Chapter 4

A comprehensive survey of the *Plasmodium* life cycle by genomic, transcriptomic, and proteomic analyses

A. Genome team

Neil Hall[1†], Jane M. Carlton[4,5,6], Taco W.A. Kooij[2], Matthew Berriman[1], Christoph S. Janssen[7], Arnab Pain[1], Keith James[1], Kim Rutherford[1], Barbara Harris[1], David Harris[1], Carol Churcher[1], Michael A. Quail[1], J. Dale Raine[3], Marianna Karras[2], Doug Ormond[1], Jon Doggett[1], Shelby L. Bidwell[4], Marie-Adele Rajandream[1], Chris J. Janse[2], Robert E. Sinden[3], Andrew P. Waters[2], C. Michael R. Turner[7] and Bart Barrell[1]

B. Transcriptome team

Marianna Karras[2†], Jane M. Carlton[4,5,6], Neil Hall[1], Georges K. Christophides[8], J. Dale Raine[3], Robert E. Sinden[3], Fotis C. Kafatos[8], Chris J. Janse[2] and Andrew P. Waters[2]

C. Proteome team

J. Dale Raine[3†], Laurence Florens[9], Holly E. Trueman[3], Jacqui Mendoza[3], Neil Hall[1], Jane M. Carlton[4,5,6], Daniel J. Carucci[10], John R. Yates III[9] and Robert E. Sinden[3]

[1]Pathogen Sequencing Unit, The Wellcome Trust Sanger Institute, Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [2]Malaria Research Group, Department of Parasitology, Centre for Infectious Diseases, Leiden University Medical Centre, Albinusdreef 2, 2333 ZA, Leiden, The Netherlands. [3]Immunology and Infection Section, Department of Biological Sciences, Imperial College London, Sir Alexander Fleming Building, Imperial College Road, London SW7 2AZ, UK. [4]The Institute for Genomic Research (TIGR), 9712 Medical Center Drive, Rockville, MD 20850, USA. [5]Department of Pathobiology, College of Veterinary Medicine, University of Florida, Gainesville, FL 32608, USA. [6]Department of Molecular Microbiology and Immunology, Johns Hopkins University, Bloomberg School of Public Health, 615 North Wolfe Street, Baltimore, MD 21205, USA. [7]Division of Infection and Immunity, Institute of Biomedical and Life sciences, University of Glasgow, Glasgow G12 8QQ, UK. [8]European Molecular Biology Laboratory (EMBL), Meyerhofstrasse 1, 69117 Heidelberg, Germany. [9]Department of Cell Biology, The Scripps Research Institute, SR-11, 10550 North Torrey Pines Road, La Jolla, CA 92037, USA. [10]Naval Medical Research Center, Malaria Program (IDD), 503 Robert Grant Avenue, Room 3A40, Silver Spring, MD 20910-7500, USA.

†These authors contributed equally to this work and are listed alphabetically. N.H. led the genome team; M.K., the transcriptome team; and J.D.R., the proteome team.

Science **307** (5706), 82-86 (2005). Reprinted with permission from the American Association for the Advancement of Science.

Introduction to Chapter 4: "A comprehensive survey of the Plasmodium life cycle by genomic, transcriptomic, and proteomic analyses".

This study was the result of multilateral effort of 30 scientists from eight different research groups from the USA, the UK, Germany and the Netherlands and presents the partial genome sequences of the rodent malaria parasites (RMPs) *Plasmodium berghei* and *Plasmodium chabaudi* in conjunction with global analyses of gene expression in the many different developmental stages of the life cycle of *P. berghei*. Genomic analyses (led by Neil Hall) were mainly performed at the Wellcome Trust Sanger Institute (WTSI, Cambridge, UK), at The Institute for Genomic Research (TIGR, MD, USA) and at the Department of Parasitology at the LUMC (Leiden, Netherlands), transcriptomic analyses (led by Marianna Karras) were mainly performed at the Department of Parasitology at the LUMC (Leiden, Netherlands) and proteomic analyses (led by J. Dale Raine) were mainly performed at the Department of Biological Sciences at the Imperial College (London, UK). My contribution to the paper consisted of an in depth analysis of the synteny between three RMPs with the completed *Plasmodium falciparum* genome.

At TIGR, the *P. berghei*, *P. chabaudi* and *Plasmodium yoelii* contigs were aligned with the *P. falciparum* genome using the MUMmer algorithm[186]. I combined these alignment data of all contigs of the three different species to construct composite RMP (cRMP) contigs based on the coordinates of alignment with the *P. falciparum* genome. This approach was feasible due to the high degree of synteny between the three RMPs and between the RMPs and *P. falciparum*. In addition, I manually analysed the 906 *P. falciparum*-specific genes that have no orthologues in the RMPs as determined by reciprocal BLAST analyses (the so-called orphan genes).

The majority of these genes are located in the subtelomeric regions abutting the syntenic core regions (575 genes) and the boundaries defining these subtelomeric regions could be sharply defined to a single intergenic region. The remaining genes located in the core regions of the chromosomes were studied further by analyses of their direct genomic location. Upstream and downstream neighbouring genes were compared by BLAST analysis with the RMP contigs and utilizing the generated tiling paths the presence of highly diverged but positionally conserved orthologues was established. This analysis identified 68 genes with low homology to RMP orthologues, undetectable by BLAST analysis, and 41 positionally conserved genes without sequence homology (including *msp* homologues and *lsa1*; Chapter 2). For another 57 genes, *P. falciparum* specificity could not be conclusively determined. This group consists of (i) 24 small genes that are not annotated in the RMP and encode proteins with less than 200 amino acids, (ii) 16 genes for which no RMP sequence was available but polymerase chain reaction (PCR) data indicated sequence gaps that were sufficiently large to contain the genes, and (iii) 17 genes for which no sequence or linkage data of the RMP existed. The remaining 161 *P. falciparum* genes were found to be species-specific and disrupt synteny, 16 of which appeared to have originated from local gene duplications.

The results of these analyses were published as Appendix 2 and SOM Tables S3 and S4 and as part of the section “Genome sequencing and annotation” of the following paper.

Abstract

Plasmodium berghei and *Plasmodium chabaudi* are widely used model malaria species. Comparison of their genomes, integrated with proteomic and microarray data, with the genomes of *Plasmodium falciparum* and *Plasmodium yoelii* revealed a conserved core of 4,500 *Plasmodium* genes in the central regions of the 14 chromosomes and highlighted genes evolving rapidly because of stage-specific selective pressures. Four strategies for gene expression are apparent during the parasites' life cycle: (i) housekeeping; (ii) host-related; (iii) strategy-specific related to invasion, asexual replication, and sexual development; and (iv) stage-specific. We observed posttranscriptional gene silencing through translational repression (TR) of messenger RNA (mRNA) during sexual development, and a 47-bp 3' untranslated region (UTR) motif is implicated in this process.

Introduction

RMPs provide model systems that allow issues to be addressed that are impossible with the human-infectious species *P. falciparum* and *Plasmodium vivax*[220]. Three closely related species, *P. chabaudi*, *P. yoelii* and *P. berghei*, are in common use in the laboratory. Comparative sequencing and analysis of the genomes of such model species, in addition to the complete genome sequence of *P. falciparum*[42], provide insights into the evolution of *Plasmodium* genes and gene families[51] (Chapter 3).

The malaria parasite differentiates into a series of morphologically distinct forms in the vertebrate and mosquito hosts. It alternates between morphologically related invasive stages (sporozoite, merozoite, and ookinete) and replicative stages (pre-erythrocytic, erythrocytic-schizont, and oocyst) interposed by a single phase of sexual development that mediates transmission from the human host to the anopheline vector[220]. This report integrates genome sequence analyses of *P. berghei* and *P. chabaudi* with transcriptome and proteome data for *P. berghei*, allowing the categorization of protein expression, the analysis of regulation mechanisms for gene expression, and the identification of species-specific gene families and genes under selective pressure.

Materials and methods

A. Genome

DNA preparation

DNA was prepared from *P. chabaudi chabaudi* (AS strain) and *P. berghei* (ANKA strain) as previously described[221].

DNA sequencing and assembly

Genomic DNA was sheared by sonication and fragments of 2-4 kb were selected. Library construction and DNA sequencing was carried out as previously described[222]. DNA sequences were assembled using the Phusion Assembler algorithm[223]. Reads from repetitive regions that were not incorporated into contigs using Phusion were assembled into contigs using Phrap (P. Green, Washington University) to remove redundancy from the final data set. The resulting contig set was screened against the *Mus musculus* genome sequence[32] using BLASTN with

a window size of 30. Contigs which were >90% identical to the mouse genome over >80% of their length were removed from the analysis. Each contig was assigned an identifier according to the assembly algorithm used to produce it. In the case of *P. berghei*, identifiers are called PB_RP0001…PB_RP3991 for contigs built using the Phusion assembler, and PB_PH0001…PB_PH5748 for contigs built by the Phrap assembler. *P. chabaudi* contigs have similar names with the prefix "PC".

Gene prediction and nomenclature

Gene models were predicted primarily using the *P. falciparum* protein set; peptides were mapped onto the genomes using the Genewise package[224]. GlimmerMExon[51] (Chapter 3) was used to predict genes that were divergent or novel to the specific genomes and therefore not represented in the peptide set. Only GlimmerMExon gene models that did not overlap GeneWise models were accepted. All *P. berghei* genes were given systematic identifiers beginning with *P. berghei* and all *P. chabaudi* genes were given identifiers starting with PC. *bir* and *cir* genes were predicted using specific hidden Markov models (HMMs) derived from alignments of 73 *cir* and *yir* genes and an alignment of 21 *bir* genes. All *bir* and *cir* models were curated by hand.

Annotation

Gene predictions were annotated based on reciprocal best matches to *P. falciparum* or *P. yoelii*. Reciprocal BLASTP hits with scores >50 were accepted. Clusters of paralogous gene families were generated using TRIBE[225] and curated by hand using Jalview (http://www.jalview.org/). HMMs of each family were built using the HMMer package (http://hmmer.wustl.edu/). For genome comparisons, orthologues were identified by reciprocal BLAST searches between the species. For each pair-wise comparison, the protein and nucleotide identities were calculated using the EMBOSS "needle" algorithm to align the DNA for each orthologous gene pair[226]. The orthologue pairs calculated by BLAST, the rate of non-synonymous versus synonymous mutation was calculated using the codeml programme, part of the PAML software package[227]. Statistical analysis of each pair of distributions shown was undertaken using a Kolmogorov-Smirnov two tailed test.

Alignment of four Plasmodium *genomes and identification of further orthologues*

Contig sequences from the three RMP genomes were concatenated and aligned to all 14 *P. falciparum* chromosomes using default options of the protein version of the local alignment programme MUMmer[186]. *P. falciparum* orphan genes, identified through the lack of reciprocal BLAST matches, were manually analysed utilizing the tiling paths generated by MUMmer. Briefly, flanking genes were subject to TBLASTN searches and orthology of orphan genes confirmed by the location of the corresponding RMP contig in the tiling path.

B. Transcriptome

Target amplification and labelling

RNA was extracted from blood-stage parasites of two clones of *P. berghei*, a non-gametocyte producer clone HPE[228] and a gametocyte producer clone HP

(reference clone 15cy1)[228], grown in highly synchronized *in vitro* cultures[221]. Gametocyte RNA was extracted from immature and mature gametocytes that were obtained from synchronous *in vivo* infections of the HP clone, and purified from other blood stages by Nycodenz density centrifugation, as described[221]. Nycodenz purification resulted in 94% pure gametocytes contaminated with 6% schizonts, as determined from Giemsa stained blood films. Purity of the isolated blood stages was determined by examination of Giemsa stained blood films and by fluorescence activated cell sorter (FACS) analysis. The input levels of parasite RNA were normalized, as early time points had a strong contamination from mouse RNA. A total of 5μg *P. berghei* RNA was used for cDNA synthesis. RNA was primed with the T7-dT(24) primer for 10 min at 70°C and first strand synthesis was carried out using Superscript II RT[229]. The reaction was incubated for 1 hour at 42°C. Second strand cDNA synthesis was performed by adding the second strand synthesis buffer, *E. coli* DNA ligase, *E. coli* DNA polymerase I and *E. coli* RNase H, incubating for 2 hours at 16°C and the reaction stopped by adding 5nM ethylenediaminetetraacetic acid (EDTA). cDNA was purified by phenol:chloroform:isoamyl-alcohol extraction. *In vitro* transcription was performed using the Ambion MEGAscript T7 RNA synthesis kit according to the manufacturer's instructions. The resulting cDNA was purified using the RNeasy kit (Qiagen). Complementary cDNA probes were synthesized and labelled with Cy3-dUTP or Cy5-dUTP fluorescent nucleotide analogues, in a random primed first-strand reverse-transcription reaction. After removal of unincorporated dNTPs with a Qiagen PCR purification kit, two differentially labelled probes were combined, lyophilized and resuspended in hybridization buffer containing 50% formamide, 6X SSC, 0.5% SDS, 5X Denhardt's reagent and 0.5mg/ml poly(A) DNA. Arrays were prehybridized in 6X SSC, 0.5% SDS and 1% BSA in 42°C for 1 hour, hybridized overnight at 42°C in humidified hybridization chambers, washed twice in 0.1X SSC, 0.1% SDS (30 min), twice in 0.1X SSC (30 min) at room temperature, rinsed with de-ionized water and dried. Microarrays were scanned using an Agilent scanner, and image analysis was performed using GenePix Pro 4.0 software (Axon). Spots of the array with obvious blemishes were manually flagged and excluded from subsequent analyses. Normalized data were further analysed with the CLUSTER and TREEVIEW programmes[230]. Hierarchical clustering analysis ordered the selected genes according to similarities in their pattern of expression throughout experiments, and genes could be divided into clusters.

Library and microarray design

The 6.3 K *P. berghei* DNA microarray was generated from a *P. berghei* genomic DNA library, supplied by J. B. Dame and J. M. Carlton (University of Florida). Briefly, genomic DNA obtained from the blood stages of clone 15cy1 of *P. berghei* ANKA was digested with mung bean nuclease, as described[230]. Mung bean nuclease digestion generates DNA fragments that contain intact genes rather than intergenic regions, thereby reducing the complexity of the library. The library consists of 6,354 clones size selected in the range 500-2,000 bp, and each clone has been sequenced at the 5' end to generate a genome-survey sequence (GSS). Each GSS was searched against a database of the assembled *P. berghei* contigs

and homology that covered ≥90% of the GSS was noted. If the homology was outside a coding sequence prediction the nearest downstream coding sequence in the direction of the GSS was reported. If no homology was observed, the GSS was searched against the *P. yoelii* contig database with a cutoff of ≥20% of the GSS. The *P. yoelii* coding sequences were mapped on the *P. falciparum* genome[42,182] and thus linked to their annotation and gene ontology assignment. The *P. yoelii* coding sequences were also used to search back against the *P. berghei* genome data. The 6,354 individual sequence tagged *P. berghei* clones correspond to at least 3,987 different gene models of which 2,045 match annotated gene models in the *P. berghei* genome, 1,941 have a direct orthologue only in *P. yoelii* and 687 sequence tags do not correspond to an annotated gene model in *Plasmodium*. Only 57 gene tags on the array were found to be mouse-specific (90% identity over 50 bp). Based on the number of protein coding genes identified in *P. yoelii*[51] (Chapter 3) it was estimated that the GSS library represents 68% of the *P. berghei* coding sequences. A total of 150 gene models represent members of the *bir* family and a further 28 are specific to the five new *P. berghei* gene families described in this study (14 specific to *pbst-a*, 12 specific to *pbst-b*, two specific to *pbst-c*). The *P. berghei* gDNA inserts were amplified from the gridded library by standard PCR using universal plasmid primers, purified through NucleoSpin columns, and resuspended in spotting buffer (ArrayIt Microspotting solution, Telechem International). A total of 25 known *P. berghei* genes and ten mouse, mosquito and bacterial genes were amplified from genomic or plasmid DNA and used as controls on the microarray. The GSS library and the control genes were spotted on aminosilane-coated glass slides, using the Omnigrid microarray spotter (GeneMachines). DNA was crosslinked onto the glass slides by baking for 3 hours at 60°C and for 10 min at 100°C. The reliability of the *P. berghei* DNA microarray was proven in various manners. The library was spotted in duplicate on the array and values for identical spots compared and confirmed to be consistent within each hybridization. Also identical experiments with a different preparation of target starting material and the same target on arrays spotted independently gave virtually identical results. Moreover, duplicate competitive hybridizations in which the Cy3 and Cy5 labels were swapped ("dye swap experiments") proved that bias due to preferential dye incorporation had no influence on the data presented. Genes with known expression patterns were used as controls throughout experiments and conformed to expectation. Lastly, cluster analysis consistently grouped independent GSS clones that contained either the same sequence or partial fragments of the same gene.

Selection criteria (blood-stage transcripts)

Data presented in Figure S19 were analysed using the CLUSTER and TREEVIEW programmes[230]. Using CLUSTER analysis, we eliminated low intensity signals and genes displaying at least a two-fold change in regulation in at least one pair-wise comparison only were selected for presentation. Relatively broad selection criteria were employed (2-fold difference in expression level) in order to include genes, which might be even weakly regulated. Moreover only genes that are detected in at least 80% of the pair-wise comparisons were selected. This selection procedure excluded several known stage-specific markers when expression was below the

two-fold threshold but managed to include genes that undergo prolonged transcription that peaks during one or more stages of the parasite's development. All gene identifications for the genes mentioned in the text are given in SOM Table S11.

C. Proteome

Parasite and mosquito maintenance

P. berghei ANKA clone 234 (gametocyte producer) and clone 233 (gametocyte non-producer) parasites were maintained in Theiler's Original female mice and *Anopheles stephensi* mosquitoes as previously described[231].

Collection of P. berghei *preparations*

Asexual blood stages (clone 233), gametocytes (clone 234) and ookinetes (clone 234) were prepared as described previously[231], with the following modifications: Ookinete preparations, following enrichment on a 55% Nycodenz density cushion, were further enriched by three successive washes in phosphate buffered saline (PBS) and centrifugation at 300 g, 200 g and 160 g, each for 10 min at 4°C (M.C. Rodriguez, personal communication). Purity was determined by microscopic analysis of Giemsa-stained blood films in which at least 10 fields of view and at least 1,000 parasites were counted for every preparation. Asexual blood-stage preparations were pure, containing no other parasite stages. Gametocyte preparations contained <5% contamination from asexual blood stages. Ookinete preparations contained <1.5% contamination from asexual blood stages and <3.5% from gametes and undifferentiated zygotes. At least $3.5x10^7$ cells were used for each gametocyte and ookinete preparation, and $>2x10^8$ cells for asexual blood-stage parasite preparations. For each oocyst preparation ~1,000 whole mosquito midguts were dissected into PBS on days 9-12 post infection (p.i.). For each sporozoite preparation ~1,000 sets of salivary glands were dissected into PBS on days 20-24 p.i. Non-infected mouse blood and non-infected guts and glands from *A. stephensi* were analysed as controls. Samples were washed in PBS then pelleted and stored at -80°C.

Proteomic analysis

Thawed samples were washed and lysed as described previously[11]. Protein fractions were digested using either endoproteinase Lys-C/trypsin or proteinase K as described[11,232]. Nine asexual blood-stage, nine gametocyte, nine ookinete, three oocyst and three sporozoite fractions digested using the endoproteinase Lys-C/trypsin protocol and nine asexual blood-stage, nine gametocyte, three ookinete, six oocyst and six sporozoite fractions using the proteinase K protocol were analysed by MudPIT as described previously[11].

Tandem mass spectrometry dataset analysis

A protein sequence database was assembled that contained gene model sequences from both the *P. berghei* (this study) and *P. yoelii*[51] (Chapter 3) genome databases. The *P. yoelii* gene model sequences were included to account, in part, for missing, partial or erroneous *P. berghei* gene models. These sequences can be found at: ftp://ftp.sanger.ac.uk/pub/pathogens/P_berghei/Berg.peptides.2.7.2003 &

ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/p_yoelii/annotation_dbs/PYA1.pep. To identify contaminating host proteins the parasite database was supplemented with a contaminant (mouse, *Anopheles*, common contaminants) database as previously described[11]. A modified version of the SEQUEST algorithm[233], PEP-PROBE[234], was used to match tandem mass spectra to sequences in the assembled sequence database. In addition to the SEQUEST filters, PEP-PROBE uses a hypergeometric probability model to provide a statistical confidence for each spectrum-peptide match to be non-random. Matches were filtered as described[11] (*i.e.* minimum cross-correlation score of 1.8 for +1, 2.5 for +2, and 3.5 for +3 spectra, minimum DeltaCn of 0.08 and minimum peptide length of seven amino acids), with the additional filter that proteins were only retained if they contained spectrum-peptide matches with a statistical confidence >85%. Non-tryptic peptides were retained in the dataset due to the non-specific cleavage activity of the proteinase K enzyme. Peptide hits were deemed unambiguous only if they were not found in non-infected controls and were uniquely assigned to parasite proteins by searching against combined parasite-host databases. For low coverage loci (proteins identified by <3 peptides), peptide/spectrum matches were visually assessed based on criteria previously described[11]. Protein lists resulting from the searches against the two different parasite databases were merged using a *P. berghei-P. yoelii* reciprocal orthologues table (ftp://ftp.sanger.ac.uk/pub/pathogens/P_berghei/COMPARISONS/py_pb.reciprocal.out).

Using the described protocol, ~$2x10^8$ tandem mass spectra generated from the MudPIT analysis were searched against the combined parasite/contaminants database. Filtering spectra-to-peptide matches based on cross-correlation score and DeltCN (*i.e.* parameters also available with SEQUEST) and following removal of contaminant host/vector or other proteins, we identified peptides matching >5,000 parasite proteins (1,000-3,000 proteins per stage). Further filtering using the hypergeometric-probability model function of PEP-PROBE and manual inspection of spectra, conclusively identified 1,836 proteins (SOM Tables S6-S8). Only these high-confidence identifications are discussed in the paper. Comparing these data to two *P. falciparum* proteome studies[11,12] gave the following statistics: proteins identified by single peptides are reduced from >32%[11,12] to 21-26% per stage, and the average sequence coverage rose from ~9%[11] to ~18%.

Genome sequencing and annotation

Partial shotgun sequencing of the genomes of *P. chabaudi* (AS) and *P. berghei* (ANKA) generated assemblies of approximately 17 and 18 Mb, respectively (Table 1A, Appendix 3). Orthologous genes of these two genomes and of *P. yoelii*[51] (Chapter 3) and *P. falciparum*[42] were inferred through bi-directional BLAST searches (Table 1B). Combining the gene predictions of the three RMPs revealed that 4,391 genes had orthologues in *P. falciparum.* These orthologues represent a universal *Plasmodium* gene set (SOM Table S2), which was mainly distributed across the central “core” regions of the 14 *P. falciparum* chromosomes. For example, in the core region of *P. falciparum* chromosome 2 (Pfchr2), 144 of 158 genes had RMP orthologues (Appendix 2), whereas in the subtelomeric regions, only three of 65 genes showed (low) homology to RMP genes (see also equivalent

Table 1: Genome summary statistics. A more detailed set of statistics is given in Appendix 3.

	P. berghei	***P. chabaudi***	***P. yoelii***	***P. falciparum***
Size (bp)	17,996,878	16,866,661	23,125,449	22,853,764
No. contigs	7,497	10,679	5,687	93
Av. contig size (bp)	2,400	1,580	4,066	213,586
Sequence coverage[a]	4x	4x	5x	14.5x
No. protein coding genes	5,864[b]	5,698[b]	5,878	5,268

[a] Average number of sequence reads per nucleotide.
[b] An excessive number of gene models were predicted for *P. berghei* and *P. chabaudi* due to the fragmented nature of the genome sequence data for these species. Thus the gene numbers indicated are for gene predictions where orthologues were identified in other *Plasmodium* species only.

maps for all chromosomes in Appendix 2). In addition to BLAST analysis, orthology of gene models was manually examined based on the conservation of gene order between the RMPs and *P. falciparum,* resulting in the identification of an additional 109 orthologues (SOM Table S3). 736 *P. falciparum* genes had no orthologues in the RMP genomes and 161 of these were located in the core regions (SOM Table S3). The other 575 are located in the subtelomeric regions and Markov[225] clustering of these *P. falciparum*-specific genes revealed that almost half could be assembled into twelve distinct gene families (Appendix 2). Only five subtelomeric gene families are obviously shared between all the sequenced *Plasmodium* species (Appendix 4)[146]. Previous studies have shown that a subtelomeric gene family of *P. vivax*, the *P. vivax* interspersed repeats (*vir*)[144], had related gene families in *P. berghei* (*bir*), *P. chabaudi* (*cir*) and *P. yoelii* (*yir*)[202,235] and we suggest *pir* (*Plasmodium* interspersed repeats) to collectively describe the families. The *bir* and *cir* families code for highly variable proteins that share approximately 30% sequence identity at the amino acid level. The copy number appears to be much higher in *P. yoelii* (>800 copies) compared to *P. berghei* (180 copies) and *P. chabaudi* (138 copies).

Selective pressure

Comparison of orthologues genes of different species by means of models of nucleotide sequence evolution can be used to investigate variable (and positive or

Table 2: Genome comparisons between the four sequenced *Plasmodium* species.

	Pb-Pc	**Pb-Py**	**Pc-Py**	**Pb-Pf**	**Pc-Pf**	**Py-Pf**
Av. protein identity (%)	83.2	88.2	84.6	62.9	61.9	61.2
Av. nucleotide identity (%)	87.1	91.3	88.1	70.3	70.1	69.6
Median d_N	0.07	0.05	0.06	0.26	0.26	0.29
Median d_S	0.49	0.026	0.53	26.1	26.5	49.4
Median d_N/d_S[a]	0.13	0.16	0.11	0.009	0.009	0.008
No. orthologous gene pairs	4,641[b]	3,153	3,318	3,890	3,842	3,375

[a] The high number of orthologues inferred between *P. chabaudi* and *P. berghei* compared to pairwise comparisons of the other species most likely reflects the method of automated annotation of both genomes, which used identical gene-finding algorithms (see also Materials and Methods).
[b] Median d_N/d_S value represents the median value of d_N/d_S for every gene pair, and is not calculated from the median d_N and d_S values for each comparison. The median d_N/d_S for comparisons with *P. falciparum* are low because of the saturation of synonymous changes in the alignments, resulting in high d_S values.

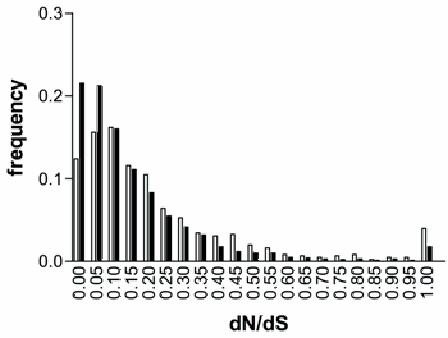
Abbreviations: Pb, *P. berghei*; Pc, *P. chabaudi*; Py, *P. yoelii*; Pf, *P. falciparum*.

negative) selective pressures[38,236]. We determined the relative number of synonymous (d_S) versus non-synonymous (d_N) substitutions between orthologues of *P. berghei* and *P. chabaudi*. In general, we found that orthologous gene pairs are under purifying selection pressure (and have $d_N/d_S < 1$) and the observed ratios of median values for genes of RMPs (Table 2) were similar to those reported for *Caenorhabditis elegans*/*Caenorhabditis briggsae* and mouse/human[32,38]. This strong divergence from $d_N/d_S = 1$ suggested that most RMP gene models code for proteins and are not mispredictions or pseudogenes. The distribution of d_N/d_S ratios of genes containing transmembrane (TM) domains or signal peptides (SPs; *i.e.* genes which may be extracellular) was greater than that of cytoplasmic proteins lacking these domains (Figure 1A) indicating reduced purifying, or increased diversifying pressure on the former, possibly as a result of selective pressure from the host. When these data are correlated with expression data from the transcriptome and proteome analysis (SOM Table S5), we observe significant difference between d_N/d_S values in SP/TM-containing and non-SP/TM-containing genes in blood-stage proteins but not vector stage proteins (Figure 1B) indicating that diversifying selection might result from the selective pressure from the host adaptive immune response, although some parasite proteins expressed in the vector are also clearly under diversifying selection. Interestingly, annotated genes with the highest d_N/d_S values included many genes that one would expect to play a role in host-parasite interactions, such as reticulocyte binding protein (0.81), rhoptry associated protein (0.94), and erythrocyte binding antigen (0.78). We have compared our dataset with the recent study of selection using codon volatility in *P. falciparum*[237]. There are 15 *P. berghei* genes with a d_N/d_S ratio > 1 which have detectable orthologues in *P. falciparum*. Not all of these have scores indicating a high volatility, a result consistent with the fact that selection will be operating at different levels in different species and that volatility and d_N/d_S values measure selection over different time scales.

Gene expression

The asexual blood-stage cycle of *P. berghei* takes 22-24 hours and gametocyte development 30 h. Gametocytes are morphologically discernable from the asexual trophozoites only after 18 hours (Figure S18). Transcriptome data were obtained from three time points during the G1 phase (rings, young and mature trophozoites) and from two time points during the S/M phase (immature and mature schizonts) as well as from purified immature (24 hours) and mature (30 hours) gametocytes. The transcription profile of these stages was compared by a series of pair-wise hybridizations to a *P. berghei* GSS amplicon DNA microarray. Proteome data were collected from mixed asexual blood stages (containing both invasive and replicative stages), gametocytes during blood-stage development, ookinetes, oocysts (day 9-12 post-infection) as well as salivary gland sporozoites and analysed by Multidimensional Protein Identification Technology[238]. The proteome analysis resulted in the identification of 1,836 parasite proteins with high confidence (SOM Tables S6-S8) and >5,000 parasite proteins with relaxed filtering. By comparing expression data for the different life cycle stages, we could categorize proteins into the following four strategies of gene expression: (i) housekeeping; (ii) host-related expression; (iii) strategy-specific expression; and (iv) stage-specific expression.

A



B

	SP or TM		Non-SP or TM			
	n	median	n	median	D	p
Genomic – all stages	726	0.174	3952	0.121	0.147	>0.001*
Proteome and transcriptome – all stages	147	0.157	501	0.100	0.217	>0.001*
- mammal expressed	106	0.159	311	0.098	0.237	>0.001*
- mosquito expressed	41	0.139	188	0.104	0.17	<0.05

Figure 1. d_N:d_S ratios between pairs of orthologous genes in *P. berghei* and *P. chabaudi* and a comparison of genes containing SP or TM domains versus those lacking such domains

(A) Frequency distribution for all orthologue pairs. Open bars represent orthologues containing SP or TM domains; solid bars represent orthologues lacking such domains. (B) Analysis of distributions for all orthologues confirmed to be transcribed using transcriptome data or expressed using proteome data (SOM Table S2), partitioned according to their expression in mammalian or mosquito phases of the life cycle. The D variable represents the Kolmogorov-Smirnov test output statistic.

Housekeeping

Of the 1,836 proteins detected, 136 were expressed in at least four of the five stages analysed (SOM Table S8). Given the lower number of proteins identified in the oocyst (277 proteins) and the sporozoite (134 proteins) compared to the other stages analysed (733 to 1,139 proteins), our analysis will have excluded some of the 301 proteins detected in asexual blood stages, gametocytes, and ookinetes (Figure 2C). Recognizing that these 301 proteins were detected in both vertebrate and mosquito stages; we anticipate that some of these will also be expressed in oocysts and sporozoites.

Host-related expression

The proteome and transcriptome datasets revealed that enzymes of the tricarboxylic acid (TCA) cycle, oxidative phosphorylation and many other

mitochondrial proteins were upregulated in the gametocyte when compared to the asexual blood stages and were even more abundant in the ookinete (SOM Figure S16, SOM Table S8). These observations suggest that, similar to trypanosomes[239], mitochondrial activity increases in the gametocyte as a pre-adaptation to life in the mosquito vector and are consistent with the more complex organization of mitochondria in gametocytes[220,240]. Mitochondrial activity apparently continues to increase in the ookinete.

Strategy-specific expression

Strategy-specific expression is related to invasion, asexual replication and sexual development. We uniquely detected 966 proteins in invasive zoite- (merozoite, ookinete, sporozoite)-containing preparations, of which 234 were shared between at least two of the three invasive stages but not with the replicative or sexual stages (Figure 2A). Gliding motility typifies the invasive stages of Apicomplexa, and many proteins with a (putative) role in this process were detected. Micronemes and rhoptries are secretory organelles specific to the invasive stages. Interestingly, while ten known rhoptry proteins were detected in blood stages and sporozoites, these rhoptry proteins were absent from ookinetes. In contrast, most known micronemal protein families were detected in all zoites but with clear stage-specific expression of different family members. Perforin-like proteins (PPLPs), first described in the micronemes of *P. yoelii* sporozoites[241], contain a membrane attack complex/perforin (MACPF)-like domain, and were found both in ookinetes and sporozoites but not in merozoites. We suggest a role for these molecules in parasite entry to and/or egress from target cells, given the role of MACPF-like domains in the formation of pores. Both the ookinete and sporozoite can traverse through several host cells[157,242] whereas a merozoite enters a target cell only once. Our data therefore supports the concept that microneme proteins mediate motility and disruption of the host cell plasma membrane and the rhoptry proteins are essential to genesis of the parasitophorous vacuole and host cell survival.

We uniquely detected 472 proteins in replicative stages, *i.e.* blood stages and oocysts (Figure 2B). Not unexpectedly and consistent with findings in *P. falciparum*[11,91,92], the majority of these genes encode proteins involved in cell growth/division, DNA replication, transcription, translation and protein metabolism. The more detailed transcriptome analysis of blood-stage gene expression confirmed a cell cycle-related timing of transcription of these genes during the G1 and S/M-phase (Figures S18 and S19) and revealed that 215 and 355 were upregulated in the G1 and the S/M phases respectively.

During the first 18 hours of development, gametocytes and asexual trophozoites share the same features of the G1 phase of growth. Subsequently, the gametocytes differentiate into either males that prepare for DNA replication and mitosis, or females preparing for post-zygotic growth. Transcriptome analysis demonstrated that 58% of the G1 proteins (125 genes) and 59.4% (199 genes) of the S/M proteins were also upregulated in gametocytes (Figure S19) and the proteome data also emphasized the similarity between protein expression in asexual blood stages and gametocytes (514 proteins were shared between these stages; SOM Table S8). Despite these similarities, the described unique morphologies indicate sexual development is a fundamental developmental switch.

This is shown by the specific upregulation of transcription of 977 genes (SOM Figure S19, SOM Table S10) including many of the known gametocyte-specific genes and the detection of 127 unique proteins in the proteome (Figure 2C).

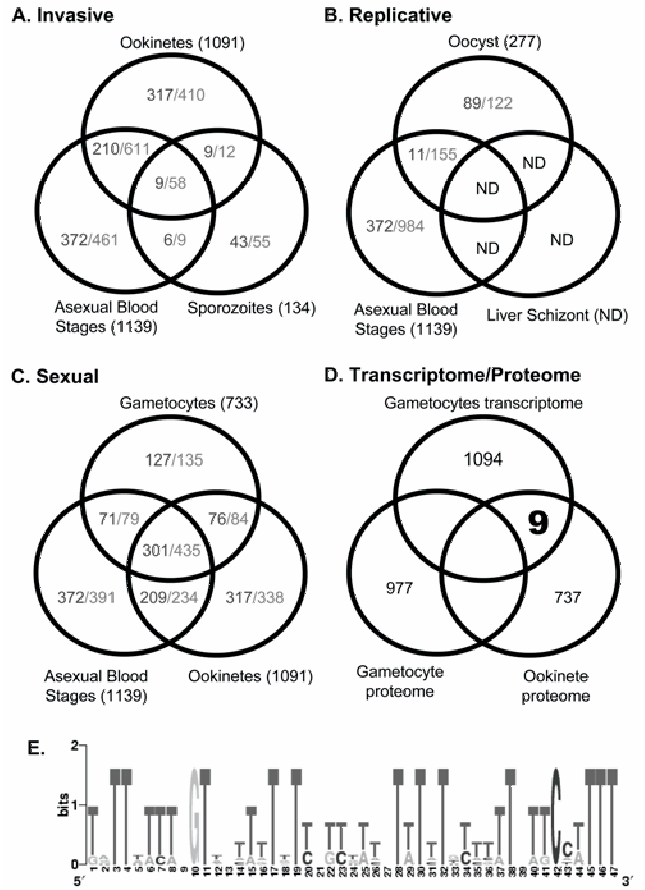


Figure 2. Different strategies of protein and gene expression during the malaria life cycle

Venn diagrams illustrate the overlap in proteins detected in the life stages involved in (A) invasion, (B) replication, and (C) sexual development. The total number of proteins detected in each stage is shown in parentheses. Numbers on the left represent proteins detected exclusively in the stages shown; numbers on the right represent proteins detected in the combination of stages shown out of the three stages included in each of the Venn diagrams (*i.e.*, these proteins could also be shared with stages not shown in the figure). ND, not done. (D) A Venn diagram representing the comparison of the gametocyte transcriptome with the proteomes of the gametocyte and the ookinete. The numbers indicate the individual transcripts and proteins in each analysis. The number of gametocyte proteins includes proteins identified in *P. berghei* during this study and proteins identified from *P. falciparum* gametocytes[11]. The bold number 9 in the intersect indicates the number of gametocyte transcripts found exclusively as ookinete proteins as a result of this study. (E) A WebLogo[243] representation of the 47-bp motif found within 500 bp downstream of the open reading frames (ORFs) of six of the nine implicated translationally repressed transcripts for which 3'UTR sequence was available. The point size of the letter is proportional to the frequency of the appearance of each nucleotide at each position.

Stage-specific expression

Just over half (948) of the proteins detected in the proteome analysis were found in one stage only, suggesting that stage-specific specialization is substantial. However, many of these stage-specific proteins belong to protein families whose expression is strategy-specific, reflecting both conserved mechanisms of parasite development between different stages and subtle molecular adaptations dictated by specific parasite-host interactions. For example, gene families encoding proteins containing MACPF-like or von Willebrand factor type A (vWA) and thrombospondin type 1 (TSP1) domains are examples of strategy (invasion)-specific expression whose members are stage specifically expressed. Unexpectedly, the PIR superfamily belongs to this category since members of the BIR protein family were detected in all stages; however, 92% were exclusive to a single stage (SOM Figure S15, SOM Table S8). Peptides were found matching 34 of ~180 predicted *P. berghei* genes and transcription of *bir* genes was detected in both asexual blood stage and gametocytes (SOM Tables S9 and S10). Although PIR are thought to play a role in immune evasion of the blood stages by antigenic variation[144], it is interesting to note that about 9% of the total BIR repertoire in our analysis is expressed only in the mosquito stages suggesting that these proteins may have other key functions.

Post transcriptional gene silencing

It has been proposed that transcripts in *Plasmodium* are essentially produced when needed[92], the so-called "transcripts to go" model[244]. However, it has been established that the abundant transcripts for *P28* in developing and mature female gametocytes are in a state of TR[167], one mechanism by which post transcriptional gene silencing is exercised. In addition, RNA binding proteins of the pumilio family (PUF proteins)[168] that play a role in TR are found in *Plasmodium* and are specifically upregulated in gametocytes and sporozoites[89,91]. Therefore, we compared the gametocyte transcriptome with the proteomes of both gametocytes and ookinetes to determine if additional gametocyte-specific transcripts might be subject to TR. Nine new genes were identified for which transcripts were detected in gametocytes but with protein products specific to the ookinete stage (Figure 2D, SOM Table S11). The analysis of the 3'UTRs of seven of these genes (for two genes there was insufficient 3'UTR sequence for analysis) and the 3'UTRs of *Pbs28* and *Pbs25* by the motif identifier programme MEME (multiple expectation maximization for motif elucidation)[245] revealed a 47-mer motif found in six of the analysed sequences within 1 kb of the 3' end of the stop codon (Figure 2E, SOM Figure S17; E-value = $4.8e^{+02}$). PUF proteins bind to a UUGU motif in 3'UTR regions[164,168] and the 3'UTR regions of all seven candidates and *Pbs28*, were enriched for this motif ($p \leq 0.001$), which was found as a sub-motif in the 47-mer motif. The 47-mer motif was used to search the entire *P. berghei* genome database using MAST[245], and 20 additional genes were identified that had the same motif within 1 kb of their 3'UTR (E-value < e^{-05}), giving a total of 29 TR candidates. Of these, 22 had orthologues in *P. falciparum*. Remarkably, 18 are upregulated in gametocytes (16 genes) and/or sporozoites (five genes) but only two were observed in gametocyte proteomes (SOM Table S11 and references therein). Analysis of 1 kb downstream of the stop codon of 20 of these *P. falciparum*

orthologues, including *pfs25* and *pfs28*, failed to identify a sequence analogous to the *P. berghei* motif. Nevertheless visual inspection identified numerous UUGU motifs at analogous positions. This lack of sequence similarity of the predicted 3'UTR binding motif is consistent with the significant sequence diversity in the predicted gene models of the *puf* orthologues of *P. falciparum* and *P. yoelii*[168]. The paucity of annotated transcription factors[51,164] (Chapter 3) and the phased expression of blood-stage transcripts have led to the proposal that post transcriptional gene silencing is a major mechanism of the regulation of gene expression in *Plasmodium*[164]. Our data suggest that at least in the gametocyte and possibly the sporozoite, TR may be an important component of these regulatory mechanisms.

The integration and initial analysis of the four datasets presented here has permitted novel insights concerning genome evolution, expression of gene families and mechanisms of post-transcriptional gene regulation in RMPs. This initial overview will be developed further and as demonstrated here will continue to emphasize the value of model systems for the study of orthologous features of human malaria parasites.

Acknowledgements

We thank J. B. Dame (University of Florida) for the gift of the *P. berghei* GSS library, J. Langhorne (National Institute of Medical Research) for providing *P. chabaudi* DNA, R. G. Sadygov (The Scripps Research Institute) for expert computer programming, and G. A. Butcher (Imperial College London) and M. J. Gardner (TIGR) for helpful advice with this manuscript. The authors acknowledge the Wellcome Trust, European Union, the Office of Naval Research, the US Army Medical Research and Material Command and the National Institutes of Health for financial support. J.D.R. and J.M. are funded by the Wellcome Trust, M.K. was supported by EU grants (RTN1-1999-00008 and QLK2-CT-1999-00753) and a grant from the NWO genomics initiative (050-10-053), and H.E.T. by the European Union MALTRANS consortium.

Notes

Supporting Online Material (SOM) accompanies the paper on the Science website (http://www.sciencemag.org/) and includes SOM Tables S1-11 and SOM Figures S1-S19. The sequences have been deposited with EMBL under the accession prefixes CAAI for *P. berghei* and CAAJ for *P. chabaudi*. All datasets are available through the official website of the *Plasmodium* genome project, PlasmoDB (http://plasmodb.org/)[165,166] and through GeneDB (http://www.genedb.org/).