

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<sup>7,8</sup>

**Annotation**

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 Genes were identified by manual curation of the output of the software packages Genefinder (P. Green, unpublished work), GlimmerM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

Gene Ontology (GO) terms<sup>28</sup> 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. Pacheban, 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 *Encelhalitozoon cucululi*. *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**, 217–2250 (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](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<sup>1</sup>. 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<sup>2–4</sup>. We report

‡ Present addresses: National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, 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<sup>5</sup>. 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*<sup>6</sup>, 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<sup>7</sup>. Conserved subtelomeric sequences<sup>2</sup> 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<sup>5</sup>. A gene finding program, phat (pretty handy annotation tool<sup>8</sup>), was developed, supplementing the GlimmerM program<sup>9</sup> 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
<b>The genome</b>					
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
<b>Exons</b>					
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
<b>Introns</b>					
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
<b>Intergenic regions</b>					
(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
<b>RNAs</b>					
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
<b>The proteome</b>					
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
<b>Gene Ontology</b>					
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
<b>Structural features</b>					
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<sup>5</sup>. Specialized searches used the following programs and databases: InterPro<sup>21</sup>, Pfam<sup>19</sup> and Gene Ontology<sup>22</sup>. Predictions of apicoplast and mitochondrial targeting were performed using TargetP<sup>20</sup> and MitoProtII<sup>25</sup>; transmembrane domains, TMHMM<sup>24</sup>; 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<sup>29,30</sup>.

§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<sup>10</sup>. 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<sup>2</sup>.

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<sup>11</sup>. 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<sup>12</sup> 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<sup>13,14</sup>. 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<sup>15</sup> and the Wellcome Trust Malaria Genome Mapping Project<sup>16</sup>. 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<sup>5</sup> (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<sup>5</sup>. Ordering of contigs was facilitated by the use of sequence tagged sites<sup>16</sup> and microsatellite markers<sup>17</sup>. The final assembly of each chromosome was verified by comparison with *Bam*HI and *Nhe*I optical restriction maps<sup>18</sup>. 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<sup>9</sup> and phat<sup>8</sup> 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<sup>19</sup>, PROSITE<sup>20</sup> and InterPro<sup>21</sup> 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<sup>22</sup> terms. Signal peptides and signal anchors were predicted with SignalP-2.0 (ref. 23). Transmembrane helices were predicted with TMHMM<sup>24</sup>. Mitochondrial- and apicoplast-targeted proteins were predicted by MitoProtII<sup>25</sup>, TargetP<sup>26</sup> and PATS<sup>27</sup>. tRNA-ScanSE<sup>28</sup> was used to identify transfer RNAs.

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

- 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).
- Gardner, M. J. *et al.* Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**, 498–511 (2002).
- Hall, N. *et al.* Sequence of *Plasmodium falciparum* chromosomes 1, 3–9 and 13. *Nature* **419**, 527–531 (2002).
- Hyman, R. W. *et al.* Sequence of *Plasmodium falciparum* chromosome 12. *Nature* **419**, 534–537 (2002).
- Gardner, M. J. *et al.* Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science* **282**, 1126–1132 (1998).
- Glockner, G. *et al.* Sequence and analysis of chromosome 2 of *Dictyostelium discoideum*. *Nature* **418**, 79–85 (2002).
- Bowman, S. *et al.* The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*. *Nature* **400**, 532–538 (1999).
- Cawley, S. E., Wirth, A. I. & Speed, T. P. Phat—a gene finding program for *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **118**, 167–174 (2001).
- Salzberg, S. L., Perlea, M., Delcher, A., Gardner, M. J. & Tettelin, H. Interpolated Markov models for eukaryotic gene finding. *Genomics* **59**, 24–31 (1999).
- Huestis, R. & Fischer, K. Prediction of many new exons and introns in *Plasmodium falciparum* chromosome 2. *Mol. Biochem. Parasitol.* **118**, 187–199 (2001).
- Maniatis, T. & Tasic, B. Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* **418**, 236–243 (2002).
- Lawrence, C. E. *et al.* Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262**, 208–214 (1993).
- 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).
- 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., Welles, 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., Schmucker, 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. Lasander, 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](mailto: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 PlasmDB (<http://plasmdb.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<sup>1</sup>. 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<sup>2–4</sup>. 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<sup>7</sup>. (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<sup>7</sup>, 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<sup>8,9</sup>. 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,