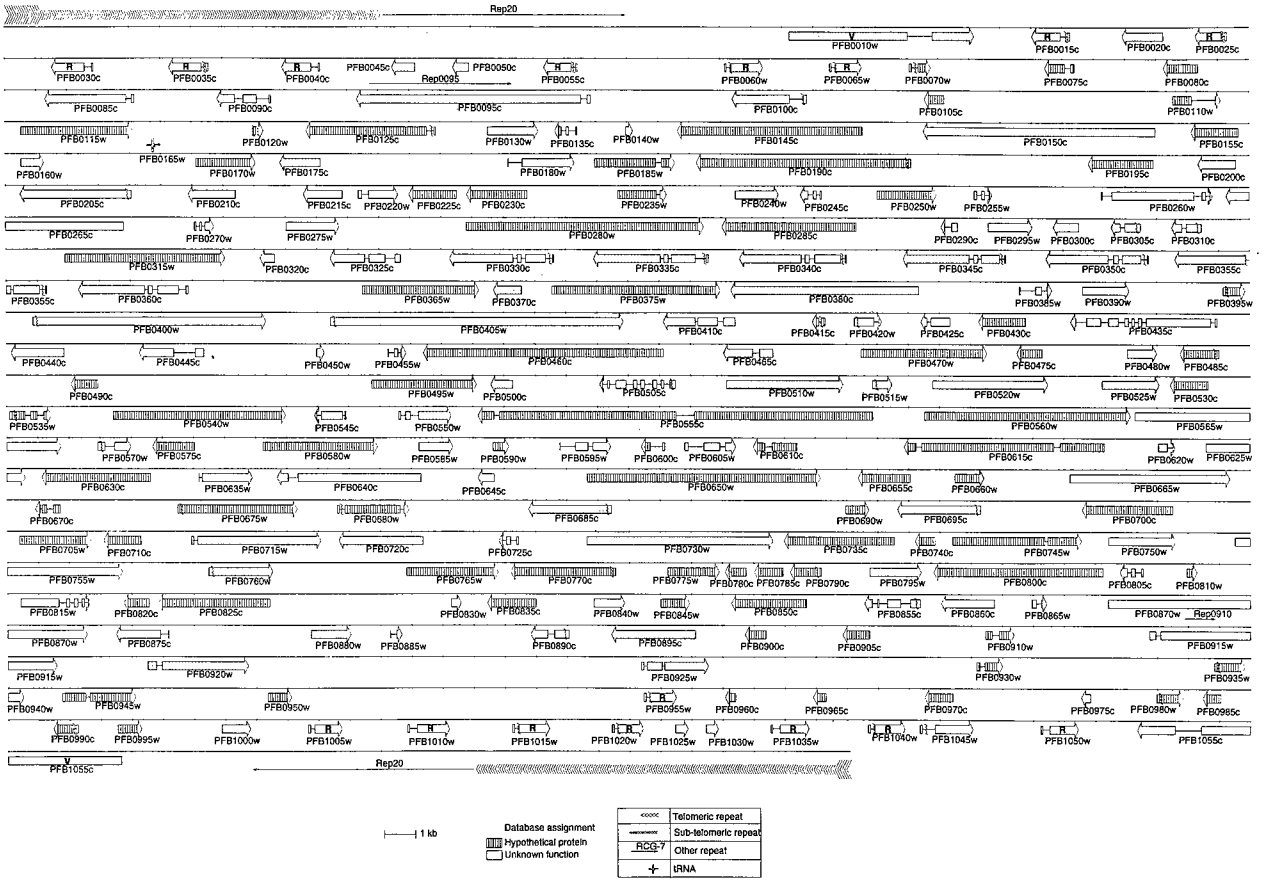


Fig. 1. Map of *P. falciparum* chromosome 2 (clone 3D7). Exons are shown as boxes or arrows, with introns represented by thin lines connecting the exons. Other features such as telomeric and subtelomeric repeats are indicated as shown in the legend. Chromosome 2 genes with similarity to known genes in the sequence databases and for which putative functional assignments could be made are stippled; hypothetical genes with no detectable similarity to known genes are indicated by vertical stripes; genes with similarity to previously sequenced genes of unknown function are indicated as open arrows. The rifin and *var* genes are labeled with 'R' and 'V', respectively. Genes were given systematic names using a scheme similar to that devised for the *S. cereviseae* genome (Mewes *et al.*, 1997). For a complete description of the genes encoded on chromosome 2, including details of functional assignments, see Gardner *et al.* (1998).

Most of the 87 evolutionarily-conserved proteins encoded on chromosome 2 show the greatest similarity to eukaryotic homologs or belong to specifically eukaryotic protein families. Many of these genes code for proteins that participate in replication, repair, transcription, or translation, and include the origin recognition complex subunit 5, two proteins involved in excision repair proteins, several proteins involved in chromatin dynamics, RNA-binding proteins, and a putative transcription factor. Other evolutionarily conserved proteins are involved in secretion, such as the SEC61 gamma subunit, the coated pit coatamer subunit, and syntaxin, suggesting early emergence of the eukaryotic secretory system. Five proteins contained DnaJ domains; in other organisms DnaJ proteins have been shown to act as cofactors for the HSP70-type molecular chaperones and to participate in a variety of processes such as protein folding and trafficking, complex assembly, organelle biogenesis, and initiation of translation (Cyr *et al.*, 1994). Chromosome 2 encodes 90 predicted membrane proteins, some of which appear to be transporters of amino acids or sugars. Five putative protein kinases were also identified, suggesting that the *P. falciparum* genome may encode about 150 protein kinases. This prominence of regulators is in striking contrast to the situation in bacterial pathogens, which appear to have shed most of the regulatory systems, and is probably a reflection of the complex life cycle. For example, phosphorylation and dephosphorylation reactions are known to be involved in the development and sexual differentiation of malaria parasites (Bracchi *et al.*, 1996). A cluster of 8 tandemly arranged genes encoding putative proteases was also found; 3 of these genes were known previously and were called SERAs (SErine Repeat Antigens). The expansion of this protease gene family suggests an important function, possibly in merozoite release from schizonts or processing of merozoite surface proteins.

While most of the evolutionarily conserved proteins were more similar to eukaryotic homologs, 16 proteins were significantly more similar to bacterial homologs and 4 other proteins were the first eukaryotic representatives of conserved bacterial protein families. These proteins may have been transferred to the nuclear genome from an organellar genome after the divergence of the apicomplexa from the other eukaryotic lineages. Several of these proteins contained N-terminal sequences that resembled organellar import peptides, which suggested that these proteins may be imported into and function within either the apicoplast or the mitochondrion. Of particular interest were 3 genes encoding proteins involved in fatty acid metabolism. One of these proteins, 3-ketoacyl-ACP synthase III (FabH), catalyzes the condensation of acetyl-CoA and malonyl-ACP in Type II (dissociated) fatty acid synthase systems. Type II synthase systems are restricted to bacteria and the plastids of plants, and the discovery of a Type II fatty acid synthase system in *Plasmodium* reinforced previous hypotheses that the apicoplast contains plant-like metabolic pathways distinct from those of the host (Wilson *et al.*, 1991; Slabas and Fawcett, 1992). Some of the biochemical processes that occur within this organelle may therefore be good drug targets (Soldati, 1999). Recent work has confirmed that at least some of the predicted import peptides can direct translocation of reporter proteins into the apicoplast in *Toxoplasma*, and in addition, that thiolactomycin, a specific inhibitor of bacterial FabH, can inhibit the growth of *P. falciparum* in vitro (Waller *et al.*, 1998).

As mentioned previously, more than half of all proteins encoded on chromosome 2 did not have detectable homologs in other species. Many of the *Plasmodium* specific genes were located in the subtelomeric regions of the chromosome. Two members of the *var* gene family were identified on chromosome 2, one in each sub-telomeric region. The *var* genes encode large proteins, collectively known as PfEMP1s, that are located on the surface of infected red cells, exhibit extensive sequence diversity, and are involved in antigenic variation, cytoadherence, and rosetting (Baruch *et al.*, 1995; Smith *et al.*, 1995; Su *et al.*, 1995; Rowe *et al.*, 1997). Most *var* genes are located in sub-telomeric regions, and *var* gene diversity is thought to be generated by recombination between alleles, a process which might be facilitated by the sub-telomeric repeats (Rubio *et al.*, 1996). Six small ORFs that had similarity to *var* sequences were also found in the sub-telomeric regions. Five of these ORFs resembled the *var* exon II cDNAs or the Pf60.1 sequences that were reported previously (Su *et al.*, 1995; Bonnefoy *et al.*, 1997). However, the largest gene family identified on chromosome 2 encoded proteins of 27-35 kD that were named rifins, after the RIF-1 repetitive elements (Weber, 1988). These proteins contained a N-terminal signal sequence, a central region of variable length and an amino acid sequence containing conserved cysteine residues, a transmembrane domain, and a C-terminus rich in basic amino acids, and were predicted to be expressed on the surface of infected red cells. All eighteen of the rifin genes were in the subtelomeric regions, centromere proximal to the *var* genes. Clusters of rifin genes have been detected on other chromosomes (Cheng *et al.*, 1998), and if the number of rifins found on chromosome 2 is representative of the other chromosomes, the *P. falciparum* genome may contain more than 500 rifin genes. While the function of the rifins is not known, the extensive sequence diversity of the rifins suggests that, like the *var* gene products, they may be clonally variant. Further studies are underway in a number of laboratories to confirm the subcellular localization of the rifins and to determine their function.

Future prospects

The completion of the first *P. falciparum* chromosome and the rapid progress being made by all three genome centers on the remaining chromosomes (Table 1) suggests that the entire *P. falciparum* genome will be completed within 2-3 years. In fact, it is quite likely that most of the parasite's genes will have been identified within 18-24 months, with the additional time being spent on the closing of gaps in the sequence. Ideally, the completion of the *P. falciparum* genome sequence will be followed by the sequencing of a second *Plasmodium* species so as to provide valuable comparative information. The human parasite *P. vivax* and several rodent malaria parasites used as model systems for vaccine and drug development are currently viewed as candidates for sequencing. In addition, information derived from expressed sequence tag (EST) or genome sequencing projects for other apicomplexa such as *Toxoplasma* (Ajioka *et al.*, 1998) will help to identify parasite-specific metabolic pathways that will be useful for development of new drugs against these organisms. Recent technological advances such as the stable transfection of several *Plasmodium* species (van Dijk *et al.*, 1995; Wu *et al.*, 1995; Crabb and Cowman, 1996; van der Wel *et al.*, 1997) and the ability to knock-out specific genes (Menard *et al.*, 1996; Crabb *et al.*, 1997), and the development of microarray technologies for global measurements of gene expression (Schena *et al.*, 1995), will help in the interpretation of the genome sequence. This is important in view of the fact that less than one-half of all the genes identified on the first *P. falciparum* chromosome to be sequenced could be assigned functional roles. Clearly, there is much exciting research to be done and researchers studying *Plasmodium* and related parasites can look forward to Bloom's post-genomic era of microbe biology.

Acknowledgements

The Malaria Genome Sequencing Project is supported by The Wellcome Trust, the US Department of Defense, The Burroughs Wellcome Fund and the National Institutes of Health. This work

was supported by a supplement to NIH grant R01-AI40125-01; the Naval Medical Research and Development Command work units 61102A.S13.00101.BFX.1431, 612787A.870.00101.EFX.1432, 623002A.810.00101.HFX.1433 and STEP C611102A0101 BCX; Department of the Army Cooperative Agreement DAMD17-98-2-8005; and NIH-NMRC interagency agreement No. Y1AI-6091-01. The opinions and assertions herein are those of the authors and are not to be construed as official or as reflecting the views of the US Navy or Department of the Army.

References

Ajioka JW, Boothroyd JC, *et al.* (1998). Gene discovery by EST sequencing in *Toxoplasma gondii* reveals sequences restricted to the Apicomplexa. Genome Res 8: 18-28.

Baruch DI, Pasloske BL, *et al.* (1995). 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.

Bloom BR (1995). A microbial minimalist. Nature 378: 236.

Bonnefoy S, Bischoff E, *et al.* (1997). Evidence for distinct prototype sequences within the *Plasmodium falciparum* Pf60 multigene family. Mol Biochem Parasitol 87: 1-11.

Bowman S, Lawson D, *et al.* (1999). The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*. Nature (in press).

Bracchi V, Langsley G, *et al.* (1996). PfKIN, an SNF1 type protein kinase of *Plasmodium falciparum* predominantly expressed in gametocytes. Mol Biochem Parasitol 76: 299-303.

Blattner FR, Plunket G, *et al.* (1997). The complete genome sequence of *Escherichia coli* K-12. Science 277: 1453-1474.

Cheng Q, Cloonan N, *et al.* (1998). *stevor* and *rif* are *Plasmodium falciparum* multicopy gene families which potentially encode variant antigens. Mol Biochem Parasitol 97: 161-176.

Cole ST, Brosch R, *et al.* (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 393: 537-544.

Crabb BS, Cooke BM, *et al.* (1997). Targeted gene disruption shows that knobs enable malaria-infected red cells to cytoadhere under physiological shear stress. Cell 89: 287-296.

Crabb BS, Cowman AF (1996). Characterization of promoters and stable transfection by homologous and nonhomologous recombination in *Plasmodium falciparum*. Proc Natl Acad Sci USA 93: 7289-7294.

Cyr DM, Langer T, *et al.* (1994). DnaJ-like proteins: molecular chaperones and specific regulators of Hsp70. Trends Biochem Sci 19: 176-181.

Dame JB, Arnot DE, *et al.* (1996). Current status of the *Plasmodium falciparum* genome project. Mol Biochem Parasitol 79: 1-12.

Devine SE, Boeke JD (1994). Efficient integration of artifical transposons into plasmid targets in vitro: a useful tool for DNA mapping, sequencing, and genetic analysis. Nucleic Acids Res 22: 3765-3772.

Fleischmann RD, Adams MD, *et al.* (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269: 496-512.

Foster J, Thompson J (1995). The *Plasmodium falciparum* genome project: a resource for researchers. Parasitol Today 11: 1-4.

Fraser CM, Casjens S, *et al.* (1997). Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. Nature 390: 580-586.

Fraser CM, Norris SJ, *et al.* (1998). Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. Science 281: 375-388.

Gardner MJ, Tettelin H, *et al.* (1998). Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. Science 282: 1126-1132.

Hoffman SL, Bancroft WH, *et al.* (1997). Funding for malaria genome sequencing. Nature 387: 647.

Jing J, Aston C, *et al.* (1999). Optical mapping of *Plasmodium falciparum* chromosome 2. Genome Res 9: 175-181.

Lanzer M, de Bruin D, *et al.* (1993). Transcriptional differences in polymorphic and conserved domains of a completed cloned *P. falciparum* chromosome. Nature 361: 654-657.

Menard R, *et al.* (1996). Circumsporozoite protein is required for development of malaria sporozoites in mosquitos. Nature 385: 336-340.

Mewes HW, Albermann K, *et al.* (1997). Overview of the yeast genome. Nature 387: 7S-65S.

Rowe JA, Moulds JM, *et al.* (1997). *P. falciparum* rosetting mediated by a parasite-variant erythrocyte membrane protein and complement-receptor 1. Nature 388: 292-295.

Rubio JP, Thompson JK, *et al.* (1996). The *var* genes of *Plasmodium falciparum* are located in the subtelomeric region of most chromosomes. EMBO J 15: 4069-4077.

Salzberg SL, Pertea M, *et al.* (1999). Interpolated Markov models for eukaryotic gene finding. Genomics 59: 24-31.

Schena M, Shalon D, *et al.* (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270: 467-470.

Slabas AR, Fawcett T (1992). The biochemistry and molecular biology of plant lipid biosynthesis. Plant Mol Biol 19: 169-191.

Smith JD, Chitnis CE, *et al.* (1995). Switches in expression of *Plasmodium falciparum var* genes correlate with changes in antigenic and cytoadherent phenotypes of infected erythrocytes. Cell 82: 101-110.

Soldati D (1999). The apicoplast as a potential therapeutic target in *Toxoplasma* and other apicomplexan parasites. Parasitol Today 15: 5-7.

Stephens RS, Kalman S, *et al.* (1998). Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. Science 282: 754-759.

Su Z, Heatwole VM, *et al.* (1995). The large diverse gene family *var* encodes proteins involved in cytoadherence and antigenic variation of *Plasmodium falciparum*-infected erythrocytes. Cell 82: 89-100.

Sutton GS, White O, *et al.* (1995). TIGR Assembler: a new tool for assembling large shotgun sequencing projects. Genome Sci Technol 1: 9-19.

Tomb JF, White O, *et al.* (1997). The complete genome sequence of the gastric pathogen *Helicobacter pylori*. Nature 388: 539-547.

van der Wel AM, Tomas AM, *et al.* (1997). Transfection of the primate malaria parasite *Plasmodium knowlesi* using entirely heterologous constructs. J Exp Med 185: 1499-1503.

van Dijk MR, Waters AP, *et al.* (1995). Stable transfection of malaria parasite blood stages. Science 268: 1358-1362.

Waller RF, Keeling PJ, *et al.* (1998). Nuclear-encoded proteins target to the plastid in *Toxoplasma gondii* and *Plasmodium falciparum*. Proc Natl Acad Sci USA 95: 12352-12357.

Walliker D, Quayki I, *et al.* (1987). Genetic analysis of the human malaria parasite *Plasmodium falciparum*. Science 236: 1661-1666.

Weber JL (1988). Interspersed repetitive DNA from *Plasmodium falciparum*. Mol Biochem Parasitol 29: 117-124.

Wilson RJM, Gardner MJ, *et al.* (1991). Have malaria parasites three genomes? Parasitol Today 7: 134-136.

Wootton JC and Federhen S (1996). Analysis of compositionally biased regions in sequence databases. Methods Enzymol 266: 554-571.

Wu Y, Sifri CD, *et al.* (1995). Transfection of *Plasmodium falciparum* within human red cells. Proc Natl Acad Sci USA 92: 973-977.

Appendix K
Final Report
DAMD17-82-2-8005

The genome of the malaria parasite

Malcolm J Gardner

The genome of the human malaria parasite *Plasmodium falciparum* is being sequenced by an international consortium. Two of the parasite's 14 chromosomes have been completed and several other chromosomes are nearly finished. Even at this early stage of the project, analysis of the genome sequence has provided promising new leads for drug and vaccine development.

Addresses
The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850, USA; e-mail: gardner@tigr.org

Current Opinion in Genetics & Development 1999, **9**:704–708

0959-437X/99/$ – see front matter

Abbreviations

CTL	cytotoxic T lymphocyte
EST	expressed sequence tag
GST	gene sequence tag
STS	sequence-tagged site
YAC	yeast artificial chromosome

Introduction

Over one-third of the world's population is at risk of contracting malaria, a mosquito-borne disease caused by apicomplexan parasites of the genus *Plasmodium*. There are ~300–500 million new cases and ~1.5–2.7 million deaths from malaria annually. Most deaths due to malaria occur among children in sub-Saharan Africa [1]. At present there is no effective, practical vaccine that can be used to prevent malaria, and although there are effective antimalarial drugs, resistance to one of more of these drugs has developed in many parts of the world. Development of new drugs and vaccines has been only moderately successful, limited by the financial resources that are available and the difficulty of working with a complex intracellular parasite. (A comprehensive collection of review articles on all aspects of *Plasmodium* biology can be found here [2•].)

Completion of the first microbial genome sequences demonstrated the benefits that accrue from genome sequencing [3]. For a pathogenic organism, the genome sequence provides the sequence of every potential drug or vaccine target; for difficult to study organisms like *Plasmodium*, sequencing of the genome may be the only way to identify these targets. The *Plasmodium falciparum* genome is approximately 28 megabase pairs (Mb) in length and contains 14 chromosomes ranging in size from ~0.6–3.4 Mb. Chromosome sizes can vary markedly between wild isolates as a result of recombination events involving the repeat-rich subtelomeric regions of the chromosome. The genome is extremely A+T rich (~80%), which might account for the instability of large fragments of *P. falciparum* DNA in *E. coli*. The DNA is more stable in yeast; large insert yeast artificial chromosome (YAC) libraries have been constructed and used to generate STS (sequence-tagged site) maps of most of the chromosomes [4]. In addition, a linkage map of the genome consisting of more than 900 microsatellite markers and having a resolution of 30 kb has been produced [5••]. Expressed sequence tags (ESTs) from blood stage parasites and gene sequence tags (GSTs) have also been prepared [6,7]. Techniques for manipulation of the genome have been developed including stable transfection and gene knockouts [8••]. This review summarizes recent progress in the sequencing of the *P. falciparum* genome, and outlines how the genome sequence information produced in this effort is contributing to the development of new drugs and vaccines against malaria.

The *Plasmodium falciparum* genome sequencing project

P. falciparum is the most lethal of the four *Plasmodium* species that cause malaria in humans. Fortunately, all stages of the *P. falciparum* life cycle can be maintained in the laboratory, blood stages can be cultured routinely, and cloned parasites are available. In late 1996, a consortium of funding agencies, genome centers, and malaria investigators was formed to sequence the *Plasmodium falciparum* genome [9,10]. A strategy was adopted whereby individual chromosomes assigned to each genome center were resolved by pulsed field gel electrophoresis and subjected to shotgun sequencing. STS markers [4], the microsatellite linkage map [5••,11], and optical restriction maps [12••,13] of the chromosomes were used for ordering of the contiguous sequences during the gap closure phase and for verification of the final sequence assembly. Chromosomes 2 and 3, which comprise about 7% of the genome, have been completed [14••,15••]; preliminary data at various stages of completion are available for the remaining chromosomes (Table 1). One difficulty faced by the sequencing groups was the identification of genes in the A+T-rich sequence. Gene finding algorithms developed for higher eukaryotes, which have a much lower gene density than *Plasmodium*, were not optimal for the prediction of coding regions in *Plasmodium* DNA, and prokaryotic gene finders were unable to predict introns. GlimmerM gene finding software was developed during the chromosome 2 project; it uses interpolated Markov models constructed from a training set of well-characterized genes for prediction of coding regions and a separate module for prediction of splice sites [16].

The chromosome sequences revealed that 20–30 kb of each chromosome end was composed of telomeric, rep20, and other repeats [14••,15••]. Centromeric to these repeats, members of multigene families involved in antigenic variation and or pathogenesis were found [17•], including *var* genes that encode the PfEMP1 proteins [18–21], open reading frames with similarity to the 3′ exon of *var* genes

Table 1

Web sites related to the malaria genome sequencing project.

Web site	Content	URL
P. falciparum chromosomes 2, 10, 11, 14, TIGR/Naval Medical Research Center	Chromosome 2 annotation [14••] Preliminary data	http://www.tigr.org/tdb/mdb/pfdb/pfdb.html
P. falciparum chromosomes 1, 3, 4, 5–9, 13, The Sanger Centre	Chromosome 3 annotation [15••] Preliminary data	http://www.sanger.ac.uk/Projects/P_falciparum/
P. falciparum chromosome 12, Stanford University	Preliminary data for chromosome 12	http://sequence www.stanford.edu/group/malaria/index.html
P. falciparum Gene Sequence, Tag Project University of Florida	A collection of ESTs and GSTs for *P. falciparum* [6,7]	http://parasite.arf.ufl.edu/malaria.html
Malaria Database, Monash University, Walter and Eliza Hall Institute	A collection of genetic information on malaria parasites	http://www.wehi.edu.au/MalDB-www/who.html
Malaria Genetics and Genomics, National Center for Biotechnology Information (NCBI)	BLAST searches on Apicomplexan sequence data, including *P. falciparum*; *P. falciparum* linkage maps, etc.	http://www.ncbi.nlm.nih.gov/Malaria/
Parasite Genomes Blast Server, European Bioinformatics Institute	BLAST searches on sequence data from many parasites, including *Plasmodium*	http://www.embl-ebi.ac.uk/parasites/parasite_blast_server.html
Malaria Foundation	General information on malaria and many links to malaria-related sites	http://www.malaria.org/index.htm
TIGR Microbial Database	A comprehensive listing of microbial genome projects	http://www.tigr.org/tdb/mdb/mdb.html

BLAST, basic local alignment search tool; dbEST, database of expressed sequence tags; GSTs, genome sequence tags.

that may represent a distinct gene family [22], and members of the *rif* and *STEVOR* gene families (see below). Gene density was about 1 gene per 4.7 kb and almost one-half of genes were predicted to contain introns. Depending upon the methods used for annotation, up to two-thirds of the genes identified had no detectable orthologs in other organisms, which suggests that our current understanding of malaria parasite biology is woefully incomplete.

The investment in sequencing of the genome has already paid handsome dividends. A large gene family (*rif*) was identified on chromosome 2 [14••] (the *STEVOR* family was proposed to be a family related to, but distinct from, the *rif* family [23••]). The *rif* genes encoded polypeptides of 27–35 kDa (rifins) that were predicted to be located on the red cell surface and which contained a region variable in length and amino acid sequence. The sequence polymorphism of the rifins, their presumed cell surface localization, and the distribution of the *rif* genes in subtelomeric regions near the *var* genes suggested that rifins might be a new class of variant surface antigen. Laboratory studies have now proven that the rifins are expressed on the surface of the infected red cell and that they are clonally-variant but the function of these proteins has not been determined [24••]. Like the PfEMP1 proteins encoded by the *var* gene family, which mediate cytoadherence and rosetting, the rifins might have a role in host–parasite interaction.

Other major findings included the discovery of genes encoding enzymes of the type II fatty acid biosynthetic pathway that were previously found only in plants and bacteria [14••,25••], a cluster of four genes of unknown function that was repeated on one end of both chromosomes 2 and 3, and the identification of putative centromere sequences [15••]. The predicted centromeres (~2–3 kb in length) were located in the most A+T rich region of each chromosome (>97% A+T), which in both cases were under represented in the plasmid shotgun libraries used for sequencing and were the most difficult regions to sequence. Proof that these regions actually are centromeres awaits improvements in transfection technology; however, if these are centromeres they could be useful for the stable maintenance of minichromosomes in transfected parasites.

Identification of new chemotherapeutic targets using the genome sequence

Investigation of a 35 kb extra-chromosomal DNA with features characteristic of plastid DNA by Wilson and

colleagues [26] led to the identification of an organelle in *Plasmodium*, *Toxoplasma*, and related parasites called the apicoplast [27–30]. Early studies revealed that organellar protein synthesis and DNA replication were targets of antibiotics with antimalarial activity (for reviews see [31,32]). Analysis of the complete sequence of the 35 kb DNA provided few clues to the function of the organelle but, like plastids of higher plants, the organelle was hypothesized to contain biochemical pathways essential for cell survival. If such pathways were parasite specific they would make attractive drug targets.

Because most proteins in the plastids of other organisms are encoded by nuclear DNA and imported into plastids, it was clear that the genes encoding the enzymes of these pathways were to be found in the nuclear genome. Analysis of the genome sequence in conjunction with transfection studies in *Toxoplasma* have led to the identification of nuclear-encoded proteins that are imported into the apicoplast and the amino-terminal sequences that direct the transport of these proteins into the organelle [25••]. The *fab*H gene encoding 3-ketoacyl acyl carrier protein synthase III — an enzyme involved in type II fatty acid biosynthesis — was identified on chromosome 2 in *P. falciparum* and shown to contain a putative apicoplast-targeting peptide. The antibiotic thiolactomycin, which inhibits the orthologous bacterial enzyme, was shown to possess growth-inhibitory activity against *P. falciparum in vitro* [25••].

Most recently, genes encoding enzymes of the non-mevalonate pathway of isoprenoid biosynthesis were identified in preliminary data from the chromosome 14 sequencing project and the enzymes were predicted to be localized in the apicoplast [33••,34]). Inhibitors of one enzyme in the pathway (1-deoxy-D-xylulose 5-phosphate reductoisomerase) were found to inhibit the activity of the recombinant enzyme expressed in bacteria and to possess antiparasite activity *in vitro* and *in vivo*. These examples validate the early interest in plastid-localized pathways as drug targets, and demonstrate the rapidity with which potential drug targets can be identified with genome sequence information.

Investigators searching for new drug targets have also found plant-like biochemical pathways in apicomplexan parasites that may not be located in the apicoplast (e.g. the shikimate pathway) using more conventional approaches [35•,36]. Other potential targets not related to the apicoplast have also been identified via gene sequence information [37]. As the sequencing of the genome proceeds it will be possible to construct an increasingly comprehensive view of parasite metabolism (the 'metabolome'), which should permit the identification of many more novel drug targets. Successful exploitation of these novel targets may reduce reliance on current antimalarials to which resistance has developed and permit the development of multi-drug therapies that may slow the development of resistance in the future.

Genome sequence and vaccine development

The *P. falciparum* genome sequence will also provide the amino acid sequences of all potential vaccine antigens. Characterization of the hundreds or thousands of antigens to be identified from the genome sequence and their formulation into effective vaccines will be a formidable task — one made more difficult by the requirement that each vaccine must elicit the appropriate immune response for targeting of the different stages of the parasite life cycle [38,39]. One proposed approach is to clone individual *P. falciparum* genes or long open reading frames into DNA vaccines, generate antisera to the encoded proteins in mice, and use immunofluorescence assays to determine the expression patterns and subcellular localization of the candidate antigens in the parasite [40••]. Antigens expressed only within infected hepatocytes, which are targeted primarily by CD8+ T cell responses, could be screened via computer algorithms to predict cytotoxic T lymphocyte (CTL) epitopes. The CTL epitopes could be combined into a series of multi-epitope DNA vaccine constructs and multicomponent DNA vaccines encoding many full-length liver stage antigens could also be prepared. Blood stage antigens accessible to antibodies could also be formulated into DNA vaccines. Clinical trials to establish immunogenicity and protective efficacy of the vaccines would follow. Pilot projects using genes from the two completed chromosomes could be used to validate this approach prior to its application on a large scale. Other approaches to the use of genome data for vaccine development are also possible, including scaling-up of the current antigen-by-antigen strategy using rodent malaria orthologs to *P. falciparum* antigens, or targeted expression library immunization techniques [41].

Comparative genomics

Four species of *Plasmodium* are currently known to infect humans. *P. falciparum* is by far the most lethal of the four species, but *P. vivax*, *P. malariae*, and *P. ovale* cause significant morbidity. *P. vivax* is the most prevalent of these and is of increasing concern because of the development of chloroquine resistance. Apart from the sequencing of genes encoding potential vaccine antigens and drug targets, comparatively little molecular biology has been done with these parasites, primarily because they are extremely difficult or impossible to culture continuously *in vitro* [42] and must be maintained in primates. Carlton *et al.* [43•] have produced karyotype maps of the three other human *Plasmodia*. Like *P. falciparum*, these species appear to have 14 chromosomes but their genomes may be 10–15 Mb larger than the *P. falciparum* genome, possibly as a result of differences in the amount of subtelomeric non-coding DNA. Four synteny groups common to all four species were identified, which suggests that gene order has been preserved across species in many cases. Because *P. vivax* is the second most important human malaria and exhibits numerous biological characteristics that differ from *P. falciparum*, it is quite likely that the *P. vivax* genome will be sequenced; an EST gene discovery project has already

been initiated. Comparison of the *P. falciparum* and *P. vivax* genomes should enable the identification of genes responsible for the biological and pathogenicity differences between the two species. In addition, sequence data from murine *Plasmodia* and related parasites such as *Toxoplasma* (Table 1) and *Theileria* [44•] will help to define apicomplexan specific genes.

Conclusions

Tremendous progress towards an understanding of *Plasmodium* biology has been made over the past decade. We can expect the rate of progress to increase in the next decade once the complete genome sequence of *P. falciparum* is determined. This information, coupled with improvements in areas such as informatics, transfection technology, and the development of oligonucleotide [45] and glass slide microarrays [46] for examination of gene expression on a genome-wide scale, will allow investigators to delve into areas of *Plasmodium* biology that are so far unexplored. These discoveries will provide a much more complete picture of malaria parasite biology and facilitate the development of new drugs and vaccines to combat malaria.

Note added in proof

An important new work on *P. falciparum* restriction mapping has just been published [48••].

Acknowledgements

I thank my colleagues at The Institute for Genomic Research (TIGR) and the Naval Medical Research Center (NMRC) for their support. Sequencing of the *P. falciparum* genome at TIGR and the NMRC is supported by the National Institutes of Health, the Burroughs Wellcome Fund, and the Departments of the Navy and Army.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1 World Health Organization: **World malaria situation in 1994: population at risk.** *Wkly Epidemiol Rec* 1997, **72**:269-276.

2. • Sherman IW (Ed): *Malaria Parasite Biology, Pathogenesis, and Protection.* Washington, D.C.: ASM Press; 1998.
This book contains a collection of reviews on most aspects of malaria parasite biology and research.

3. Clayton RA, White O, Fraser CM: **Findings emerging from complete microbial genome sequences.** *Curr Opin Microbiol* 1998, **1**:562-566.

4. Dame JB, Arnot DE, Bourke PF, Chakrabarti D, Christodoulou Z, Coppel RL, Cowman AF, Craig AG, Fischer K, Foster J *et al.*: **Current status of the *Plasmodium falciparum* genome project.** *Mol Biochem Parasitol* 1996, **79**:1-12.

5. •• Su XZ, Wellems TE: ***Plasmodium falciparum*: assignment of microsatellite markers to chromosomes by PFG-PCR.** *Exp Parasitol* 1999, **91**:367-369.
A description of the *P. falciparum* linkage map produced using microsatellite markers covering most of the genome.

6. Reddy GR, Chakrabarti D, Schuster SM, Ferl RJ, Almira EC, Dame JB: **Gene sequence tags from *Plasmodium falciparum* genomic DNA fragments prepared by the 'genease' activity of mung bean nuclease.** *Proc Natl Acad Sci USA* 1993, **90**:9867-9871.

7. Chakrabarti D, Reddy GR, Dame JB, Almira EC, Laipis PJ, Ferl RJ, Yang TP, Rowe TC, Schuster SM: **Analysis of expressed sequence tags from *Plasmodium falciparum.*** *Mol Biochem Parasitol* 1994, **66**:97-104.

8. •• Wellems TE, Su X, Ferdig M, Fidock DA: **Genome projects, genetic analysis, and the changing landscape of malaria research.** *Curr Opin Microbiol* 1999, **2**:415-419.
A review article from one of the leading malaria research groups providing their view of the anticipated impact of recent technological developments, including genome sequencing, on research leading to new drugs and vaccines against malaria.

9. Butler D: **Funding assured for international malaria sequencing project.** *Nature* 1997, **388**:701.

10. Hoffman SL, Bancroft WH, Gottlieb M, James SL, Bond EC, Stephenson JR, Morgan MJ: **Funding for malaria genome sequencing.** *Nature* 1997, **387**:647.

11. Su XZ, Wellems TE: **Toward a high-resolution *Plasmodium falciparum* linkage map: polymorphic markers from hundreds of simple sequence repeats.** *Genomics* 1996, **33**:430-444.

12. •• Jing J, Aston C, Zhongwu L, Carucci DJ, Gardner MJ, Venter JC, Schwartz DC: **Optical mapping of *Plasmodium falciparum* chromosome 2.** *Genome Res* 1999, **9**:175-181.
A report describing the rapid generation of restriction maps for an entire chromosome by direct visualization of restriction enzyme digested chromosome fragments on glass slides.

13. Aston C, Mishra B, Schwartz DC: **Optical mapping and its potential for large-scale sequencing projects.** *Trends Biotechnol* 1999, **17**:297-302.

14. •• Gardner MJ, Tettelin H, Carucci DJ, Cummings LM, Aravind L, Koonin EV, Shallom S, Mason T, Yu K, Fujii C *et al.*: **Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum.*** *Science* 1998, **282**:1126-1132.
This article and the following article by Bowman *et al.* [15••] describe the methods used to sequence the first two *P. falciparum* chromosomes and summarize the major findings.

15. •• Bowman S, Lawson D, Basham D, Brown D, Chillingworth T, Churcher CM, Craig A, Davies RM, Devlin K, Feltwell T *et al.*: **The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum.*** *Nature* 1999, **400**:532-538.
See annotation [14••].

16. Salzberg SL, Pertea M, Delcher A, Gardner MJ, Tettelin H: **Interpolated Markov models for eukaryotic gene finding.** *Genomics* 1999, **59**:24-31.

17. • Newbold CI: **Antigenic variation in *Plasmodium falciparum*: mechanisms and consequences.** *Curr Opin Microbiol* 1999, **2**:420-425.
A concise summary of the progress being made towards understanding the process of antigenic variation in malaria parasites.

18. Baruch DI, Pasloske BL, Singh HB, Bi X, Ma XC, Feldman M, Taraschi TF, Howard RJ: **Cloning the *P. falciparum* gene encoding PfEMP1, a malarial variant antigen and adherence receptor on the surface of parasitized human erythrocytes.** *Cell* 1995, **82**:77-87.

19. Su Z, Heatwole VM, Wertheimer SP, Guinet F, Herrfeldt JA, Peterson DS, Ravetch J, Wellems TE: **The large diverse gene family *var* encodes proteins involved in cytoadherence and antigenic variation of *Plasmodium falciparum*-infected erythrocytes.** *Cell* 1995, **82**:89-100.

20. Smith JD, Chitnis CE, Craig AG, Roberts DJ, Hudson-Taylor DE, Peterson DS, Pinches R, Newbold CI, Miller LH: **Switches in expression of *Plasmodium falciparum var* genes correlate with changes in antigenic and cytoadherent phenotypes of infected erythrocytes.** *Cell* 1995, **82**:101-110.

21. Borst P, Bitter W, McCulloch R: **Antigenic variation in malaria.** *Cell* 1995, **82**:1-4.

22. Carcy B, Bonnefoy S, Guillotte M, Le SC, Grellier P, Schrevel J, Fandeur T, Mercereau-Puijalon O: **A large multigene family expressed during the erythrocytic schizogony of *Plasmodium falciparum.*** *Mol Biochem Parasitol* 1994, **68**:221-233.

23. •• Cheng Q, Cloonan N, Fischer K, Thompson J, Waine G, Lanzer M, Saul A: ***stevor* and *rif* are *Plasmodium falciparum* multicopy gene families which potentially encode variant antigens.** *Mol Biochem Parasitol* 1998, **97**:161-176.
A report characterizing novel multigene families encoding proteins involved in antigenic variation.

24. •• Kyes SA, Rowe JA, Kriek N, Newbold CI: **Rifins: a second family of clonally variant proteins expressed on the surface of red cells infected with *Plasmodium falciparum.*** *Proc Natl Acad Sci USA* 1999, **96**:9333-9338.
An article that provides the biological and immunological evidence supporting the classification of the rifins as a novel family of variant surface antigens.

25. Waller RF, Keeling PJ, Donald RGK, Striepen B, Handman E, Lang
•• Unnasch N, Cowman AF, Besra GS, Roos DS, McFadden GI: **Nuclear-encoded proteins target to the plastid in *Toxoplasma gondii* and *Plasmodium falciparum*.** *Proc Natl Acad Sci USA* 1998, **95**:12352-12357.

Several nuclear-encoded proteins from *Toxoplasma* and *Plasmodium* that are transported into the apicoplast are described and the apicoplast targeting sequences are identified. Also, a new drug target expressed in the apicoplast is described.

26. Wilson RJM, Denny PW, Preiser PR, Rangachari K, Roberts K, Roy A, Whyte A, Strath M, Moore DJ, Moore PW *et al.*: **Complete gene map of the plastid-like DNA of the malaria parasite *Plasmodium falciparum*.** *J Mol Biol* 1996, **261**:155-172.

27. Kohler S, Delwiche CF, Denny PW, Tilney LG, Webster P, Wilson RJM, Palmer JD, Roos DS: **A plastid of probable green algal origin in apicomplexan parasites.** *Science* 1997, **275**:1485-1489.

28. Roos DS, Crawford MJ, Donald RG, Kissinger JC, Klimczak LJ, Striepen B: **Origin, targeting, and function of the apicomplexan plastid.** *Curr Opin Microbiol* 1999, **2**:426-432.

29. Denny P, Preiser P, Williamson D, Wilson I: **Evidence for a single origin of the 35 kb plastid DNA in apicomplexans.** *Protist* 1998, **149**:51-59.

30. Lang-Unnasch N, Reith ME, Munholland J, Barta JR: **Plastids are widespread and ancient in parasites of the phylum Apicomplexa.** *Int J Parasitol* 1998, **28**:1743-1754.

31. Soldati D: **The apicoplast as a potential therapeutic target in *Toxoplasma* and other apicomplexan parasites.** *Parasitol Today* 1999, **15**:5-7.

32. McFadden GI, Roos DS: **Apicomplexan plastids as drug targets.** *Trends Microbiol* 1999, **7**:328-333.

33. Jomaa H, Wiesner J, Sanderbrand S, Altincicek B, Weidemeyer C,
•• Hintz M, I Tr, Eberl M, Zeidler J, Lichtenthaler HK *et al.*: **Inhibitors of the nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs.** *Science* 1999, **285**:1573-1576.

This report documents the discovery, using genome sequence information, of a novel drug target in the apicoplast. This is perhaps the most spectacular example to date of the use of genome sequence data to identify new drug targets in *Plasmodium* and related parasites.

34. Ridley RG: **Planting the seeds of new antimalarial drugs.** *Science* 1999, **285**:1502-1503.

35. Roberts F, Roberts CW, Johnson JJ, Kyle DE, Krell T, Coggins JR,
• Coombs GH, Milhous WK, Tzipori S, Ferguson DJ *et al.*: **Evidence for the shikimate pathway in apicomplexan parasites.** *Nature* 1998, **393**:801-805.

The shikimate pathway of chorismate biosynthesis is found in plants, algae, fungi, and bacteria. Chorismate is essential for folate biosynthesis in these organisms and compounds that inhibit enzymes of the shikimate pathway have antimicrobial and herbicidal properties. This paper demonstrates that there is a functional shikimate pathway in *Plasmodium* and related parasites. Because this pathway is not found in mammals, the enzymes of the shikimate pathway may represent new chemotherapeutic targets for antimalarial drugs.

36. Ridley RG: **Planting new targets for antiparasitic drugs.** *Nat Med* 1998, **4**:894-995.

37. Woodrow CJ, Penny JI, Krishna S: **Intraerythrocytic *Plasmodium falciparum* expresses a high affinity facilitative hexose transporter.** *J Biol Chem* 1999, **274**:7272-7277.

38. Miller LH, Hoffman SL: **Research toward vaccines against malaria.** *Nat Med* 1998, **4**:520-524.

39. Good MF, Doolan DL: **Immune effector mechanisms in malaria.** *Curr Opin Immunol* 1999, **11**:412-419.

40. Hoffman SL, Rogers WO, Carucci DJ, Venter JC: **From genomics to
•• vaccines: malaria as a model system.** *Nat Med* 1998, **4**:1351-1353.

This article points out that new, high-throughput strategies for the identification and testing of potential vaccine targets must be devised if the full benefits from the enormous amounts of sequence data being generated by the Malaria Genome Project are to be realized. It is proposed that DNA vaccine technology can be used to determine the stage-specificity and subcellular localization of proteins predicted from the genome sequence, and that this information can be used to select antigens for vaccine development.

41. Barry MA, Lai WC, Johnston SA: **Protection against mycoplasma infection infection using expression-library immunization.** *Nature* 1995, **377**:632-635.

42. Golenda CF, Li J, Rosenberg R: **Continuous *in vitro* propagation of the malaria parasite *Plasmodium vivax*.** *Proc Natl Acad Sci USA* 1997, **94**:6786-6791.

43. Carlton JM, Galinski MR, Barnwell JW, Dame JB: **Karyotype and
• synteny among the chromosomes of all four species of human malaria parasite.** *Mol Biochem Parasitol* 1999, **101**:23-32.

This article describes the use of pulsed field gel electrophoresis to produce karyotype maps of all four species of malaria parasites that infect humans. All four species appeared to contain 14 chromosomes, but the chromosomes of *P. vivax*, *P. ovale*, and *P. malariae* were found to be larger than those of *P. falciparum*. Four of five synteny groups that are conserved between *P. falciparum* and rodent malarias were conserved in all four human malaria parasites despite the differences in chromosome size.

44. Nene V, Morzaria S, Bishop R: **Organisation and informational
• content of the *Theileria parva* genome.** *Mol Biochem Parasitol* 1998, **95**:1-8.

Theileria parva is a cattle parasite that is transmitted by ticks and is related to *Plasmodium* and *Toxoplasma*. *T. parva* infects the lymphocytes of the host and causes a lymphoproliferative disorder called East Coast Fever. This article reviews current knowledge of the *T. parva* genome, which is one-third the size of the *P. falciparum* genome. Sequencing of *T. parva* genome would assist the characterization of the malaria parasite genome and may also shed light on the mechnisms of *T. parva*-induced host-cell transformation.

45. Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ: **High density synthetic oligonucleotide arrays.** *Nat Genet* 1999, **21**:20-24.

46. Debouck C, Goodfellow PN: **DNA microarrays in drug discovery and development.** *Nat Genet* 1999, **21**:48-50.

47. Ajioka JW, Boothroyd JC, Brunk BP, Hehl A, Hillier L, Manger ID, Marra M, Overton GC, Roos DS, Wan KL *et al.*: **Gene discovery by EST sequencing in *Toxoplasma gondii* reveals sequences restricted to the Apicomplexa.** *Genome Res* 1998, **8**:18-28.

48. Lai, Z, Jing J, Aston C, Clarke V, Apodaca J, Dimalanta ET, Carucci DJ,
•• Gardner MJ, Mishra B, Anantharaman TS *et al.*: **A shotgun optical map of the entire *Plasmodium falciparum* genome.** *Nat Genet* 1999, **23**:309-313.

This is an extension of the work reported by Jing *et al.* [12••], where optical restriction mapping was used to rapidly prepare a restriction map of *P. falciparum* chromosome 2. In this paper, an optical restriction map of the complete *P. falciparum* genome was constructed. Optical maps of *P. falciparum* chromosomes have proven very useful for gap closure and sequence verification. Optical restriction maps may be very useful in the sequencing of *Plasmodium* genomes for which other physical or genetic maps are not available.

Reprint Series
6 November 1998, Volume 282, pp. 1126–1132

SCIENCE

Appendix L

Final Report
DAMD17-82-2-8005

Chromosome 2 Sequence of the Human Malaria Parasite *Plasmodium falciparum*

Malcolm J. Gardner, Hervé Tettelin, Daniel J. Carucci, Leda M. Cummings, L. Aravind, Eugene V. Koonin, Shamira Shallom, Tanya Mason, Kelly Yu, Claire Fujii, James Pederson, Kun Shen, Junping Jing, Christopher Aston, Zhongwu Lai, David C. Schwartz, Mihaela Pertea, Steven Salzberg, Lixin Zhou,* Granger G. Sutton,† Rebecca Clayton, Owen White, Hamilton O. Smith,† Claire M. Fraser, Mark D. Adams,† J. Craig Venter,† and Stephen L. Hoffman‡

Chromosome 2 Sequence of the Human Malaria Parasite *Plasmodium falciparum*

Malcolm J. Gardner, Hervé Tettelin, Daniel J. Carucci, Leda M. Cummings, L. Aravind, Eugene V. Koonin, Shamira Shallom, Tanya Mason, Kelly Yu, Claire Fujii, James Pederson, Kun Shen, Junping Jing, Christopher Aston, Zhongwu Lai, David C. Schwartz, Mihaela Pertea, Steven Salzberg, Lixin Zhou,* Granger G. Sutton,† Rebecca Clayton, Owen White, Hamilton O. Smith,† Claire M. Fraser, Mark D. Adams,† J. Craig Venter,† Stephen L. Hoffman‡

Chromosome 2 of *Plasmodium falciparum* was sequenced; this sequence contains 947,103 base pairs and encodes 210 predicted genes. In comparison with the *Saccharomyces cerevisiae* genome, chromosome 2 has a lower gene density, introns are more frequent, and proteins are markedly enriched in nonglobular domains. A family of surface proteins, rifins, that may play a role in antigenic variation was identified. The complete sequencing of chromosome 2 has shown that sequencing of the A+T-rich *P. falciparum* genome is technically feasible.

Malaria, a disease caused by protozoan parasites of the genus *Plasmodium*, is one of the most dangerous infectious diseases affecting human populations. Approximately 300 million to 500 million people are infected annually, and 1.5 million to 2.7 million lives are lost to malaria each year, with most deaths occurring among children in sub-Saharan Africa (*1*). Of the four species that cause malaria in humans, *P. falciparum* is the greatest cause of morbidity and mortality. The resistance of the malaria parasite to drugs and the resistance of mosquitoes to insecticides have resulted in a resurgence of malaria in many parts of the world and a pressing need for vaccines and new drugs. The identification of new targets for vaccine and drug development is dependent on the expansion of our understanding of parasite biology; this understanding is hampered by the complexity of the parasite life cycle. The sequencing of the *Plasmodium* genome may circumvent many of these difficulties and rapidly increase our knowledge about these parasites.

The *P. falciparum* genome is ~30 Mb in size; has a base composition of 82% A+T; and contains 14 chromosomes, which range from 0.65 to 3.4 Mb. Chromosomes from different wild isolates exhibit extensive size polymorphism. Mapping studies have indicated that the chromosomes contain central domains that are conserved between isolates and polymorphic subtelomeric domains that contain repeated sequences. *P. falciparum* also contains two organellar genomes. The mitochondrial genome is a 5.9-kb, tandemly repeated DNA molecule; a 35-kb circular DNA molecule, which encodes genes that are usually associated with plastid genomes, is located within the apicoplast [an organelle of uncertain function in *Plasmodium* and the related parasite *Toxoplasma* (*2*)].

Chromosome 2 (GenBank accession number AE001362) was sequenced with the shotgun sequencing approach, which was previously used to sequence several microbial genomes (*3, 4*), with modifications to compensate for the A+T richness of *P. falciparum* DNA (*5*). These modifications included the following: the extraction of DNA from agarose under high-salt conditions to prevent the DNA from melting at a high temperature, the avoidance of ultraviolet (UV) light, the use of the "vector plus insert" protocol for library construction, sequencing with dye-terminator chemistry, the use of a reduced extension temperature in polymerase chain reactions (PCRs), and the use of a transposon-insertion method for the closure of gaps that are very rich in AT. The assembly software was also modified to minimize the misassembly of A+T-rich sequences. The complete sequence included portions of both telomeres and had an average redundancy of 11-fold; colinearity of the final sequence and genomic DNA was proven with optical restriction and yeast artificial chromosome (YAC) maps.

Chromosome 2 of *P. falciparum* (clone 3D7) is 947 kb in length and has an overall base composition of 80.2% A+T. The chromosome contains a large central region that encodes single-copy genes and several duplicated genes, subtelomeric regions that contain variant antigen genes (*var*) (*6–8*), repetitive interspersed family (RIF)–1 elements (*9*) and other repeats, and typical eukaryotic telomeres (Fig. 1). The terminal 23-kb portions of the chromosome are noncoding and exhibit 77% identity in opposite orientations. The left and right telomeres consist of tandem repeats of the sequence TT(TC)AGGG (*10*) and total 1141 and 551 nucleotides (nt), respectively. The subtelomeric regions do not exhibit repeat oligomers until ~12 to 20 kb into the chromosome, where rep20 (*11*) (a 21-bp tandem direct repeat found exclusively in these regions) occurs 134 and 96 times in the left and right ends of the chromosome, respectively. The sequence similarity that was observed between the subtelomeric regions supports previous suggestions that recombination between chromosome ends may be one mechanism by which genetic diversity is generated. A region with centromere functions could not be identified on the basis of sequence similarity to *S. cerevisiae* or other eukaryotic centromeres (*12*). However, several regions of up to 12 kb are devoid of large open reading frames (ORFs) and might contain the centromere. Alternatively, centromeric functions may be defined by higher order DNA structures and chromatin-associated protein complexes (*13*).

Two hundred and nine protein-encoding genes and a gene for $tRNA^{Glu}$ (Fig. 1 and Table 1) were predicted (*14*) on chromosome 2, giving a gene density of one gene per 4.5 kb, which is a value between that observed in yeast (one gene per 2 kb) and in *Caenorhabditis elegans* (one gene per 7 kb). Of the 209 protein-encoding genes, 43% contain at least one intron. This percentage is an estimate

M. J. Gardner, H. Tettelin, L. M. Cummings, S. Shallom, T. Mason, K. Yu, C. Fujii, J. Pederson, K. Shen, L. Zhou, G. G. Sutton, R. Clayton, O. White, H. O. Smith, C. M. Fraser, M. D. Adams, J. C. Venter, The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA. D. J. Carucci and S. L. Hoffman, Malaria Program, Naval Medical Research Institute, 12300 Washington Avenue, Rockville, MD 20852, USA. L. Aravind, Department of Biology, Texas A & M University, College Station, TX 70843, USA, and National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA. E. V. Koonin, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA. J. Jing, C. Aston, Z. Lai, D. C. Schwartz, W. M. Keck Laboratory for Biomolecular Imaging, Department of Chemistry, New York University, New York, NY 10003, USA. M. Pertea, Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA. S. Salzberg, The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA, and Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA.

*Present address: ARIAD Pharmaceuticals, 26 Landsdowne Street, Cambridge, MA 02139, USA.
†Present address: Celera Genomics, 45 West Gude Drive, Rockville, MD 20850, USA.
‡To whom correspondence should be addressed. E-mail: hoffmans@nmripo.nmri.nnmc.navy.mil

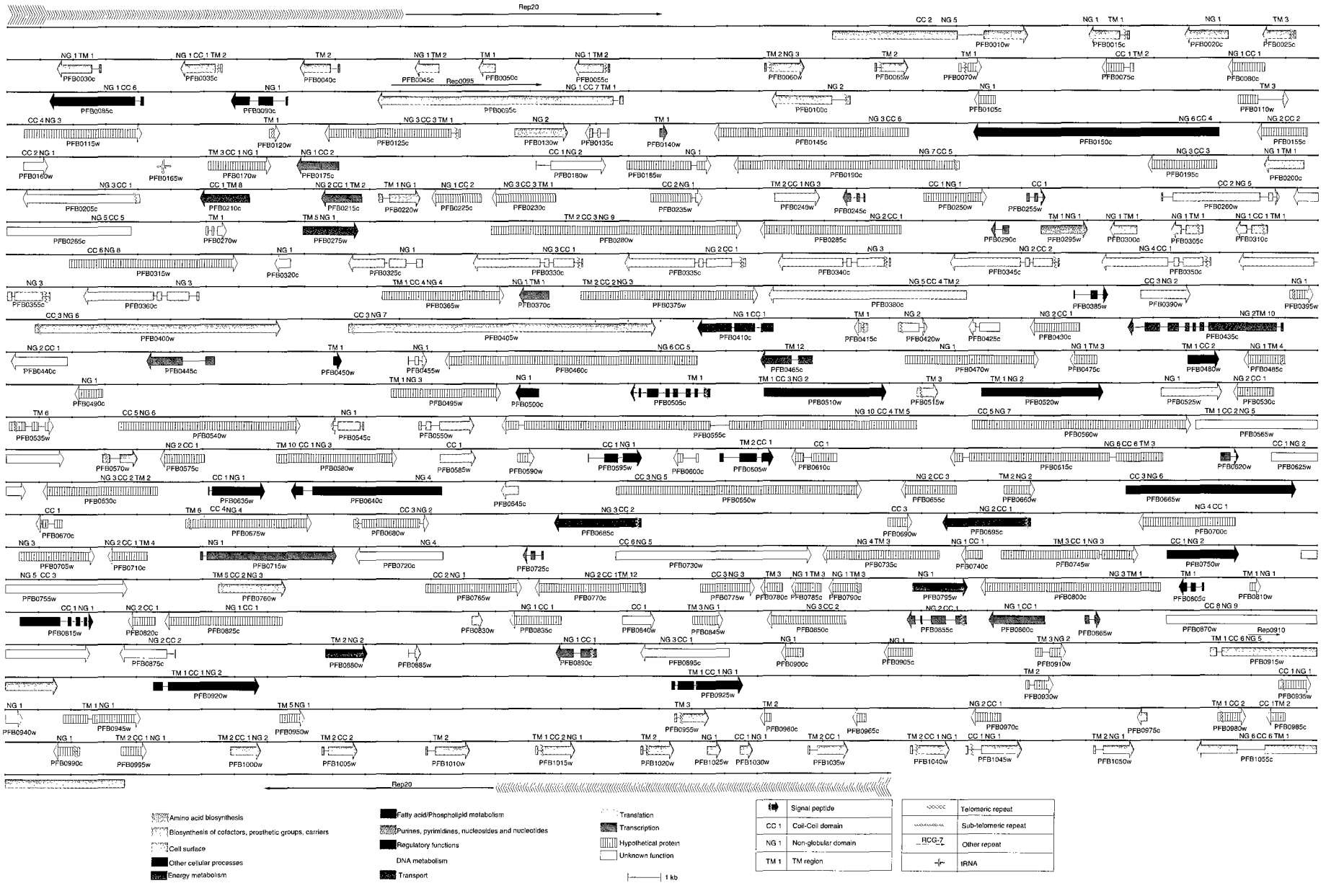


Fig. 1. Gene map of *P. falciparum* chromosome 2. Predicted coding regions are shown on each strand. Exons of protein-encoding genes are indicated by rectangles, and lines linking rectangles represent introns. The single $tRNA^{Glu}$ gene is indicated by a cloverleaf structure. Genes are color-coded according to broad role categories as shown in the key. Genes identification numbers correspond to PF numbers in Table 2. The letters CC, NG, and TM followed by numerals indicate the number of predicted coiled-coil, nonglobular, and transmembrane domains in the proteins, respectively.

because some introns may have been missed by the gene-finding method. Most spliced genes consist of two or three exons. In terms of intron content and gene density, the *Plasmodium* genome, which was assessed by the analysis of the first completed chromosome sequence, appears to be intermediate between the condensed yeast genome and the intron-rich genomes of multicellular eukaryotes.

The proteins encoded in chromosome 2 (Table 2) fall into the following three categories: (i) 72 proteins (34%) are conserved in other genera and contain one or more distinct globular domains; (ii) 47 proteins (23%) belong to *Plasmodium*-specific families with identifiable structural features and, in some cases, known functions; and (iii) 90 predicted proteins (43%) have no detectable homologs, although many contain structural features such as signal peptides and transmembrane domains. Homologs outside *Plasmodium* were detected for 87 (42%) of the 209 predicted proteins. These include proteins in the first category, in addition to those proteins in the second category that possess a conserved domain or domains that are arranged in a manner unique to *Plasmodium*. The percentage of evolutionarily conserved proteins is about two times lower than that found for other genomes, mainly because most of the remaining proteins were predicted to consist primarily of nonglobular domains (*15*) (Table 1). The abundance of nonglobular domains in *Plasmodium* proteins is very unusual; the proportion of proteins with predicted large nonglobular domains in other eukaryotes, such as *S. cerevisiae* (Table 1) or *C. elegans* (*16*), is approximately half that observed in *Plasmodium*. Furthermore, 13 of the 87 conserved proteins on chromosome 2 appear to contain large nonglobular structures (>30 amino acids) that are inserted directly into globular domains, as determined by alignment with homologs from other species.

To determine whether nonglobular domains and proteins are expressed in *P. falciparum*, we performed a reverse transcriptase (RT)–PCR on 11 nonglobular domains and on two genes that encoded predominantly nonglobular proteins, using total blood-stage RNA as a template. In all cases, RT-PCR products were the same size as those that were amplified from genomic DNA, and the sequence of RT-PCR products matched the genomic DNA sequence (*17*). Thus, it is likely that most, if not all, predicted nonglobular domains in chromosome 2 genes are expressed. One example of the insertion of a nonglobular domain into a well-defined globular domain is seen in a protein containing a 5′-3′ exonuclease (Fig. 2). The alignment of the *Plasmodium* sequence with four bacterial exonucleases revealed a 176–amino acid insertion in a region between a strand and a helix in the three-dimensional structure of this protein (*18*). This suggests that eukaryotic proteins can accommodate inserts that may be excluded from the protein core folding without impairing the protein function. The propagation of nonglobular domains in *Plasmodium* suggests that such proteins provide specific selective advantages to the parasite. A structural analysis of *Plasmodium* proteins that contain nonglobular inserts may be valuable for understanding the general principles of protein folding.

Of the 87 conserved proteins that are encoded on chromosome 2, 71 (83%) show the greatest similarity to eukaryotic homologs (Table 2). In contrast, the remaining 16 proteins are most similar to bacterial proteins, and 4 of these represent the first eukaryotic members of protein families that have previously been seen only in bacteria. At least some of these 16 genes may have been transferred to the nuclear genome from an organellar genome after the divergence of the phylum Apicomplexa from other eukaryotic lineages. Several of these proteins appear to contain NH_2-terminal organellar import peptides (*19*) and may function within the apicoplast or the mitochondrion. One such gene encodes 3-ketoacyl–acyl carrier protein (ACP) synthase III (FabH), which catalyzes the condensation of acetyl–coenzyme A and malonyl-ACP in type II (dissociated) fatty acid synthase systems. Type II synthase systems are restricted to bacteria and the plastids of plants, confirming previous hypotheses that the *Plasmodium* apicoplast contains metabolic pathways that are distinct from those of the host (*20*, *21*).

Because the phylum Apicomplexa represents a deep branch in the eukaryotic tree, the presence of eukaryotic-specific genes in *P. falciparum* suggests the appearance of these genes early in eukaryotic evolution. Most of these genes code for proteins that are involved in DNA replication, repair, transcription, or translation (Table 2) and include the origin recognition complex subunit 5, excision repair proteins ERCC1 and RAD2, and proteins involved in chromatin dynamics (such as the BRAHMA helicase, an ortholog of the DRING protein containing the RING finger domain, and chromatin protein SNW1). Furthermore, several eukaryotic proteins involved in secretion are encoded in chromosome 2 (such as the SEC61 γ subunit, the coated pit coatamer subunit, and syntaxin), suggesting an early emergence of the eukaryotic secretory system.

Proteins of the DnaJ superfamily act as cofactors for HSP70-type molecular chaperones and participate in protein folding and trafficking, complex assembly, organelle biogenesis, and initiation of translation (*22*). Five proteins containing DnaJ domains are present on chromosome 2, which suggests multiple roles for this domain in the *Plasmodium* life cycle. Two of these proteins consist primarily of the DnaJ domain, whereas three of the five proteins also contain a large nonglobular domain. Several proteins containing a DnaJ domain have been detected on other chromosomes, indicating that this is a large gene family in *Plasmodium* (*23*). One of its members, the ring-infected erythrocyte surface antigen, binds to the cytoplasmic side of the erythrocyte membrane, suggesting that DnaJ domains perform chaperone-like functions in the formation of protein complexes at this location (*24*). DnaJ domains in some *P.*

Table 1. Summary of features of *P. falciparum* chromosome 2 (*P. f.* chr 2) and comparison to *S. cerevisiae* chromosome 3 (*S. c.* chr 3). Protein structural features were predicted as described (*14*). ND, not determined. Numbers in parentheses indicate the percentage of the total genes or proteins with the specified properties.

Description	Number	
	P. f. chr 2	*S. c.* chr 3
Chromosome length (kb)	947	315
Percent G+C content	19.7	38.6
Exons	24.3	40.0
Introns	13.3	ND
Kilobases per gene	4.50	1.73
Number of predicted protein-coding regions	209	171
Number of genes with introns (%)	90 (43)	4 (2.2)
tRNA genes	1	10
Class of proteins		
Total	209	171
Secreted (%)	22 (11)	11 (6)
Integral membrane (%)	90 (43)	42 (24)
Integral membrane with multiple predicted transmembrane domains (%)	27 (13)	21 (12)
Containing coiled-coil domains (%)	111 (53)	32 (19)
Containing other large compositionally biased regions with predicted nonglobular structure (%)	155 (74)	71 (41)
Completely nonglobular (%)	17 (8)	6 (3.5)
With detectable homologs in other species	87 (42)	145 (85)

Table 2. Identification of genes on *P. falciparum* chromosome 2. The PF number is the systematic name assigned according to a method adapted from *S. cerevisiae* (*14*). The description contains the name (if known) and prominent features of the gene. The table includes genes with homologs in other species and members of *Plasmodium* gene families. An expanded version of this table with additional information is available on the World Wide Web at www.tigr.org/tdb/mdb/pfdb/pfdb.html. Prt, protein; OO, organellar origin; TP, transit peptide; ATP, adenosine triphosphate; euk., eukaryotic; nt, nucleotide.

PF number	Description
Amino acid biosynthesis	
PFB0200c	Aspartate aminotransferase
Biosynthesis of cofactors, prosthetic groups, and carriers	
PFB0130w	Prenyl transferase
PFB0220w	Ubiquinone biosynthesis methyltransferase
Fatty acid and phospholipid metabolism	
PFB0385w	Acyl-carrier prt
PFB0410c	Phospholipase A2-like a/b fold hydrolase
PFB0505c	3-ketoacyl carrier prt synthase III, FabH (OO, TP)
PFB0685c	ATP-dependent acyl-CoA synthetase (TP)
PFB0695c	ATP-dependent acyl-CoA synthetase (TP)
Purines, pyrimidines, nucleosides, and nucleotides	
PFB0295w	Adenylosuccinate lyase (OO)
DNA metabolism	
PFB0160w	ERCC1-like excision repair prt
PFB0180w	Prt with 5′-3′ exonuclease domain (OO, TP)
PFB0205c	Prt with 5′-3′ exonuclease domain (Kem-1 family)
PFB0265c	RAD2 endonuclease
PFB0440c	Chromatinic RING finger prt, DRING ortholog
PFB0720c	Origin recognition complex subunit 5 (ATPase)
PFB0730w	BRAHMA ortholog (DNA helicase superfamily II)
PFB0840w	Replication factor C, 40-kDa subunit (replication activator)
PFB0875c	Chromatin-binding prt (SKI/SNW family)
PFB0895c	Replication factor C, 140-kDa subunit (ATPase)
Energy metabolism	
PFB0795w	ATP synthase alpha chain
PFB0880w	FAD-dependent oxidoreductase (OO)
Transcription	
PFB0140w	Metal-binding prt (DHHC domain)
PFB0175c	Prt of the MAK16 family
PFB0215c	Prt with Egl-like 3′-5′ exonuclease domain
PFB0245c	RNA polymerase 16-kD subunit, RPB4-like
PFB0255w	RRM-type RNA-binding prt
PFB0290c	Zn-ribbon transcription factor (TFIIS family)
PFB0370c	RNA-binding prt (KH domain)
PFB0445c	eIF-4A–like DEAD family RNA helicase
PFB0620w	YOU2-like small euk. C2C2 Zn finger prt
PFB0715w	DNA-directed RNA polymerase subunit 2
PFB0725c	Meta-binding prt (DHHC domain)
PFB0855c	rRNA methylase (SpoU family) (OO, TP)
PFB0860c	RNA helicase
PFB0865w	Small nuclear ribonucleoprt. (SNRNP family)
PFB0890c	Pseudouridine synthetase (RsuA family); first euk. member (OO)
Translation and post-translational modification	
PFB0165w	tRNA-Glu
PFB0240w	PINT domain prt (proteasomal subunit)
PFB0260w	PSD2-like 26S proteasomal subunit
PFB0325c	SERA antigen/protease with active Cys
PFB0330c	SERA antigen/protease with active Cys
PFB0335c	SERA antigen/protease with active Cys
PFB0340c	SERA antigen/protease with active Ser
PFB0345c	SERA antigen/protease with active Ser
PFB0350c	SERA antigen/protease with active Ser
PFB0355c	SERA antigen/protease with active Ser
PFB0360c	SERA antigen/protease with active Ser
PFB0380c	phosphatase (acid phosphatase family)
PFB0390w	Ribosome releasing factor (OO, TP)
PFB0455w	Ribosomal prt L37A
PFB0515w	Glycosyl transferase (novel euk. family)
PFB0525w	Asparaginyl-tRNA synthetase (OO, TP)
PFB0545c	Ribosomal prt L7/L 12 (OO)
PFB0550w	Euk. peptide chain release factor
PFB0585w	Leu/Phe-tRNA prt transferase, first euk. member (OO)
PFB0645c	Ribosomal prt L13 (OO)
PFB0830w	Ribosomal prt S26
PFB0885w	Ribosomal prt S30
Regulatory functions	
PFB0150c	Ser/Thr prt kinase
PFB0510w	GAF domain prt (cyclic nt signal transduction)
PFB0520w	Novel prt kinase
PFB0605w	Ser/Thr prt kinase
PFB0665w	Ser/Thr prt kinase
PFB0815w	Calcium-dependent prt kinase (C-terminus EF hand)
Transport	
PFB0210c	Monosaccharide transporter
PFB0275w	Membrane transporter
PFB0435c	Predicted amino transporter
PFB0465c	Membrane transporter
Cell surface	
PFB0010w	*var* gene
PFB0015c	Rifin
PFB0020c	*var* gene fragment
PFB0025c	Rifin
PFB0030c	Rifin
PFB0035c	Rifin
PFB0040c	Rifin
PFB0045c	*var* gene fragment
PFB0050c	Rifin pseudogene
PFB0055c	Rifin
PFB0060w	Rifin
PFB0065w	Rifin
PFB0100c	Knob-associated His-rich prt
PFB0300c	Merozoite surface antigen MSP-2
PFB0305c	Merozoite surface antigen MSP-5 (EGF domain)
PFB0310c	Merozoite surface antigen MSP-4 (EGF domain)
PFB0400w	PfS230 paralog (predicted secreted prt)
PFB0405w	Transmission-blocking target antigen PfS230
PFB0570w	Predicted secreted prt (thrombospondin domain)
PFB0760w	Mtn3/RAG1IP-like prt
PFB0915w	RESA-H3 antigen
PFB0955w	Rifin
PFB0975c	*var* gene fragment
PFB1000w	Rifin pseudogene
PFB1005w	Rifin
PFB1010w	Rifin
PFB1015w	Rifin
PFB1020w	Rifin
PFB1025w	*var* gene fragment
PFB1030w	*var* gene fragment
PFB1035w	Rifin
PFB1040w	Rifin
PFB1045w	*var* gene fragment
PFB1050w	Rifin
PFB1055c	*var* gene
Other cellular processes	
PFB0085c	Prt with DnaJ domain (RESA-like)
PFB0090c	Prt with DnaJ domain
PFB0450w	Prt translocation complex, SEC61 γ chain
PFB0480w	Syntaxin
PFB0500c	RAB GTPase
PFB0595w	Prt with DnaJ domain, DNJ1/SIS1 family
PFB0635w	T-complex prt 1 (HSP60 fold superfamily)
PFB0640c	WEB-1 ortholog, WD40
PFB0750w	VPS45-like prt (STXBP/UNC-18/SEC1 family)
PFB0805c	Clathrin coat assembly prt
PFB0920w	Prt with DnaJ domain (RESA-like)
PFB0925w	Prt with DnaJ domain (RESA-like)
Unknown function	
PFB0270w	SLR1419 family prt (OO)
PFB0320c	HesB family prt (possible redox activity, OO, TP)
PFB0420w	YgdB prt first euk. member (OO, TP)
PFB0425c	YMR7 family prt

Fig. 2. Multiple alignment of the predicted 5′-3′ exonuclease (PFB0180w) encoded in chromosome 2 with homologous bacterial exonuclease domains showing the large nonglobular insert in *Plasmodium*. The alignment was constructed with the profile alignment option of CLUSTALW (*34*). The alignment column shading is based on a 100% consensus, which is shown underneath the alignment; **h** indicates hydrophobic residues (A, C, F, I, L, M, V, W, and Y), **u** indicates "tiny" residues (G, A, and S), **o** indicates hydroxy residues (S and T), **c** indicates charged residues (D, E, K, R, and H), and + indicates positively charged residues (K and R) (*35*). The aspartates involved in metal coordination have a red background and inverse type. Secondary structure elements derived from the crystal structure of *Thermus aquaticus* DNA polymerase (*18*) are shown above the alignment (H indicates α helix, and E indicates extended conformation, or β strand). 5′-3′-exo_Aae is a stand-alone exonuclease from *Aquifex aeolicus*, and the remaining bacterial sequences are the NH_2-terminal domains of DNA polymerase I.

falciparum proteins contain substitutions in the His-Pro-Asp signature that is required for interaction with HSP-70–type proteins, which may indicate a modification of the typical chaperone function.

Chromosome 2 contains five protein families that are unique to *Plasmodium* in terms of their distinct domain organization, although three of them contain domains that are conserved in other genera. The genes encoding the *Plasmodium*-specific families are primarily located near the ends of the chromosome. A single *var* gene was identified in each subtelomeric region. The *var* genes encode large transmembrane proteins (PfEMP1) expressed in knobs on the surface of schizont-infected red cells. PfEMP1 proteins exhibit extensive sequence diversity; are clonally variant; and are involved in antigenic variation, cytoadherence, and rosetting (*6*–*8*). In addition to the full-length *var* genes, six small ORFs were identified in the subtelomeric regions that were similar to *var* sequences. Five of these ORFs resembled the *var* exon II cDNAs or the Pf60.1 sequences that were reported previously (*7*, *25*).

The largest *Plasmodium*-specific family found on chromosome 2 encodes proteins that were dubbed rifins, after the RIF-1 repetitive element. RIF-1 contained a 1-kb ORF but no initiation codon, was found on most chromosomes, and was transcribed in late blood-stage parasites (*9*). The function of the RIF-1 element was unknown. Eighteen ORFs with similarities to RIF-1 were found in the subtelomeric regions of chromosome 2, centromeric to the *var* genes. An inspection of the sequence upstream of these ORFs revealed exons encoding signal peptides, which indicated that the RIF-1 elements were actually genes consisting of two exons. These genes encode potential transmembrane proteins of 27 to 35 kD, with an extracellular domain that contains conserved Cys residues that might participate in disulfide bonding, a transmembrane segment, and a short basic COOH-terminus. The extracellular domain also contains a highly variable region (Fig. 3). RT-PCR with schizont RNA showed that one of six rifin genes that were tested was transcribed. The function of the rifins is unknown, but their sequence diversity, predicted cell surface localization, and expression in schizont stages suggest that, like *var* genes, they may be clonally-variant. Multiple rifin genes were detected in the telomeric regions of chromosomes 3 and 14, suggesting that rifin genes have propagated as clusters in the course of *Plasmodium* evolution (*26*). If the number found on chromosome 2 is representative of other chromosomes, there may be 500 or more rifin genes in the *P. falciparum* genome (~7% of all protein-coding genes), making it the most abundant gene family in this organism. The presence of *var* and rifin genes and other ORFs in subtelomeric regions of *P. falciparum* chromosomes confirms that the subtelomeric regions are not transcriptionally silent (*27*).

Another family of membrane-associated proteins, serine repeat antigens (SERAs), contains a papain protease-like domain. A cluster of three SERA genes, which were all transcribed in the same direction (from centromere to telomere), was known to be on chromosome 2 (*28*); at least one SERA has been evaluated for use in blood-stage vaccines. These genes are part of an eight-gene cluster; seven genes have a similar four-exon structure, but the gene at the 3′ end of the cluster contains only three exons. The protease domains in these proteins are unusual because five of the eight contain serine instead of cysteine in the active nucleophile position, suggesting that they are serine proteases with a structure that is typical of cysteine proteases (*29*).

Two proteins (MSP-4 and MSP-5) that contain an epidermal growth factor (EGF) module in their extracellular domains were identified (*30*, *31*). In organisms that are not classified in the animal kingdom, MSP-4,

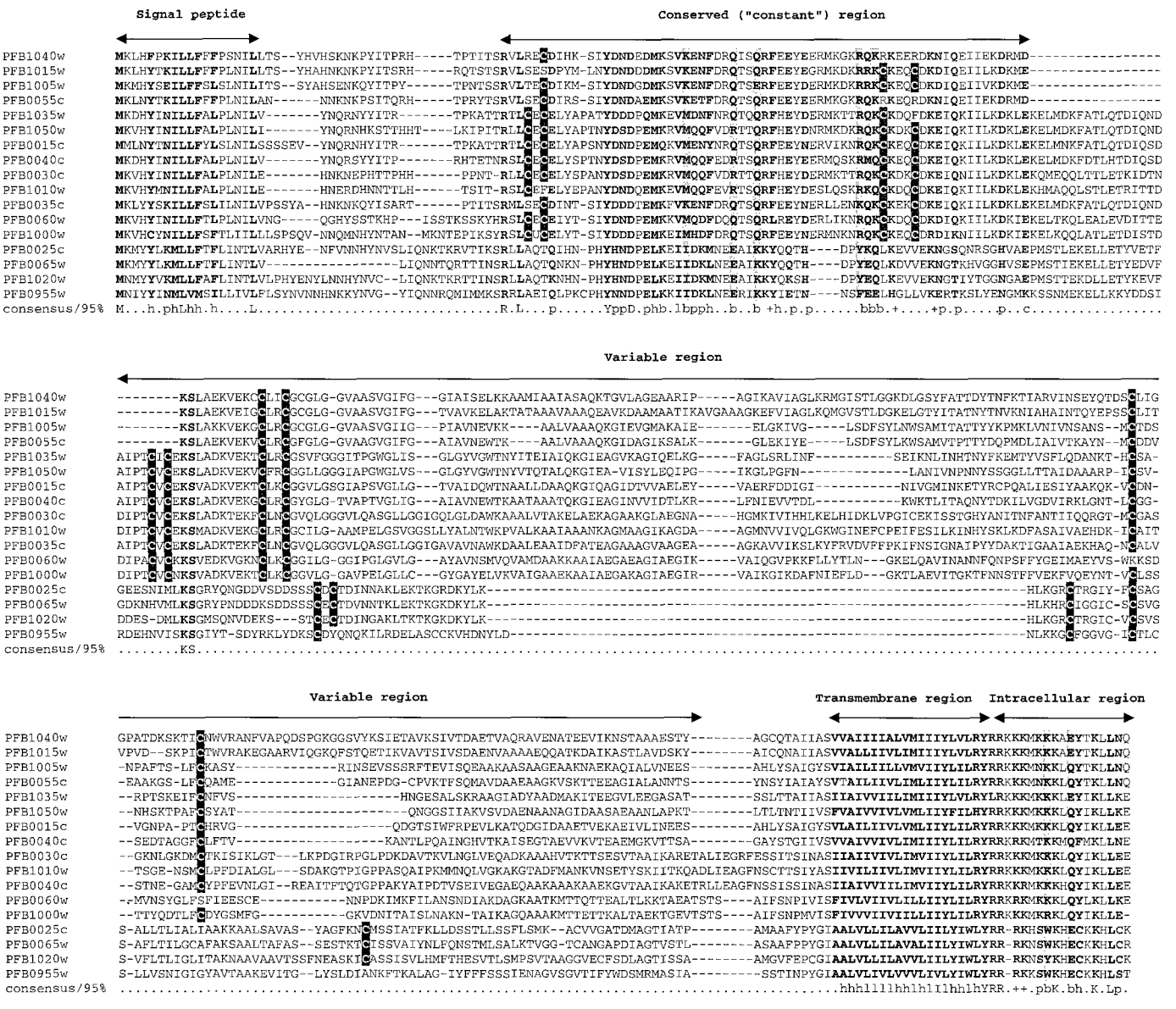


Fig. 3. Multiple sequence alignment of rifins encoded on chromosome 2. The predicted coding regions were aligned with CLUSTALW (*34*) using the default settings. The alignment column shading is based on a 95% consensus, which is shown underneath the alignment; **h** indicates hydrophobic residues (A, C, F, I, L, M, V, W, and Y), **p** indicates polar residues (D, E, H, K, N, Q, R, S, and T), **b** indicates "big" residues (F, I, L, M, V, W, Y, K, R, Q, and E), and + indicates positively charged residues (K and R) (*35*). The cysteines conserved in subsets of rifins are shown by inverse type.

MSP-5, and MSP-1 (a multi-EGF domain protein encoded on chromosome 3) and two *Plasmodium* sexual-stage antigens (*32*) are the only proteins that contain EGF repeats, which suggests that *Plasmodium* obtained the sequence for this domain from its animal host. The plasmodial EGF domains may be involved in parasite adhesion to host cells.

In addition to the families of *Plasmodium*-specific proteins, chromosome 2 contains genes for many secreted and membrane proteins. One of these genes encodes a protein with a modified thrombospondin domain and was transcribed in blood-stage parasites (*17*). Other *Plasmodium* proteins containing thrombospondin domains, such as sporozoite surface protein 2/TRAP and circumsporozoite protein, are involved in the parasitic invasion of host cells (*33*), suggesting that this protein may be involved in the binding of infected red cells to host-cell ligands.

Determination of the first *P. falciparum* chromosome sequence demonstrates that the A+T richness of *P. falciparum* DNA will not prevent the sequencing of the genome. Although technical difficulties not observed during the sequencing of other microbial genomes were encountered, solutions to these problems were found that will facilitate sequencing of the remaining chromosomes. The genome sequence should be of value in the study of *Plasmodium* biology and in the development of new drugs and vaccines for the treatment and prevention of malaria. In addition to these practical benefits, the *Plasmodium* genome sequence should provide broader biological insights, particularly in regard to the plasticity of the eukaryotic genome that is manifest in the preponderance of the predicted nonglobular domains in plasmodial proteins.

References and Notes

1. World Health Organization, *Wkly. Epidemiol. Rec.* **72**, 269 (1997).
2. J. B. Dame *et al.*, *Mol. Biochem. Parasitol.* **79**, 1 (1996); M. Lanzer, D. de Bruin, J. V. Ravetch, *Nature* **361**, 654 (1993); K. Suplick, R. Akella, A. Saul, A. B. Vaidya, *Mol. Biochem. Parasitol.* **30**, 289 (1988); M. J. Gardner, D. H. Williamson, R. J. M. Wilson, *ibid.* **44**, 115 (1991); R. J. M. Wilson *et al.*, *J. Mol. Biol.* **261**, 155 (1996); S. Köhler *et al.*, *Science* **275**, 1485 (1997).
3. R. D. Fleischmann *et al.*, *Science* **269**, 496 (1995).
4. C. M. Fraser *et al.*, *ibid.* **270**, 397 (1995); C. J. Bult *et al.*, *ibid.* **273**, 1058 (1996); C. M. Fraser *et al.*, *Nature* **390**, 580 (1997); J.-F. Tomb *et al.*, *ibid.* **388**, 539

(1997); H. P. Klenk *et al.*, *ibid.* **390**, 364 (1997); C. M. Fraser *et al.*, *Science* **281**, 375 (1998).

5. *P. falciparum* clone 3D7 was selected because it can complete all stages of the life cycle and because 3D7 was used in a genetic cross [D. Walliker *et al.*, *Science* **236**, 1661 (1987)] and in The Wellcome Trust Malaria Genome Mapping Project [J. Foster, J. Thompson, *Parasitol. Today* **11**, 1 (1995)]. Parasites were grown in vitro [W. Trager and W. Jensen, *Nature* **273**, 621 (1978)] and embedded in agarose [D. J. Kemp *et al.*, *ibid.* **315**, 347 (1985)]. Chromosomes were resolved on preparative pulsed-field gels (the process used 1.2% SeaPlaque GTG agarose, a Bio-Rad DRIII apparatus, a 180- to 250-s switch time, a 120° field angle, and 3.7 V/cm for 90 hours at 14°C). Chromosome 2 bands from five gels were adjusted to 0.3 M sodium acetate to prevent melting of the AT-rich DNA and were digested with agarase. The exposure of the DNA to UV light was minimized. A shotgun library of 1- to 2-kb fragments was prepared in pUC18 as described (3), except that treatment with *Escherichia coli* DNA polymerase I was performed (0.5 mM deoxynucleoside triphosphates at 37°C for 10 min) after the second ligation step to close nicks before electroporation into DH10B cells. The gel-purified chromosome 2 DNA was only ~85% pure because of the co-migration of sheared DNA from other chromosomes. To compensate for this ~85% purity and to provide excess coverage to compensate for the possible nonrandomness of the shotgun library, we obtained 23,768 sequences (a coverage of about 10-fold). FS+ dye-terminator chemistry (Perkin-Elmer Applied Biosystems, Foster City, CA) was superior to dye-primer chemistry for the sequencing of AT-rich DNA. Sequences were assembled with The Institute for Genomic Research (TIGR) Assembler [G. S. Sutton, O. White, M. D. Adams, A. R. Kerlavage, *Genome Sci. Tech.* **1**, 9 (1995)], which was modified to assemble A+T-rich sequences. Neighboring contigs were identified with the program GROUPER (A. D. Mays, TIGR, Rockville, MD), and 10 groups of 114 contigs were mapped on the chromosome by comparison to sequence-tagged site (STS) markers [M. Lanzer, D. de Bruin, J. V. Ravetch, *Nature* **361**, 654 (1993)]. The closure of physical and sequence gaps was performed as described (3). Physical gaps were closed by PCR reactions with a genomic DNA template with primers from adjacent mapped groups or with primers from one mapped group and each of the unmapped groups. PCR reactions (Expand Long Template PCR System, Boehringer Mannheim) contained 100 ng of genomic DNA and 15 pmol of each primer (BioServe Biotechnologies, Laurel, MD) in a 50-ml reaction. Cycling conditions (Perkin-Elmer GeneAmp PCR Systems 9600 or 9700) were as follows: 94°C for 2 min; 10 cycles at 94°C for 1 min, at 50° or 55°C for 1 min, and at 60°C for 2 min; 20 cycles at 94°C for 1 min, at 50° or 55°C for 1 min, and at 60°C for 2 min plus 20 s per cycle; and 1 cycle at 60°C for 10 min. PCR products were purified (QIAquick PCR Purification Kit; QIAGEN, Chatsworth, CA) and sequenced with dye-terminator chemistry. Sequence gaps that were too rich in A+T for primer synthesis and walking were closed by the insertion of the artificial transposon AT-2 [S. E. Devine and J. D. Boeke, *Nucleic Acids Res.* **22**, 3765 (1994)] into the plasmid templates that spanned each sequence gap; multiple transposon-containing subclones of each template were sequenced to close the gaps. The coverage criteria were that every position required at least double-clone coverage (or sequence from a PCR product that was amplified from genomic DNA) and either sequence from both strands or coverage with two different sequencing chemistries. The sequence was edited manually with TIGR Editor, and additional sequencing reactions were performed to improve coverage and to resolve sequence ambiguities. To independently confirm the colinearity of the assembled sequence and genomic DNA, we prepared Nhe I and Bam HI optical restriction maps of chromosome 2 DNA [J. Jing *et al.*, in preparation] and compared them with restriction maps that were predicted from the sequence. The relative errors of predicted and observed fragment sizes were 4.3 and 5.8% for the Nhe I and Bam HI maps, respectively, indicating that the assembled sequence was an accurate representation of the chromosome. Further proof of colinearity was obtained by a comparison of the sequence to a scaffold of YAC-end sequences from chromosome 2 YACs that were isolated from a library provided by K. Hinterberg [J. Foster and J. Thompson, *Parasitol. Today* **11**, 1 (1995); L. Cummings *et al.*, in preparation].
6. D. I. Baruch *et al.*, *Cell* **82**, 77 (1995).
7. Z. Su *et al.*, *ibid.*, p. 89.
8. J. D. Smith *et al.*, *ibid.*, p. 101 (1995); J. A. Rowe, J. M. Moulds, C. I. Newbold, L. H. Miller, *Nature* **388**, 292 (1997).
9. J. L. Weber, *Mol. Biochem. Parasitol.* **29**, 117 (1988).
10. K. D. Vernick and T. F. McCutchan, *ibid.* **28**, 85 (1988).
11. P. Oquendo *et al.*, *ibid.* **18**, 89 (1986); J. Pataraporikul and G. Langsley, *Nucleic Acids Res.* **16**, 4331 (1988).
12. S. Saitoh, K. Takahashi, M. Yanagida, *Cell* **90**, 131 (1997); M. M. Smith *et al.*, *Mol. Cell Biol.* **16**, 1017 (1996); M. M. Mahtani and H. F. Willard, *Genome Res.* **8**, 100 (1998); R. D. Shelby, O. Vafa, K. F. Sullivan, *J. Cell Biol.* **136**, 501 (1997); D. du Sart *et al.*, *Nature Genet.* **16**, 144 (1997).
13. J. Lechner and J. Ortiz, *FEBS Lett.* **389**, 70 (1996); A. A. Hyman and P. K. Sorger, *Annu. Rev. Cell Dev. Biol.* **11**, 471 (1995).
14. The nonredundant (NR) protein sequence database at the National Center for Biotechnology Information (NCBI) (NIH, Bethesda, MD) was searched with the gapped BLAST and PSI-BLAST programs. Coding regions were predicted with GlimmerM, a eukaryotic gene-finding program based on Glimmer [S. L. Salzberg, A. L. Delcher, S. Kasif, O. White, *Nucleic Acids Res.* **26**, 544 (1998)], trained on a set of 117 *P. falciparum* sequences. Gene models based on GlimmerM predictions, similarity of ORFs to known proteins, and prediction of putative signal peptides and transmembrane domains were constructed with ANNOTATOR (L. Xhou, TIGR). In cases where a putative gene had no database match and multiple GlimmerM predictions of gene structure, the highest scoring model was reported. After the first set of models was inspected, it was added to the training set, and GlimmerM was retrained. Gene models should be regarded as preliminary until confirmed by other methods. Protein structural features were delineated with the UniPred program of the SEALS package [D. R. Walker and E. V. Koonin, *Ismb* **5**, 333 (1997)]. Signal peptides were predicted with SignalP [H. Nielsen, J. Engelbrecht, S. Brunack, G. von Heijne, *Protein Eng.* **10**, 1 (1997)], and transmembrane helices were predicted with PHThtm [B. Rost, R. Casadio, P. Fariselli, C. Sander, *Protein Sci.* **4**, 521 (1995)]. Coiled-coil domains were predicted with COILS (J. Kuzio, NCBI). Nonglobular structures were predicted with SEG [J. C. Wooton and S. Federhen, *Methods Enzymol.* **266**, 554 (1996)]. Multiple sequence alignments were constructed with CLUSTALW or with the Gibbs-sampling option of the MACAW program [G. D. Schuler, S. F. Altschul, D. J. Lipman, *Proteins* **9**, 180 (1991); A. F. Neuwald, J. S. Liu, C. E. Lawrence, *Protein Sci.* **4**,1618 (1995)]. Transfer RNAs were identified with tRNAscan [T. M. Lowe and S. R. Eddy, *Nucleic Acids Res.* **25**, 955 (1997)]. Systematic gene names based on a scheme for *S. cerevisiae* [H. W. Mewes *et al.*, *Nature* **387** (suppl.), 7 (1997)] were assigned with the convention PF (for *P. falciparum*), a letter for the chromosome (A for chromosome 1, B for chromosome 2, and so forth), a three-digit code ordering the genes from left to right in increments of five (to allow for the addition of new genes), and a letter denoting the coding strand (w or c, for Watson or Crick strand, respectively).
15. The term "nonglobular" refers to proteins or domains of proteins that do not assume compact, folded structures [J. C. Wootton, *Comput. Chem.* **18**, 269 (1994)]. There is a strong inverse correlation between compositional bias in protein sequences and their ability to fold into a compact, globular domain [J. C. Wootton and S. Federhen, *Methods Enzymol.* **266**, 554 (1996)]. Accordingly, the compositional complexity of a sequence can be used to partition it into predicted globular and nonglobular domains. In this analysis, the prediction was performed with the SEG program with the following parameters: window length, 45; trigger complexity, 3.4; and extension complexity, 3.75.
16. L. Aravind and E. Koonin, unpublished data.
17. D. J. Carucci *et al.*, data not shown.
18. Y. Kim *et al.*, *Nature* **376**, 612 (1995).
19. V. Haucke and G. Schatz, *Trends Cell Biol.* **7**, 103 (1997).
20. A. R. Slabas and T. Fawcett, *Plant Mol. Biol.* **19**, 169 (1992); R. J. M. Wilson, M. J. Gardner, J. E. Feagin, D. H. Williamson, *Parasitol. Today* **7**, 134 (1991).
21. After this manuscript was submitted for publication, we learned of work that confirmed the identification of the 3-ketoacyl-ACP synthase III gene in *Plasmodium* and the importation of nuclear-encoded proteins into the apicoplast in the related parasite *Toxoplasma* [R. F. Waller *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 12352 (1998)].
22. D. M. Cyr, T. Langer, M. G. Douglas, *Trends Biochem Sci.* **19**, 176 (1994).
23. L. Aravind *et al.*, data not shown.
24. P. Bork, C. Sander, A. Valencia, B. Bukau, *Trends Biochem. Sci.* **17**, 129 (1992); J. Watanabe, *Mol. Biochem. Parasitol.* **88**, 253 (1997); R. L. Coppel *et al.*, *Nature* **310**, 789 (1984); I. A. Quakyi *et al.*, *Infect. Immun.* **57**, 833 (1989); M. Foley, L. Corcoran, L. Tilley, R. Anders, *Exp. Parasitol.* **79**, 340 (1994).
25. S. Bonnefoy, E. Bischoff, M. Guillotte, O. Mercereau-Puijalon, *Mol. Biochem. Parasitol.* **87**, 1 (1997).
26. Sequence data for *P. falciparum* chromosome 3 was obtained from the Sanger Centre (available at http://www.sanger.ac.uk/Projects/P_falciparum/). Sequencing of *P. falciparum* chromosome 3 was accomplished as part of the Malaria Genome Project Consortium with support by the Wellcome Trust.
27. R. R. Hernandez *et al.*, *Mol. Cell Biol.* **17**, 604 (1997); K. Fischer *et al.*, *ibid.*, p. 3679 (1997).
28. B. Knapp, E. Hundt, U. Nau, H. A. Kupper, *Mol. Biochem. Parasitol.* **32**, 73 (1989); B. Knapp, U. Nau, E. Hundt, H. A. Kupper, *ibid.* **44**, 1 (1991); W. B. Li, D. J. Bzik, T. Horii, J. Inselburg, *ibid.* **33**, 13 (1989); B. A. Fox and D. J. Bzik, *ibid.* **68**, 133 (1994).
29. D. G. Higgins, D. J. McConnell, P. M. Sharp, *Nature* **340**, 604 (1989); A. E. Eakin, J. M. Higaki, J. H. McKerrow, C. S. Craik, *ibid.*, **342**, 132 (1989).
30. V. M. Marshall *et al.*, *Infect. Immun.* **65**, 4460 (1997).
31. V. M. Marshall, W. Tieqiao, R. L. Coppel, *Mol. Biochem. Parasitol.* **94**, 13 (1998).
32. L. Aravind, unpublished observations; M. J. Blackman, I. T. Ling, S. C. Nicholls, A. A. Holder, *Mol. Biochem. Parasitol.* **49**, 29 (1991); D. C. Kaslow *et al.*, *Nature* **333**, 74 (1988); P. E. Duffy, P. Pimenta, D. C. Kaslow, *J. Exp. Med.* **177**, 505 (1993).
33. K. J. Robson *et al.*, *Nature* **335**, 79 (1988); C. Cerami *et al.*, *Cell* **70**, 1021 (1992); W. O. Rogers *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 9176 (1992).
34. J. D. Thompson, D. G. Higgins, T. J. Gibson, *Nucleic Acids Res.* **22**, 4673 (1994).
35. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
36. We thank the members of the Malaria Genome Sequencing Consortium for the open discussion of data during the development of the effort to sequence the *P. falciparum* genome; D. J. Lipman and L. H. Miller for helpful discussions; M. Gottlieb for support and encouragement; A. Craig for providing the 3D7 clone and for suggestions on pulsed-field gel electrophoresis; P. de la Vega for the culturing of parasites; M. Lanzer for providing STS data; K. Hinterberg for providing the 3D7 YAC library; and the TIGR faculty, sequencing core, bioinformatics staff, and systems administrators for expert advice and assistance. This work was supported by a supplement to the National Institute of Allergy and Infectious Diseases grant R01 AI40125-01, which was made possible with funds from NIH's Office for Research on Minority Health; Department of the Army Cooperative Agreement grant DAMD17-98-2-8005 (to J.C.V.); and Naval Medical Research and Development Command Work Units 61102A.S13.00101.BFX1431, 612787A.870.00101.EFX.1432, 623002A.810.00101.HFX.1433, and STEP C611-102A0101BCX. The opinions and assertions herein are those of the authors and are not to be construed as official or as reflecting the views of the U.S. Navy or Department of the Army.

29 June 1998; accepted 29 September 1998

Protist, Vol. 149, 109–112, May 1998 © Gustav Fischer Verlag

PROTIST NEWS

The Malaria Genome Sequencing Project

Malaria is caused by apicomplexan parasites of the genus *Plasmodium*. It is a major public health problem in many parts of the world. In 1994 the World Health Organization estimated that there were 300–500 million cases and up to 2.7 million deaths caused by malaria each year, and because of increased parasite resistance to chloroquine and other antimalarials the situation is expected to worsen considerably (WHO 1997). These dire facts have stimulated efforts to develop an international, coordinated strategy for malaria research and control (Butler et al. 1997). Development of new drugs and vaccines against malaria will undoubtedly be an important factor in control of the disease. However, despite recent progress, drug and vaccine development has been a slow and difficult process, hampered by the complex life cycle of the parasite, a limited number of drug and vaccine targets, and our incomplete understanding of parasite biology and host-parasite interactions.

The advent of microbial genomics, i.e. the ability to sequence and study the entire genomes of microbes, should accelerate the process of drug and vaccine development for microbial pathogens. As pointed out by Bloom, the complete genome sequence provides the "sequence of every virulence determinant, every protein antigen, and every drug target" in an organism (Bloom 1995), and establishes an excellent starting point for this process. Today, the complete genome sequences of 13 microbes have been published, including several human pathogens, and many more microbial genomes are in the works (a listing of microbial genome projects completed or underway can be found at www.tigr.org/mdb).

Two main strategies have been used in these projects. One approach, pioneered at TIGR (Fleischmann et al. 1995), is the whole genome shotgun method, in which a genomic library of sheared 1–2 kb fragments is prepared in a plasmid vector, and clones are picked at random and sequenced. Special software is then used to assemble the overlapping fragments into a contiguous sequence. The whole genome method is dependent on high-quality shotgun libraries and robust software for fragment assembly. The second method, used to sequence the *E. coli* genome, for example, involves sequencing of large-insert clones from cosmid or lambda libraries (Blattner et al. 1997). Although not so dependent upon computational resources as the whole genome shotgun method, sequencing of large-insert clones does require a physical map of the genome to guide selection of the clones to be sequenced.

At first, it was unclear how best to proceed in sequencing the genome of *P. falciparum*, the human malaria parasite responsible for the most morbidity and mortality. The *P. falciparum* genome is about 30 Mb in size, about 8- to 10-fold larger than a typical eubacterial genome, and its size was thought to preclude the whole-genome approach due to the computational limitations inherent in the assembly process, and difficulties in closing gaps that usually persist after assembly. The large-insert library approach was ruled out by the fact that *P. falciparum* has an overall base composition of approximately 82% AT. This unusual base composition is thought responsible for the fact that *P. falciparum* DNA is notoriously unstable in *E. coli*, such that representative large-insert (> 20 kb) genomic libraries in plasmid, lambda, and cosmid vectors that could be used for sequencing cannot be prepared. Yeast artificial chromosome (YAC) libraries of *P. falciparum* (Foster and Thompson 1995) have been constructed, however, and while these appear to stably maintain large inserts, YACs are not very well suited to high-throughput sequencing projects.

Abbreviations: EBI: European Bioinformatics Institute; EST: expressed sequence tag; DoD: US Department of Defense; GST: genomic sequence tag; NCBI: National Center for Biotechnology Information; NMRI: Naval Medical Research Institute; PFGE: pulsed field gel electrophoresis; TIGR: The Institute for Genomic Research; TDR: Special Programme for Research and Training in Tropical Diseases; YAC: yeast artificial chromosome.

These problems led to development of a third approach to genome sequencing, namely shotgun sequencing of individual chromosomes purified by pulsed field gel electrophoresis (PFGE). *P. falciparum* has 14 chromosomes ranging from 0.8 to 3.4 Mb in length. Most of the chromosomes of *P. falciparum* clone 3D7 (the clone selected for sequencing) can be resolved in PFGE gels, except for chromosomes 5–9 which co-migrate as a "blob" in the middle of the gel. Chromosomes are resolved on preparative PFGE gels and chromosomal DNA is extracted by agarase digestion. The chromosomal DNA is then sheared into 1–2 kb fragments, cloned into plasmid or M13 vectors, and randomly-picked clones are sequenced. The sequences are assembled to form contigs, and techniques such as PCR from genomic DNA with primers derived from the ends of contigs are used to close gaps in the sequence. Some laboratories also perform limited sequencing of shotgun libraries prepared from YACs previously localized on the chromosomes (Foster and Thompson 1995). The YAC-derived sequences help to group contigs from the same part of the chromosome and assist in gap closure.

Three groups are sequencing the *P. falciparum* genome: TIGR and the Malaria Program of the US Naval Medical Research Institute (NMRI); the Sanger Centre in the UK; and Stanford University. An international consortium including the genome laboratories, bioinformatics centers, and funding agencies was formed to oversee the project, facilitate collaboration, and ensure that the data will be provided to the scientific community in a timely and useful manner (Hoffman et al. 1997). Members of the consortium meet every 6 months to review progress and plan future work. The current status of the project is summarized in Table 1. The strategy of sequencing on a chromosome-by-chromosome basis led naturally to assignment of individual chromosomes to the different genome centers, with the "blob" of currently unresolved chromosomes being undertaken rather heroically by the Sanger Centre. Progress in the first pilot projects, namely chromosome 2 by TIGR/NMRI and chromosome 3 by the Sanger Center, has after initial technical difficulties been good such that both chromosomes are expected to be completed shortly, and the Stanford group has begun work on chromosome 12. Preliminary, unedited data have been released into the public domain and are available for downloading, browsing or searching on web sites maintained at each laboratory (Table 2), the National Center for Biotechnology Information (NCBI), and the European Bioinformatics Institute (EBI). The Sanger Centre and TIGR have started work on the other chromosomes.

Thus despite initial scepticism in the malaria research community that the AT-rich *P. falciparum* genome could be sequenced, the success achieved on chromosomes 2 and 3 proves that it is technically feasible, and malaria researchers should soon have access to the complete genome sequence. Recent technological advances such as stable transfection of *Plasmodium* spp., and microarray technologies for global measurement of gene expression, in combination with the genome sequence, will facilitate research to understand *Plasmodium* biology. In addition, sequencing efforts planned or underway for other *Plasmodium* species and other Apicomplexa such as *Toxoplasma* (Table 2) will provide useful complementary data. Although

Table 1. Chromosome assignments and sequencing status for the Malaria Genome Sequencing Project.

Chromosome(s)[a]	Size (Mb)	Laboratory	Funding[b]	Status (as of 3/98)
1	0.8	Sanger Centre	Wellcome Trust	random sequencing
2	1.0	TIGR/NMRI	NIAID, DoD	annotation
3	1.2	Sanger Centre	Wellcome Trust	closure
4	1.4	Sanger Centre	Wellcome Trust	random sequencing
5–9	1.6–1.8	Sanger Centre	Wellcome Trust	library preparation
10	2.1	TIGR/NMRI	NIAID, DoD	library preparation
11	2.3	TIGR/NMRI	NIAID, DoD	library preparation
12	2.4	Stanford University	BWF	random sequencing
13	3.2	Sanger Centre	Wellcome Trust	library preparation
14	3.4	TIGR/NMRI	BWF, DoD	random sequencing

[a]Estimated sizes for *P. falciparum* clone 3D7 taken from Dame et al. (1996).
[b]NIAID, National Institute for Allergy and Infectious Diseases; DoD, US Department of Defense; BWF, Burroughs Wellcome Fund.

Table 2. Internet resources related to the Malaria Genome Sequencing Project.

Web Site	Content	URL
P. falciparum chromosome 2 TIGR	Preliminary sequence data for chromosome 2.	http://www.tigr.org/tdb/mdb/pfdb/pfdb.html
P. falciparum chromosomes 1, 3, 4 The Sanger Centre	Preliminary sequence data for chromosomes 1, 3, and 4.	http://www.sanger.ac.uk/Projects/P_falciparum/
P. falciparum chromosome 12 Stanford University	Preliminary sequence data for chromosome 12.	http://sequence-www.stanford.edu/group/malaria/index.html
P. falciparum Gene Sequence Tag Project, University of Florida	A collection of ESTs and GSTs for *P. falciparum*.	http://parasite.arf.ufl.edu/malaria.html
Malaria Database Monash Univ., Walter and Eliza Hall Institute	A collection of genetic information on malaria parasites. Sponsored by WHO/TDR.	http://www.wehi.edu.au/MalDB-www/who.html
Malaria Genetics and Genomics National Center for Biotechnology Information (NCBI)	BLAST searches on Apicomplexan sequence data, including *P. falciparum*; *P. falciparum* linkage maps, etc.	http://www.ncbi.nlm.nih.gov/Malaria/
Parasite Genomes Blast Server European Bioinformatics Institute	BLAST searches on sequence data from many parasites, including *Plasmodium.*	http://www.embl-ebi.ac.uk/parasites/parasite_blast_server.html
Malaria Foundation	General information on malaria and many links to malaria-related sites.	http://www.malaria.org/index.htm
Toxoplasma Database University of Pennsylvania	Toxoplasma ESTs clustered with ESTs from dbEST.	http://daphne.humgen.upenn.edu:1024/toxodb/ver_1/
TIGR Microbial Database	A comprehensive listing of microbial genome projects.	http://www.tigr.org/tdb/mdb/mdb.html

it is a long way from laboratory research to the fielding of new drugs or vaccines, with the advent of microbial genomics we can expect the process to be speeded up considerably.

Acknowledgements

The Malaria Genome Sequencing Project is supported by The Wellcome Trust, the US Department of Defense, The Burroughs Wellcome Fund and the National Institutes of Health.This work was supported by a supplement to NIH grant R01-AI40125-01, the Naval Medical Research and Development Command work unit M00101S131720, and NIH-NMRI interagency agreement number Y1AI-6091-01. The opinions and assertions herein are those of the authors and are not to be construed as official or as reflecting the views of the US Navy or naval service at large.

References

Blattner FR, Plunkett Gr, Bloch CA, Perna NT, Burland V, Riley M, Collado VJ, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B and Shao Y (1997) The complete genome sequence of *Escherichia coli* K-12. Science **277**: 1453–74

Bloom BR (1995) A microbial minimalist. Nature **378**: 236

Butler D, Maurice J and O'Brien C (1997) Briefing malaria. Nature **386**: 535–540

Dame JB, Arnot DE, Bourke PF, Chakrabarti D, Christodoulou Z, Coppel RL, Cowman AF, Craig AG, Fischer K, Foster J, Goodman N, Hinterberg K, Holder AA, Holt DC, Kemp DJ, Lanzer M, Lim A, Newbold CI, Ravetch JV, Reddy GR, Rubio J, Schuster SM, Su XZ, Thompson JK and Werner EB (1996) Current status of the *Plasmodium falciparum* genome project. Mol Biochem Parasitol **79**: 1–12

Fleischman RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science **269**: 496–512

Foster J and Thompson J (1995) The *Plasmodium falciparum* genome project: a resource for researchers. Parasitol Today **11**: 1–4

Hoffman SL, Bancroft WH, Gottlieb M, James SL, Bond EC, Stephenson JR and Morgan MJ (1997) Funding for malaria genome sequencing. Nature **387**: 647

WHO (1997) World malaria situation in 1994. Wkly Epidemiol Rec **72**: 269–276

Malcolm J. Gardner[a,1], Hervé Tettelin[a], Daniel J. Carucci[b], Leda M. Cummings[a], Mark D. Adams[a], Hamilton O. Smith[a], J. Craig Venter[a], and Stephen L. Hoffman[b]

[a] The Institute for Genomic Research,
9712 Medical Center Drive,
Rockville, MD 20850, USA
[b] Malaria Program,
Naval Medical Research Institute,
12300 Washington Avenue,
Rockville, MD 20852, USA

[1] Corresponding author;
fax 1-301-838-0208
e-mail gardner@tigr.org