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

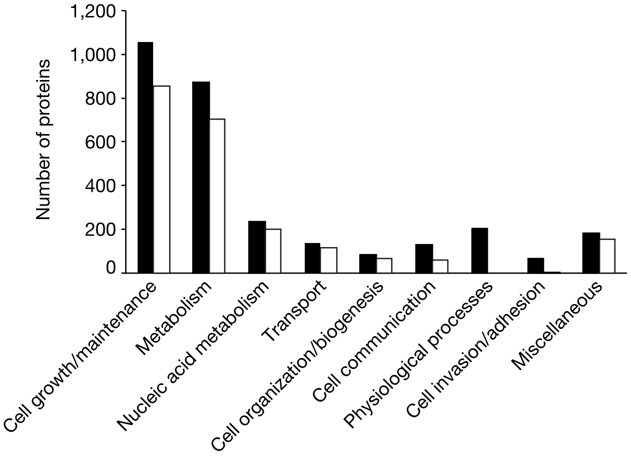

Figure 1 Functional classification comparison between *P. falciparum* and *P. y. yoelii* proteins. We compared the GO terms of proteins assigned to 'biological process' for the orthologous genes identified between the two species. The process group contains 3,041 *P. falciparum* annotations (filled bars), 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-*

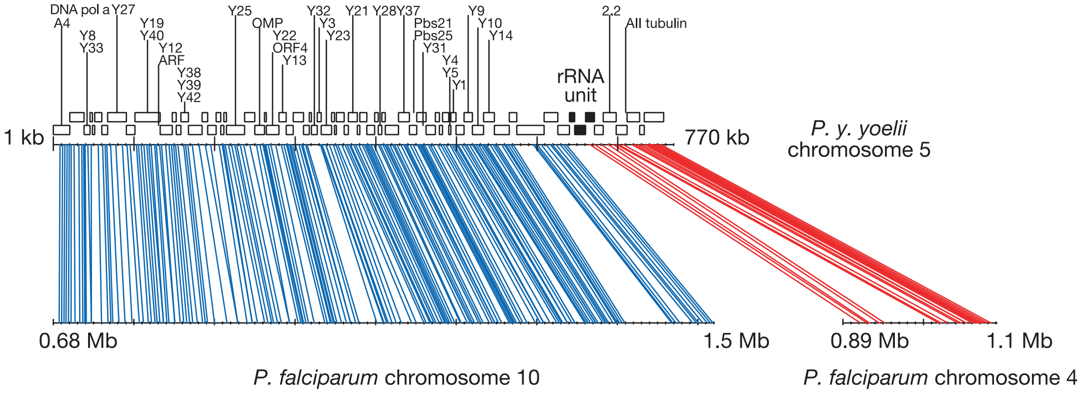

Figure 2 Conservation of gene synteny between *P. y. yoelii* chromosome 5 and *P. falciparum* chromosomes 4 and 10. Physical marker data used to confirm contig order in the tiling path of *P. y. yoelii* chromosome 5 are shown above the contigs (open boxes). Each coloured line represents a pair of orthologous 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.

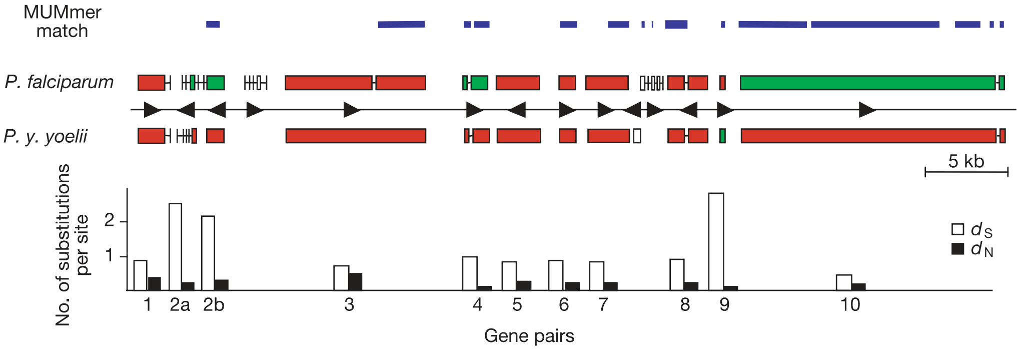

Figure 3 Global alignment scheme of a syntenic region between *P. falciparum* and *P. y. yoelii* encompassing ten orthologous gene pairs and nine intergenic regions. White boxes represent genes that have no orthologue and were excluded from analysis; green boxes represent gene models that were refined; 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.