

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. Zuegg, 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. 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). 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 year¹. 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 published⁷. (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 data⁷, 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.

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 putative 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 WTSI². *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 Lasorder *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* genes² (Table 1).

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 ¹	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: InterPro¹⁶; Pfam¹⁹; Gene Ontology²⁰. Predictions of apicoplast and mitochondrial targeting were performed using TargetP²¹ and MitoProtII²²; transmembrane domains, TMHMM²³; and signal peptides and signal anchors, SignalP-2.0²⁴.

*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}.

¹Hypothetical proteins are proteins with insufficient similarity to characterized proteins in other organisms to justify provision of functional assignments.

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 described⁷. 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 described¹⁴. 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 reads¹⁵. Consed was used to edit the assembled sequence¹⁶. 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 program¹⁷. 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. Bream, 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 erythrocyte⁴. 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 parasite⁵. 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 sequences⁷. 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