Rodent Malaria Parasites:

Genome Organization & Comparative Genomics

Taco W.A. Kooij

Rodent Malaria Parasites: Genome Organization & Comparative Genomics
Kooij, Taco Wilhelmus Antonius
Thesis Leiden University, with summary in Dutch

ISBN-10: 90-9020399-0
ISBN-13: 978-90-9020399-7

Cover: Detail of the painting "Mosquito dance" by Hans Kanters

Rodent Malaria Parasites:

Genome Organization & Comparative Genomics

Proefschrift

ter verkrijging van

de graad van Doctor aan de Universiteit Leiden,

op gezag van de Rector Magnificus Dr. D.D. Breimer,

hoogleraar in de faculteit der Wiskunde en

Natuurwetenschappen en die der Geneeskunde,

volgens besluit van het College voor Promoties

te verdedigen op donderdag 9 maart 2006

klokke 15.15 uur

door

Taco Wilhelmus Antonius Kooij

geboren te Warmond in 1976

Promotiecommissie

Promotor:	Prof. dr. A.P. Waters
Referent:	Prof. dr. J.H. Adams *University of Notre Dame, IN, USA*
Overige Leden:	Prof. dr. P. ten Dijke Prof. dr. M.A. Huynen *Radboud Universiteit Nijmegen* Prof. dr. H.P. Spaink Prof. dr. H.G. Stunnenberg *Radboud Universiteit Nijmegen* Dr. A.W. Thomas *Biomedical Primate Research Centre, Rijswijk*

The printing of this thesis was financially supported by:
J.E. Jurriaanse Stichting
GlaxoSmithKline BV

You sleep in the light
Yet the night and the silent water
Still so dark…

Opeth

So much has been done,
exclaimed the soul of Frankenstein
- more, far more, will I achieve:
treading in the steps already marked,
I will pioneer a new way,
explore unknown powers,
and unfold to the world the deepest mysteries of creation.

Mary Shelley's Frankenstein

Contents

Abbreviations

Pb	=	*Plasmodium berghei*
Pc	=	*Plasmodium chabaudi*
Pf	=	*Plasmodium falciparum*
Py	=	*Plasmodium yoelii*
ACP	=	acyl-CoA binding protein
ACS	=	acyl-CoA synthetase
AMA1	=	apical membrane antigen 1
CAT	=	centrally located AT-rich
chr	=	chromosome
CLAG	=	cytoadherence-linked asexual gene
CS	=	circumsporozoite protein
CTRP	=	circumsporozoite TRAP-related protein
DHFR-TS	=	dihydrofolate reductase-thymidilate synthetase
EMP	=	erythrocyte membrane protein
EST	=	expressed sequence tag
ETRAMP	=	early transcribed membrane protein
ETS	=	external transcribed spacer
indel	=	insertions and deletions
GBP	=	glycophorin-binding protein
GFP	=	green fluorescent protein
GLURP	=	glutamate-rich protein
GPI-AP	=	glycosylphosphatidyl inositol-anchored protein
GSS	=	genome-survey sequence
HMM	=	hidden Markov model
ITS	=	internal transcribed spacer
LSA	=	liver-stage antigen
LSU	=	large subunit
MRCA	=	most recent common ancestor
mRNA	=	messenger RNA
MSP	=	merozoite surface protein
My	=	million years
OMP-DC	=	orotidine 5'-monophosphate decarboxylase
ORF	=	open reading frame
(RT-)PCR	=	(reverse transcription-) polymerase chain reaction
PEXEL	=	*Plasmodium* export element
PFGE	=	pulsed-field gel electrophoresis
PIR	=	*Plasmodium* interspersed repeat
PUF	=	RNA-binding proteins of the pumilio family
RAP	=	rhoptry-associated protein
(c)RMP	=	(composite) rodent malaria parasite
rRNA	=	ribosomal RNA
SB	=	synteny block
SBP	=	synteny breakpoint
SERA	=	serine-repeat antigen
SOM	=	supporting online material
SP	=	signal peptide
SSU	=	small subunit
STS	=	sequence tagged site
TAP	=	tandem affinity purification
TGF-β	=	transforming growth factor β
TM	=	transmembrane
TR	=	translational repression
TRAP	=	thrombospondin-related adhesive protein
tRNA	=	transfer RNA
TSTK	=	TGF-β receptor-like serine/threonine protein kinase
UTR	=	untranslated region
VICAR	=	*var* internal cluster-associated repeat
VTS	=	vacuolar transport signal

Chapter 1

Introduction

Malaria

Malaria is a devastating disease that has already been described by Hippocrates in ancient Greece, roughly 2,500 years ago. For long it was thought that the bad air (mal aria) from marshes was causing the disease. In 1880, Alphonse Laveran, a French surgeon working in Algeria discovered the malarial parasite in the blood of a patient suffering from malaria and in 1897, Ronald Ross, an English doctor born and working in British India demonstrated the transmission of avian malaria parasites by feeding female anopheline mosquitoes, which two years later was confirmed for humans by the Italian investigator Giovanni Grassi. There are four species of parasitic protozoa that cause malaria in humans of which *Plasmodium falciparum* is the most devastating and is responsible for the majority of deaths. The second most important malaria parasite for humans is *Plasmodium vivax*, which is found mainly in South-East Asia and South America but which is absent from large parts of Africa. In 1955, the World Health Organization (WHO) initiated an ambitious and intensive eradication programme. With a combination of mosquito control using dichlorodiphenyl-trichloroethane (better known as DDT) to prevent transmission and extensive treatment of malaria cases using anti-malarial drugs such as the highly effective and affordable drug chloroquine, one hoped to be able to deal with the malaria problem once and for all. Despite all the efforts, tropical countries are dealing with a strong resurgence of malaria during the past decades, such that more people are now suffering from malaria than ever before[1]. Different aspects have contributed to this resurgence, including (i) the emergence and rapid spread of drug-resistant malaria parasites[2-4] and insecticide-resistant mosquitoes[5]; (ii) factors that affect the public-health system, such as continuing political instability and war, unrelenting poverty and natural disasters; and (iii) more frequent transmission due to an increased (more than doubled) human population[6]. A recent extensive survey has shown that, in 2002, roughly 2.2 billion people were at risk of contracting *P. falciparum*, while a conservative estimate of 515 million became infected[1]. The majority of these cases (70%) occurred in Africa, while a significant 25% of the cases were reported in the densely populated South-East Asian region. The same authors estimated that, in 2000, 1.1 million Africans, mainly children under five years old, died from malaria[7], a number only challenged by tuberculosis. Apart from the human suffering, malaria is responsible for a significant economic burden and has been estimated to decrease economic growth by 1.3% annually[8].

Symptoms of the disease are a consequence of proliferation of the parasites in the blood where they infect red blood cells, resulting in complications such as anaemia, hypoglycemia, cerebral and placental malaria. The blood-stage infection is only one part of the complex life cycle shared by all parasites of the genus *Plasmodium*. The malaria life cycle is summarized below; a detailed description is provided in Chapter 2.

Parasites in the form of sporozoites are introduced into the blood with the saliva of a feeding mosquito and rapidly invade a liver cell where they develop and replicate to form over 10,000 daughter parasites (merozoites). Upon release merozoites will infect red blood cells where they will develop and divide into 16-32 new merozoites. This stage of the infection is responsible for the pathology typical

of the disease. Small numbers of the merozoites stop replication after erythrocyte invasion and develop into either a male or female sexually committed cell. After ingestion by another mosquito, these sexually committed parasites escape from the red blood cells transforming into gametes, fertilization takes place and the parasite traverses the midgut epithelium. In the midgut lining an oocyst is formed in which over 10,000 sporozoites develop that migrate to the salivary glands of the mosquito, ready to continue the cycle.

The problem of malaria and the aim of this study

Malaria has been under investigation for over a century, but despite the intensive research efforts no effective vaccine is available yet and people are still dependent on the few cheap and effective drugs in use that are losing efficacy rapidly due to drug resistance. It is vital to continue research efforts to generate drugs against previously successful targets and to identify and exploit new targets. The availability of an effective vaccine is generally seen as an essential tool to successfully combat this devastating infection, while alternative strategies may be developed and also employed to prevent transmission of the parasite.

The *P. falciparum* genome sequencing project was initiated with these goals in mind. The real-time release of partial genome sequences during the course of this 6-year project enabled researchers to identify unique *P. falciparum* genes that can serve as novel drug and vaccine targets[9,10]. The completion of the *P. falciparum* genome provided the malaria research community with an unprecedented opportunity to identify more *P. falciparum*-specific genes or genes that differ sufficiently from the host genes such that they may serve as targets for chemotherapeutic interventions with a decreased risk of side effects. In addition, the genome is predicted to encode a large number of proteins that would be dispersed to the surface of the parasite offering a much expanded range of potential vaccine candidates. Proteomic studies experimentally validated 70% of the predicted genes providing an insight in the evolution of the metabolic pathways utilized by the parasite and its unique features as compared with the human host[11,12]. The comparison of the *P. falciparum* and *Plasmodium yoelii* genomes provided a wealth of information on differences in genome organization stressing the importance of the subtelomeric regions in the generation of diversity in genes that allow the parasite to change and thereby evade recognition by the host immune system. The biggest advance has been made in the understanding of the biology of the parasite, its complex life cycle and the strategies employed for its survival in the variable environments. Whether one studies molecular evolution or gene transcription, population genetics or developmental biology, cellular mechanics or signal transduction, whole-genome information is what defines the playing field[13].

The aim of the studies described in this thesis was to compare and exploit the conservation of organization and gene content of the genome of malaria parasites that infect rodents with those of the human parasite *P. falciparum*. These rodent malaria parasites (RMPs) are widely used as research models and one of these, *Plasmodium berghei*, is the model for malaria research used in the Leiden Malaria Research Group (see also our website, http://www.lumc.nl/1040/research/malaria/malaria.html). A high level of conservation of genome organization and gene

content would validate the use of the RMPs for investigations to identify and characterize new vaccine and drug targets.

RMPs: models in malaria research

There are over 200 different *Plasmodium* species described infecting a wide range of hosts, including reptiles, birds, rodents, non-human primates and humans[14]. Only four species infect humans: *P. falciparum*, *P. vivax*, *Plasmodium ovale* and *Plasmodium malariae*, while the first two are a common cause of infection, the latter two are relatively rare.

It is possible to culture *P. falciparum* to study the disease-causing blood stages of the parasite. Technical and ethical considerations render the study of other stages of the infection, such as invasion of liver cells and transmission through mosquito feeding, virtually impossible. The use of model malaria species, for example infecting birds, rodents or non-human primates, can provide access to these less accessible stages of the malaria life cycle. Additionally, these systems allow the study of the infection *in vivo* and thus with all the complications associated with it, such as cerebral malaria (Ref. [15] for review) or responsiveness of the immune system. For example, the differential effects on the biochemistry, bioenergetics and gene expression in mice brains were examined following infection with cerebral and non-cerebral *P. berghei* strains[16]. Analysis of the competitive ability of *Plasmodium chabaudi* strains of variable virulence suggested that within-host competition is a driving force in parasite evolution where transmission efficiency is related directly to blood-stage parasite numbers, which may explain why many parasites harm their hosts[17]. Four *Plasmodium* species have been identified that have African tree rats as their natural host but which can also infect other rodents such as laboratory mice and rats. Three of these (*P. berghei*, *P. yoelii*, *P. chabaudi*) are widely in use as models in malaria research, mainly because of the relatively low costs and acceptible ethical concerns of *in vivo* experimentation (in comparison with model *Plasmodium* species that infect non-human primates). Many aspects of the biology, life cycle, and morphology of RMPs show a high level of similarity with the human parasites, validating their use as models for human infection. There is a high degree of conservation of metabolic pathways, which is reflected in a similar molecular basis of drug-sensitivity and resistance. In addition, many surface antigens of human parasites that are prime-candidate vaccine targets are also present in RMPs, such as TRAP and CS of sporozoites; CTRP, P25 and P28 of ookinetes; AMA1 and MSP1 of merozoites; and P45/48, P47 and P230 of gametes. *In vitro* culture techniques for large-scale production and manipulation of different life cycle stages are available, including the parts of the life cycle less accessible in the human parasites, such as the liver and mosquito stages. RMPs further allow *in vivo* investigations of parasite-host interactions as well as *in vitro* and *in vivo* drug testing, while the possibility to genetically modify parasites and the availability of well-characterized genetic background of mouse and rat and transgenic lines are invaluable for immunological studies. A high level of conservation of genome organization and gene content between *P. falciparum* and the RMPs would further validate the use of the RMPs for investigations to identify and characterize new vaccine and drug targets in these models. It was expected that a considerable proportion of the genome would be

conserved, reflecting the conservation of the complex life cycle of all *Plasmodium* species that infect mammals. Morphologically, little to no differences can be observed between the corresponding life cycle stages of different *Plasmodium* species. Many of the processes are shared; these include but are not restricted to the invasion of liver cells and red blood cells (though from different hosts), the sexual development necessary for transmission, fertilization, penetrating the mosquito midgut epithelium and migration to the salivary glands. Differences in gene content and genome organization will most likely be related to the adaptation of the parasites to their respective hosts. Relatively small differences between human and mouse or rat liver and red blood cell architecture, but more importantly, differences in the immune defence systems, will have forced the parasites to adapt, thus generating differences that we expect to find back in the genomic organization and gene content.

P. falciparum pre-genomics

The first malaria parasite genes were cloned at the beginning of the 1980s. Many of these genes encoded surface proteins that are exposed to the host immune system. In the early days of recombinant DNA technology, hopes were high that cloning important *Plasmodium* antigens would rapidly lead to development and production of an effective vaccine. Cloning of the first *Plasmodium* antigen, encoding a surface protein of the infective sporozoites (circumsporozoite protein) from a *Plasmodium knowlesi* cDNA library was a milestone in malaria research in 1983[18]. One year later, the *P. falciparum* orthologue of this gene was cloned[19], followed by other genes encoding potential vaccine candidates chiefly of blood stages of the parasite life cycle. Thereafter, there was a rapid increase in the amount of *Plasmodium* DNA sequence available in GenBank (http://www.ncbi.nlm.nih.gov/) - in 1990, there were roughly 70 entries and by 1995 that number had grown to more than 1,000, mainly from *P. falciparum* but also from *Plasmodium vivax* and other model *Plasmodium* parasites.

The individual chromosomes of *Plasmodium* could not be visualized by conventional microscopy, but their separation by pulsed-field gel electrophoresis (PFGE) revealed that the genome comprises 14 linear chromosomes in the size range of 0.5-3.5 Mb, resulting in an early estimate of the total genome size of about 25-30 Mb (~2.5x that of the yeast *Saccharomyces cerevisiae*[20,21]). PFGE analysis also revealed large variations in the size of the subtelomeric regions (Refs. [22,23] for reviews), which contain numerous and varied DNA repeats. The subtelomeric regions also contain species-specific gene families encoding proteins that are transported to the surface of infected erythrocytes and are involved in antigenic variation and immune evasion (Ref. [24] for review). By contrast, initial results of comparative mapping of housekeeping genes on the individual chromosomes revealed a high level of organizational conservation (synteny) between the core regions of the chromosomes of different *Plasmodium* species[25,26]. It was also demonstrated that *Plasmodium* possesses two non-nuclear genomes - a compact but anticipated mitochondrial genome of 6 kb and, surprisingly, a plastid-like 35-kb circular genome that was ultimately shown to reside in an organelle now known as the apicoplast[27].

Despite such advances, problems such as non-protective immune responses evoked by the chosen antigens and difficulties with vaccine antigen production all hindered the production of the hypothesized multi-stage cocktail vaccine[28]. In addition, many areas of the biology of the parasite remained insufficiently characterized. It was generally hoped that the sequencing of the genome might sidestep several of these problems. Following a successful multi-centre genome-mapping exercise[20], a genome sequencing consortium was established in 1997[29] working on the principle of real-time data deposition to allow the scientific community to benefit from the work-in-progress.

The genomics era and comparative genomics

The genomics era for eukaryotes started in 1996 with the publication of the complete genome sequence of the yeast *Saccharomyces cerevisiae*[21]. Since then, roughly 20 eukaryotic genomes have been sequenced, including those of mammals (that can be infected by malaria parasites) such as human[30], rat[31] and mouse[32], the tiger pufferfish[33], the sea squirt[34], insects like the fruit fly *Drosophila melanogaster*[35] and the malaria mosquito *Anopheles gambiae*[36], the nematode worms *Caenorhabditis elegans*[37] and *Caenorhabditis briggsae*[38] and plants including two rice species[39,40] and *Arabidopsis thaliana*[41]. Besides genome sequences of numerous prokaryotes, many of which cause disease in humans, complete genome sequences are now also available for a number of eukaryotic parasites. These include five apicomplexan parasites: *P. falciparum*[42], *Cryptosporidium parvum*[43] (infecting both humans and other mammals), *Cryptosporidium hominis*[44] (restricted to humans), *Theileria parva*[45], and *Theileria annulata*[46] (which both infect African cattle); three kinetoplastid parasites: *Trypanosoma brucei*[47] (which causes African sleeping sickness), *Trypanosoma cruzi*[48] (Chagas disease), and *Leishmania major*[49] (Leishmaniasis); and finally *Entamoeba histolytica*[50]. These data together with substantial amounts of partial genome sequence data, including partial genome sequences of three RMPs, *P. yoelii*[51] (Chapter 3), *P. berghei* and *P. chabaudi*[52] (Chapter 4), have been made publicly available through the websites of the sequencing centres and consortiums. It is expected that the volume of released sequence data will increase rapidly if not exponentially over the coming years.

The sequences of all these individual genomes contain a wealth of information that can be used by the scientific communities that study the different organisms. Analysis of single genomes has enabled the generation of hypotheses such as the whole-genome duplication of yeast[53] but comparative genome analysis has proved to be a powerful tool to test these theories. Comparative genomics can further improve our understanding of the genetic principles underlying the differences between both closely and more distantly related species and their evolutionary relationships by shedding light on the mechanisms that helped reshape their respective genomes including but not limited to: (i) micro-rearrangements such as single gene deletions, inversions and duplications; (ii) gross chromosomal rearrangements like translocations and deletions, inversions and duplications of entire segments resulting in loss of synteny; (iii) whole-genome duplications as shown for the yeast; (iv) ectopic exchanges in the (sub)telomeric regions of the chromosomes; and (v) the presence of recombination hotspots. Furthermore,

comparative genome analysis can significantly improve: (vi) the identification of both highly and less conserved orthologues either through homology or the analysis of syntenic segments; (vii) the identification of species-specific gene content which might be related to specific adaptations to environmental conditions; (viii) the identification of conserved non-coding elements, regulatory or structural; (ix) the annotation of genes, especially of small or complex multi-exon genes; and (x) assigning putative functions to hypothetical proteins. Many of these aspects can also aid in *Plasmodium* research as will be demonstrated below using examples from recently published comparative genome studies.

All comparative genome analysis is based upon the assumption that the two studied genomes originate from a common ancestor and that the respective genome sequences are the result of evolution acting on the ancestral genome sequence, *i.e.* a combination of the accumulation of random mutations and subsequent selection for example due to environmental pressures. Therefore, the resolving power of a two-sided whole-genome comparison to a large extent depends upon the proximity of the phylogenetic relationship between the species.

Global alignment comparisons between vertebrates (450 million years [My] divergence)[33] and comparisons between two Diptera, the fruit fly *D. melanogaster* and malaria mosquito *A. gambiae* (250 My divergences)[54] revealed that roughly 75% and 50% of the genes in the respective genomes being compared have orthologues. Gene contents of the fruit fly were also compared with the more distantly related nematode *C. elegans* and the yeast *S. cerevisiae* demonstrating that nearly 20% of the *D. melanogaster* genes have a putative orthologue in both other species reflecting a core eukaryotic gene set, however, all three genomes also appeared to contain approximately one-third of unique, species-specific genes without a homologue in either of the other species or itself[55]. Comparison between species from a single genus revealed that *C. elegans* and *C. briggsae* share 60% orthologues[38], while for four *Saccharomyces* species the amount of orthologues is as high as 95%[56]. Comparison of the *Plasmodium* gene content with that of the human host may provide insight the unique gene content and biological pathways that may be employed for the development new drugs, while knowledge of the common genes shared by parasite and host should help understand and hopefully prevent potential side effects induced by the treatment. The first comparison of the *P. falciparum* gene content with all sequences available in the public databases revealed that in contrast with other eukaryotes, as much as two-thirds of the *Plasmodium* gene content are unique[42]. This may reflect the larger evolutionary distance between *Plasmodium* and other eukaryotes, increased even more by the exceptionally high AT content of the genome. With the availability of genomes of more closely related species such as the Apicomplexa *C. parvum*[43]-*C. hominis*[44] and *T. parva*[45]-*T. annulata*[46] and more apicomplexan genome sequences like that of *Toxoplasma gondii* underway, this number is expected to decrease considerably. The identification of a *Plasmodium* core gene (Chapters 3 and 4) set should help validate the model species used in malaria research and identify common targets for drug interventions aiming to cure all four different malaria species infecting humans. The differences in gene content between different *Plasmodium* species will hopefully shed light on the molecular basis underlying species-specific traits such as host-specificity, differences in virulence (including

pathologically important phenotypes like sequestration, rosetting, or clumping) or transmission efficiency, hypnozoite formation in *P. vivax* or reticulocyte-preference. This is further exemplified by the comparative analysis of *Listeria monocytogenes*, the etiologic agent of listeriosis, a severe food-borne disease, with the non-pathogenic *Listeria innocua*. The presence of 270 *L. monocytogenes*- and 149 *L. innocua*-specific genes (clustered in 100 and 63 indels, respectively) suggests that virulence in *Listeria* results from multiple gene acquisition and deletion events[57]. Such a clear relation between gene content and virulence is not obvious from the comparison of the genome sequences of *Bacillus cereus*, an opportunistic pathogen causing food poisoning, and the animal and human pathogen *Bacillus anthracis*, which indicated the conservation of numerous factors for invasion, establishment and propagation of bacteria within the host expected for *B. anthracis* but not *B. cereus*[58]. Comparative analysis of the genome *Bordetella bronchiseptica*, which causes a chronic infection of the respiratory tract in a variety of animals, with the genomes of two closely-related bacteria causing whooping cough in humans (*Bordetella pertussis* and *Bordetella parapertussis*) demonstrated relations between genome organization and host-specificity[59]. During evolution of the host-restricted species, there has been extensive gene loss and inactivation. The authors also suggest a link between virulence and loss of regulatory functions.

The closest available pairs of eukaryotic genomes that are most fully sequenced to date are, for multicellular organisms, the free living nematodes, *C. elegans* and *C. briggsae* that diverged 80-110 My ago[38] and, for unicellular species, *S. cerevisiae* and three related yeast species, *Saccharomyces paradoxus*, *Saccharomyces mikatae* and *Saccharomyces bayanus*, that are thought to have a 5-20 My evolutionary distance[56]. In both comparisons the chromosomal ends appear to diverge more rapidly than the core regions of the chromosomes and the compared genomes show extensive co-linearity. The comparison of such closely related eukaryotic genomes reveals dynamic processes such as the constitution and perhaps origins of gene families. In *Caenorhabditis*, 96% of functionally organized gene clusters are conserved and the majority of diverged sequence consists of rapidly evolving repetitive DNA elements, which also account for the 4% difference in genome size. Furthermore, a direct comparison at the nucleotide level suggested a predicted increase of 1,300 *C. elegans* genes based purely on the identification of conserved regions between the two genomes[38]. The analyses of the different *Saccharomyces* genomes suggested the elimination of ~500 gene models of *S. cerevisiae* and further enhanced previous annotations by identifying 43 new putative, small genes (encoding 50-99 amino acids) and by redefining intron-exon boundaries, start and stop codons. Screening of the intergenic regions of the four yeast species for sequence motifs further revealed 72 genome-wide elements, including most known regulatory motifs but various new ones were also defined[56]. On a smaller scale comparative analysis has also been shown to help improve *Plasmodium* gene annotation by identifying genes in a complex and gene-dense region that were initially missed by the annotation algorithms[60]. This study also showed that using genome comparison it was possible to improve definitions of the intron-exon boundaries. The availability of multiple *Plasmodium* genomes can help train the annotation algorithms through identification of short but conserved coding regions.

Micro-rearrangements result from insertions, deletions and duplications of genes, repeat elements or other short DNA segments. The rate at which rearrangements occur seems to depend on the genomic location[32,61]. These local changes are most prominently present in (sub)telomeric and (peri)centromeric regions and happen at a much higher rate than gross chromosomal rearrangements. These highly recombinant regions mainly consist of tandem arrays of repetitive elements, including species-specific transposable elements (like in *D. melanogaster* and *A. gambiae*)[35,36] and recent gene duplications amongst primates. In the case of *P. falciparum*, subtelomeric regions are the main location for members of three gene families involved in immune evasion, the *var*, *rif*, and *stevor* families. Indeed, it has been suggested that the subtelomeric location of the *var* genes is essential for the process of antigenic variation in *P. falciparum*[62]. Though the nature of the repeats varies amongst different organisms, a relation with the genomic instability of these regions seems obvious.

A majority of the synteny breakpoints (SBPs) between human chromosome 19 and the mouse genome are located in regions with many repeats elements or clustered gene families[63]. Similar associations were found when different primate genomes were compared and, strikingly, many of the segmental duplications also seem to play a role in chromosomal rearrangements involved in human genetic diseases and polymorphisms[64,65]. *transfer rna* (*trna*) genes flank inversion breakpoints between four yeast genomes[56] and several repetitive elements in *C. elegans* could be associated with translocation and transposition events[66].

Gross chromosomal rearrangements helped reshape the organization of large synteny blocks (SBs). SBs are regions of conserved gene content and organization between different species, with the exception of micro-rearrangements like gene insertions, deletions or inversions. Within these syntenic regions the resolving power of comparisons can be greater facilitating the identification of both novel genes and conserved non-coding elements that control gene expression. While the genomes of four yeast species exhibit a relatively small number of one to five translocations[56], those of the nematodes *C. elegans* and *C. briggsae* are arranged in as many as 4,837 syntenic clusters[38]. Human and mouse have a predicted gene content that is 80% orthologous[32] arranged in 281 SBs larger than 1 Mb[67]. The presence of a large number of short “hidden” SBs, which are defined by closely located SBPs, led to the suggestion that mammalian genomes are mosaics of fragile regions with high propensity for rearrangements and solid regions with low propensity for rearrangements[68]. It has been estimated that at least 245 rearrangements of these SBs have occurred since the divergence of human and mouse[67]. Establishing if similar fragile regions exist in the genomes of malaria parasites could demonstrate additional mechanisms through which genetic diversity can be created as well as confirm the known generation of genetic diversity in the subtelomeric regions. Alternatively, the conservation of large genome segments between different *Plasmodium* species could indicate that there is a selective disadvantage to these gross chromosomal rearrangements perhaps because of some higher order organization of the genome[69].

Eichler and Sankoff[70] wrote a clear review on chromosomal dynamics of eukaryotic chromosome evolution also containing a synteny map of the mouse genome overlaid on top of the human genome while a more detailed description of

yeast evolution and comparative genomics was published recently by Liti and Louis[71]. Using the general principles set out in these two papers, we attempted to put our findings on the evolution *Plasmodium* genome organization and gene content in the perspective of what is known about eukaryotic genome evolution, noting the remarkable similarities as well as differences that exist between *Plasmodium* genomes and both extremes of eukaryote genome landscape.

* * *

Outline of this thesis

Malaria parasites that infect rodents are widely used models in the study of the biology of human malaria parasites and for the identification and characterization of targets for drugs and vaccines. The value of such studies using RMPs is dependent on the level of similarity between RMPs and the malaria parasites that infect man. The aim of the studies described in this thesis was to investigate the genome organization of the RMPs, with specific emphasis on *P. berghei*, in more detail and compare and exploit the organization and gene content of RMP genomes with those of the human parasite *P. falciparum*.

In **Chapter 2**, a review is given describing the current status of genomic and post-genomic research in *Plasmodium* summarizing the different genome sequencing projects and our understanding of the genome organization of different *Plasmodium* species, including the conclusions from the comparative genome analyses between RMPs and *P. falciparum* resulting from the investigations described in this thesis. In addition, this chapter contains a detailed description of the complex life cycle of the malaria parasite and many useful links to websites containing information on both genome and post-genome research in general and about malaria in particular.

Prior to the publication of the *P. berghei* genome sequence, investigations on the genome organization of *P. berghei* started with the characterization of the 14 chromosomes by separation using PFGE[26]. We in particular focussed on unravelling the genome organization of *P. berghei* chromosome 5 (Pbchr5), since several genes expressed in the sexual stages appeared to be clustered on Pbchr5[60,72]. Using a long-range restriction map of Pbchr5 and 15 markers, a physical map was generated. Simultaneously with the publication of the complete *P. falciparum* genome in 2002, partial sequence data for another RMP *P. yoelii* were released enabling the first-ever comparative genome analysis of genome sequences of two species belonging to the same genus. In this study presented in **Chapter 3**, the physical map of Pbchr5 was used to demonstrate the high level of conservation of the core region of this chromosome of the RMPs *P. yoelii* and *P. berghei* with that of large parts of only two chromosomes (Pfchr4 and 10) of the human parasite *P. falciparum*. This showed for the first time in detail the high level of synteny (conservation gene content and organization with the exception of microrearrangements) between rodent and human malaria parasites. This study also provided the first clues that the subtelomeric regions of chromosomes of RMPs are highly divergent from those of *P. falciparum* and that these regions are

separated by distinct boundaries from the core regions that show a high level of synteny between the rodent and human malaria parasites. Interestingly, in these variable regions many species-specific gene families are located.

After publication of the first RMP genome of *P. yoelii*, the partial genome sequences of two additional RMPs, *P. berghei* and *P. chabaudi*, have been published, which is presented in **Chapter 4**. Comparison of these genome sequences with that of the other RMP *P. yoelii* and that of the human parasite *P. falciparum* showed a high level of conservation of gene content. At least 4,500 of the 5,300 genes of *P. falciparum* have an RMP orthologue (the core *Plasmodium* gene set) and are localized in the core regions of the chromosomes (the central, non-subtelomeric regions). A majority of the 736 *P. falciparum* genes without an RMP orthologue belong to one of the *P. falciparum*-specific gene families; 161 are located within the core regions disrupting synteny while 575 are located in the subtelomeric regions of the chromosomes. These subtelomeric genes could be assembled into 12 distinct gene families only five of which are shared with the RMPs.

The availability of the genome sequences of three RMPs and a completely annotated *P. falciparum* genome made it possible to generate a detailed genome-wide synteny map of four *Plasmodium* species. This study is described in **Chapter 5** and shows that the organization of the core regions of the RMP and *P. falciparum* genomes are highly conserved in as little as 36 SBs. Analysis of these SBs showed that the organization of *P. falciparum* genome could be generated from that of the composite RMP (cRMP) genome in a minimum of 15 chromosomal recombination events and *vice versa*. This relatively low number of only 15 rearrangements suggests that gross chromosomal rearrangements resulting in the loss of or change in synteny is infrequent in *Plasmodium*. Moreover, the locations of both centromeres and boundaries between the conserved core regions and variable subtelomeric regions are conserved between the RMP and *P. falciparum*. The 168 non-subtelomeric *P. falciparum*-specific genes (161 genes reported in Chapter 4 plus seven genes of the newly discovered *vicar* family) disrupting synteny were analysed in more detail. Of these genes, 42 are located between the SBs at SBPs while 126 are located in so-called indels disrupting the syntenic regions. Interestingly, 68% of these genes are potentially exported to the surface of the parasite or infected erythrocyte and several belong to gene families, including two newly discovered gene families. These results show that not only subtelomeric regions but also SBPs and indels can be foci for species-specific genes with a role in host-parasite interaction and immune evasion and suggest involvement of gross chromosomal rearrangements in the generation of *P. falciparum*-specific gene families. This is exemplified by the discovery of *P. falciparum*-specific gene family consisting of 21 copies that encode transforming growth factor β (TGF-β) receptor-like serine/threonine protein kinases (PfTSTK) with only a single syntenic orthologue in the RMPs. Combination of the suggested 15 recombination events with phylogenetic analysis of the TSTK protein sequences provided insights in the mechanisms underlying the generation of this gene family.

The studies described in Chapters 2 to 5 have been initiated with the characterization of the genome organization of Pbchr5 since this chromosome contained a number of genes that are exclusively expressed during sexual development[72]. A detailed analysis of a 13.6-kb region, the B9 locus, of *P. berghei* containing six tightly clustered genes, three of which are exclusively expressed during the sexual stages of the parasite, revealed high levels of conservation with its *P. falciparum* counterpart on Pfchr10. The gene number, organization of the intron-exon boundaries of the four multi-exon genes and expression patterns are entirely conserved[60]. We have been trying to investigate these genes in more detail by gene modification technologies. The results of these studies have not been published yet but will be discussed briefly in Chapter 7. Analysis of the gene content of Pbchr5 revealed the gene encoding α-tubulin II, which is also expressed during sexual development. Malaria parasites have two genes that encode α-tubulins, one of which, α-tubulin I, is expressed constitutively and is located on Pbchr4, while the second one, α-tubulin II on Pbchr5, is highly expressed in male gametocytes and there is evidence for a specific function in the formation of the axoneme of the male gamete. We have characterized both *P. berghei* genes and tried to analyse the precise role of α-tubulin II in sexual development of particularly the male gametocytes by gene modification, which is described in **Chapter 6**. Surprisingly and despite its importance for male gamete formation, α-tubulin II is not exclusively expressed during sexual development but is also essential for normal asexual development of the blood stages.

In **Chapter 7**, the results of our studies on the genome organization of *P. berghei*, that were initiated with investigation of the organization of Pbchr5, and the comparative genomics studies are summarized and discussed. In addition, some studies are mentioned that were aimed at characterization of individual sex-specific genes that are located in the gene-dense B9 locus on Pbchr5 that is highly conserved between the RMP and *P. falciparum*.

Chapter 2

Plasmodium post-genomics - better the bug you know?

Taco W.A. Kooij, Chris J. Janse and Andrew P. Waters

Malaria Research Group, Department of Parasitology, Centre for Infectious Diseases, Leiden University Medical Centre, Albinusdreef 2, 2333 ZA, Leiden, The Netherlands.

Preface

Since the publication of the sequence of the genome of the major causative agent of human malaria, *Plasmodium falciparum*, numerous post-genomic studies have been completed. Invaluably, these data can now be analysed comparatively due to the availability of a significant amount of genome sequence data from several closely related model species of *Plasmodium* and accompanying global proteome and transcriptome studies. This review summarizes current knowledge and how this has been - and may be - exploited in the search for vaccines and drugs against this most significant infectious disease of the tropics.

Introduction

It cannot be restated too often that malaria, principally through infection by the protozoan apicomplexan parasite *Plasmodium falciparum*, is responsible for more than one million deaths each year. A recent survey has shown that, in 2002, roughly 2.2 billion people were at risk of contracting *P. falciparum* infection, while a conservative estimate of 515 million individuals became infected[1]. 70% of these cases occurred in Africa and it was estimated that, in 2000, 1.1 million Africans, mainly children under five years old, died from malaria[7].

Clearly a disease of poverty, the dwindling effectiveness of frontline affordable drugs and the continuing lack of a vaccine mean that malaria poses a greater problem now than at any time since the failure of the WHO's eradication drive in the 1950s. It is in this context that malaria research must act. Basic biological investigation of malaria parasites has always offered the promise of the development of new therapeutics (chiefly envisaged as both vaccines and drugs). Although biologists initially worked on single genes of therapeutic interest, the landmark publication of the complete *P. falciparum* genome with first-pass annotation in late 2002[42] irrevocably changed the face and practices of malaria research. This comprehensive dataset, combined with the availability of substantial sequence tracts from the rodent malaria parasite (RMP) *Plasmodium yoelii*, has made it possible to embrace the latest global genome-survey technologies[51].

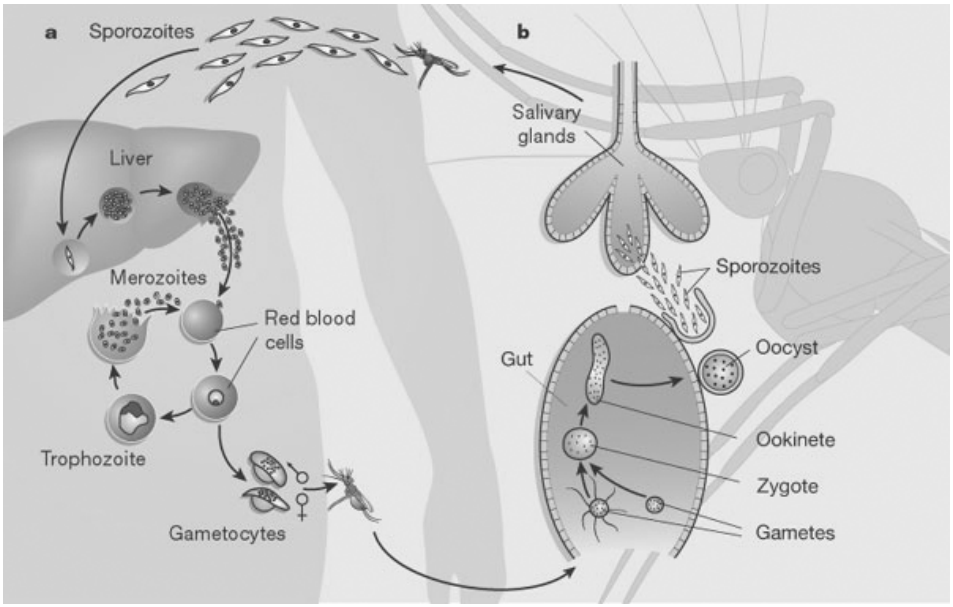
Thanks to the admirable real-time release policy of the three sequencing centres involved, this embrace was immediate and two detailed proteomic studies accompanied the genome into press[11,12]. We will attempt to review here the current status of *Plasmodium* genomic and post-genomic research, indicating trends and the future requirements necessary to build a detailed "virtual parasite" in which we are fully informed of the molecular biology that underpins its biology and interactions with both host and vector.

The *P. falciparum* genome

The major pre-genomic milestones in *P. falciparum* research are summarized in the accompanying timeline (Figure 1), and the complex life cycle of a *Plasmodium* parasite is illustrated in Box 1.

The *P. falciparum* genome comprises 14 linear chromosomes and two non-nuclear genomes - a compact mitochondrial genome of 6 kb and a plastid-like, 35-kb circular genome that resides in an organelle now known as the apicoplast[27]. The genome of the 3D7 strain of *P. falciparum* was the first parasite genome to be sequenced to completion[42]. A chromosome-by-chromosome shotgun sequencing

Box 1: The *Plasmodium* life cycle



Plasmodium species all share the same life cycle, whose stages are depicted in the figure. The haploid sporozoites enter the bloodstream with the saliva of a feeding anopheline mosquito. After invasion of a liver cell, the parasite undergoes a period of growth (G1 phase of the cell cycle), followed by multiple genome replications and mitotic divisions (S/M phase) resulting in the production of 20-40,000 daughter parasites, merozoites, that invade erythrocytes after release from the hepatocyte in the circulation. After a period of intra-erythrocytic growth several mitotic divisions result in the production of 16 to 32 new merozoites. The duration of the erythrocytic cycle and the number of merozoites produced are species-specific. The asexual blood-stage parasites are the cause of pathology.

In the blood, a small percentage of parasites develop into sexually committed cells, the female and male gametocytes that are the precursor cells of the gametes. These cells are arrested in G0 phase and are only activated to produce gametes in the midgut when ingested as part of the bloodmeal of the female anopheline mosquito. Gamete formation is followed by fertilization, resulting in the diploid zygote. After one round of meiotic division, the zygote develops into a motile ookinete, penetrates the cells of the midgut and traverses the midgut wall to form an oocyst on the basolateral lamina. After growth and multiplication, >10,000 sporozoites develop in a single oocyst. When the motile sporozoites are released, they migrate via the haemolymph to invade the salivary glands, from which they are ready to infect a new victim and continue the cycle. (Figure reproduced from Wirth, *Nature*, 2002[73].)

strategy was employed based on pulsed-field gel electrophoresis (PFGE)-separated chromosomes. Initially, *P. falciparum* chromosomes 2[74] and 3[75] were published but, chiefly due to the extreme AT bias (80%) of the genome, it took almost seven years before the genome was complete. The full genome sequence revealed that the 22.8-Mb nuclear genome harbours 5,268 predicted protein-encoding genes on 14 chromosomes ranging in size from 0.64-3.3 Mb at an average frequency of one gene every 4.3 kb. Matching expressed sequence tags

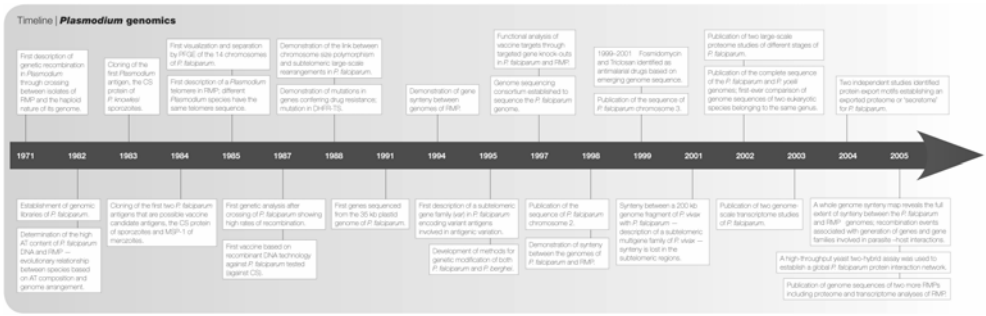


Figure 1. Timeline: *Plasmodium* genomics

(ESTs) and proteomic analyses were used to experimentally validate approximately 70% of these genes. At the time of publication, there were still 79 gaps in the sequence. The largest contiguous DNA sequence (contig) was chromosome 12 (2.3 Mb). Now, just three chromosomes (7, 8, and 13) are still awaiting closure, with fewer than ten gaps remaining; chromosome 14, at 3.3 Mb, is the largest contig (http://www.sanger.ac.uk/Projects/P_falciparum/status.shtml, November 2005; M. Berriman, personal communication).

Comparison of the first-pass annotation of the *P. falciparum* nuclear genome with other genomes showed that 60% of the predicted genes could not be assigned functions. The products of at least 1.3% of the *P. falciparum* genes are known to be involved in cell-to-cell adhesion or invasion of host cells and a further 3.9% are postulated to play a role in evasion of the host immune response; many of these 250-300 proteins possess host-like extracellular adhesion domains. Curiously, only 8% of the *P. falciparum* genes could be assigned functions in metabolism, in contrast to 17% of the genes of the yeast *Saccharomyces cerevisiae*[21]. This is possibly due to sequence divergence or a consequence of the parasitic life style of *Plasmodium*. From the first-pass annotation, the *P. falciparum* complement of transport molecules appeared to be similarly reduced compared with free-living organisms.

Approximately 10% of the nuclear-encoded proteins are predicted to target to the apicoplast[42,76,77]. The apicoplast is thought to serve as an organization centre for certain metabolic pathways (isoprenoid and fatty acid biosynthesis), was probably acquired from non-green algae by endosymbiosis, and is found in many apicomplexan parasites (Refs. [78,79] for reviews). The evolutionary origin of the apicoplast immediately suggested potential drug targets based on anti-fungal agents.

Numerous gene families accounted for 5-10% of all genes, and include gene families involved in immune evasion and sequestration (*var*[80-82], 59 members) as well as putative variant antigens (*rif*; 149 members, and *stevor*, 28 members[83,84]). The majority of these genes are distributed in the subtelomeric regions of the chromosomes, internal to the complex species-specific repeats that abut the telomere repeats[42]. All chromosomes harbour some copies of one or more of the different gene families, but the composition and order vary. *P. falciparum*

erythrocyte membrane protein 1 (PfEMP1), encoded by the *var* genes, is demonstrably associated with clonal antigenic variation, a strategy used in the erythrocytic stages to evade the adaptive host immune response. PfEMP1 also plays an important role in the sequestration of infected erythrocytes in capillaries in the brain (resulting in cerebral malaria) and other tissues, including the placenta. The fact that 60% of the *var* genes have a subtelomeric location should facilitate the generation of diversity in the gene repertoire of this family[85,86] and the high frequency of (ectopic) recombination in these regions could also contribute to the observed size variation in the subtelomeric regions. There are also seven non-subtelomeric *var* loci containing between one and seven copies, regularly interrupted by *rif* genes (Figure 2 and Appendix 1).

Smaller gene families encoding diverse functions are also found (Appendix 4) - for example, genes encoding acyl CoA synthetases (11 members) and receptor-associated protein kinases[87] (21 members; Appendix 1). The genome sequence also expanded our knowledge of the extent of gene families encoding vaccine candidates and may yet reveal new antigens for vaccine development. Before the sequencing initiative, the 6-Cys family, containing transmission-blocking vaccine candidates, was thought to have three members (P48/45, P12, and P230); we now know that ten members are expressed at different stages of the life cycle, but principally in gametocytes (six members; Ref. [88] for review).

The *P. falciparum* transcriptome

Several DNA microarray studies have been published, ranging from analysis of gene transcription using random clones selected from genomic DNA (gDNA) libraries[89,90] to more recent quasi-global surveys of transcription[91,92] based on oligonucleotides designed using the emerging genome sequence. The arrays were used to probe RNA from defined parasite stages. Additionally, gene transcription in defined parasite stages has been investigated by several EST surveys[93-96], by different techniques creating and analysing stage-specific enriched cDNA libraries[97,98], and through serial analyses of gene expression (SAGE)[99,100].

Two microarray studies, analysing transcription of the blood stages, revealed that the parasite transcribes a large core of the genome during blood-stage development (88%[91] or 60%[92] of the genes present on the chips) and there are clear patterns of stage-specific gene transcription. Both studies showed remarkable concordance in gene transcription patterns and pinpointed genes encoding surface proteins that might serve as vaccine candidates for specific parasite stages. The more detailed analysis carried out by Bozdech *et al.*[92] suggested a cascade of transcriptional activation during blood-stage development, where transcripts are produced in an ordered manner, as and when needed, to fulfil the demands of the cell (the “transcripts-to-go” model). This raises the hope that inhibition of a few key early transcription factors might provide a means to arrest parasite development, a concept that remains generally valid despite the implications of the more recent discovery of translational repression (TR) in certain life cycle stages[52,101] (Chapter 4). The authors of both microarray studies proposed that the groups of genes with similar transcription profiles might be involved in similar functions or cellular processes, perhaps giving insight into the role of those genes whose function was not revealed by annotation. A study based on SAGE technology revealed an

unusual feature of blood-stage transcription - the significant accumulation of anti-sense transcripts in a stage-specific fashion[99,100]. At present, this is thought to represent some level of gene regulation, but the mechanism and indeed proof of function of anti-sense RNA in *Plasmodium* remain obscure at present.

By including two invasive forms of the parasite in their analysis (sporozoites and merozoites), Le Roch *et al.*[91] were able to identify gene groups associated with cell invasion, emphasizing the similarities of these stages with stage-specific modulation of gene transcription patterns according to context (invasion of erythrocytes by the merozoite and invasion of either the salivary glands of the mosquito or the liver by the sporozoite).

Continued data mining of the published *P. falciparum* transcriptome, in addition to new transcriptome studies of defined developmental stages or mutant parasites, will provide a better understanding of the biology of the malaria parasite. Some examples of such studies are the analysis of the antioxidant defence system[102], the pentose phosphate pathway[103], the transcription of variable antigens[104], and detailed analysis of gametocyte development[105-107].

The *P. falciparum* proteome

Several detailed high-throughput mass-spectrometry studies of the *P. falciparum* proteome have been published and have given an insight into the biology of five different parasite stages. Reassuringly, the protein contents of the different blood stages agree well with the presence of transcripts of the genes encoding these proteins[91,92] and therefore the proteome data generally support the “transcripts-to-go” model.

Florens *et al.* characterized the proteome of four stages: sporozoites, merozoites, trophozoites, and gametocytes[11]. Of 2,415 proteins identified, only 6% were expressed in all four stages, whereas more than half were unique to one particular stage. Almost half of the sporozoite proteins were stage-specific, whereas for the blood stages the numbers range from 20-33%. This demonstrates the highly specialized nature of the different life cycle stages that are adapted to interact with different cell types of two different hosts.

The proteome of the sexually developing parasite was described in more detail by Lasonder *et al.*[12]. Comparison of these data sets, with the help of annotation, identified candidate proteins for transmission-blocking vaccines, such as a family of genes containing limulus coagulation factor C (LCCL) domains and predicted lectin domains. These proteins, initially thought to be exclusive to gametocytes and gametes, are also expressed in ookinetes and are essential for oocyst development[108,109], suggesting they have a role in the interactions of the parasite with the mosquito midgut epithelium.

The annotation of the genome has benefited considerably from the proteome studies; for example, Lasonder *et al.* reported a set of peptides with significant matches in the *P. falciparum* genome that were not predicted by computational methods[12] and further analysis of these peptides is ongoing (E. Lasonder, personal communication). Both transcriptome and proteome data revealed no tendency for the genome to be compartmentalized into regions containing genes that are co-ordinately expressed, ruling out an operon-like organization; however, a tandem

array of five genes encoding proteins located in the Maurer's clefts (MCs) has been reported[110].

In addition to the reported proteomes of the whole life cycle stages, proteome studies have also focussed on specific organelles and structures. For examples, the proteomes of infected erythrocyte membranes[111] and the MCs[110] of mature trophozoites and schizonts have been investigated, revealing 36 and 50 candidate proteins, respectively. MCs are parasite-derived membranous structures in the erythrocyte cytosol that are thought to be involved in parasite protein transport to the erythrocyte surface[112]. Perhaps surprisingly, the two datasets share only four proteins, which could be a reflection of the different methods used to detect these proteins. Alternatively, this lack of proteome overlap suggests that the parasite proteins on the erythrocyte surface are only transiently associated with the MC, or that proteins residing in the MCs are more easily detected. Comparison of the *P. falciparum* genes encoding MC proteins with the genes of RMPs using the *P. falciparum*-RMP synteny maps revealed that 36 of the 50 genes (72%) are syntenic with the RMPs or belong to locally expanded gene families shared between the different species (T.W.A.K., unpublished observations). The relatively large proportion of syntenic orthologues encoding MC proteins indicates that a considerable part of the protein export machinery is conserved between *P. falciparum* and the RMPs.

Furthermore, the proteome of gradient-purified detergent-resistant membranes of mature blood-stage parasites (late schizonts/merozoites) has been analysed[113]. These membranes are greatly enriched in glycosylphosphatidyl inositol-anchored proteins (GPI-APs) and their putative interacting partners. GPI-APs coat the surface of extracellular *P. falciparum* merozoites and several are validated candidates for inclusion in a blood-stage malaria vaccine. In addition to detecting confirmed GPI-APs, this study identified new GPI-APs and several other novel, potentially GPI-AP-interacting proteins that are predicted to localize to the merozoite surface and/or apical, invasion-associated organelles (rhoptries and micronemes).

P. falciparum protein interaction networks and protein structure

Understanding the interactions between proteins can provide insights into the function of, and functional relationships between, these proteins. Recently, the first studies have been published on large-scale analysis of interaction between proteins of the asexual blood stages of *P. falciparum*[114,115]. Using a high-throughput yeast two-hybrid assay, 2,846 interactions were identified involving 1,312 largely uncharacterized proteins. By combining information on protein interactions with patterns of co-expression and putative function, informed by annotation and the presence of specific domains, groups of interacting proteins were identified that play a role in chromatin modification, transcription, messenger RNA (mRNA) stability and ubiquitination, and invasion of host cells. Comparing the *P. falciparum* protein networks with those of yeasts, nematodes, bacteria, and insects showed little conservation of the complexes between *P. falciparum* and these other organisms (three in yeast, none in the others), whereas yeasts, insects, and nematodes share substantial numbers of conserved complexes with each other. However, 29 highly connected *P. falciparum*-specific protein complexes were

identified, suggesting that the patterns of protein interaction in *Plasmodium*, like its genome sequence, are quite different from other species, although it is anticipated that many will prove to be conserved in other Apicomplexa.

Improved insights into structure-function relationships of increasing numbers of proteins might reveal new drug targets. Several groups have begun initial attempts to achieve a larger-scale protein structure analysis by generating expression libraries of soluble proteins. The Structural Genomic Consortium has initiated an admirable initiative to attempt a high-throughput elucidation of 3D structures of *Plasmodium* proteins. The structural data produced are freely available (http://www.sgc.utoronto.ca/) and so far 19 proteins from different *Plasmodium* species and other apicomplexan parasites have been resolved.

The *P. falciparum* "secretome" and "permeome"

Malaria parasites secrete proteins across the vacuolar membrane into the erythrocyte cytosol or to the erythrocyte membrane, inducing modifications of the erythrocyte that are necessary for parasite survival, but which are also associated with disease. Two studies have independently identified a conserved sequence motif in such secreted proteins, termed either the *Plasmodium* export element (PEXEL)[116] or the vacuolar transport signal (VTS)[117]. Bioinformatics using the PEXEL/VTS signal sequence predicts a "secretome" of 300-400 proteins for *P. falciparum* (~8% of all genes). In addition to 225 *var*, *rif*, and *stevor* genes, the secretome includes 160 genes encoding proteins likely to be involved in remodelling of the host erythrocyte, including heat-shock proteins, kinases, phosphatases, and putative transporters[116], thus vastly expanding the number of potential vaccine and drug targets. The PEXEL/VTS motif seems to be distinct from known cellular transport signals, which suggests that it might be a novel eukaryotic secretion signal associated with intracellular parasites.

Besides the transport of parasite proteins to the erythrocyte, the intra-erythrocytic parasites need to take up nutrients from the erythrocyte cytosol and excrete metabolic waste products. Membrane-transport proteins mediate these processes but are also implicated in antimalarial drug resistance. Furthermore, the parasites will need ion channels to maintain their ion homeostasis. The initial annotation of the *P. falciparum* genome identified only a limited number of transporters and no channels. By combining different bioinformatic approaches, based on the hydropathy plots of proteins, several putative ion channels and more than 100 membrane transport proteins were identified, including equal numbers of known and candidate transporters for a range of organic and inorganic nutrients that had not been annotated previously[118]. The term "permeome" was used to describe the total complement of proteins involved in membrane permeability.

Global genome polymorphisms

Microsatellite sequences are small simple polymorphic tandem repeat sequences distributed throughout the genome that provide sensitive fingerprints for all genomic loci. The *P. falciparum* genome demonstrates a microsatellite frequency of one every 1-2 kb, effectively creating a high-resolution linkage map that can be exploited for further genetic characterization. This revolutionized the application of microsatellites to population genetics studies, which had previously been restricted

to an examination of a few loci of specific interest. The study of microsatellite polymorphisms can now involve a whole-genome survey that can be used to infer conclusions about subjects as broad as the evolutionary history of the parasite (sampling multiple genomes and microsatellite loci), or as specific as the isolation of genes associated with traits such as drug resistance (through the analysis of the progeny resulting from the crossing of distinct cloned parasite lineages). Thus distinct population structures associated with geographical distribution have been established[3,119] and such populations can continue to be monitored to measure rates and patterns of gene flow.

Microsatellite distributions were successfully used to isolate *pfcrt*, a gene important for chloroquine resistance[120,121] and linkage-disequilibrium studies demonstrated that this gene was under strong selective pressure[3]. Combining single-nucleotide polymorphism (SNP) studies with linkage disequilibrium can also help to identify genome loci that influence complex phenotypes, such as chloroquine resistance. For example, Mu *et al.*[122] demonstrated that 11 of 49 genes encoding putative ATP-binding cassette (ABC) transporters exhibited significant linkage disequilibrium associated with decreased sensitivity to chloroquine and/or quinine. Analysis of SNPs spanning chromosome 3 of 99 field isolates of *P. falciparum* showed high variation in recombination rates among populations and along the chromosome. Between the different populations conserved recombination hotspots were found at the chromosome ends[123]. Further genome-wide studies of microsatellite and SNP frequencies analysing recombination and linkage disequilibrium within parasite populations might reveal novel associations between genes and phenotypes.

Comparative genomics

While analysis of a single genome provides tremendous biological insights for any given organism, comparative analysis of multiple genomes can provide substantially more information on the physiology and evolution of genomes, and expand the ability to identify and assign putative functions to predicted coding regions. Orthology recognition is becoming increasingly sophisticated and bioinformatic methodologies to improve *Plasmodium* annotation through the recognition of global orthologies have been developed to discover and annotate biosynthetic pathways[93,124]. Comparative genomics can also help to identify orthologous genes or refine gene predictions through local alignments, thus substantially improving multi-exon gene models[51,60]. When closely related species within a single genus are compared, this should provide additional levels of insight, for example, into the repertoire of species-specific genes that might be associated with differences in life style, such as the invasion of reticulocytes versus normocytes by *Plasmodium* merozoites, and even into speciation.

Animal models of malaria have long been established as alternative means to gain insights into the biology underlying the parasite-host/vector interactions that cannot be obtained readily or ethically working with the human malarias *P. falciparum* or *P. vivax*. In addition to several primate malarias (for example, *Plasmodium reichenowi*, a close relative of *P. falciparum* that infects chimpanzees, and *P. knowlesi*, which is more closely related to *P. vivax*) and a chicken parasite (*Plasmodium gallinaceum*, that has an intriguing phylogenetic relationship with all

four human malarias[125,126]), much work is done using RMPs as they are cheaper to maintain *in vivo* and there are fewer ethical concerns in the handling of their host organisms. Significant amounts of genome data are available not only for all of the aforementioned parasites (Table 1), but also for the second most important human malaria parasite, *P. vivax*. Its genome sequence has almost been completed and annotation and analyses are drawing to a close (http://www.tigr.org/tdb/tgi/, November 2005; Jane Carlton, personal communication); an update on the status of the project was published in May 2003[127]. These extensive genome datasets of different *Plasmodium* species have not only facilitated comparative genomics, but have also given rise to significant post-genome studies, characterizing both the transcriptome and proteome of different life cycle stages. Comparison of the *P. falciparum* genome with other genome data available in 2002 showed that 60% of the annotated *P. falciparum* genes could not be assigned functions and hence could encode functions that are unique to *P. falciparum* or to the genus *Plasmodium*. More recently, the genomes of several other unicellular parasites have been published, allowing cross-genus genome comparisons of closely related parasites to be performed. The list of unicellular parasites for which significant amounts of genome sequence are now available includes two apicomplexan parasites infecting humans, *Cryptosporidium parvum*[43] and *Cryptosporidium hominis*[44], two apicomplexan parasites infecting cattle, *Theileria parva*[45] and *Theileria annulata*[46], *Entamoeba histolytica*[50], and three kinetoplastid parasites, *Trypanosoma brucei*[47], *Trypanosoma cruzi*[48], and *Leishmania major*[49], while the genome sequence of *Toxoplasma gondii* is nearing completion.

The first comparative analysis of the genomes of two apicomplexan species, *P. falciparum* and *C. parvum*, showed that both lineages have acquired protein-adhesion domains, originating from proteins of their animal hosts, and identified at least 145 apicomplexan-specific genes[128]. Initially, comparative genome analyses of *Theileria*, *Cryptosporidium*, and *Plasmodium* species with other public genome databases indicated that the genomes of all three apicomplexan lineages have an

Table 1: Current status of the *Plasmodium* genome projects (November 2005).

Species	Genome Size	Coverage[a]	Max Contig Size	Annotated full length genes	Sequencing Centre
P. falciparum (3D7 strain)	22.8 Mb	14.5x	3.3 Mb	5,260	SGTC,TIGR, WTSI
P. falciparum (Ghanaian clinical isolate)	23 Mb	8x	NA	NA	WTSI
P. falciparum (IT strain)	23 Mb	Ongoing (1x)	NA	NA	WTSI
P. vivax	30 Mb	10x	2.1 Mb	5,431	TIGR
P. reichenowi	25-27 Mb	Ongoing	NA	NA	WTSI
P. knowlesi	25 Mb	8x	NA	NA	WTSI
P. berghei	25 Mb	8x	37 kb	4,617	WTSI
P. chabaudi	25 Mb	8x	17 kb	4,100	WTSI
P. yoelii	23.1 Mb	5x	51 kb	4,034	TIGR
P. gallinaceum	25 Mb	3x	NA	NA	WTSI

[a] Average number of sequence reads per nucleotide.

Abbreviations: NA, not available; SGTC, Stanford Genome Technology Center; TIGR, The Institute for Genomic Research; WTSI, Wellcome Trust Sanger Institute.

unexpected paucity of specific transcription factors, despite their complex life cycles. However, a new apicomplexan protein family of genes with apetala 2 (AP2)-integrase DNA-binding domains was found, which is predominantly found in transcription factors of plants[129]. Further discussion of (comparative) genomics of the other apicomplexan and kinetoplastid parasites is beyond the scope of this review but it is clear that detailed comparisons of the genomes of these species in the near future will help to unravel the function of many hypothetical genes of *Plasmodium* and will lead to new insights into the complex parasitic life styles.

Comparative genomics between *Plasmodium* species

Extensive synteny between genomes

Comparative genomics between *Plasmodium* species was initiated by gene-mapping studies on separated chromosomes[25,26,130], followed by more detailed analysis of (small) fragments of individual chromosomes[72,131]. In general, these studies demonstrated significant conservation of gene-linkage groups (high levels of synteny) between different species. By definition, synteny is the conservation of gene association in organized blocks. Within the blocks there can be deletions and changes in gene order but the syntenic relationship of the genes remains unaltered. However, the complete picture of the degree of synteny in *Plasmodium* remained unclear until sufficient sequence data were available.

The high degree of synteny between more distantly related *Plasmodium* species was demonstrated with the publication of extensive genome shotgun sequence of the RMP *P. yoelii*[51]. Contigs covering >70% of the *P. yoelii* genome could be aligned along the scaffolds of the 14 *P. falciparum* chromosomes (except at the subtelomeric ends). The similarity of these two *Plasmodium* genomes was not only demonstrated by the high level of synteny but also mirrored by the predicted gene content. This was the first-ever comparison of gene content of two eukaryotic species within a single genus and it identified more than 3,300 *P. yoelii* orthologues of 5,268 predicted *P. falciparum* genes. Although the orthologues were predominantly housekeeping genes, orthologues of many vaccine candidate antigens involved in parasite-host/vector interaction were also described (for example, CS, and the MSP and 6-Cys families; Box 2).

Such high levels of orthology were perhaps unsurprising, given that the majority of the features of the life cycle are conserved between different *Plasmodium* species (Box 1). However, the validation of model malaria species that is provided by their genetic similarity to human infectious species has emphasized the fact that structure-function studies on *P. falciparum* vaccine candidates could be carried out with the more accessible, tractable model species, where appropriate. Therefore, the molecular mechanisms underlying gamete fertilization[132] and the motility of sporozoites[133] could reasonably be studied in model systems. However, non-primate models might be less appropriate to investigate adaptive processes of human parasites, such as the ability of *P. falciparum* to successfully invade human erythrocytes via several independent routes. The complexity and flexibility of erythrocyte invasion by *P. falciparum* may well have evolved as part of a selection and counter-selection “arms race” that model species clearly cannot recreate.

Box 2: *Plasmodium* post-genome analyses and vaccines

Vaccination is feasible against all of the extracellular forms of the parasite: (i) the sporozoite as it enters the bloodstream and before it invades cells of the liver, (ii) the merozoite as it seeks an erythrocyte, and (iii) the gametes and the ookinete in the mosquito midgut may all be targeted by antibodies induced by prior immunization with purified stage-specific components produced by recombinant technologies. In addition, the hepatocyte presents parasite components in association with MHC class II, which may be utilized for effective cell-mediated immunity. An anti-disease immunization may seek to prevent the interaction of the adhesive ligands (including PfEMP1) with ligands at the endothelial cell surface of capillaries lining organs where parasite sequestration is known to occur and induce pathology. The ideal vaccine is currently expected to consist of multiple components that would induce activity against all of these aspects of the parasite life cycle. A brief and non-exhaustive list of leading current (largely pre-genomic) candidates is given here.

The 6-Cys family exemplifies the potential of genomics to expand the range of vaccine candidates. The family has expanded from three known members (P48/45, P230, and P12) in the pre-genomic era to its current status of ten. All members are predicted to be surface proteins, six of which are expressed in gametes, three in schizonts, and one in sporozoites. Current genome annotation (November 2005) predicts 579 proteins with SPs, illustrating the potential to uncover proteins that might come in direct contact with the host immune system.

Plasmodium vaccine candidates.

Parasite stage	Vaccine candidate gene	References
Sporozoite	CS, TRAP	[134,135]
Liver stage	LSA1	[136]
Merozoite	MSP1, AMA1, MSP2	[137-139]
Gamete	P48/45, P230	[132,140]
Ookinete	P25, P28	[141-143]

The initial comparisons of the genomes of different *Plasmodium* species have been recently extended with the publication of two additional partial shotgun genome sequence datasets (4x coverage each) from the RMPs *Plasmodium berghei* and *Plasmodium chabaudi*[52] (Chapter 4). The virtually complete synteny and high levels of sequence identity (88-92%) between the genomes of the three RMPs are such that it proved possible to compile composite RMP (cRMP) contigs, extending the contig size by 400% on average[87] (Chapter 5). The use of these contigs when combined with chromosome-mapping studies has enabled complete comparative synteny maps to be compiled for the “prototype” RMP genome and that of *P. falciparum*[52,87] (Appendix 1). These maps showed again the extensive synteny of the internal chromosome regions. Interestingly, a minimum of only 15 gross chromosomal rearrangements reshuffling the 36 synteny blocks (SBs) is needed to convert the *P. falciparum* genome into the composite RMP genome and vice versa[87] (Chapter 5). Clearly, as comparative genomics is expanded to more *Plasmodium* species, it should be possible to reconstruct the organization of the minimum genome of the most recent common ancestor (MRCA) of the genus.

The combined sequence data from the different RMPs improved the *P. falciparum* orthologue predictions and revealed a conserved set of 2,125 genes with orthologues present in the datasets of all four species. Perhaps more telling is

the fact that roughly 4,500 of the 5,268 *P. falciparum* full-length protein-encoding genes had an orthologue in at least one of the three RMP genome datasets.

Lack of synteny in the subtelomeric regions

The subtelomeric regions of *Plasmodium* chromosomes lack synteny, which stems from their generally distinct content of numerous gene families (for example, *var*, *rif*, and *stevor* genes in *P. falciparum*) that are replaced by other families in RMPs and *P. vivax*, and from the presence of large numbers of species-specific non-coding repeat sequences. However, the gene content of the subtelomeric regions of different *Plasmodium* species is not completely species-specific. The RMPs and *P. vivax* share a distinct family of subtelomeric variant genes, collectively known as the *pir* (*Plasmodium* interspersed repeats) superfamily, originally described in *P. vivax*, where they are known as *vir* genes[144]. The *pir* superfamily is predicted to be large, with ~150 to ~850 members in each species. The proteins encoded by certain members of the *pir* superfamily have been localized to the surface of erythrocytes[145], suggesting a role in antigenic variation and immune evasion, but proteome data indicate this might not be the exclusive function of the family[52] (Chapter 4).

Our knowledge of subtelomeric gene families in other species than *P. falciparum*, however, remains incomplete. Although some general conservation might be anticipated between species, the true picture of repeat family diversity, organization, and relationship to expression and function can only emerge from increased genome sequencing. What is clear though is that the subtelomeric localization of these gene families should promote recombination, in turn generating diversity and hence confusing synteny[85,86]. This tendency to diversify is exemplified by the recent analysis of members of a subtelomeric gene family present in RMPs and *P. falciparum*. These genes were first identified as two different species-specific families through BLAST analyses within the RMPs (*pyst-b*) and *P. falciparum* (*pf-fam-b*), yet could be classified as members of the same inter-species gene superfamily (renamed *pfmc-2tm*) only through shared predicted protein structure (basic proteins with two transmembrane [TM] domains), as they lacked obvious sequence similarities[146]. Shared structural features of proteins encoded by subtelomeric gene families also suggest the existence of a gene superfamily within *P. falciparum* that includes both *rif* and *stevor* genes[146]. Interestingly, this superfamily might well be extended to include the subtelomeric *pir* genes found in other *Plasmodium* species, again indicating the rapid gene evolution that is one consequence of their subtelomeric location. Appendix 4[52] (Chapter 4) provides an overview of all *P. falciparum*-specific, RMP-specific and their common gene families.

Disruption of synteny by species-specific genes

Analysis of the 168 chromosome-internal *P. falciparum*-specific genes, not present in the RMP genomes, revealed that 126 of these genes disrupt synteny within the SBs (intrasyntenic genes), while 42 are located at the SBPs between SBs (intersyntenic genes; see Figure 2 for an example of each, Chapter 5)[87]. Curiously, synteny breakpoints in the RMPs only harboured five intersyntenic genes. The majority of the *P. falciparum* intra- (62%) and intersyntenic (88%) genes encode

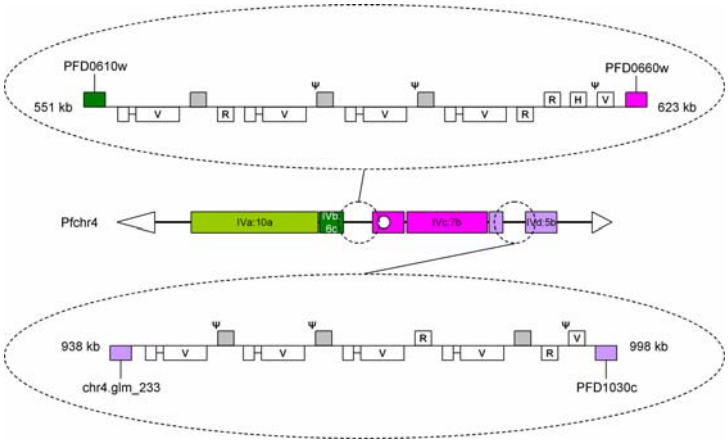


Figure 2. Inter- and intrasyntenic *var* clusters of Pfchr4

Comparison of the *P. falciparum* and cRMP genomes revealed the presence of *P. falciparum*-specific gene clusters either marking SBPs or disrupting SBs, termed inter- and intrasyntenic regions, respectively. Here two chromosome-internal *var* clusters of Pfchr4 are shown, one linking regions with synteny to cRMP chromosomes 6 (cRMPchr6; green) and 7 (pink), the other disrupting a region syntenic to cRMPchr5 (lavender). The gene annotations of the syntenic (coloured) genes flanking the regions are provided and the *P. falciparum*-specific genes (white) are designated as follows: V = *var* genes; R = *rif* genes; H = hypothetical protein. The *vicar* elements are shown in grey. All pseudogenes are marked with a Ψ.

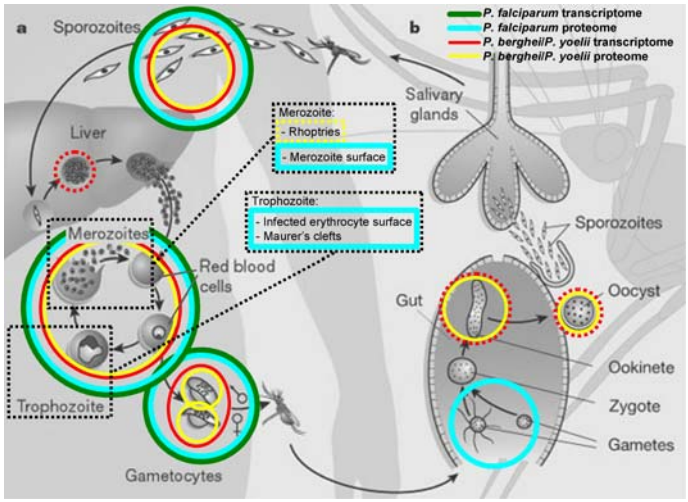


Figure 3. Transcriptomes and proteomes of different *P. falciparum* and RMP life cycle stages

Transcriptome and proteome studies have already covered a wide range of *Plasmodium* life cycle stages, although no detailed studies have been published yet for the less accessible liver stages, the short-living zygotes or oocyst/hemocoel sporozoites. The coloured circles indicate stages for which *P. falciparum* and RMP transcription and expression data are available (see the figure for colour-coding legends). Incomplete datasets or datasets of less pure stages are indicated by dotted lines. (Figure reproduced and adapted from Wirth, *Nature*, 2002[73].)

predicted exported proteins destined for the membrane surface of the merozoite or the infected erythrocyte (including *var* and *rif* genes; Figure 2), and hence are most likely involved in parasite-host interactions. The presence of species-specific genes at SBPs suggests that gross chromosomal rearrangements shape the species-specific gene content of the genomes of *Plasmodium* species. Evidence for the association of such gross chromosomal rearrangements and the generation of species-specific gene families has been found for a gene family encoding transforming growth factor β (TGF-β) receptor-like serine/threonine protein kinases (*pftstk*; Chapter 5)[87], consisting of 21 copies in *P. falciparum* (and possibly *P. reichenowi*) compared to one for all other malaria species. This gene family is the first gene family for which a single progenitor gene in other *Plasmodium* species has been identified and that appears to have expanded relatively recent in only *P. falciparum* and *P. reichenowi*, possibly as the result of a specific adaptation to the (MRCA of) human and chimpanzee hosts.

The analysis of the location of *P. falciparum*-specific genes using the synteny maps revealed that, in addition to recombination in the more frequently recombining subtelomeric regions, also chromosome-internal rearrangements may influence diversity and complexity of the *Plasmodium* genome, increasing the ability of the parasite to successfully interact with its vertebrate host. Furthermore, it indicates that determination of the SBPs may help to rapidly identify the species-specific gene content of future *Plasmodium* genomes.

Transcriptomes and proteomes of other *Plasmodium* species

Although global transcription profiles might only be correctly interpreted when a whole-genome database is available, intelligent applications of cDNA-based technologies were initiated well before the publication of homologous genome data. Several RMP and one *P. vivax*[147] EST libraries have been produced, generating tens of thousands of sequences[148-151] that can be compiled separately or in a common database allowing investigations of transcript-specific features such as splicing. In addition, several stage-specific enriched suppression subtractive hybridization (SSH) libraries for RMPs throughout development in the mosquito have been generated[108,152]. These data have not only pinpointed stage-specific transcripts, but also confirmed the conserved nature of invasive organelles associated with the invasion of host cells by different stages of *Plasmodium.* In-keeping with their morphological similarities, certain invasive organelle proteins are expressed in more than one invasive stage[148].

DNA microarray studies covering ~70% of the genes of the model parasite *P. berghei* have been performed on blood-stage parasites and generally support the “transcripts-to-go” model[92]. Transcription profiles of purified immature and mature gametocytes demonstrated that these forms share many of the cellular processes common to asexual blood-stage parasites, but enter G0 (G1 arrest)[52] (Chapter 4). The switch from asexual to sexual development involves a significant reprogramming of the transcriptional activity of the parasite[153] (~25% of the genes on the array were upregulated) carried out on the background of ongoing basic cellular processes[52] (Chapter 4).

An extensive high-throughput proteome survey has been carried out on five stages of *P. berghei*, including the first survey of *Plasmodium* ookinetes and

oocysts[52] (Chapter 4). The study uncovered numerous predicted ookinete surface proteins that can be explored for their transmission-blocking potential. In addition, this study revealed that the variable antigenic PIR proteins are not only expressed in the blood stages but are expressed as virtually non-overlapping subsets in many different life cycle stages, such as gametocytes and sporozoites. This expression pattern is more reminiscent of RIFINs in *P. falciparum*, whose expression is also not exclusive to blood stages, than of PfEMP1, suggesting that PIR and RIFIN proteins have multiple functions in their respective hosts and are not only involved in antigenic variation of blood stages. Through comparison of the proteomes of the five different stages a dichotomous strategy of protein expression was visible: the stage-specific expression of proteins that are directly involved in the interaction between the parasite and the different host cells coupled to a more constitutive expression of proteins underlying the conserved cellular machinery of the parasite in most of the different life cycle stages.

In addition to the proteomes of ookinetes and oocysts of *P. berghei*, which have not been reported yet for a human malaria parasite, the individual proteomes of the male and the female gametocytes have been analysed in *P. berghei*. The two proteomes contained 36% (236 of 650) and 19% (101 of 541) sex-specific proteins, respectively[154]. The protein content of the male gametocytes was the most distinct of all proteomes reported for other life cycle stages and shared only 69 proteins with the female gametocyte, showing the diverged features of both sexes. This proteome analysis revealed the presence of sex-specific phophatases and protein kinases that are involved in gender-specific signalling pathways. Figure 3 provides a schematic overview of the available transcriptome and proteome data sets of the different *P. falciparum* and RMP life cycle stages.

Synergy of the data and practical applications

The datasets of genome sequences from various malaria parasites and significant proteome and transcriptome surveys from at least two species provide a unique opportunity to perform comparative analyses and examine aspects of the biology of *Plasmodium* that simply would not be possible with datasets from a single species. Several studies have already been published that show how the use of different combinations of global databases of *Plasmodium* can generate novel insights into parasite-host interactions with potential therapeutic value. Comparative post-genomics of RMP genomes allowed additional detail to be teased out of the predicted protein sequences of orthologous genes. Calculation of the non-synonymous (d_N) versus synonymous (d_S) nucleotide substitutions can reveal genes encoding more rapidly evolving proteins (high d_N/d_S values) compared with more conserved proteins (for example, housekeeping)[155,156]. Not surprisingly, in RMPs proteins containing predicted signal peptide (SP) sequences and/or TM domains showed the highest d_N/d_S ratios. Analysis of the expression data generated by both transcriptome and proteome studies showed that a significantly greater number of blood-stage SP/TM proteins had high d_N/d_S ratios compared with mosquito-stage SP/TM proteins. This difference could reflect amino acid changes that have accumulated as a consequence of interactions with the host immune response and, therefore, identify genes that are under selective immune pressure.

Although most methods to detect genes under natural selection are based on the comparative analysis of sequences within and between species, Plotkin *et al.*[157] developed a single-genome-based method called codon volatility, which defines the proportion of point mutations resulting in codons encoding a different amino acid. Although the approach has been questioned[158,159], observations confirmed that genes under selective pressure, such as *var*, contain relatively more volatile codons as opposed to genes that are not under selective pressure but are instead under a strong purifying selection to maintain their protein sequence, such as housekeeping genes[157]. In part due to the extreme AT richness of the *P. falciparum* genome, this parasite has a different codon usage compared with most other organisms. Indeed, the majority of the codon triplets end with an A or U. Highly expressed proteins appear to preferentially use energetically less expensive amino acids[160], which could be another drive towards non-synonymous mutations.

An elegant method combining genome and proteome data to identify novel *P. falciparum* antigens has been described by Doolan *et al.*[161]. They used a strategy to mine genomic sequence databases using epitope predictions for the identification of novel sporozoite antigens and epitopes recognized by experimentally vaccinated humans. Such an approach could lead to the generation of an antigen map of sporozoite/liver stages ("immunosome"). Another novel method to identify antigens using genomic data has been described for *P. chabaudi*. This approach, termed linkage-group selection is based on crossing two genetically different *Plasmodium* lines followed by applying a selective pressure, in this study immune pressure, on the recombinant progeny. Subsequent analysis of the decrease in the frequency of parental alleles in the progeny after immune pressure by using quantitative genome-wide molecular markers can identify genome loci containing genes encoding proteins that were under immune selective pressure[162]. A third method to identify new antigens based on the genome and proteome data has been developed using *P. yoelii*. Exons of genes encoding sporozoite proteins were cloned in a DNA immunization vector using high-throughput methods. These vectors were then used to immunize mice that were subsequently analysed for their protection against sporozoite infection[163].

Combined analysis of transcriptome and proteome data reveals insight into regulation of transcription and protein expression. The genome of *P. falciparum* contains only a limited number of genes encoding transcription-associated proteins (only a third of the number usually found in the genomes of free-living eukaryotes[164]). However, proteins containing CCCH-type zinc finger motifs, that are often associated with modulation of mRNA decay and translation rates, are abundant[164], suggesting that post-transcriptional processes play a significant role in the regulation of *P. falciparum* protein levels. Bioinformatic analysis comparing mRNA transcript and protein abundance levels for seven different stages of *P. falciparum* indeed implied mechanisms of post-transcriptional control, either involving interplay between mRNA stability and degradation, gene-specific control of mRNA translation, or a combination of both[101]. Also the combination of transcriptome and proteome data of *P. berghei* demonstrated the presence of post-transcriptional control of gene expression in gametocytes through the mechanism of TR. TR was known to affect the expression of two gametocyte-specific transcripts that encode vaccine candidate antigens (P28 and P25) translated only

Box 3: Databases: towards the "virtual parasite"

Supplementary online data files and the general public databases such as GenBank (http://www.ncbi.nlm.nih.gov/) aside, most genome projects require and enable an associated database specifically tailored to the features of the organism. PlasmoDB (http://plasmodb.org/)[165,166] is the malaria community's resource that attempts to present the chromosomes, their genes, primary annotation, and all relevant information pertaining to the deeper post-genomic characterization of the expression and function of the gene product. The format is logic based and enables intuitive browsing through the genome, utilizing a graphical interface representing the *P. falciparum* genome. The extensive amount of available data can also be explored through dedicated, complex queries. On top of that, PlasmoDB integrates the ever-increasing amount of comparative information from other *Plasmodium* genomes described here. In addition, the genome sequencing centres also have their own curated databases, GeneDB (WTSI, http://www.genedb.org/) and Gene Indices (TIGR, http://www.tigr.org/tdb/tgi/_protist.shtml) that provide comprehensive information at the level of the individual genes, but without at present providing the more global synthesis attempted by PlasmoDB. Further integration of (post-)genomic and experimental data and enhancement of algorithms to predict not only gene structure, but also expression profiles and protein structures, could ultimately facilitate the generation of a "virtual parasite". See the next page for a list of useful links to genome, transcriptome and/or proteome databases and several other interesting malaria websites.

in the zygote just after fertilization[142,167]. Comparison of transcriptomes and proteomes of gametocytes with the *P. berghei* ookinete proteome identified nine genes undergoing TR and a sequence motif, putatively involved in TR, was subsequently identified in the 1-kb region downstream of these genes. This motif is not conserved in *P. falciparum* but shares a conserved sub-motif, nanos response element (NRE) to which RNA-binding proteins of the pumilio family (PUF) can bind that play a role in TR[168]. A similar analysis of *P. falciparum* identified two genes that contain an NRE in their 3' untranslated region (UTR), which have abundant transcripts in the gametocyte stage, while the proteins they encode are significantly more abundant in the gamete stage[101]. A detailed understanding of the specific mechanisms of transcriptional and translational control in *Plasmodium* might reveal novel therapeutic targets and strategies. For example, targeting the unlocking (derepressing) of TR in gametocytes circulating in the blood might lead to inappropriate expression of gametocyte-specific translational repressed transcripts, possibly resulting in both the inhibition of further development of gametocytes and exposure of their protein products (including current vaccine candidates) to the host immune system and thus generating transmission-blocking immune responses.

The *Plasmodium* life cycle involves intimate interactions with cells of both host and vector. A variety of cell types must be recognized and colonized by the parasite and the stage-specific gene expression of the parasite parallels this need, furnishing the parasite with appropriate cell-surface ligands and morphologies. In addition, both host and vector offer immune responses as a line of defence, which could generate a co-evolutionary "arms race". Clearly, any drive to therapeutics

Box 3: Databases: towards the "virtual parasite" - *continued*

Databases with genome data of Plasmodium

PlasmoDB, the official site of the *Plasmodium* genome project, includes many more information on expression, protein features etc.
http://plasmodb.org/

Protozoan genomes at the Wellcome Trust Sanger Institute (WTSI)
http://www.sanger.ac.uk/Projects/Protozoa/

GeneDB, the curated database of the WTSI
http://www.genedb.org/

Parasite databases at The Institute for Genomic Research (TIGR)
http://www.tigr.org/parasiteProjects.shtml

Gene Indices, the curated database of TIGR
http://www.tigr.org/tdb/tgi/protist.shtml

Malaria projects at the Stanford Genome Technology Center (SGTC)
http://sequence-www.stanford.edu/group/malaria/index.html

Databases with transcription data of Plasmodium

Resources of the Winzeler Laboratory, The Scripps Research Institute
http://www.scripps.edu/cb/winzeler/resources.htm

Microarray database of the DeRisi Laboratory, University of California San Fransisco
http://malaria.ucsf.edu/index.php

Malaria full-length cDNA database of the University of Tokyo
http://fullmal.ims.u-tokyo.ac.jp/index.html

Global networks and databases for malaria research(ers)

NCBI malaria pages with sequence data, literature, search tools etc.
http://www.ncbi.nlm.nih.gov/projects/Malaria/

WHO/TDR malaria pages with information from sequences to conferences
http://www.wehi.edu.au/MalDB-www/

Malaria Foundation International (MFI)
http://www.malaria.org/

Multilateral Initiative on Malaria (MIM)
http://www.mim.su.se/english/index.asp

Malaria Research and Reference Reagent Resource Center (MR4)
http://www.malaria.mr4.org/

More sites with information on the biology of Plasmodium

Malaria pages of the World Health Organization (WHO)
http://www.who.int/health_topics/malaria/en/

Biology, biochemistry & physiology of *Plasmodium* at the Hebrew University of Jerusalem
http://sites.huji.ac.il/malaria/index.html

Basic cellular and molecular biological processes of *Plasmodium* at Tulane University
http://www.tulane.edu/~wiser/malaria/cmb.html

Genome atlas maps of the 14 *P. falciparum* chromosomes at the Center for Biological Sequence analysis of the Technical University of Denmark
http://www.cbs.dtu.dk/services/GenomeAtlas/

3D structures of 14 *Plasmodium* proteins at the Toronto-based Structural Genomics Consortium website
http://www.sgc.utoronto.ca/

And last but not least our own website: the *P. berghei* model parasite pages of the Leiden University Medical Centre (LUMC)
http://www.lumc.nl/1040/research/malaria/malaria.html

must be informed by an understanding of these processes, and has been enhanced by the availability of full genome sequences and an increasing amount of post-genome analyses for humans[30], the major African vector, *Anopheles gambiae*[36], and the major model host, the mouse[32] (the latter two may also be genetically manipulated).

Concluding remarks

Malaria research is in a period of intense data collection, ensuring that the "labels" on each gene in the *Plasmodium* genomes and the proteins they encode are correct. Clearly, the initial phase is a "stamp-collecting" exercise but is essential as it is only through the lens of full and accurate annotation and protein characterization that we will be able to make sense of - and exploit - the genome. Although drug and vaccine discovery programmes are already (and rightly) underway as a result of the availability of the *Plasmodium* genomes, the hard choices will be at the level of inclusion or exclusion of drug-targets or vaccine-candidate antigens for further development. However, the increased knowledge of structural and functional properties of a large number of *Plasmodium* proteins as well as antigenic properties of vaccine candidates will highly benefit the decision making process.

HIV/AIDS is a devastating disease that the world has only known for 30 years and where the prospect exists to combat and control the disease and its transmission, should financial resources be made available to provide the drugs that have been developed. Conversely, malaria is an ancient disease, acknowledged for thousands of years, whose aetiological agent was first recognized over 100 years ago. Nevertheless, malaria is a steadily worsening scourge with possible and unproven new therapeutics some distance away. Whilst it has improved, significant investment is still required at all levels of investigation, development and application in order to realize the potential of the *P. falciparum* genome and translate the promise into a tangible effect.

Notes

Supporting Online Material (SOM) accompanies the paper on the Nature website (http://www.nature.com/nature/) and includes SOM Tables S1 and S2.

Chapter 3

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

Jane M. Carlton[1], Samuel V. Angiuoli[1], Bernard B. Suh[1], Taco W.A. Kooij[2], Mihaela Pertea[1], Joana C. Silva[1], Maria D. Ermolaeva[1], Jonathan E. Allen[1], Jeremy D. Selengut[1], Hean L. Koo[1], Jeremy D. Peterson[1], Mihai Pop[1], Daniel S. Kosack[1], Martin F. Shumway[1], Shelby L. Bidwell[1], Shamira J. Shallom[1], Susan E. van Aken[1], Steven B. Riedmuller[1], Tamara V. Feldblyum[1], Jennifer K. Cho[1], John Quackenbush[1], Martha Sedegah[3], Azadeh Shoaibi[1], Leda M. Cummings[1], Laurence Florens[4], John R. Yates[4], J. Dale Raine[5], Robert E. Sinden[5], Michael A. Harris[6], Deirdre A. Cunningham[7], Peter R. Preiser[7], Lawrence W. Bergman[8], Akhil B. Vaidya[8], Leo H. van Lin[2], Chris J. Janse[2], Andrew P. Waters[2], Hamilton O. Smith[6], Owen R. White[1], Steven L. Salzberg[1], J. Craig Venter[9], Claire M. Fraser[1], Stephen L. Hoffman[3], Malcolm J. Gardner[1] and Daniel J. Carucci[3]

[1]The Institute for Genomic Research (TIGR), 9712 Medical Center Drive, Rockville, MD 20850, USA. [2]Malaria Research Group, Department of Parasitology, Centre for Infectious Diseases, Leiden University Medical Centre, Albinusdreef 2, 2333 ZA, Leiden, The Netherlands. [3]Naval Medical Research Center, Malaria Program (IDD), 503 Robert Grant Avenue, Room 3A40, Silver Spring, MD 20910-7500, USA. [4]Department of Cell Biology, The Scripps Research Institute, SR-11, 10550 North Torrey Pines Road, La Jolla, CA 92037, USA. [5]Immunology and Infection Section, Department of Biological Sciences, Imperial College London, Sir Alexander Fleming Building, Imperial College Road, London SW7 2AZ, UK. [6]Celera Genomics, 45 West Gude Drive, Rockville, MD 20850, USA. [7]Division of Parasitology, National Institute for Medical Research, London, UK. [8]Division of Molecular Parasitology, Department of Microbiology & Immunology, Drexel University College of Medicine, Philadelphia, PA 191929, USA. [9]The Center for the Advancement of Genomics, 1901 Research Boulevard, Rockville, MD 20850, USA.

Introduction to Chapter 3: "Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*".

This study was the result of a multilateral effort of 44 scientists from nine different research groups from the USA, the UK and the Netherlands and presents the partial genome sequence of the rodent malaria parasite (RMP) *Plasmodium yoelii*. The genome sequence of *P. yoelii* has been published in 2002 in Nature alongside papers describing the complete genome sequence of *Plasmodium falciparum*, which was the first eukaryotic parasite for which the genome was sequenced to completion. The availability of genomes of these two species allowed, for the first time, the comparison of two eukaryotic species within a single genus. The majority of the experimental work on *P. yoelii* was performed at The Institute for Genomic Research (TIGR, MD, USA) and at the Department of Parasitology at the LUMC (Leiden, Netherlands). My contribution to the paper consisted of an in depth analysis of the synteny between *P. yoelii* with the completed *P. falciparum* genome.

One of the key interests of the malaria research group in Leiden is the sexual development of malaria parasites (gametocyte and gamete production; fertilization) and processes involved in the transmission of the parasites to the mosquito. Previous studies had shown that a specific chromosome of *Plasmodium berghei*, chromosome 5 (Pbchr5), contains various gametocyte-specific genes and genes whose transcription is upregulated in gametocytes, for example *p25* and *p28*[169], *α-tubulin II*[170] (Chapter 6), the *c-ribosomal rna* (*c-rrna*) gene unit[171] and three genes located in a complex, gene-dense region[60] (T.W.A.K. and A.P.W., unpublished data). Furthermore, during mechanical passage of a gametocyte producing line of *P. berghei* (clone 8417HP) a parasite clone was isolated with a 70-kb subtelomeric deletion of Pbchr5 that was unable to produce gametocytes (clone HPE)[172]. To further analyse Pbchr5 a long range restriction map was generated[72] and a detailed study of a 13.6-kb region containing tightly clustered genes demonstrated that this locus is highly conserved between *P. berghei* and *P. falciparum*[60].

To extend the studies on the conservation of this chromosome of *P. berghei* and *P. falciparum*, Leonard H.M. van Lin and I selected and characterized a range of sequence tagged site (STS) markers in addition to the already published genes located on Pbchr5. Selection was performed by hybridization of the markers to chromosomes of clones 8417HP and HPE that were separated by field inversion gel electrophoresis (PFGE)[26]. The location of the STS markers on Pbchr5 was demonstrated by long-range restriction mapping (Southern analysis of *Apa*I, *Bgl*I, and *Sma*I whole-genome digests) and comparison with the existing long-range restriction map available of Pbchr5[72]. All STS markers were sequenced and compared with the preliminary genome sequences of *P. falciparum* by BLASTN and TBLASTN analyses. Most of the markers appeared to be syntenic with *P. falciparum* chromosome 10 (Pfchr10), while a small proportion of the markers were shown to be syntenic with Pfchr4, demonstrating a high degree of synteny in these regions of the genome between *P. berghei* and *P. falciparum*.

This high degree of synteny caught the attention of Jane M. Carlton (TIGR), who was involved in sequencing of the *P. yoelii* genome and we agreed to extend our synteny study to *P. yoelii* using the contigs aligned along the entire length of Pfchr4 and 10. For this study, I compared the *P. yoelii* sequence data with the STS

marker map of *P. berghei* and linked the grouped *P. yoelii* contigs by polymerase chain reactions (PCR). A specialized PCR strategy for AT-rich genomes[173] using Platinum *Taq* DNA polymerase High-Fidelity (Invitrogen) was used to close large gaps between the sequenced contigs. No STS markers were available for the regions of Pfchr4 and 10 that are not syntenic to the RMP chromosome 5 (RMPchr5). The exact chromosomal location of the *P. yoelii* contigs in these regions was assigned by hybridization of the PCR probes (used to link the contigs) to chromosomes of the four RMPs that were separated by PFGE.

The results of the synteny comparison of *P. yoelii* with chromosomes 4 and 10 of *P. falciparum* were published as part of the main foldout figure (Figure 5) of the accompanying paper: "Genome sequence of the human malaria parasite *Plasmodium falciparum*"[42] and were described and illustrated in more detail in the section "A genome-wide synteny map" and Figure 6 of the following paper.

Abstract

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*, and comparative studies with the genome of the human malaria parasite *Plasmodium falciparum* clone 3D7. A synteny map of 2,212 *P. 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. 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.

Introduction

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 *P. falciparum* is particularly problematic as the complete life cycle cannot be maintained *in vitro*. Four RMPs (*P. yoelii*, *Plasmodium berghei*, *Plasmodium chabaudi*, and *Plasmodium vinckei*) isolated from wild thicket rats in Africa have been adapted to grow in laboratory rodents[174]. 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 RMPs, a prime example being stable genetic transformation[175]. 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[176]. We describe here the genome sequence of the RMP *P. 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. yoelii* and *P. falciparum* genomes, sequencing and annotation of the former can substantially improve the accuracy and efficiency of annotation of the latter.

Materials and methods

Genome and EST sequencing

P. yoelii yoelii 17XNL line[177], selected from an isolate taken from the blood of a wild-caught thicket rat in the Central African Republic[178], 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[179]. BLASTN of the *P. yoelii* contigs and singletons against the complete set of Celera mouse contigs[180], using a cutoff of 90% identity over 100 nucleotides, identified contaminating mouse sequences that were subsequently removed. Contigs were assigned to groups using Grouper[74]. Each contig was assigned an identifier in the format "MALPY00001".

Proteomic analysis

MudPIT technology and methods were as described[11]. Sporozoites of *P. yoelii* were dissected from infected *Anopheles stephensi* mosquito salivary glands, and *P. yoelii* gametocytes were prepared as described[181]. Cellular debris from uninfected mosquitoes and mouse erythrocytes were analysed as controls. Tandem mass spectrometry data sets were searched against several databases: the complete set of *P. yoelii* full and partial proteins (7,860 total); 791,324 *P. 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. yoelii* expressed sequence tags (ESTs). Phat and the exon recognition module of GlimmerMExon were trained on *P. falciparum* data as described[182]. Combiner was used to generate a final ranked list of *P. yoelii* gene models, and TIGR's Eukaryotic Genome Control suite of programmes was used for automated annotation of these as described[182]. 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 gene families were identified by a domain-based clustering algorithm developed at TIGR. Families were regarded as potentially *Plasmodium*- or *P. yoelii*-specific if they were not described by any Pfam[183] or TIGRFAM[184] 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[185]. 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[186] with a minimum exact match ranging from 30-40 bases. Tandem Repeat Finder[187] 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 Pfchr7 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[188], and nucleotide sequences of predicted gene models were aligned using CLUSTALW[189] with default parameters, and refined manually. The number of substitutions per synonymous (d_S) and non-synonymous (d_N) sites were estimated using the Nei and Gojobori method[190]. Conservation of gene order was established using Position Effect (http://www.tigr.org/software/), where matches between *P. falciparum* and *P. 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. yoelii*.

P. 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[191], to achieve fivefold sequence coverage of the genome of a clone of the 17XNL line of *P. yoelii* (Table 1). This level of coverage is expected to comprise 99% of the genome[192] assuming random library representation. As with *P. falciparum*, the genomes of RMPs are highly AT-rich[193], which adversely affects DNA stability in plasmid libraries. Consequently, all 220,000 reads were produced from clones originating from small (2-3 kb) insert libraries. Contigs were assembled using TIGR

Table 1: *P. yoelii* genome coverage statistics.

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

Abbreviations: EST, expressed sequence tag.

Assembler[179]. 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[179]. 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"[180], and emphasizes the use of such linking libraries in partial genome projects. The genome size of *P. yoelii* is estimated to be 23 Mb, in agreement with karyotype data[26].

Expression data from the *P. yoelii* transcriptome and proteome were generated to aid in gene identification and annotation of the contigs (Table 1). A total of 13,080 EST sequences generated from clones of an asexual blood-stage *P. yoelii* complementary DNA library[194], in combination with other *P. yoelii* ESTs and transcript sequences available from public databases, were assembled and used to compile a gene index[195] of expressed *P. 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. 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 programmes, GlimmerMExon and Phat[196], to predict coding regions in *P. yoelii*. GlimmerMExon is based on the eukaryotic gene finder GlimmerM[197], 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 programmes. 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.

Comparative genome analysis

A comparison of several genome features of *P. falciparum* and *P. yoelii* is shown in Table 2 and Appendix 3, demonstrating that many similarities exist between the genomes. Besides the similarly extreme GC compositions, both genomes contain a comparable number of predicted full-length genes, with the higher figure in *P. yoelii* due to an extremely high copy number of variant antigen genes. Where differences between the genomes do exist, such as the GC content of the coding portion of the genomes, incompleteness of the *P. 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. yoelii* proteomic data identified 83 regions of the genome apparently expressed during sporozoite and/or

Table 2: Genome summary statistics. A more detailed set of statistics is given in Appendix 3.

	P. yoelii	***P. falciparum***
Size (bp)	23,125,449	22,853,764
No. contigs	5,687	93
Av. contig size (bp)	4,066	213,586
Sequence coverage[a]	5x	14.5x
No. protein coding genes	5,878	5,268

[a] Average number of sequence reads per nucleotide.

gametocyte stages but not assigned to a *P. yoelii* gene model (unpublished data). 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. 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[197] with a cutoff E-value of 10^{-15} and no low-complexity filtering, 3,310 bidirectional orthologues (defined as genes related to each other through vertical evolutionary descent) can be identified in the full protein complement of *P. falciparum* (5,268 proteins) and the protein complement of *P. 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. 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 *6-Cys* superfamily identified previously[88] were confirmed.

We identified several *P. yoelii* orthologues of *P. falciparum* biochemical pathway components under study as targets for drug design (Table 3), most notably: (i) 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*[9]; (ii) enoyl-acyl carrier protein reductase (FABI) whose product is inhibited by triclosan in *P. falciparum in vitro* cultures and mice infected with *P. berghei*[10]; and (iii) a gene encoding farnesyl transferase (FTASE), which is inhibited in cultures of *P. falciparum* treated with custom-designed peptidomimetics[198]. The rodent models of malaria have proved 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 gene classification system[199], 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. yoelii* (Figure 1). A similar proportion of proteins were identified for most of the gene ontology classes between the two species, with the caveat that fewer total numbers of proteins were identified in *P. 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. yoelii*. These classes contain members

Table 3: *P. yoelii* (Py) orthologues of *P. falciparum* (Pf) candidate vaccine & drug interaction genes[a].

***P. falciparum* gene**	**Pf chr**	**ST[b]**	**Pf locus**	**Py locus**
Candidate vaccine antigens				
Ring-infected erythrocytic surface antigen 1, *resa1*	1	Y	PFA0110w	NI
Merozoite surface protein 4, *msp4*	2	N	PFB0310c	PY07543[c]
Merozoite surface protein 5, *msp5*	2	N	PFB0305c	PY07543[c]
Liver-stage antigen 3, *lsa3*	2	N	PFB0915w	NI
Merozoite surface protein 2, *msp2*	2	N	PFB0300c	NI
Transmission-blocking target antigen 230, *pfs230*	2	N	PFB0405w	PY03856
Circumsporozoite protein, *cs*	3	N	MAL3P2.11	PY03168
Rhoptry-associated protein 2, *rap2*	5	Y	PFE0080c	PY03918
Sporozoite surface antigen, *starp*	7	Y	PF07_0006	NI
Merozoite surface protein 1, *msp1*	9	N	PFI475w	PY05748
Liver-stage antigen 1, *lsa1*	10	N	PF10_0356	NI
Merozoite surface protein 3, *msp3*	10	N	PF10_0345	NI
Glutamate-rich protein, *glurp*	10	N	PF10_0344	NI
Ookinete surface protein 25, *pfs25*	10	N	PF10_0303	PY00523
Ookinete surface protein 28, *pfs28*	10	N	PF10_0302	PY00522
Erythrocyte membrane-associated 332 antigen, *pf332*	11	N	PF11_0507	PY06496
Apical membrane antigen 1, *ama1*	11	N	PF11_0344	PY01581
Exported protein 1, *exp1*	11	N	PF11_0224	NI
Surface sporozoite protein 2, *ssp2*	13	N	PF13_0201	PY03052
Sexual-stage-specific surface antigen 48/45, *pfs48/45*	13	N	PF13_0247	PY04207
Rhoptry-associated protein 1, *rap1*	14	Y	PF14_0637	PY00622
Candidate drug interaction genes				
Dihydrofolate reductase, *dhfr*	4	N	PFD0830w	PY04370
Multidrug resistance protein 1, *pfmdr1*	5	N	PFE1150w	PY00245
Translationally controlled tumour protein, *tctp*	5	N	PFE0545c	PY04896
Farnesyl transferase, *ftase*	5	N	PFE0970w	PY06214
Enoyl-acyl carrier reductase, *fabi*	6	N	MAL6P1.275	PY03846
Dihydro-protate dehydrogenase, *dhod*	6	N	MAL6P1.36	PY02580
Chloroquine-resistance transporter, *pfcrt*	7	N	MAL7P1.27	PY05061
Dihydropteroate synthase, *dhps*	8	N	PF08_0095	PY02226
Lactate dehydrogenase, *ldh*	13	N	PF13_0141	PY03885
DOXP reductoisomerase, *doxpr*	14	N	PF14_0641	PY05578

[a] A full listing of all orthologues can be found as Table A in the Supplementary Information on the Nature website (http://www.nature.com/nature/).

[b] Subtelomeric (ST) location is defined as >75% of the distance from the centre to the end of the *P. falciparum* chromosome.

[c] Homologue of *P. falciparum msp4* and *msp5* genes found as a single gene *msp4/5* in *P. yoelii* and other RMPs[200].

Abbreviations: chr, chromosome; ST, subtelomeric; NI, not identified.

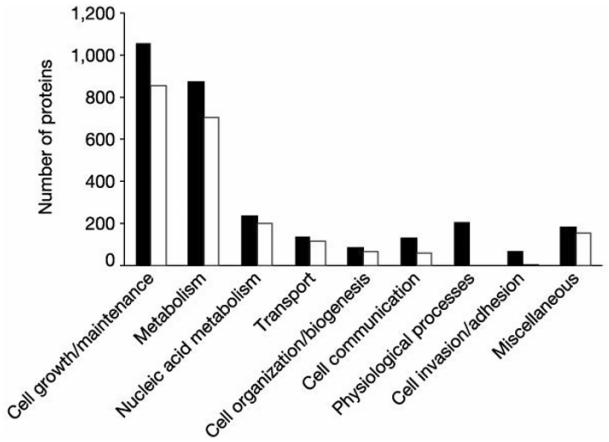


Figure 1. Functional classification comparison between *P. falciparum* and *P. yoelii* proteins

We compared the gene ontology 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. yoelii* (open bars). Ten gene ontology classes with similar numbers of *P. falciparum* and *P. 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.

of three gene families whose genes are found predominantly in the subtelomeric regions of *P. falciparum* chromosomes: PfEMP1, the protein product of the *var* 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 (Ref. [201] for review). 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. yoelii* genome.

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

The largest family of genes identified in the *P. yoelii* genome is the *yir* family, homologues of the *vir* family recently described in the human malaria parasite *Plasmodium vivax*[144] and in other RMPs[202]. 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. yoelii* genome data, 838 *yir* genes (693 full genes and 145 partial genes) are present (Figures 2 and 3, Table 4). Almost 75% of the annotated contigs identified as containing subtelomeric sequences 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[11]. Preliminary results using antibodies developed against the conserved regions of the protein have confirmed protein localization at the surface of the infected red blood cell[203]. The number of gene copies in the

Table 4: Paralogous gene families in *P. yoelii* (Py).

Gene family	No.	Name	HMM ID	ST Py	Py exp.[a]	*P. falciparum* locus	SP/ TM[b]
yir/bir/cir	838	Variant antigen family	TIGR01590	Y	S,B,G	None	A/P
235 kDa	14	Reticulocyte binding family	TIGR01612	Y	S,B,G	PFD0110w MAL13P1.176 PF13_0198 PFL2520w PFD0110w	A/P
pyst-a	168	Hypothetical	TIGR01599	Y	S,G	PF14_0604	A/A
pyst-b	57	Hypothetical	TIGR01597	Y	B	None	A/P
pyst-c	21	Hypothetical	TIGR01601 TIGR01604	Y	B	None	P/P
pyst-d	17	Hypothetical	TIGR01605	Y	G	None	P/P
etramp	11	Early transcribed membrane protein family	TIGR01495	Y	S,B,G	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	Y	S,G	PFL2530w PF10_0379 PF14_0738 PF14_0017 PF14_0737 PFI800w PFI1775w PF07_0040 PF07_0005 PFA0120c	A/A
rhoph1/clag	2	Rhoptry H1/ cytoadherence-linked asexual gene family	PF03805	Y	B,G	PFC0110w PFC0120w PFI1730w PFI1710w PFB0935w	P/A

[a] Genes were found to be expressed in the listed life cycle stages but expression may not be limited to these stages.
[b] Signal peptide (SP) and transmembrane (TM) domain predictions were identical for *P. falciparum* and *P. yoelii* members of the same gene family (Ref. [42] for details regarding SP and TM prediction algorithms).

Abbreviations: HMM ID, hidden Markov model identifier; ST Py, subtelomeric location in *P. yoelii*; exp., expression; SP, signal peptide; TM, transmembrane domain; S, sporozoite; B, asexual blood stage; G, gametocyte; A, absent; P, predicted.

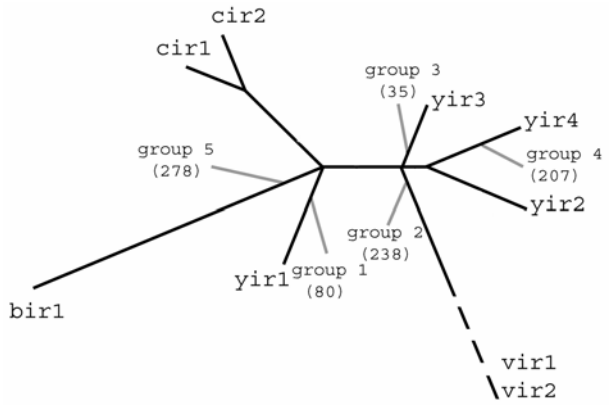


Figure 2. Phylogenetic tree of the PIR superfamily

Phylogenetic tree containing previously identified *P. yoelii* proteins YIR1, YIR2, YIR3, YIR4 (GenBank accession numbers AJ320478-AJ320481), *P. chabaudi* proteins CIR1, CIR2 (AJ315472, AJ315473), *P.berghei* proteins BIR1 (AJ320482), *P.vivax* proteins VIR1, VIR2 (AL360354) built using a maximum-likelihood method[204], from the conserved 3' terminal region of the protein. Based on this analysis, 240 genes were assigned to one of five groups. Proteins in group 2 appear close to the VIR branch but are in fact more similar to other RMP homologues since the VIR branch is 10-fold longer than the next longest branch. Numbers in parentheses refer to total number of genes in each group, and include the remaining *P. yoelii yir* homologues which were assigned to groups based on their closest homologue (minimum BlastN E-value) among the 240 resolved genes.

P. 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* family can be identified among the *P. 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[205]. Closely related homologues of the *py235* family have been found in other RMPs, and more distantly related homologues have been found in *P. vivax*[206] and *P. falciparum*[207]. The gene copy number identified in the current data set is less than has been predicted in other *P. 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. yoelii* genome (Figure 4).

Four further paralogous gene families, *pyst-a* to *-d*, are specific to *P. yoelii* (Table 4). The *pyst-a* family deserves mention, as it is homologous to a *P. chabaudi* GLURP[208] and to a single hypothetical gene on Pfchr14, suggesting expansion of this family in the RMPs from a common ancestral *Plasmodium* gene. Two paralogous gene families containing multiple members are homologous to

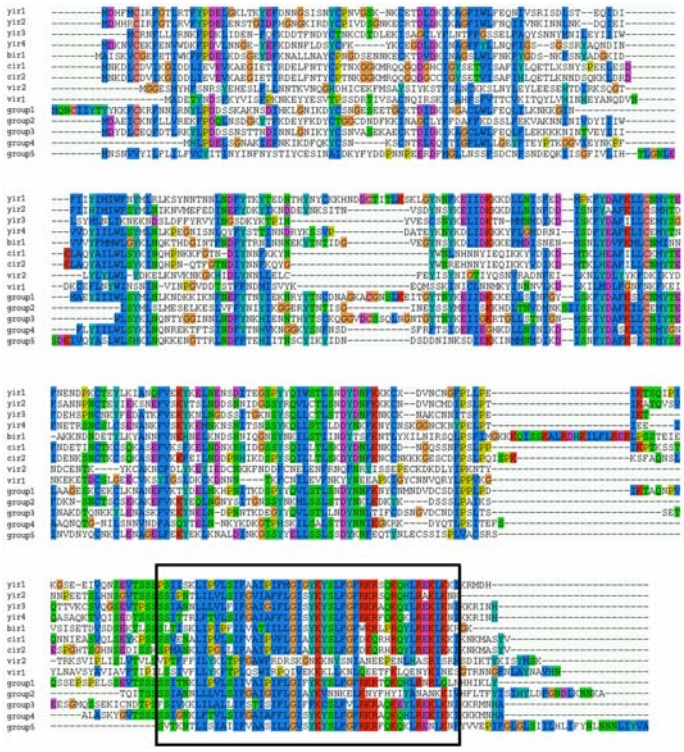


Figure 3. CLUSTALW-generated multiple sequence alignment[189] of genes from Figure 2

One gene was chosen as representative of each of the five groups. The boxed area delineates the conserved 3' region used for phylogenetic tree construction.

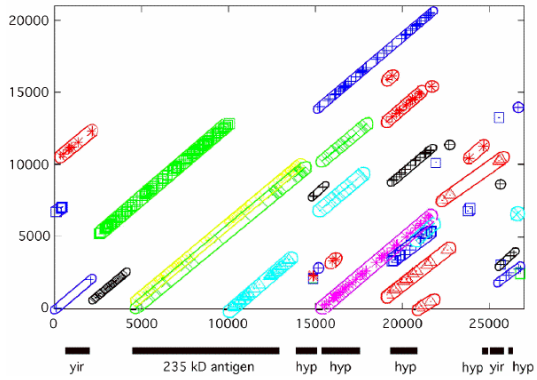


Figure 4. Alignment of subtelomeric *P. yoelii* contigs

Alignment of subtelomeric *P. yoelii* contig MALPY00313 to eleven other contigs, indicated by different colours and shapes. Each point in the plot corresponds to a sequence of 25 bp or longer that is identical in both contigs, and each diagonal indicates a longer region of shared homology. Schematics of the seven predicted genes on this contig are indicated below the x-axis. Note the juxtaposition of a *yir* gene with a member of the *py235* family.

gene families identified in *P. falciparum*. Gene members of one family, *etramp* (early transcribed membrane protein), have previously been identified in *P. falciparum*[209] and in *P. chabaudi* where a single member has been identified and localized to the parasitophorous vacuole membrane[210].

Telomeres and chromosomal exchange in subtelomeric regions

The telomeric repeat in *P. 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. 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 bp (45 copies) and a 31 bp (up to ten 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. yoelii* contigs show extensive large-scale similarity, indicating that these regions undergo chromosomal exchange similar to that reported for *P. falciparum*[42]. The longest subtelomeric contig is approximately 27 kb (Figure 4) 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 (Ref. [24] for review). The suggestion of extensive chromosome exchange in *P. 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 (My) ago[211], and species of the parasite are known to infect birds, mammals and reptiles[212]. On the basis of the analysis of *small subunit* (*ssu*)-*rrna* sequences, the closest relative to *P. falciparum* is *Plasmodium reichenowi*, a parasite of chimpanzees, with the RMPs forming a distinct clade[125,126]. Early gene-mapping studies have shown that regions of gene synteny exist between RMPs[26] and between human malaria species[130,131], despite extensive chromosome size polymorphisms between homologous chromosomes[23]. This level of gene synteny seems to decrease as the phylogenetic distance between *Plasmodium* species increases[25]. 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. yoelii* genome data, we have constructed a genome-wide syntenic map between the species. To avoid confounding factors inherent in DNA-based analyses of AT-rich genomes, we first calculated the protein similarity between all possible protein-coding regions in both data sets using

MUMmer[186]. 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[180]. 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. yoelii* genome (Table 5). The per cent of each *P. falciparum* chromosome covered with *P. yoelii* matches varies from 12% (*P. falciparum* chromosome 4 [Pfchr4]) to 22% (Pfchr1 and 14), with an average of about 18%. The spatial arrangement of the tiling paths (Figure 5) confirms previous suggestions[26] 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. yoelii* genome.

Although the tiling paths indicate the degree of conservation of gene order between *P. falciparum* and *P. yoelii*, longer stretches of contiguous *P. yoelii* sequence are necessary to examine this feature in depth. Accordingly, we carried out linkage of many *P. 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. 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. yoelii* sequences. Third, amplification of the sequence between adjacent contigs in the tiling paths linked a further 817 assemblies. Linkage of *P. yoelii* sequences by these methods resulted in the

Table 5: *P. falciparum* (Pf) and *P. yoelii* (Py) synteny map statistics.

Pfchr (length kb)	Cum. MUMs (kb)/ Cum. Py contigs (kb)	% Pfchr covered	No. Pf genes with MUMs/Total (%)	No. cons. interg. reg.[a]	No. Py syntenic groups (No. cont.)	Pychr[b]
1 (643)	141/668	22%	59/143 (41%)	4	4 (53)	2
2 (947)	124/506	13%	95/223 (43%)	4	8 (81)	NI
3 (1,060)	172/682	16%	140/239 (59%)	1	16 (99)	4,8
4 (1,204)	141/668	12%	107/237 (45%)	3	4 (67)	5,6,7,10
5 (1,343)	264/1,030	20%	198/312 (63%)	12	23 (156)	11,12
6 (1,378)	268/985	19%	189/312 (61%)	4	25 (142)	11
7 (1,350)	230/996	17%	161/277 (58%)	9	35 (118)	8
8 (1,323)	252/981	19%	194/295 (66%)	21	20 (149)	13
9 (1,542)	268/966	17%	216/365 (59%)	2	23 (145)	4,8
10 (1,694)	280/1,135	17%	216/403 (54%)	4	2 (150)	5,12
11 (2,035)	387/1,523	19%	303/492 (62%)	11	79 (232)	9
12 (2,271)	464/1,690	20%	344/526 (65%)	3	73 (254)	NI
13 (2,747)	582/2,057	21%	448/672 (67%)	8	51 (299)	11,13,14
14 (3,291)	722/2,474	22%	503/769 (65%)	10	94 (344)	10,13,14
Total = 14	4,295/16,361	18%	3,137/5,268 (60%)	95	457 (2,289[c])	12

[a] Number of conserved intergenic regions; may represent additional exons missed during annotation, or conserved non-coding/regulatory elements that require further investigation.
[b] Limited chromosome localization data for each syntenic group available.
[c] Some *P. yoelii* sequences appear more than once in the chromosome tiling paths; actual total no. contigs in tiling path is 2,212.

Abbreviations: chr, chromosome; MUM, maximal unique match; cont., contig; NI, not identified.

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. yoelii* chromosome where possible through the use of a partial physical map[26]. Thus, long contiguous sections of the *P. yoelii* genome with accompanying *P. yoelii* chromosomal location can be assigned to each *P. falciparum* chromosome (Figure 5). 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. 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 synteny map between *P. falciparum* and *P. yoelii* - associated with Pfchr4 and 10 and *P. yoelii* chromosome 5 (Pychr5) - provides a detailed snapshot of synteny between the species. Pychr5 has received particular attention owing to the localization of a number of sexual-stage-specific genes to it[72], and because truncated versions of the chromosome are found in lines of the RMP *P. berghei*, which is defective in gametocytogenesis[172]. Genomic resources available for Pbchr5 include chromosome markers and long-range restriction maps[72]. Exploiting the high level of synteny of RMP chromosomes[26], these tools were applied in combination with further mapping studies to close the syntenic map of Pychr5 (Figure 6).

Approximately 0.8Mb of Pychr5 (estimated total length of 1.5 Mb) could be linked into one group that is syntenic to Pfchr10 and 4. From a total of 243 genes predicted in the syntenic region of Pfchr10, and 34 genes predicted in the syntenic region of Pfchr4, 171 (70%) and 22 (65%) of these, respectively, have homologues along Pychr5 that appear in the same order. Pairs of homologous genes that map to regions of conserved synteny between *P. yoelii* and *P. falciparum* are probably orthologues, which is confirmed by the finding that most of these homologous pairs are also reciprocal best matches between the *P. falciparum* and *P. yoelii* proteins. Genes in the synteny gap on Pfchr10 (Figure 6) include S antigen, GLURP, MSP3, MSP6 and LSA1, several of which are prime vaccine antigen candidates in *P. falciparum*. Genes in the synteny gap on Pfchr4 include four *var* and two *rif* genes, which make up one of the four chromosome-internal *var*/*rif* clusters found in *P. falciparum*[42]. A series of uncharacterized hypothetical genes occur on the contigs that overlap these regions in *P. yoelii*.

An intriguing finding from the study of RMPchr5 has been the analysis of the synteny breakpoint (SBP) between Pfchr4 and 10. The final *P. yoelii* contig in the tiling path with significant synteny to Pfchr10 also contains the *external transcribed spacer* (*ets*) of the *c-ssu-rrna* gene unit. The synteny resumes on Pfchr4 in a *P. yoelii* contig that also contains the *ets* of the *large subunit* (*lsu*) of the same *rrna* gene unit. (No *rrna* gene unit sequences are located on Pfchr4 and 10; matches to contigs containing these genes occur in coding regions of other genes.) Both *P. yoelii* contigs are linked to each other through a third contig that contains the remaining elements (*ssu*, *5.8s*, *lsu*, and *internal transcribed spacers its1* and *its2*) of the complete *rrna* gene unit (Figure 6). Thus it seems that the break in synteny between *Plasmodium* chromosomes has occurred within a single *rrna* gene unit, a phenomenon first reported in prokaryotes[213]. Six *rrna* gene units reside as individual operons on Pfchr1, 5, 7, 8, 11 and 13 respectively[42], in contrast to RMPs that have four[215]. Intriguingly, breaks in the synteny between *P. yoelii* and

Figure 5. Schematic representation of the *P. falciparum* 3D7 genome

Protein-encoding genes are indicated by open diamonds. All genes are depicted at the same scale regardless of their size or structure. The labels indicate the name for each gene. The rows of coloured rectangles represent, from top to bottom for each chromosome, the high-level gene ontology assignment for each gene in the "biological process", "molecular function", and "cellular component" ontologies[214]; the life cycle stage(s) at which each predicted gene product has been detected by proteomics techniques[11,12]; and *P. yoelii* genes that exhibit conserved sequence and organization with genes in *P. falciparum*, as shown by a position effect analysis. Rectangles surrounding clusters of *P. yoelii* genes indicate genes shown to be linked in the *P. yoelii* genome[51] (this chapter). Boxes containing coloured arrowheads at the ends of each chromosome indicate subtelomeric blocks.

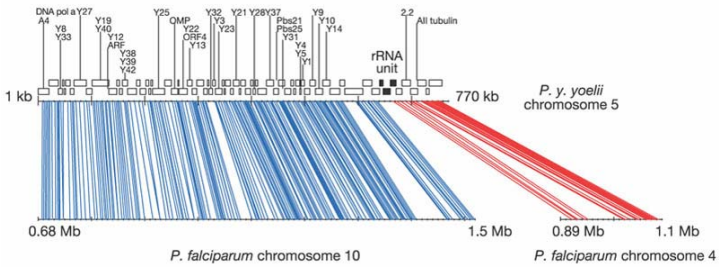


Figure 6. Conservation of gene synteny between Pychr5 and Pfchr4 and 10

Physical marker data used to confirm contig order in the tiling path of Pychr5 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. yoelii rrna* gene unit are shown as filled boxes.

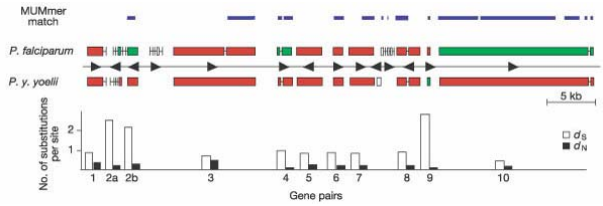


Figure 7. Global alignment scheme of a syntenic region between *P. falciparum* and *P. yoelii* encompassing ten orthologous gene pairs and nine intergenic regions

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.

P. falciparum can be mapped to almost all *rrna* gene unit loci on the *P. falciparum* chromosomes (Figure 5). 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* gene 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[60], and that such comparisons are useful for gene prediction and evolutionary studies. Accordingly, we used a comparison of the longest assembly of *P. yoelii* (MALPY00395, 51.3 kb) and its syntenic region in *P. falciparum* (Pfchr7, at coordinates 1,131-1,183 kb) as a case study for a preliminary evolutionary analysis of the two genomes. Gene prediction programmes run against these two regions identified 11 genes in the syntenic region of both species (Figure 7), 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. 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 (Figure 7) reveals the similarity between species in intergenic regions to be almost negligible, as mirrored in similar syntenic comparisons of mouse and human[216,217]. 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[218]. 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.

RMPs as models for *P. falciparum* biology

The usefulness of RMPs 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 (Ref. [219] for review). 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. 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. 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. 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*.

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. 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; 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.

Notes

Supporting Online Material (SOM) accompanies the paper on the Nature website (http://www.nature.com/nature/) and includes SOM Tables A-C and SOM Figures S1-S3. The sequences have been deposited with DDBJ/EMBL/GenBank under the accession prefix AABL. The version described in this paper is the first version, AABL01000000. All datasets are available through the official website of the *Plasmodium* genome project, PlasmoDB (http://plasmodb.org/)[165,166] and through the TIGR Eukaryotic Projects website (http://www.tigr.org/).

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.

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.

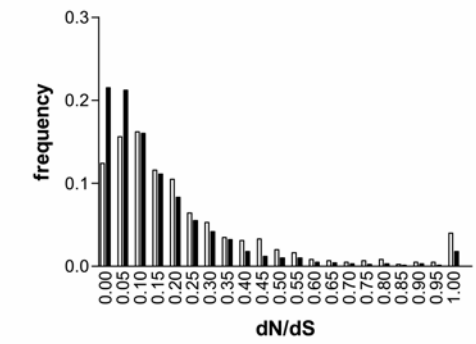
Abbreviations: Pb, *P. berghei*; Pc, *P. chabaudi*; Py, *P. yoelii*; Pf, *P. falciparum*.

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).

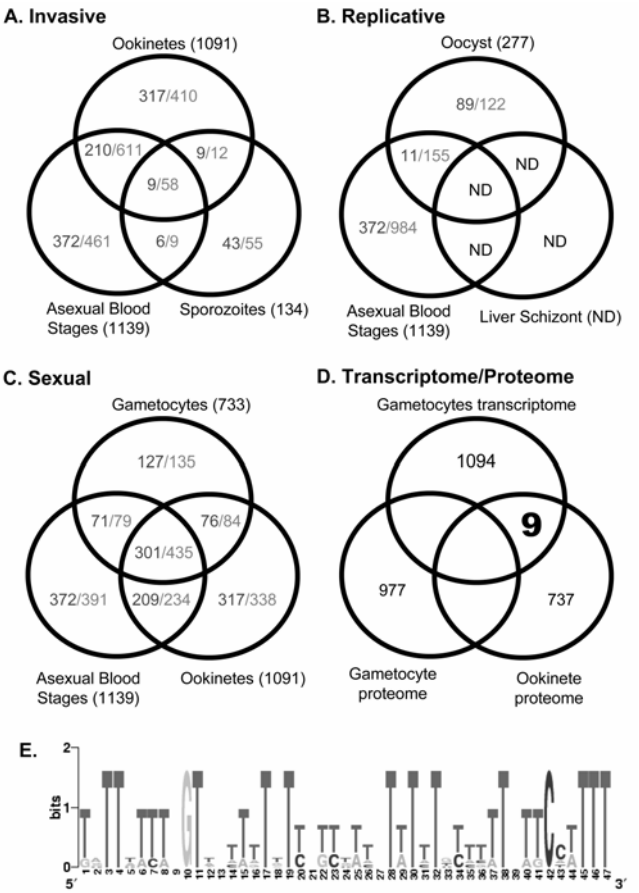


Figure 2. Different strategies of protein and gene expression during the malaria life cycle

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/).

Chapter 5

A *Plasmodium* whole-genome synteny map: indels and synteny breakpoints as foci for species-specific genes

Taco W.A. Kooij[1], Jane M. Carlton[2], Shelby L. Bidwell[2], Neil Hall[2,3], Jai Ramesar[1], Chris J. Janse[1] and Andrew P. Waters[1]

[1]Malaria Research Group, Department of Parasitology, Centre for Infectious Diseases, Leiden University Medical Centre, Albinusdreef 2, 2333 ZA, Leiden, The Netherlands. [2]The Institute for Genomic Research (TIGR), 9712 Medical Center Drive, Rockville, MD 20850, USA. [3]Pathogen Sequencing Unit, The Wellcome Trust Sanger Institute, Genome Campus, Hinxton, Cambridge CB10 1SA, UK.

Abstract

Whole-genome comparisons are highly informative regarding genome evolution and can reveal the conservation of genome organization and gene content, gene regulatory elements, and presence of species-specific genes. Initial comparative genome analyses of the human malaria parasite *Plasmodium falciparum* and rodent malaria parasites (RMPs) revealed a core set of 4,500 *Plasmodium* orthologues located in the highly syntenic central regions of the chromosomes that sharply defined the boundaries of the variable subtelomeric regions. We used composite RMP (cRMP) contigs, based on partial DNA sequences of three RMPs, to generate a whole-genome synteny map of *P. falciparum* and the RMPs. The core regions of the 14 chromosomes of *P. falciparum* and the RMPs are organized in 36 synteny blocks (SBs), representing groups of genes that have been stably inherited since these malaria species diverged, but whose relative organization has altered as a result of a predicted minimum of 15 recombination events. Species-specific genes and gene families are found in the variable subtelomeric regions (575 genes), at synteny breakpoints (SBPs; 42 genes) and as intrasyntenic indels (126 genes). Of the 168 non-subtelomeric genes, including two newly discovered gene families, 68% are predicted to be exported to the surface of the blood-stage parasite or infected erythrocyte. Chromosomal rearrangements are implicated in the generation and dispersal of *P. falciparum*-specific gene families, including one encoding receptor-associated protein kinases. The data show that both SBPs and intrasyntenic indels can be foci for species-specific genes with a predicted role in host-parasite interactions and suggest that, besides rearrangements in the subtelomeric regions, chromosomal rearrangements may also be involved in the generation of species-specific gene families. A majority of these genes are expressed in blood stages suggesting, that the vertebrate host exerts a greater selective pressure than the mosquito vector, resulting in the acquisition of diversity.

Introduction

Comparative genomics enables inferences to be drawn concerning the coding potential of related genomes and the evolutionary forces that have influenced genome organization[70]. The resolving power of whole-genome comparisons to a large extent depends upon the proximity of the phylogenetic relationship between the species. Comparative eukaryotic genome studies of several species from a wide range of lineages and different times of divergence have revealed that the level of both the conservation of organization and the recombination rates are relatively variable. Human and mouse, which diverged ~75 million years (My) ago, have a predicted gene content that is 80% orthologous[32] arranged in 281 SBs larger than 1 Mb[67]. Three-way alignment of the human genome with that of mouse and rat confirmed the conservation of ~280 SBs between human and each of the rodent genomes, while the more closely related rat and mouse genomes are ~90% orthologous with a reduced number of 105 shared SBs of larger average size[31]. Subsequent publication of the chicken genome, which diverged from the mammalian genomes ~310 My ago, provided the first non-mammalian amniote genome sequence and allowed a four-way whole-genome comparison[246] revealing 586 smaller, conserved SBs. Here, roughly 50% of the human genes have a chicken orthologue reducing to 35% that have orthologues in both chicken and

pufferfish (estimated time of divergence ~450 My). These data show that, in terms of the extent of organization and gene homology, the level of genomic conservation can generally be considered to be relatively proportional to the time of divergence, within these species. However, a more recent comparison of genome sequences from eight mammals demonstrated that the rates of chromosomal rearrangements can vary both between species and in time (about 0.2-2 breaks/My)[247].

In contrast with the relatively slow evolution of mammalian and chicken chromosome structure, gene order and linkage in Diptera species has altered at a much higher rate. Although 50% of the genes are orthologues, little conservation of synteny could be observed in comparisons of the genomes of the fruit fly with two different malaria mosquitoes, which diverged ~250 My ago[54,248]. Even in the more closely related Diptera[248,249], extensive reshuffling and inversion have altered the gene order and organization, although genes were found to be located on the same chromosome arms. Similarly, the genomes of the nematodes *C. elegans* and *C. briggsae*, which diverged ~100 My ago, share 60% gene orthology but are arranged as 4,837 microsyntenic clusters[38].

The continuing efforts to sequence a variety of unicellular parasites has resulted in the publication of a comparison of the genome sequences of three human protozoan pathogens, *Trypanosoma brucei*, *Trypanosoma cruzi* and *Leishmania major*[250], and two apicomplexan parasites infecting cattle, *Theileria annulata* and *Theileria parva*[46]. The two *Theileria* species are very closely related, with 81% (*T. annulata*) and 86% (*T. parva*) orthologous genes and no interchromosomal rearrangements[46], comparable to the well-conserved genomes of four yeast species that diverged only 5-20 My ago and show relatively few (1-5) translocations[56]. The trypanosomatid species *T. brucei* and *L. major* share 68% and 75% gene orthology, respectively, organized in 110 SBs, despite having diverged as long as 200-500 My ago (chromosomal recombination rate of ~0.2-0.5 breaks/My)[250]. In conclusion, these comparative genome studies indicate that effective recombination rates and levels of gene orthology can vary greatly between species but are relatively low in protozoa.

In both pathogenic bacteria and certain unicellular eukaryotes (for example, the trypanosomatids listed above), including members of the genus *Plasmodium* that are the etiological agents of malaria, the organization and gene content of the subtelomeric regions of chromosomes are highly variable and typically contain large gene families encoding proteins that may be involved in host-pathogen interactions and antigenic variation[251]. The subtelomeric regions of *P. falciparum*, for example, harbour a repertoire of unique gene families, including 59 *var*[80-82], 149 *rif*, and 28 *stevor*[83,84]. The *var* family encodes the erythrocyte membrane protein 1 (PfEMP1), which is a variant antigen expressed at the erythrocyte surface. PfEMP1 is involved in the binding of parasite-infected erythrocytes to receptors of host endothelial cells, erythrocytes, lymphocytes, and blood platelets[251], is subject to antigenic variation, and is thought to play a role in virulence. Other *Plasmodium* species lack the *P. falciparum*-specific *var, rif*, and *stevor* families but the subtelomeric regions of their chromosomes also harbour (species-specific) gene families. For example, the human parasite *Plasmodium vivax*; *Plasmodium knowlesi*, which infects primates; and three RMPs (*Plasmodium berghei*, *Plasmodium chabaudi*, and *Plasmodium yoelii*) share the *pir*

superfamily[52,145]. Proteins encoded by the *pir* superfamily are also found on the surface of infected erythrocytes and may be implicated in antigenic variation[145]. It is generally believed that the subtelomeric location of gene families confers an enhanced capacity for gene diversification and amplification through mechanisms of ectopic recombination that may be between different chromosomes[252]. Such recombination may be facilitated through the clustering of telomeres at the nuclear periphery[62].

Genome sequence data for *Plasmodium* species are extensive and include a complete genome sequence for the major human pathogen *P. falciparum*[42] and 5x coverage of the genome of a RMP, *P. yoelii*[51]. The *P. yoelii* contigs, when aligned with the 14 *P. falciparum* chromosomes, demonstrated extensive similarity over the relatively short length of these contigs. However, similarity was evident only in the core regions of the chromosomes mainly containing conserved genes (4,500) that are present in all characterized *Plasmodium* species[52] and which are bounded by the variable subtelomeric regions that contain the different gene families. In addition to the genome sequence of *P. yoelii*, partial genome sequence and analysis have been published for two other RMPs, *P. berghei* and *P. chabaudi*, whose core genome sequence and organization are so similar[25,26,69] that it has proved possible to merge the sequenced DNA contigs of the three RMPs to form cRMP contigs that cover 90% of the core RMP genomes[51,52]. In this study, the cRMP contigs and 138 sequence tagged site (STS) markers (SOM Table S1) have been used to produce a whole-genome synteny map for the three RMPs that, when compared with the *P. falciparum* genome, identified 36 SBs describing the core genome. This synteny map shows that species-specific genes - including rapidly evolving *P. falciparum* gene families - are found not only in the subtelomeric regions but also at SBPs and as intrasyntenic indels. Our data suggest that chromosomal rearrangements in the core regions might be involved in the generation and subsequent dispersal of one such *P. falciparum*-specific gene family. These results show that not only recombination in the more frequently recombining subtelomeric regions but also chromosome-internal rearrangements may influence diversity and complexity of the *Plasmodium* genome, increasing the ability of the parasite to successfully interact with its vertebrate host.

Materials and methods

Creation of a cRMP genome

7,215 contigs of three RMP genomes, *P. yoelii yoelii* (17XNL line)[51], *P. berghei* (ANKA strain), and *P. chabaudi chabaudi* (AS strain)[52] were previously aligned with the *P. falciparum* genome using MUMmer to identify annotation-independent protein similarities[186]. We manually aligned an additional 177 contigs using linkage data from the *P. yoelii* genome publication and by performing BLASTN analyses with ~500-bp sized sequences from the ends of the RMP contigs, thus closing gaps in the synteny map and "walking" towards the telomeric ends. Linking of these 7,392 contigs through identification of overlapping contigs resulted in the generation of 910 cRMP contigs (see Figure 1A for an example of the procedure to generate cRMP contigs). The high level of nucleotide identity between the genomes of the three RMPs (*P. yoelii* versus *P. berghei*, 91.3%; *P. yoelii* versus *P. chabaudi*, 88.1%; and *P. berghei* versus *P. chabaudi*, 87.1%) facilitated this

process. The cRMP contigs that showed MUMmer hits to two different *P. falciparum* chromosomes revealed SBPs. Linkage between adjacent *P. y. yoelii* contigs had previously been established using Grouper[74], through the alignment of overlapping *P. yoelii* expressed sequence tags (ESTs) and by polymerase chain reactions (PCR)[51]. Combining these data with the 910 cRMP contigs resulted in the generation of 243 scaffolds of linked cRMP contigs. STS markers were used to determine chromosomal locations of the linked cRMP contigs. These markers included 79 previously described and 59 new markers strategically chosen based on the position of the SBPs (SOM Table S1). All markers were hybridized to chromosomes of *P. yoelii*, *P. berghei*, *P. chabaudi*, and *Plasmodium vinckei* that had been separated by pulsed-field gel electrophoresis (PFGE)[26].

Analysis of the synteny map of the cRMP and P. falciparum *genomes*

Intergenic sequences flanking the SBs at all 22 *P. falciparum* SBPs as well as the five subtelomere linked ends that are chromosome-internal in the RMPs (92 kb in total) were analysed for repetitive motifs using MEME[253]. The intergenic sequences of the 20 RMP SBPs for which sequence was available were also analysed. Non-syntenic genes were compared with the genome data of the different *Plasmodium* species by TBLASTN analysis, and the expression profiles and putative functions of these genes were investigated using data available from PlasmoDB[91,92,165,166]. The predicted protein sequences of the *tstk* family members were analysed for functional domains by SMART[254].

GRIMM (genome rearrangements in man and mouse)[67] was used to confirm the suggested minimum 15 recombination events. To test the significance of the association between SBPs and *P. falciparum*-specific gene content, we used computer simulations to reassign the 22 chromosome-internal SBPs to random positions in the core genome of *P. falciparum*, thus excluding the subtelomeric regions. We used two different approaches: the first approach utilized the sizes of the entire SBP regions, including the species-specific gene content, while the second approach utilized fixed SBP sizes (5 kb, slightly larger than the largest non-coding intergenic, intersyntenic regions). For both approaches, we counted the number of associations of the virtual SBPs of 1,000 random distributions with the locations of all inter- and intrasyntenic genes.

Phylogenetic analyses of members of the TSTK and SERA families were performed using manually corrected ClustalW alignments[189]. Protein parsimonies, pairwise distances and maximum likelihood distances were calculated using different regions of alignment with algorithms and matrices from the phylogeny inference package (PHYLIP)[255] and gave comparable results. For the final tree construction, 100 bootstrap trees were generated (each with 10x jumbling) of a manually corrected alignment of roughly 400 amino acids of the C-terminal ends of all TSTKs containing the serine/threonine protein kinase domain using SEQBOOT[256]. Maximum likelihood distances[204] were calculated using default parameter settings and 10x jumbling. The 100 bootstrap trees thus constructed were combined using CONSENSE[257]. The tree was rooted using the clade of non-*Plasmodium* TSTKs as the outgroup with RETREE, and the final tree was drawn using DRAWTREE, both also available from PHYLIP[255].

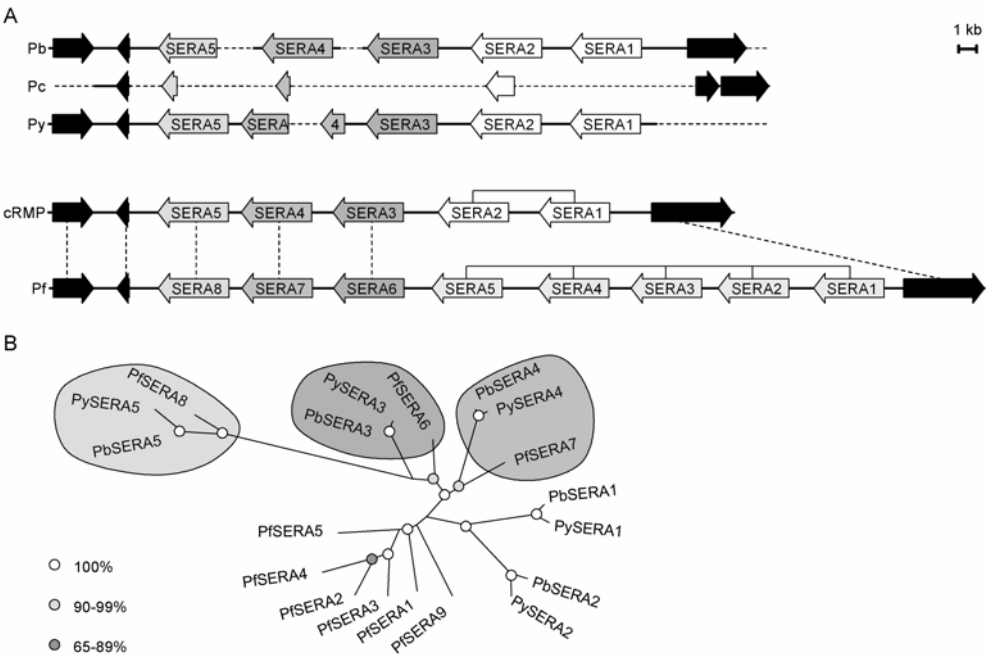


Figure 1. Deduced organization of the cRMP *sera* locus

(A) The combination of three *P. berghei*, six *P. chabaudi*, and two *P. yoelii* contigs (thick black lines) in a region of Pfchr2 containing eight *sera* copies demonstrates the strength of the "composite genome approach". Syntenic genes (black, linked by dashed vertical lines; left, PFB0315w and PFB0320c; right, PFB0365w) flank the *sera* clusters and reveal the presence of five *sera* genes in the RMPs. (B) Phylogenetic analysis revealed a close relation between *pfsera8*, *pfsera7*, and *pfsera6* and their syntenic orthologues in the RMPs (shaded grey, linked by dashed vertical lines in [A]). Other *sera* copies (*pfsera1-5*, *pbsera1-2*, and *pysera1-2*) clustered in species-specific groups (linked by solid horizontal lines in [A]). Circles represent branch points with bootstrap values of 100% (white), 90-99% (light grey) and 65-89% (dark grey).

A whole-genome synteny map of four *Plasmodium* species

A total of 7,392 contigs of the three RMPs, aligned with the *P. falciparum* genome, were used to generate 910 cRMP contigs (see Materials and Methods, Figure 1, Table 1, SOM Table S2). The tiling paths of all cRMP contigs are shown for both the individual *P. falciparum* and RMP chromosomes (SOM Tables S3-S30). The cRMP contigs that were syntenic with the *P. falciparum* genome totalled 17.2 Mb (75%) of the 22.9-Mb *P. falciparum* genome, equivalent to 90% of the predicted total region of synteny. After linkage of the aligned cRMP contigs 229 gaps remained. No synteny could be observed in the subtelomeric regions of chromosomes between RMPs and *P. falciparum*[51], largely due to divergence of subtelomeric repeat sequences and gene families, but also to the poor assembly of these regions in the RMP genome projects[52].

When the alignment of the cRMP contigs with the *P. falciparum* genome was examined, 19 were identified with MUMmer hits to two different *P. falciparum*

chromosomes, indicating that these contigs covered a SBP between the cRMP and the *P. falciparum* genomes. In addition, three SBPs were determined by chromosome mapping of STS markers and confirmed by PCR analysis, linking the cRMP contigs on either side of the SBP (unpublished data). In total, we found 22 SBPs in the core regions of the *P. falciparum* genome when compared to the core cRMP genome. Since the cRMP and *P. falciparum* genomes comprise 14 chromosomes, these 22 SBPs define a total of 36 SBs. Chromosome mapping of 138 *P. berghei* and *P. yoelii* STS markers (SOM Table S1) confirmed the 22 SBPs and the chromosomal location of the 36 SBs in the RMPs. The majority (23 of 28) of *P. falciparum* subtelomeric regions coincided with putative locations of cRMP subtelomeric regions, while the remaining five *P. falciparum* subtelomeric linked SBs were linked to SBPs in the cRMP genome. Conversely, five SBs that are linked to SBPs in *P. falciparum* were linked to subtelomeric regions in the cRMP genome. Appendix 1 shows the reciprocal synteny maps of the *P. falciparum* and cRMP genomes.

Centrally located AT-rich (CAT) regions of 2-3 kb (average >97% AT) found on all *P. falciparum* chromosomes (with the exception of *P. falciparum* chromosome 13; Pfchr13) have been predicted to be centromeres[222], and functional proof for their centromere function is accruing (S. Iwanaga, C.J.J., and A.P.W., unpublished data). While no CAT regions had been sequenced in the RMP genomes, genes immediately up- and downstream of 11 of the *P. falciparum* CAT regions were syntenic and located at 11 different cRMP chromosomes (Appendices 1 and 5). The predicted centromere of Pfchr7 is located in a SBP and therefore cannot be syntenic, and RMP sequences aligning with the predicted centromere of Pfchr6 did not show an elevated AT content in the cRMP chromosome. Assuming complete synteny of the CAT regions, we suggested new positions for the CAT regions of Pfchr6, 7, and 13 in the regions syntenic with cRMP chromosome 1 (cRMPchr1), 6, and 13, respectively. Unpublished releases of the latest *P. falciparum* sequences

Table 1: Summary of the characteristics of the cRMP contigs, scaffolds, SBs and SBPs.

	Pb contig	Pc contig	Py contig	All contig	cRMP contig	Contig gaps	cRMP scaffolds	Scaffold gaps	SBs	SBPs
Number	2,264	2,721	2,407	7,392	910	896	243	229	36	22
Av. size (kb)	4.8	3.0	6.4	4.7	18.9	1.9	75	3.0	533	16.2
Min. size (kb)	<1	<1	<1	<1	<1	<1	<1	<1	42	<1
Max. size (kb)	37	80	51	80	125	23	380	23	1,792	106
Total size (kb)	10,888	8,228	15,346	34,462	17,217	1,722	18,250	689	19,180	356
Increase contig size	393%	626%	297%	406%						
% syntenic region					90%		95%		100%	
% Pf genome size					75%		80%		84%	

Abbreviations: cRMP, composite rodent malaria parasite; SBs, synteny blocks; SBPs, synteny breakpoints; Pb, *P. berghei*; Pc, *P. chabaudi*; Py, *P. yoelii*; Pf, *P. falciparum*.

confirmed these predictions (M. Berriman, personal communication). These results indicate that each of the 14 cRMP chromosomes contained one of the syntenic regions surrounding the *P. falciparum* CAT regions. Cloning and sequencing of two 1.5-kb regions of cRMPchr5 and 13 (GenBank accession numbers DQ054838 and DQ054839) that aligned with the CAT regions of Pfchr10 and Pfchr13, respectively, revealed these were also extremely AT-rich (>97%) and consistent with the size and gene paucity of the *P. falciparum* CAT regions.

Comparison of the organization and location of common orthologous gene families of RMPs and *P. falciparum* allowed species-specific features of these families to be defined. For example, *P. falciparum* possesses a cluster of eight genes encoding putative serine proteases known as *sera* (PFB0325c to PFB0360c; *P. falciparum* gene models referred to in the text are available from PlasmoDB, http://plasmodb.org/). The *P. berghei* and *P. yoelii* databases both contain five *sera* (PB000649.01.0, PB000352.01.0, PB000107.03.0, PB107093.00.0, PB000108.03.0 for *P. berghei*; and PY02063, PY02062+PY00294, PY00293, PY00292, PY00291 for *P. yoelii*; *P. berghei* and *P. yoelii* gene models referred to in the text are available from GeneDB, http://www.genedb.org/, and GeneIndices, http://www.tigr.org/tdb/tgi/protist.shtml, respectively), whose organization in the individual RMP genomes was unresolved, yet could be reconstructed using the cRMP contigs, demonstrating one utility of the cRMP contig construction (Figure 1A). Combining the synteny analysis with standard phylogenetic analysis (Figure 1B) indicated that all RMP *sera* cluster at a single locus on cRMPchr3, which aligns with the *P. falciparum sera* cluster on Pfchr2. Within these clusters, direct orthologues for three *sera* (RMP *sera3-5* and *P. falciparum sera6-8*) were immediately adjacent and thus syntenic. The remaining RMP *sera1-2* and the *pfsera1-5* are also immediately adjacent to one another and each positioned similarly within the *sera* cluster in both genomes but form different phylogenetic clades and can be considered species-specific.

Inferring the pathway of synteny rearrangements between the cRMP and *P. falciparum* genomes

The organization of the three RMP genomes is highly conserved, and only one or two chromosomal rearrangements were noted when the genomes of the individual RMP species were compared with the cRMP genome (Appendix 1). The organization of the *P. berghei* genome is identical to that of the cRMP genome, suggesting it is also most similar to the genome structure of the most recent common ancestor (MRCA) of the RMPs.

The *P. falciparum* genome organization could be generated from the cRMP genome in a minimum of 15 recombination events when the following assumptions were made: (i) that the resulting genome always consists of 14 chromosomes; (ii) that all chromosomes always contain only one of the SBs containing a CAT region; and (iii) that a recombination event generating a subtelomeric from a chromosome-internal region (or vice versa, collectively termed telomere conversions) has happened only once. These 15 recombination events included eight single crossover events, five telomere conversions, one inversion of an entire SB, and one insertion involving an intersyntenic *var* cluster (Figure 2). This most parsimonious pattern of gross chromosomal rearrangements was supported by

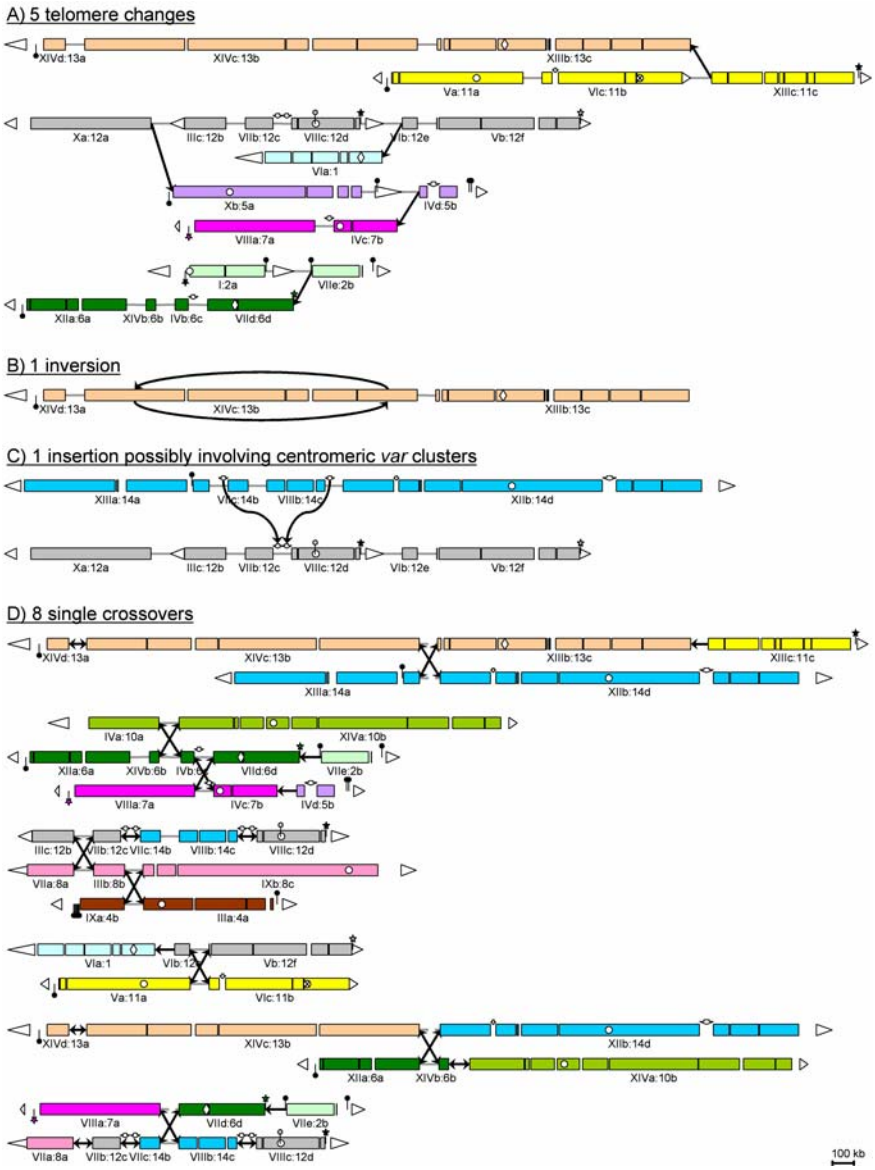


Figure 2. Schematic representation of the 15 recombination events

Schematic representation of the 15 recombination events that would permit the 36 SBs to be rearranged to generate the *P. falciparum* genome from the cRMP genome. See Appendix 1 for the numbering of the SBs and the symbols used in this figure. Grey lines between SBs represent links as present in the cRMP genome; grey dashed lines indicate intermediate links, and black arrows show links corresponding to the *P. falciparum* genome. Five subtelomeric regions of the cRMP genome must become chromosome-internal in the *P. falciparum* genome (A), thereby generating five subtelomeric regions in *P. falciparum* that are linked to SBPs in the cRMP genome. SB "XIVc:13b" is inverted (B), and SBs "VIIc:14b" and "VIIIb:14c" are inserted between SBs "VIIb:12c" and "VIIIc:12d", a process likely to involve chromosome-internal clusters of *var* and *rif* genes possibly mediated by *vicar* genes (C). Eight single crossover events generate the remaining links between the remaining SBs (D).

Table 2: Summary of inter- and intrasyntenic gene content of *P. falciparum* (Pf) and comparison to intersyntenic gene content of the rodent malaria parasites (RMPs).

	RMP intersyntenic	Pf intersyntenic	Pf intrasyntenic
Genes total	5	42	126
Gene families (%)	1 (20%)	30 (71%)	43 (34%)
Putative exported (%)	1 (20%)	37 (88%)	78 (62%)
var genes	-	14	9
rif genes	-	6	4
Other PEXEL/VTS genes	-	5	6
Genes with SP & TM-N[a]	-	7	40
Genes with TM-N[a]	1	5	19
Pseudogenes	0	11	12
Indels total	5	8	82
Indel sizes[b]	1 (1)	1-13 (5.3)	1-9 (1.5)
Indel sizes incl. pseudogenes[b]	1 (1)	1-20 (6.6)	1-10 (1.7)
Single gene indels	5	2	65
Multiple gene indels	-	6	17
Cluster sizes[b]	-	2-13 (6.7)	2-9 (3.6)
Cluster sizes incl. pseudogenes[b]	-	2-20 (8.5)	2-10 (4.3)

[a] Genes with a signal peptide (SP) and/or a transmembrane domain in their N-terminal ends (TM-N) were considered encoding potentially exported or surface proteins.
[b] The ranges and average gene numbers (with or without pseudogenes) per indel are shown.

analysis using the GRIMM algorithm[67] that identified one inversion and 15 translocations, counting the *var* cluster insertion as two single translocation events (unpublished data). The relatively low number of 15 rearrangements events suggests that gross chromosomal rearrangements resulting in the loss of or change in synteny is infrequent in *Plasmodium*. However, the same recombination events could be associated with the formation and dispersal of (members of) species-specific gene families.

P. falciparum-specific genes are found both at SBPs and in intrasyntenic indels

The average size of species-specific DNA regions located between SBs (intersyntenic regions) is significantly smaller in the cRMP genome (~2.5 kb, range 0.4-15 kb) than in the *P. falciparum* genome (~16 kb, range 0.7-106 kb). Only four of the 19 intersyntenic regions in the cRMP genome for which sequence data are available contain a species-specific open reading frame (ORF), but only the non-syntenic *c-ribosomal rna* (*c-rrna*) gene unit on cRMPchr5 is known to be expressed (Table 2, Appendix 6). In contrast, eight of the 22 intersyntenic regions in *P. falciparum* contain clusters of one to 13 genes without RMP orthologues (Table 2, Appendix 7). These 42 intersyntenic genes include 14 *var* and six *rif* genes, as well as five other genes, which all encode proteins containing the *Plasmodium* export element/vacuolar transport signal motif (PEXEL/VTS)[116,117] - for example, glycophorin-binding protein 130 precursor: GBP130, PF10_0159[258]; and two receptor-associated protein kinases: PfTSTK7a, MAL7P1.144, and PfTSTK10a, PF10_0160. The PEXEL/VTS motif is one element that is associated with transport of the proteins to the surface of the infected erythrocyte. A further 12 genes encode proteins with a transmembrane (TM) domain at the N-terminal end (for example, MAL7P1.58 of the *pfmc-2tm* family, which encodes proteins localized to the Maurer's clefts[146]), seven of which also have a signal peptide (SP; for example,

PF10_0164 of the *etramp* family[259] and five *var internal cluster associated repeat* [*vicar*] genes). Figure 3A provides a detailed example of the SBP on Pfchr10 and alignment of the flanking syntenic regions with *P. yoelii* contigs. In conclusion, it seems that the majority of the intersyntenic, *P. falciparum*-specific, SBP-associated genes encode predicted exported proteins destined for the membrane surface of the cell-free parasite or the infected erythrocyte.

In addition to the species-specific genes located at SBPs, *P. falciparum*-specific genes were also found clustered in small intrasyntenic regions that interrupt the SBs (*i.e.* indels; Table 2, Appendix 8). These 82 indels, including four *var* clusters, range in size from one to nine genes but are generally less gene-rich than the intersyntenic regions (1.5 genes/indel compared to 5.3 genes/SBP). Whereas only two of eight SBPs contain a single *P. falciparum*-specific gene, 65 of 82 of the intrasyntenic indels contain only one gene. The 126 intrasyntenic, *P. falciparum*-specific genes include nine *var* and four *rif* genes as well as an additional six genes with the PEXEL/VTS motif[116,117] including *pftstk13* (MAL13P1.109; see Discussion). Another 59 of these genes encode proteins with an N-terminal TM domain, 40 of which also contain a SP, giving a total of 78 genes encoding potential secreted or surface proteins. For example, a multigenic indel on Pfchr10 (Figure 3B) contains a cluster of six *P. falciparum*-specific genes that are all expressed in merozoites[11,91,92] and encode three known merozoite surface protein paralogues (MSP3, PF10_0345; MSP6, PF10_0346; and H101, PF10_0347), glutamate-rich protein (GLURP, PF10_0344), S-antigen (PF10_0343), and a hypothetical protein (PF10_0342) containing a SP sequence. The presence of a fourth *msp* paralogue H103 (PF10_0352) in the neighbouring syntenic region suggests that the gene content of this indel might have arisen in part through local gene duplication[260].

Evolution of gene families associated with recombination events at SBPs

In order to analyse whether recombination events in the core regions that resulted in the loss of synteny are associated with the dispersal and formation of species-specific gene families, all intersyntenic genes of *P. falciparum* and the RMPs were analysed for the presence and location of orthologous genes in their respective genomes. In addition to members of the *var*, *rif*, and *rrna* families, one intrasyntenic (*pftstk13*) and two intersyntenic (*pftstk7a* and *pftstk10a*) *P. falciparum* genes were identified that belong to a gene family encoding 21 transforming growth factor β (TGF-β) receptor-like serine/threonine protein kinases (PfTSTK)[261-263]. In addition to these three genes, 17 members are located in the subtelomeric regions of ten different chromosomes (Appendix 9) and one member is located adjacent to the Pfchr8 CAT region (M. Berriman, personal communication). In the RMP genome there is a single member of this family on cRMPchr12 syntenic to the copy near the Pfchr8 CAT region. Phylogenetic analysis groups these syntenic kinases in the same clade as the unique members of all other characterized *Plasmodium* species, with exception of the proteins encoded by the multiple *tstk* genes found in *Plasmodium reichenowi*, a very close relative of *P. falciparum* infecting chimpanzees[126]. These findings suggest that the syntenic *pftstk* on Pfchr8 could be the progenitor gene of this *P. falciparum*-specific gene family (Figure 4A).

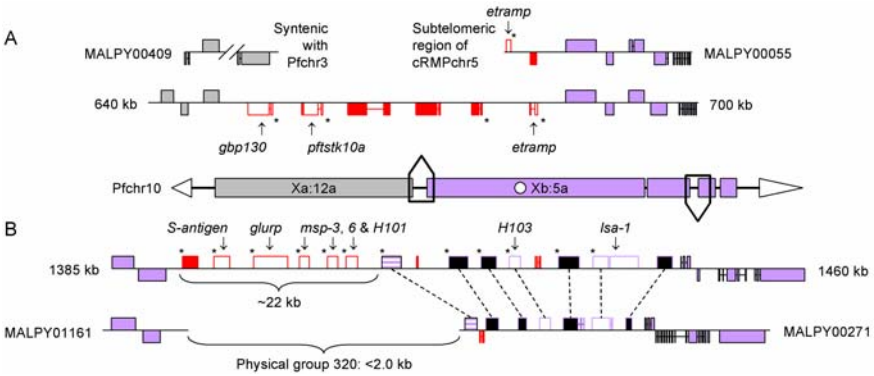
Two different recombination pathways that would generate the *pftstk* family are consistent with the data. (i) A copy of the syntenic, orthologous progenitor *pftstk* on

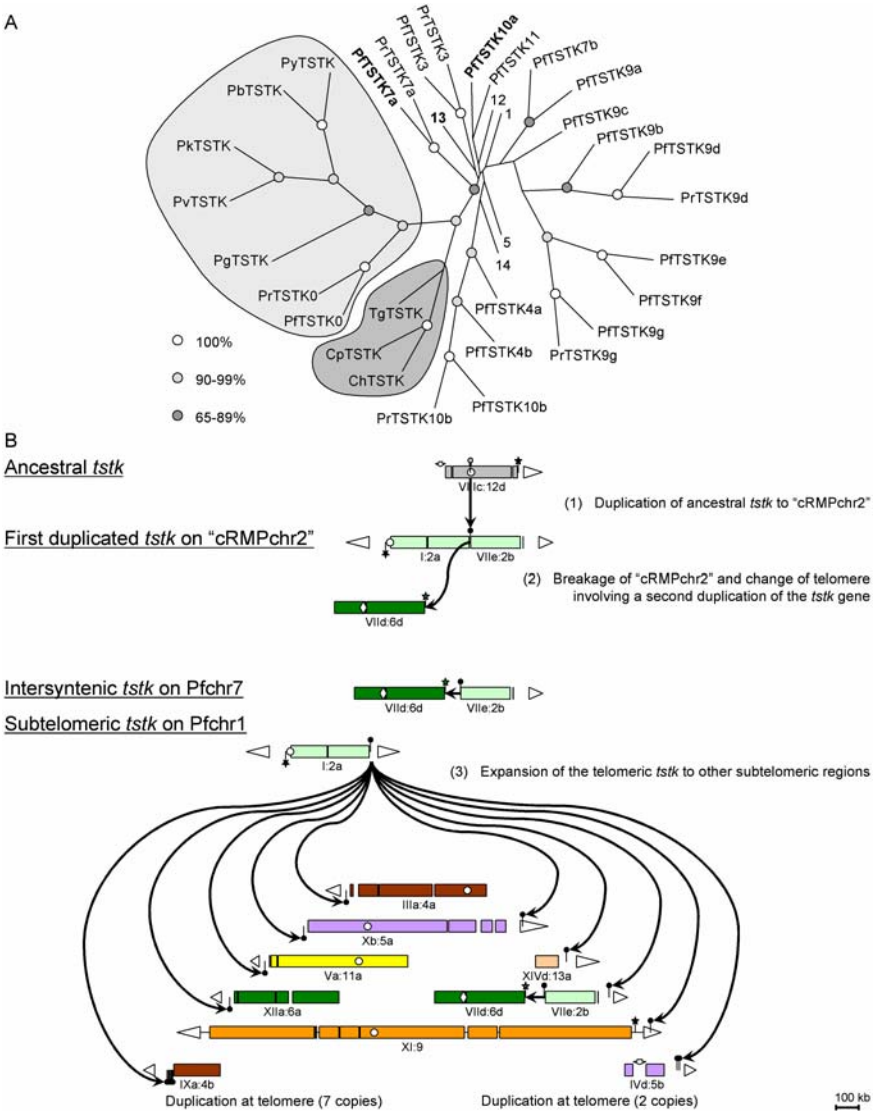
Figure 3. Inter- and intrasyntenic indels contain clusters of *P. falciparum*-specific genes

(A) A detailed illustration of the SBP between the SBs "Xa:12a" and "Xb:5a" (Pfchr10) flanked by *P. yoelii* contigs MALPY00409 (grey) and MALPY00055 (purple). The last gene on MALPY00409 is located on Pfchr3 ("IIIc:12b") and defines the SBP; MALPY00055 is the last syntenic contig flanking a subtelomeric region that contains a cRMP-specific gene encoding a hypothetical protein (red) and a non-syntenic *etramp* (white with red outline). The *P. falciparum* intersyntenic region contains three annotated genes (white with red outline): *gbp130*, *pftstk10a*, and *etramp*; and three genes encoding hypothetical proteins (red). Interestingly, four of six genes encode putative secreted proteins with N-terminal TM domains destined for the parasite surface or infected host cell membrane (asterisks; Appendix 7). (B) A detailed illustration of a ~22-kb indel within SB "Xb:5a" that contains *P. falciparum*-specific genes directly upstream of a region containing genes that are highly diverged in the RMPs. Only four of 12 genes annotated on MALPY00271 have a clear orthologue (purple) and the last gene (PY01020), which encodes a hypothetical protein, shows low similarity at the N-terminal end with PF10_0348 (horizontal purple lines). Comparison of the *P. yoelii* and *P. falciparum* annotations revealed the presence in both species of six genes with the same orientation and comparable size, including four genes that encode hypothetical proteins (black with purple outline) and two annotated genes (white with purple outline): the putative *P. yoelii lsa1* (PY01014/PB101910.00.0+PB105996.00.0/PF10_0356), and the putative *msp* paralogue *P. yoelii H103* (PY01016/PB105993.00.0/PF10_0352). MALPY01161 and MALPY00271 are physically linked as determined with Grouper software and are therefore between 500 and 2,000 bp apart, leaving no space for the remaining genes in the ~22-kb *P. falciparum* indel that include *S-antigen*, *glurp*, *msp3*, *msp6*, and *H101* (all white with red outline) and one gene encoding a hypothetical protein (red). In the entire regions, 12 of 15 genes encode putative secreted proteins destined for the parasite surface or infected host cell membrane (asterisks; Appendix 8).

Figure 4. Origin and putative mechanism of expansion of the *tstk* family in *P. falciparum*

(A) Analysis of *P. falciparum*-specific genes at the SBPs revealed a gene family encoding receptor-associated protein kinases (TSTK). Maximum likelihood distances were calculated for the C-terminal 400 amino acids of all TSTKs, including those found for other *Plasmodium* species, *Toxoplasma gondii*, *Cryptosporidium parvum*, and *Cryptosporidium hominis*. The tree was rooted using the clade with the three non-*Plasmodium* sequences as the outgroup (shaded dark grey). The syntenic progenitor genes clearly form one clade (shaded light grey), while the clustering of the other 20 mainly subtelomeric *pftstk* is more ambiguous (the three non-subtelomeric copies are shown in bold and include *pftstk7a*, which appears most closely related to the clade of progenitor genes). Circles represent branch points with bootstrap values of 100% (white), 90-99% (light grey) and 65-89% (dark grey). (B) See Appendix 1 for the numbering of the SBs and the symbols used in this figure. Based on the 15 recombination events described in Figure 2 and the phylogenetic analysis of the *tstk* family, we suggest the origin and putative evolution of the *pftstk* family as shown here. Phylogenetic analysis suggests that the intersyntenic *pftstk7a* is most closely related to the progenitor founder gene, *pftstk0*. Interestingly, this gene is the first non-syntenic gene upstream of SB "VIIe:2b". This SB is linked in the cRMP genome to SB "I:2a" that in *P. falciparum* is also flanked by a member of the *tstk* family, the subtelomeric *pftstk1*. Based on these observations we suggest that the founder gene *pftstk0* was duplicated after the split of *P. falciparum* from the other *Plasmodium* species but before SBs "VIIe:2b" and "I:2a" were separated (1). This gene was then directly involved in the breakage of this link, creating Pfchr1 ("I:2a") and destroying the telomere of "VIId:6d" by addition of "VIIe:2b" (2). During this recombination process, the gene was duplicated and is now present not only as two chromosome-internal copies on "VIIIc:12d" (*pftstk0*) and between "VIId:6d" and "VIIe:2b" (*pftstk7a*) but also as a first telomeric copy on the newly formed telomere of Pfchr1 (*pftstk1*). From here the gene could expand to the other subtelomeric regions (3). Local gene duplications resulted in the generation of seven copies on Pfchr9 and two copies on Pfchr4. After a copy of *pftstk* ended up at the left-hand cRMP subtelomeric end of SB "Xb:5a", the telomere conversion linked SB "Xa:12a" to SB "Xb:5a", which turned this telomeric copy into an intersyntenic gene (*pftstk10a*). The last non-subtelomeric copy, *pftstk13*, most likely resulted from a different process of mobility of *P. falciparum*-specific elements creating the intrasyntenic genes.


A
MALPY00409
Syntenic with Pfchr3
Subtelomeric region of cRMPchr5
etramp
MALPY00055
640 kb
700 kb
gbp130
pftstk10a
etramp
Pfchr10
Xa:12a
Xb:5a
B
S-antigen
glurp
msp-3, 6 & H101
H103
lsa-1
1385 kb
1460 kb
~22 kb
MALPY01161
MALPY00271
Physical group 320: <2.0 kb


A
PyTSTK
PbTSTK
PkTSTK
PvTSTK
PgTSTK
PrTSTK0
PfTSTK0
TgTSTK
CpTSTK
ChTSTK
PrTSTK10b
PfTSTK10b
PfTSTK4b
PfTSTK4a
PrTSTK9g
PfTSTK9g
PfTSTK9f
PfTSTK9e
PrTSTK9d
PfTSTK9d
PfTSTK9b
PfTSTK9c
PfTSTK9a
PfTSTK7b
PfTSTK11
PfTSTK10a
PrTSTK3
PfTSTK3
PrTSTK7a
PfTSTK7a
13
12
1
5
14
100%
90-99%
65-89%
B
Ancestral tstk
VIIc:12d
(1) Duplication of ancestral tstk to "cRMPchr2"
First duplicated tstk on "cRMPchr2"
I:2a
VIIe:2b
(2) Breakage of "cRMPchr2" and change of telomere involving a second duplication of the tstk gene
VIId:6d
Intersyntenic tstk on Pfchr7
Subtelomeric tstk on Pfchr1
(3) Expansion of the telomeric tstk to other subtelomeric regions
IIIa:4a
Xb:5a
Va:11a
XIVd:13a
XIIa:6a
XI:9
IXa:4b
IVd:5b
Duplication at telomere (7 copies)
Duplication at telomere (2 copies)
100 kb

Pfchr8 relocated to a subtelomeric region, where it underwent extensive gene duplication and redistribution. The centrally located *pftstk* genes could then have originated from telomere changes. (ii) Combining the information on the location and phylogeny of the *pftstk* family with the predicted 15 synteny rearrangements suggests that both chromosome-internal rearrangements resulting in the loss of synteny and subtelomeric recombination are associated with the evolution and distribution of this family (Figure 4B). *P. falciparum*-specific duplication/translocation of the ancestral *tstk* to an ancestral "cRMPchr2" followed by chromosome breakage and recombination may have led to the translocation of a *tstk* copy to a subtelomeric position (Pfchr1). Additional subtelomeric copies may be translocated to the nine additional subtelomeric locations by ectopic recombination events between different chromosomes similar to the events suggested to play a crucial role in the generation of *var* gene diversity[62]. The intersyntenic copy on Pfchr10 might be the result of a subsequent recombination event leading to the internalization of this gene. The intrasyntenic *pftstk13* may have originated independently of the mechanism that generated this gene family in a similar (if obscure) mechanism to other intrasyntenic genes with apparent subtelomeric origin, including the *var* and *rif* genes. All the predicted duplication and translocation events required to distribute the *pftstk* family could be linked to the proposed rearrangement pathway that converts the RMP genome organization to that of *P. falciparum*. Since there are alternative pathways for the order of the suggested SBP recombination events (also indicated by the GRIMM algorithm analysis; SOM Table S36), further elucidation of the pathway of recombination from the genome organization of the MRCA of *Plasmodium* awaits the availability of the genome of a third species[264].

Identification of a new putative gene family associated with chromosome-internal *var* clusters

Since repetitive sequences might be associated with recombination events between SBs, the intergenic regions flanking SBPs were examined using the MEME algorithm. This analysis resulted in the identification of a highly conserved *P. falciparum*-specific gene family consisting of seven putative genes and eight pseudogenes termed *var internal cluster associated repeat* (*vicar*) genes. These genes were found to be associated with five of seven chromosome-internal *var* clusters. Of these seven genes, five have a SP and five genes have one or two TM domains; only one of these genes is identified in the current annotation (MAL7P1.39) and is supported by transcriptome data[91]. The sequences correspond to the previously described GC-rich elements that were suggested to serve as regulatory elements for *var*-related genetic processes[222]. No other repetitive sequences were identified that, in the light of current knowledge, could be associated with chromosomal recombination events.

Discussion

The generation of composite contigs from three closely related *Plasmodium* species infecting rodents greatly facilitated the construction of a synteny map between the RMPs and *P. falciparum* and significantly reduced the need for experimental data from PCR and STS-mapping studies. Current contig assembly

algorithms rely upon a minimum of 95% sequence identity between sequence reads[179], a criterion not met by the RMP sequences. The high degree of synteny and similarity of gene content of the core *Plasmodium* genome enabled the compilation of cRMP contigs using sequences of the three RMPs with a lower sequence identity by aligning them to the assembled *P. falciparum* sequence. With only 229 gaps remaining and the location of 138 STS markers identified, the synteny map is a comprehensive tool for identifying the location of most genes. Individually, cRMP contigs are not sufficient to build an entire composite genome, since coverage and linkage of the cRMP scaffolds are incomplete. An unknown proportion of small rearrangements such as single gene insertions, inversions, or deletions will have been missed. Thus the need for continued sequencing to completion of at least one RMP genome remains. Approximately 4,500 (85%) of the 5,300 predicted *P. falciparum* genes have an orthologue in at least one of the RMPs, and these likely represent the core set of *Plasmodium* genes[52]. A similar level of orthology is seen in the genome organization, since the 36 SBs cover 84% of both genomes.

The synteny maps of *P. falciparum* and cRMP demonstrated that only a minimum of 15 recombination events are needed to generate the *P. falciparum* genome from the 36 SBs of the RMPs, compared with 245 events needed to convert the human genome organization to that of the mouse[67]. This relatively low number of *Plasmodium* genome rearrangements suggests either that divergence of *P. falciparum* and the RMPs might be relatively recent or that chromosomal rearrangements in *Plasmodium* are infrequent, either as a result of unknown (intrinsic) features of the DNA or due to some higher order organization of the genome[69]. Because the evolutionary relationships and the time of divergence between *P. falciparum* and other *Plasmodium* species is unclear[14,125,126,265-268], it is not yet possible to draw conclusions on the rate of chromosomal rearrangements in *Plasmodium*. A rough estimate consistent with published data would be that *P. falciparum* diverged and developed separately between 50 and 200 My ago. Thus the effective chromosomal recombination rate would be between 0.08 and 0.3 breaks/My. In comparison, the recombination rate in yeast species appears to be ~0.2 breaks/My[56]. Both are at the lower end of the range of rates observed for different mammalian species[247]. The genomes of different trypanosomatid species were also suggested to have a low recombination rate[250].

In many species, centromeres have been associated with chromosomal rearrangements and have proven to be positionally dynamic, with transposable elements often found to function in centromere relocation[70]. *Plasmodium* centromeres have not been functionally characterized but based on previous predictions, preliminary functional evidence (S. Iwanaga, C.J.J., and A.P.W.), and the distribution of the CAT regions as demonstrated by the *Plasmodium* synteny map, it is tempting to suggest that the predicted centromeres of *Plasmodium* are positionally static. One of the assumptions upon which the initial intuitive derivation of the minimum 15 recombination events was based was that each chromosome at any time always contains one CAT region and one only, in keeping with their still-hypothetical function as centromeres. The GRIMM analysis did not include such an assumption, yet it predicted the same number of rearrangements, while maintaining a single SB containing a CAT region in each newly formed

chromosome, emphasizing their predicted lack of involvement in the recombination events identified in this study. Furthermore, these recombination events are also unlikely to involve transposable elements, since these were not found in a cross-species comparison of the sequences in the vicinity of SBPs, consistent with previous studies[42].

In contrast to the low number of chromosomal rearrangements in the *Plasmodium* genomes, a relatively large proportion (15%) of the *P. falciparum* genes have no readily identifiable orthologue in any of the RMPs. These genes (including the well known *var, rif*, and *stevor* families) are mainly located in the subtelomeric regions, which appear to have a higher rate of gene evolution in many organisms, including *Plasmodium*[70,252]. However, this study shows that a significant proportion of *P. falciparum*-specific genes and members of gene families are not restricted to the subtelomeric region of the chromosomes but can be found as intrasyntenic indels and at SBPs. The majority (115 genes [68%]) of these 168 genes encode predicted or known surface or secreted proteins that are predominantly expressed in asexual blood-stage parasites (both infected erythrocytes and merozoites) and thus are involved in parasite interactions with the human host and possibly associated with immune selection/evasion. Interestingly, several of the larger clusters of genes, such as the indel containing *msp3* and *msp6*, appear to be coordinately expressed and may even be transcribed in an operon-like manner[269], despite earlier analyses that did not find evidence for the existence of such clusters[11]. Perhaps surprisingly, indels containing RMP-specific genes were not readily found, and although this may be in part due to the incomplete RMP genome sequence data that are currently available, the depth of coverage of the cRMP genome suggests that RMP indels are not as frequent as in *P. falciparum*. However, indels are not absent from the RMP genomes, and evidence is accumulating for RMP indels that contain members of the *pir* superfamily normally found in the subtelomeric regions reminiscent of the organization of the *var* family in the *P. falciparum* genome[52,145] (SOM Tables S3-S30).

To test whether SBPs are significantly more associated with chromosome-internal *P. falciparum*-specific genes than what might be expected based on a random distribution of the SBPs, we used computer simulations to generate randomly distributed SBPs in the genome and compared these with the inter- and intrasyntenic gene content. Using a conservative and a more relaxed approach (see Materials and Methods), we showed that based on a random breakage model, between 1.9 and 3.0 of the 22 SBPs on average could be expected to be associated with *P. falciparum*-specific gene clusters. This is significantly different ($p < 0.001$) from the observed association of eight (36%) of the 22 SBPs with *P. falciparum*-specific genes. This result indicates a non-random distribution of *P. falciparum*-specific genes associating with a higher frequency to SBPs and, therefore, with chromosomal rearrangements that have led to loss of synteny. Interestingly, from comparisons of the human and mouse genomes, evidence has emerged for a similar non-random distribution of repeat sequences in the genome and their association with SBPs[270,271].

The presence of members of species-specific gene families at the SBPs suggests that recombination events resulting in loss of synteny helped shape

species-specific gene content. SBPs and the intrasyntenic indels might therefore distinguish islands where variations in gene content occur (and then evolve) between the different *Plasmodium* species. The location and phylogeny of the *pftstk* family and the chromosomal rearrangements between SBs were consistent with different possible recombination pathways and mechanisms. Interestingly, the processes of gene duplication and translocation described for the *tstk* family could also be associated with the generation of two other gene families in *P. falciparum* encoding acyl-CoA binding proteins (ACP; four *P. falciparum* genes and one cRMP gene) and acyl-CoA synthetases (ACS; 11 *P. falciparum* genes and three cRMP genes). Both families have one syntenic copy in *P. falciparum* and the RMPs that are located in the *P. falciparum* genome next to an indel. The syntenic *acp* is located next to an indel on Pfchr8, and the syntenic *acs* next to an indel on Pfchr2 (PFB0685c). This latter gene appears to have undergone local gene duplication, followed by relocalization and expansion to seven subtelomeric copies in *P. falciparum* (unpublished data). In conclusion, our data show that both SBPs and intrasyntenic indels can be foci for species-specific genes with a predicted role in host-parasite interactions and indicate that not only rearrangements in the subtelomeric regions but also chromosomal rearrangements are involved in the generation of species-specific gene families. The majority are expressed in blood stages (complete list in Appendix 8), suggesting that the vertebrate host exerts a greater selective pressure than the mosquito vector, resulting in the acquisition of diversity.

It is already evident that a single recombinational mechanism underlying the origin of the inter- and intrasyntenic gene content or the generation of gene families in *P. falciparum* cannot be postulated. The 42 SBP-associated genes of *P. falciparum* can be classified into three groups: (i) two single genes that are associated with single crossover events; (ii) three clusters of genes (total 12 genes) that might have their origin in subtelomeric regions that became chromosome-internal after a telomere change (these include the SBPs containing *pftstk* genes); and (iii) three *var* clusters, two associated with the insertion of SBs “VIIc:14b” and “VIIb:14c” and one associated with a single crossover event (total 28 genes; Appendix 7). Thus it is clear that different recombination mechanisms were involved in shaping the *P. falciparum* genome. Evidence from both the 15 SBP-associated recombination events and previous *var* gene classifications[272] cannot be reconciled with an origin of central *var* clusters associated with telomere recombination changes and subsequent internalization of subtelomeric *var* genes. Both SBP and intrasyntenic *var* clusters are associated with the *vicar* genes identified in this study and previously described as the GC-rich elements[222]. The position of *vicar* elements is consistent with an as yet unproven role in recombination.

The pairwise whole-genome comparison presented here, while indicating that 15 chromosomal rearrangements can create the *P. falciparum* genome organization from that of the RMP, does not resolve the organization of the MRCA, which requires more complete *Plasmodium* genomes. Genome-wide comparison of the location and distribution of SBPs between different *Plasmodium* species should provide a reliable dataset enabling construction of a definitive phylogeny of the genus and resolving issues of precise clade topology[264]. In addition, whole-genome

comparisons and the identification of SBPs might prove to be an effective means of identifying species-specific genes and members of gene families that are involved in host-parasite interactions and immune evasion, including antigenic variation.

Acknowledgements

We would like to thank Matthew Berriman and The Wellcome Trust Sanger Institute for kindly providing pre-publication *P. falciparum* sequences and Ross Coppel for constructive criticism. We would like to thank the anonymous reviewers for their constructive criticism that resulted in a significant reshaping of this manuscript.

Notes

Supporting Online Material (SOM) accompanies the paper on the PLoS website (http://www.plos.org/) and includes SOM Tables 1-36. The sequences of two putative *P. yoelii* centromeres (Pychr5 and 13) have been deposited with GenBank (http://www.ncbi.nlm.nih.gov/) under the accession numbers DQ054838 and DQ054839, respectively. All datasets will become available through the official website of the *Plasmodium* genome project, PlasmoDB (http://plasmodb.org/)[165,166].

Chapter 6

Plasmodium berghei α-tubulin II: a role in both male gamete formation and asexual blood stages

Taco W.A. Kooij, Blandine Franke-Fayard, Jasper Renz, Hans Kroeze, Maaike W. van Dooren, Jai Ramesar, Kevin D. Augustijn, Chris J. Janse and Andrew P. Waters

Malaria Research Group, Department of Parasitology, Centre for Infectious Diseases, Leiden University Medical Centre, Albinusdreef 2, 2333 ZA, Leiden, The Netherlands.

Abstract

Plasmodium falciparum contains two genes encoding different isotypes of α-tubulin, *α-tubulin I* and *α-tubulin II*. α-tubulin II is highly expressed in male gametocytes and forms part of the microtubules of the axoneme of male gametes. Here we present the characterization of *Plasmodium berghei α-tubulin I* and *α-tubulin II* that encode proteins of 453 and 450 amino acids, respectively. α-tubulin II lacks the well-conserved three amino acid C-terminal extension including a terminal tyrosine residue present in α-tubulin I. Investigation of transcription by Northern analysis and reverse transcription-polymerase chain reaction (RT-PCR) and analysis of promoter activity by green fluorescent protein (GFP) tagging showed that α-tubulin I is expressed in all blood and mosquito stages. As expected, α-tubulin II was highly expressed in the male gametocytes but transcription was also observed in the asexual blood stages, female gametocytes, ookinetes and oocysts. Gene disruption experiments using standard transfection technologies did not produce viable parasites indicating that both α-tubulin isotypes are essential for the asexual blood stages. Targeted modification of α-tubulin II by the addition of the three C-terminal amino acids of α-tubulin I did not affect either blood-stage development nor male gamete formation. Attempts to modify the C-terminal region by adding a tandem affinity purification (TAP) tag to the endogenous *α-tubulin II* gene were not successful. Introduction of a transgene, expressing TAP-tagged α-tubulin II, next to the endogenous *α-tubulin II* gene, had no effect on the asexual blood stages but strongly impaired formation of male gametes. These results show that α-tubulin II not only plays an important role in the male gamete but is also expressed in and essential for asexual blood-stage development.

Introduction

Microtubules are subcellular components present in all eukaryotes that are central to a wide range of cellular processes including chromosome separation during mitosis, intracellular transport of organelles and cell motility. Furthermore, microtubules maintain the structural integrity and cytoplasmic architecture of the cell[273]. The major component of microtubules is tubulin, a heterodimer of two 50-55 kDa subunits: α- and β-tubulin. Many organisms express multiple α- and β-tubulin isotypes, the discrete functions of which are uncertain. Further diversity of tubulin proteins is generated by post-translational modifications, which might affect function[274]. In analysing the differences among tubulin isotypes, some appear to have no functional significance, some increase the overall adaptability of the organism to environmental challenges and some appear to perform specific functions, including formation of particular organelles and interactions with specific proteins[274]. Although the significance of the covalent modifications of tubulin is not fully understood, some of them may influence the stability of modified microtubules as well as interactions with certain proteins. Furthermore, they may help to determine the functional role of microtubules in the cell. Despite the variety of functions of the microtubules, the tubulins are highly conserved proteins, both between different isotypes within one species as well as between tubulins of different species[275].

Different species of *Plasmodium* express only one *β*- and two *α-tubulin* genes, *α-tubulin I* and *α-tubulin II*[276-282]. The nucleotide identity between the two *α-tubulin* genes is 85% and amino acid sequences are 95% and 40% identical when compared with each other and *β-tubulin*, respectively[277]. The most notable difference between the two predicted α-tubulin isoforms is that α-tubulin II lacks a terminal tyrosine residue[277,282], which is present in the great majority of *α-tubulin* genes. Interestingly, *P. falciparum α-tubulin II* was reported to be highly and specifically transcribed in male gametocytes[283,284] and studies using specific anti-α-tubulin II monoclonal antibodies showed the localization of α-tubulin II to the axoneme of the male gamete[283]. In contrast to other motile parasite forms that use a unique actomyosin motor to drive locomotion and host cell invasion (Refs. [133,285] for reviews), male gametes have microtubular axonemes that allow flagellar movement[286,287]. The molecular structure and function in motility and signalling of axonemes have been most extensively described for sperm flagella (Ref. [288] for review). The specific localization of α-tubulin II in the male gamete and its reported absence in asexual blood stages, female gametocytes and sporozoites has led to the suggestion that α-tubulin II has a specific and exclusive role in formation of the axoneme and motility of the male gamete[283].

In this study, we characterized the expression of the two *α-tubulin* genes of *P. berghei* in more detail through analysis of transcription by promoter tagging and genetic modification strategies. We show that, as expected, *P. berghei* α-tubulin II is highly expressed in male gametes and plays an essential role in gamete formation but unexpectedly its role appears not to be exclusive to the male gamete. Expression of α-tubulin II also occurs during asexual blood and mosquito stages and functional disruption of the *α-tubulin II* gene in blood stages was not possible.

Materials and methods

Parasites

The gametocyte producing reference clone, cl15cy1 (HP) of the ANKA strain of *P. berghei* was used. In addition, the non-gametocyte producer clone (HPE) of the ANKA strain was used[228,289].

Characterization of the two α-tubulin *genes*

To isolate DNA clones containing *α-tubulin* sequences, a partial *Sau*3AI-digested genomic *P. berghei* library in phage lambda zap-SK and a *P. berghei* cDNA library (kindly provided by M. Ponzi, Instituto di Sanitate Superiore, Roma, Italy) were screened with a *α-tubulin*-specific probe (L281/L282, 459 bp; this probe is based on a consensus *Plasmodium α-tubulin* sequence resulting from the comparison of published *P. falciparum* and *Plasmodium yoelii α-tubulin* sequences, Table 1). Plasmids from positive phages were isolated as described previously[290]. Selected genomic and cDNA clones were sequenced manually according to the dideoxynucleotide chain termination method using the T7 sequenase Kit version 2 (Amersham Biosciences, UK) or sequenced by BaseClear Molecular Biology Services BV (Leiden, The Netherlands). DNA sequences were analysed with the ClustalW alignment algorithm[189] using default settings.

Chromosomal locations of the *α-tubulin* genes were determined by hybridization using probes specific for the 3' untranslated region (UTR) of *α-tubulin I* (L389/L391,

290 bp) or the 5'UTR of *α-tubulin II* (L561/L562, 396 bp; Table 1) to pulsed-field gel electrophoresis (PFGE)-separated chromosomes[291]. Transcription of the *α-tubulin* genes was analysed by standard Northern blotting[167] of RNA isolated form synchronized asexual blood stages from HP and HPE parasites and gametocytes. The RNA was hybridized to probes specific for the 3'UTR of *α-tubulin I* (L389/L391) or the 5'UTR of *α-tubulin II* (L561/L562). In addition, stage-specific cDNA of the same *P. berghei* stages, as well as from maturing oocysts (day 7-10 after mosquito infection) and sporozoites from the HP clone, was produced from 1-2 µg DNAse treated RNA using both hexanucleotides and oligo d(T) primers with the Reverse Transcription System (Promega, The Netherlands) according to the manufacturers instructions. The cDNA was then used for standard RT-PCR analysis using primers L420/L484 (*α-tubulin I*, 1,035 bp gDNA, 355 bp cDNA) and L443/L444 (*α-tubulin II*, 717 bp gDNA, 388 bp cDNA; Table 1).

Analysis of promoter activity of α-tubulin I *and* α-tubulin II

Activity of the *α-tubulin* promoters was analysed through transgene GFP expression under the control of the two promoter regions. Promoter regions were amplified using primers L2207/L2208 (*α-tubulin I*, 1,471 bp) and L1516/L1517 (*α-tubulin II*, 1,259 bp; Table 1) and cloned into double-digested *Eco*RV/*Bam*HI pPbGFP$_{CON}$ vector. The construction of the pPbGFP$_{CON}$ vector for expression of GFP under the control of the *elongation factor-1α* (*ef-1α*) promoter has been described previously[292] and formed the basis for the generation of the two *α-tubulin*

Table 1: Primers used for the construction of transfection vectors, probes and for checking correct integration of DNA vectors.

Primer	Nucleotide sequence[a]	Restriction site	Construct	S or α-S[b]	Target[b]
L190	CGGGATCCATGCATAAACCGGTGTGTC	-	-	S	*pyrR2*
L191	CGGGATCCAAGCTTCTGTATTTCCGC	-	-	α-S	*pyrR2*
L281	TTTATGTTRTCWTCATATGCTCC	-	-	S	*Pf* & *Py α-tub I* & *II*
L282	CTAAATTCWCCTTCTTCCATACC	-	-	α-S	*Pf* & *Py α-tub I* & *II*
L389	AAAAAGCATATTAGATGTCTAAG	-	-	S	3'UTR *Pb α-tub I*
L391	GTAGAGAAAACATATTTTTATGG	-	-	α-S	3'UTR *Pb α-tub I*
L420	ACACATCAATGACTTCTTTACC	-	-	α-S	Exon 3 *Pb α-tub I*
L443	AGTTATTAGCATCCATGTTGG	-	-	S	Exon 1 *Pb α-tub II*
L444	TAAACCTGTACAATTGTCAGC	-	-	α-S	Exon 3 *Pb α-tub II*
L466	CCCAAGCTTGGATCCCACAGCATATGCTAATTATATAT	*Hin*dIII, *Bam*HI	pL0102	S	5'UTR *Pb α-tub I*
L467	CCCAAGCTTCATGTATACTTATTACTTCTCTC	*Hin*dIII	pL0102	α-S	5'UTR *Pb α-tub I*
L468	CCCAAGCTTGGATCCTGTAGATATATCCACATTTTACA	*Hin*dIII, *Bam*HI	pL0103	S	5'UTR *Pb α-tub II*
L469	CCCAAGCTTGCTAATAACTTCTCTCATTTTCG	*Hin*dIII	pL0103	α-S	5'UTR *Pb α-tub II*
L470	GGATATCCATCACCACAGGTTTCTACTGC	*Eco*RV	pL0102, pL0103	S	Exon 3 *Pb α-tub I* & *II*
L471	GGAATTCAAACTTCAGCAATAGCAGTTGAG	*Eco*RI	pL0102, pL0103	α-S	Exon 3 *Pb α-tub I* & *II*
L484	GAAGTAATAAGTATACATGTAGG	-	-	S	Exon 1 *Pb α-tub I*
L561	TGTGTACAGATATATTTTTCCAC	-	-	S	5'UTR *Pb α-tub II*

Primer	Nucleotide sequence[a]	Restriction site	Construct	S or α-S[b]	Target[b]
L562	TATCATATTATTGTAAAATGTCGG	-	-	α-S	5'UTR *Pb α-tub II*
L635	TTTCCCAGTCACGACGTTG	-	-	S	Plasmid backbone
L636	GGATAACAATTTCACACAGG	-	-	α-S	Plasmid backbone
L1382	GGAGGATCCATGGAAAAGAGAAG	*Bam*HI	pBSp48TAP	S	CBP-TAP tag
L1383	CCGCTCGAGGGTTGACTTCCCCGCGG AATTC	*Xho*I	pBSp48TAP	α-S	CBP-TAP tag
L1384	GCTCTAGATGAAAGAAGATCAGTAATA TGTAG	*Xba*I	pBSp48TAP	S	5'UTR *Pb p48/45*
L1385	CGCGGATCCACCAATTTTAATATTCATA AAACCAG	*Bam*HI	pBSp48TAP	α-S	5'UTR *Pb p48/45*
L1386	CCGCTCGAGGGTTCCGCATATTATGCT TTTC	*Xho*I	pBSp48TAP	S	3'UTR *Pb p48/45*
L1387	CGGGGTACCGATATCCGCATATCGAAA TGATGCTATC	*Kpn*I, *Eco*RV	pBSp48TAP	α-S	3'UTR *Pb p48/45*
L1482	CATGGATCGTCATCGGATCCTCACTAG TGTCTAGATAGC	Multiple	pb3DTAP	S	Multiple linker
L1483	GGCCGCTATCTAGACACTAGTGAGGAT CCGATGACGATC	Multiple	pb3DTAP	α-S	Multiple linker
L1516	CCGGATATCGGTAAGAGACTCCTGATG TGC	*Eco*RV	pL0106	S	5'UTR *Pb α-tub II*
L1517	CGCGGATCCTTTCGAATAAATTTATCTA AAATAG	*Bam*HI	pL0106	α-S	5'UTR *Pb α-tub II*
L1664	AAACTAGTAAGGTAAGAGACTCCTGAT GTGC	*Spe*I	pL0221	S	5'UTR *Pb α-tub II*
L1665	AAACTAGTCAACCAGATGGTCAAATGC	*Spe*I	pL1004	S	Exon 2 *Pb α-tub II*
L1666	TTCCATGGCTTCATATCCTTCATCTTCT CCTTC	*Nco*I	pL1004, pL0221	α-S	Exon 3 *Pb α-tub II*
L1681	CAAGTGCCCCGGAGGATG	-	-	α-S	TAP tag
L1888	GCAAAGGAGTATGAATCCTAG	-	-	S	5'UTR *Pb α-tub II*
L1889	CGGTGTAAACACATTTTATGTG	-	-	α-S	3'UTR *Pb α-tub II*
L2122	AAGGATCCAGGTATTCAAATCGGAAAT GC	*Bam*HI	pL1006	S	Exon 1 Pb α-tub II
L2123	GGGATATCACAGTGGGTTCTAAGTCAA CG	*Eco*RV	pL1006	α-S	Exon 3 *Pb α-tub II*
L2124	TTCTAAGCTTGCATTAAATGTTGATGTT ACCG	*Hin*dIII	pL1006	S	Exon 3 *Pb α-tub II*
L2125	CAGGTACCACTAAATTCTCCTTCTTCC ATACCC	*Asp*718I	pL1006	α-S	Exon 3 *Pb α-tub II*
L2130	TTGGATCCTACCCGGTGGAGACTTAGC	*Bam*HI	pL1007	S	Exon 3 *Pb α-tub I* & *II*
L2131	AAGAAAGCTTGTTTAATAGTCTGCCTC ATATCC	*Hin*dIII	pL1007	α-S	Exon 3 *Pb α-tub I*
L2132	ATGAAGGATATGAATAAACAAGC	(*Hin*dIII)	pL1007	S	Exon 3 *Pb α-tub II*
L2133	ACGATATCTATTATTATCCCTATACATA CGC	*Eco*RV	pL1007	α-S	3'UTR *Pb α-tub II*
L2134	ATGATATCGTTACCTTGATGGTATAC	*Eco*RV	-	α-S	3'UTR *Pb α-tub II*
L2136	TTCTAAGCTTTTAGAGTTTCAATATGA GCATAGTAGG	*Hin*dIII	pL1007	S	3'UTR *Pb α-tub II*
L2137	AAGGTACCAGCTCCACACAAAAATAAA TGG	*Asp*718I	pL1007	α-S	3'UTR *Pb α-tub II*
L2207	GGGATATCGCTGAGAAATTATAACATA CTTTGTAG	*Eco*RV	pL0255	S	5'UTR *Pb α-tub I*
L2208	CCGGATCCTTTTACTTGTATATTATAAA ATAAACAATTG	*Bam*HI	pL0255	α-S	5'UTR *Pb α-tub I*

[a] Underlined sequences indicate the restriction sites.
[b] The orientation of primers, sense (S) or antisense (α-S), is shown as compared to the target sequences of *P. berghei* (Pb), *P. yoelii* (Py) and *P. falciparum* (Pf).

vectors, pL0255 and pL0106 in which the *gfp* is placed under the control of the *α-tubulin-I* and *α-tubulin-II* promoters, respectively. Both vectors were linearized using the unique *Apa*I site in the *d-small subunit-ribosomal rna* (*d-ssu-rrna*) target sequence for integration into the genome by single-crossover homologous recombination into either the *c-* or *d-rrna* unit[292] (Figure 1C-E). Transfection and generation of parasite lines that express GFP under the *α-tubulin* promoters was performed as described below. GFP-fluorescence was visualized using fluorescence MDR microscopy (Leica; GFP filter settings) and images recorded using a DC500 digital camera.

Disruption and modification of the α-tubulin *genes*

Three standard replacement vectors that contain PCR-amplified target fragments of the *α-tubulin* genes for integration by homologous recombination were made, two for the disruption of *α-tubulin II* (pL0103, pL1006) and one for *α-tubulin I* (pL0102). For the construction of vectors pL0102 and pL0103, the first target sequences were amplified using primers L466/L467 (*α-tubulin I* 5'UTR, 699 bp) and L468/L469 (*α-tubulin II* 5'UTR, 677 bp; Table 1), respectively, which were then cloned into *Hin*dIII digested pb3D.$D_{T\Delta H}.\Delta D_b$, which contains the selectable marker cassette with the pyrimethamine-resistant *Toxoplasma gondii dihydrofolate reductase-thymidilate synthetase* (*tgdhfr-ts*)[293]. The resulting vectors were subsequently double-digested with *Eco*RI/*Eco*RV and a PCR fragment amplified with primers L470/L471 (*α-tubulin I* and *α-tubulin II* exon 3, 645 bp; Table 1) was introduced in both. The vectors were linearized for transfection using two *Bam*HI sites. For the construction of vector pL1006, the first target sequence was amplified using primers L2124/L2125 (*α-tubulin II* exon 3, 519 bp; Table 1) and cloned into *Hin*dIII/*Asp*718I double-digested pb3D.$D_{T\Delta H}.\Delta D_b$. The resulting vector was subsequently double-digested with *Bam*HI/*Eco*RV and a PCR fragment amplified with primers L2122/L2123 (*α-tubulin II* exon 2 and introns 1 and 2, 515 bp; Table 1) was introduced. The vector was linearized for transfection using the *Bam*HI/*Asp*718I sites. Transfection was performed as described below.

A DNA construct (pL1007) was made to convert the C-terminus of the *α-tubulin II* gene into that of *α-tubulin I* (an addition of nine base pairs encoding three amino acids, ADY). First, a 585-bp fragment that lies 903 bp downstream of *α-tubulin II* was PCR-amplified (L2136/L2137) and cloned into *Hin*dIII/*Asp*718I double-digested vector pb3D.$D_{T\Delta H}.\Delta D_b$. Subsequently, a fragment of 276 bp of the C-terminal sequence of *α-tubulin I* (L2130/L2131 double-digested with *Hin*dIII/*Bam*HI) and a fragment of 307 bp of the 3'UTR region of *α-tubulin II* (L2132/L2133 double-digested with *Hin*dIII/*Eco*RV) were simultaneously ligated into the second cloning site of the vector double-digested with *Bam*HI/*Eco*RV. The vector was linearized for transfection using the *Bam*HI/*Asp*718I sites. Transfection was performed as described below. The resulting vector contained the C-terminal sequence of *α-tubulin I* linked to the first 307 bp of the *α-tubulin II* 3'UTR sequence, followed by the *tgdhfr-ts* selectable marker cassette and a second fragment of the *α-tubulin II* 3'UTR (Figure 2A-C).

Two DNA constructs were made for the introduction of a modified *α-tubulin II* in the *α-tubulin II* locus encoding a tubulin protein that has a tandem TAP tag introduced at the C-terminal end. First, a generic vector (pb3DTAP) containing the

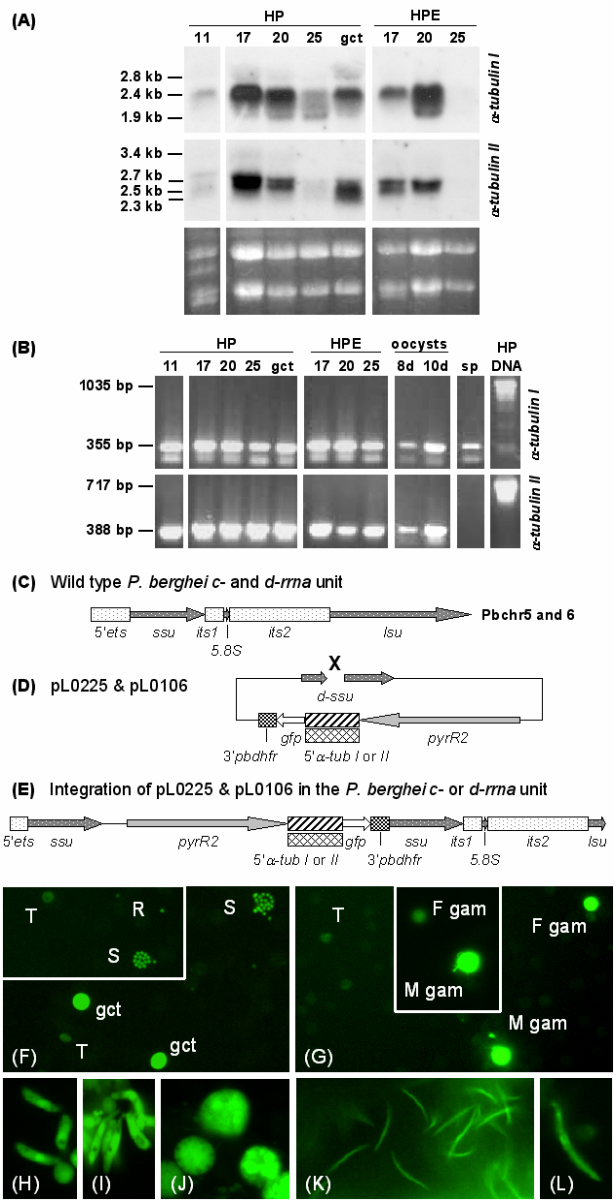


Figure 1. Transcription of *P. berghei α-tubulin I* and *α-tubulin II*

(A) Northern analysis of *α-tubulin I* (upper panel) and *α-tubulin II* (middle panel) messenger RNA (mRNA) of synchronized blood stages of HP and HPE parasites at 11 (mid-trophozoites), 17 (old trophozoites), 20 (young schizonts) and 25 (mature schizonts) hours after injection (hpi) of merozoites and mRNA of enriched HP gametocytes (gct). The lower panel shows the RNA loading in the agarose gel. (B) RT-PCR analysis of *α-tubulin I* (upper panel) and *α-tubulin II* (lower panel) in the same blood stages as in (A) and in oocysts (day 7-10 after mosquito feeding) and sporozoites (sp) both from HP parasites. (C) Schematic representation of the integration locus (the non-essential *c-* or *d-rrna* gene unit on Pbchr5 and 6) for the *gfp* gene under the control of either the *α-tubulin I* or the *α-tubulin II* promoter region. The *5' external transcribed spacer* (*5'ets*), *small* and *large subunit* (*ssu* and *lsu*), *5.8S* and both *internal transcribed spacers* (*its1* and *its2*) are shown. (D) Schematic representation of the two vectors (pL0225 and pL0106) that were used to introduce the transgene *gfp* under the control of the *α-tubulin* promoters into the genome of *P. berghei*. The target for integration *d-ssu* (part of the *d-ssu-rrna* unit) and the pyrimethamine-resistant *tgdhfr-ts* selectable marker cassette (*pyrR2*) are shown. (E) Schematic representation of the integration of vectors pL0225 and pL0106 in the *c-* or *d-ssu-rrna* unit. (F, H) GFP fluorescence of different blood stages (F) and ookinetes (H) of transgenic parasites expressing GFP under the control of the *α-tubulin I* promoter. (G, I-L) GFP fluorescence of different blood stages (G), ookinetes (I), oocysts (J, 14 day old; 40um) and sporozoites (K, L) of transgenic parasites expressing GFP under the control of the *α-tubulin II* promoter. R (ring stage; 2 μm), T (trophozoite; 2-5 μm), S (schizont; 6 μm), gct (gametocyte; 6-12 μm), F gam (activated female gametocyte; 6-12 μm), M gam (activated male gametocyte; 6-12 μm).

tgdhfr-ts selectable marker cassette was constructed with a TAP tag[294] linked at its 5' end to a multiple cloning site and at its 3' end to the *P. berghei p48/45* 3'UTR (for more information on the TAP tag vector see http://www-db.embl-heidelberg.de/jss/servlet/de.embl.bk.wwwTools.GroupLeftEMBL/ExternalInfo/serap hin/TAP.html). In three subsequent steps, the *P. berghei p48/45* 5'UTR (L1384/L1385, 1,105 bp, *Xba*I/*Bam*HI), TAP tag (L1382/L1383, 555 bp *Bam*HI/*Xho*I) and *P. berghei p48/45* 3'UTR (L1386/L1387, 1,006 bp, *Xho*I/*Kpn*I) were cloned in the pBSKS vector (pBSp48TAP). Subsequently, pBSp48TAP was

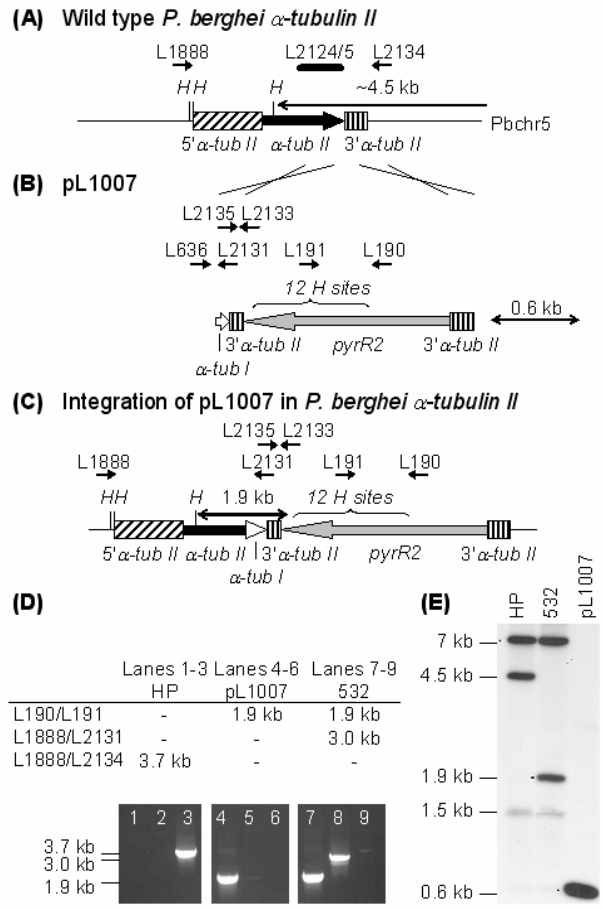


Figure 2. Replacement of the C-terminal sequence of *α-tubulin II* by that of *α-tubulin I* in the genome of *P. berghei*

(A) Schematic representation of the *α-tubulin II* locus on Pbchr5. The location of the probe L2124/L2125 is shown which is used for Southern analysis of *Hin*fI (*H*) restricted DNA (see E). This probe recognizes a ~4.5-kb fragment of *α-tubulin II* and a 7-kb fragment of *α-tubulin I* in wild type parasites. Small arrows show the primers used for PCR (see D). (B) Schematic representation of vector pL1007. This vector contains 276 bp of the C-terminal sequence of *α-tubulin I* and 307 bp of the 3'UTR of *α-tubulin II* followed by the pyrimethamine-resistant *tgdhfr-ts* selectable marker cassette (*pyrR2)* and 585 bp of the 3'UTR further downstream of *α-tubulin II*. (C) Schematic representation of the integration of pL1007 in the *α-tubulin II* locus resulting in the replacement of the C-terminal end of *α-tubulin II* by that of *α-tubulin I*. (D) Successful integration of pL1007 in parasite line 532 as shown by PCR. Details of the primers, DNA samples and the expected band sizes are listed above the figure. See (A) and (C) for the location of the primers. (E) Successful integration of pL1007 in parasite line 532 as shown by Southern analysis of *Hin*fI restricted genomic DNA hybridized to probe L2124/L2125. The wild type fragment of ~4.5 kb of *α-tubulin II* changes into a 1.9kb fragment after integration. The wild type band of 7 kb (*α-tubulin I*) is the same in wild type parasites and in 532. The linearized vector pL1007 shows the fragment of 0.6kb.

*Xho*I/*Not*I double-digested to replace the *P. berghei p48/45* 5'UTR sequence with a multiple cloning site containing a *Nco*I, *Bam*HI, *Spe*I, *Xba*I and *Not*I site that was amplified with two complementary oligonucleotide primers L1482/L1483 (39 bp). Finally, complete TAP tag cassette with multiple cloning site and *P. berghei p48/45* 3'UTR was cloned into *Xba*I/*Eco*RV double-digested pb3D.$D_{T\Delta H}$.ΔD_b containing the *tgdhfr-ts* selectable marker cassette (pb3DTAP).

The pb3DTAP vector was used for the construction of two vectors for the introduction of a TAP-tagged *α-tubulin II*. For the first construct (pL1004), 1,417 bp of the C-terminal sequence of *α-tubulin II* was amplified (primers L1665/L1666; Table 1) and cloned into *Spe*I/*Nco*I double-digested pb3DTAP. The vector was linearized for transfection using the unique *Mlu*NI site. Successful single cross-over integration of this construct in the *α-tubulin II* locus would result in the addition of the C-terminal half of the amplified *α-tubulin II* fragment linked to the TAP tag to the endogenous *α-tubulin II* gene, introduction of the selectable marker cassette and the generation of an incomplete second *α-tubulin II* lacking the promoter region, exon 1, intron 1 and half of exon 2. For the second construct (pL0221) the complete *α-tubulin II* gene including 1,261 bp of 5'UTR sequence was PCR-amplified (primers L1664/L1666; Table 1). The resulting 2,891-bp fragment was cloned into *Spe*I/*Nco*I double-digested pb3DTAP. After linearization at the unique *Mlu*NI site, integration of this construct by single cross-over in the *α-tubulin II* locus should result in the addition of the C-terminal half of the amplified *α-tubulin II* fragment linked to the TAP tag to the endogenous *α-tubulin II* gene, introduction of the selectable marker cassette and the generation of an additional complete *α-tubulin II* gene including the amplified promoter region (Figure 3A-C).

Transfection of parasites and selection of mutant parasites was performed as described[295]. Parasites transfected with constructs pL0221 and pL1007 were checked for correct integration by PCR and Southern analysis using an *α-tubulin II*-specific probe (L2124/L2125, Figure 2D-E, Figure 3D-E; Table 1).

TAP-tagged α-tubulin II was visualized in fixed thin blood films of blood stages and of gametocytes that were activated to form gametes[132]. Blood films were fixed in ice-cold methanol and incubated 1 hour at room temperature with human IgG (Sigma-Aldrich, The Netherlands) that binds to the IgG-binding domain of the TAP tag[294]. FITC-labelled goat anti-human IgG antibody (Sigma-Aldrich, The Netherlands) was used as a secondary antibody. In addition, we used human IgG-FITC conjugate (Sigma-Aldrich, The Netherlands) directly. Parasite nuclei were stained using 4,6-diamidino-2-phenylindole (DAPI, Sigma-Aldrich, The Netherlands) using the manufacturers instructions. Fluorescence was visualized using fluorescence MDR microscopy (Leica; GFP and DAPI filter settings) and images recorded using a DC500 digital camera.

Gametocyte and gamete production, fertilization and ookinete development of transgenic parasites were analysed using *in vitro* cultures as described[296,297] and fertility of gametes was analysed using *in vitro* cross-fertilization assays[132].

Characterization of the two *α-tubulin* genes

Sequencing of DNA clones, obtained from cDNA and gDNA libraries demonstrated that the *α-tubulin I* and *α-tubulin II* genes of *P. berghei* encode typical α-tubulin proteins of 453 and 450 amino acids, respectively, with an estimated size of

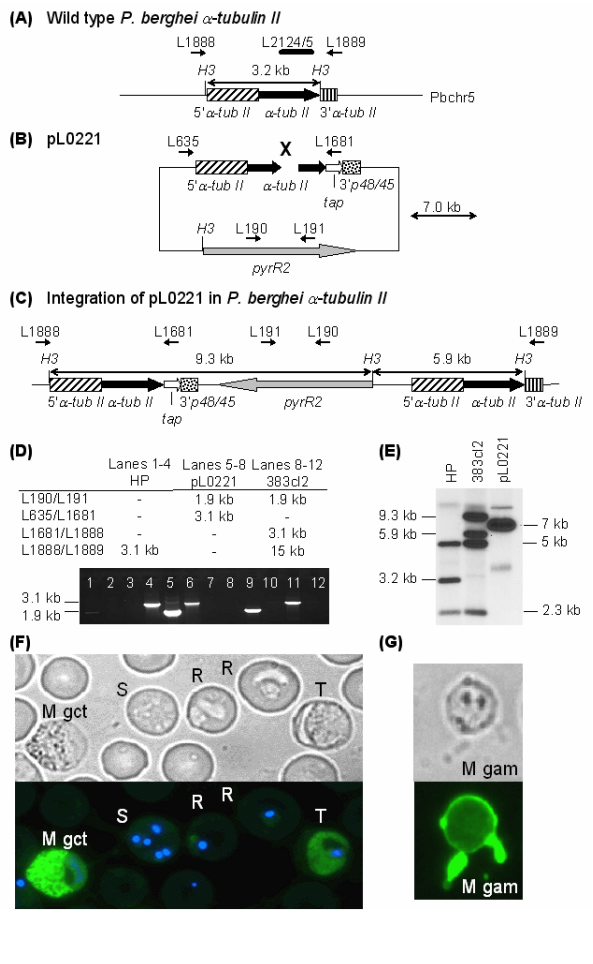


Figure 3. Introduction of a TAP-tagged *α-tubulin II* into the genome of *P. berghei*

(A) Schematic representation of the *α-tubulin II* locus on Pbchr5 in which the TAP-tagged *α-tubulin II* was introduced. The location of the probe L2124/L2125, which is used for Southern analysis of *Hin*dIII (*H3*) restricted DNA is shown (see E). This probe recognizes a 3.2-kb fragment of *α-tubulin II* and a 5-kb fragment of *α-tubulin I* in wild type parasites. Small arrows show the primers used for PCR (see D). (B) Schematic representation of vector pL0221 that was used to introduce a TAP-tagged *α-tubulin II* gene into the genome. The vector contains the complete *α-tubulin II* gene including 1,261-bp 5'UTR promoter sequence the 3'UTR of *P. berghei p48/45* as well as the pyrimethamine-resistant *tgdhfr-ts* selectable marker cassette (*pyrR2*). (C) Schematic representation of the integration event by which the TAP-tagged *α-tubulin II* gene is introduced into the genome. This event results in the duplication of the *α-tubulin II* gene and its promoter region thereby leaving both a wild type and a TAP-tagged copy. Small arrows show the primers used to show correct integration (see D). In addition, the *Hin*dIII (*H3*) restriction fragments of 5.9 and 9.3 kb that are recognized by probe L2124/L2125 are shown (see E). (D) Successful integration of pL0221 in parasite line 383cl2 as shown by PCR. Details of the primers, DNA samples and the expected band sizes are listed above the figure. See (A) and (C) for the location of the primers. (E) Successful integration of pL0221 in parasite line 383cl2 as shown by Southern analysis of *Hin*dIII restricted genomic DNA hybridized to probe L2124/L2125. The wild type fragment of 3.2 kb of *α-tubulin II* changes in two fragments of 5.9 and 9.3 kb after integration. The wild type bands of 5 and 2.3 kb (*α-tubulin I*) are the same in wild type parasites and in 383cl2. The linearized vector pL0221 shows the fragment of 7kb. (F, G) Immunofluorescent detection of the TAP-tagged α-tubulin II in blood stages (lower panel). Blood stages from tail blood or activated gametocytes were fixed with methanol and the TAP tag visualized by incubation with human IgG followed by staining with FITC-labelled goat anti-human IgG antibody. Male gametocytes (F) and activated male gametocytes (G) showed strong fluorescence, whereas old trophozoites (T; 2-5 μm) and female gametocytes show lower intensity. Interestingly, the activated male gametocytes that are not able to produce fertile gametes (see Results section) often showed strong fluorescent protrusions (G), most probably indicating impaired and aberrant formation of gametes/flagella. R (ring stage; 2 μm), T (trophozoite; 2-5 μm), S (schizont; 6 μm), M gct (male gametocyte; 6-12 μm) and M gam (activated male gametocyte; 6-12 μm). The upper panel shows a phase-contrast picture of the erythrocytes/parasites.

50 kDa. The genomic DNA sequences of the *α-tubulin I* and *α-tubulin II* gene loci have been deposited with GenBank (accession numbers DQ070855 and DQ070856). Partial sequences were also found in the *P. berghei* genome databases containing 1,936 bp and 926 bp (including both intron sequences) of the N-terminal sequences of both *α-tubulin I* and *α-tubulin II* as well as 1,048 bp and 1,305 bp of the respective 5'UTR sequences. Putative *α-tubulin I* and *α-tubulin II* genes were annotated and assigned the gene models PB000857.00.0 (*α-tubulin I*) and PB001519.02.0 (*α-tubulin II*). Both genes contain two introns (Figure 4A) as expected based on the presence of introns in the *α-tubulin* genes of *P. falciparum*[277-279] and *P. yoelii*[282]. A comparison of the α-tubulin proteins of *P. berghei*, *P. yoelii* and *P. falciparum* shows a high level of conservation (Figure 4B). The most marked difference between α-tubulin I and α-tubulin II in all three species is that α-tubulin I has three more C-terminal amino acids including a terminal tyrosine residue (ADY), which are absent from α-tubulin II. The sequencing of multiple cDNA clones of *α-tubulin II* revealed that transcripts extend at least 1.1 kb upstream of the start codon (Figure 4C) and the sequencing of the 3'UTR of six *α-tubulin II* positive cDNA clones indicates that (at least) three different putative polyadenylation sites exist, located 277 (3x), 286 (2x) and 295 (1x) nucleotides downstream of the stop codon (Figure 4C). Alignment with the genomic *α-tubulin II* sequence also demonstrated the presence of a 364-bp intron in the 5'UTR region only 16 bp upstream of the start codon (Figure 4C).

α-tubulin I and *α-tubulin II* are located on *P. berghei* chromosome 4 (Pbchr4) and 5, respectively, as shown by hybridization of probes to separated chromosomes that are specific to the *α-tubulin I* 3'UTR (L389/L391) or to *α-tubulin II* 5'UTR (L561/L562, unpublished data). Analysis and comparison of *P. yoelii* contigs MALPY00304 (containing *α-tubulin I*, PY01155) and MALPY01222 (containing *α-tubulin II*, PY04063) with the equivalent regions in *P. falciparum* chromosomes 9 (Pfchr9, PFI0180w) and 4 (PFD1050w) respectively, showed conservation of local gene organization and the absence of annotated genes within 1.3 kb of *α-tubulin I* and 2.4 kb of *α-tubulin II* respectively (Figure 4A). The lack of predicted genes in the vicinity of *α-tubulin II* is consistent with the large UTR regions characterized in the cDNA and may be important for genetic modification of the locus since in chromosome areas with a compact gene density, modification of the locus of interest might affect expression of neighbouring genes[60].

Transcription and promoter activity of the *α-tubulin* genes

Transcription of the *α-tubulin* genes was determined by Northern analysis, RT-PCR and promoter activity analysis. Northern analysis using probes specific for the *α-tubulin I* 3'UTR (L389/L391) and *α-tubulin II* 5'UTR (L561/L562) demonstrated that both *α-tubulin* genes are transcribed during blood-stage development, both in blood stages of a gametocyte producer clone, HP, as well as in blood stages of a non-gametocyte producer clone, HPE (Figure 1A). Prolonged exposure also revealed low transcript levels of *α-tubulin I* and *α-tubulin II* in both HP and HPE parasites at 25 hours (mature schizonts; unpublished data). In addition, the Northern analysis shows the production of transcripts with different sizes. A predominant transcript of 2.4 kb of *α-tubulin I* is present in all blood stages and a larger transcript of 2.8 kb present in HP that appears to be absent in the asexual

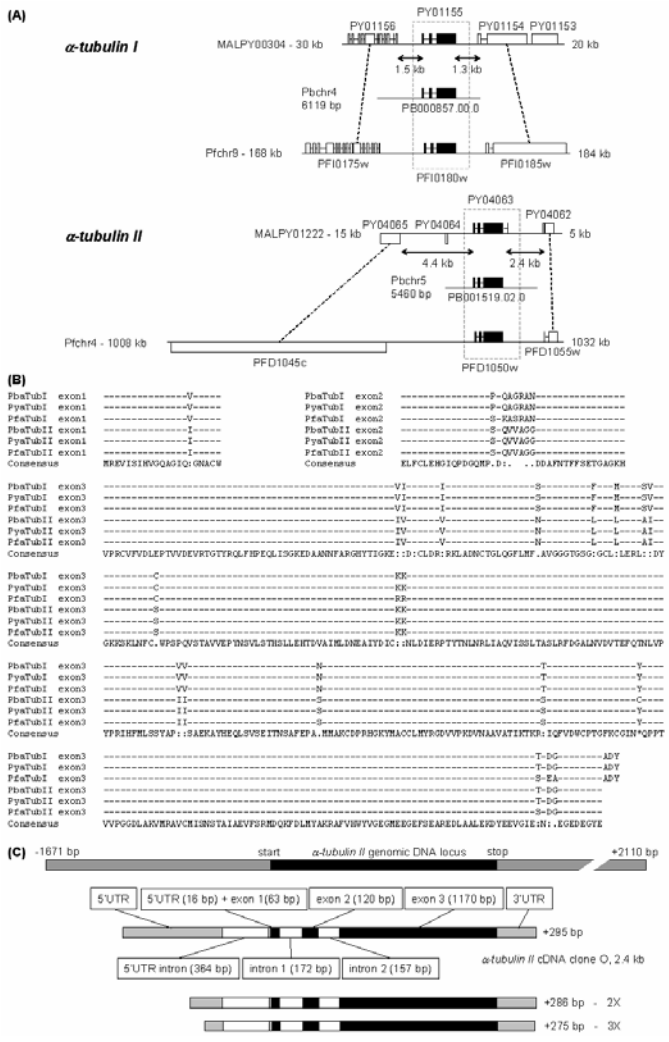


Figure 4. Comparison of the genomic location (A) and sequence (B) of *Plasmodium α-tubulin* genes and schematic representation of the *P. berghei* cDNA clones

(A) Sequencing cloned *P. berghei* DNA positive for *α-tubulin I* resulted in a 6,119-bp contig located on Pbchr4 (GenBank accession number DQ070855). The *P. yoelii α-tubulin I* orthologue (PY01155 on contig MALPY00304) is positionally conserved with *P. falciparum α-tubulin I* (PFI0180w on Pfchr9). Sequencing of *α-tubulin II* positive clones of *P. berghei* resulted in a 5,460-bp contig containing *P. berghei α-tubulin II* located on Pbchr5 (GenBank accession number DQ070856). *P. yoelii α-tubulin II* (PY04063 on contig MALPY01222) is positionally conserved with *P. falciparum α-tubulin II* (PFD1050w on Pfchr4). *Plasmodium α-tubulin* genes are shown in black and flanking syntenic genes in white (linked by dotted lines). (B) Sequences of *P. berghei*, *P. yoelii* and *P. falciparum* α-tubulin I and α-tubulin II were aligned with ClustalW[189] using standard settings. To highlight the differences, only the consensus sequence and those amino acid positions that are not identical between all six aligned sequences are shown. (C) Schematic representation of the six *P. berghei α-tubulin II* positive cDNA clones compared with the 5,460-bp genomic contig containing *α-tubulin II* (black). Partial sequencing of the largest cDNA clone, the 2.4-kb *α-tubulin II* cDNA clone O, confirmed the size and location of the three exons (black) and two introns (white) of *α-tubulin II*. It also demonstrated the presence of approximately 1,100-bp 5'UTR sequence (grey), a polyadenylation site 295 bp downstream of the stop codon and a 364-bp intron (white) in the 5'UTR region 16 bp upstream of the start codon of *α-tubulin II*.

blood stages of HPE. In all blood stages, transcripts of *α-tubulin II* are present with estimated sizes of 2.5 kb and 2.7 kb. In gametocytes, a smaller transcript of about 2.3 kb is also present, which is not detected in the asexual blood stages of the HPE. RT-PCR analysis confirmed the transcription of both *α-tubulin* genes in the different blood stages and demonstrated that, whereas transcription of both *α-tubulin* genes occurs in oocysts, only *α-tubulin I* transcripts are present in sporozoites (Figure 1B).

Promoter regions of *α-tubulin I* (1,471 bp) and of *α-tubulin II* (1,259 bp) were used to drive expression of GFP after stably introducing the *gfp*-promoter constructs in the non-essential *c-* or *d-rrna* units. GFP under the control of the promoter region of *α-tubulin I* is present in all blood stages and ookinetes (Figure 1F and H), with higher GFP fluorescence in female gametocytes then in males (unpublished data). As expected, GFP is highly expressed in male gametocytes under the control of the *α-tubulin II* promoter. However, low GFP expression was again observed in asexual blood stages (old trophozoites and developing schizonts) and also in female gametocytes (Figure 1G). We cannot formally exclude that this low expression of GFP in blood stages is the result of low, non-specific transcription of the introduced *gfp* gene, independent of the *α-tubulin II* promoter. However, the introduction of *gfp* in the same locus under the control of several other different sex-specific genes did not result in GFP expression in blood stages[154] (C.J.J. and A.P.W., unpublished data). Moreover, analysis of expression of TAP-tagged *α-tubulin II* also confirmed low expression of this gene in blood stages. The ratio of GFP expression in males and females under the control of the *α-tubulin II* promoter is 6:1[154]. Consistent with the RT-PCR study, GFP under control of the *α-tubulin II* promoter is also produced in different mosquito stages, such as ookinetes and oocysts (Figure 1I-L). Since GFP has a relatively long half-life, the GFP observed in sporozoites driven by the *α-tubulin II* promoter may well be carried over from the oocyst given that transcription of *α-tubulin II* was not observed sporozoites (Figure 1B).

Disruption and modification of the *α-tubulin* genes

Different standard DNA constructs were made to disrupt the *α-tubulin* genes by homologous recombination after transfection. We were unable to select for mutant parasites deficient in expression of either α-tubulin I (one construct, two independent experiments) or α-tubulin II (two constructs, seven independent experiments). These results indicate that not only α-tubulin I but also α-tubulin II is essential for asexual blood-stage development.

However, it was possible to modify the C-terminal sequence of *α-tubulin II* by homologous recombination, demonstrating that the *α-tubulin II* locus is accessible to genetic modifications. One construct used replaced the C-terminal 276 bp of *α-tubulin II* with those of *α-tubulin I*, thus introducing an additional 9 bp encoding the three additional C-terminal amino acids (ADY) of α-tubulin I. This modification had no detectable effect on asexual blood-stage development, gametocyte production or formation of male gametes and fertilization (Table 2).

Three independent transfection experiments with a DNA construct aimed at modifying the endogenous copy of *α-tubulin II* by introduction of a TAP tag were unsuccessful. Introduction of a TAP-tagged *α-tubulin II* gene, next to the

endogenous *α-tubulin II* gene was successful, generating a parasite with a normal development of the asexual stages. In the light of the normal development of the asexual stages, it was perhaps unexpected that male gamete formation was strongly impaired. In the parasites containing an extra copy of a TAP-tagged *α-tubulin II*, exflagellation was impaired (Figure 3G, Table 2), no mature, motile male gametes were produced and fertilization of female gametes was absent (Table 2). Cross-fertilization of the females containing a TAP-tagged *α-tubulin II* with fertile male gametes of the *p47*-knockout parasite line (line 270cl1)[154] demonstrated that the fertility of these females was not affected (Table 2). TAP-tagged α-tubulin II could be detected clearly in asexual trophozoites, gametocytes and activated gametocytes and at very low levels in ring stage parasites by immunofluorescence microscopy (Figure 3F-G).

Table 2: Gametocyte production, male exflagellation and ookinete production of the different *α-tubulin* mutant lines of *P. berghei*.

Parasite line (Plasmid)	Parasite	Gametocytes (CR)[a]	Exflag.[b]	Ookinetes (CR)[c]
cl15cy1	wild type	15-24% (6 exp)	82-98% (6 exp)	52-94% (6 exp)
544 (pL0225)	*gfp* under control of the *α-tub I* promoter, integrated into the *c-* or *d-rrna* gene unit	16% (2 exp)	88% (2 exp)	73% (2 exp)
606 (pL0225)	*gfp* under control of the *α-tub I* promoter, integrated into the *c-* or *d-rrna* gene unit	22% (1 exp)	83% (1 exp)	ND
357cl1 (pL0106)	*gfp* under control of the *α-tub II* promoter, integrated into the *c-rrna* gene unit	19% (2 exp)	94% (2 exp)	68% (2 exp)
357cl2 (pL0106)	*gfp* under control of the *α-tub II* promoter, integrated into the *d-rrna* gene unit	21% (2 exp)	89% (2 exp)	ND
532 (pL1007)	*α-tub II* with the C-terminal end replaced by the 3 amino acid extension of *α-tub I*	22% (2 exp)	91% (2 exp)	65% (2 exp)
382 (pL0221)	extra transgene *α-tub II* with TAP tag, integrated into the *α-tub II* gene locus	18% (2 exp)	4% (2 exp)[d]	0% (2 exp)
383cl2 (pL0221)	extra transgene *α-tub II* with TAP tag, integrated into the *α-tub II* gene locus	17% (2 exp)	2% (2 exp)[d]	0% (2 exp)
454 (pL0221)	extra transgene *α-tub II* with TAP tag, integrated into the *α-tub II* gene locus	23% (2 exp)	1% (2 exp)[d]	0% (2 exp)
270cl1	P47 deficient parasite[e]	20% (2 exp)	85% (2 exp)	0 (2 exp)
383cl2 crossed with 270cl1[e]	extra transgene *α-tub II* with TAP tag, CROSSED with P47 deficient parasite[e]	18% (383cl2); 20% (270cl1)	2% (383cl2); 90% (270cl1)	58% (4 exp)

[a] The gametocyte conversion rate (CR) is the percentage of blood-stage parasites that develop into gametocytes under standardized conditions [154].
[b] The percentage of exflagellating male gametocytes is the percentage of exflagellations *in vitro* under standardized conditions [154].
[c] The ookinete conversion rate (CR) is the percentage of female gametes/gametocytes that develop into mature ookinetes *in vitro* under standardized conditions [154].
[d] The very low numbers of exflagellating males were different from wild type exflagellations. They showed slow moving gametes that were in most cases unable to become detached from the host cell or the exflagellating male.
[e] The female gametes of 383cl2 were crossed with P47 deficient gametes of 270cl1 that produce fertile males but infertile females [154]. The ookinetes observed (CR of 58%) are thus formed by fertilization of the females of 383cl2 by the males of 270cl1.

Abbreviations: CR, conversion rate; ND, not done.

Discussion

The genomes of different *Plasmodium* species contain two genes encoding different isotypes of α-tubulin, *α-tubulin I* and *α-tubulin II*[277-279,282], which are differentially expressed[282-284]. The high expression of α-tubulin II in the male gametocytes of *P. falciparum* in which the protein is part of the microtubules of the axoneme of the male gametes[283], provided strong evidence that *α-tubulin II* is a male-specific gene and that the protein it encodes might be specific for axonemal microtubules. Many organisms have multiple genes encoding different isotypes of α-tubulin and often these genes are differentially expressed in certain cell types. However, in many cases the discrimination between functional significance of (small) structural differences between the isotypes that might affect the formation of microtubules of different organelles or a dose-dependent association with gene copy number is not clear[274]. The same is true for the possible functional effects of the many different post-translational modifications of α-tubulin that has been described in different organisms[274]. Here we show that *P. berghei*, like *P. falciparum* and *P. yoelii,* has two *α-tubulin* genes encoding two different α-tubulin isotypes. We characterized these genes in more detail with the aim to determine whether *P. berghei* α-tubulin II is a male-specific protein. Investigation of transcription of male-specific genes might provide insight into male-specific promoter elements and might lead to the development of tools to specifically express transgenes in male gametes.

We found that, as expected, α-tubulin I is expressed in all parasite forms examined including gametocytes with the higher GFP expression observed in female compared to male gametocytes. In contrast and also as expected, the promoter activity of *α-tubulin II* is very high in the male gametocytes but low GFP expression was also observed in the female gametocytes (6:1 ratio). This gender-specific activity has been exploited to physically separate the male and female gametocytes by differential flow sorting of gametocytes, based on differences in GFP expression[154]. In addition to the unexpected activity of the *α-tubulin II* promoter in female gametocytes, we found that the *α-tubulin II* promoter is also active in asexual blood stages and during oocyst development in the mosquito. Northern analysis revealed both *α-tubulin I* and *α-tubulin II* produce differently sized transcripts, a phenomenon previously described for a number of *Plasmodium* genes including the *α-tubulins* and *β-tubulin*[276,284]. Delves *et al.* described differentially sized transcripts for *P. falciparum α-tubulin I* with comparable sizes to our findings (2.5 kb and 2.9 kb)[284]. We report here that both *α-tubulin I* and *α-tubulin II* transcripts produced in gametocytes seem to differ in size from the transcripts present in asexual stages. A similar situation exists with *set* that encodes a histone associated protein required in both asexually dividing parasites and in male gametocytes where alternative promoters are used and splicing of gametocyte introns produce transcripts that differ in the size of their 5'UTR between asexual and sexual stages[298]. Extensive sequencing of *α-tubulin II* cDNA has yet to reveal alternative splicing as all cDNA clones isolated have the 5'UTR intron spliced out. It is promising that of two previously described *P. falciparum α-tubulin II* expressed sequence tags (ESTs) one appears to include the intron sequence (AU086114) while the other has the intron spliced out (AU088025)[299]. The *P. falciparum* 5'UTR intron is smaller (235 versus 364 bp) but located further

upstream (80 versus 16 bp) than the *P. berghei* intron. Analysis of promoter activity by GFP tagging of a modified *α-tubulin II* promoter lacking the 5'UTR intron and a modified *α-tubulin I* promoter with the intron introduced close to the start codon may shed more light on the role of this 5'UTR intron.

Expression of α-tubulin II is not restricted to the male gametocytes but also occurs during asexual blood-stage development and in female gametocytes (at low levels), ookinetes and oocysts suggesting that it does not have an exclusive function in the formation of microtubules of the axoneme of the male gamete. The unsuccessful attempts to disrupt *α-tubulin II* by standard technologies for generation of gene knockouts in *P. berghei* indeed strongly suggest that the protein is an essential component during asexual blood-stage development. This idea is supported by both the lack of closely linked genes in the genomic locus of *α-tubulin II* and the fact that *α-tubulin II* locus is amendable to genetic modification. Addition of the three amino acids (ADY) from α-tubulin I, including a C-terminal tyrosine residue, to α-tubulin II had no detectable effect on its function. In contrast, modifying the C-terminal region of the endogenous copy of *α-tubulin II* by adding a TAP tag did not result in viable parasites, indicating that the TAP tag inhibits the essential function of α-tubulin II during asexual blood-stage development. The TAP tag might directly affect the interaction of tubulin with other proteins or may affect essential but as yet unidentified post-translational modifications, since it is known that many post-translational modifications occur in the C-terminal region of tubulin proteins[274]. Surprisingly, the introduction of an extra TAP-tagged copy of *α-tubulin II* next to the endogenous (unchanged) copy appeared to have no effect on blood-stage development nor on female gamete fertility, whereas male gamete formation was strongly impaired. This differential effect in blood stages and male gametes is likely a dose effect, since the expression of α-tubulin II is considerably higher in male gametes than in the other blood stages. However, it cannot be excluded that due to differences in the molecular architecture of the male gamete axonemes and the microtubules of asexual parasites the TAP-tag only affects the former. Microtubules in yeast and a mammalian cell line have been successfully labelled with N-terminal GFP-tagged α-tubulins expressed as transgenes at relatively low levels (17%), whereas higher levels of expression proved toxic[300,301], supporting that a dose effect could be the cause of the differential effect in blood stages and male gametocytes.

In conclusion, *P. berghei* α-tubulin II is an essential protein expressed in every life cycle stage examined and merely over-expressed in the male gametocytes where it participates in the formation of the axonemal microtubules. The functional significance of the missing amino acids at the C terminus of α-tubulin II is not clear and apparently minimal. Further functional studies may benefit from the inclusion of small epitopes tags (Ref. [302] for review) that can be used for visualization of the precise structures in which the different tubulin isoforms participate.

Notes

Supporting Online Material (SOM) accompanies the paper on the Elsevier online library (http://www.sciencedirect.com/) and includes SOM Table 2 and SOM Figure 4. The sequences of *P. berghei α-tubulin I* and *α-tubulin II* loci have been deposited with GenBank (http://www.ncbi.nlm.nih.gov/) under the accession numbers DQ070855 and DQ070856.

Chapter 7

General discussion

High level of conservation of the organization and gene content of the RMP and *P. falciparum* genomes

Rodent malaria parasites (RMPs) are widely used models for the study of the biology of malaria parasites and especially for those life cycle stages that are technically or ethically less accessible for study in the clinically most important malaria parasite, *Plasmodium falciparum*, which infects over half a billion people world-wide and kills at least a million children in sub-Saharan Africa each year. In addition, RMPs have been extensively used for drug discovery and testing and for the identification and further characterization of proteins that are vaccine candidate antigens.

Although many characteristics of the morphology and biology of rodent and human malaria parasites show striking similarities, before the "genomics era" not much was known about the conservation of the molecular and biochemical mechanisms underlying these similarities. In the Leiden Malaria Research Group, studies had already been initiated prior to the genome sequencing initiatives to compare the genome organization and gene content of rodent and human malaria parasites by mapping studies of genes to pulsed-field gel electrophoresis (PFGE)-separated chromosomes[25,26], by long-range restriction mapping of individual chromosomes[72] and by comparing in detail the gene content and organization of specific genomic areas[60].

In this thesis, these studies have been extended to whole-genome analyses making use of the genome sequence initiatives that resulted in the publication of the genome sequences of the human malaria parasite *P. falciparum*[42] and three RMPs[51,52] (Chapters 3 and 4). The emphasis of this study was on the investigation of the level of conservation of genome organization and gene content between the RMPs and *P. falciparum*.

The genome sequences of the different *Plasmodium* species result from whole-genome shotgun sequencing projects. Genomic DNA is digested with a regular cutting enzyme and the resulting DNA fragments are cloned into vectors. Using standard plasmid-based primers, the cloned DNA fragments are sequenced and the single read sequences are assembled into contigs using an assembly algorithm. This resulted in the assembly of the complete genome of *P. falciparum*, however, sequencing coverage of the three RMPs (4-5x coverage compared with 14.5x coverage for *P. falciparum*) was too low to be able to assemble a complete genome sequence for any individual RMP. Assuming that the RMP genomes are very highly conserved, one could imagine that combining the sequences of the three species could enable the assembly of a composite RMP (cRMP) genome. The assembly algorithms typically require a minimum sequence identity of 95% between the single read sequences[179]. Although the RMP genome sequences are highly identical (88-92%, Chapter 4), this 95% criterion is not met and it is therefore not feasible to assemble a cRMP genome based only on the single read sequences of the three individual RMPs. Using the finished *P. falciparum* genome as a template, we aligned all individual contigs of the three RMPs, which enabled us to construct cRMP contigs of the overlapping contigs (Chapters 4 and 5). This approach was possible as the result of the high level of synteny both between the genomes of the three RMPs and between the *P. falciparum* and cRMP genomes.

After construction and alignment of all cRMP contigs to the *P. falciparum* template, only 228 gaps remained in the assembly of the cRMP genome. In combination with mapping 138 sequence tagged site (STS) markers to the RMP chromosomes, we demonstrated that the cRMP contigs were organized into 36 blocks that were syntenic with the *P. falciparum* genome. These 36 synteny blocks (SBs) represent 84% of the *P. falciparum* genome equivalent to at least 4,500 genes of the roughly 5,300 *P. falciparum* genes (85%), which can be considered the core set of *Plasmodium* genes.

Between the genomes of the three RMPs only one or two chromosomal translocations were found that disrupt synteny, suggesting that gross chromosomal rearrangements are infrequent in *Plasmodium*. The *Plasmodium berghei* genome was identically organized to the assembled cRMP genome, suggesting that it is most closely related to the genome of a most recent common ancestor (MRCA) of the RMPs. Due to the incompleteness of the genome sequence data of the RMPs and the impossibility to assemble a complete genome from one of the RMPs, small differences between the genomes of the different species, for example as the result of single gene insertions, inversions or deletions will have been missed. A completed genome sequence for at least one of the RMPs will be required to shed light on such small differences. It is tempting to speculate that *P. berghei* is the most suitable candidate for whole-genome sequencing since it has a genome organization, which most closely resembles that of a MRCA of the RMPs and would therefore be the most suitable standard RMP genome both for comparison with other RMPs and other *Plasmodium* species infecting primates and humans. Despite these possible small differences undetectable with the available genome sequences, our analysis shows a high level of conservation between the RMPs and *P. falciparum* genomes in the core regions of the chromosomes that are organized in only 36 SBs. In addition, the gene content in these regions is highly conserved with up to 97% of the centrally located *P. falciparum* gene content sharing an orthologue with at least one of the RMPs (in other words, the 85% of the total gene content of *P. falciparum* that is considered to be the core *Plasmodium* gene set).

The subtelomeric regions of chromosomes are not conserved between the RMPs and *P. falciparum*

In contrast with the highly conserved core regions of chromosomes, the subtelomeric regions appear to be highly variable both in organization and in gene content. Subtelomeric regions have been previously reported to vary in length through changing numbers of subtelomeric repeats, thus contributing to chromosome size polymorphisms in *P. falciparum*[303] and *P. berghei*[304-306]. Gene families colonizing the subtelomeric regions have long been understood to form an additional source of variability in malaria parasite chromosome size, organization and gene content. A majority of the identified species-specific genes located in the subtelomeric regions are thought to encode proteins distributed to the surfaces of the parasites or infected erythrocytes and hence are thought to be involved in antigenic variation, immune evasion or other host-parasite interactions. These gene families include amongst others the *var*[80-82], *rif*, and *stevor*[83,84] families in *P. falciparum* and members of the *pir* superfamily in *Plasmodium vivax*[144]

Plasmodium knowlesi and the RMPs[145,202]. RMP subtelomeric regions contain additional gene families typified by an 80-kb subtelomeric sequence of *Plasmodium chabaudi* that contains at least ten gene families, five of which have homologues in simian and human parasites, while the other RMPs have homologues of all ten gene families[235]. A first indication of the sharp boundaries separating the *Plasmodium* species-specific subtelomeric regions from the conserved core regions came from a comparison of a 200-kb fragment of a *P. vivax* chromosome with the genome of *P. falciparum*[131]. This sequence demonstrated a high degree of synteny with an internal fragment of *P. falciparum* chromosome 3 (Pfchr3) but synteny was lost entirely in the subtelomeric region harbouring arrays of *P. vivax*-specific *vir* genes. The availability of the genome sequences of *P. falciparum* and *Plasmodium yoelii* further strengthened the theory that species-specific subtelomeric sequences flank the highly conserved core regions, but the exact structure and gene content of *P. yoelii* (or any of the other RMPs) remains obscure to this date. Later analyses indicated that this initial conclusion was premature and several gene families located in the subtelomeres are conserved between numerous *Plasmodium* species including *P. falciparum* and the RMPs.

Despite the extreme variability in organization and gene content of the subtelomeric regions of the different *Plasmodium* species, the first clues are starting to emerge that many of the gene families that at first sight show no homology indeed perform similar functions and can be thought of as highly diverged paralogues rather than different gene families. One such an example is the *pir* superfamily[145,202] (Chapter 4), which is not only thought to exist of the *vir*, *kir*, *bir*, *cir* and *yir* families (of *P. vivax*, *P. knowlesi*, *P. berghei*, *P. chabaudi*, and *P. yoelii*, respectively), but may also include the *P. falciparum rif* genes. Structural comparison revealed another example of such a highly diverged gene superfamily, termed *pfmc-2tm*, that were found to encode proteins located in the Maurer's clefts[146]. In contrast, there appears to be no conservation of subtelomeric repeat sequences. The 21-bp repeat sequences (Rep20) found in *P. falciparum* subtelomeric regions[303] are not present in the RMPs and even between the RMPs there seems to be little conservation of these repetitive elements, exemplified by the 2.3-kb subtelomeric repeat elements that are unique for *P. berghei*[291,304,306].

Our comparative genome analysis of the RMPs and *P. falciparum* sharply defined all boundaries between the highly conserved core regions and the variable subtelomeric regions, which could be localized to a single intergenic region. Interestingly, the majority of these boundaries (23 of 28) was conserved between *P. falciparum* and the RMPs (Chapter 5). Unfortunately, due to the loss in synteny in the subtelomeric regions, RMP contigs could not be aligned in these regions and it was therefore not possible to construct subtelomeric cRMP contigs. However, manual BLAST analyses did reveal overlapping RMP contigs that crossed some of the subtelomere boundaries ensuring that the last cRMP contig in the tiling path did contain a short stratch of the subtelomeric sequence, thereby also indicating that there is at least some degree of synteny in the RMP subtelomeric regions. The extent of this subtelomeric synteny remains to be seen, since PFGE separation of RMP chromosomes indicates that both the sizes and gene content of the subtelomeric regions vary considerably not only between *P. falciparum* and RMPs but also between the different RMPs themselves.

Explanations for these differences in size and organization of the subtelomeric regions of the RMPs can be found in both the variation in number and sequence of subtelomeric repeat elements and variation in the copy number of members of the *pir* superfamily[145,202] (Chapter 4). This large gene family, which was first discovered in the human parasite *P. vivax*[144] and which has also been found in the primate-infecting *P. knowlesi*[145], is, as noted above, mainly located in the subtelomeric regions of the chromosomes but there is a great variety in estimated copy numbers between the different *Plasmodium* species. In order to be able to characterize the genomic organization and evolution of this important gene family, which is thought to play a role in antigenic variation and host-parasite interactions[144,145,202,307], it is essential to continue sequencing until at least one RMP genome is finished.

Further evidence of some degree of homology between the subtelomeric regions of *P. falciparum* and the RMPs came from an analysis of the 743 *P. falciparum*-specific genes without an RMP orthologue (the 736 genes reported in Chapter 4 plus the seven *vicar* genes described in Chapter 5). We found that 575 (11% of the total gene content of *P. falciparum*) are located in the variable subtelomeric regions (Chapters 4 and 5). These genes could be classified into 12 distinct gene families, of which five are shared with the RMPs. Based on the presence of a large number of *P. falciparum*-specific genes that are involved in host-parasite interactions and antigenic variation one could suggest that different species of *Plasmodium* have striking differences in their immune evasion strategies, however, in our opinion it is more likely that different *Plasmodium* species use the same mechanisms of immune evasion and that the lack of clear orthologues is merely due to host-specific adaptations and the extreme rates of recombination observed in the subtelomeric regions. Indeed, it has been suggested that the subtelomeric location of gene families is an essential factor in the generation of diversity in antigenic and adhesive phenotypes[62]. Clustering of telomeres at the nuclear periphery in asexual and sexual stages of *P. falciparum* facilitates ectopic recombination thus stimulating rapid evolution and diversification of genes encoding proteins involved in immune evasion and adaptation to the different hosts[24,62]. In this light, it is interesting to see if similar mechanistics to generate antigenic diversity in the RMPs might be in place. Continuing efforts to identify homologies between apparently unrelated gene families from different *Plasmodium* species as suggested for the *pir* and *pfmc-2tm* superfamilies mentioned above[145,146,202] (Chapter 4) should further improve our understanding of these important aspects of malarial infection.

Subtelomeres of *Plasmodium* chromosomes are rich in repetitive DNA sequences. Such repetitive DNA sequences have been postulated to play a significant role in karyotypic (chromosome) evolution and genome organization. In many organisms, including bacteria[308], yeast[309], plants[310], insects[311], worms[66] and mammals[61,63] reciprocal translocation and inversion breakpoints are associated with segmental duplications and are thought to be mediated mainly by homologous recombination of transposable elements, dispersed repeats and gene family members. In most eukaryotes, telomeres and centromeres consist of repeat sequences and are flanked by subtelomeric and pericentromeric regions, respectively, that have a tendency to accumulate (micro)rearrangements, *i.e.*

insertions, deletions, duplications and inversions[70]. Eukaryotic genomes with less than 10% repeats, including that of *Dictyostelium discoideum* (that like *P. falciparum* has an AT content of nearly 80%), show a bias towards the accumulation of transposable elements in these heterochromatic regions[312-315]. However, to date not one transposable element has been reported in the genome of any species of *Plasmodium*. Though the nature of the subtelomeric repeat-sequences varies amongst different organisms, an association with genome instability of the subtelomeric regions mediated by various forms of recombination is apparent. In *Plasmodium,* the subtelomeric instability and recombination activity are thought at least in part to serve a productive purpose in the generation of (diversity in) gene families encoding proteins involved in antigenic variation and thereby creating antigenic diversity[42,235] (Chapter 4). Although the generation of antigenic diversity could simply reflect the general instability of subtelomeric regions, clustering of telomeres at the nuclear periphery as reported for *P. falciparum* supports this idea[24,62].

In general, centromeres are not only composed of highly repetitive sequences but have proved positionally dynamic. This is exemplified by a comparative study amongst primates showing that even in relatively short evolutionary time frames centromere locations can change radically[316] possibly through the generation of new centromeres[317]. In contrast, centromere sequences, their positions and their binding proteins in highly diverged yeast species are conserved[318]. The *Plasmodium* synteny map presented in this thesis (Chapter 5) indicates that pericentromeric regions and even the putative *Plasmodium* centromeres, defined as gene-poor and AT-rich (typically >97%) regions of 1.5-2.5 kb, are completely syntenic, providing further support for the apparent absence of transposable elements from the *Plasmodium* genomes and indicating that the mechanisms for generating gene diversity in the subtelomeric regions might be different from those in other eukaryotes with transposable elements.

To explain the paradox of the highly conserved function of centromeres and their rapidly evolving, highly repetitive and complex sequences in plants and animals, a theory of meiotic drive during female meiosis was postulated[319]. During female meiosis, only one of each pair of chromosomes will be include in the egg nucleus, allowing for evolutionary competition between chromosomes. This drive is absent from yeast that has highly stable centromeres and possibly this might also be the case in *Plasmodium* species. In *Plasmodium* the haploid female gamete is fertilized by the haploid male gamete resulting in the formation of a diploid zygote in which meiosis occurs immediately after fusion of the male and female nuclei. Meiotic genome replication results in the presence of four haploid copies of the genome in the zygote within a single nucleus since nuclear division does not follow immediately after genome replication. Nuclear division and the formation of daughter cells occur only in the oocyst stage 10-12 days after meiosis and after multiple rounds of genome replications. It is unknown if all four genome copies or only a single one is involved in these multiple rounds of genome replication in the oocyst stage. If all four genome sets are used or when selection of a single set occurs after rather than during the meiotic division of the DNA, meiotic drive in *Plasmodium* might be absent as has been proposed for yeast. To support the theory of meiotic drive and its absence in yeast, Henikoff and colleagues showed

the adaptive evolution of centromere protein C (CENPC) in animals and plants but not in yeast[320]. Unfortunately, an initial attempt to identify orthologues of this protein in *Plasmodium* by motif searches with the CENPC motif did not reveal any candidate genes.

***P. falciparum*-specific genes are not only located in the subtelomeric regions but are also found at SBPs and in indels**

Through analysis of all 743 *P. falciparum*-specific genes and comparing their location in the genome using the synteny maps, we found that a significant proportion of *P. falciparum*-specific genes (168) is not located in the variable subtelomeric regions. Of these 168 *P. falciparum*-specific genes, 42 are identified at synteny breakpoints (SBPs) in eight intersyntenic indels and 126 are located in 82 intrasyntenic indels interrupting synteny. Interestingly, several SBPs and indels contain clusters of genes with similar orientation and expression profiles that may in part arise from gene duplication, such as the intrasyntenic cluster on Pfchr10 presented in Figure 4 of Chapter 5 containing merozoite-expressed genes including *msp3* and *msp6*. These genes may even be transcribed in an operon-like manner[269], despite earlier analyses which did not find evidence for the existence of such clusters[11].

Over two-thirds of the 168 non-subtelomeric *P. falciparum*-specific genes encode proteins that are predominantly expressed in asexual blood stages and contain an N-terminal transmembrane (TM) domain and henceforth are potentially secreted or exported to the surface of the parasite or infected erythrocyte. These include several known surface or secreted proteins as well as two newly discovered gene families. It is therefore likely that the *P. falciparum*-specific genes interrupting synteny play a role in immune evasion and host-parasite interactions indicating that not only recombination in the more volatile subtelomeric regions but also chromosome-internal rearrangements may influence diversity and complexity of the *Plasmodium* genome, increasing the ability of the parasite to successfully interact with its vertebrate host.

Interestingly, there is significantly more sequence information located at the SBPs that also contain considerably more genes in *P. falciparum* than in the 19 of 22 RMP SBPs for which sequence is available. In addition, indels containing RMP-specific genes were not readily found and although this may be in part due to the incomplete RMP genome sequence data that are currently available, the depth of coverage of the cRMP genome indicates that RMP indels are not as frequent as in *P. falciparum*. Despite the incomplete genome sequences of the RMPs, evidence is accumulating for the presence of indels in the cRMP genome containing up to 50 RMP-specific genes, ~40% of which appears to belong to the *pir* superfamily that are usually found in the subtelomeric regions reminiscent of the organization of the *var* and *rif* families in the *P. falciparum* genome[145] (Chapter 4). Despite these similarities, the data suggest some differences in the underlying mechanisms that are the cause of the micro-rearrangements and the generation of the species-specific gene content. Whole-genome synteny maps of other human and primate malarias, such as *P. vivax*[127] and *P. knowlesi* will reveal if intersyntenic genes are a *P. falciparum*-specific phenomenon.

The genome organization of the *P. falciparum* could be generated from the cRMP genome in a minimum of 15 gross chromosomal rearrangements

The level of synteny that exists between genomes of several related species appears to be proportional to the estimated evolutionary time separating them[32,67]. However, this is not always the case, possibly as a result of adaptations to environmental changes and alterations in life strategies that may influence the rate of rearrangements affecting synteny[54]. Two Diptera, the fruit fly *Drosophila melanogaster* and the mosquito malaria vector *Anopheles gambiae*, that diverged 250 million years (My) ago share roughly 50% orthologues[54]. Despite general conservation of chromosomal linkage of these genes, extensive reshuffling of genes within the chromosome resulted in just 34% of the genes to colocalize in microsyntenic clusters. This conservation of chromosomal linkage in combination with extensive reshuffling of gene order within the chromosomes was confirmed by comparison of *A. gambiae* with a second malaria vector, *Anopheles funestus*[249]. The two most closely related eukaryotic genomes sequenced to date are those of two nematodes, *Caenorhabditis elegans* and *Caenorhabditis briggsae*, that diverged approximately 110 My ago share 63% clear orthologues but as little as 4% of the *C. briggsae* genes do not have any homologue in *C. elegans*[38]. The genes were organized into 4,837 SBs larger than 1.8 kb (mean 37 kb) comprising 85 and 81% of their respective genomes. Changes in gene order were attributed to 244 putative translocation events as well as almost 1,400 inversions and just over 2,700 transpositions. Comparable to the levels of orthologues found between *P. falciparum* and the RMPs, the genomes of their respective vertebrate hosts, which diverged between 65 and 100 My ago, demonstrated roughly 80% one-to-one orthologues, organized into 281 SBs larger than 1 Mb that result from a minimum of 245 chromosomal rearrangements[67]. This means that the average rate of syntenic rearrangement since the divergence of human and mouse was roughly 2.5 breaks/My.

The time of divergence between *P. falciparum* and the other human infectious *Plasmodium* species as well as the RMPs is roughly 50-200 My[14]. By comparison of the synteny maps of *P. falciparum* and the RMPs, we demonstrated that a minimum of 15 gross chromosomal rearrangements are needed to generate the *P. falciparum* (core) genome from the 36 SBs of the RMPs (Chapter 5). This suggests that the average rate of syntenic rearrangement in *Plasmodium* is about 0.08-0.3 breaks/My, indicating that the core *Plasmodium* genome is considerably more stable than that of its host organisms and of the nematodes. Interestingly, only 1% of human genes have no homologue in the mouse genome, while 15% of the *P. falciparum* genes had no clear orthologues in any of the RMPs. 77% (575 of 743) of these genes are located in the subtelomeric regions many of which are members of gene families, suggesting that the rate of gene evolution in *Plasmodium* subtelomeres is significantly higher as opposed to the core regions of the chromosomes. Only between one and five translocations reshaped the chromosomal organization of four yeast species, *Saccharomyces cerevisiae*, *Saccharomyces paradoxus*, *Saccharomyces mikatae*, and *Saccharomyces bayanus* (that diverged around 5-20 My ago), which share a minimum of 95% one-to-one orthologues. The divergence between RMPs has been estimated at 18 My[14], which is in the order of the split between the four yeast species and like for

the yeast species as little as one or two translocations reshaped the genome organization of the RMP genomes.

Although a rearrangement pathway to generate the *P. falciparum* genome from the cRMP genome could be deduced, the availability of genome sequences of just two species did not yet allow us to generate a putative genome of the MRCA of *Plasmodium* for which at least a third genome is required[264]. Preliminary results of a comparison of the SBPs discovered between RMPs and *P. falciparum* with the contigs of the primate malaria parasite, *P. knowlesi* (http://www.sanger.ac.uk/Projects/P_knowlesi/), indicated that the cRMP genome organization was more similar to that of the primate malaria parasite (with five shared SBPs) than that of *P. falciparum* (one shared SBP; T.W.A.K. and A.P.W., unpublished data). With the expected completion of another human malaria parasite, *P. vivax*[127], it should prove possible to deduce the genome organization of the MRCA. As more genomes will become available, we can expect the construction of a more definitive phylogenetic tree for the *Plasmodium* genus based upon whole-genome organization. This will also enable the elucidation of the full pathway of gross chromosomal rearrangements that have generated the SB configuration of the genome of each present day species and might give insight into the role of these rearrangements in the generation and shaping of gene families and also reveal the progenitor genes that served as a template for further expansion into gene families. This possibility was illustrated in Chapter 5 of this thesis by the demonstration that the generation of a *P. falciparum*-specific gene family of 21 genes, encoding transforming growth factor β (TGF-β) receptor-like serine/threonine protein kinases (PfTSTKs), from a single progenitor gene shared by all other species of *Plasmodium*, could be linked to the gross chromosomal rearrangements that resulted in the loss of synteny.

***P. falciparum*-specific gene families and gross chromosomal rearrangements**

Most *P. falciparum*-specific gene families are located in the subtelomeric regions of the chromosomes. In previous studies on the location of members of such subtelomeric gene families, it had been shown that *var* and *rif* genes are not exclusively located in the subtelomeric regions but are also arranged in clusters in the internal regions of chromosomes. These clusters can vary considerably in size and were found to be as small as a single gene associated with two pseudogenes (Pfchr12) or as large as eight genes plus four pseudogenes (Pfchr7)[42].

Analysis of the synteny map that was generated to compare the genomes of *P. falciparum* and the RMPs revealed a number of genes belonging to *P. falciparum*-specific gene families that are located at SBPs in the core regions of the chromosomes. The presence of such species-specific genes at the SBPs indicated that recombination events resulting in gross chromosomal rearrangements of the core regions and loss of synteny are involved in the generation and shaping of species-specific gene (family) content and mark islands where species-specific variation in gene content can occur. In addition, we found that it is not uncommon that members of these gene families are located in intrasyntenic indels, which regularly contain more than one copy.

Though marginal, the amount of indels located close to the subtelomeric boundaries seems somewhat higher than in the more central regions of the

chromosomes. One such a nearly-subtelomeric indel contains a *pseudo-var* gene as well as two copies of the *cytoadherence-linked asexual gene* (*clag*)[321]. There are four *pfclag* genes, which are located in the subtelomeric regions of Pfchr2 and 9 and, as mentioned above, in a nearly subtelomeric indel on Pfchr3. None of these genes appears to be directly syntenic with any of the three RMP *clags*, although these were all shown to be located on chromosome 8 (cRMPchr8) that contains a region syntenic with Pfchr9 and that is flanked by the subtelomeric region containing the *clag* gene (D.L. Gardiner, personal communication). These data suggest that this gene family originated prior to the split between *P. falciparum* and the RMPs and may have been formed by local gene duplication in the subtelomeric region in a MRCA of *P. falciparum* and the RMPs and subsequent redistribution of *clag* genes in *P. falciparum*. Alternatively, the *clag* family might have formed in the MRCA followed by species-specific gene loss after the split between rodent and human malaria species.

For two other gene families specifically expanded in *P. falciparum*, we found that all the RMP genes are syntenic with one of the members of the *P. falciparum* gene family. These are the gene family encoding ACPs (four *P. falciparum* genes, one RMP gene) and the gene family encoding ACSs (11 *P. falciparum* genes, three RMP genes). In *P. falciparum*, one syntenic copy of each of these gene families is located next to an indel. One of these, the syntenic *acs* located on Pfchr3, appears to have undergone local gene duplication generating a *P. falciparum*-specific intrasyntenic copy that may have undergone subsequent relocalization and expansion to the seven *P. falciparum*-specific subtelomeric copies.

We could also associate four of seven chromosome-internal *var* clusters that are located in the core regions of the chromosomes with the gross chromosomal rearrangements that affected synteny, suggesting that gross chromosomal recombination also influences copy numbers and gene content of this important gene family encoding proteins that are involved in antigenic variation and immune evasion. Conversely, chromosome-internal *var* clusters may have facilitated gross chromosomal rearrangements. Interestingly, the analysis of the intergenic regions flanking *P. falciparum* SBPs revealed a yet undiscovered putative new gene family, which we named the *var internal cluster associated repeat* (*vicar*) genes. The location of these genes suggested that these genes could be linked to the recombination events that are involved in the generation of the chromosome-internal *var* clusters. The positions of two *vicar* genes on the opposing flanks of the intersyntenic *var* clusters of Pfchr7 and 8 (Figure 1) could suggest that *vicar* genes are involved in a recombination event resulting in the initial insertion of a single, large *var* cluster (bounded by the two copies of *vicar*) that was later split creating the two intersyntenic *var* clusters that now reside on Pfchr7 and 8.

Another intriguing gene family specific for *P. falciparum* that was discovered by analysing the genes at SBPs is the *tstk* family. In *P. falciparum*, this gene family consists of 21 copies rather than the 20 reported previously[261-263], most of which have a subtelomeric location. All other *Plasmodium* species, except for *Plasmodium reichenowi* that is closely related to *P. falciparum*, contain only a single copy that is located in the core regions of the chromosomes and is syntenic with a *P. falciparum tstk* located on Pfchr8.

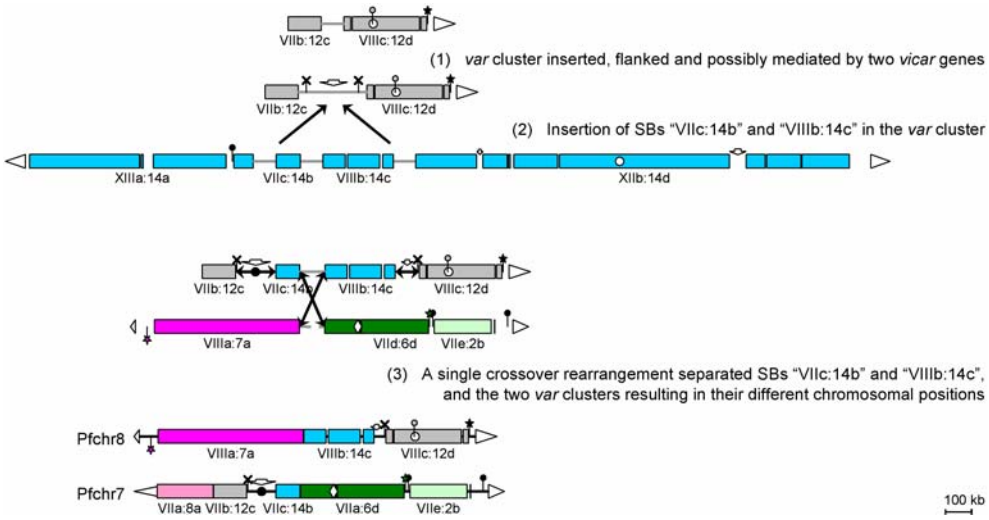


Figure 1. Putative mechanism of expansion of chromosome-internal *var* clusters through mediation of a new family of *var internal cluster associated repeat* (*vicar*) genes

See Appendix 1 for the numbering of the SBs and the symbols used in this figure. All 15 *vicar* genes are located within the chromosome-internal *var* cluster and three of these form the border with the regions syntenic to the RMPs. Two SBs ("VIIb:12c" and "VIIIc:12d") that are linked in the RMPs are both flanked by one of these *vicar* genes in *P. falciparum* and separated by an chromosome-internal *var* cluster from two other SBs that are linked in the RMPs ("VIIc:14b" and "VIIb:14c"). A possible explanation for these observations could be the insertion of a chromosome-internal *var* cluster mediated by one or more *vicar* genes separating SBs "VIIb:12c" and "VIIIc:12d" (1). This was followed by the insertion of SBs "VIIc:14b" and "VIIb:14c" within this cluster (2). Subsequently, a single crossover event could have caused the separation of these two clusters to different chromosomes of the *P. falciparum* genome (3).

By combining information on the location and phylogeny of the members of the *pftstk* family and the gross chromosomal rearrangements between SBs, we provided evidence that the formation of this gene family might originate from a recombination event that locates a copy of the "core" founder gene in the subtelomeric regions that may then have been amplified and translocated to subtelomeric regions of other chromosomes. All the predicted duplication and translocation events required to distribute the *pftstk* family could be linked to the proposed rearrangement pathway that converts the cRMP genome organization to that of *P. falciparum*.

At this moment it is completely unknown why *P. falciparum* and *P. reichenowi* have multiple copies of this kinase while all other species need only one copy. It is clear from experiments with both monkey and human parasites that the expression of variant antigens at the RBC surface[322] as well as switching between different family members[323] is controlled at least in part by host factors. The molecular mechanisms underlying these processes are currently unexplored but it may be expected that active signalling between host and parasite molecules will be involved. It is tempting to speculate that the *P. falciparum*-specific gene family, *pftstk*, could play a role in this process. Several features make this gene family of

particular interest for studying host-parasite interactions at a molecular level. Like many proteins involved in host-parasite interactions they: (i) are encoded by genes that are predominantly located in subtelomeric regions; (ii) are highly divergent; (iii) all have TM domains, a predicted signal peptide (SP), and a *Plasmodium* export element/vacuolar transport signal (PEXEL/VTS)[116,117]; and (iv) are encoded by genes that are transcribed at the late ring and (early) trophozoite stages, just prior to the onset of other genes involved in antigenic variation such as the *var* genes. Sera from humans living in endemic areas were shown to recognize one of the more highly-expressed *pftstk* family members[324].

The general structure of the PfTSTKs resembles that of serine/threonine protein kinase TGF-β receptors that are active in signal transduction via SMAD proteins in various human tissues as well as in many other invertebrates (Refs. [325,326] for reviews). Initial attempts to identify genes encoding SMAD-like proteins in the *P. falciparum* genome using motif-based searches have not revealed any candidates thus far but this could also reflect that parasites recruit and utilize host signalling factors instead[327]. Apart from the identification of a gene family structurally resembling TGF-β receptors, there are other indications supporting that TGF-β signalling could occur in *Plasmodium*. Firstly, functional polymorphism in both promoter and coding regions of the otherwise highly conserved human TGF-β suggest a link with malaria. Secondly, TGF-β production by spleen cells and levels of circulating TGF-β are constitutive in mice infected with non-lethal *Plasmodium* strains, whereas they drop considerably upon infection with lethal parasite lines[328], giving further support for a link between TGF-β and the immunological balance in malaria infection (Ref. [329] for review). The limited strength of the protein-protein interactions involved in TGF-β signalling makes this pathway a suitable target for drug or vaccine interventions since competitive binding may be achieved relatively easily[325].

As mentioned above, only a single *tstk* orthologue is present in all other *Plasmodium* species analysed, with the exception of the chimpanzee parasite, *P. reichenowi*. Phylogenetic analyses revealed that the syntenic copy of *P. falciparum* (*pftstk0*) is the most conserved member of this gene family (Chapter 5). Attempts to knock out the *tstk* gene of *P. berghei* by targeted gene disruption were unsuccessful indicating that this *tstk* gene is essential for asexual blood-stage development (T.W.A.K. and A.P.W., unpublished data).

TGF-β could stimulate the expression and switching of genes involved in antigenic variation. It will therefore be interesting to test the effects of increased TGF-β levels and antagonists of the TGF-β signalling pathway, such as the immunophilin FKBP12[330], cystatin C[331] and the small synthetic compound SB-431542[332,333] on the expression profiles and switching rates of *var*, *rif* and *stevor* and transport of the proteins they encode in different *P. falciparum* strains. To get a better understanding of the “ancestral” function of TSTK, it will be interesting to test the effects of administration of exogenous TGF-β to mice as well as deprivation of activated TGF-β by injection of recombinant latency associated protein (LAP)[334] or other TGF-β signalling antagonists prior or during infection with virulent and a-virulent strains of *P. berghei* on the course of infection. Direct BLAST analysis did not identify SMAD-related proteins in *P. falciparum* but continued searching for less obvious structural homologues based on alternative computational approaches,

such as hidden Markov model (HMM) profiling[185], could prove fruitful as was previously shown[146]. As mentioned above, the parasite might even utilize host-derived signalling molecules or alternative signalling pathways like in the case of the MAP kinase pathway. Using tags suitable for affinity purification will help identify such and other proteins the PfTSTKs might form complexes with.

Analysis of gametocyte-specific genes that are conserved between *P. falciparum* and the RMPs

The global studies on the conservation of genome organization and gene content reported in this thesis started in our laboratory on a small scale by the investigation of the organization of Pbchr5[72]. The focus on Pbchr5 was the result of the possible existence of a link between the organization of this chromosome and sexual development. It had been found that several genes specifically expressed during sexual development were located on Pbchr5 and that large-scale deletions in the subtelomeric regions were associated with the loss of the capacity of sexual differentiation, which might point to clustering and coordinate expression of sex-specific genes.

Although both the small-scale studies and subsequent global analyses did not provide evidence for the existence of large clusters of coordinately expressed sex-specific genes, these studies demonstrated that many sex-specific genes and their genomic organization are highly conserved between the RMPs and *P. falciparum* despite the significant differences in the morphology and duration of development of the gametocytes, which are the precursor cells of the gametes. Examples are the high level of conservation of the organization between *P. falciparum* and the RMPs of several sex-specific genes in the B9 locus[60] and the *6-Cys* superfamily, encoding proteins involved in fertilization[88]. In addition, recent global analyses of the proteomes of male and female gametocytes showed that >99% of the male- and female-specific proteins of *P. berghei* had orthologues in *P. falciparum*[154]. This high similarity of the organization and expression of sex-specific genes strengthens the use of RMP models to study the biology of sexual development and for the characterization of sex-specific antigens that may be used as targets for transmission-blocking vaccines (Ref. [335] for review) with relevance for human malaria.

Reverse genetics is a powerful approach that in malaria research is used to specifically alter the parasite genome to explore its biology and gain new insights into gene function and expression. In a post-genomic setting, it is one of the principle technologies that will be applied to increase our understanding of parasite biology with the potential of a full genome sequence. For example, it has been used to investigate the function in both RMPs and *P. falciparum* of P48/45, a transmission-blocking vaccine candidate[132]. Disruption of *p48/45* severly affected male gamete fertility, greatly reducing zygote formation and transmission to mosquitoes, demonstrating the conserved and essential role of the gamete surface protein P48/45 in fertilization of both *P. falciparum* and the RMPs.

For this thesis, we initiated studies to characterize the genes in the B9 locus located on Pbchr5, of which three are specifically expressed in gametocytes, and *α-tubulin II*, which is likewise highly expressed in gametocytes and located on Pbchr5. A second *α-tubulin* gene located on chromsome 4, *α-tubulin I*, was

analysed alongside. The studies on the *α-tubulins* are reported in Chapter 6 but since the work on the B9 genes has not been published yet, we will give some more details on these studies below.

Genes expressed in gametocytes: genes located in the B9 locus and *α-tubulin II*

The genomic organization of a 13.6-kb, complex, and gene-dense region containing three gametocyte-specific genes, termed the B9 locus, has been characterized previously[60]. The B9 locus provides an excellent example of the extreme level of conservation within the SBs and contains the gene encoding orotidine 5'-monophosphate decarboxylase (*omp-dc*) and five open reading frames (ORF1-5), encoding proteins of unknown function that are conserved between different *Plasmodium* species. These shared no homology with other prokaryote or eukaryote proteins, except for ORF2 that shows homology (E-value = $3.8e^{-12}$) to the human mitotic/meiotic spindle checkpoint protein (MAD2). The adjacent genes, transcribed from complementary strands, overlap in their untranslated regions (UTRs) and even introns and exons, resulting in a tight clustering and overlap of both regulatory and coding sequences. This tight clustering and overlapping of genes might hamper the analysis of individual genes using gene-disruption technologies.

We attempted to disrupt the following genes from the B9 locus by standard double cross-over technologies for the generation of gene knockouts in *P. berghei*[175]: two genes expressed during asexual blood stages, *omp-dc* and *orf2*, and three gametocyte-specific genes, *orf1*, *orf3*, and *orf4*. We were unable to select viable parasites with a disrupted *omp-dc* (three transfection experiments; L.H.M. van Lin and A.P.W., unpublished data), *orf1* (three transfection experiments; T.W.A.K. and A.P.W., unpublished data) and *orf2* (one transfection experiment; T.W.A.K. and A.P.W., unpublished data). This is perhaps not surprising for *omp-dc* and *orf2*, since both genes are transcribed during asexual blood-stage development, which may indicate that these genes are essential for blood-stage parasites. Given that ORF2 shows homology to a human mitotic/meiotic spindle checkpoint protein and the role of OMP-DC in DNA synthesis, it is conceivable that these are essential genes for asexual proliferation of the parasite. The failure to knock out *orf1* may be due to different reasons. Although the gene was shown to be highly upregulated in gametocytes, low but essential expression in asexual blood-stage parasites may have been missed in the analysis (analogous to *α-tubulin II*). Expression data available from the PlasmoDB website indicate that *pforf1* is expressed during trophozoite stages supporting asexual expression of this gene[92]. Another reason could be the organizational complexity and gene density of the region. The *orf1* gene has its 3'UTR including the last exon overlap with the 3'UTR of *omp-dc*, while the 5'UTR and first two exons of *orf1* overlap with the 5'UTR of *orf2*. Despite the careful choice of the integration sites, interference with as yet unknown, additional downstream polyadenylation sites of *omp-dc*, or upstream transcription initiation sites of *orf2* cannot be excluded. Finally, the failure to knock out *orf1* could be the result of the inefficiency of the transfection technology, which has not been repeated since the recent improvements in this technology (C.J.J., unpublished

data). We managed to obtain parasites with disrupted *orf3* (one experiment) and *orf4* (three experiments) but these parasites showed no distinct phenotype with regard to asexual blood-stage development, to production of gametocytes, and to the capacity of these parasites to fertilize and develop into ookinetes. In addition, both could be transmitted by mosquitoes, suggesting that they have no essential role during mosquito development and development in the liver of the vertebrate host.

Plasmodium species contain two genes that encode α-tubulins, *α-tubulin I* and *α-tubulin II*. It has been reported that *α-tubulin II* is highly expressed in gametocytes[282,283] and evidence has been reported that it plays an exclusive role in the formation of the axoneme of the male gamete[283]. In the light of the observation that clusters of gametocyte-specific genes were located on Pbchr5, it was interesting that we found that the gene encoding *α-tubulin II* was located on Pbchr5. We characterized the two *α-tubulin* genes in more detail with the aim to determine whether *P. berghei α-tubulin II* is a male-specific protein (Chapter 6). Investigation of transcription of male-specific genes might provide insight into male-specific promoter elements and lead to the development of tools to specifically express transgenes in male gametes.

This analysis of the *α-tubulin* genes in *P. berghei* again showed the conservation of gene content and organization between RMPs and *P. falciparum*, and the high transcription of *α-tubulin II* in male gametocytes and gametes could be confirmed[282,283]. However, additional low transcription of *α-tubulin II* was demonstrated in many other stages, such as asexual blood stages, female gametocytes, ookinetes, and oocysts. In addition, *α-tubulin II* could not be disrupted, whereas its C-terminal region could be modified with standard genetic modification technologies. This indicates that *α-tubulin II*, like *α-tubulin I*, is essential for asexual blood-stage development. One of the major defining characteristics of α-tubulin II of all *Plasmodium* species is the absence of three C-terminal amino acids (ADY), including a terminal tyrosine residue present in α-tubulin I. Unexpectedly, replacement of the C-terminal sequence of *α-tubulin II* by that of *α-tubulin I* generated a parasite line that had completely normal development of asexual blood stages, gametocytes and male gametes.

Notwithstanding the expression in asexual blood stages, our characterization of the promoter region of *α-tubulin II* enabled the generation of parasite lines that highly express green fluorescent protein (GFP), under the control of the *α-tubulin II* promoter, in male gametocytes enabling for the first time separation of pure male populations from both female gametocytes and asexual blood stages for proteome analysis[154], which shed new light on the sex-specific biology of malaria parasites. Furthermore, the knockout studies of expected gametocyte-specific genes presented in this thesis have highlighted several interesting aspects. Firstly, several of these gametocyte-specific genes could not be disrupted, possibly because low-level expression in asexual blood stages proved essential for their development as was clearly demonstrated in the case of *α-tubulin II*. Secondly, genes might be difficult to knock out when they are located in gene-dense regions like the B9 locus, where genes can overlap even in their coding sequences. Lastly, gene knockouts of a variety of genes were generated without resulting in an

obvious phenotype, these include *orf3* and *orf4* of the B9 locus but also *p25* and *p28*[169] and many other as yet unreported genes (C.J.J. and A.P.W., unpublished data). This may be the result of either redundancy of the genes or expression of the protein in later stages of the parasite life cycle, for example in sporozoites as shown for *crm3* and *crm4* (K.D.A., J. Thompson, and A.P.W., unpublished data), and awaits further investigation.

Perspective

In an era of rapidly increasing amounts of sequenced genomes, additional post-genomic analyses are essential to explore the wealth of information provided by these genome sequences and gain increasing interest and importance. In the studies described in this thesis, comparative genomics was used to investigate similarities and differences between the organization and gene content of the *P. falciparum* and RMPgenomes. First, our studies showed the feasibility and power of a composite genome approach, which uses partial genome sequences of three closely related RMP species to construct one cRMP genome.

Following the automated alignment of the single RMP contigs to the *P. falciparum* genome, the generation of the cRMP contigs was performed manually through combining overlapping single RMP contigs. In the future, the development of an algorithm to construct a composite DNA sequence from contigs of closely related species based on the alignment along a finished genome would significantly increase the speed of such an approach. The availability of two assembled genomes of closely related species will be beneficial for the prediction of coding regions, especially for genes that are difficult to predict, such as multi-exon genes. This approach can only be successful if the genomes under analysis have a low rate of recombination that is the case for the core regions of *Plasmodium*. Indeed, the approach failed to assemble the subtelomeric regions of the RMP genomes, which are thought to be highly recombinogenic.

The comparison of the cRMP genome with the human malaria genome demonstrated a high degree of conservation of gene content and organization, which strengthens the use of RMP models in future post-genomic research to investigate the biology of malaria parasites and to identify and characterize drug and vaccine targets with relevance for human parasites. In addition, in showing the similarities between the genomes of the RMPs and *P. falciparum*, our studies also revealed the differences, in particular the organization of species-specific genes. Further study of these genes may reveal differences in the biology of different species that are the result of specific adaptations to the different hosts, since many of these genes appear to play a role in host-parasite interactions, such as invasion of erthrocytes and the interaction of infected erthrocytes with microvascular endothelial cells. In addition, further study of the organization of species-specific genes in genomes of other *Plasmodium* species may provide more insight into mechanisms underlying the generation of diversity.

The expected release of the whole-genome shotgun sequence of the human malaria parasite *P. vivax*[127] will provide the “third” complete *Plasmodium* genome sequence (*P. falciparum*, cRMP, and *P. vivax*). At least three genome organizations are required to derive the genome organization of the MRCA[264] and one may hope that the availability of the genome organizations of *P. falciparum*,

P. vivax and the RMPs will enable the deduction of a genome organization of this "ancient malaria". We have investigated whether the SBPs between *P. falciparum* and the RMPs also exist in the available genome sequence of the non-human primate malaria parasite *P. knowlesi* that is closely related to *P. vivax*. We found preliminary evidence that this species has a (large-scale) genome organization that resembles more the RMP genome, which may suggest that also the genome of *P. vivax* is more similar to the RMP genome than to that of *P. falciparum*. It will be very interesting to see whether *P. vivax*, like the RMPs, lacks species-specific genes at most SBPs. If this were the case, it would point towards fundamental differences between *P. falciparum* and other mammalian malaria parasites in the generation of species-specific gene content. A complete *P. vivax* genome and increasing sequence information of the RMPs will also reveal the possible presence of indels containing species-specific genes, for example indels containing members of the *pir* superfamily that is present in both *P. vivax* and the RMPs but absent from *P. falciparum*. Interestingly, exhaustive analyses of yeast genomes indicate that SBPs and gross chromosomal rearrangements are not a driving evolutionary force for speciation and the generation of species-specific gene content[71].

More distantly related apicomplexan species such as *Toxoplasma gondii*, *Cryptosporidium parvum*, *Cryptosporidium hominis*, *Theileria parva*, *Theileria annulata*, *Babesia bovis*, and *Eimeria tenella* may provide additional information on chromosome evolution in these parasites. Possible traits such as the location of species-specific genes in subtelomeric regions, at SBPs or in indels interrupting synteny may be found, especially when comparison are made within the same genus (for example, *C. parvum*-*C. hominis* or *T. parva*-*T. annulata*). Such analyses could also improve the identification of rapidly evolving genes.

In conclusion, continued investigation of the species-specific genes and gene families identified in this study and future comparative analyses, might provide a better insight into the specific adaptations to the different host cells. In addition, the high conservation of gene content and organization of the genomes of the RMPs and *P. falciparum* emphasize the value of RMPs for further post-genomic analyses to identify and characterize new drug and vaccine candidates.

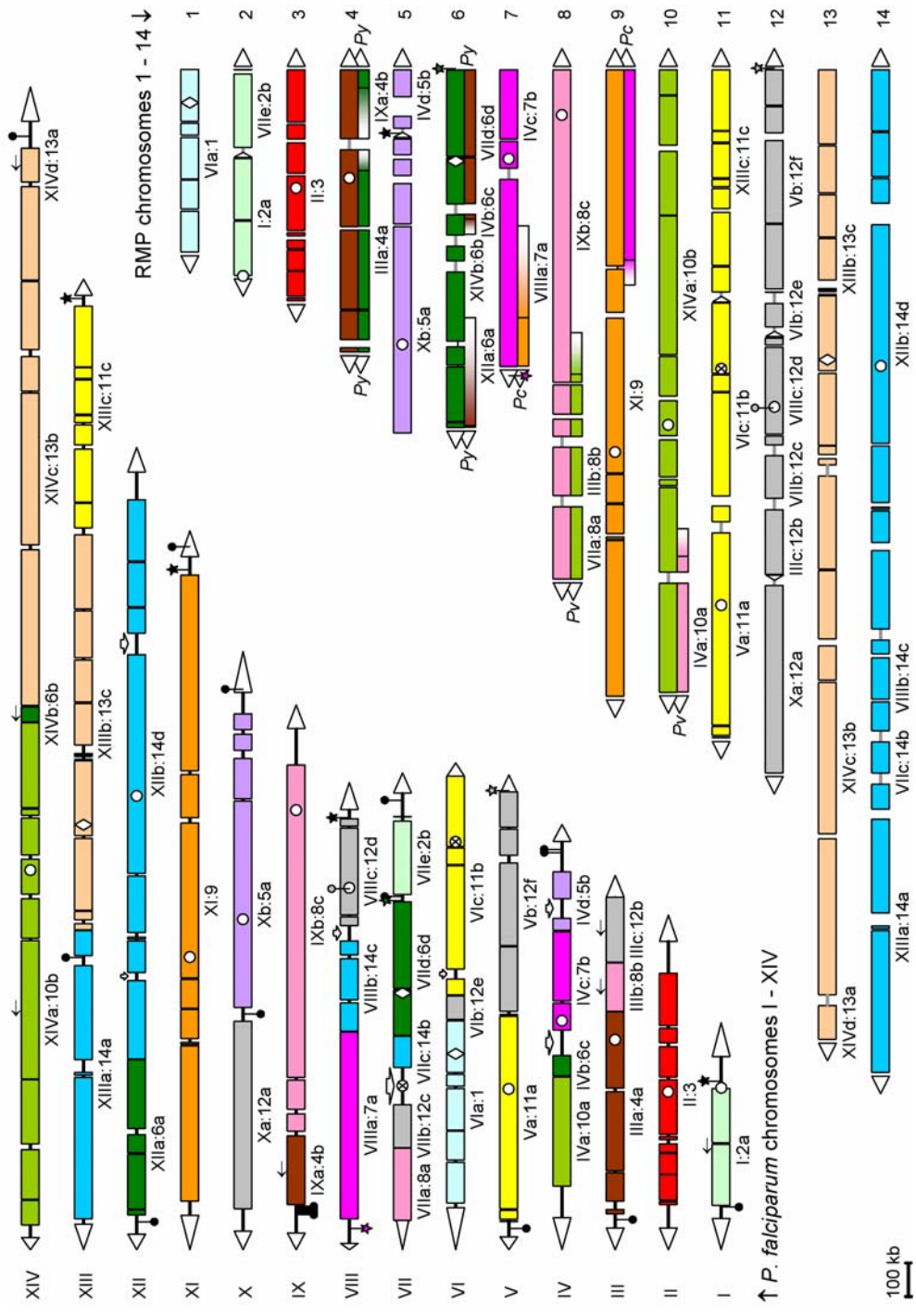
Appendices

Appendix 1. A whole-genome synteny map of *P. falciparum* and three RMPs (page 133)

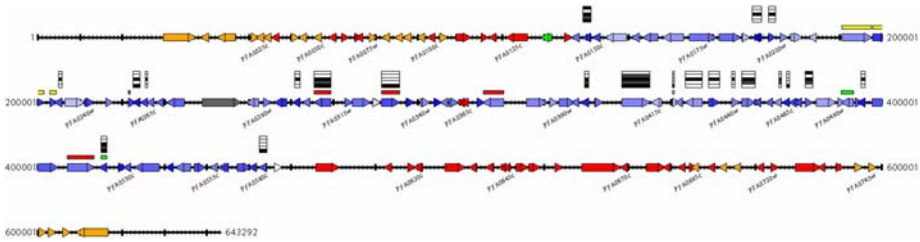
Synteny map of the core regions of all chromosomes of *P. falciparum* (left) and the RMPs (right), showing the 36 SBs, 22 SBPs, 14 CAT regions, *P. falciparum*-specific indels, and translocations in the RMP chromosomes. The 36 SBs, coloured according to their chromosomal location in the cRMP genome, are named with a Roman and an Arabic number referring to the corresponding chromosome location in *P. falciparum* and the cRMP genome, respectively. Letters give the order in which the SBs are connected. Small arrows indicate the inverted orientation of a SB in *P. falciparum* relative to the cRMP genome. Indels containing *P. falciparum*-specific intrasyntenic genes are indicated through interruption of the coloured SBs. *P. falciparum* telomeres are shown as white arrowheads (▷). SBs forming the cRMP chromosomes are linked by grey lines. In the cRMP genomes, the 23 coinciding subtelomeric linked ends are shown as white arrowheads (▷) and the five *P. falciparum* subtelomeric ends that are chromosome-internal in the cRMP chromosomes are indicated by small white arrowheads (◊). The 11 syntenic *P. falciparum* CAT regions[222] are shown as white circles (○), two inconsistent CAT regions as white circles with a cross (⊗), and three newly recognized CAT regions as white diamonds (◇). Chromosome-internal *var* clusters are shown as white arrows (⇦); stars and circles on sticks indicate *rrna* gene units (★) and *tstk* genes (●); black stars and circles represent non-syntenic genes; while syntenic genes (three *rrna* gene units and one *tstk* gene) are coloured according to their chromosomal location in the RMPs. Bars under the cRMP chromosomes represent the differences in the organization of the SBs of *P. yoelii*, *P. chabaudi*, and *P. vinckei* as a result of translocations. Colours indicate the cRMP chromosome with which recombination has taken place, while colour gradients represent the ill-defined regions of the translocation breakpoints.

Appendix 2. Schematic maps of *P. falciparum* chromosomes 1-14 (pages 134-139)

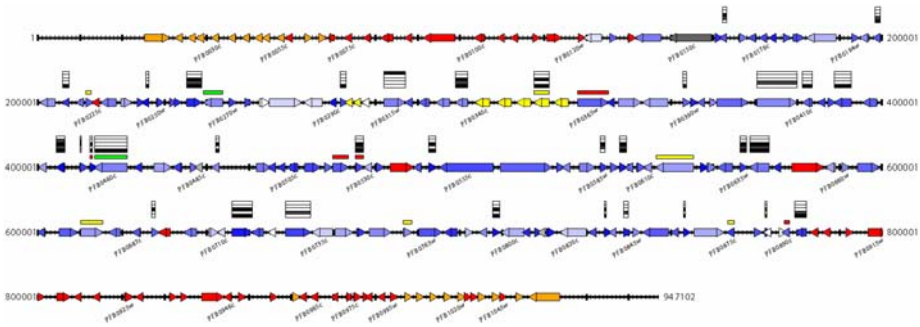
Arrowheads and boxes represent genes and their orientation on the DNA molecule. Thin and thick vertical lines represent 1-kb and 10-kb intervals, respectively. *P. falciparum* genes with orthologues in the RMP genomes are marked in shades of blue according to their degree of similarity, from light blue (indicating 40% identity) through dark blue (indicating 100% identity); white genes show <40% identity to their closest ortholog. Weak orthologues not detected by reciprocal BLAST analyses are indicated in dark grey and in light grey if the gene is absent in all RMP genomes. *P. falciparum* genes with no detectable orthologue are classified as follows: green, *pftstk* family; orange, *var*, *rif*, and *stevor* families; yellow, centrally located expanded gene families shared with the RMP; red, all other *P. falciparum* orphan genes. A full list of these genes and their classification can be found in SOM Table S3. Shaded areas of the map indicate the boundaries of the conserved chromosome core. Transcriptome and proteome data are marked above each gene where available. Transcripts that are up-regulated in asexual and gametocyte stages are shown as red or green horizontal lines, respectively; yellow lines denote genes that are up-regulated in both stages. Protein expression data are indicated by use of a bar code in which shading of each level indicates the following: top bar, sporozoite; second bar, oocyst; third bar, ookinete; fourth bar, gametocyte; and lowest bar, asexual stages. The identifier for every fifth gene (for example, PFI0025c) is indicated.


RMP chromosomes 1 - 14 ↓
↑ P. falciparum chromosomes I - XIV
100 kb
XIVa:10b
XIVb:6b
XIVc:13b
XIVd:13a
XIIIa:14a
XIIIb:13c
XIIIc:11c
XIIa:6a
XIIb:14d
XI:9
Xa:12a
Xb:5a
IXa:4b
IXb:8c
VIIIa:7a
VIIIb:14c
VIIIc:12d
VIIa:8a
VIIb:12c
VIIc:14b
VIId:6d
VIIe:2b
VIa:1
VIb:12e
VIc:11b
Va:11a
Vb:12f
IVa:10a
IVb:6c
IVc:7b
IVd:5b
IIIa:4a
IIIb:8b
IIIc:12b
II:3
I:2a
VIa:1
I:2a
VIIe:2b
II:3
IIIa:4a
IXa:4b
Xb:5a
IVd:5b
XIVb:6b
XIIa:6a
VIId:6d
IVb:6c
VIIIa:7a
IVc:7b
IXb:8c
IIIb:8b
XI:9
IVa:10b
XIVa:10b
Va:11a
Vc:11b
XIIIc:11c
IIIc:12b
VIIb:12c
VIIIc:12d
VIb:12e
Vb:12f
XIVc:13b
XIIIb:13c
XIVd:13a
XIIIa:14a
VIIc:14b
VIIIb:14c
XIIb:14d
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Py
Pc
Pv

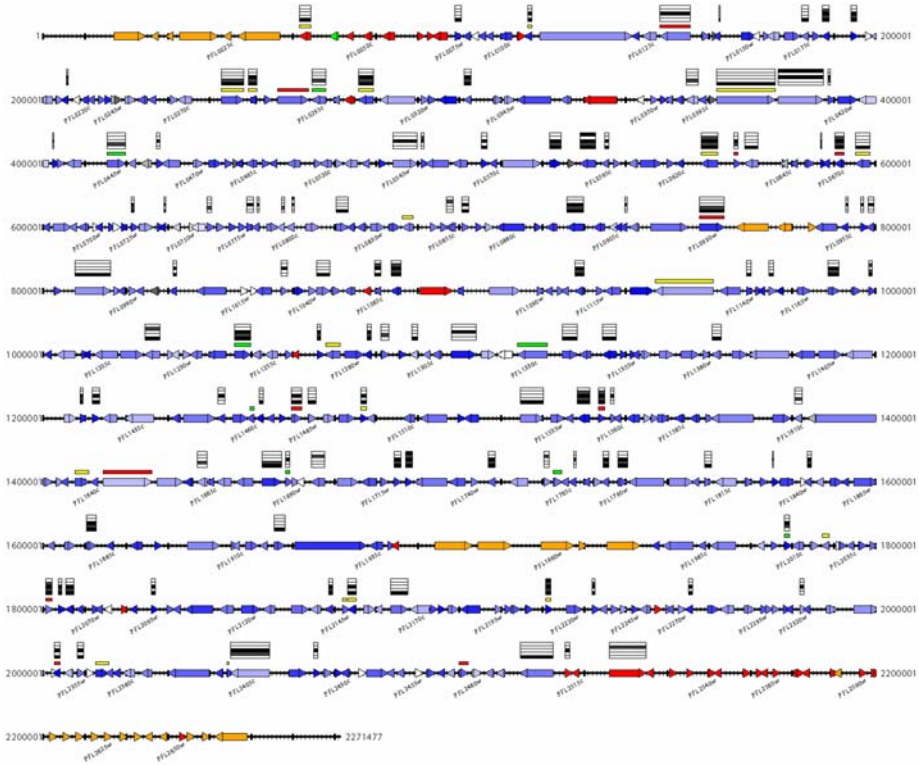
A



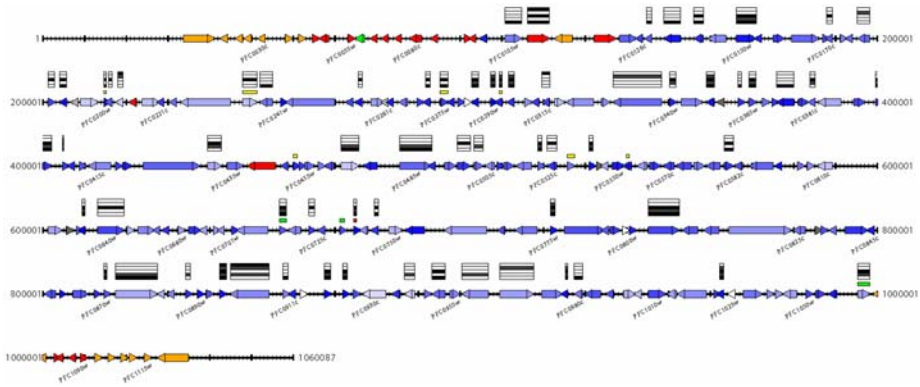
***P. falciparum* chromosome 1**



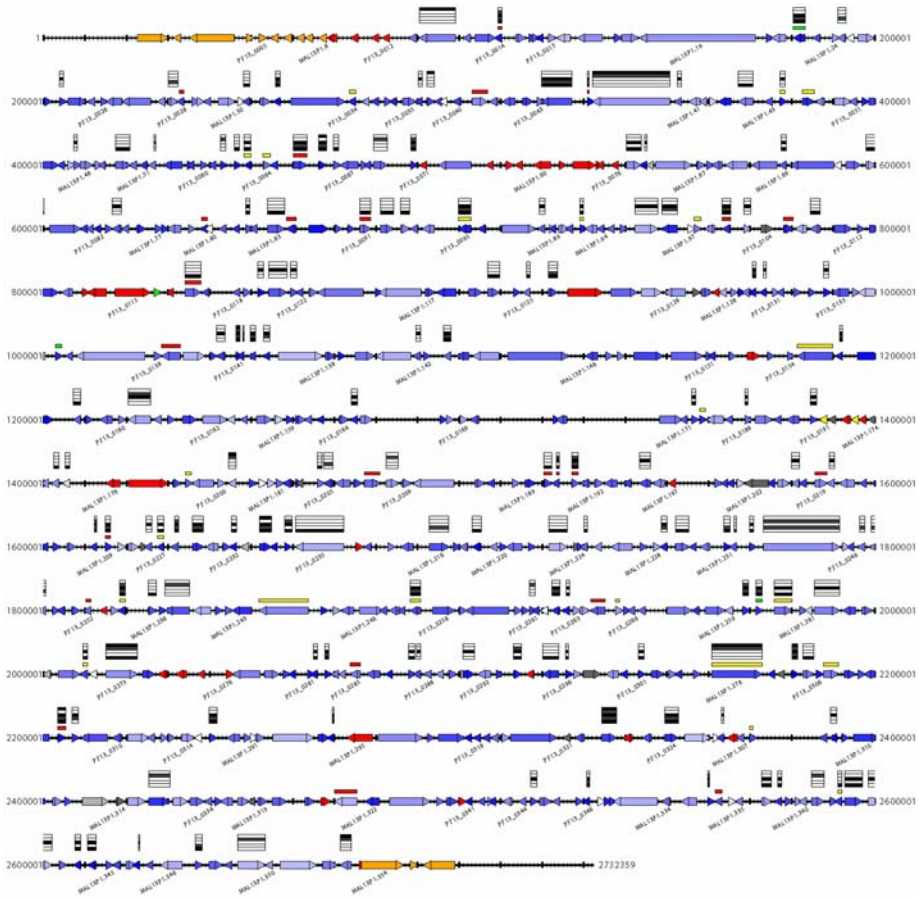
***P. falciparum* chromosome 2**



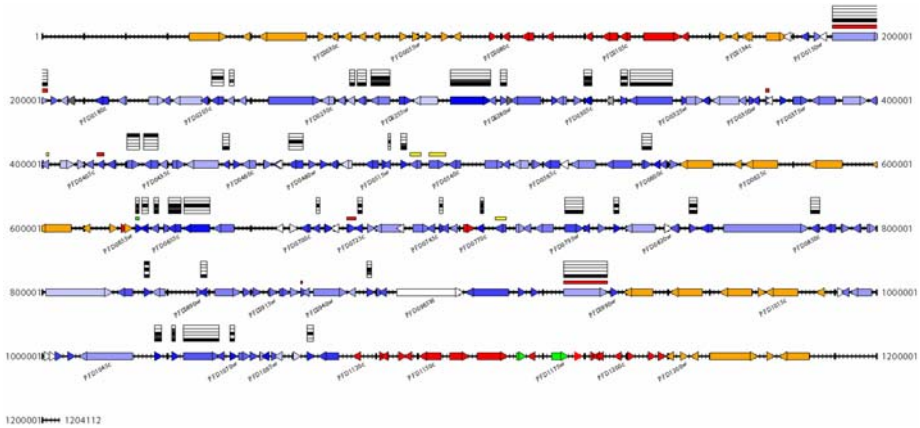
***P. falciparum* chromosome 12**



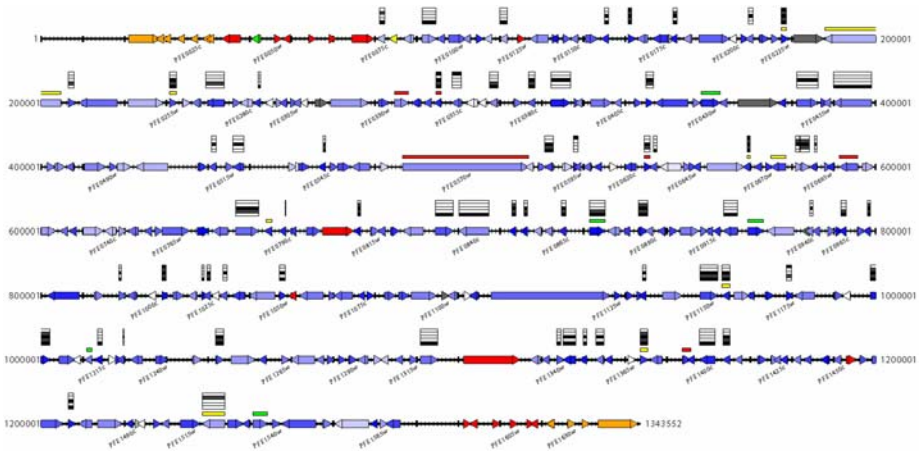
***P. falciparum* chromosome 3**



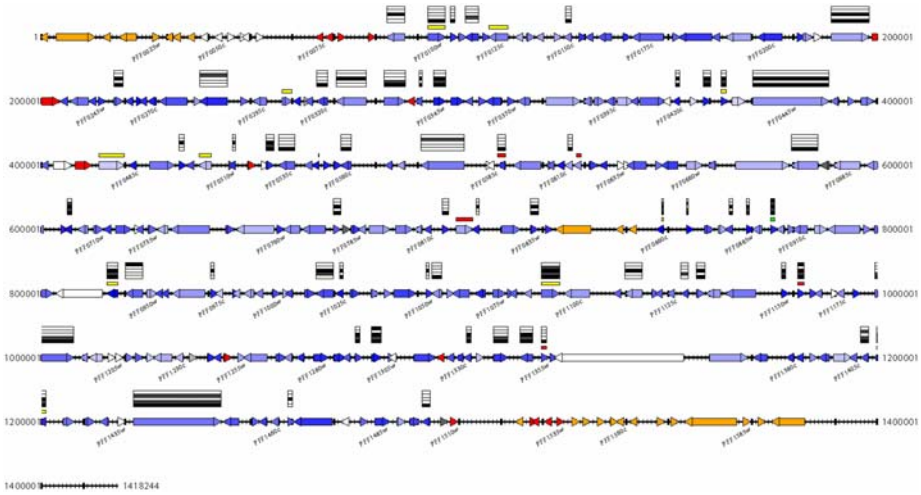
***P. falciparum* chromosome 13**



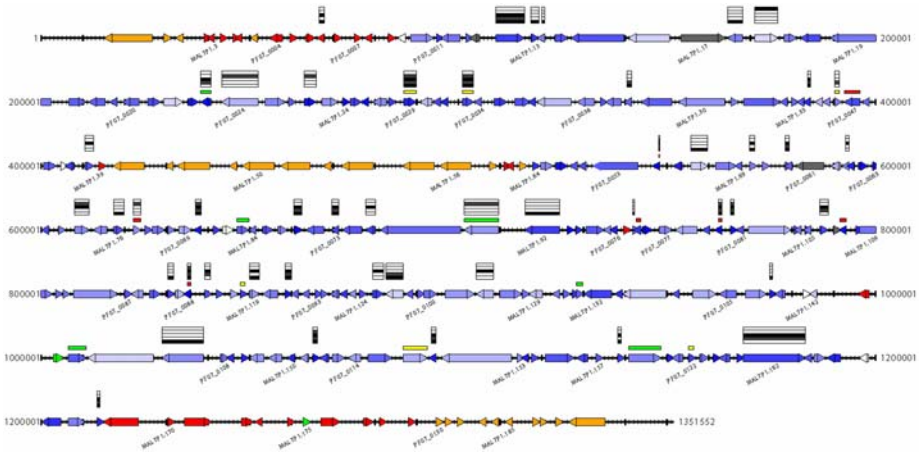
***P. falciparum* chromosome 4**



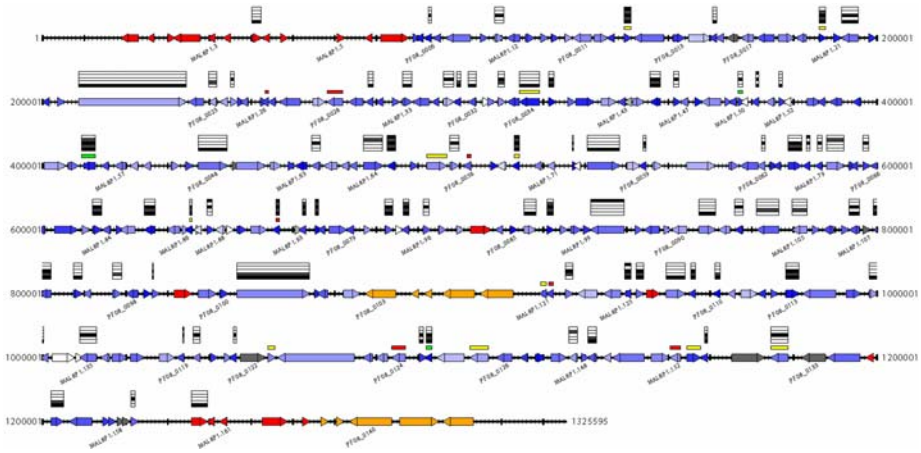
***P. falciparum* chromosome 5**



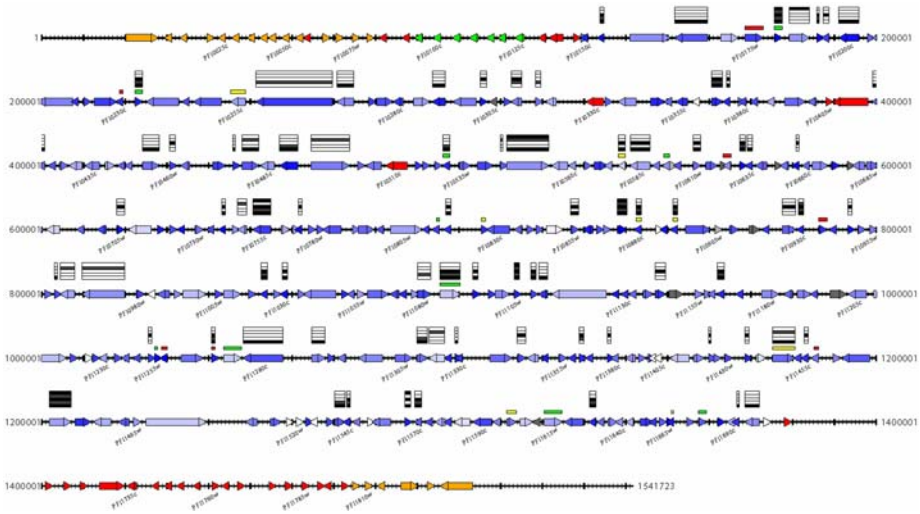
***P. falciparum* chromosome 6**



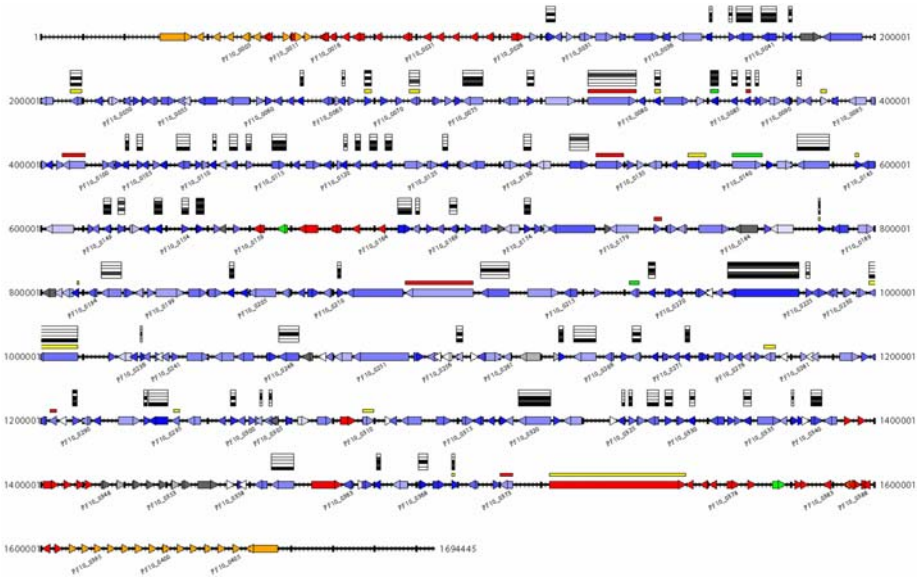
***P. falciparum* chromosome 7**



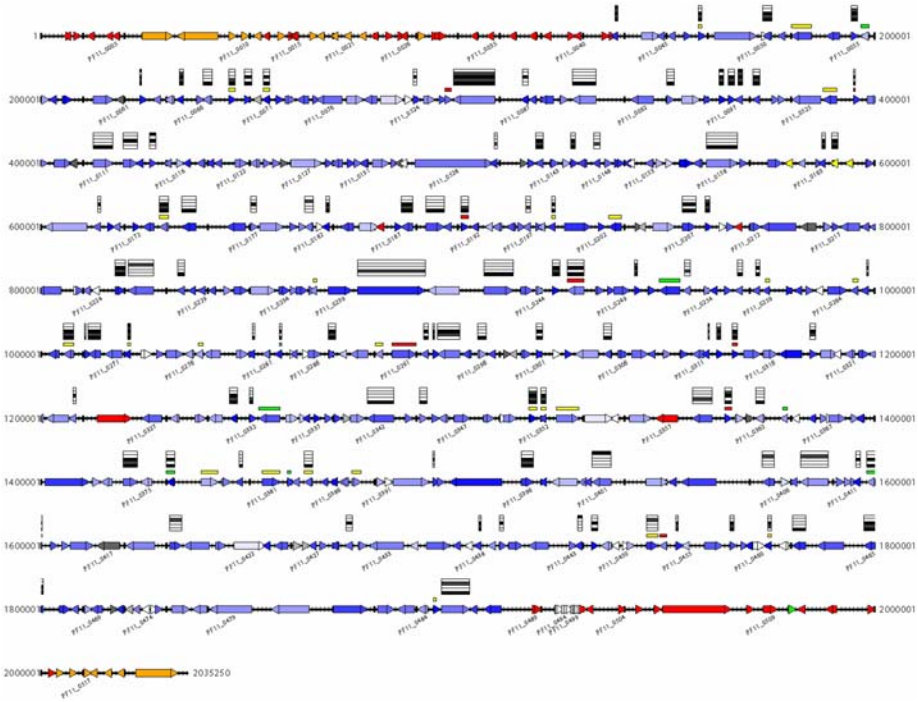
***P. falciparum* chromosome 8**



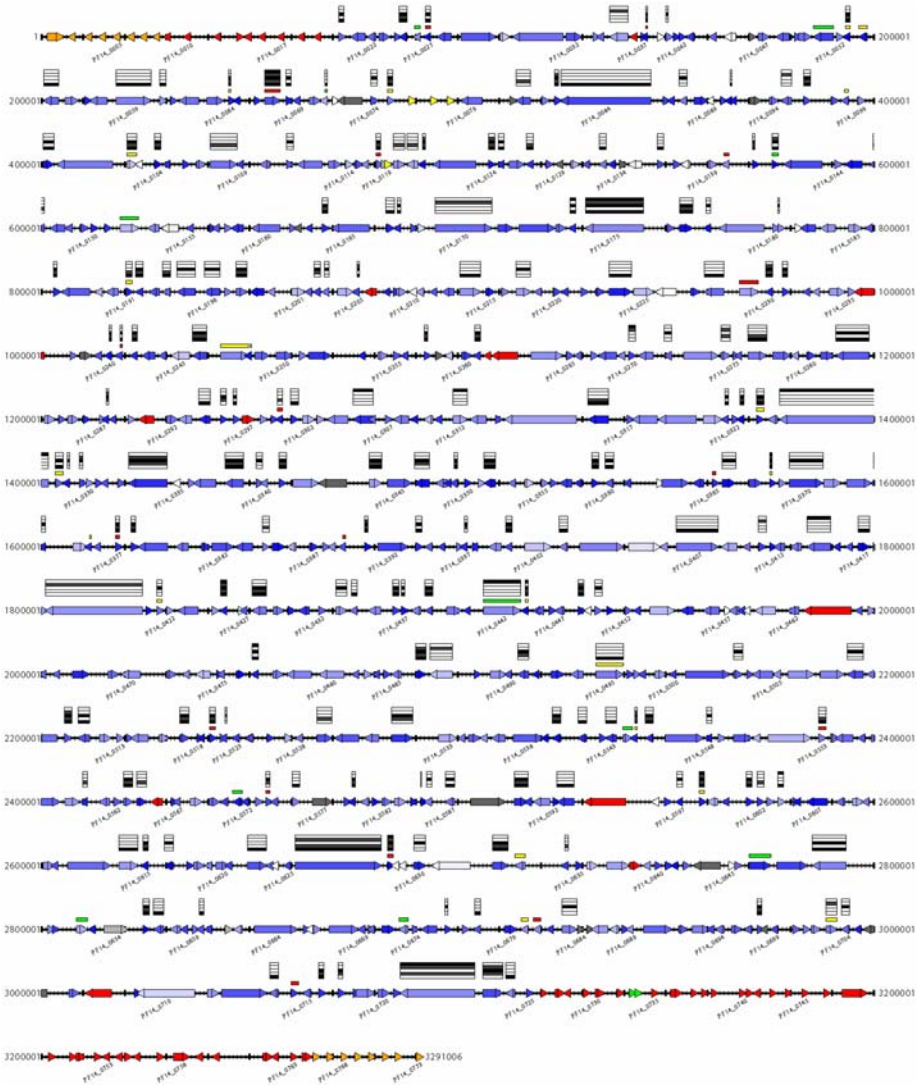
***P. falciparum* chromosome 9**



***P. falciparum* chromosome 10**



***P. falciparum* chromosome 11**



***P. falciparum* chromosome 14**

Appendix 3: A detailed set of genome summary statistics.

	P. berghei	***P. chabaudi***	***P. yoelii***	***P. falciparum***
Genome				
Size	17,996,878	16,866,661	23,125,449	22,853,764
No. of contigs	7,497	10,679	5,687	93
Av. contig size (bp)	2,400	1,580	4,066	213,586
Max. contig size (bp)	37,075	16,829	51,480	2,271,477
No. scaffolds	4,700	3,615	2,906	NA
Coverage[a]	4x	4x	5x	14.5x
Transcriptome				
No. of ESTs	12,277	none	17,014	21,371
Av. singleton length (bp)	541	NA	499	444
No. of EST clusters	1,939	NA	5,191	7,956
Av. cluster length (bp)	737	NA	513	424
Genome content				
GC content (%)	23.7	24.3	22.6	19.4
Total no. of predicted protein coding genes (* with orthologues)	12,208 (* 5,864)	14,970 (* 5,698)	5,878	5,268
Total no. full length genes	4,617	4,100	4,034	5,260
Total no. partial genes	7,591	10,870	1,844	8
No. of tRNAs	65	59	39	43
No. of 5.8s,18s,28s rRNA units	4	4	4	7
No. of 5s rRNA genes	3	3	3	3
Gene density	1,476	1,126	2,556	4,338
% coding	56.7	58.6	50.6	52.6
% genes with introns	40	50	54	52
Genes with EST hits	1,365	NA	3,120	3,301
Gene products detected by proteome	1,847	NA	NA	2,415
Mean no. exons per gene	1.74	1.57	2.0	2.4
GC content of exons (%)	24.7	25.6	24.8	23.7
Mean length of exons (bp)	422	585	641	944
GC content of introns (%)	22.8	21.1	21.1	13.5
Mean length of introns (bp)	136	132	209	179
Av. length of intergenic regions (bp)	736	1,063	853	1,654
GC content of intergenic regions (%)	21.2	23.5	20.7	13.4

[a] Average number of sequence reads per nucleotide.

Abbreviations: EST, expressed sequence tag; NA, not available.

Appendix 4: *Plasmodium* gene families.

	Gene family[a]	Pb	Pc	Py	Pf	Function	Protein Expression[b]
Common gene families	*235kDa rhoptry protein*	10	4	14	4	Reticulocyte binding protein	S,OCY
	HAD hydrolase	3	19	5	5	HAD hydrolase	B,G,OKN
	pst-a	6	41	12	10	Alpha beta hydrolase	G,OKN
	rhoph1/clag	3	3	3	4	Cytoadherence linked asexual protein	S,B,GAM
	etramp	7	3	11	11	Early transcribed membrane protein	B,G,GAM
RMP gene families	*yir/cir/bir*	180	138	838	0	Variant erythrocyte surface antigen?	S,B,G,OKN,OCY
	pyst-a	45	108	168	1	Unknown	S,B
	pyst-b	34	10	57	0	Unknown	S,B
	pyst-c	4	5	21	0	Unknown	ND
	pyst-d	1	0	17	0	Unknown	ND
	pcst-f	2	10	1	0	Unknown	B,G,OKN
	pcst-g	11	75	7	0	Unknown	S
	pcst-h	6	5	4	1	Unknown	ND
Pf gene families	*var*	0	0	0	59	Variant erythrocyte surface antigen (PfEMP1)	S,B,GAM
	rif	0	0	0	149[c]	Variant erythrocyte surface antigen	S,B,GAM
	stevor	0	0	0	28[c]	Variant erythrocyte surface antigen	S,B,GAM
	pf-fam-a	0	0	0	32	DNAJ domain protein - Linked to MC?	S,B,GAM
	pf-fam-b	0	0	0	13	Unknown	S,B
	pf-fam-c (*tstk*)	1	1	1	21	Receptor-associated protein kinase	S,B,GAM
	pf-fam-d	0	0	0	16	Unknown	ND
	pf-fam-e	0	0	0	6	Unknown	S,B,G,GAM
	pf-fam-f	0	0	0	10	Unknown	S,B,G
	pf-fam-g (*pfmc-2tm*)	0	0	0	7	Linked to MC	S,B,G,GAM
	pf-fam-h	2	0	2	19	Unknown	S,B,G,GAM
	pf-fam-i	3	2	3	11	Acyl-CoA synthetase	S,B,G,GAM

[a] Families were identified as described. Some but not all of these correspond to subtelomeric gene families that have previously been identified[51,235] (Chapter 3). Gene family designations are taken as follows: from Carlton *et al.*[51] (*py*) (Chapter 3), Fischer *et al.*[235] (*pc*) and Gardner *et al.*[42] (*pf*). *pf-fam-a* and *pf-fam-h* share six genes due to shared domains. *pf-fam-g* has recently been described using the name *pfmc-2tm*[146].
[b] *P. falciparum* protein expression data are from Lasonder *et al.*[12] and Florens *et al.*[11].
[c] Numbers of *var*, *rif* and *stevor* genes in *P. falciparum* are taken from Gardner *et al.*[42].

Abbreviations: Pb, *P. berghei*; Pc, *P. chabaudi*; Py, *P. yoelii*; Pf, *P. falciparum*; RMP, rodent malaria parasite; MC, Maurer's clefts; S, sporozoite; B, asexual blood stage; G, gametocyte; GAM, gamete; OKN, ookinete; OCY, oocyst; ND, not done.

Appendix 5: Centromere predictions[a].

Pfchr	Pf gene downstream	RMP gene downstream	RMP contig downstream	End of synteny	Pf gene upstream	RMP gene upstream	RMP contig upstream	Start of synteny	Distance between flanking syntenic genes	Method of linkage	Size of PCR (bp)	RMP chr
1	PFA0585w	PY01445	MALPY00384	457,239	PFA0590w	NI	not syntenic	-	-	SLE	-	2
2	PFB0490c	PB000558.01.0	Pb.c001904474.Contig1	443,244	PFB0495w	PY05091	MALPY01600	449,666	6,422	PCR	4,000	3
3	PFC0610c	PY02287	MALPY00628	587,740	PFC0615w	PY07280	MALPY02636	601,130	13,390	None	-	4
4	PFD0690c	PY00550	MALPY00150	641,902	PFD0695w	PY02205	MALPY00605	659,376	17,474	None	-	7
5	chr5-tRNA-Leu-1	tRNA-Leu-1	MALPY02997	452,699	PFE0530w	PY04900	MALPY01520	458,119	5,420	PCR	4,000	11
6	MAL6P1.152	PY04637	MALPY01429	1,151,208	MAL6P1.151	PY02706	MALPY00748	1,152,410	1,202	None	-	11
6[b]	**MAL6P1.116**	**PY03119**	**MALPY00884**	**526,726**	**MAL6P1.309**	**PY02914**	**MALPY00821**	**528,552**	**1,826**	**None**	**-**	**1**
7	PF07_0050	NI	not syntenic	411,168	MAL7P1.52	NI	not syntenic	517,160	105,992	None	-	SBP
7[b]	**MAL7P1.91**	**PY02993**	**MALPY00846**	**708,966**	**MAL7P1.92**	**PY05845**	**MALPY01919**	**714,099**	**5,133**	**None**	**-**	**6**
8	MAL8P1.136	PY05152	MALPY01625	1,020,763	PF08_0118	PY05558	MALPY01784	1,023,858	3,095	None	-	12
9	PFI1500w	PY06205	MALPY02081	1,240,525	PFI1505c	PY07561	MALPY02813	1,242,181	1,656	PCR	2,200	8
10	PF10_0216	PY07622	MALPY02840	934,786	PF10_0217	PY01659	MALPY00449	939,652	4,866	PCR	3,500	5
11	PF11_0226	PY06052	MALPY02015	826,272	PF11_0227	PY06967	MALPY02459	831,942	5,670	PCR	2,000	9
12	PFL1505c	PY05255	MALPY01663	1,282,467	PFL1510c	PY05931	MALPY01956	1,285,021	2,554	PCR	3,500	14
13[b]	**MAL13P1.151**	**PY04447**	**MALPY01358**	**1,200,716**	**PF13_0157**	**PY01467**	**MALPY00394**	**1,205,374**	**4,568**	**PCR**	**4,000**	**13**
14	PF14_0252	PY05795	MALPY01896	1,068,157	PF14_0253	PY02120	MALPY00583	1,075,409	7,252	PCR	9,000	10

[a] Predictions of the location of centromeres in the genome of *P. falciparum* and RMP based on comparison of the synteny blocks. For all previous centromere predictions[222], the flanking *P. falciparum* genes and their positions on the chromosome as well as the RMP genes and contigs both up- and downstream of the centromere are given.
[b] Newly predicted centromeres are shown in bold.

Abbreviations: Pf, *P. falciparum*; chr, chromosome; RMP, rodent malaria parasites; NI, not identified; SLE, subtelomere linked end; SBP, synteny breakpoint.

Appendix 6: Characteristics of the five RMP-specific intersyntenic genes.

A list of five RMP-specific genes located in indels between SBs (intersyntenic genes). Proteins with a TM domain in the first 100 amino acids (TM-N) are considered proteins potentially targeted to the surface membrane of the parasite or erythrocyte (one of five genes marked with an asterisk).

RMP chr	Pb gene	Pb contig	Pc gene	Pc contig	Py gene	Py contig	Product	Size (aa)	SP	AP	TM-N	TM
5	*c-rrna* gene unit	Contig5161	*c-rrna* gene unit	Contig4882	*c-rrna* gene unit	MALPY00527	28S, 5.8S, 18S rRNA	-	-	-	-	-
6	PB101816.00.0	Pb.Contig31	PC404234.00.0	Pch1282h06.q1k UNCHAB.0.29007	not annotated	MALPY02445	hypothetical protein	203	-	X	-	-
10	not annotated	Pb.Contig126	not annotated	Pch1539b06.q1k UNCHAB.0.23105	PY02305	MALPY00633	hypothetical protein	46	-	-	-	-
12	NI	Pb.c00170279 3.Contig1	NI	no contig	PY01538	MALPY00409	hypothetical protein	73	-	-	-	-
12 *	PB108348.00.0 - PB000847.03.0	Pb.c00030147 1.Contig1.0/1	not annotated	Pc.Contig10	PY01786	MALPY00483	putative dentin phosphoryn	965	-	-	X	X

Abbreviations: RMP, rodent malaria parasites; SB, synteny block; chr, chromosome; Pb, *P. berghei*; Pc, *P. chabaudi*; Py, *P. yoelii*; aa, amino acids; SP, predicted signal peptide; AP, predicted apicoplast transit peptide; TM, transmembrane domain; TM-N, N-terminal TM-domain; NI, not identified.

Appendix 7: Characteristics of the 53 *P. falciparum*-specific intersyntenic genes.

A list of 53 *P. falciparum*-specific genes located in indels between SBs (intersyntenic genes) including all intersyntenic pseudogenes. Groups of genes forming a cluster (eight indels) are separated by a black line Proteins with a TM domain in the first 100 amino acids (TM-N) as well as *var* and *rif* genes are considered proteins potentially targeted to the surface membrane of the parasite or erythrocyte (37 of 42 genes and 11 of 11 pseudogenes marked with an asterisk, genes with a PEXEL/VTS[116,117] are marked with two asterisks).

Pf gene	Product	*pf-fam*[a]	Size indel (bp)	trans.	exp.	SP	AP	TM-N	TM
PFD0615c **	erythrocyte membrane protein 1 (PfEMP1)	*var*	70,195						
chr4.phat_143/ chr4.glm_148 *	var internal cluster associated repeat gene 4a	*vicar*	70,195	-	-	**X**	**-**	**X**	**X**
PFD0620c **	rifin--related protein	*rif*	70,195						
PFD0625c **	erythrocyte membrane protein 1 (PfEMP1)	*var*	70,195						
- *	vicar pseudogene	*vicar*	70,195						
PFD0630c **	erythrocyte membrane protein 1 (PfEMP1)	*var*	70,195						
- *	vicar pseudogene	*vicar*	70,195						
PFD0635c **	erythrocyte membrane protein 1 (PfEMP1)	*var*	70,195						
PFD0640c **	rifin	*rif*	70,195						
PFD0645w **	rifin	*rif*	70,195						
PFD0650w	hypothetical protein	-	70,195	-	-	**-**	**-**	**-**	**-**
PFD0655w **	var-related pseudogene	*var*	70,195						
MAL7P1.39 *	var internal cluster associated repeat gene 7a	*vicar*	105,155	S,M	-	**X**	**-**	**X**	**X**
PF07_0048 **	erythrocyte membrane protein 1 (PfEMP1)	*var*	105,155						
chr7.phat_103/ chr7.glm_101 *	var internal cluster associated repeat gene 7b	*vicar*	105,155	-	-	**X**	**-**	**X**	**X**
MAL7P1.43 **	rif pseudogene	*rif*	105,155						
PF07_0049 **	erythrocyte membrane protein 1 (PfEMP1)	*var*	105,155						
chr7.phat_108/ chr7.glm_105 *	var internal cluster associated repeat gene 7c	*vicar*	105,155	-	-	**X**	**-**	**X**	**X**
MAL7P1.47 **	rif pseudogene	*rif*	105,155						
MAL7P1.50 **	erythrocyte membrane protein 1 (PfEMP1)	*var*	105,155						
PF07_0050 **	erythrocyte membrane protein 1 (PfEMP1)	*var*	105,155						
MAL7P1.52 **	rif pseudogene	*rif*	105,155						
PF07_0051 **	erythrocyte membrane protein 1 (PfEMP1)	*var*	105,155						
chr7.phat_114 *	vicar pseudogene	*vicar*	105,155						
MAL7P1.55 **	erythrocyte membrane protein 1 (PfEMP1)	*var*	105,155						
MAL7P1.56 **	erythrocyte membrane protein 1 (PfEMP1)	*var*	105,155						
MAL7P1.57 **	rifin	*rif*	105,155						
MAL7P1.58 *	hypothetical protein, conserved in Pf	*pfmc-2tm*	105,155	ET	-	**-**	**-**	**X**	**X**
MAL7P1.59	hypothetical protein, conserved in Pf	-	105,155	-	-	**-**	**-**	**-**	**-**

Pf gene	Product	*pf-fam*[a]	Size indel (bp)	trans.	exp.	SP	AP	TM-N	TM
MAL7P1.60 **	var fragment, pseudogene	*var*	105,155						
MAL7P1.62 **	var fragment, pseudogene	*var*	105,155						
MAL7P1.63 **	var fragment, pseudogene	*var*	105,155						
PF07_0107 *	hypothetical protein	-	13,799	M-LR	-	-	-	X	X
MAL7P1.144 **	pftstk7a	*tstk*	13,799	ER-LT,G	S	-	-	-	X
PF08_0103 **	erythrocyte membrane protein 1 (PfEMP1)	*var*	43,407						
- *	vicar pseudogene	*vicar*	43,407						
PF08_0104 **	rifin	*rif*	43,407						
PF08_0105 **	rifin	*rif*	43,407						
PF08_0106 **	erythrocyte membrane protein 1 (PfEMP1)	*var*	43,407						
PF08_0107 **	erythrocyte membrane protein 1 (PfEMP1)	*var*	43,407						
chr8.phat_232/ chr8.glm_253 *	var internal cluster associated repeat gene 8	*vicar*	43,407	-	-	X	-	X	X
PFI0330c	hypothetical protein	-	11,729	LR-LT	-	-	-	-	X
PF10_0159 **	glycophorin-binding protein 130 precursor	*gbp*	37,843	ER-LS	T	-	-	X	X
PF10_0160 **	3.8 protein - pftstk10a	*tstk*	37,843	ER-ES	T	-	-	X	X
PF10_0161	hypothetical protein	-	37,843	ER-LR	S	-	-	-	-
PF10_0162	hypothetical protein	-	37,843	ET,M	-	-	-	-	-
PF10_0163 **	hypothetical protein	-	37,843	LS-ER	-	X	X	X	X
PF10_0164 *	hypothetical protein	*etramp*	37,843	S,G	G	X	-	X	X
MAL13P1.268 *	hypothetical protein	-	18,712	M-ER	-	X	-	X	X
MAL13P1.269 *	hypothetical protein	-	18,712	B,G	T	-	-	X	X
PF13_0275 **	hypothetical protein	-	18,712	M-ET	T,G,M	-	-	X	X
PF13_0276 *	hypothetical protein	-	18,712	M-LR	-	-	-	X	X
PF14_0708 *	hypothetical protein	-	13,194	G	-	-	-	X	X

[a] References to the papers describing the gene families are as follows: *var*[80-82]; *vicar*, *var* internal cluster associated repeat gene[87] (Chapter 5); *rif*[84]; *pfmc-2tm*, *P. falciparum* Maurer's cleft localized two transmembrane domain protein[146]; *tstk*, transforming growth factor β (TGF-β) receptor-like serine/threonine protein kinase[87] (Chapter 5); *gbp*, glycophorin-binding protein[258,336,337]; *etramp*, early transcribed membrane protein[259].

Abbreviations:SB, synteny block; PEXEL, *Plasmodium* export element; VTS, vacuolar transport signal; Pf, *P. falciparum*; *pf-fam*, described *P. falciparum* gene family; SP, predicted signal peptide; AP, predicted apicoplast transit peptide; TM, transmembrane domain; TM-N, N-terminal TM-domain; trans., approximate stage of (highest) transcription for all genes except the *var*, *rif* and *vicar* genes based on transcriptome data[91,92]; M, merozoite; ER, early ring; LR, late ring; ET, early trophozoite; LT, late trophozoite; ES, early schizont; LS, late schizont; G, gametocyte; S, sporozoite; B, asexual blood stage; exp., expression of those genes based on proteome data[11,12]; M, merozoite; T, trophozoite; G, gametocyte; S, sporozoite.

Appendix 8: Characteristics of the 138 *P. falciparum*-specific intrasyntenic genes.

A list of 138 *P. falciparum*-specific genes located in indels within SBs (intrasyntenic genes) including all intrasyntenic pseudogenes. Groups of genes forming a cluster (82 indels) are separated by a black line. Proteins with a TM domain in the first 100 amino acids (TM-N) as well as *var* and *rif* genes are considered proteins potentially targeted to the surface membrane of the parasite or erythrocyte (78 of 126 genes and 11 of 12 pseudogenes marked with an asterisk, genes with a PEXEL/VTS[116,117] are marked with two asterisks).

Pf gene	Product	*pf-fam*[a] or homologue	Size indel (bp)	trans.	exp.	SP	AP	TM-N	TM
PFA0360c **	hypothetical protein	PFA0365c	2,246	S,G	-	X	X	X	X
PFA0365c *	hypothetical protein	PFA0360c	2,246	-	-	X	X	X	X
PFB0140w *	hypothetical protein	-	1,237	LT-LS	-	-	-	X	X
PFB0225c	hypothetical protein	-	1,526	G	S	-	-	-	-
PFB0300c *	merozoite surface protein 2 precursor	*msp*	2,805	ES-M	M	X	X	X	X
PFB0305c *	merozoite surface protein 5	*msp*	2,805	LS-M	-	X	-	X	X
PFB0340c *	cysteine protease, putative	*protease*	22,130	ES-LS	M	X	-	X	X
PFB0345c *	cysteine protease, putative	*protease*	22,130	LT-LS	S	X	-	X	X
PFB0350c *	cysteine protease, putative	*protease*	22,130	LT-LS	G,T	X	-	X	X
PFB0355c	cysteine protease, putative	*protease*	22,130	ES-LS,S	-	-	-	-	X
PFB0360c *	cysteine protease, putative	*protease*	22,130	ES-LS	-	X	-	X	X
PFB0540w	hypothetical protein	-	5,534	LT-ES	S	-	-	-	-
PFB0650w	hypothetical protein	-	7,502	LT-M	-	-	-	-	-
PFB0695c *	acyl-CoA synthetase	*acyl-CoA*	2,666	M-LR	-	X	X	X	X
PFC0110w *	cytoadherence linked asexual protein	*clag*	21,201	ES-M	-	X	-	X	X
PFC0115w **	pseudo var gene	*var*	21,201						
PFC0120w *	cytoadherence linked asexual protein	*clag*	21,201	ES-M	M	X	-	X	X
PFC0215c *	hypothetical protein	-	1,856	-	-	X	X	X	X
PFC0440c	helicase, putative	-	6,809	LR-ES	S,T	-	-	-	X
PFD0765w	RING finger protein, putative	-	2,627	M	M,S	-	-	-	X
PFD0995c **	erythrocyte membrane protein 1 (PfEMP1)	*var*	52,417						
*	vicar pseudogene	*vicar*	52,417						
PFD1000c **	erythrocyte membrane protein 1 (PfEMP1)	*var*	52,417						
*	vicar pseudogene	*vicar*	52,417						
PFD1005c **	erythrocyte membrane protein 1 (PfEMP1)	*var*	52,417						
PFD1010w **	rifin	*rif*	52,417						
PFD1015c **	erythrocyte membrane protein 1 (PfEMP1)	*var*	52,417						
chr4.phat_229 *	var internal cluster associated repeat gene 4b	*vicar*	52,417	-	-	X	-	X	X
PFD1020c **	rifin	*rif*	52,417						
PFD1025c **	var pseudogene	*var*	52,417						
PFE0080c *	rhoptry-associated protein 2	PFE0075c	1,196	ES-M	M	X	-	X	X
PFE0125w *	hypothetical protein	-	1,784	M,G	-	X	-	X	X
PFE0805w *	cation-transporting ATPase 1	-	7,375	LR-ES,G	S	-	-	X	X
PFE1055c	hypothetical protein	-	1,690	ER-ES,G	-	-	-	-	-
PFE1325w	hypothetical protein	-	13,574	-	-	-	-	-	X

Pf gene	Product	*pf-fam*[a] or homologue	Size indel (bp)	trans.	exp.	SP	AP	TM-N	TM
PFE1455w *	sugar transporter, putative	-	2,359	S	-	-	-	X	X
MAL6P1.49	DNA helicase, putative	-	5,493	ET	M,T	-	-	-	X
MAL6P1.71 *	hypothetical protein	-	899	-	-	X	-	X	X
MAL6P1.99	hypothetical protein	-	3,961	-	M	-	-	-	X
MAL6P1.108	calcium-dependent protein kinase	-	1,907	ET-LS,G	-	-	-	-	X
MAL6P1.252 **	erythrocyte membrane protein 1 (PfEMP1)	*var*	19,289						
MAL6P1.251 **	rifin	*rif*	19,289						
MAL6P1.250 **	rifin	*rif*	19,289						
MAL6P1.170	hypothetical protein	-	1,928	S	-	-	-	-	X
MAL6P1.156 *	troponin c-like protein, putative	-	1,248	-	-	-	-	X	X
MAL7P1.97 *	hypothetical protein	-	904	-	-	-	-	X	X
MAL7P1.167	hypothetical protein	-	8,321	LS	S	-	-	-	-
MAL8P1.97	hypothetical protein	-	4,830	S,G,M	-	-	-	-	X
MAL8P1.111	hypothetical protein	-	4,204	B,G	G	-	-	-	X
MAL8P1.126 *	serine protease, putative	-	3,139	S	-	X	-	X	X
MAL8P1.155 *	hypothetical protein	-	1,292	S	-	-	-	X	X
PFI0405w *	hypothetical protein	-	10,089	LT-LS	-	X	X	X	X
PFI0410c	hypothetical protein	-	10,089	ES-LS	T	-	-	-	X
PFI0510c *	hypothetical protein	-	5,093	LS	S	-	-	X	X
PF10_0309	hypothetical protein (helicase, putative)	-	3,616	S,B,G	-	-	-	-	X
PF10_0342 *	hypothetical protein	-	19,110	ES-M,G	-	X	-	X	X
PF10_0343 *	S-antigen	-	19,110	ES-M	-	X	-	X	X
PF10_0344 *	glutamate-rich protein	-	19,110	ES-M	G,S	X	X	X	X
PF10_0345 *	merozoite surface protein 3	*msp*	19,110	ES-M	M,T	X	-	X	X
PF10_0346 *	merozoite surface protein 6	*msp*	19,110	LS-M	M	-	-	X	X
PF10_0347 *	hypothetical protein	*msp*	19,110	ES-M	-	-	-	X	X
PF10_0362 *	DNA polymerase zeta catalytic subunit, putative	*msp*	7,204	LT-LS	-	-	-	X	X
PF11_0161 *	falcipain-2 precursor, putative	*protease*	1,448	ER-LT	T	X	X	X	X
PF11_0165 *	hypothetical protein	*protease*	5,073	LR-LT	T	X	X	X	X
PF11_0166	hypothetical protein	-	5,073	LS-LR	M,T	-	-	-	-
PF11_0186 *	hypothetical protein	-	756	B,G	S	-	-	X	X
PF11_0211 *	hypothetical protein, conserved	-	1,348	M-LR	-	-	-	X	X
PF11_0326	hypothetical protein	-	8,291	-	S	-	-	-	X
PF11_0357	hypothetical protein	-	5,168	G	M	-	-	-	-
PFL0105w	hypothetical protein	-	2,136	G	M,G,T	-	-	-	-
PFL0305c	hypothetical protein	-	2,422	G,S	G	-	-	-	-
PFL0360c	hypothetical protein	-	8,171	LT	-	-	-	-	X
PFL0935c **	erythrocyte membrane protein 1 (PfEMP1)	*var*	16,848						
PFL0940c **	var fragment, pseudogene	*var*	16,848						
PFL0945w **	var fragment, pseudogene	*var*	16,848						
PFL1060c **	hypothetical protein, conserved	-	1,709	-	-	X	X	X	X
PFL1075w	Apicomplexan AP2-integrase DNA-binding domain containing protein	*apiap2*	7,676	ET-LS,G	G,S	-	-	-	X

Pf gene	Product	*pf-fam*[a] or homologue	Size indel (bp)	trans.	exp.	SP	AP	TM-N	TM
PFL1265c	hypothetical protein	-	1,751	G	-	-	-	-	-
PFL1945c *	hypothetical protein	*etramp*	58,313	M	-	X	X	X	X
PFL1950w **	erythrocyte membrane protein 1 (PfEMP1)	*var*	58,313						
PFL1955w **	erythrocyte membrane protein 1 (PfEMP1)	*var*	58,313						
chr12.phat_410/ chr12.glm_457 *	var internal cluster associated repeat gene 12	*vicar*	58,313	-	-	X	-	X	X
PFL1960w **	erythrocyte membrane protein 1 (PfEMP1)	*var*	58,313						
PFL1965w **	rif pseudogene	*rif*	58,313						
*	vicar pseudogene	*vicar*	58,313						
PFL1970w **	var pseudogene	*var*	58,313						
*	vicar pseudogene	*vicar*	58,313						
PFL2085w	hypothetical protein	-	1,331	LT-M	S	-	-	-	-
PFL2255w	hypothetical protein	-	1,819	G,S	-	-	-	-	-
PF13_0071	hypothetical protein	-	1,688	ET-LT,G	S	-	-	-	X
MAL13P1.58	hypothetical protein	-	30,826	ER-LT	-	-	-	-	-
PF13_0073 **	hypothetical protein	-	30,826	ER-ET	T	X	-	X	X
MAL13P1.59 *	hypothetical protein	*pf-fam-f*	30,826	-	-	-	-	X	X
MAL13P1.60 *	erythrocyte binding antigen 140	*dbl-ebp*	30,826	LS-M	M,T	X	-	X	X
MAL13P1.61 **	hypothetical protein	-	30,826	M-LR	-	X	X	X	X
PF13_0074	hypothetical protein	*pf-fam-b*	30,826	M	-	-	-	-	X
PF13_0075	hypothetical protein, conserved in Pf	*pf-fam-b*	30,826	M-ER	-	-	-	-	-
MAL13P1.62 *	hypothetical protein	-	30,826	-	-	X	X	X	X
PF13_0076 **	hypothetical protein	-	30,826	M-ET	-	X	-	X	X
MAL13P1.106 *	hypothetical protein	-	21,907	G	G	X	-	X	X
MAL13P1.107 *	hypothetical protein	-	21,907	ET-LT	G,S	X	-	X	X
PF13_0115 *	frameshifted ebl1, pseudogene	*dbl-ebp*	21,907	LS-M	-	X	X	X	X
MAL13P1.109 **	pftstk13	*tstk*	21,907	-	-	X	-	X	X
MAL13P1.110 *	hypothetical protein	-	21,907	LR-LT	-	X	-	X	X
MAL13P1.122	hypothetical protein	-	8,478	S,B,G	T	-	-	-	-
PF13_0127	hypothetical protein	-	1,538	LT-LS	-	-	-	-	X
PF13_0153	hypothetical protein	-	3,014	S,B,G	M	-	-	-	X
PF13_0191 *	hypothetical protein	*msp*	10,811	ET-LT	S	X	X	X	X
PF13_0192 *	hypothetical protein	*msp*	10,811	ET-ES	-	X	-	X	X
PF13_0193 *	MSP7-like protein	*msp*	10,811	ES-LS	-	X	-	X	X
PF13_0194 *	hypothetical protein	-	10,811	M-LR	T	X	-	X	X
PF13_0195	MSP7 fragment, pseudogene	*msp*	10,811	S,B,G	-	-	-	-	-
MAL13P1.176	reticulocyte binding protein 2 homologue b	PF13_0198	14,894	LS-M	-	-	-	-	X
PF13_0198 *	reticulocyte binding protein 2 homologue a	MAL13P1.176	14,894	LS-M	S,T	-	-	X	X
MAL13P1.197	hypothetical protein	-	1,092	-	-	-	-	-	-
MAL13P1.214	phosphoethanolamine N-methyltransferase, putative	-	1,219	S,B,G	M	-	-	-	-
MAL13P1.235 *	hypothetical protein	-	1,227	-	-	-	-	X	X
PF13_0295	hypothetical protein	-	1,870	G,S	-	-	-	-	-
MAL13P1.295	hypothetical protein	-	6,149	ET-ES,G	S	-	-	-	X

Pf gene	Product	*pf-fam*[a] or homologue	Size indel (bp)	trans.	exp.	SP	AP	TM-N	TM
MAL13P1.303 *	polyadenylate-binding protein, putative	-	2,259	ET-LS,G	G,T,M	-	-	X	X
MAL13P1.306	hypothetical protein	-	2,491	ES-M	-	-	-	-	-
PF13_0338 *	hypothetical protein	-	2,390	LT-LS,G	M	X	-	X	X
MAL13P1.325	hypothetical protein	-	1,094	ET-LT	-	-	-	-	-
PF14_0036 *	acid phosphatase, putative	-	1,728	S,B,G	G,T,M	-	-	X	X
PF14_0076 *	plasmepsin 1 precursor	*protease*	10,526	M-LR	T,G,M	X	X	X	X
PF14_0077 *	plasmepsin 2 precursor	*protease*	10,526	LR-LT	T,M,G	-	-	X	X
PF14_0078 *	HAP protein	*protease*	10,526	ER-ES	T,G,M	X	X	X	X
PF14_0119 *	hypothetical protein, conserved	PF14_0117	962	ES-M	G,S	X	-	X	X
PF14_0206	hypothetical protein	-	2,861	G	-	-	-	-	X
PF14_0236	hypothetical protein	-	5,345	G	S,M	-	-	-	-
PF14_0262 *	hypothetical protein	-	6,925	S	-	-	-	X	X
PF14_0263	hypothetical protein	-	6,925	-	G	-	-	-	X
PF14_0291	hypothetical protein	-	3,677	M-LR	-	-	-	-	X
PF14_0297 *	putative ecto-nucleoside triphosphate diphosphohydrolase 1,	-	2,846	ET-ES	-	X	X	X	X
PF14_0463	chloroquine resistance marker protein	-	11,385	LT,LS	S,M	-	-	-	X
PF14_0565	hypothetical protein	-	2,867	B,G	M	-	-	-	-
PF14_0594	hypothetical protein	-	9,968	M	S	-	-	-	-
PF14_0638 *	hypothetical protein	-	2,582	-	G	X	-	X	X

[a] In the column *pf-fam* or homologue, various gene families are listed as well as the gene names of single homologous genes. References to the papers describing the gene families are as follows: *msp*, merozoite surface protein[338]; *protease*[339]; *acyl-CoA*, acyl-CoA synthetase[52] (Chapter 4); *clag*, cytoadherence-linked asexual gene[340]; *var*[80-82]; *vicar*, *var* internal cluster associated repeat gene[87] (Chapter 5); *rif*[84]; *apiap2*, apicomplexan apetala 2-integrase DNA-binding domain containing protein[129]; *etramp*, early transcribed membrane protein[259]; *pf-fam-f*, *P. falciparum* gene family f[52] (Chapter 4); *dbl-ebp*, Duffy-binding-like erythrocyte-binding protein[341]; *pf-fam-b*, *P. falciparum* gene family b[52] (Chapter 4); *tstk*, transforming growth factor β (TGF-β) receptor-like serine/threonine protein kinase[87] (Chapter 5).

Abbreviations: SB, synteny block; PEXEL, *Plasmodium* export element; VTS, vacuolar transport signal; Pf, *P. falciparum*; *pf-fam*, described *P. falciparum* gene family; SP, predicted signal peptide; AP, predicted apicoplast transit peptide; TM, transmembrane domain; TM-N, N-terminal TM-domain; trans., approximate stage of (highest) transcription for all genes except the *var*, *rif* and *vicar* genes based on transcriptome data[91,92]; S, sporozoite; M, merozoite; ER, early ring; LR, late ring; ET, early trophozoite; LT, late trophozoite; ES, early schizont; LS, late schizont; B, asexual blood stage; G, gametocyte; exp., expression of those genes based on proteome data[11,12]; M, merozoite; T, trophozoite; G, gametocyte; S, sporozoite.

Appendix 9: The 575 *P. falciparum*-specific subtelomeric genes.

A list of 575 *P. falciparum*-specific genes located in the subtelomeric regions including all pseudogenes. Genes from the two different subtelomeric regions of one chromosome are separated by a black line. Black bars separate the genes from different chromosomes. Core-telomere boundaries are given as the coordinates of the stop codon of the last subtelomeric gene.

Pf gene	Product	Core-Telomere Boundary	*pf-fam*[a] or homologue
PFA0005w	erythrocyte membrane protein 1 (PfEMP1)		*var*
PFA0010c	rifin		*rif*
PFA0015c	var-like protein		*var*
PFA0020w	rifin		*rif*
PFA0025c	var fragment, pseudogene		*var*
PFA0030c	rifin		*rif*
PFA0035c	hypothetical protein		*pf-fam-d*
PFA0040w	rifin		*rif*
PFA0045c	rifin		*rif*
PFA0050c	rifin		*rif*
PFA0055c	hypothetical protein, conserved in Pf		
PFA0060w	hypothetical protein, conserved in Pf		
PFA0065w	hypothetical protein, conserved in Pf		*pfmc-2tm*
PFA0070c	pseudogene, Pf-specific conserved gene family		
PFA0075w	var fragment, pseudogene		*var*
PFA0080c	rifin		*rif*
PFA0085c	truncated var-related protein		*var*
PFA0090c	stevor		*stevor*
PFA0095c	rifin		*rif*
PFA0100c	hypothetical protein		*pf-fam-d*
PFA0105w	stevor		*stevor*
PFA0110w	ring-infected erythrocyte surface antigen precursor		*pf-fam-a+h*
PFA0115w	hypothetical protein		
PFA0120c	hypothetical protein		
PFA0125c	Ebl-1 like protein, putative		
PFA0130c	pftstk1		*tstk*
PFA0135w	hypothetical protein	125,719	
PFA0590w	ABC transporter, putative	465,876	
rRNA1	putative, 18s; ITS1; 5.8s; ITS2; 28s rRNA		*rrna*
PFA0610c	hypothetical protein, conserved in Pf		PFB0110w
PFA0615w	hypothetical protein		
PFA0620c	glutamic acid-rich protein (garp)		
PFA0625w	hypothetical protein		*pf-fam-b*
PFA0630c	hypothetical protein		
PFA0635c	hypothetical protein		
PFA0640c	hypothetical protein		
PFA0645c	hypothetical protein		
PFA0650w	hypothetical protein		*pf-fam-b*
PFA0655w	hypothetical protein		*pf-fam-b*
PFA0660w	protein with DNAJ domain, dnj1/sis1 family		
PFA0665w	hypothetical protein		
PFA0670c	hypothetical protein		MAL13P1.61
PFA0675w	Pf RESA-like protein with DnaJ domain		*pf-fam-h*
PFA0680c	hypothetical protein, conserved in Pf		*pfmc-2tm*
PFA0685c	hypothetical protein, conserved in Pf		
PFA0690w	pseudogene, Pf-specific gene family		
PFA0695c	var-like pseudogene		*var*

Pf gene	Product	Core-Tel. Boundary	*pf-fam*[a] or homologue
PFA0700c	hypothetical protein, conserved in Pf		
PFA0705c	stevor pseudogene		*stevor*
PFA0710c	rifin		*rif*
PFA0715c	hypothetical protein		*pfst-2tm*
PFA0720w	hypothetical protein		
PFA0725w	hypothetical protein		*pf-fam-b*
PFA0730c	hypothetical protein		
PFA0735w	hypothetical protein		
PFA0740w	rifin		*rif*
PFA0745w	rifin		*rif*
PFA0750w	stevor		*stevor*
PFA0755w	var-related pseudogene		*var*
PFA0760w	rifin		*rif*
PFA0765c	erythrocyte membrane protein 1 (PfEMP1)		*var*
PFB0010w	erythrocyte membrane protein 1 (PfEMP1)		*var*
PFB0015c	rifin		*rif*
PFB0020c	erythrocyte membrane protein 1 (PfEMP1), truncated		*var*
PFB0025c	stevor, putative		*stevor*
PFB0030c	rifin		*rif*
PFB0035c	rifin		*rif*
PFB0040c	rifin		*rif*
PFB0045c	erythrocyte membrane protein 1 (PfEMP1), truncated		*var*
PFB0050c	stevor isoform gam beta		*stevor*
PFB0055c	rifin		*rif*
PFB0056c	hypothetical protein		*pf-fam-d*
PFB0060w	rifin		*rif*
PFB0065w	stevor, putative		*stevor*
PFB0070w	hypothetical protein		
PFB0075c	hypothetical protein		
PFB0080c	hypothetical protein		*pf-fam-a*
PFB0085c	hypothetical protein		*pf-fam-a+h*
PFB0090c	hypothetical protein, conserved		
PFB0095c	erythrocyte membrane protein 3 (PfEMP3)		
PFB0100c	knob associated histidine-rich protein		
PFB0105c	hypothetical protein		
PFB0106c	hypothetical protein		
PFB0110w	hypothetical protein		PFA0610c
PFB0115w	hypothetical protein		
PFB0120w	hypothetical protein	128,314	
PFB0900c	hypothetical protein	783,721	
PFB0905c	hypothetical protein		
PFB0910w	hypothetical protein		
PFB0915w	liver-stage antigen 3		
PFB0920w	hypothetical protein		*pf-fam-a+h*
PFB0921c	hypothetical protein		
PFB0923c	hypothetical protein		
PFB0925w	hypothetical protein		*pf-fam-h*
PFB0926c	hypothetical protein		
PFB0930w	hypothetical protein		
PFB0932w	hypothetical protein		
PFB0935w	cytoadherence linked asexual protein 2		*clag*
PFB0946c	hypothetical protein		
PFB0950w	hypothetical protein		
PFB0953w	hypothetical protein		

Pf gene	Product	Core-Tel. Boundary	*pf-fam*[a] or homologue
PFB0955w	stevor, degenerate, putative		*stevor*
PFB0960c	hypothetical protein		*pfmc-2tm*
PFB0965c	hypothetical protein		
PFB0970c	hypothetical protein		
PFB0972w	hypothetical protein		
PFB0973c	hypothetical protein		
PFB0974c	erythrocyte membrane protein 1 (PfEMP1), truncated, degenerate		*var*
PFB0975c	erythrocyte membrane protein 1 (PfEMP1), truncated		*var*
PFB0976w	hypothetical protein		
PFB0980w	hypothetical protein		*pf-fam-h*
PFB0985c	hypothetical protein		*pfmc-2tm*
PFB0990c	hypothetical protein		
PFB0995w	hypothetical protein		
PFB1000w	rifin		*rif*
PFB1005w	rifin		*rif*
PFB1010w	rifin		*rif*
PFB1015w	rifin		*rif*
PFB1020w	stevor, putative		*stevor*
PFB1025w	erythrocyte membrane protein 1 (PfEMP1), truncated, degenerate		*var*
PFB1030w	hypothetical protein		
PFB1035w	rifin		*rif*
PFB1040w	rifin		*rif*
PFB1045w	erythrocyte membrane protein 1 (PfEMP1), truncated		*var*
PFB1050w	rifin		*rif*
PFB1055c	erythrocyte membrane protein 1 (PfEMP1)		*var*
PFC0005w	var protein, putative		*var*
PFC0010c	rifin		*rif*
PFC0015c	varc pseudogene		*var*
PFC0025c	stevor, putative		*stevor*
PFC0030c	rifin		*rif*
PFC0035w	rifin		*rif*
PFC0040w	rifin		*rif*
PFC0045w	hypothetical protein		
PFC0050c	long chain fatty acid ligase, putative		*pf-fam-i*
PFC0055w	hypothetical protein		
PFC0060c	pftstk3		*tstk*
PFC0065c	alpha/beta hydrolase protein, putative		
PFC0070c	hypothetical protein		
PFC0075c	hypothetical protein		
PFC0080c	hypothetical protein		
PFC0085c	hypothetical protein		
PFC0090w	hypothetical protein		
PFC0095c	hypothetical protein	103,789	
PFC1070c	varc pseudogene	999,860	*var*
PFC1075w	hypothetical protein		*pf-fam-h*
PFC1080c	hypothetical protein, conserved in Pf		*pfmc-2tm*
PFC1085c	hypothetical protein, conserved		
PFC1090w	hypothetical protein, conserved in Pf		
PFC1095w	rifin (3D7-rifT3-5)		*rif*
PFC1100w	rifin (3D7-rifT3-6)		*rif*
PFC1105w	stevor (3D7-stevorT3-2)		*stevor*
PFC1110w	varc pseudogene		*var*

Pf gene	Product	Core-Tel. Boundary	*pf-fam*[a] or homologue
PFC1115w	rifin (3D7-rifT3-7)		*rif*
PFC1120c	var (3D7-varT3-2)		*var*
PFD0005w	erythrocyte membrane protein 1 (PfEMP1)		*var*
PFD0010w	unknown		
PFD0015c	rifin		*rif*
PFD0020c	erythrocyte membrane protein 1 (PfEMP1)		*var*
PFD0025w	rifin		*rif*
PFD0030c	rifin		*rif*
PFD0035c	stevor		*stevor*
PFD0040c	rifin		*rif*
PFD0045c	rifin		*rif*
PFD0050w	rifin		*rif*
PFD0055w	rifin		*rif*
PFD0060w	rifin		*rif*
PFD0065w	stevor pseudogene		*stevor*
PFD0070c	rifin		*rif*
PFD0075w	hypothetical protein, conserved in Pf		
PFD0080c	hypothetical protein		*pf-fam-a*
PFD0085c	ATP-dept. acyl-CoA synthetase, putative		*pf-fam-i*
PFD0090c	hypothetical protein		*pf-fam-f*
PFD0095c	hypothetical protein		*pf-fam-a*
PFD0100c	hypothetical protein		*pf-fam-b*
PFD0105c	hypothetical protein		*pf-fam-b*
PFD0110w	reticulocyte binding protein, putative		
PFD0115c	hypothetical protein		
PFD0120w	rif pseudogene		*rif*
PFD0125c	rifin		*rif*
PFD0130c	rif pseudogene		*rif*
PFD0135w	var pseudogene	178,412	*var*
PFD1120c	predicted integral membrane protein	1,075,873	
PFD1125c	hypothetical protein		
PFD1130w	hypothetical protein		
PFD1135c	hypothetical protein		
PFD1140w	hypothetical protein		*pf-fam-a*
PFD1145c	hypothetical protein		
PFD1150c	hypothetical protein		
PFD1155w	erythrocyte binding antigen, putative		
PFD1160w	hypothetical protein		*pf-fam-b*
PFD1165w	pftstk4a		*tstk*
PFD1170c	RESA-like protein, truncated		*pf-fam-a*
PFD1175w	pftstk4b		*tstk*
PFD1180w	trophozoite antigen r45-like protein,truncated		*pf-fam-a*
PFD1185w	hypothetical protein		*pf-fam-f*
PFD1190c	hypothetical protein		
PFD1195c	hypothetical protein		
PFD1200c	hypothetical protein		
PFD1205w	hypothetical integral membrane protein,conserved in Pf		
PFD1210w	hypothetical protein		
PFD1215w	hypothetical protein		
PFD1220c	rifin		*rif*
PFD1225w	rif pseudogene		*rif*
PFD1230c	rifin		*rif*
PFD1235w	erythrocyte membrane protein 1 (PfEMP1)		*var*
PFD1240w	rifin		*rif*

Pf gene	Product	Core-Tel. Boundary	*pf-fam*[a] or homologue
PFD1245c	erythrocyte membrane protein 1 (PfEMP1)		*var*
PFE0005w	erythrocyte membrane protein 1 (PfEMP1)		*var*
PFE0010c	var pseudogene		*var*
PFE0015c	rif pseudogene		*rif*
PFE0020c	rifin		*rif*
PFE0025c	rifin		*rif*
PFE0030c	stevor pseudogene		*stevor*
PFE0035c	rif pseudogene		*rif*
PFE0040c	Mature parasite-infected erythrocyte surface antigen (MESA) or PfEMP2		
PFE0045c	pftstk5		*tstk*
PFE0050w	hypothetical protein		
PFE0055c	heat shock protein, putative		
PFE0060w	hypothetical protein		
PFE0065w	skeleton binding protein		
PFE0070w	interspersed repeat antigen, putative	79,842	
PFE1590w	early transcribed membrane protein	1,301,219	
PFE1595c	hypothetical protein		
PFE1600w	hypothetical protein		*pf-fam-a*
PFE1605w	DNAJ protein		*pf-fam-a*
PFE1610w	hypothetical protein		
PFE1615c	hypothetical protein		
PFE1620c	var fragment, pseudogene		*var*
PFE1625c	var fragment, pseudogene		*var*
PFE1630w	rifin		*rif*
PFE1635w	rif pseudogene		*rif*
PFE1640w	erythrocyte membrane protein 1 (PfEMP1), truncated		*var*
MAL6P1.1	erythrocyte membrane protein 1 (PfEMP1)		*var*
MAL6P1.2	rifin		*rif*
MAL6P1.3	rif pseudogene		*rif*
MAL6P1.4	erythrocyte membrane protein 1 (PfEMP1)		*var*
MAL6P1.5	rifin		*rif*
MAL6P1.6	rifin		*rif*
MAL6P1.7	rifin		*rif*
MAL6P1.8	rifin		*rif*
MAL6P1.9	rifin		*rif*
MAL6P1.10	stevor		*stevor*
MAL6P1.11	rifin		*rif*
MAL6P1.12	rif pseudogene		*rif*
MAL6P1.13	hypothetical protein, conserved in Pf		
MAL6P1.14	hypothetical protein, conserved in Pf		
MAL6P1.15	hypothetical protein, conserved in Pf		*pfmc-2tm*
MAL6P1.16	hypothetical protein, conserved in Pf		*pf-fam-h*
MAL6P1.17	var fragment, pseudogene		*var*
MAL6P1.18	var pseudogene		*var*
MAL6P1.19	hypothetical protein		*pf-fam-a*
MAL6P1.20	hypothetical protein, conserved		MAL6P1.117
MAL6P1.21	hypothetical protein		*pf-fam-d+e*
MAL6P1.22	hypothetical protein	128,781	
MAL6P1.310	rifin-related protein	1,341,840	*rif*
MAL6P1.311	rifin		*rif*
MAL6P1.312	erythrocyte membrane protein 1 (PfEMP1), truncated		*var*
MAL6P1.313	rifin		*rif*
MAL6P1.314	Pf var-like protein		*var*

Pf gene	Product	Core-Tel. Boundary	*pf-fam*[a] or homologue
MAL6P1.315	rifin		*rif*
MAL6P1.316	erythrocyte membrane protein 1 (PfEMP1)		*var*
MAL6P1.317	var-like protein, truncated		*var*
MAL7P1.1	erythrocyte membrane protein 1 (PfEMP1)		*var*
MAL7P1.2	rifin		*rif*
PF07_0001	rifin		*rif*
MAL1P1.3a	pseudogene, predicted protein		
MAL7P1.3	hypothetical protein, conserved in Pf		
MAL7P1.4	hypothetical protein, conserved in Pf		
MAL7P1.5	hypothetical protein, conserved in Pf		*pfmc-2tm*
PF07_0002	hypothetical protein, conserved in Pf		*pf-fam-h*
PF07_0003	rifin		*rif*
PF07_0004	hypothetical protein		
MAL7P1.6	hypothetical protein		
MAL7P1.7	RESA-like protein		*pf-fam-a*
PF07_0005	lysophospholipases-like protein, putative		
PF07_0006	starp antigen		
PF07_0007	hypothetical protein		
PF07_0008	hypothetical protein		
PF07_0009	chitinase precursor fragment, truncated	82,976	
MAL7P1.170	ring stage expressed protein	1,229,225	
MAL7P1.171	hypothetical protein		
MAL7P1.172	hypothetical protein		
MAL7P1.173	hypothetical protein		
MAL7P1.174	hypothetical protein		*pf-fam-a*
MAL7P1.175	pftstk7b		*tstk*
PF07_0128	erythrocyte binding antigen		
PF07_0129	ATP-dept. acyl-CoA synthetase		*pf-fam-i*
MAL7P1.177	predicted integral membrane protein, conserved in Pf		
MAL7P1.178	hypothetical protein		
PF07_0130	stevor		*stevor*
PF07_0131	var pseudogene		*var*
PF07_0132	rifin		*rif*
PF07_0133	rifin		*rif*
PF07_0134	rifin		*rif*
PF07_0135	rifin		*rif*
PF07_0136	rifin		*rif*
PF07_0137	rifin		*rif*
PF07_0138	rifin		*rif*
PF07_0139	erythrocyte membrane protein 1 (PfEMP1)		*var*
MAL8P1.1	conserved hypothetical protein, conserved in Pf		*pf-fam-b*
MAL8P1.2	hypothetical protein with DNAJ domain		*pf-fam-a*
PF08_0001	hypothetical protein		
PF08_0002	hypothetical protein, conserved in Pf		*pf-fam-b*
MAL8P1.3	integral membrane protein, conserved in Pf		
MAL8P1.4	hypothetical protein		
PF08_0003	tryptophan/threonine-rich antigen		
PF08_0004	hypothetical protein		
PF08_0005	hypothetical protein		
MAL8P1.5	hypothetical protein		
MAL8P1.6	hypothetical protein		
MAL8P1.7	hypothetical protein	87,253	
rRNA8b	rRNA	1,222,682	*rrna*
PF08_0137	hypothetical protein		

Pf gene	Product	Core-Tel. Boundary	*pf-fam*[a] or homologue
MAL8P1.160	hypothetical protein		*pfst-2tm*
MAL8P1.161	hypothetical protein		*pfst-2tm*
MAL8P1.162	hypothetical protein		*pf-fam-b*
MAL8P1.163	hypothetical protein		
PF08_0138	rifin		*rif*
PF08_0139	rifin		*rif*
PF08_0140	erythrocyte membrane protein 1 (PfEMP1)		*var*
PF08_0141	erythrocyte membrane protein 1 (PfEMP1)		*var*
PF08_0142	erythrocyte membrane protein 1 (PfEMP1)		*var*
PFI0005w	erythrocyte membrane protein 1 (PfEMP1)		*var*
PFI0010c	rifin		*rif*
PFI0015c	rifin		*rif*
PFI0020w	rifin		*rif*
PFI0025c	rifin		*rif*
PFI0030c	rifin		*rif*
PFI0035c	rifin		*rif*
PFI0040c	varc-like pseudogene		*var*
PFI0045c	stevor		*stevor*
PFI0050c	rifin		*rif*
PFI0055c	rifin		*rif*
PFI0060c	hypothetical protein		*pf-fam-d*
PFI0065w	rifin		*rif*
PFI0070w	rifin		*rif*
PFI0075w	rifin		*rif*
PFI0080w	stevor		*stevor*
PFI0085c	hypothetical protein		
PFI0090c	hypothetical protein		
PFI0095c	pftstk9a		*tstk*
PFI0100c	pftstk9b		*tstk*
PFI0105c	pftstk9c		*tstk*
PFI0110c	pftstk9d		*tstk*
PFI0115c	pftstk9e		*tstk*
PFI0120c	pftstk9f		*tstk*
PFI0125c	pftstk9g		*tstk*
PFI0130c	hypothetical protein		*pf-fam-a*
PFI0135c	papain family cysteine protease, putative		
PFI0140w	hypothetical protein	127,723	
PFI1710w	cytoadherence-linked protein	1,377,942	
PFI1715w	hypothetical protein		
PFI1720w	hypothetical protein		
PFI1725w	hypothetical protein		
PFI1730w	cytoadherence linked asexual protein		*clag*
PFI1735c	hypothetical protein		
PFI1740c	hypothetical protein		
PFI1745c	hypothetical protein		
PFI1750c	hypothetical protein		
PFI1755c	hypothetical protein		
PFI1760w	hypothetical protein		
PFI1765c	hypothetical protein		
PFI1770w	hypothetical protein		*pf-fam-f*
PFI1775w	hypothetical protein		
PFI1780w	hypothetical protein		
PFI1785w	hypothetical protein		*pf-fam-a*
PFI1790w	hypothetical protein		*pf-fam-a*

Pf gene	Product	Core-Tel. Boundary	*pf-fam*[a] or homologue
PFI1795c	hypothetical protein		
PFI1800w	enzyme, putative		
PFI1805w	rifin		*rif*
PFI1810w	rifin		*rif*
PFI1815c	rifin		*rif*
PFI1820w	erythrocyte membrane protein 1 (PfEMP1)		*var*
PFI1825w	rifin		*rif*
PFI1830c	erythrocyte membrane protein 1 (PfEMP1)		*var*
PF10_0001	erythrocyte membrane protein 1 (PfEMP1)		*var*
PF10_0002	rifin		*rif*
PF10_0003	rifin		*rif*
PF10_0004	rifin		*rif*
PF10_0005	rifin		*rif*
PF10_0006	rifin		*rif*
PF10_0007	hypothetical protein		*pf-fam-d*
PF10_0008	hypothetical protein		
PF10_0009	pseudogene, stevor, putative		*stevor*
PF10_0011	erythrocyte membrane protein 1 (PfEMP1), truncated, degenerate		*var*
PF10_0012	erythrocyte membrane protein 1 (PfEMP1), truncated		*var*
PF10_0013	hypothetical protein		
PF10_0014	hypothetical protein		
PF10_0015	acyl CoA binding protein, putative		
PF10_0016	acyl CoA binding protein, putative		
PF10_0017	hypothetical protein		*pf-fam-d+f*
PF10_0018	hypothetical protein		
PF10_0019	early transcribed membrane protein		
PF10_0020	hypothetical protein		
PF10_0021	hypothetical protein		
PF10_0022	hypothetical protein		
PF10_0023	hypothetical protein		
PF10_0024	hypothetical protein		
PF10_0025	PF70 protein		
PF10_0026	hypothetical protein	116,016	
PF10_0374	gene 11-1 protein precursor	1,522,017	
PF10_0375	hypothetical protein		
PF10_0376	hypothetical protein		
PF10_0377	hypothetical protein		
PF10_0378	hypothetical protein		*pf-fam-a+h*
PF10_0379	phospholipase, putative		
PF10_0380	pftstk10b		*tstk*
PF10_0381	hypothetical protein		
PF10_0382	hypothetical protein		
PF10_0383	hypothetical protein, conserved		
PF10_0384	hypothetical protein		
PF10_0385	PfEMP-1, truncated, putative		*var*
PF10_0386	hypothetical protein		
PF10_0387	hypothetical protein		
PF10_0388	hypothetical protein		
PF10_0389	hypothetical protein		
PF10_0390	hypothetical protein		*pfmc-2tm*
PF10_0391	hypothetical protein		
PF10_0392	hypothetical protein		
PF10_0393	rifin		*rif*

Pf gene	Product	Core-Tel. Boundary	*pf-fam*[a] or homologue
PF10_0394	rifin		*rif*
PF10_0395	stevor, putative		*stevor*
PF10_0396	rifin		*rif*
PF10_0397	rifin		*rif*
PF10_0398	rifin		*rif*
PF10_0399	rifin		*rif*
PF10_0400	rifin		*rif*
PF10_0401	rifin		*rif*
PF10_0402	rifin		*rif*
PF10_0403	rifin		*rif*
PF10_0404	rifin		*rif*
PF10_0405	rifin		*rif*
PF10_0406	erythrocyte membrane protein 1 (PfEMP1)		*var*
PF11_0001	hypothetical protein		
PF11_0002	hypothetical protein		
PF11_0003	hypothetical protein		
PF11_0004	hypothetical protein		
PF11_0005	hypothetical protein		
PF11_0006	hypothetical protein		
PF11_0007	erythrocyte membrane protein 1 (PfEMP1)		*var*
PF11_0008	erythrocyte membrane protein 1 (PfEMP1)		*var*
PF11_0009	rifin		*rif*
PF11_0010	rifin		*rif*
PF11_0011	rifin		*rif*
PF11_0012	hypothetical protein		*pf-fam-d*
PF11_0013	stevor, putative, degenerate		*stevor*
PF11_0014	hypothetical protein		*pfmc-2tm*
PF11_0015	hypothetical protein		
PF11_0016	hypothetical protein		
PF11_0522	pseudogene, erythrocyte membrane protein 1 (PfEMP1), truncated		*var*
PF11_0529	rifin		*rif*
PF11_0020	rifin		*rif*
PF11_0021	rifin		*rif*
PF11_0022	pseudogene, rifin, degenerate, putative		*rif*
PF11_0023	hypothetical protein		
PF11_0024	hypothetical protein		
PF11_0025	hypothetical protein		*pfmc-2tm*
PF11_0026	hypothetical protein		*pf-fam-h*
PF11_0523	erythrocyte membrane protein 1 (PfEMP1), truncated		*var*
PF11_0032	hypothetical protein		
PF11_0033	hypothetical protein		
PF11_0034	hypothetical protein		*pf-fam-h*
PF11_0035	hypothetical protein		
PF11_0036	hypothetical protein, conserved		
PF11_0037	hypothetical protein		*pf-fam-a*
PF11_0038	hypothetical protein		
PF11_0039	early transcribed membrane protein 11.1		
PF11_0040	early transcribed membrane protein 11.2		
PF11_0041	hypothetical protein		
PF11_0042	hypothetical protein	136,651	
PF11_0489	hypothetical protein	1,917,905	
PF11_0490	hypothetical protein		
rRNA11	rrna		*rrna*

Pf gene	Product	Core-Tel. Boundary	*pf-fam*[a] or homologue
PF11_0503	hypothetical protein		
PF11_0504	hypothetical protein		
PF11_0505	hypothetical protein		
PF11_0506	hypothetical protein		
PF11_0507	antigen 332, putative		
PF11_0508	hypothetical protein		*pf-fam-a*
PF11_0509	ring-infected erythrocyte surface antigen, putative		*pf-fam-a+h*
PF11_0510	pftstk11		*tstk*
PF11_0511	hypothetical protein		
PF11_0512	ring-infected erythrocyte surface antigen 2, RESA-2 - malaria parasite (Pf)-related		*pf-fam-a*
PF11_0513	hypothetical protein		*pf-fam-h*
PF11_0514	hypothetical protein		*pf-fam-d*
PF11_0515	rifin		*rif*
PF11_0516	stevor, putative		*stevor*
PF11_0517	rifin		*rif*
PF11_0518	rifin, putative, truncated		*rif*
PF11_0519	rifin		*rif*
PF11_0520	rifin		*rif*
PF11_0521	erythrocyte membrane protein 1 (PfEMP1)		*var*
PFL0005w	erythrocyte membrane protein 1 (PfEMP1)		*var*
PFL0010c	rifin		*rif*
PFL0015c	rifin		*rif*
PFL0020w	erythrocyte membrane protein 1 (PfEMP1)		*var*
PFL0025c	rifin		*rif*
PFL0030c	erythrocyte membrane protein 1 (PfEMP1)		*var*
PFL0035c	octapeptide-repeat antigen, putative		*pf-fam-i*
PFL0040c	pftstk12		*tstk*
PFL0045c	hypothetical protein		
PFL0050c	hypothetical protein		*pf-fam-a*
PFL0055c	protein with DNAJ domain (resa-like), putative		*pf-fam-a+h*
PFL0060w	hypothetical protein		
PFL0065w	hypothetical protein		
PFL0070c	hypothetical protein	97,146	
PFL2510w	chitinase	2,125,420	
PFL2515c	hypothetical protein		
PFL2520w	hypothetical protein		
PFL2525c	hypothetical protein		
PFL2530w	hypothetical protein		
PFL2535w	RESA-like protein, putative		*pf-fam-a*
PFL2540w	hypothetical protein		*pf-fam-a*
PFL2545c	hypothetical protein		
PFL2550w	hypothetical protein, conserved in Pf		*pf-fam-h*
PFL2555w	hypothetical protein		*pf-fam-f*
PFL2560c	hypothetical protein		
PFL2565w	hypothetical protein		*pf-fam-e*
PFL2570w	acyl-coa ligase antigen		*pf-fam-i*
PFL2575c	hypothetical protein		
PFL2580w	hypothetical protein		
PFL2585c	rifin		*rif*
PFL2590w	hypothetical protein		*pf-fam-f*
PFL2595w	hypothetical protein		*pf-fam-e*
PFL2600w	hypothetical protein		*pf-fam-d*
PFL2605w	rifin		*rif*

Pf gene	Product	Core-Tel. Boundary	*pf-fam*[a] or homologue
PFL2610w	stevor		*stevor*
PFL2615w	rifin		*rif*
PFL2620w	stevor		*stevor*
PFL2625w	rifin		*rif*
PFL2630w	rifin		*rif*
PFL2635w	stevor		*stevor*
PFL2640c	rifin		*rif*
PFL2645c	rifin		*rif*
PFL2650w	hypothetical protein		*pf-fam-d*
PFL2655w	rifin		*rif*
PFL2660w	rifin		*rif*
PFL2665c	erythrocyte membrane protein 1 (PfEMP1)		*var*
PF13_0001	erythrocyte membrane protein 1 (PfEMP1)		*var*
PF13_0002	rifin		*rif*
PF13_0003	erythrocyte membrane protein 1 (PfEMP1)		*var*
PF13_0004	rifin		*rif*
PF13_0005	rifin		*rif*
PF13_0006	rifin		*rif*
PF13_0007	rifin		*rif*
PF13_0008	var pseudogene		*var*
PF13_0009	stevor		*stevor*
MAL13P1.8	rif pseudogene		*rif*
MAL13P1.11a	hypothetical protein, conserved in Pf		*pf-fam-d*
MAL13P1.11	hypothetical protein		*pf-fam-e*
PF13_0010	gbph2		
PF13_0011	Pf gamete antigen 27/25		
PF13_0012	hypothetical protein	83,091	
MAL13P1.353	hypothetical protein	2,672,731	
MAL13P1.354	erythrocyte membrane protein 1 (PfEMP1), pseudogene		*var*
PF13_0363	rifin		*rif*
PF13_0364	erythrocyte membrane protein 1 (PfEMP1)		*var*
rRNA13	rrna		*rrna*
PF14_0001	erythrocyte membrane protein 1 (PfEMP1), truncated, pseudogene		*var*
PF14_0002	rifin		*rif*
PF14_0003	rifin		*rif*
PF14_0004	rifin		*rif*
PF14_0005	rifin		*rif*
PF14_0006	rifin		*rif*
PF14_0007	stevor, putative		*stevor*
PF14_0008	rifin		*rif*
PF14_0009	hypothetical protein		*pf-fam-d+e*
PF14_0010	glycophorin binding protein-related antigen		
PF14_0013	hypothetical protein		*pf-fam-h*
PF14_0014	hypothetical protein		
PF14_0015	aminopeptidase, putative		
PF14_0016	hypothetical protein		
PF14_0017	lysophospholipase, putative		
PF14_0018	hypothetical protein		*pf-fam-a*
PF14_0019	hypothetical protein	67,081	
PF14_0726	hypothetical protein	3,119,924	
PF14_0727	hypothetical protein		*pf-fam-a*
PF14_0728	hypothetical protein		

Pf gene	Product	Core-Tel. Boundary	*pf-fam*[a] or homologue
PF14_0729	hypothetical protein		
PF14_0730	hypothetical protein		
PF14_0731	hypothetical protein		*pf-fam-f*
PF14_0732	hypothetical protein		*pf-fam-a*
PF14_0733	pftstk14		*tstk*
PF14_0734	pftstk14		*tstk*
PF14_0735	hypothetical protein		
PF14_0736	hypothetical protein		
PF14_0737	lysophospholipase, putative		
PF14_0738	lysophospholipase, putative		
PF14_0739	hypothetical protein		
PF14_0740	hypothetical protein		
PF14_0741	hypothetical protein		
PF14_0742	hypothetical protein		
PF14_0743	hypothetical protein		
PF14_0744	hypothetical protein		
PF14_0745	hypothetical protein		
PF14_0746	hypothetical protein		*pf-fam-a*
PF14_0747	hypothetical protein		*pf-fam-b*
PF14_0748	hypothetical protein		
PF14_0749	acyl CoA binding protein		
PF14_0751	fatty acyl coenzyme A synthetase-1, putative		*pf-fam-i*
PF14_0752	hypothetical protein		*pf-fam-f*
PF14_0753	hypothetical protein		
PF14_0754	hypothetical protein		
PF14_0755	hypothetical protein		
PF14_0756	hypothetical protein		
PF14_0757	hypothetical protein		*pf-fam-d*
PF14_0758	hypothetical protein		
PF14_0759	hypothetical protein		
PF14_0760	hypothetical protein		
PF14_0761	fatty acyl CoA synthetase 1		*pf-fam-i*
PF14_0762	hypothetical protein		
PF14_0763	hypothetical protein		*pf-fam-f*
PF14_0764	hypothetical protein		*pf-fam-e*
PF14_0765	hypothetical protein		*pf-fam-d*
PF14_0766	rifin		*rif*
PF14_0767	stevor, putative		*stevor*
PF14_0768	rifin		*rif*
PF14_0769	rifin		*rif*
PF14_0770	rifin		*rif*
PF14_0771	stevor, putative		*stevor*
PF14_0772	rifin		*rif*
PF14_0773	erythrocyte membrane protein 1 (PfEMP1), truncated, pseudogene		*var*

[a] In the column *pf-fam* or homologue, various gene families are listed as well as the gene names of single homologous genes. References to the papers describing the gene families are as follows: *clag*, cytoadherence-linked asexual gene[340]; *var*[80-82]; *vicar*, *var* internal cluster associated repeat gene[87] (Chapter 5); *rif*[84]; *pf-fam-a/i*, *P. falciparum* gene family a/i[52] (Chapter 4); *tstk*, transforming growth factor β (TGF-β) receptor-like serine/threonine protein kinase[87] (Chapter 5).

Abbreviations: Pf, *P. falciparum*; *pf-fam*, described *P. falciparum* gene family.

References

References

1. Snow RW, Guerra CA, Noor AM, Myint HY & Hay SI. The global distribution of clinical episodes of Plasmodium falciparum malaria. *Nature* **434**, 214-217 (2005).
2. Hastings IM, Bray PG & Ward SA. Parasitology. A requiem for chloroquine. *Science* **298**, 74-75 (2002).
3. Wootton JC, Feng X, Ferdig MT, Cooper RA, Mu J *et al*. Genetic diversity and chloroquine selective sweeps in Plasmodium falciparum. *Nature* **418**, 320-323 (2002).
4. Sidhu AB, Verdier-Pinard D & Fidock DA. Chloroquine resistance in Plasmodium falciparum malaria parasites conferred by pfcrt mutations. *Science* **298**, 210-213 (2002).
5. Hemingway J, Field L & Vontas J. An overview of insecticide resistance. *Science* **298**, 96-97 (2002).
6. Hartl DL. The origin of malaria: mixed messages from genetic diversity. *Nat. Rev. Microbiol.* **2**, 15-22 (2004).
7. Hay SI, Guerra CA, Tatem AJ, Atkinson PM & Snow RW. Urbanization, malaria transmission and disease burden in Africa. *Nat. Rev. Microbiol.* **3**, 81-90 (2005).
8. Gallup JL & Sachs JD. The economic burden of malaria. *Am. J. Trop. Med. Hyg.* **64**, 85-96 (2001).
9. Jomaa H, Wiesner J, Sanderbrand S, Altincicek B, Weidemeyer C *et al*. Inhibitors of the nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs. *Science* **285**, 1573-1576 (1999).
10. Surolia N & Surolia A. Triclosan offers protection against blood stages of malaria by inhibiting enoyl-ACP reductase of Plasmodium falciparum. *Nat. Med.* **7**, 167-173 (2001).
11. Florens L, Washburn MP, Raine JD, Anthony RM, Grainger M *et al*. A proteomic view of the Plasmodium falciparum life cycle. *Nature* **419**, 520-526 (2002).
12. Lasonder E, Ishihama Y, Andersen JS, Vermunt AM, Pain A *et al*. Analysis of the Plasmodium falciparum proteome by high-accuracy mass spectrometry. *Nature* **419**, 537-542 (2002).
13. Doolittle RF. The grand assault. *Nature* **419**, 493-494 (2002).
14. Rich SM & Ayala FJ. Progress in malaria research: the case for phylogenetics. *Adv. Parasitol.* **54**, 255-280 (2003).
15. Lou J, Lucas R & Grau GE. Pathogenesis of cerebral malaria: recent experimental data and possible applications for humans. *Clin. Microbiol. Rev.* **14**, 810-20, table (2001).
16. Rae C, McQuillan JA, Parekh SB, Bubb WA, Weiser S *et al*. Brain gene expression, metabolism, and bioenergetics: interrelationships in murine models of cerebral and noncerebral malaria. *FASEB J.* **18**, 499-510 (2004).
17. de Roode JC, Pansini R, Cheesman SJ, Helinski ME, Huijben S *et al*. Virulence and competitive ability in genetically diverse malaria infections. *Proc. Natl. Acad. Sci. U. S. A* **102**, 7624-7628 (2005).
18. Ellis J, Ozaki LS, Gwadz RW, Cochrane AH, Nussenzweig V *et al*. Cloning and expression in E. coli of the malarial sporozoite surface antigen gene from Plasmodium knowlesi. *Nature* **302**, 536-538 (1983).
19. Dame JB, Williams JL, McCutchan TF, Weber JL, Wirtz RA *et al*. Structure of the gene encoding the immunodominant surface antigen on the sporozoite of the human malaria parasite Plasmodium falciparum. *Science* **225**, 593-599 (1984).
20. Dame JB, Arnot DE, Bourke PF, Chakrabarti D, Christodoulou Z *et al*. Current status of the Plasmodium falciparum genome project. *Mol. Biochem. Parasitol.* **79**, 1-12 (1996).
21. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B *et al*. Life with 6000 genes. *Science* **274**, 546, 563-546, 567 (1996).
22. Foote SJ & Kemp DJ. Chromosomes of malaria parasites. *Trends Genet.* **5**, 337-342 (1989).
23. Janse CJ. Chromosome size polymorphisms and DNA rearrangements in Plasmodium. *Parasitol. Today* **9**, 19-22 (1993).
24. Scherf A, Figueiredo LM & Freitas-Junior LH. Plasmodium telomeres: a pathogen's perspective. *Curr. Opin. Microbiol.* **4**, 409-414 (2001).
25. Carlton JM, Vinkenoog R, Waters AP & Walliker D. Gene synteny in species of Plasmodium. *Mol. Biochem. Parasitol.* **93**, 285-294 (1998).
26. Janse CJ, Carlton JM, Walliker D & Waters AP. Conserved location of genes on polymorphic chromosomes of four species of malaria parasites. *Mol. Biochem. Parasitol.* **68**, 285-296 (1994).

27. Waller RF & McFadden GI. The apicoplast. **In:** Malaria parasites, genomes and molecular biology - p289-338. Waters AP & Janse CJ (eds.) Caister Academic Press, Wymondham (2004).
28. Moorthy VS, Good MF & Hill AV. Malaria vaccine developments. *Lancet* **363**, 150-156 (2004).
29. Hoffman SL, Bancroft WH, Gottlieb M, James SL, Burroughs EC *et al*. Funding for malaria genome sequencing. *Nature* **387**, 647 (1997).
30. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC *et al*. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
31. Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ *et al*. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493-521 (2004).
32. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF *et al*. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-562 (2002).
33. Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM *et al*. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science* **297**, 1301-1310 (2002).
34. Dehal P, Satou Y, Campbell RK, Chapman J, Degnan B *et al*. The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins. *Science* **298**, 2157-2167 (2002).
35. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD *et al*. The genome sequence of Drosophila melanogaster. *Science* **287**, 2185-2195 (2000).
36. Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R *et al*. The genome sequence of the malaria mosquito Anopheles gambiae. *Science* **298**, 129-149 (2002).
37. The C.elegans Sequencing Consortium. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282**, 2012-2018 (1998).
38. Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR *et al*. The Genome Sequence of Caenorhabditis briggsae: A Platform for Comparative Genomics. *PLoS. Biol.* **1**, E45 (2003).
39. Goff SA, Ricke D, Lan TH, Presting G, Wang R *et al*. A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). *Science* **296**, 92-100 (2002).
40. Yu J, Hu S, Wang J, Wong GK, Li S *et al*. A draft sequence of the rice genome (Oryza sativa L. ssp. indica). *Science* **296**, 79-92 (2002).
41. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**, 796-815 (2000).
42. Gardner MJ, Hall N, Fung E, White O, Berriman M *et al*. Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* **419**, 498-511 (2002).
43. Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G *et al*. Complete genome sequence of the apicomplexan, Cryptosporidium parvum. *Science* **304**, 441-445 (2004).
44. Xu P, Widmer G, Wang Y, Ozaki LS, Alves JM *et al*. The genome of Cryptosporidium hominis. *Nature* **431**, 1107-1112 (2004).
45. Gardner MJ, Bishop R, Shah T, de Villiers EP, Carlton JM *et al*. Genome Sequence of Theileria parva, a Bovine Pathogen That Transforms Lymphocytes. *Science* **309**, 134-137 (2005).
46. Pain A, Renauld H, Berriman M, Murphy L, Yeats CA *et al*. Genome of the Host-Cell Transforming Parasite Theileria annulata Compared with T. parva. *Science* **309**, 131-133 (2005).
47. Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H *et al*. The genome of the African trypanosome Trypanosoma brucei. *Science* **309**, 416-422 (2005).
48. El Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G *et al*. The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease. *Science* **309**, 409-415 (2005).
49. Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G *et al*. The genome of the kinetoplastid parasite, Leishmania major. *Science* **309**, 436-442 (2005).
50. Loftus B, Anderson I, Davies R, Alsmark UC, Samuelson J *et al*. The genome of the protist parasite Entamoeba histolytica. *Nature* **433**, 865-868 (2005).
51. Carlton JM, Angiuoli SV, Suh BB, Kooij TW, Pertea M *et al*. Genome sequence and comparative analysis of the model rodent malaria parasite Plasmodium yoelii yoelii. *Nature* **419**, 512-519 (2002).
52. Hall N, Karras M, Raine JD, Carlton JM, Kooij TW *et al*. A comprehensive survey of the Plasmodium life cycle by genomic, transcriptomic, and proteomic analyses. *Science* **307**, 82-86 (2005).
53. Wolfe KH & Shields DC. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**, 708-713 (1997).

54. Zdobnov EM, von Mering C, Letunic I, Torrents D, Suyama M *et al*. Comparative genome and proteome analysis of Anopheles gambiae and Drosophila melanogaster. *Science* **298**, 149-159 (2002).
55. Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR *et al*. Comparative genomics of the eukaryotes. *Science* **287**, 2204-2215 (2000).
56. Kellis M, Patterson N, Endrizzi M, Birren B & Lander ES. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241-254 (2003).
57. Glaser P, Frangeul L, Buchrieser C, Rusniok C, Amend A *et al*. Comparative genomics of Listeria species. *Science* **294**, 849-852 (2001).
58. Ivanova N, Sorokin A, Anderson I, Galleron N, Candelon B *et al*. Genome sequence of Bacillus cereus and comparative analysis with Bacillus anthracis. *Nature* **423**, 87-91 (2003).
59. Parkhill J, Sebaihia M, Preston A, Murphy LD, Thomson N *et al*. Comparative analysis of the genome sequences of Bordetella pertussis, Bordetella parapertussis and Bordetella bronchiseptica. *Nat. Genet.* **35**, 32-40 (2003).
60. van Lin LH, Pace T, Janse CJ, Birago C, Ramesar J *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).
61. Chiaromonte F, Yang S, Elnitski L, Yap VB, Miller W *et al*. Association between divergence and interspersed repeats in mammalian noncoding genomic DNA. *Proc. Natl. Acad. Sci. U. S. A* **98**, 14503-14508 (2001).
62. Freitas-Junior LH, Bottius E, Pirrit LA, Deitsch KW, Scheidig C *et al*. Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of P. falciparum. *Nature* **407**, 1018-1022 (2000).
63. Dehal P, Predki P, Olsen AS, Kobayashi A, Folta P *et al*. Human chromosome 19 and related regions in mouse: conservative and lineage-specific evolution. *Science* **293**, 104-111 (2001).
64. Eichler EE. Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends Genet.* **17**, 661-669 (2001).
65. Stankiewicz P & Lupski JR. Genome architecture, rearrangements and genomic disorders. *Trends Genet.* **18**, 74-82 (2002).
66. Coghlan A & Wolfe KH. Fourfold faster rate of genome rearrangement in nematodes than in Drosophila. *Genome Res.* **12**, 857-867 (2002).
67. Pevzner P & Tesler G. Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. *Genome Res.* **13**, 37-45 (2003).
68. Pevzner P & Tesler G. Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proc. Natl. Acad. Sci. U. S. A* **100**, 7672-7677 (2003).
69. van Lin LH, Janse CJ & Waters AP. The conserved genome organisation of non-falciparum malaria species: the need to know more. *Int. J. Parasitol.* **30**, 357-370 (2000).
70. Eichler EE & Sankoff D. Structural dynamics of eukaryotic chromosome evolution. *Science* **301**, 793-797 (2003).
71. Liti G & Louis EJ. Yeast Evolution and Comparative Genomics. *Annu. Rev. Microbiol.* (2004).
72. van Lin LH, Pace T, Janse CJ, Scotti R & Ponzi M. A long range restriction map of chromosome 5 of Plasmodium berghei demonstrates a chromosome specific symmetrical subtelomeric organisation. *Mol. Biochem. Parasitol.* **86**, 111-115 (1997).
73. Wirth DF. Biological revelations. *Nature* **419**, 495-496 (2002).
74. Gardner MJ, Tettelin H, Carucci DJ, Cummings LM, Aravind L *et al*. Chromosome 2 sequence of the human malaria parasite Plasmodium falciparum. *Science* **282**, 1126-1132 (1998).
75. Bowman S, Lawson D, Basham D, Brown D, Chillingworth T *et al*. The complete nucleotide sequence of chromosome 3 of Plasmodium falciparum. *Nature* **400**, 532-538 (1999).
76. Foth BJ, Ralph SA, Tonkin CJ, Struck NS, Fraunholz M *et al*. Dissecting apicoplast targeting in the malaria parasite Plasmodium falciparum. *Science* **299**, 705-708 (2003).
77. Zuegge J, Ralph S, Schmuker M, McFadden GI & Schneider G. Deciphering apicoplast targeting signals--feature extraction from nuclear-encoded precursors of Plasmodium falciparum apicoplast proteins. *Gene* **280**, 19-26 (2001).
78. Foth BJ & McFadden GI. The apicoplast: a plastid in Plasmodium falciparum and other Apicomplexan parasites. *Int. Rev. Cytol.* **224**, 57-110 (2003).

79. Ralph SA, van Dooren GG, Waller RF, Crawford MJ, Fraunholz MJ *et al*. Tropical infectious diseases: Metabolic maps and functions of the Plasmodium falciparum apicoplast. *Nat. Rev. Microbiol.* **2**, 203-216 (2004).
80. Baruch DI, Pasloske BL, Singh HB, Bi X, Ma XC *et al*. Cloning the P. falciparum gene encoding PfEMP1, a malarial variant antigen and adherence receptor on the surface of parasitized human erythrocytes. *Cell* **82**, 77-87 (1995).
81. Smith JD, Chitnis CE, Craig AG, Roberts DJ, Hudson-Taylor DE *et al*. Switches in expression of Plasmodium falciparum var genes correlate with changes in antigenic and cytoadherent phenotypes of infected erythrocytes. *Cell* **82**, 101-110 (1995).
82. Su XZ, Heatwole VM, Wertheimer SP, Guinet F, Herrfeldt JA *et al*. The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of Plasmodium falciparum-infected erythrocytes. *Cell* **82**, 89-100 (1995).
83. Cheng Q, Cloonan N, Fischer K, Thompson J, Waine G *et al*. stevor and rif are Plasmodium falciparum multicopy gene families which potentially encode variant antigens. *Mol. Biochem. Parasitol.* **97**, 161-176 (1998).
84. Kyes SA, Rowe JA, Kriek N & Newbold CI. Rifins: a second family of clonally variant proteins expressed on the surface of red cells infected with Plasmodium falciparum. *Proc. Natl. Acad. Sci. U. S. A* **96**, 9333-9338 (1999).
85. Figueiredo LM, Freitas-Junior LH, Bottius E, Olivo-Marin JC & Scherf A. A central role for Plasmodium falciparum subtelomeric regions in spatial positioning and telomere length regulation. *EMBO J.* **21**, 815-824 (2002).
86. Scherf A, Hernandez-Rivas R, Buffet P, Bottius E, Benatar C *et al*. Antigenic variation in malaria: in situ switching, relaxed and mutually exclusive transcription of var genes during intra-erythrocytic development in Plasmodium falciparum. *EMBO J.* **17**, 5418-5426 (1998).
87. Kooij TW, Carlton JM, Bidwell SL, Hall N, Ramesar J *et al*. A Plasmodium Whole-Genome Synteny Map: Indels and Synteny Breakpoints as Foci for Species-Specific Genes. *PLoS Pathog.* **1**, e44 (2005).
88. Thompson J, Janse CJ & Waters AP. Comparative genomics in Plasmodium: a tool for the identification of genes and functional analysis. *Mol. Biochem. Parasitol.* **118**, 147-154 (2001).
89. Hayward RE, Derisi JL, Alfadhli S, Kaslow DC, Brown PO *et al*. Shotgun DNA microarrays and stage-specific gene expression in Plasmodium falciparum malaria. *Mol. Microbiol.* **35**, 6-14 (2000).
90. Mamoun CB, Gluzman IY, Hott C, MacMillan SK, Amarakone AS *et al*. Co-ordinated programme of gene expression during asexual intraerythrocytic development of the human malaria parasite Plasmodium falciparum revealed by microarray analysis. *Mol. Microbiol.* **39**, 26-36 (2001).
91. Le Roch KG, Zhou Y, Blair PL, Grainger M, Moch JK *et al*. Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science* **301**, 1503-1508 (2003).
92. Bozdech Z, Llinas M, Pulliam BL, Wong ED, Zhu J *et al*. The Transcriptome of the Intraerythrocytic Developmental Cycle of Plasmodium falciparum. *PLoS. Biol.* **1**, E5 (2003).
93. Li L, Brunk BP, Kissinger JC, Pape D, Tang K *et al*. Gene discovery in the apicomplexa as revealed by EST sequencing and assembly of a comparative gene database. *Genome Res.* **13**, 443-454 (2003).
94. Li L, Crabtree J, Fischer S, Pinney D, Stoeckert CJ, Jr. *et al*. ApiEST-DB: analyzing clustered EST data of the apicomplexan parasites. *Nucleic Acids Res.* **32 Database issue**, D326-D328 (2004).
95. Watanabe J, Sasaki M, Suzuki Y & Sugano S. Analysis of transcriptomes of human malaria parasite Plasmodium falciparum using full-length enriched library: identification of novel genes and diverse transcription start sites of messenger RNAs. *Gene* **291**, 105-113 (2002).
96. Watanabe J, Suzuki Y, Sasaki M & Sugano S. Full-malaria 2004: an enlarged database for comparative studies of full-length cDNAs of malaria parasites, Plasmodium species. *Nucleic Acids Res.* **32 Database issue**, D334-D338 (2004).
97. Cui L, Rzomp KA, Fan Q, Martin SK & Williams J. Plasmodium falciparum: Differential Display Analysis of Gene Expression during Gametocytogenesis. *Exp. Parasitol.* **99**, 244-254 (2001).
98. Fidock DA, Nguyen TV, Ribeiro JM, Valenzuela JG & James AA. Plasmodium falciparum: generation of a cDNA library enriched in sporozoite-specific transcripts by directional tag subtractive hybridization. *Exp. Parasitol.* **95**, 220-225 (2000).

99. Munasinghe A, Patankar S, Cook BP, Madden SL, Martin RK *et al*. Serial analysis of gene expression (SAGE) in Plasmodium falciparum: application of the technique to A-T rich genomes. *Mol. Biochem. Parasitol.* **113**, 23-34 (2001).
100. Patankar S, Munasinghe A, Shoaibi A, Cummings LM & Wirth DF. Serial analysis of gene expression in Plasmodium falciparum reveals the global expression profile of erythrocytic stages and the presence of anti-sense transcripts in the malarial parasite. *Mol. Biol. Cell* **12**, 3114-3125 (2001).
101. Le Roch KG, Johnson JR, Florens L, Zhou Y, Santrosyan A *et al*. Global analysis of transcript and protein levels across the Plasmodium falciparum life cycle. *Genome Res.* **14**, 2308-2318 (2004).
102. Bozdech Z & Ginsburg H. Antioxidant defense in Plasmodium falciparum--data mining of the transcriptome. *Malar. J.* **3**, 23 (2004).
103. Bozdech Z & Ginsburg H. Data mining of the transcriptome of Plasmodium falciparum: the pentose phosphate pathway and ancillary processes. *Malar. J.* **4**, 17 (2005).
104. Ralph SA, Bischoff E, Mattei D, Sismeiro O, Dillies MA *et al*. Transcriptome analysis of antigenic variation in Plasmodium falciparum--var silencing is not dependent on antisense RNA. *Genome Biol.* **6**, R93 (2005).
105. Gissot M, Refour P, Briquet S, Boschet C, Coupe S *et al*. Transcriptome of 3D7 and its gametocyte-less derivative F12 Plasmodium falciparum clones during erythrocytic development using a gene-specific microarray assigned to gene regulation, cell cycle and transcription factors. *Gene* **341**, 267-277 (2004).
106. Silvestrini F, Bozdech Z, Lanfrancotti A, Di Giulio E, Bultrini E *et al*. Genome-wide identification of genes upregulated at the onset of gametocytogenesis in Plasmodium falciparum. *Mol. Biochem. Parasitol.* **143**, 100-110 (2005).
107. Young JA, Fivelman QL, Blair PL, De L, V, Le Roch KG *et al*. The Plasmodium falciparum sexual development transcriptome: a microarray analysis using ontology-based pattern identification. *Mol. Biochem. Parasitol.* **143**, 67-79 (2005).
108. Dessens JT, Margos G, Rodriguez MC & Sinden RE. Identification of differentially regulated genes of Plasmodium by suppression subtractive hybridization. *Parasitol. Today* **16**, 354-356 (2000).
109. Pradel G, Hayton K, Aravind L, Iyer LM, Abrahamsen MS *et al*. A Multidomain Adhesion Protein Family Expressed in Plasmodium falciparum Is Essential for Transmission to the Mosquito. *J. Exp. Med.* **199**, 1533-1544 (2004).
110. Vincensini L, Richert S, Blisnick T, Van Dorsselaer A, Leize-Wagner E *et al*. Proteomic analysis identifies novel proteins of the maurer's clefts, a secretory compartment delivering plasmodium falciparum proteins to the surface of its host cell. *Mol. Cell Proteomics.* (2005).
111. Florens L, Liu X, Wang Y, Yang S, Schwartz O *et al*. Proteomics approach reveals novel proteins on the surface of malaria-infected erythrocytes. *Mol. Biochem. Parasitol.* **135**, 1-11 (2004).
112. Przyborski JM, Wickert H, Krohne G & Lanzer M. Maurer's clefts--a novel secretory organelle? *Mol. Biochem. Parasitol.* **132**, 17-26 (2003).
113. Sanders PR, Gilson PR, Cantin GT, Greenbaum DC, Nebl T *et al*. Distinct protein classes including novel merozoite surface antigens in raft-like membranes of Plasmodium falciparum. *J. Biol. Chem.* (2005).
114. LaCount DJ, Vignali M, Chettier R, Phansalkar A, Bell R *et al*. A protein interaction network of the malaria parasite Plasmodium falciparum. *Nature* **438**, 103-107 (2005).
115. Suthram S, Sittler T & Ideker T. The Plasmodium protein network diverges from those of other eukaryotes. *Nature* **438**, 108-112 (2005).
116. Marti M, Good RT, Rug M, Knuepfer E & Cowman AF. Targeting malaria virulence and remodeling proteins to the host erythrocyte. *Science* **306**, 1930-1933 (2004).
117. Hiller NL, Bhattacharjee S, van Ooij C, Liolios K, Harrison T *et al*. A host-targeting signal in virulence proteins reveals a secretome in malarial infection. *Science* **306**, 1934-1937 (2004).
118. Martin RE, Henry RI, Abbey JL, Clements JD & Kirk K. The 'permeome' of the malaria parasite: an overview of the membrane transport proteins of Plasmodium falciparum. *Genome Biol.* **6**, R26 (2005).

119. Anderson TJ, Su XZ, Roddam A & Day KP. Complex mutations in a high proportion of microsatellite loci from the protozoan parasite Plasmodium falciparum. *Mol. Ecol.* **9**, 1599-1608 (2000).
120. Fidock DA, Nomura T, Talley AK, Cooper RA, Dzekunov SM *et al*. Mutations in the P. falciparum digestive vacuole transmembrane protein PfCRT and evidence for their role in chloroquine resistance. *Mol. Cell* **6**, 861-871 (2000).
121. Su X, Kirkman LA, Fujioka H & Wellems TE. Complex polymorphisms in an approximately 330 kDa protein are linked to chloroquine-resistant P. falciparum in Southeast Asia and Africa. *Cell* **91**, 593-603 (1997).
122. Mu J, Ferdig MT, Feng X, Joy DA, Duan J *et al*. Multiple transporters associated with malaria parasite responses to chloroquine and quinine. *Mol. Microbiol.* **49**, 977-989 (2003).
123. Mu J, Awadalla P, Duan J, McGee KM, Joy DA *et al*. Recombination hotspots and population structure in Plasmodium falciparum. *PLoS Biol.* **3**, e335 (2005).
124. McConkey GA, Pinney JW, Westhead DR, Plueckhahn K, Fitzpatrick TB *et al*. Annotating the Plasmodium genome and the enigma of the shikimate pathway. *Trends Parasitol.* **20**, 60-65 (2004).
125. Waters AP, Higgins DG & McCutchan TF. Plasmodium falciparum appears to have arisen as a result of lateral transfer between avian and human hosts. *Proc. Natl. Acad. Sci. U. S. A* **88**, 3140-3144 (1991).
126. Escalante AA & Ayala FJ. Phylogeny of the malarial genus Plasmodium, derived from rRNA gene sequences. *Proc. Natl. Acad. Sci. U. S. A* **91**, 11373-11377 (1994).
127. Carlton J. The Plasmodium vivax genome sequencing project. *Trends Parasitol.* **19**, 227-231 (2003).
128. Templeton TJ, Iyer LM, Anantharaman V, Enomoto S, Abrahante JE *et al*. Comparative analysis of apicomplexa and genomic diversity in eukaryotes. *Genome Res.* **14**, 1686-1695 (2004).
129. Balaji S, Babu MM, Iyer LM & Aravind L. Discovery of the principal specific transcription factors of Apicomplexa and their implication for the evolution of the AP2-integrase DNA binding domains. *Nucleic Acids Res.* **33**, 3994-4006 (2005).
130. Carlton JM, Galinski MR, Barnwell JW & Dame JB. Karyotype and synteny among the chromosomes of all four species of human malaria parasite. *Mol. Biochem. Parasitol.* **101**, 23-32 (1999).
131. 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).
132. van Dijk MR, Janse CJ, Thompson J, Waters AP, Braks JA *et al*. A central role for P48/45 in malaria parasite male gamete fertility. *Cell* **104**, 153-164 (2001).
133. Menard R. Gliding motility and cell invasion by Apicomplexa: insights from the Plasmodium sporozoite. *Cell Microbiol.* **3**, 63-73 (2001).
134. Yoshida N, Nussenzweig RS, Potocnjak P, Nussenzweig V & Aikawa M. Hybridoma produces protective antibodies directed against the sporozoite stage of malaria parasite. *Science* **207**, 71-73 (1980).
135. Robson KJ, Hall JR, Jennings MW, Harris TJ, Marsh K *et al*. A highly conserved amino-acid sequence in thrombospondin, properdin and in proteins from sporozoites and blood stages of a human malaria parasite. *Nature* **335**, 79-82 (1988).
136. Guerin-Marchand C, Druilhe P, Galey B, Londono A, Patarapotikul J *et al*. A liver-stage-specific antigen of Plasmodium falciparum characterized by gene cloning. *Nature* **329**, 164-167 (1987).
137. Hall R, Hyde JE, Goman M, Simmons DL, Hope IA *et al*. Major surface antigen gene of a human malaria parasite cloned and expressed in bacteria. *Nature* **311**, 379-382 (1984).
138. Peterson MG, Marshall VM, Smythe JA, Crewther PE, Lew A *et al*. Integral membrane protein located in the apical complex of Plasmodium falciparum. *Mol. Cell Biol.* **9**, 3151-3154 (1989).
139. Smythe JA, Coppel RL, Brown GV, Ramasamy R, Kemp DJ *et al*. Identification of two integral membrane proteins of Plasmodium falciparum. *Proc. Natl. Acad. Sci. U. S. A* **85**, 5195-5199 (1988).
140. Quakyi IA, Carter R, Rener J, Kumar N, Good MF *et al*. The 230-kDa gamete surface protein of Plasmodium falciparum is also a target for transmission-blocking antibodies. *J. Immunol.* **139**, 4213-4217 (1987).

141. Hisaeda H, Stowers AW, Tsuboi T, Collins WE, Sattabongkot JS *et al*. Antibodies to malaria vaccine candidates Pvs25 and Pvs28 completely block the ability of Plasmodium vivax to infect mosquitoes. *Infect. Immun.* **68**, 6618-6623 (2000).
142. del Carmen RM, Gerold P, Dessens J, Kurtenbach K, Schwartz RT *et al*. Characterisation and expression of pbs25, a sexual and sporogonic stage specific protein of Plasmodium berghei. *Mol. Biochem. Parasitol.* **110**, 147-159 (2000).
143. Siden-Kiamos I, Vlachou D, Margos G, Beetsma A, Waters AP *et al*. Distinct roles for pbs21 and pbs25 in the in vitro ookinete to oocyst transformation of Plasmodium berghei. *J. Cell Sci.* **113 Pt 19**, 3419-3426 (2000).
144. del Portillo HA, Fernandez-Becerra C, Bowman S, Oliver K, Preuss M *et al*. A superfamily of variant genes encoded in the subtelomeric region of Plasmodium vivax. *Nature* **410**, 839-842 (2001).
145. Janssen CS, Phillips RS, Turner CM & Barrett MP. Plasmodium interspersed repeats: the major multigene superfamily of malaria parasites. *Nucleic Acids Res.* **32**, 5712-5720 (2004).
146. Sam-Yellowe TY, Florens L, Johnson JR, Wang T, Drazba JA *et al*. A Plasmodium Gene Family Encoding Maurer's Cleft Membrane Proteins: Structural Properties and Expression Profiling. *Genome Res.* **14**, 1052-1059 (2004).
147. Merino EF, Fernandez-Becerra C, Madeira AM, Machado AL, Durham A *et al*. Pilot survey of expressed sequence tags (ESTs) from the asexual blood stages of Plasmodium vivax in human patients. *Malar. J.* **2**, 21 (2003).
148. Kappe SH, Gardner MJ, Brown SM, Ross J, Matuschewski K *et al*. Exploring the transcriptome of the malaria sporozoite stage. *Proc. Natl. Acad. Sci. U. S. A* **98**, 9895-9900 (2001).
149. Kaiser K, Matuschewski K, Camargo N, Ross J & Kappe SH. Differential transcriptome profiling identifies Plasmodium genes encoding pre-erythrocytic stage-specific proteins. *Mol. Microbiol.* **51**, 1221-1232 (2004).
150. Abraham EG, Islam S, Srinivasan P, Ghosh AK, Valenzuela JG *et al*. Analysis of the Plasmodium and Anopheles transcriptional repertoire during ookinete development and midgut invasion. *J. Biol. Chem.* **279**, 5573-5580 (2004).
151. Srinivasan P, Abraham EG, Ghosh AK, Valenzuela J, Ribeiro JM *et al*. Analysis of the Plasmodium and Anopheles transcriptomes during oocyst differentiation. *J. Biol. Chem.* **279**, 5581-5587 (2004).
152. Matuschewski K, Ross J, Brown SM, Kaiser K, Nussenzweig V *et al*. Infectivity-associated changes in the transcriptional repertoire of the malaria parasite sporozoite stage. *J. Biol. Chem.* **277**, 41948-41953 (2002).
153. Hayward RE. Plasmodium falciparum phosphoenolpyruvate carboxykinase is developmentally regulated in gametocytes. *Mol. Biochem. Parasitol.* **107**, 227-240 (2000).
154. Khan SM, Franke-Fayard B, Mair GR, Lasonder E, Janse CJ *et al*. Proteome analysis of separated male and female gametocytes reveals novel sex-specific Plasmodium biology. *Cell* **121**, 675-687 (2005).
155. Kimura M. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* **267**, 275-276 (1977).
156. Kafatos FC, Efstratiadis A, Forget BG & Weissman SM. Molecular evolution of human and rabbit beta-globin mRNAs. *Proc. Natl. Acad. Sci. U. S. A* **74**, 5618-5622 (1977).
157. Plotkin JB, Dushoff J & Fraser HB. Detecting selection using a single genome sequence of M. tuberculosis and P. falciparum. *Nature* **428**, 942-945 (2004).
158. Friedman R & Hughes AL. Codon volatility as an indicator of positive selection: data from eukaryotic genome comparisons. *Mol. Biol. Evol.* **22**, 542-546 (2005).
159. Sharp PM. Gene "volatility" is most unlikely to reveal adaptation. *Mol. Biol. Evol.* **22**, 807-809 (2005).
160. Peixoto L, Fernandez V & Musto H. The effect of expression levels on codon usage in Plasmodium falciparum. *Parasitology* **128**, 245-251 (2004).
161. Doolan DL, Southwood S, Freilich DA, Sidney J, Graber NL *et al*. Identification of Plasmodium falciparum antigens by antigenic analysis of genomic and proteomic data. *Proc. Natl. Acad. Sci. U. S. A* **100**, 9952-9957 (2003).

162. Martinelli A, Cheesman S, Hunt P, Culleton R, Raza A *et al*. A genetic approach to the de novo identification of targets of strain-specific immunity in malaria parasites. *Proc. Natl. Acad. Sci. U. S. A* (2005).
163. Haddad D, Bilcikova E, Witney AA, Carlton JM, White CE *et al*. Novel antigen identification method for discovery of protective malaria antigens by rapid testing of DNA vaccines encoding exons from the parasite genome. *Infect. Immun.* **72**, 1594-1602 (2004).
164. Coulson RM, Hall N & Ouzounis CA. Comparative genomics of transcriptional control in the human malaria parasite Plasmodium falciparum. *Genome Res.* **14**, 1548-1554 (2004).
165. Bahl A, Brunk B, Crabtree J, Fraunholz MJ, Gajria B *et al*. PlasmoDB: the Plasmodium genome resource. A database integrating experimental and computational data. *Nucleic Acids Res.* **31**, 212-215 (2003).
166. Kissinger JC, Brunk BP, Crabtree J, Fraunholz MJ, Gajria B *et al*. The Plasmodium genome database. *Nature* **419**, 490-492 (2002).
167. Paton MG, Barker GC, Matsuoka H, Ramesar J, Janse CJ *et al*. Structure and expression of a post-transcriptionally regulated malaria gene encoding a surface protein from the sexual stages of Plasmodium berghei. *Mol. Biochem. Parasitol.* **59**, 263-275 (1993).
168. Cui L, Fan Q & Li J. The malaria parasite Plasmodium falciparum encodes members of the Puf RNA-binding protein family with conserved RNA binding activity. *Nucleic Acids Res.* **30**, 4607-4617 (2002).
169. Vervenne RA, Dirks RW, Ramesar J, Waters AP & Janse CJ. Differential expression in blood stages of the gene coding for the 21-kilodalton surface protein of ookinetes of Plasmodium berghei as detected by RNA in situ hybridisation. *Mol. Biochem. Parasitol.* **68**, 259-266 (1994).
170. Kooij TW, Franke-Fayard B, Renz J, Kroeze H, van Dooren MW *et al*. Plasmodium berghei alpha-tubulin II: a role in both male gamete formation and asexual blood stages. *Mol. Biochem. Parasitol.* **144**, 16-26 (2005).
171. Waters AP, van Spaendonk RM, Ramesar J, Vervenne RA, Dirks RW *et al*. Species-specific regulation and switching of transcription between stage-specific ribosomal RNA genes in Plasmodium berghei. *J. Biol. Chem.* **272**, 3583-3589 (1997).
172. Janse CJ, Ramesar J, van den Berg FM & Mons B. Plasmodium berghei: in vivo generation and selection of karyotype mutants and non-gametocyte producer mutants. *Exp. Parasitol.* **74**, 1-10 (1992).
173. Su XZ, Wu Y, Sifri CD & Wellems TE. Reduced extension temperatures required for PCR amplification of extremely A+T-rich DNA. *Nucleic Acids Res.* **24**, 1574-1575 (1996).
174. Carter R & Diggs CL. **In:** Parasitic Protozoa - p359-465. Academic Press, New York/San Fransisco/London (1977).
175. van Dijk MR, Waters AP & Janse CJ. Stable transfection of malaria parasite blood stages. *Science* **268**, 1358-1362 (1995).
176. Carlton JM & Carucci DJ. Rodent models of malaria in the genomics era. *Trends Parasitol.* **18**, 100-102 (2002).
177. Weinbaum FI, Evans CB & Tigelaar RE. An in vitro assay for T cell immunity to malaria in mice. *J. Immunol.* **116**, 1280-1283 (1976).
178. Landau I & Chabaud AG. [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).
179. Sutton GG, White OR, Adams MD & Kerlavage AR. TIGR Assembler: A new tool for assembling large shotgun sequencing projects. *Genome Science & Technology* **1**, 9-19 (1995).
180. Mural RJ, Adams MD, Myers EW, Smith HO, Miklos GL *et al*. A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science* **296**, 1661-1671 (2002).
181. Beetsma AL, van de Wiel TJ, Sauerwein RW & Eling WM. Plasmodium berghei ANKA: purification of large numbers of infectious gametocytes. *Exp. Parasitol.* **88**, 69-72 (1998).
182. Gardner MJ, Shallom SJ, Carlton JM, Salzberg SL, Nene V *et al*. Sequence of Plasmodium falciparum chromosomes 2, 10, 11 and 14. *Nature* **419**, 531-534 (2002).
183. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L *et al*. The Pfam protein families database. *Nucleic Acids Res.* **30**, 276-280 (2002).

184. Haft DH, Loftus BJ, Richardson DL, Yang F, Eisen JA *et al*. TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Res.* **29**, 41-43 (2001).
185. Eddy SR. Profile hidden Markov models. *Bioinformatics.* **14**, 755-763 (1998).
186. Delcher AL, Phillippy A, Carlton J & Salzberg SL. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.* **30**, 2478-2483 (2002).
187. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573-580 (1999).
188. Ogurtsov AY, Roytberg MA, Shabalina SA & Kondrashov AS. OWEN: aligning long collinear regions of genomes. *Bioinformatics.* **18**, 1703-1704 (2002).
189. Thompson JD, Higgins DG & Gibson TJ. 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).
190. Nei M & Gojobori T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**, 418-426 (1986).
191. Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP *et al*. A whole-genome assembly of Drosophila. *Science* **287**, 2196-2204 (2000).
192. Lander ES & Waterman MS. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2**, 231-239 (1988).
193. McCutchan TF, Dame JB, Miller LH & Barnwell J. Evolutionary relatedness of Plasmodium species as determined by the structure of DNA. *Science* **225**, 808-811 (1984).
194. Daly TM, Long CA & Bergman LW. 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).
195. Quackenbush J, Cho J, Lee D, Liang F, Holt I *et al*. The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Res.* **29**, 159-164 (2001).
196. Cawley SE, Wirth AI & Speed TP. Phat-a gene finding program for Plasmodium falciparum. *Mol. Biochem. Parasitol.* **118**, 167-174 (2001).
197. Salzberg SL, Pertea M, Delcher AL, Gardner MJ & Tettelin H. Interpolated Markov models for eukaryotic gene finding. *Genomics* **59**, 24-31 (1999).
198. Ohkanda J, Lockman JW, Yokoyama K, Gelb MH, Croft SL *et al*. Peptidomimetic inhibitors of protein farnesyltransferase show potent antimalarial activity. *Bioorg. Med. Chem. Lett.* **11**, 761-764 (2001).
199. The GO Consortium. Creating the gene ontology resource: design and implementation. *Genome Res.* **11**, 1425-1433 (2001).
200. Black CG, Wang L, Hibbs AR, Werner E & Coppel RL. Identification of the Plasmodium chabaudi homologue of merozoite surface proteins 4 and 5 of Plasmodium falciparum. *Infect. Immun.* **67**, 2075-2081 (1999).
201. Cooke BM, Mohandas N & Coppel RL. The malaria-infected red blood cell: structural and functional changes. *Adv. Parasitol.* **50**, 1-86 (2001).
202. Janssen CS, Barrett MP, Turner CM & Phillips RS. A large gene family for putative variant antigens shared by human and rodent malaria parasites. *Proc. R. Soc. Lond B Biol. Sci.* **269**, 431-436 (2002).
203. Cunningham DA, Jarra W, Koernig S, Fonager J, Fernandez-Reyes D *et al*. Host immunity modulates transcriptional changes in a multigene family (yir) of rodent malaria. *Mol. Microbiol.* **58**, 636-647 (2005).
204. Strimmer K & von Haeseler A. Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. *Molecular Biological Evolution* **13**, 964-969 (1996).
205. Preiser PR, Jarra W, Capiod T & Snounou G. A rhoptry-protein-associated mechanism of clonal phenotypic variation in rodent malaria. *Nature* **398**, 618-622 (1999).
206. Galinski MR, Xu M & Barnwell JW. 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).
207. Rayner JC, Galinski MR, Ingravallo P & Barnwell JW. 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. U. S. A* **97**, 9648-9653 (2000).

208. Wiser MF, Giraldo LE, Schmitt-Wrede HP & Wunderlich F. Plasmodium chabaudi: immunogenicity of a highly antigenic glutamate-rich protein. *Exp. Parasitol.* **85**, 43-54 (1997).
209. Spielmann T & Beck HP. 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).
210. Favaloro JM, Culvenor JG, Anders RF & Kemp DJ. A Plasmodium chabaudi antigen located in the parasitophorous vacuole membrane. *Mol. Biochem. Parasitol.* **62**, 263-270 (1993).
211. Mu J, Duan J, Makova KD, Joy DA, Huynh CQ *et al*. Chromosome-wide SNPs reveal an ancient origin for Plasmodium falciparum. *Nature* **418**, 323-326 (2002).
212. Garnham PCC. Malaria Parasites and Other Haemosporidia. Blackwell Scientific, Oxford (1966).
213. Liu SL & Sanderson KE. Rearrangements in the genome of the bacterium Salmonella typhi. *Proc. Natl. Acad. Sci. U. S. A* **92**, 1018-1022 (1995).
214. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25-29 (2000).
215. Dame JB & McCutchan TF. 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).
216. 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).
217. Shabalina SA, Ogurtsov AY, Kondrashov VA & Kondrashov AS. Selective constraint in intergenic regions of human and mouse genomes. *Trends Genet.* **17**, 373-376 (2001).
218. Makalowski W & Boguski MS. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. *Proc. Natl. Acad. Sci. U. S. A* **95**, 9407-9412 (1998).
219. Carlton JM, Fidock DA, Djimde A, Plowe CV & Wellems TE. 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).
220. Sinden RE. **In:** Rodent Malaria - p85-168. Killick-Kendrick R & Peters W (eds.) Academic Press, London (1978).
221. Janse CJ & Waters AP. Plasmodium berghei: The application of cultivation and purification techniques to molecular studies of malaria parasites. *Parasitol. Today* **11**, 138-143 (1995).
222. Hall N, Pain A, Berriman M, Churcher C, Harris B *et al*. Sequence of Plasmodium falciparum chromosomes 1, 3-9 and 13. *Nature* **419**, 527-531 (2002).
223. Mullikin JC & Ning Z. The phusion assembler. *Genome Res.* **13**, 81-90 (2003).
224. Birney E & Durbin R. Using GeneWise in the Drosophila annotation experiment. *Genome Res.* **10**, 547-548 (2000).
225. Enright AJ, Van Dongen S & Ouzounis CA. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575-1584 (2002).
226. Rice P, Longden I & Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276-277 (2000).
227. Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**, 555-556 (1997).
228. Janse CJ, Boorsma EG, Ramesar J, van Vianen P, van der MR *et al*. Plasmodium berghei: gametocyte production, DNA content, and chromosome-size polymorphisms during asexual multiplication in vivo. *Exp. Parasitol.* **68**, 274-282 (1989).
229. Dimopoulos G, Christophides GK, Meister S, Schultz J, White KP *et al*. Genome expression analysis of Anopheles gambiae: responses to injury, bacterial challenge, and malaria infection. *Proc. Natl. Acad. Sci. U. S. A* **99**, 8814-8819 (2002).
230. Eisen MB, Spellman PT, Brown PO & Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A* **95**, 14863-14868 (1998).
231. Sinden RE, Butcher GA & Beetsma AL. Maintenance of the Plasmodium berghei life cycle. *Methods Mol. Med.* **72**, 25-40 (2002).
232. Wu CC, MacCoss MJ, Howell KE & Yates JR, III. A method for the comprehensive proteomic analysis of membrane proteins. *Nat. Biotechnol.* **21**, 532-538 (2003).
233. Eng JK, McCormack AL & Yates JR, III. *Journal of the American Society of Mass Spectrometry* **5**, 976 (1994).

234. Sadygov RG & Yates JR, III. A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. *Anal. Chem.* **75**, 3792-3798 (2003).
235. Fischer K, Chavchich M, Huestis R, Wilson DW, Kemp DJ *et al*. Ten families of variant genes encoded in subtelomeric regions of multiple chromosomes of Plasmodium chabaudi, a malaria species that undergoes antigenic variation in the laboratory mouse. *Mol. Microbiol.* **48**, 1209-1223 (2003).
236. Endo T, Ikeo K & Gojobori T. Large-scale search for genes on which positive selection may operate. *Mol. Biol. Evol.* **13**, 685-690 (1996).
237. Han YS, Thompson J, Kafatos FC & Barillas-Mury C. Molecular interactions between Anopheles stephensi midgut cells and Plasmodium berghei: the time bomb theory of ookinete invasion of mosquitoes. *EMBO J.* **19**, 6030-6040 (2000).
238. Washburn MP, Wolters D & Yates JR, III. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19**, 242-247 (2001).
239. Vickerman K. Polymorphism and mitochondrial activity in sleeping sickness trypanosomes. *Nature* **208**, 762-766 (1965).
240. Krungkrai J, Prapunwattana P & Krungkrai SR. Ultrastructure and function of mitochondria in gametocytic stage of Plasmodium falciparum. *Parasite* **7**, 19-26 (2000).
241. Kaiser K, Camargo N, Coppens I, Morrisey JM, Vaidya AB *et al*. A member of a conserved Plasmodium protein family with membrane-attack complex/perforin (MACPF)-like domains localizes to the micronemes of sporozoites. *Mol. Biochem. Parasitol.* **133**, 15-26 (2004).
242. Mota MM, Hafalla JC & Rodriguez A. Migration through host cells activates Plasmodium sporozoites for infection. *Nat. Med.* **8**, 1318-1322 (2002).
243. Crooks GE, Hon G, Chandonia JM & Brenner SE. WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188-1190 (2004).
244. Waters AP. Parasitology. Guilty until proven otherwise. *Science* **301**, 1487-1488 (2003).
245. Bailey TL & Gribskov M. Methods and statistics for combining motif match scores. *J. Comput. Biol.* **5**, 211-221 (1998).
246. Hillier LW, Miller W, Birney E, Warren W, Hardison RC *et al*. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**, 695-716 (2004).
247. Murphy WJ, Larkin DM, Everts-van der Wind A, Bourque G, Tesler G *et al*. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science* **309**, 613-617 (2005).
248. Severson DW, DeBruyn B, Lovin DD, Brown SE, Knudson DL *et al*. Comparative genome analysis of the yellow fever mosquito Aedes aegypti with Drosophila melanogaster and the malaria vector mosquito Anopheles gambiae. *J. Hered.* **95**, 103-113 (2004).
249. Sharakhov IV, Serazin AC, Grushko OG, Dana A, Lobo N *et al*. Inversions and gene order shuffling in Anopheles gambiae and A. funestus. *Science* **298**, 182-185 (2002).
250. El Sayed NM, Myler PJ, Blandin G, Berriman M, Crabtree J *et al*. Comparative genomics of trypanosomatid parasitic protozoa. *Science* **309**, 404-409 (2005).
251. Kyes S, Horrocks P & Newbold C. Antigenic variation at the infected red cell surface in malaria. *Annu. Rev. Microbiol.* **55**, 673-707 (2001).
252. Barry JD, Ginger ML, Burton P & McCulloch R. Why are parasite contingency genes often associated with telomeres? *Int. J. Parasitol.* **33**, 29-45 (2003).
253. Bailey TL & Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2**, 28-36 (1994).
254. Schultz J, Milpetz F, Bork P & Ponting CP. SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl. Acad. Sci. U. S. A* **95**, 5857-5864 (1998).
255. Felsenstein J. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods Enzymol.* **266**, 418-427 (1996).
256. Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**, 783-791 (1985).
257. Day WH & McMorris FR. A consensus program for molecular sequences. *Comput. Appl. Biosci.* **9**, 653-656 (1993).

258. Perkins ME. Surface proteins of Plasmodium falciparum merozoites binding to the erythrocyte receptor, glycophorin. *J. Exp. Med.* **160**, 788-798 (1984).
259. Spielmann T, Fergusen DJ & Beck HP. etramps, a new Plasmodium falciparum gene family coding for developmentally regulated and highly charged membrane proteins located at the parasite-host cell interface. *Mol. Biol. Cell* **14**, 1529-1544 (2003).
260. Pearce JA, Mills K, Triglia T, Cowman AF & Anders RF. Characterisation of two novel proteins from the asexual stage of Plasmodium falciparum, H101 and H103. *Mol. Biochem. Parasitol.* **139**, 141-151 (2005).
261. Anamika, Srinivasan N & Krupa A. A genomic perspective of protein kinases in Plasmodium falciparum. *Proteins* **58**, 180-189 (2005).
262. Schneider AG & Mercereau-Puijalon O. A new Apicomplexa-specific protein kinase family : multiple members in Plasmodium falciparum, all with an export signature. *BMC. Genomics* **6**, 30 (2005).
263. Ward P, Equinet L, Packer J & Doerig C. Protein kinases of the human malaria parasite Plasmodium falciparum: the kinome of a divergent eukaryote. *BMC. Genomics* **5**, 79 (2004).
264. Nadeau JH & Sankoff D. Landmarks in the Rosetta Stone of mammalian comparative maps. *Nat. Genet.* **15**, 6-7 (1997).
265. Escalante AA, Barrio E & Ayala FJ. Evolutionary origin of human and primate malarias: evidence from the circumsporozoite protein gene. *Mol. Biol. Evol.* **12**, 616-626 (1995).
266. McCutchan TF, Kissinger JC, Touray MG, Rogers MJ, Li J *et al*. Comparison of circumsporozoite proteins from avian and mammalian malarias: biological and phylogenetic implications. *Proc. Natl. Acad. Sci. U. S. A* **93**, 11889-11894 (1996).
267. Rathore D, Wahl AM, Sullivan M & McCutchan TF. A phylogenetic comparison of gene trees constructed from plastid, mitochondrial and genomic DNA of Plasmodium species. *Mol. Biochem. Parasitol.* **114**, 89-94 (2001).
268. Kissinger JC, Souza PC, Soarest CO, Paul R, Wahl AM *et al*. Molecular phylogenetic analysis of the avian malarial parasite Plasmodium (Novyella) juxtanucleare. *J. Parasitol.* **88**, 769-773 (2002).
269. Carlton JM. Gene synteny across Plasmodium spp: could 'operon-like' structures exist? *Parasitol. Today* **15**, 178-179 (1999).
270. Bailey JA, Baertsch R, Kent WJ, Haussler D & Eichler EE. Hotspots of mammalian chromosomal evolution. *Genome Biol.* **5**, R23 (2004).
271. Armengol L, Pujana MA, Cheung J, Scherer SW & Estivill X. Enrichment of segmental duplications in regions of breaks of synteny between the human and mouse genomes suggest their involvement in evolutionary rearrangements. *Hum. Mol. Genet.* **12**, 2201-2208 (2003).
272. Kraemer SM & Smith JD. Evidence for the importance of genetic structuring to the structural and functional specialization of the Plasmodium falciparum var gene family. *Mol. Microbiol.* **50**, 1527-1538 (2003).
273. Hyams JS & Lloyd CW. Microtubules. Wiley-Liss, New York (1994).
274. Luduena RF. Multiple forms of tubulin: different gene products and covalent modifications. *Int. Rev. Cytol.* **178**, 207-275 (1998).
275. Little M & Seehaus T. Comparative analysis of tubulin sequences. *Comp Biochem. Physiol B* **90**, 655-670 (1988).
276. van Belkum A, Janse C & Mons B. Nucleotide sequence variation in the beta-tubulin genes from Plasmodium berghei and Plasmodium falciparum. *Mol. Biochem. Parasitol.* **47**, 251-254 (1991).
277. Holloway SP, Gerousis M, Delves CJ, Sims PF, Scaife JG *et al*. The tubulin genes of the human malaria parasite Plasmodium falciparum, their chromosomal location and sequence analysis of the alpha-tubulin II gene. *Mol. Biochem. Parasitol.* **43**, 257-270 (1990).
278. Sen K & Godson GN. Isolation of alpha- and beta-tubulin genes of Plasmodium falciparum using a single oligonucleotide probe. *Mol. Biochem. Parasitol.* **39**, 173-182 (1990).
279. Holloway SP, Sims PF, Delves CJ, Scaife JG & Hyde JE. Isolation of alpha-tubulin genes from the human malaria parasite, Plasmodium falciparum: sequence analysis of alpha-tubulin. *Mol. Microbiol.* **3**, 1501-1510 (1989).
280. Wesseling JG, Dirks R, Smits MA & Schoenmakers JG. Nucleotide sequence and expression of a beta-tubulin gene from Plasmodium falciparum, a malarial parasite of man. *Gene* **83**, 301-309 (1989).

281. Delves CJ, Ridley RG, Goman M, Holloway SP, Hyde JE *et al*. Cloning of a beta-tubulin gene from Plasmodium falciparum. *Mol. Microbiol.* **3**, 1511-1519 (1989).
282. Akella R, Arasu P & Vaidya AB. Molecular clones of alpha-tubulin genes of Plasmodium yoelii reveal an unusual feature of the carboxy terminus. *Mol. Biochem. Parasitol.* **30**, 165-174 (1988).
283. Rawlings DJ, Fujioka H, Fried M, Keister DB, Aikawa M *et al*. Alpha-tubulin II is a male-specific protein in Plasmodium falciparum. *Mol. Biochem. Parasitol.* **56**, 239-250 (1992).
284. Delves CJ, Alano P, Ridley RG, Goman M, Holloway SP *et al*. Expression of alpha and beta tubulin genes during the asexual and sexual blood stages of Plasmodium falciparum. *Mol. Biochem. Parasitol.* **43**, 271-278 (1990).
285. Kappe SH, Buscaglia CA, Bergman LW, Coppens I & Nussenzweig V. Apicomplexan gliding motility and host cell invasion: overhauling the motor model. *Trends Parasitol.* **20**, 13-16 (2004).
286. Aikawa M, Carter R, Ito Y & Nijhout MM. New observations on gametogenesis, fertilization, and zygote transformation in Plasmodium gallinaceum. *J. Protozool.* **31**, 403-413 (1984).
287. Sinden RE, Canning EU & Spain B. Gametogenesis and fertilization in Plasmodium yoelii nigeriensis: a transmission electron microscope study. *Proc. R. Soc. Lond B Biol. Sci.* **193**, 55-76 (1976).
288. Inaba K. Molecular architecture of the sperm flagella: molecules for motility and signaling. *Zoolog. Sci.* **20**, 1043-1056 (2003).
289. Dearsly AL, Sinden RE & Self IA. Sexual development in malarial parasites: gametocyte production, fertility and infectivity to the mosquito vector. *Parasitology* **100 Pt 3**, 359-368 (1990).
290. Vinkenoog R, Veldhuisen B, Speranca MA, del Portillo HA, Janse C *et al*. Comparison of introns in a cdc2-homologous gene within a number of Plasmodium species. *Mol. Biochem. Parasitol.* **71**, 233-241 (1995).
291. Ponzi M, Janse CJ, Dore E, Scotti R, Pace T *et al*. Generation of chromosome size polymorphism during in vivo mitotic multiplication of Plasmodium berghei involves both loss and addition of subtelomeric repeat sequences. *Mol. Biochem. Parasitol.* **41**, 73-82 (1990).
292. Franke-Fayard B, Trueman H, Ramesar J, Mendoza J, van der KM *et al*. A Plasmodium berghei reference line that constitutively expresses GFP at a high level throughout the complete life cycle. *Mol. Biochem. Parasitol.* **137**, 23-33 (2004).
293. van der Wel AM, Tomas AM, Kocken CH, Malhotra P, Janse CJ *et al*. Transfection of the primate malaria parasite Plasmodium knowlesi using entirely heterologous constructs. *J. Exp. Med.* **185**, 1499-1503 (1997).
294. Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M *et al*. A generic protein purification method for protein complex characterization and proteome exploration. *Nat. Biotechnol.* **17**, 1030-1032 (1999).
295. Koning-Ward TF, Janse CJ & Waters AP. The development of genetic tools for dissecting the biology of malaria parasites. *Annu. Rev. Microbiol.* **54**, 157-185 (2000).
296. Janse CJ, Mons B, Rouwenhorst RJ, van der Klooster PF, Overdulve JP *et al*. In vitro formation of ookinetes and functional maturity of Plasmodium berghei gametocytes. *Parasitology* **91 (Pt 1)**, 19-29 (1985).
297. Janse CJ, Rouwenhorst RJ, van der Klooster PF, van der Kaay HJ & Overdulve JP. Development of Plasmodium berghei ookinetes in the midgut of Anopheles atroparvus mosquitoes and in vitro. *Parasitology* **91 (Pt 2)**, 219-225 (1985).
298. Pace T, Birago C, Janse CJ, Picci L & Ponzi M. Developmental regulation of a Plasmodium gene involves the generation of stage-specific 5' untranslated sequences. *Mol. Biochem. Parasitol.* **97**, 45-53 (1998).
299. Watanabe J, Sasaki M, Suzuki Y & Sugano S. FULL-malaria: a database for a full-length enriched cDNA library from human malaria parasite, Plasmodium falciparum. *Nucleic Acids Res.* **29**, 70-71 (2001).
300. Rusan NM, Fagerstrom CJ, Yvon AM & Wadsworth P. Cell cycle-dependent changes in microtubule dynamics in living cells expressing green fluorescent protein-alpha tubulin. *Mol. Biol. Cell* **12**, 971-980 (2001).
301. Straight AF, Marshall WF, Sedat JW & Murray AW. Mitosis in living budding yeast: anaphase A but no metaphase plate. *Science* **277**, 574-578 (1997).
302. Jarvik JW & Telmer CA. Epitope tagging. *Annu. Rev. Genet.* **32**, 601-618 (1998).

303. Corcoran LM, Thompson JK, Walliker D & Kemp DJ. Homologous recombination within subtelomeric repeat sequences generates chromosome size polymorphisms in P. falciparum. *Cell* **53**, 807-813 (1988).
304. Dore E, Pace T, Ponzi M, Picci L & Frontali C. Organization of subtelomeric repeats in Plasmodium berghei. *Mol. Cell Biol.* **10**, 2423-2427 (1990).
305. Pace T, Ponzi M, Dore E, Janse C, Mons B *et al*. Long insertions within telomeres contribute to chromosome size polymorphism in Plasmodium berghei. *Mol. Cell Biol.* **10**, 6759-6764 (1990).
306. Janse CJ & Mons B. Deletion, insertion and translocation of DNA sequences contribute to chromosome size polymorphism in Plasmodium berghei. *Mem. Inst. Oswaldo Cruz* **87 Suppl 3**, 95-100 (1992).
307. del Portillo HA, Lanzer M, Rodriguez-Malaga S, Zavala F & Fernandez-Becerra C. Variant genes and the spleen in Plasmodium vivax malaria. *Int. J. Parasitol.* **34**, 1547-1554 (2004).
308. Romero D, Martinez-Salazar J, Ortiz E, Rodriguez C & Valencia-Morales E. Repeated sequences in bacterial chromosomes and plasmids: a glimpse from sequenced genomes. *Res. Microbiol.* **150**, 735-743 (1999).
309. Fischer G, James SA, Roberts IN, Oliver SG & Louis EJ. Chromosomal evolution in Saccharomyces. *Nature* **405**, 451-454 (2000).
310. Zhang J & Peterson T. Genome rearrangements by nonlinear transposons in maize. *Genetics* **153**, 1403-1410 (1999).
311. Caceres M, Ranz JM, Barbadilla A, Long M & Ruiz A. Generation of a widespread Drosophila inversion by a transposable element. *Science* **285**, 415-418 (1999).
312. Kaminker JS, Bergman CM, Kronmiller B, Carlson J, Svirskas R *et al*. The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. *Genome Biol.* **3**, RESEARCH0084 (2002).
313. Hoskins RA, Smith CD, Carlson JW, Carvalho AB, Halpern A *et al*. Heterochromatic sequences in a Drosophila whole-genome shotgun assembly. *Genome Biol.* **3**, RESEARCH0085 (2002).
314. Glockner G, Szafranski K, Winckler T, Dingermann T, Quail MA *et al*. The complex repeats of Dictyostelium discoideum. *Genome Res.* **11**, 585-594 (2001).
315. Glockner G, Eichinger L, Szafranski K, Pachebat JA, Bankier AT *et al*. Sequence and analysis of chromosome 2 of Dictyostelium discoideum. *Nature* **418**, 79-85 (2002).
316. Ventura M, Archidiacono N & Rocchi M. Centromere emergence in evolution. *Genome Res.* **11**, 595-599 (2001).
317. Amor DJ & Choo KH. Neocentromeres: role in human disease, evolution, and centromere study. *Am. J. Hum. Genet.* **71**, 695-714 (2002).
318. Dujon B, Sherman D, Fischer G, Durrens P, Casaregola S *et al*. Genome evolution in yeasts. *Nature* **430**, 35-44 (2004).
319. Henikoff S, Ahmad K & Malik HS. The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* **293**, 1098-1102 (2001).
320. Talbert PB, Bryson TD & Henikoff S. Adaptive evolution of centromere proteins in plants and animals. *J. Biol.* **3**, 18 (2004).
321. Trenholme KR, Gardiner DL, Holt DC, Thomas EA, Cowman AF *et al*. clag9: A cytoadherence gene in Plasmodium falciparum essential for binding of parasitized erythrocytes to CD36. *Proc. Natl. Acad. Sci. U. S. A* **97**, 4029-4033 (2000).
322. Barnwell JW, Howard RJ, Coon HG & Miller LH. Splenic requirement for antigenic variation and expression of the variant antigen on the erythrocyte membrane in cloned Plasmodium knowlesi malaria. *Infect. Immun.* **40**, 985-994 (1983).
323. Brown KN. Antibody induced variation in malaria parasites. *Nature* **242**, 49-50 (1973).
324. Bonnefoy S, Guillotte M, Langsley G & Mercereau-Puijalon O. Plasmodium falciparum: characterization of gene R45 encoding a trophozoite antigen containing a central block of six amino acid repeats. *Exp. Parasitol.* **74**, 441-451 (1992).
325. Massague J. TGF-beta signal transduction. *Annu. Rev. Biochem.* **67**, 753-791 (1998).
326. Josso N & di Clemente N. Serine/threonine kinase receptors and ligands. *Curr. Opin. Genet. Dev.* **7**, 371-377 (1997).
327. Harrison T, Samuel BU, Akompong T, Hamm H, Mohandas N *et al*. Erythrocyte G protein-coupled receptor signaling in malarial infection. *Science* **301**, 1734-1736 (2003).

328. Omer FM & Riley EM. Transforming growth factor beta production is inversely correlated with severity of murine malaria infection. *J. Exp. Med.* **188**, 39-48 (1998).
329. Omer FM, Kurtzhals JA & Riley EM. Maintaining the immunological balance in parasitic infections: a role for TGF-beta? *Parasitol. Today* **16**, 18-23 (2000).
330. Wang T, Li BY, Danielson PD, Shah PC, Rockwell S *et al*. The immunophilin FKBP12 functions as a common inhibitor of the TGF beta family type I receptors. *Cell* **86**, 435-444 (1996).
331. Sokol JP & Schiemann WP. Cystatin C antagonizes transforming growth factor beta signaling in normal and cancer cells. *Mol. Cancer Res.* **2**, 183-195 (2004).
332. Inman GJ, Nicolas FJ, Callahan JF, Harling JD, Gaster LM *et al*. SB-431542 is a potent and specific inhibitor of transforming growth factor-beta superfamily type I activin receptor-like kinase (ALK) receptors ALK4, ALK5, and ALK7. *Mol. Pharmacol.* **62**, 65-74 (2002).
333. Laping NJ, Grygielko E, Mathur A, Butter S, Bomberger J *et al*. Inhibition of transforming growth factor (TGF)-beta1-induced extracellular matrix with a novel inhibitor of the TGF-beta type I receptor kinase activity: SB-431542. *Mol. Pharmacol.* **62**, 58-64 (2002).
334. Bottinger EP, Factor VM, Tsang ML, Weatherbee JA, Kopp JB *et al*. The recombinant proregion of transforming growth factor beta1 (latency-associated peptide) inhibits active transforming growth factor beta1 in transgenic mice. *Proc. Natl. Acad. Sci. U. S. A* **93**, 5877-5882 (1996).
335. Carter R. Transmission blocking malaria vaccines. *Vaccine* **19**, 2309-2314 (2001).
336. Nolte D, Hundt E, Langsley G & Knapp B. A Plasmodium falciparum blood stage antigen highly homologous to the glycophorin binding protein GBP. *Mol. Biochem. Parasitol.* **49**, 253-264 (1991).
337. Rudolph B, Nolte D & Knapp B. Isolation of a third member of the Plasmodium falciparum glycophorin-binding protein gene family. *Mol. Biochem. Parasitol.* **68**, 173-176 (1994).
338. Cowman AF, Baldi DL, Healer J, Mills KE, O'Donnell RA *et al*. Functional analysis of proteins involved in Plasmodium falciparum merozoite invasion of red blood cells. *FEBS Lett.* **476**, 84-88 (2000).
339. Bourgon R, Delorenzi M, Sargeant T, Hodder AN, Crabb BS *et al*. The serine repeat antigen (SERA) gene family phylogeny in Plasmodium: the impact of GC content and reconciliation of gene and species trees. *Mol. Biol. Evol.* **21**, 2161-2171 (2004).
340. Holt DC, Gardiner DL, Thomas EA, Mayo M, Bourke PF *et al*. The cytoadherence linked asexual gene family of Plasmodium falciparum: are there roles other than cytoadherence? *Int. J. Parasitol.* **29**, 939-944 (1999).
341. Adams JH, Sim BK, Dolan SA, Fang X, Kaslow DC *et al*. A family of erythrocyte binding proteins of malaria parasites. *Proc. Natl. Acad. Sci. U. S. A* **89**, 7085-7089 (1992).

Samenvatting

Malaria

Het malariaprobleem

Ondanks verwoede pogingen van de Wereld Gezondheids Organisatie (WHO) in de jaren ’50 en ’60 om malaria uit te roeien, wordt elk jaar zo’n 10% van de gehele wereldbevolking opnieuw geïnfecteerd en sterven meer dan een miljoen mensen, voornamelijk kleine kinderen in Afrika, aan deze parasitaire infectie. Naast het directe leed voor de bevolking, zorgt deze ziekte ook voor een vertraging van de economische groei van de vaak toch al arme landen waar deze ziekte veel voorkomt. Een oplossing van het malariaprobleem lijkt nog altijd ver weg; er is nog steeds geen effectief vaccin tegen malaria, de parasieten worden resistent tegen de gangbare medicijnen, muskieten die de malariaparasiet overbrengen worden resistent tegen chemische bestrijdingsmiddelen en veel Afrikaanse landen hebben te weinig geld voor de bestrijding en behandeling van malaria.

De veroorzaker van malaria

Malaria wordt veroorzaakt door eencellige parasieten van het geslacht *Plasmodium*. Er zijn veel verschillende *Plasmodium* soorten die verschillende gastheren parasiteren, zoals reptielen, vogels, knaagdieren en apen. Bij de mens vinden we vier soorten, waarvan *Plasmodium falciparum* de meeste slachtoffers maakt.

De parasieten worden overgebracht door *Anopheles* muggen. Tijdens een beet van een vrouwelijke mug komen de parasieten, via het speeksel van de mug, in het bloed van de mens. Deze zogeheten sporozoieten komen via het bloed in de lever terecht, waar zij een levercel binnendringen en zich vervolgens beginnen te vermenigvuldigen. Zo’n geïnfecteerde levercel kan meer dan 10.000 parasieten bevatten, de zogenoemde merozoieten die, na openbarsten van de levercel, in het bloed terechtkomen en rode bloedcellen binnen dringen. In de rode bloedcel deelt één parasiet zich verder, daarbij 16 tot 32 nieuwe parasieten vormend, die opnieuw rode bloedcellen kunnen infecteren. Deze aseksuele vermenigvuldiging in de rode bloedcellen zorgt voor een snelle toename van parasieten in het bloed en is verantwoordelijk voor de kenmerkende ziekteverschijnselen van malaria, zoals koorts en beschadigingen van organen zoals milt, lever en hersenen, hetgeen kan leiden tot coma en de dood. Sommige van de parasieten in het bloed stoppen met de aseksuele deling en ontwikkelen zich tot de seksuele stadia, de zogeheten gametocyten. Deze vormen van de parasiet zijn noodzakelijk voor een succesvolle transmissie van de parasiet via de mug naar een nieuwe gastheer. Wanneer een mug zich voedt op een malaria-geïnfecteerde gastheer dan ontwikkelen deze gametocyten zich verder in de muggenmaag tot mannelijke en vrouwelijke gameten. Na bevruchting dringt de nieuwgevormde zygoot (de ookineet) door de maagwand en nestelt zich aan de buitenkant om zich verder te ontwikkelen tot een oocyste waarin door deling vele duizenden sporozoieten gevormd worden. Deze sporozoieten migreren vervolgens naar de speekselklieren van de mug, vanwaar de cyclus opnieuw kan beginnen.

Knaagdiermalariaparasieten als model

Malariaparasieten die Afrikaanse boomratjes infecteren worden wereldwijd als model gebruikt voor onderzoek aan malaria. Het werken met deze modellen heeft een aantal voordelen. De parasieten infecteren laboratoriumknaagdieren, waardoor experimenteel *in vivo* onderzoek goed toegankelijk is, hetgeen door technische en ethische aspecten moeilijker uitvoerbaar is met malariaparasieten die mensen en andere primaten infecteren. Zo worden deze modellen bijvoorbeeld veel gebruikt bij het onderzoek naar de ontwikkeling van de parasiet in de lever en de seksuele ontwikkeling die gedeeltelijk in de mug plaats vindt en bij onderzoek naar immunologische aspecten van malaria. Het gebruik van deze malariaparasieten als model is echter alleen zinvol wanneer deze parasieten veel overeenkomsten vertonen met de malariaparasieten die mensen infecteren, waardoor het onderzoek resultaten oplevert die relevant zijn voor de humane situatie.

Doel van het onderzoek

Knaagdiermalariaparasieten vertonen veel overeenkomsten met humane malariaparasieten. De levenscyclus van de verschillende malariaparasieten is hetzelfde, net als de morfologie van de verschillende stadia. Daarnaast blijkt ook uit moleculair en biochemisch onderzoek dat er een hoge mate van overeenkomsten bestaat. Zo hebben beide soorten 14 lineaire chromosomen, zijn een groot aantal genen coderend voor antigenen, die gezien worden als vaccinkandidaten, geconserveerd. Bovendien werken veel medicijnen tegen beide malariaparasieten hetzelfde door middel van herkenning van dezelfde targets en is resistentie tegen deze medicijnen het gevolg van overeenkomstige mutaties in deze targets. Het anti-malariamiddel pyrimethamine kan bijvoorbeeld gebruikt worden tegen zowel knaagdiermalaria als humane malaria en in beide soorten kan resistentie optreden als gevolg van overeenkomstige mutaties in hetzelfde gen.

Het doel van het onderzoek zoals beschreven in dit proefschrift was om meer inzicht te krijgen in de mate van overeenkomst tussen de organisatie en “gen-inhoud” van de genomen van malariaparasieten knaagdieren en mensen. Het onderzoek begon op kleine schaal met studies die de organisatie van enkele genen en delen van de chromosomen in kaart brachten. Door het vrijkomen van de genoomsequenties van de humane malariaparasiet *P. falciparum*[42] en van verschillende knaagdiermalariaparasieten (*Plasmodium yoelii*[51] - Hoofdstuk 3; *Plasmodium berghei* en *Plasmodium chabaudi*[52] - Hoofdstuk 4) is het onderzoek uitgebreid tot het vergelijken van de volledige genomen van deze soorten. Omdat voor de knaagdiermalariaparasieten slechts partiële genoomsequenties beschikbaar waren, is het onderzoek in eerste instantie gericht op het creëren van één samengesteld genoom van de drie knaagdiermalariaparasieten. Dit werd bereikt door de korte DNA-sequenties (contigs) van deze drie soorten langs het *P. falciparum* genoom te leggen en met elkaar te verbinden. Vervolgens is de organisatie van dit samengestelde knaagdiermalariagenoom in detail vergeleken met dat van *P. falciparum*.

Deze vergelijkende studies lieten zien dat er een hoge mate van overeenkomst is in de organisatie en gen-inhoud van de genomen van malariaparasieten van knaagdieren en mensen. Dit is een belangrijke vinding ter ondersteuning van de

waarde van knaagdiermalariaparasieten als modellen in het malariaonderzoek en het gebruik van dit soort modellen voor de verdere functionele karakterisering van genen en eiwitten in het post-genoomtijdperk, bijvoorbeeld met behulp van “reverse genetics” methoden.

In deze studie zijn, naast de genoomvergelijkingen, dergelijke “reverse genetics” methoden gebruikt om de functie van bepaalde genen, die geconserveerd zijn tussen malariaparasieten van knaagdieren en mensen, verder te onderzoeken.

Vergelijkend genoomonderzoek

Een hoge mate van overeenkomst tussen de organisatie en gen-inhoud van de genomen van de humane malariaparasiet P. falciparum *en knaagdiermalariaparasieten*

Gelijktijdig met de publicatie van het *P. falciparum* genoom in oktober 2002 werd de partiële genoomsequentie van *P. yoelii* gepubliceerd, wat het mogelijk maakte om voor de eerste keer in de geschiedenis de genomen van twee organismen van eenzelfde geslacht te vergelijken (Hoofdstuk 3). Het computer programma MUMmer werd gebruikt om 2.212 *P. yoelii* contigs in lijn te leggen met het *P. falciparum* genoom. Door middel van polymerase kettingreacties (PCR) werd vervolgens voor zoveel mogelijk aan elkaar grenzende contigs aangetoond of ze daadwerkelijk fysiek dicht bij elkaar liggen in het *P. yoelii* genoom. Voortbouwend op het werk dat was aangevangen door L.H.M. van Lin in Leiden, gebruikten we een fysieke kaart van chromosoom 5 van *P. berghei* om in meer detail aan te tonen dat de organisatie en gen-inhoud van het centrale deel van het vergelijkbare chromosoom 5 van *P. yoelii* in grote mate overeenkomt met delen van chromosomen 4 en 10 van *P. falciparum*. De reden waarom wij bijzonder geïnteresseerd waren in dit specifieke chromosoom is, dat er een link zou kunnen bestaan tussen de organisatie van dit chromosoom en de seksuele ontwikkeling van de parasiet. Zo waren er verscheidene genen bekend, gelegen op dit chromosoom, die tot expressie kwamen tijdens de seksuele ontwikkeling, terwijl grootschalige deleties in een subtelomeer gebied geassocieerd waren met het verlies van het vermogen om de seksuele stadia te vormen. Deze waarneming duidde op de mogelijkheid van clustering en gecoördineerde expressie van genen, specifiek voor de seksuele stadia. Daarnaast toonden we aan dat de twee delen van chromosoom 5 van *P. berghei* en *P. yoelii*, die overeenkomen met chromosomen 4 en 10 van *P. falciparum*, aan elkaar gekoppeld zijn via één van de vier *ribosomale rna* genen, die specifiek is voor de knaagdiermalariaparasieten, wat tevens een eerste aanwijzing was dat genfamilies mogelijk betrokken zijn bij grootschalige genoomreorganisaties.

Partiële genoomsequenties van twee andere knaagdiermalariaparasieten, *P. berghei* en *P. chabaudi*, werden gecombineerd met de *P. yoelii* contigs en vergeleken met elkaar en met het *P. falciparum* genoom. Door de hoge mate van synteny, tussen de genomen van de drie knaagdiermalariaparasieten onderling enerzijds en tussen deze genomen en dat van *P. falciparum* anderzijds, was het mogelijk om grote, samengestelde knaagdiermalariacontigs te maken van overlappende *P. yoelii*, *P. berghei* en *P. chabaudi* contigs. Hierdoor werd het mogelijk om gedetailleerde genoomkaarten samen te stellen, die de

overeenkomsten en de verschillen in organisatie van het samengestelde knaagdiermalariagenoom en het *P. falciparum* genoom laten zien, de zogenoemde syntenykaarten (Hoofdstukken 4 en 5).

Nadat alle samengestelde contigs van de knaagdiermalariaparasieten langs het *P. falciparum* genoom waren gelegd, bleken er nog slechts 228 gaten in het samengestelde genoom te zitten. In combinatie met 138 DNA-markers konden we aantonen, dat de samengestelde knaagdiermalariacontigs verdeeld liggen over 36 blokken die synteny vertonen met 84% van het *P. falciparum* genoom (Hoofdstuk 5). Tweezijdige analyses met het computerprogramma BLASTP hadden reeds aangetoond dat tenminste 3.300 *P. yoelii* genen homologie vertonen met een van de 5.300 *P. falciparum* genen, zogenoemde orthologen (Hoofdstuk 3). Het beschikbaar komen van DNA-sequenties van *P. berghei* en *P. chabaudi* maakte het mogelijk om, in combinatie met de genoomdata van *P. yoelii*, een kernset van tenminste 4.500 orthologen van de ongeveer 5.300 *P. falciparum* genen (85%) te beschrijven, die worden gedeeld tussen de genomen van *P. falciparum* en tenminste een van de knaagdiermalariasoorten (Hoofdstuk 4).

Slechts twee grootschalige, chromosomale translocaties onderscheiden de organisatie van de genomen van de drie knaagdiermalariaparasieten, wat suggereert dat grootschalige genoomreorganisaties, niet vaak voorkomen in *Plasmodium*. De organisatie van het *P. berghei* genoom is identiek aan dat van het samengestelde knaagdiermalariagenoom en komt daarmee waarschijnlijk het meest overeen met het genoom van de meest recente gemeenschappelijke voorouder van de drie knaagdiermalariaparasieten.

De subtelomere gebieden van de chromosomen zijn niet geconserveerd tussen humane en knaagdiermalariaparasieten

Tijdens de verschillende genoomstudies werd het steeds duidelijker dat de centrale delen van alle malaria chromosomen sterk zijn geconserveerd en dat de grenzen die deze centrale delen scheiden van de subtelomere delen scherp zijn afgetekend. Deze subtelomere gebieden van chromosomen van *Plasmodium* variëren sterk in lengte doordat ze variabele aantallen korte, zichzelf herhalende stukjes DNA (“repeat-sequenties”) bevatten. Daarnaast bevatten deze gebieden een groot aantal genen die behoren tot variabele genfamilies, die voor een extra bron van variëteit in lengte, organisatie en gen-inhoud van de subtelomere gebieden zorgen. Veel van deze subtelomere “repeat-sequenties” en genfamilies zijn niet geconserveerd tussen de knaagdiermalariaparasieten en *P. falciparum*. Eén zo’n familie in de knaagdiermalariaparasieten wordt gevormd door de *yir*, *bir* en *cir* genen, waarvan >800, 180 en 138 kopieën voorkomen in de genomen van respectievelijk *P. yoelii*, *P. berghei* en *P. chabaudi*. Deze genen zijn mogelijk betrokken bij antigene variatie en liggen voornamelijk in de subtelomere gebieden.

Ondanks de extreme mate van variabiliteit in zowel organisatie als gen-inhoud van de subtelomere gebieden van de verschillende *Plasmodium* soorten, zijn er aanwijzingen dat een aantal genfamilies, die op het eerste gezicht geen homologie vertonen tussen malariaparasieten van knaagdieren en mensen, vergelijkbare functies kunnen vervullen. Bij nauwkeurige analyse blijkt dat genen van deze families toch een zekere mate van homologie kunnen vertonen in bepaalde gebieden (domeinen) of structuur. Genen van dergelijke families van verschillende

malariaparasieten kunnen dan mogelijk beschouwd worden als snel evoluerende paralogen in plaats van ongerelateerde genen. Een voorbeeld vormt de *pir* superfamilie[145,202] (Hoofstuk 4), die naast de *yir*, *bir* en *cir* genen in de knaagdiermalariaparasieten bestaat uit de *vir* genen in de humane parasiet *Plasmodium vivax* en de *kir* genen in de primatenparasiet *Plasmodium knowlesi*, maar die wellicht ook de *rif* genen van *P. falciparum* omvat.

Door het vergelijken van de genomen van de knaagdiermalariaparasieten en *P. falciparum* konden de grenzen tussen de variabele subtelomere en geconserveerde centrale gebieden worden gedefinieerd als korte DNA-sequenties die twee naast elkaar gelegen genen scheiden. De meerderheid (23 van de 28) van deze grensgebieden bleken geconserveerd tussen de verschillende soorten (Hoofdstuk 5). Helaas maakte het gebrek aan synteny in de subtelomere gebieden het onmogelijk om knaagdiermalariacontigs langs het *P. falciparum* genoom te leggen en was het daardoor tevens onmogelijk om samengestelde knaagdiermalariacontigs te genereren van deze subtelomere gebieden.

Voor 743 *P. falciparum* genen kon geen ortholoog worden geïdentificeerd in de datasets, beschikbaar voor de knaagdiermalariaparasieten. Het merendeel van deze *P. falciparum* specifieke genen (575, 11% van alle *P. falciparum* genen) bleek gelegen te zijn in de subtelomere gebieden van de chromosomen en veel van deze genen konden onderverdeeld worden in 12 genfamilies. Genen van vijf van deze families vertoonden homologie met genen van de knaagdiermalariaparasieten. Dit laat zien dat er tenminste enige, zij het geringe, homologie bestaat tussen de subtelomere gebieden van humane en knaagdiermalariagenomen (Hoofdstukken 4 en 5).

Tenslotte bleek uit de syntenykaart dat de locaties van potentiële centromeren, DNA-sequenties met een lengte van 1.5-2.5 kb die geen genen bevatten en voornamelijk bestaan uit de basen A en T (ongeveer 97%), en de gebieden die aan deze potentiële centromeren grenzen sterk overeenkomen tussen *P. falciparum* en de knaagdiermalariaparasieten.

P. falciparum *specifieke genen liggen niet alleen in de subtelomere gebieden maar kunnen ook gevonden in syntenybreekpunten en in indels in de syntenyblokken*

Door de organisatie en locatie van de 743 *P. falciparum* specifieke genen te analyseren met behulp van de syntenykaart, konden we aantonen dat een significant deel hiervan (168 genen) niet gelegen is in de subtelomere gebieden. Van deze 168 genen bleken er 42 in de syntenybreekpunten tussen de geconserveerde syntenyblokken te liggen (in acht indels), terwijl de meerderheid (126 genen) aanwezig was als indels gelegen in de syntenyblokken (82 indels). Het is interessant dat verscheidene van deze indels clusters van meerdere genen bevatten, die dezelfde oriëntatie en vergelijkbare expressie patronen hebben. Zulke clusters zijn mogelijk ontstaan door locale genduplicatie en worden wellicht als operon afgeschreven[269], een proces waar in eerdere studies geen bewijs voor is gevonden[11] (twee voorbeelden, waaronder een cluster van *msp* genen zijn te zien in Hoofdstuk 5, Figuur 4). Wellicht nog belangrijker voor het begrijpen van de biologie van malariaparasieten was de observatie dat tenminste twee-derde van de 168 *P. falciparum* specifieke genen voor eiwitten codeert met N-terminale transmembrane domeinen, hetgeen een indicatie is dat deze eiwitten gelocaliseerd

zijn aan de oppervlakte van de parasiet of de geinfecteerde rode bloedcelmembraan en dus een rol kunnen spelen bij parasiet-gastheer interacties.

Opvallend is dat er significant meer genen gelegen zijn tussen de syntenyblokken (in de syntenybreekpunten) in het *P. falciparum* genoom dan in de 19 van de 22 gekarakteriseerde syntenybreekpunten van knaagdiermalaria-parasieten. Ook zijn er tot nu toe maar weinig indels gevonden specifiek voor knaagdiermalariaparasieten. Dit kan echter ook het gevolg kan zijn van de nog incomplete genoomsequenties van de drie knaagdiermalariaparasieten. De hoeveelheid sequentie die op dit moment beschikbaar is duidt er echter wel op dat indels in de knaagdiermalariaparasieten niet zo frequent voorkomen als in *P. falciparum*.

In slechts 15 stappen kan de genoomorganisatie van P. falciparum *uit die van de knaagdiermalariaparasieten worden gevormd*

Naar schatting zo'n 50-200 miljoen jaren evolutie scheiden de humane parasiet *P. falciparum* en de knaagdiermalariaparasieten[14]. Door de syntenykaarten van beide soorten te vergelijken, konden we aantonen dat er tenminste 15 chromosomale recombinaties nodig zijn om het (centrale) genoom van de knaagdiermalariaparasieten en dat van *P. falciparum* om te zetten en vice versa. Hieruit volgt dat er gemiddeld slechts 0,08-0,3 grootschalige recombinatie plaatsvinden elke miljoen jaar, wat lijkt te duiden dat *Plasmodium* soorten een "stabieler kerngenoom" hebben dan hun gastheren of wormen (zie ook de algemene discussie, Hoofdstuk 7). Daarentegen heeft 14% van de *P. falciparum* genen geen duidelijk homoloog gen in één van de knaagdiermalariaparasieten, terwijl bijvoorbeeld slechts 1% van alle humane genen geen homoloog heeft in het muisgenoom. Zoals hierboven aangegeven ligt het merendeel van deze genen (77%) in de subtelomere gebieden en behoort tot (grote) genfamilies.

Alhoewel het mogelijk was om de reorganisatie te simuleren waarmee het *P. falciparum* genoom gevormd kan worden uit het knaagdiermalariagenoom, liet een verdere analyse van de DNA-sequenties die syntenybreekpunten flankeren geen direkte verbanden zien tussen DNA-structuren en mogelijke recombinaties. Ook is het niet mogelijk om met twee genomen een voorspelling te doen over de organisatie van het genoom van de meest recente voorouder. Hiervoor is de kennis van de organisatie van minimaal nog een derde *Plasmodium* genoom nodig[264].

Het beschikbaar komen van meer genoomsequenties van onder andere *P. vivax* en *P. knowlesi* kan leiden tot het genereren van een meer definitieve stamboom van het geslacht *Plasmodium*, gebaseerd op de genoomorganisaties en mate van synteny. Dit kan ook meer inzicht geven in het "recombinatie schema" dat geleid heeft tot de organisatie van syntenyblokken zoals die in hedendaagse *Plasmodium* soorten wordt gevonden en in de processen die hebben bijgedragen en mogelijk nog steeds bijdragen aan de vorming van genfamilies. Dit principe is beschreven voor de vorming van een *P. falciparum* specifieke genfamilie bestaande uit 21 receptor proteïne kinases (*pftstk*) die slechts een enkel "voorouder" gen heeft in alle andere malariasoorten en wiens ontstaansgeschiedenis voor een gedeelte gelinkt is met grootschalige recombinaties die de synteny beïnvloed hebben.

P. falciparum *specifieke genfamilies en grootschalige genoomreorganisaties*

De meeste *P. falciparum* specifieke genfamilies liggen in de subtelomere gebieden van de chromosomen. In eerdere studies was al aangetoond dat genen van dergelijke genfamilies, zoals de *var* en *rif* families, niet slechts in de subtelomere gebieden liggen maar ook als clusters in de centrale gebieden van de chromosomen. Door de analyses van de syntenykaart en de locaties van de *P. falciparum* specifieke genen, konden wij vier van de zeven centrale *var* clusters in verband brengen met grootschalige recombinaties en werd een nieuwe genfamilie (*vicar*) gevonden, die specifiek geassocieerd lijkt te zijn met de centrale *var* clusters en mogelijk een rol gespeeld heeft bij het ontstaan van deze interne clusters.

Twee syntenybreekpunten bevatten genen die behoren tot een intrigerende familie van 21 serine/threonine receptor kinases (*pftstk*), hetgeen er op duidt dat grootschalige recombinaties zijn geassocieerd met de vorming van deze genfamilie, waarvan de meeste in de subtelomere gebieden zijn gelegen. Slechts één lid van deze *P. falciparum* specifieke familie dat centraal gelegen is op chromosoom 8 heeft een homoloog gen in de andere *Plasmodium* soorten. Het combineren van de syntenykaart en phylogenetische analyses van de *pftstk* familie geeft inzicht in ontstaansgeschiedenis van deze familie. Na duplicatie van het centraal gelegen “voorouder” gen is één copy in een subtelomeer gebied terecht gekomen waarna deze familie verder is geëxpandeerd door amplificatie en uitwisseling met subtelomere gebieden van andere chromosomen.

Functionele analyse van genen met verhoogde expressie tijdens de seksuele stadia die geconserveerd zijn tussen P. falciparum *en knaagdiermalariaparasieten*

Zoals hierboven al werd genoemd is, hadden wij bijzondere interesse in de organisatie van chromosoom 5 van *P. berghei*, omdat het vermoeden bestond dat deze mogelijk verrijkt is in genen specifiek voor de seksuele stadia van de parasiet. Zowel in de studies beschreven in dit proefschrift als in andere studies zijn geen sterke aanwijzingen gevonden dat genen die verhoogd of exclusief tot expressie komen tijdens deze seksuele stadia specifiek geclusterd zijn op chromosoom 5, maar dat deze verspreid liggen over de 14 chromosomen. In dit proefschrift zijn een aantal studies beschreven met als doel de verdere karakterisatie van “sex-specifieke” genen. Zo onderzochten wij een aantal genen van een locus gelegen op chromosoom 5 (de B9 locus), waarvan drie van de zes genen specifiek in gametocyten tot expressie komen. In de algemene discussie (Hoofdstuk 7) is meer informatie te vinden over deze studies. Gezien het voorlopige karakter van deze studies, zijn de gegevens hiervan nog niet gepubliceerd.

Daarnaast is een analyse uitgevoerd van de expressie van de twee *α-tubulin* genen van *Plasmodium*. Eén van deze genen, *α-tubulin II*, ligt op chromosoom 5 en eerder is gerapporteerd dat dit gen specifiek in mannelijke gametocyten tot expressie komt en onderdeel is van de axonemen van de mannelijke gameten (Hoofdstuk 6). Tegen de verwachting in bleek uit onze studies dat α-tubulin II niet specifiek is voor de mannelijke gameten maar dat het ook een essentiële rol vervult tijdens de aseksuele ontwikkeling van de bloedstadia.

Concluderende opmerkingen

Samenvattend kan gesteld worden dat de vergelijkende genoomanalyses die in dit proefschrift staan beschreven aantonen dat er een hoge mate van overeenkomst bestaat tussen de organisatie en gen-inhoud van de genomen van de knaagdiermalariaparasieten en de humane malariaparasiet *P. falciparum*. Dit is een belangrijke vinding ter ondersteuning van de waarde van knaagdiermalariaparasieten als modellen in het malariaonderzoek en het gebruik van dit soort modellen voor de verdere functionele karakterisering van genen en eiwitten in het post-genoomtijdperk, bijvoorbeeld met behulp van “reverse genetics” methoden. Naast deze hoge mate van conservatie, heeft het onderzoek interessante verschillen aan het licht gebracht met betrekking tot de organisatie van genfamilies. Het belang van deze verschillen ligt vooral in het feit, dat veel van de eiwitten die gecodeerd worden door de genen van deze families betrokken zijn bij de interacties van de parasiet met cellen en het immuunsysteem van de gastheer. Verdere bestudering van deze interacties en de genfamilies die hierin betrokken zijn kan meer inzicht geven, niet alleen in de moleculaire mechanismen die ten grondslag liggen aan het ontstaan van deze soortspecifieke genen, maar ook in soortspecifieke adaptatie aan de verschillende gastheren en de mechanismen die parasieten in staat stellen om te ontsnappen aan het immuunsysteem.

Acknowledgements

I would like to take this opportunity to express my gratitude to all that have in one way or another contributed to the realization of this thesis, to begin with all the people of the Department of Parasitology, and in particular those of the Malaria Research Group. First of all, I would like to thank my colleague, great friend, and paranimf Hans Kroeze, for his protocols in molecular biology and support both within and outside the lab. I would also like to thank Jai Ramesar, for guiding me through my first steps in malaria research and for the invaluable support with the culturing of parasites. Special thanks to Blandine Franke-Fayard, for helping with visualizing those parasites and excellent discussions after parasitology meetings. Apart from being great sports and supports in the lab the following people helped by contributing probes or chromosomal locations of their partiular genes of interest: Anneke Braks, Milly van Dijk, Don Gardiner, Alireza Haghparast, Joke de Jong, Marianna Karras, Shahid Khan, Gunnar Mair, Resie van Spaendonk, Joanne Thompson, and Dina Vlachou. Furthermore, I would like to thank Kevin Augustijn and Maaike van Dooren, for providing me with the TAP-tag vector; Leo van Lin, for introducing me to *Plasmodium berghei* chromosome 5; Jasper Renz, for providing me with a flying start on the tubulin project; and Auke van Wigcheren, for helping me out when I wanted to crash my computer. Finally, I would like to thank our visiting Ph.D. students, Vasco Correia, Marta Tufet-Bayona, and Pierrick Uzureau for sharing fate, and my student Abdallah Musa Abdallah for his brave efforts trying to map the *pir* superfamily.

There are a couple of other people from the department that deserve some attention: Caroline Remmerswaal and Jantien Guldemond-Nieuwland helped me keep the process of getting my promotion organized running smoothly, while I was in Oxford; Rene van Zeijl, for computer support and cryptanalysis; and my fellow Ph.D. students on the floor, Anita van den Biggelaar, Hanneke de Gruijter, Desirée van der Kleij, Alexandra van Remoortere, Marjolein Robijn, Marike van Roon, and Koen van den Vijver, for mental support and understanding.

The radioactive work was done at the Department of Virology. I would like to thank Richard (Rimo) Molenkamp for his help in and around the Geiger-counter; and Erwin van den Born and Patrick Wanningen for their support in and around the labs.

Much of what I have learned about biology of parasitism was brought to me in eight intense weeks in Woods Hole, MA, USA.. Therefore, I would like to express my gratitude to the organizers Chris Tschudi and Elisabetta Ullu for their amazing job and enthusiasm; co-director Jay Bangs, for his outrageous enthusiasm; Rick Tarleton, for making me see the light of immunology, however short; and last but not least Geoff McFadden and his side-kick Stuart Ralph, for all the midnight, FRAP-man hours. Finally, a big thanks to my fellow students, the eminent boppers and boppettes.

The majority of my thesis, which covers the comparative genome analysis of rodent malaria species, would not have been possible without a number of people: Jane Carlton, Shelby Bidwell, and Neil Hall (The Institute for Genomic Research, MD, USA; also thanks to Jane for hosting me at my visit to TIGR); Matt Berriman (Wellcome Trust Sanger Institute); Tomasso Pace (Istituto Superiore di Sanità); and Mike Turner (University of Glasgow).

Tenslotte wil ik graag al mijn vrienden bedanken voor hun begrip en ondersteuning. Ik had dit zeker niet voor elkaar gekregen zonder jullie steun of zonder mijn uitlaatklep met de mannen van Işgot.

Papa, mama, Eelke, Pepijn en paranimf Sanne, bedankt voor jullie interesse, onvoorwaardelijke steun, blind vertrouwen en liefde (en voor de spellingscontrole natuurlijk).

Jojanneke, bedankt voor je eindeloze geduld, advies en onuitputtelijke liefde, ik heb je dan toch eindelijk bijgehaald en nu kunnen we samen “uitrusten”.

Curriculum vitae

Taco Wilhelmus Antonius Kooij was born on the 20th of March 1976 in Warmond (The Netherlands). In 1994, he passed his gynasium β exams (*cum laude*) at the Thomas à Kempis College in Zwolle and started his study chemistry at Utrecht University. From September 1997 to March 1998 he worked on a EU-funded research project in bioorganic chemistry at the Department of Pure and Applied Chemistry, University of Strathclyde (Glasgow, UK), under the supervision of Prof. dr. C.J. Suckling. Following this European adventure he continued to specialize in biochemistry at the Department of Biochemistry of Lipids, Utrecht University (The Netherlands), under the supervision of Dr. ing. G.P.C. Drummen and Prof. dr. J.A.F. Op den Kamp, and obtained his master's degree (*met genoegen*) in September 1999. After a short break (working nightshifts at the Dutch Parcel Service [NPD]), he continued in science as a Ph.D. student funded by the University of Leiden. From March 2000 to July 2004 he studied molecular biology and genomics of the rodent model malaria parasite *Plasmodium berghei* under the supervision of Prof. dr. A.P. Waters and Dr. C.J. Janse at the Department of Parasitology, Leiden University Medical Centre (LUMC, The Netherlands). Most of this work was performed in close collaboration with The Institute for Genomic Research (TIGR, Rockville, MD, USA) and the Wellcome Trust Sanger Institute (Cambridge, UK). As part of his training, he followed the 8-week course Biology of Parasitism (Woods Hole, MA, USA) in 2002. Currently, he is working for the Nuffield Department of Clinical Laboratory Sciences (University of Oxford) as a post-graduate researcher in the laboratory of Prof. dr. D.J. Roberts at the National Blood Service (John Radcliffe Hospital, Oxford, UK), studying the human malaria parasite *Plasmodium falciparum*.

Publications

Eward H.W. Pap, Gregor P.C. Drummen, Victor J. Winter, **Taco W.A. Kooij**, Phylip J. Rijken, Karel W.A. Wirtz, Jos A.F. Op den Kamp, Willem J. Hage and Jan A. Post, "Ratio-fluorescence microscopy of lipid oxidation in living cells using C_{11}-BODIPY$^{581/591}$", *FEBS Lett.* **453** (3), 278-282 (1999).

Jane M. Carlton, Samuel V. Angiuoli, Bernard B. Suh, **Taco W.A. 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 and Daniel J. Carucci, "Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*", *Nature* **419** (6906), 512-519 (2002).

Neil Hall, Marianna Karras, J. Dale Raine, Jane M. Carlton, **Taco W.A. Kooij**, Matthew Berriman, Laurence Florens, Christoph S. Janssen, Arnab Pain, Georges K. Christophides, Keith James, Kim Rutherford, Barbara Harris, David Harris, Carol Churcher, Michael A. Quail, Doug Ormond, Jon Doggett, Holly E. Trueman, Jacqui Mendoza, Shelby L. Bidwell, Marie-Adele Rajandream, Daniel J. Carucci, John R. Yates III, Fotis C. Kafatos, Chris J. Janse, Bart Barrell, C. Michael R. Turner, Andrew P. Waters and Robert E. Sinden, "A comprehensive survey of the *Plasmodium* life cycle by genomic, transcriptomic, and proteomic analyses", *Science* **307** (5706), 82-86 (2005).

Taco W.A. Kooij, Blandine Franke-Fayard, Jasper Renz, Hans Kroeze, Maaike W. van Dooren, Jai Ramesar, Kevin D. Augustijn, Chris J. Janse and Andrew P. Waters, "*Plasmodium berghei* α-tubulin II: a role in both male gamete formation and asexual blood stages", *Mol. Biochem. Parasitol.* **144** (1), 16-26 (2005).

Taco W.A. Kooij, Jane M. Carlton, Shelby L. Bidwell, Neil Hall, Jai Ramesar, Chris J. Janse and Andrew P. Waters, "A *Plasmodium* whole-genome synteny map: indels and synteny breakpoints as foci for species-specific genes", *PLoS Pathog.* **1** (4), e44 (2005).

Taco W.A. Kooij, Chris J. Janse and Andrew P. Waters, "*Plasmodium* post-genomics - better the bug you know?", *Nat. Rev. Microbiol.* submitted (2006).