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.

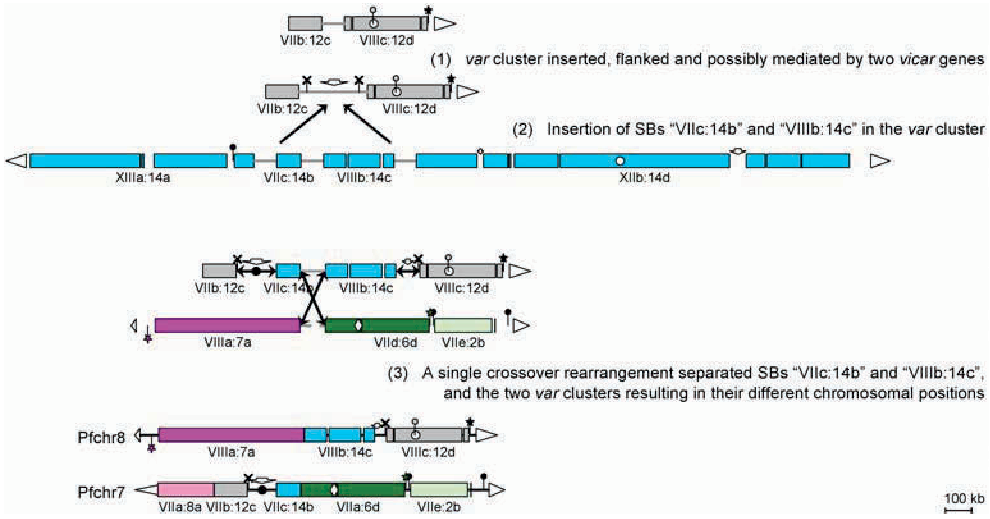

Figure 1. Putative mechanism of expansion of chromosome-internal *var* clusters through mediation of a new family of *var internal cluster associated repeat* (*vicar*) genes

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.