A comprehensive evaluation of rodent malaria parasite genomes and gene expression

Thomas D Otto1, Ulrike Böhme1, Andrew P Jackson2, Martin Hunt1, Blandine Franke-Fayard3, Wieteke A M Hoeijmakers4, Agnieszka A Religa3, Lauren Robertson1, Mandy Sanders1, Solabomi A Ogun6, Deirdre Cunningham6, Annette Erhart7, Oliver Billker1, Shahid M Khan3, Hendrik G Stunnenberg4, Jean Langhorne6, Anthony A Holder6, Andrew P Waters5, Chris I Newbold8,9, Arnab Pain10, Matthew Berriman1* and Chris J Janse3*

Abstract

Background: Rodent malaria parasites (RMP) are used extensively as models of human malaria. Draft RMP genomes have been published for Plasmodium yoelii, P. berghei ANKA (PbA) and P. chabaudi AS (PcAS). Although availability of these genomes made a significant impact on recent malaria research, these genomes were highly fragmented and were annotated with little manual curation. The fragmented nature of the genomes has hampered genome wide analysis of Plasmodium gene regulation and function.

Results: We have greatly improved the genome assemblies of PbA and PcAS, newly sequenced the virulent parasite P. yoelii YM genome, sequenced additional RMP isolates/lines and have characterized genotypic diversity within RMP species. We have produced RNA-seq data and utilised it to improve gene-model prediction and to provide quantitative, genome-wide, data on gene expression. Comparison of the RMP genomes with the genome of the human malaria parasite P. falciparum and RNA-seq mapping permitted gene annotation at base-pair resolution. Full-length chromosomal annotation permitted a comprehensive classification of all subtelomeric multigene families including the ‘Plasmodium interspersed repeat genes’ (pir). Phylogenetic classification of the pir family, combined with pir expression patterns, indicates functional diversification within this family.

Conclusions: Complete RMP genomes, RNA-seq and genotypic diversity data are excellent and important resources for gene-function and post-genomic analyses and to better interrogate Plasmodium biology. Genotypic diversity between P. chabaudi isolates makes this species an excellent parasite to study genotype-phenotype relationships. The improved classification of multigene families will enhance studies on the role of (variant) exported proteins in virulence and immune evasion/modulation.

Keywords: Plasmodium chabaudi, Plasmodium berghei, Plasmodium yoelii, Genomes, RNA-seq, Genotypic diversity, Multigene families, pir, Phylogeny

Background

Rodent malaria parasites (RMP) are used extensively as models of human malaria [1,2]. Four different species that infect African rodents have been adapted for laboratory use: Plasmodium berghei, P. yoelii, P. chabaudi and P. vinckei. Small differences exist in the biology of the different RMP in laboratory mice and this makes them particularly attractive models to investigate different aspects of human malaria. Specifically, P. chabaudi is a model to investigate mechanisms of drug resistances and immune evasion, in particular antigenic variation [3,4]. It invades normocytes and reticulocytes and mostly produces chronic, non-lethal, infections. In contrast, P. berghei preferentially invades reticulocytes and usually produces infections in mice that induce severe pathology [2]. In combination with different mouse strains it has been used as a model to study immunopathology, experimental cerebral malaria, pregnancy-associated malaria and lung pathology [2]. P. yoelii is widely used in studies on the biology

* Correspondence: mb4@sanger.ac.uk; c.j.janse@lumc.nl
† Equal contributors
1 Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK
2 Leiden Malaria Research Group, Department of Parasitology, Leiden University Medical Center, Leiden, The Netherlands
Full list of author information is available at the end of the article

© 2014 Otto et al; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.
of liver stages and on innate and acquired immunity against liver stages [5,6]. Blood stage \( P. yoelii \) parasites of some lines are restricted to reticulocytes whereas others can invade all red blood cells and have been used to study receptors for erythrocyte binding [7,8]. The availability of efficient reverse genetics technologies for \( P. berghei \) and \( P. yoelii \) [9-11] and the ability to analyse these parasites throughout the complete life cycle have made these two species the preferred models for analysis of \( Plasmodium \) gene function [12-14]. For these two species more than 600 different genetically modified mutants have been reported [15].

The first draft RMP genome was published in 2002 for \( P. yoelii \) YM 17XL [16]. This was followed by publication of draft genomes of \( P. berghei \) ANKA (\( PbA \)) and \( P. chabaudi \) chabaudi AS (\( PcAS \)) in 2005 [17]. Comparisons with the genome of the human parasite \( P. falciparum \) and other primate malaria species defined a large set of core genes that are shared between RMPs and primate malarias [18-20]. Although availability of draft RMP genomes made a significant impact in applying post-genomic technologies for understanding malaria biology [18] and were used in many follow-up functional genomics studies to analyse gene regulation and function [9,10], these RMP genomes were highly fragmented and were annotated with little or no manual curation. The fragmented nature of the genomes has hampered genome wide analysis of gene regulation and function, especially of the (subtelomeric) multigene families. To utilise RMP models to their full potential, we therefore undertook production of high quality reference genomes: for \( PbA \) and \( PcAS \) large-scale improvement of their existing genomes, with re-sequencing, re-analysis and manual re-annotation, and for \( P. y. yoelii \) a genome sequence was produced de novo from the virulent YM line using the latest sequencing technologies and computational algorithms. In addition, we have utilised comprehensive RNA-seq data derived from a number of life-cycle stages to both improve gene model prediction and to provide genome-wide, quantitative data on gene expression. By sequencing additional isolates/lines of \( P. berghei \), \( P. yoelii \) and \( P. chabaudi \) (including the subspecies \( P. c. adami \)) we have documented genotypic diversity that exists within different RMP species. The availability of RMP reference genomes in combination with the RNA-seq and genotypic diversity data will serve as excellent resources for gene function and post-genomic analyses and, therefore, better interrogation of \( Plasmodium \) biology and development of anti-malaria interventions.

The genomes of RMP contain a number of multigene families located in the subtelomeric chromosomal regions. These include a large family of so-called ‘\( Plasmodium \) interspersed repeat genes’ (\( pir \)) [16], that are present also in other human/primate \( Plasmodium \) species [20-23]. Most of these gene families are expressed in blood stages and these proteins show features that have been reported to contribute to immune evasion through antigenic variation [24-26] and may play a role in the sequestration of infected red blood cells and virulence [26,27]. As a result of the improved annotation, we have been able to define all multigene families in the RMP genomes. Comparative phylogenetic analyses of the \( pir \) genes and analyses of \( pir \) expression patterns in blood stages of \( P. berghei \) provide evidence of functional diversification within this gene family. The improved classification of multigene families will enhance studies on the role of (variant) exported proteins in virulence and evasion and modulation of the immune system.

Results
Generation of high-quality RMP reference genomes
With a combination of Sanger and second generation sequencing (that is, Illumina and 454), automated scaffolding, gap closure, error correction and annotation transfer, followed by manual inspection, we obtained highly accurate and almost complete reference genomes of \( PbA \), \( PcAS \) and \( P. y. yoelii \) YM (\( PyYM \)). This resulted in a significant reduction in contig number for the \( PbA \) and \( PcAS \) genomes (Table 1) compared with existing highly fragmented drafts [16,17]. The new assemblies contain 4,979, 5,139 and 5,675 protein-coding genes for \( PbA \), \( PcAS \) and \( PyYM \), respectively, with 487 and 409 novel genes in the genomes of \( PbA \) and \( PcAS \) that were absent in the draft genomes (Table 1, Additional file 1). More than 98% of the predicted genes are now present as full-length gene models and we were able to ascribe putative functions (that is, they are not annotated as encoding hypothetical proteins of unknown function) to 56% to 61% of these. This percentage is comparable to the 60% of the \( P. falciparum \) genes that have annotated functions. As a result of eliminating incomplete gene models and merging multiple incorrect gene models into single gene models and by removing mouse DNA sequence contamination, only 63% and 77% of the previously annotated \( PbA \) and \( PcAS \) genes were mapped back to the new genomes. The RMP reference genomes have a size of 18.5 to 21.9 Mb (Table 1), confirming the smaller genome sizes of RMPs compared with primate malaria species but both the mitochondrial and apicoplast RMP genomes are highly comparable in size and gene content to those of \( P. falciparum \) (Table 1). The predicted proteomes were analysed for the presence of PEXEL-motifs, a characteristic of host-exported proteins, using ExportPred v2.0 [28]. Between 97 and 119 PEXEL-positive proteins were predicted for the different RMP. This indicates that, like \( P. berghei \), the other RMP also contain three times more PEXEL-positive proteins than was previously predicted [29] (see Additional file 2).
Conserved chromosome organization and gene orthology between RMP and primate malaria parasites

The improved genomes confirmed the extensive conservation of the RMP genomes and the presence of only a single synteny breakpoint ([19]; Table 1). Despite the highly conserved internal regions of the 14 chromosomes, species-specific paralogous expansion and diversification of certain genes has occurred in each genome. Additional file 3A shows an example of such an expanded locus within a region of conserved synteny between \textit{Pc} AS and \textit{Py} YM. In \textit{Pb} A only a single copy (PBANKA\_091920) is present, whereas the genomes of \textit{Pc} AS and \textit{Py} YM contain multiple copies (\textit{fam-d}; Table 2, Additional file 4), organized as a single gene cluster on chromosome 9.

The re-annotation resulted in a better characterization of several non-coding and coding features of the chromosomes such as the centromeric and subtelomeric regions. As an example we show in Additional file 3B the size, location and GC-content of a positionally conserved centromere-containing region of chromosome 7 of \textit{Pb} A and \textit{Pc} AS. Figure 1A shows an example of the organization of RMP subtelomeric regions, visualizing the location of genes of multigene families. Although these regions contain members of multigene families that are shared between RMP, they are highly variable as a result of variation in gene copy number and the presence of species-specific genes and (non-coding) repeat sequences. For example, several \textit{Pb} A subtelomeric regions contain many copies of a large, 2.3 kb, repeat element that is \textit{P. berghei}-specific (Figure 1B; [30]). Based on the coverage-depth of Illumina sequence data mapped onto the 2.3 kb repeats, we estimated that the \textit{Pb} A genome contains about 400 copies, representing approximately 4% of the total nucleotide content. In the \textit{Pb} A genome only 47 of these repeats have been assembled and the remaining copies (approximately 350) are either located as clusters in the sequence gaps that still exist in the \textit{Pb} A subtelomeric regions or are located within the existing current 2.3 kb repeat-arrays. These 2.3 kb repeats contain telomeric repeat sequences and many contain a copy of a highly degenerate \textit{pir} pseudogene (Figure 1B; [30]) suggesting that the expansion of this repeat may have originally been driven by an expansion of \textit{pir} gene numbers.

We compared all predicted RMP protein-coding genes with those of three primate malaria species, \textit{P. falciparum}, \textit{P. knowlesi} and \textit{P. vivax} using OrthoMCL and divided the predicted RMP proteome into three different categories: (1) RMP proteins with orthologs in any of the primate malarials; (2) RMP-specific proteins with no orthologs in primate malarials; and (3) primate malaria-specific proteins.
with no orthologs in any of the RMP (see Additional file 5).

Between the predicted RMP proteomes (15,793 proteins in total) and primate malaria proteomes (15,853 proteins in total), approximately 87% of the RMP proteins had detectable orthologs in at least one of the primate malarias and only 2,104 proteins (13.3%) were predicted to be RMP-specific. Of those 2,104 proteins, 1,854 (88.1%) are from gene families, as defined in Additional file 4.

For 2,306 primate malaria proteins (14.6%) no orthologs have been detected in the RMP. Of these primate malaria specific genes, approximately 1,635 (70.9%) are subtelomeric genes or members of subtelomeric gene families (see Additional file 5).

Genotypic diversity within RMP isolates: *P. chabaudi* isolates exhibit high level polymorphism amongst their genes

The availability of multiple isolates of RMP with different phenotypic traits offers the possibility of using genetics to study phenotype/genotype associations. To quantify the level of genotypic diversity across multiple RMP isolates we produced genome sequence data at a 79- to 437-fold coverage, from isolates of *P. berghei* (NK65, K173, SP11 and its pyrimethamine resistant descendant SP11 RLL), *P. c. chabaudi* (AJ, CB) and *P. c. adami* (DK, DS). In addition, we sequenced *P. y. yoelii* 17X (see Additional file 6 for parasite selection rationale and Additional file 7 for parasite origins). Single nucleotide polymorphisms (SNPs) in the genomes of these parasites were called by mapping the reads against their respective reference genomes (Table 3, Additional file 8) after excluding repetitive or low complexity regions of genes and members of multigene families (genes in Additional file 4). The level of polymorphism between the four *P. berghei* isolates is surprisingly low with SNPs detected in only 4 to 469 genes (Table 3, Additional file 8). In *P. berghei* the highest SNP numbers were found in two lines, SP11 RLL and K173cl1, which have been maintained in the laboratory for prolonged periods by mechanical blood passage between mice. Comparison of the genomes of SP11 and its pyrimethamine resistant descendant SP11 RLL, revealed the point mutation in the dihydrofolate reductase-thymidylate synthase gene, known to be involved in pyrimethamine resistance (see Additional file 8; [31]). Comparing the genomes of the non-lethal *P. y. yoelii* 17X isolate and its virulent descendant laboratory line YM showed limited polymorphism and revealed SNPs in the Duffy-binding protein that have been implicated in the different invasion and virulence phenotypes of these two lines (see Additional file 8; [7]). In contrast to the low numbers of *P. y. yoelii* genes with SNPs (eight genes), large differences exist in gene copy number of subtelomeric multigene families (Table 3) which accounts for the difference in genome size between the two laboratory lines. In contrast to the *P. berghei* isolates, the *P. chabaudi* isolates and subspecies have much higher SNP densities with 4,300 to 4,500 (out of 4,576) non-subtelomeric genes having at least one SNP (Table 3, Additional file 8). The high genotypic diversity is not only evident between the subspecies *P. c. chabaudi* and *P. c. adami*, but also between isolates of the same subspecies. For example, we found 94,668 and 71,074 unique SNPs (in 3,978 and 4,166 genes) in the *P. c adami* DK and DS isolates, respectively. Between different *P. chabaudi* isolates differences exist in virulence- and invasion phenotypes of blood stage infections (see Additional file 6). We detected multiple SNPs in

### Table 2 Different (subtelomeric) multigene families in the RMP genomes

| Gene family (new name) | Other (previous) names | PBA | PcAS | PyYM |
|------------------------|------------------------|-----|------|------|
|                        |                        | CG  | PSG  | FG   | CG  | PSG  | FG   | CG  | PSG  | FG   |
| *pir*                  | *pir, bir, cir, yir*   | 100 | 88   | 12   | 194 | 3    | 4    | 583 | 40   | 172  |
| *RMP-fam-a*            | *Pb-fam-1; Pc-fam-1; fam-a; PYSTA* | 23  | 16   | 3    | 132 | 2    | 0    | 94  | 8    | 11   |
| *RMP-fam-b*            | *Pb-fam-3; PYSTB*      | 34  | 1    | 5    | 26  | 0    | 0    | 48  | 2    | 4    |
| *RMP-fam-c*            | *PYSTC*                | 6   | 0    | -    | 10  | 0    | -    | 22  | 0    | -    |
| *RMP-fam-d*            | *Pc-fam*               | 1   | 0    | -    | 17  | 4    | -    | 5   | 0    | -    |
| *Early transcribed membrane protein* | *etramp*               | 7   | -    | -    | 13  | -    | -    | 12  | -    | -    |
| *Reticulocyte binding protein, putative* | *P235; 235kDA protein* | 6   | -    | 8    | 8   | -    | 0    | 11  | -    | 3    |
|                        | *rhoptry protein, putative* |       |       |       |       |       |       |       |       |       |
| *Lysophospholipase*    |                        | 4   | 1    | -    | 28  | 0    | -    | 11  | 1    | -    |
| *RMP-erythrocyte membrane antigen (RMP-EMA1)* | *pcema1*               | 1   | 0    | 0    | 13  | 1    | 1    | 11  | 0    | 0    |
| *haloacid dehalogenase-like hydrolase, putative* |       | 1   | -    | -    | 9   | -    | -    | 1   | -    | -    |
| *‘Other subtelomeric genes’* |                        | 46  | -    | -    | 67  | -    | -    | 46  | -    | -    |

See Additional file 4 for details of all genes of *P. berghei* ANKA (*PbA*), *P. c. chabaudi* AS (*PcAS*) and *P. y. yoelii* YM (*PyYM*). CG: complete gene; FG: fragment; PSG: pseudogene.
*P. chabaudi* genes involved in binding to red blood cells (RBC) such as the Duffy-binding protein and reticulocyte binding proteins (see Additional file 8), genes that are associated with differences in virulence between *P. y. yoelii* lines. Isolate-specific protective immunity between *P. c. chabaudi* isolates has been linked to the merozoite surface protein 1 (MSP1; PCHAS_083130) [32,33]. Our analyses revealed an excess of non-synonymous substitutions (reflected in high Ka/Ks values) in *msp1* of all *P. chabaudi* isolates (see Additional file 8).

**High resolution, genome-wide expression data from different RMP life cycle stages**

To further improve gene annotation and to provide foundational data for gene-function studies, we generated RNA-seq data from several life-cycle stages. RNA was analysed...
Table 3 Sequence diversity and number of multigene families from different RMP isolates/lines

| Isolate/line | Genome coverage | SNPs<sup>b</sup> | Genes with SNPs<sup>b</sup> | Assembly size (Mb) | Contigs | pir genes<sup>c</sup> | fam-a genes<sup>c</sup> | fam-b genes<sup>c</sup> | fam-c genes<sup>c</sup> | fam-d genes<sup>c</sup> |
|--------------|-----------------|-----------------|--------------------------|-------------------|---------|-----------------|-----------------|-----------------|-----------------|-----------------|
| *P. berghei* ANKA<sup>a</sup> | -               | -               | -                        | -                 | -       | -               | -               | -               | -               | -               |
| *P. berghei* NK65 E | 437x             | 294             | 21                       | 18.56             | 220     | 200             | 42              | 40              | 6               | 1               |
| *P. berghei* NK65NY | 161x             | 127             | 14                       | 18.47             | 310     | 191             | 54              | 47              | 5               | 1               |
| *P. berghei* SP11 A | 268x             | 95              | 4                        | 18.44             | 333     | 182             | 54              | 45              | 7               | 1               |
| *P. berghei* SP11 RLL A | 401x           | 2,098           | 345                      | 18.32             | 275     | 170             | 58              | 50              | 6               | 1               |
| *P. berghei* K173cl1 | 262x             | 2,759           | 469                      | 18.72             | 123     | 224             | 54              | 41              | 7               | 1               |
| *P. y. yoelli* YM<sup>a</sup> | -               | -               | -                        | -                 | 22.03   | 195             | 795             | 54              | 22              | 5               |
| *P. y. yoelli* 17X | 289x             | 740             | 8                        | 22.75             | 154     | 980             | 157             | 78              | 26              | 5               |
| *P. c. chabaudi* AS<sup>a</sup> | -               | -               | -                        | -                 | 18.83   | 40              | 201             | 134             | 26              | 10              |
| *P. c. chabaudi* CB | 79x             | 144,148         | 4,321                     | 18.98             | 246     | 276             | 160             | 31              | 23              | 21              |
| *P. c. chabaudi* AJ | 127x             | 144,281         | 4,326                     | 18.89             | 240     | 277             | 161             | 28              | 18              | 19              |
| *P. c. adami* DS | 98x             | 251,984         | 4,514                     | 19.63             | 272     | 371             | 215             | 37              | 33              | 29              |
| *P. c. adami* DK | 248x             | 274,877         | 4,509                     | 19.44             | 275     | 398             | 193             | 41              | 32              | 25              |

<sup>a</sup>Reference genomes that is, *PbA, PcAS* and *PyYM* to which sequence data from other isolates were mapped and analysed. 
<sup>b</sup>Excluded from the analysis: all subtelomERICally located genes (as mentioned in Additional file 4) and repetitive and low complexity regions of genes. Only single nucleotide polymorphisms (SNPs) were counted with at least 10 high quality mapped reads, 90% allele and 20% calls on each strand (see Additional file 8 for details of the SNPs in subtelomERICally located genes (as mentioned in Additional file 4) and repetitive and low complexity regions of genes. Only single nucleotide polymorphisms (SNPs) were counted with at least 10 high quality mapped reads, 90% allele and 20% calls on each strand (see Additional file 8 for details of the SNPs in individual genes); <sup>c</sup>including pseudogenes and fragments. RMP, rodent malaria parasites.

from synchronised *PbA* asexual blood stages (ring forms, late trophozoites and schizonts) and from purified gametocytes and ookinetes. In addition, RNA-seq data was generated from multiple samples of blood stage trophozoites of *PcAS* and from blood stages of *PyYM* (see Additional file 9). To analyse the reproducibility of our RNA-seq data, we calculated Pearson correlations of the FPKM (fragments per kilo base of exon per million fragments mapped) values of RMP genes for which one-to-one ortholog relationships exist in the different RPM genomes (Figure 2A). Expression was highly correlated not only between biological replicates of the same species (r = 0.88 to 1.0), but also between comparable stages of different RMP, such as *PbA* and *PcAS* trophozoites (r = 0.77 to 0.86). Both the gametocyte and ookinete samples clustered separately from asexual blood stages, which reflects the different program of gene expression during sexual commitment and zygote development. Heat maps representing the expression of all *PbA* genes reveal clusters of genes with distinct expression patterns in the different life cycle stages (Figure 2B, left panel), consistent with both the morphological and functional differences between these stages. When *PbA* genes are ordered according to the expression levels of their *P. falciparum* orthologs [34], ring-, trophozoite- and schizont-expressed genes display the expected characteristic temporal cascade of gene expression (Figure 2B, right panel). Genome wide expression data of developing ookinetes have not been published before. We found that mature (24 hour) ookinetes have a distinct expression pattern compared to immature, developing ookinetes (16 hour). In Additional file 10 an overview is presented of all genes that are up- or down regulated in the two developmental stages of ookinetes. Genome ontology (GO)-annotation of differentially regulated genes reveals that genes encoding proteins involved in protein phosphorylation, inner membrane and myosin complex formation and ATP binding are most significantly up-regulated in 16 hour ookinetes compared to 24 hour ookinetes (see Additional file 10). In contrast, mature ookinetes show a strong up-regulation of genes encoding proteins involved in protein translation and ribosome formation (see Additional file 10), most likely in preparation for the rapid growth expansion of the oocyst after ookinete traversal of the mosquito midgut wall.

To further improve the reference genomes we mapped the RNA-seq data onto the RMP genomes and visually inspected the alignments using the Artemis Comparison Tool (ACT), a genome viewing tool [35]. A comparative analysis with the *P. falciparum* 3D7 genome allowed us to determine gene structure at base-pair (bp) resolution for at least 89% of the genes. Of the 896 newly annotated protein-coding genes that were absent in the previous genome assemblies, 70% have primate malaria orthologs, 83% have expression evidence (RNA-seq FPKM values >21) and we could ascribe functions to 75% (see Additional file 1). The different RNA-seq data sets have also been used to confirm splice sites and to identify putative alternative splice sites (see Additional file 11). This analysis resulted in the identification of 839 alternative splicing events in a total of 567 RMP genes.

Characterization of RMP multigene families

As a result of having dramatically improved the annotation of the subtelomERIC regions we were able to accurately
### A Correlation expression (RNA-seq) in RMP life-cycle stages

![Heatmap showing correlation expression](image)

### B Level of PbA gene expression

- **Ordered through ab initio clustering**
- **Ordered according to Pf gene expression**

![Heatmap showing PbA gene expression levels](image)

**Figure 2** (See legend on next page.)
define the RMP multigene families that are located there (Table 2, Additional file 4). For proteins of nearly all of these families experimental evidence exists that they are exported into the host RBC in the absence of a PEXEL motif [29]. The pir family is the most abundant multigene family (see next section) encoding exported proteins that lack a canonical PEXEL motif. The second largest gene family is the fam-a gene family, formerly identified as the pyst-a family in \textit{P. yoelii} 17XLN and named as \textit{Pb}-fam-1, \textit{Pc}-fam-1 or fam-a [16,17]. \textit{Pb} fam-a proteins are exported into the host RBC and can be transported to the RBC surface membrane [29] but lack a PEXEL-motif. Single copy orthologs have been defined in all primate malaria and the expansion of this family is RMP-specific. Most members have a subtelomeric location (see Additional files 12, 13 and 14), but all three RMP have at least one internally located copy that is positionally conserved with the primate malaria orthologs and, therefore, likely to represent the ancestral copy of this family. In order to standardise the naming of orthologous multigene families in different RMP, we have renamed the two multigene families, \textit{pyst-b}/\textit{pb}-fam-3 and \textit{pyst-c} genes [16,17,29] as fam-b and fam-c, respectively (Table 2; Additional file 4). The fam-b family is exclusively subtelomeric and is characterized by the presence of the \textit{pyst-b} domain. Most members contain a transmembrane domain (58%), a signal peptide (75%) and PEXEL-motif (76%) (see Additional files 12, 13 and 14). \textit{Pb} fam-b proteins are exported into the host RBC [29]. The fam-c is also exclusively found in the subtelomeric regions and is characterized by the presence of a \textit{pyst-c1} and/or \textit{pyst-c2} domain [16]. Most members have a transmembrane domain (60%) and a signal peptide (92%) (see Additional files 12, 13 and 14) and only a small percentage (24%) contain a predicted PEXEL-motif.

Other subtelomeric multigene families include the ‘early transcribed family of proteins’ (ETRAMPs) and ‘putative reticulocyte binding proteins’ (Table 2, Additional files 4, 12, 13 and 14). ETRAMPs are small exported proteins with a predicted signal peptide and transmembrane domain but without a PEXEL-motif. These proteins are mainly located in the parasitophorous vacuole membrane [36,37]. The genes encoding putative reticulocyte binding proteins (RBP), that were first described in \textit{P. yoelii} as Py235 and are expressed in merozoites [38], are clear orthologs of the reticulocyte binding proteins of \textit{P. vivax} [39] and the RH proteins of \textit{P. falciparum} [40]. These large proteins typically have a predicted signal sequence and at the C-terminus a transmembrane domain containing a rhomboid cleavage site and a cytoplasmic domain, although \textit{P. falciparum} RH5 contains just the signal peptide and N-terminal ligand binding domain [41]. The RMPs have genes encoding two short RRBpS reminiscent of \textit{P. falciparum} RH5 (typified by PYYM_0101400 and PYYM_0701100) and six or more full length proteins (Table 2). Compared with PYYM, Py17X contains an additional full length RBP.

In \textit{Pc}AS several other expanded gene families are present in the subtelomeric regions. These include ‘putative lysophospholipases’, ‘erythrocyte membrane antigen 1’ (EMA1), and ‘putative haloacid dehalogenase-like hydrolases’ (Table 2, Additional files 4, 12, 13 and 14). The genes encoding lysophospholipases are characterized by the ‘pyst-a’ domain [42] and all RMP have two copies with an internal chromosomal location that are syntenic with orthologs of primate malarias. For two of the five \textit{Pb} lysophospholipases evidence exists that they are exported into the RBC [29] and again they lack a PEXEL-motif. In \textit{Pc}AS this family has expanded into 28 copies (Table 2, Additional file 4). In the genome of \textit{Pb}YM and \textit{Pb}Ao n l ya several other expanded gene families are present in the subtelomeric regions. Many of these ‘other subtelomeric genes’ (46 to 67 genes; Table 2, Additional file 4) encode proteins that are RMP-specific and more than 96% of these proteins contain a predicted signal peptide, transmembrane domain or PEXEL-motif and for several proteins experimental evidence exists for their export into the host RBC cytoplasm. Combined, these observations indicate that most, if not all, RMP subtelomeric genes (apart from the RBP family) encode exported proteins and most lack a PEXEL-motif. The presence of large numbers of PEXEL-negative exported proteins in RMP indicates alternative export mechanisms possibly
common to all *Plasmodium* species and investigations with highly tractable RMP species can, therefore, be used to understand these mechanisms better.

**The RMP pir multigene family: phylogeny and expression**

We analysed the expression patterns of all members of the three largest multigene families, *fam-a*, *fam-b* and *pir* in the *PbA* life cycle stages using heat maps of the RNA-seq data. This revealed distinct transcription patterns both between the gene families and also between members within a family (Figure 3). All three families show strongly reduced transcription in ookinetes. Whereas most RMP-*fam-a* and RMP-*fam-b* members had reduced transcript levels in gametocytes compared to asexual blood stages, a large cluster of *pir* genes were up-regulated (at least a fold change of 2) in gametocytes. Expression patterns are not only different between asexual and sexual stages but also between different asexual stages, for example distinct *pir* gene clusters are up-regulated in schizonts. Distinct transcription patterns in different life cycle stages of gene clusters may indicate functional differences between members of a single gene family. With the new genome assemblies we were able to determine the total number of *pirs* and their structure and spatial organization more precisely (Table 2; Additional file 4). In *PcAS* and *PbA* the total number of *pirs* (excluding pseudogenes) is 194 and 100, respectively, whereas in *PyYM* this gene family is greatly expanded to 583 copies. In Additional files 12, 13 and 14

---

**Expression of gene families in *P. berghei* life-cycle stages**

![Heatmap of expression patterns](image)

**Figure 3** Expression of members of three large RMP multigene families in different life cycle stages. Temporal expression patterns of members of the three largest *PbA* multigene families (*pir, fam-a* and *fam-b*) in different life cycle stages as visualized by heat maps of RNA-seq data. The expression (FKPM) values of genes over the life-cycle stages (yellow-red) are normalised per gene. The min/max column values are the log minimal and log maximal FPKM values for each gene. Only genes with an FPKM above 21, in all conditions, were included. FKPM, fragments per kilo base of exon per million fragments mapped; RMP, rodent malaria parasites.
the chromosomal distribution of all pirs is shown. Most pir genes share a similar structure across the different species with a short first exon, long second exon and a third exon encoding a trans-membrane domain. They lack a PEXEL-motif and all pirs are chromosomally arranged such that they are transcribed in a centromere to telomere direction except for several members of PcAS (these are ‘long-form’ pirs of clade L1; see below). A number of pirs have long low complexity regions in the predicted extracellular domain and a few Pc-pirs have a four exon structure (see Additional file 3C). Remarkably, as stated earlier, the Pb-pir genes include a large number (88 genes; 44%) of pseudogenes and nearly half (35 genes; 18%) of these Pb-pirs are contained within the 2.3 kb subtelo-meric repeat described above (Figure 1B). To analyse whether the differential expression of groups of pir members was associated with definable sequence differences (and possibly functional differences) between pirs we undertook a detailed phylogenetic analysis of all RMP pirs (including predicted pseudogenes). Estimations of Maximum Likelihood (ML) phylogeny based on nucleotide sequences or amino acid sequences and an estimation of a Bayesian phylogeny resulted in a phylogenetic tree with a robust separation of ‘long-form’ and ‘short form’ pirs (Figure 4; Additional file 15). We identified 12 clades in the phylogeny that have robust support; four long-form clades (L1 to 4) with a mean pir length ranging from 1,062 to 2,369 aa and eight short-form clades (S1 to 8) with a mean length ranging from 786 to 952 aa. Most long-form pirs have an extended repetitive region located within the second exon, downstream of the core pir do- main and upstream of the transmembrane region. All RMP species have both short- and long-form pirs, indicat-ing that the presence or absence of an extended repetitive region has evolved once and defines a principle division in pir diversity. Many clades are dominated by pirs from one species, particularly PyYM (for example, clades S1, S4, S8; Figure 4). Yet, even such clades contain rare sequence types from PbA (for example, S1d, S2, S8) or PcAS (S1g) indicating that these lineages originate from the RMP an-cestor and probably expanded after speciation. Both PcAS and PbA appear to have experienced their own specific expansions after speciation (for example, S4, S5, S7). The maintenance of orthology within clades, in the presence of frequent gene conversion (see Discussion section), may indicate that selection pressure maintains structural differences between pirs, for example diversifying selection under immune pressure or purifying selection on functional diversity. The observation that the ratio of the different pir clades is highly similar in the five, highly di-verse, isolates of the two P. chabaudi subspecies (Figure 4) supports the presence of selective pressures that main-tain the clade structure of pir genes. We next analysed whether the pir expression patterns in PbA blood stages were correlated with structural differences between pirs, by comparing the RNA-seq expression patterns with the phylogenetic clades. We observed that pirs that are pre-dominantly expressed in gametocytes mainly belong to only two clades of the small-form pirs, S1 and S4, whereas genes up-regulated in gametocytes are mainly long-forms of clades L1, L2 and L3 (see Additional file 16). The stage-specific up- or down-regulation of expression of clusters of structurally different pirs support the hypothesis of the existence of functional diversification within the pir family and is in agreement with other observations indicating that differences in pir sequences are associated with different functional properties [44,45].

Discussion

By extensive re-sequencing and annotation we have generated three high quality RMP reference genomes with nearly all core genes as complete gene models and a much improved and almost complete representation of chromo-somal subtelomeric regions. These reference genomes will greatly enhance the use of RMP as model organisms in malaria research. We provide full-length gene models for more than 98% of predicted protein-coding genes. The approximately 60% of genes with functional annotation is comparable to the percentage of functionally annotated genes in the P. falciparum 3D7 reference genome and a high percentage (approximately 90%) of the predicted RMP proteins have orthologs in primate malaria species. It is this high level of orthology between RMP and primate malaria genomes that strongly supports RMPs as models in experimental approaches to characterize the Plasmo-dium gene function. Similarly, the genome-wide RNA-seq data from different RMP developmental stages is a valu-able resource to further analyse Plasmodium gene func-tion and the regulatory networks underlying the multiple differentiation pathways of Plasmodium. The RNA-seq studies presented here provide information on gene ex-pression at an unprecedented depth and breadth of cover-age of multiple blood stages and ookinetes. Previously, only a few large-scale transcriptome (microarray) analyses of P. berghei blood stages and ookinetes had been per-formed [17,46]. These studies were based on a highly frag-mented draft P. berghei genome and, therefore, expression data were only generated for about half of all P. berghei genes. In addition, important/valuable large scale tran-scriptome studies have been performed on RMP life-cycle stages, such as sporozoites and liver stages [47-49]. These life-cycle stages would also benefit from re-examination using the latest RMP genome assemblies we provide in this study.

Our studies reveal that large scale changes in gene ex-pression occur in ookinetes between 16 and 24 hours after fertilization, possibly required for the differentiation of (retort-form) zygotes into the mature ookinetes. The
strong up-regulation in mature ookinetes of transcripts involved in ribosome biogenesis and protein translation suggest that the mature ookinete generates transcripts for proteins required after the ookinete has traversed the mosquito midgut wall and starts its rapid transition into the oocysts, possibly using mechanisms of translational repression similar to those in gametocytes [50,51] and sporozoites [52,53]. What these three stages have in common is that they are fully differentiated cells that will undergo rapid cellular differentiation and/or growth expansion upon entering a new environment. Whether mature ookinetes store repressed transcripts requires further investigation.

The additional sequence data from multiple RMP isolates will help to further unravel gene function and establish relationships between phenotypic traits and genotypic diversity. The near absence of polymorphisms within the genomes of \textit{P. berghei} isolates was unexpected. Low sequence diversity of a limited number of genes of \textit{P. berghei} isolates had been reported previously and it was proposed that this may result from cross-contamination.
of *P. berghei* isolates in the laboratory after isolation [54,55]. However, this seems unlikely as one line would have needed to be mislabelled with the names of all other isolates, then all these mislabelled lines would have had to be sent to all the different laboratories worldwide replacing the ‘correct’ isolates that may have existed in their recipient laboratories. However, sequencing of additional stocks from these isolates, which were frozen in different laboratories soon after isolation from the natural host, may reveal whether low sequence diversity is due to cross-contamination. The *P. berghei* isolates we have sequenced were obtained from other laboratories (SP11, NK65) and they also show a similar lack of sequence polymorphism.

In contrast, the isolates of *P. chabaudi* exhibit considerable genotypic diversity. These *P. chabaudi* isolates exhibit differences in virulence, RBC invasion, growth rates and immunogenic profiles [7,56-60] and further studies, for example using linkage or quantitative trait loci (QTL) analyses [33,61-63], will facilitate identification of genes associated with defined phenotypes. For RMP species there is evidence that differences in virulence are associated with differences in RBC invasion [2,7,56,60,64]. For example, *P. yoelii* virulence has been associated with mutations in proteins involved in RBC invasion [7,65,66]. Interestingly, we found extensive sequence polymorphism in *P. chabaudi* genes encoding such proteins. While much attention is given to the role of exported proteins of multigene families and virulence in both human and RMP, further analysis of RMP proteins that regulate invasion phenotypes may reveal novel mechanisms that underlie virulence.

The new sequence data allowed for a much improved annotation of chromosomal subtelomeric regions and to better define the different subtelomeric multigene families. In addition to the large pir gene family, all three RMP contain an expanded gene family encoding exported proteins, fam-a, with orthology to a single-copy gene in primate malaria, which contains a START-domain (steroidogenic acute regulatory-related lipid transfer domain; [67]). START-containing proteins of eukaryotes are involved in the transfer of phospholipids, ceramide or fatty acids between membranes [68]. A START domain has also recently been identified in an exported, PEXEL-containing, *P. falciparum* protein that was shown to transfer phospholipids [69]. The single-copy RMP orthologs of this gene (PF3D7_0104200) also contain a PEXEL-motif, indicating that phospholipid-transporting proteins are exported into the RBC in both primate malarias and RMP. *P. chabaudi* contains an additional, highly expanded, gene family that contains domains involved in phospholipid/fatty acid metabolism. These genes, encoding putative lysophospholipases, lack a PEXEL motif; however, for several *P. berghei* orthologs as well as lysophospholipases of *P. falciparum* there is evidence for their export into the host RBC [29,70]. Combined, these observations indicate the importance of phospholipid/fatty acid metabolism/transport mediated by *Plasmodium* proteins exported into the RBC cytosol. Why such genes have been differentially expanded into multigene families in different species remains to be investigated.

The pir family is the largest RMP multigene family and is shared with human and non-human primate species *P. vivax, P. knowlesi* and *P. cynomolgi* [20-23]. PIR proteins are exported into the RBC in the absence of a PEXEL-motif, and there is evidence that they are located on, or close to, the RBC surface or dispersed in the RBC cytoplasm [24-26,29,71,72]. The function of pir is unknown and no functional domains have been identified so far. Recently, it has been shown that in *P. chabaudi* a change in virulence was associated with differential expression of members of the pir multi-gene family [27]. It has been suggested that PIRs are transported to the surface of infected RBC and play a role in RBC sequestration comparable to the role of the Pfemp1 gene family of virulence factors in *P. falciparum*. However, for several *P. berghei* PIRs a direct role in RBC sequestration is unlikely since no evidence was found for their location on the RBC surface although they were exported into the RBC cytoplasm of both sequestering asexual blood stages and non-sequestering gametocytes [29]. For *P. vivax* PIRs it has been shown that different members have distinct subcellular locations in the infected RBC [26]. These observations indicate that functional differences may exist between members of the PIR family. Phylogenetic analyses support the possibility of functional differences between the PIRs. A recent phylogenetic analysis of the newly annotated PcaS pir identified two distinct pir sub-families (A and B), which contain distinct amino acid sequence motifs [44]. Our phylogenetic analyses included pir from all three RMP species and resulted in the identification of a number of different clades. The presence of clearly distinguishable clades indicates that structural differentiation exists among pir and that this evolved prior to the separation of the RMP species. Our observations of the stagespecific up- or down-regulation of expression of clusters of structurally different pir in different blood stages supports the hypothesis that there is functional diversification within the pir family and that purifying selection plays a role in shaping this family. By including multiple species in the pir phylogeny it is clear that this gene family is subject to rapid turnover, that is, gene gain and loss, indicating the absence of strong selective forces that would result in distinct orthologous groups/clades that are shared and maintained in different species for functional reasons. Gain of pir genes in different species is evident in the multiple species-specific expansions of clades. Assuming that the common ancestor had a pir family equal in abundance and diversity, the relatively limited instances of orthology
(12 clades) indicates significant losses of ancestral sequence types. A plausible explanation for both the abundance of species-specific sequences and the paucity of ancestral sequences is a continual process of gene turnover driven by gene conversion, a mechanism that has been proposed for pirs of P. chabaudi [44] and which was evident in each of the clades revealed in this study (data not shown). The effect of frequent gene conversion is the replacement of ancestral sequence types with species-specific sequences, which results in distinct species-specific clades without orthology. Loss of orthology is only resisted when selective forces maintain structurally distinct pirs, which we propose, explains the presence of the (limited) orthology between pir clades of the different RMP species. The improved annotation and phylogeny demonstrating clusters of structurally different pirs in all RMP combined with expression profiles are powerful data that can help to further delineate function, the relationship of expression with virulence and how the (species-specific) expansion of the pirs is related to distinct selective pressures.

Conclusions
To maximise the utility of RMP we have greatly improved the genome assemblies of P. berghei and P. chabaudi, comprehensively sequenced the P. yoelii YM genome, sequenced multiple RMP isolates and generated in-depth expression data from multiple RMP life-cycle stages. Comparison of the RMP and P. falciparum genomes and RNA-seq mapping permitted gene annotation at base-pair resolution and has defined the level of orthology between RMP and human parasite genomes. The very high level orthology between RMP and human malarias (both in genome structure and gene content) supports the use of highly tractable RMPs as experimental models to characterize the function of the very many Plasmodium genes that remain uncharacterised. Only a few large-scale transcriptome (microarray) analyses of different P. berghei life-cycle stages had previously been performed. Moreover, these studies were based on highly fragmented draft RMP genomes and consequently, for example, for one of the most well studied RMP, P. berghei, gene expression data was only mapped to about half of all P. berghei genes that have now been characterised. The RNA-seq studies we present provide information on gene expression, at an unprecedented depth and breadth of coverage, of multiple life cycle stages and provide the foundational data needed for the performance of large-scale analyses of gene regulatory networks that underlie cellular differentiation.

We show that extensive genotypic diversity exists between P. chabaudi isolates making this species an excellent organism to study genotype-phenotype relationships. Differences in virulence red blood cell (RBC) invasion, growth rates and immunogenic profiles exist between parasites of these isolates. Therefore, studies, such as linkage or quantitative trait loci analysis, are now possible to help identify genes associated with these defined phenotypes. For RMP species there is evidence that differences in virulence are associated with differences in RBC invasion, and, indeed, we find extensive sequence polymorphism in P. chabaudi genes encoding proteins involved in RBC invasion. Much attention is given to the role of exported proteins of multigene families and virulence in both human and RMP (for example, var, pirs), and analysis of differences between RMP proteins, that regulate invasion phenotypes, may reveal novel mechanisms that underlie virulence.

Full-length chromosomal annotation has permitted a comprehensive classification of all RMP subtelomeric multigene families. Our analyses indicate that most, if not all, RMP subtelomeric genes (apart from the RBP family) encode proteins exported out of the parasite; however, most lack a canonical PEXEL-motif. The presence of large numbers of PEXEL-negative exported proteins indicates alternative export mechanisms possibly common to all Plasmodium species. Investigations with highly tractable RMP species can therefore be used to understand these mechanisms better.

Our analyses of the phylogeny and expression of the largest RMP multi-gene family, the pirs, indicates functional diversification between members of the pir multigene family (this gene family is conserved between human/primate and RMP malaria species). Our new pir annotation and phylogeny demonstrates that clusters of structurally different pirs are differentially expressed. This is powerful data that can help to better understand their function, the relationship of pir expression with virulence and how the (species specific) pirs expansion is related to different selective pressures.

Methods
Animal experiments and parasites
All animal experiments performed in the Leiden malaria Research Group were approved by the Animal Experiments Committee of the Leiden University Medical Center (DEC 07171, DEC 10099). The Ethics Statement for P. yoelii YM and P. yoelii 17X: all animal work protocols were reviewed and approved by the Ethical Review Panel of the MRC National Institute for Medical Research and approved and licensed by the UK Home Office as governed by law under the Animals (Scientific Procedures) Act 1986 (Project license 80/1832, Malaria parasite-host interactions). Animals were handled in strict accordance with the ‘Code of Practice Part 1 for the housing and care of animals (21/03/05)’ available at [73]. The numbers of animals used was the minimum consistent with obtaining scientifically valid data. The experimental procedures were
designed to minimize the extent and duration of any harm and included predefined clinical and parasitological endpoints to avoid unnecessary suffering. The study of *P. chabaudi* DNA and RNA was carried out in strict accordance with the UK Animals (Scientific Procedures) Act 1986 and was approved by the Ethical Committee of the MRC National Institute for Medical Research, and the British Home Office (PPL: 80/2538).

For sequencing of the RMP reference genomes the following were used: for *PbA* the cloned reference line c115cy1 of the ANKA isolate of *P. berghei* [11]; for *PcAS* the 2722 clone of the AS isolate of *P. chabaudi chabaudi* (cloned after mosquito-transmission in 1978 and obtained from D. Walliker, University of Edinburgh, Edinburgh, UK); for *PyYM* the cloned 17XYM line of the YM line of *P. yoelii yoelii* [74]. In Additional files 6 and 7 details are provided of the other RMP isolates/lines that have been sequenced.

**Sequencing, assembly and annotation**

Sequencing was performed using Sanger capillary, Illumina and 454 sequencing. Sequence assemblies were performed using different assemblers [75,76], which were improved automatically using a number of configuration tools [77-82] and manual inspection. First pass annotation was performed through a combination of *ab initio* gene finding via Augustus [83] and transfer of annotation through orthology using RATT [80]. Gene models of the three reference genomes were corrected manually using RNA-Seq and orthologous information. Details of the assemblies and annotation are provided in Additional file 6. To define the orthologous and paralogous relationships between the predicted RMP proteins and those of human/ primate malaria species OrthoMCL [84] was used. The presence of a PEXEL-motif was determined using the updated HMM algorithm ExportProl v2.0 with a cutoff value of 1.5 [28]. Classification of the RMP multigene families was done through manual inspection of conserved domains (Interpro) and gene structure. SNPs in the genomes of these parasites were called by mapping the reads against their respective reference genomes, ignoring low complexity and repetitive regions. From the SNPs the Ka/Ks ratio was calculated for the *P. chabaudi* isolates with the Bio::Align::DNAStatistics Perl module.

**Transcriptomics**

RNA was collected from multiple synchronized blood stages [85] and purified gametocytes and ookinetes [86] of *PbA*, from *PcAS* blood stages (late trophozoites), isolated from different mice as described [44] and from *PyYM* late blood stages of two parasite lines (the cloned YM line and mutant PY01365-KO) [8]. RNA was sequenced as described [8,44,87,88]. To correct gene models and to compare the expression between samples, each sample was first mapped against its reference genome using TopHat [89] (version v2.0.6, parameter -g). For the resulting output a custom Perl script was written to detect errors in the annotation and to find new or alternative splice sites. To determine transcript abundance FPKM values were calculated for all genes (FPKM: fragments per kilo base of exon per million fragments mapped) using Cufflinks [90]. Accepting 10% of the intron as real signal, a cut-off FPKM value of 21 over all RNA-seq samples was determined. See also Additional file 6 for a detailed description of the generation and analysis of the RNA-seq data.

Heatmaps were generated with FPKM values of each gene and condition, using the heatmap.2 function of the gplots package. Correlation plots were done in R (Foundation for Statistical Computing; [91]) and generated with the corrplot function of the corrplot R library. Only genes were included that had one to one orthology in the three rodent species. For differential expression we used cuffdiff [90] (v2.0.2, with parameters -u -q) to compensate for GC variation and repetitive regions. GO enrichment of differentially expressed genes was performed in R, using TopGO. As a GO-database the predicted GO-terms from the reference RMP genomes were used.

**Phylogenetic analyses of pirs**

All full-length RMP *pir* coding sequences, including predicted pseudogenes, were used. Translated nucleotide sequences for 1,160 genes were aligned in ClustalW 92; all multiple alignments were manually edited to resolve all frame-shifts. Non-homologous positions at the N-terminus were removed by curtailing the alignment to the N-terminal-most conserved cysteine position. Non-homologous repetitive motifs were removed from ‘long-form’ PIRs (that is, 188 proteins >1,200 amino acids in length). The resultant 1,266-character alignment constitutes the conserved core of all PIRs and almost the complete repertoire of ‘short-forms’ (that is, <1,200 amino acids in length and 972/1,160 genes). A Maximum Likelihood phylogeny was estimated from the nucleotide sequence alignment using RAxML v7.0.4 [93] using a GTR + G model. Node support was assessed using 100 non-parametric bootstrap replicates [94]. A Bayesian phylogeny was estimated using MrBayes v3.2.1 [95] with a GTR + G model for a subsample of *pir* nucleotide sequences (MCMC settings: Nruns = 4, Ngen = 1,000,000, sample burnin = 1,000, and default prior distribution). See also Additional file 6 for a detailed description of the phylogenetic analyses.

**Accession numbers**

All the raw data used in the assemblies of the genome and the RNA-seq data have been deposited with accession numbers shown in Additional file 17. The reference
Additional files

Additional file 1: Novel putative protein coding genes which were absent in the previous genome annotations of \( P. \) berghei ANKA and \( P. \) c. chabaudi AS.

Additional file 2: \textit{RMP} genes with a PEXEL motif as identified by using the updated HMM algorithm ExportPred v2.0.

Additional file 3: A. Chromosomal organization of the expanded \textit{fam-d} multigene family in the internal region of chromosome 9. The \textit{PcAS}, \textit{PybM} and \textit{Pyb}A genomes contain 21, 5 and 1 copy (PBANKA\_091920), respectively (shown in blue). B. An \textit{ACT} (Artemis Comparison Tool) comparison of syntenic centromeric regions (green) of chromosome 7 of \textit{PbA} and \textit{PcAS} and chromosome 6 of \textit{P.} \textit{barlum} 3D7, showing size, location and GC-content. Grey bars: forward/reverse DNA strands. The red lines represent sequence similarity (tBLASTx) (related to Table 1). C. Structural organization of several types of full-length \textit{pir} genes in \textit{PbA} (bis), \textit{PcAS} (cis) and \textit{PybM} (dis). Exons: yellow boxes; with introns: linking lines. The IDs shown represent a single example.

Additional file 4: The different (subtelomeric) multigene families in the genomes of \( P. \) \textit{berghei} ANKA, \( P. \) c. \textit{chabaudi} AS and \( P. \) y. \textit{yoelii} YM: \textit{pir}; \textit{fam-a}; \textit{fam-b}; \textit{fam-c}; \textit{fam-d}; \textit{ETRAMP}; \textit{reticulocyte binding protein}, putative; \textit{lysophospholipase}; \textit{PCEMA1}; halobacid dehalogenase-like hydrolase; ‘other subtelomeric genes’.

Additional file 5: Putative orthologs (OrthoMCL) of \textit{RMP} and primate malaria proteins: 1) \textit{RMP} proteins shared with at least one primate malaria protein; 2) \textit{RMP} specific proteins (no orthologs in primate malaria); 3) primate malaria specific proteins (no orthologs in \textit{RMP}).

Additional file 6: Additional methods.

Additional file 7: Information on \textit{RMP} isolates/lines used in the study for genome sequencing and RNA-seq analyses (see also Additional file 6).

Additional file 8: Genes with single nucleotide polymorphisms (SNPs): 1) different stabilates of \( P. \) \textit{berghei} isolates/lines (NK65NY, NK65 E, SP11 A, SP11 RLL A, K173; compared to the genome of \( P. \) \textit{berghei} ANKA); 2) in \( P. \) y. \textit{yoelii} 17X (compared to the genome of \( P. \) y. \textit{yoelii} YM); and 3) in different isolates of \( P. \) c. \textit{chabaudi} (AJ, CB) and \( P. \) c. \textit{adami} (DS, DK) (compared to the genome of \( P. \) c. \textit{chabaudi} AS).

Additional file 9: RNA-seq analysis: mRNA abundance (FPKM values) of all genes in the different life cycle stages of \( P. \) \textit{berghie} ANKA, \( P. \) c. \textit{chabaudi} AS and \( P. \) y. \textit{yoelii} YM.

Additional file 10: Differentially expressed genes in 16 hour ookinetes compared to 24 hour ookinetes of \( P. \) \textit{berghie} ANKA and GO-annotation of up- and down-regulated genes.

Additional file 11: Overview and details of spliced transcripts and alternative splicing (AS) in the \textit{RMP} genomes based on RNA-seq analyses.

Additional file 12: Distribution of members of different multigene families on chromosomes of \textit{PbA}.

Additional file 13: Distribution of members of different multigene families on chromosomes of \textit{PcAS}.

Additional file 14: Distribution of members of different multigene families on chromosomes of \textit{PybM}.

Competing interests

The authors declare that they have no competing interests.

Authors’ contributions

TDO, UB, MH, LR, SAO, CIN, AP, MB contributed to the genome assembly, annotation and analysis. TDO, BFF, WAMH, AAR, OB, SMK, HGS, APW, CJI generated and analysed RNA-seq data of \( P. \) \textit{berghei}. DC, JL, AAH provided materials for genome sequences and/or RNA-seq data of different isolates/lines of \( P. \) y. \textit{yoelii} and \( P. \) \textit{chabaudi}. AE provided materials for genomes of isolates of \( P. \) \textit{berghei}. APJ performed the phylogenetic analysis. TDO, UB, APJ, BFF, WAMH, SMK, AAH, CIN, AP, MB, CJI conceived the study, participated in its design and coordination and drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgments

The work was funded by the Wellcome Trust (grant WT 089851). C. Newbold was supported by a Wellcome Trust program grant (082130), B. Franke-Fayard and W.A.M. Hoeijmakers by grants from The Netherlands Organization for Scientific Research (ZonMW TOP Grant No. 9120_6135; NWO Toptalent 021.001.011), T.D. Otto, M. Hunt, H.G. Stunnenberg, and C.J. Janse by a grant from the European Community’s Seventh Framework Program (247772007–2013; Grant Agreement No. 242095) and A.P. Water and A.A. Religa were supported by the Wellcome Trust (Ref. 083811/2/07/2). Work in the Holder and Langhorne labs is funded by the MRC (U17532067 and U17589248) respectively. We thank the following individuals: Martine Zilversmit for producing an Augustus training set; Jai Ramesar (LUMC, Leiden) and C. van Overmeir (ITM, Antwerp) for technical support; The European Malaria Reagents Repository [96] as the source for \( P. \) \textit{chabaudi} isolates; Sandra Cheesman for DNA preparation of the \( P. \) \textit{chabaudi} isolates; Richard Carter for helpful discussion and providing information for the \( P. \) \textit{chabaudi} isolates; R. Menard for providing \( P. \) \textit{berghei} NK65 NY; Ph. van den Steen for \( P. \) \textit{berghei} NK65 E; S. J. Boddey and T. Sargeant for assisting in the \textit{PEXEL} analysis; R. Davies and Q. Lin for help with data release; L. Robertson, D. Harris, K. Segar, A. Babbage, H. Beasley, L. Clark, J. Harley, P. Heath, P. Howden, G. Kenny, S. Pelan, D. Saunders and J. Wood for manual improvement of the \( P. \) \textit{chabaudi} AS assembly; R. Rance, M. Quail and members of DNA Pipelines for sequencing libraries at WTSI; J. Keane, M. Aslett, N. de Silva for database support. We acknowledge Roche (Branford, USA) for the generation of 454 20 kb libraries. The authors have no competing financial interests.

Author details

\(^{1}\)Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK. \(^{2}\)Department of Infection Biology, Institute of Infection and Global Health, University of Liverpool, Liverpool, UK. \(^{3}\)Leiden Malaria Research Group, Department of Parasitology, Leiden University Medical Center, Leiden, The Netherlands. \(^{4}\)Department of Molecular Biology, Science faculty, Radboud Institute for Molecular Life Sciences, Radboud University, Nijmegen, The Netherlands. \(^{5}\)Institute of Infection, Immunity & Inflammation, School of Medical, Veterinary & Life Sciences, & Wellcome Centre for Molecular Parasitology, Glasgow Biomedical Research Centre, University of Glasgow, Glasgow, Scotland, UK. \(^{6}\)Division of Parasitology, MRC National Institute for Medical Research, Mill Hill, London, UK. \(^{7}\)Unit of Malarialogy, Institute of Tropical Medicine, Antwerp, Belgium. \(^{8}\)Weatherall Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Oxford, UK. \(^{9}\)Weatherall Institute of Molecular Medicine, John Radcliffe Hospital, Headington, Oxford, UK. \(^{10}\)Biological and Environmental Sciences and Engineering (BESE) Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Kingdom of Saudi Arabia.
Fischer K, Chavchich M, Huebner R, Wilson DW, Kemp DJ, Saul A: 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 2003, 48:1209–1223.

Favaloro JM, Kemp DJ: Sequence diversity of the erythrocyte membrane antigen 1 in various strains of Plasmodium chabaudi. Mol Biochem Parasitol 1994, 66:39–47.

Lawton J, Brugat T, Yarn YX, Reid AJ, Boehme U, Otto TD, Pain A, Jackson A, Berrien M, Cunningham D, Preiser P, Langhorne J: Characterization and gene expression analysis of the multi-gene family of Plasmodium chabaudi chabaudi (AS). BMC Genomics 2012, 13:125.

Ebbinghaus P, Krucken J: Characterization and tissue-specific expression patterns of the Plasmodium chabaudi cimultigene family. Malar J 2011, 10:272.

Vontas J, Siden-Kiamos I, Papagiannakis G, Karras M, Walliker D: Characterization and tissue-specific expression patterns of the Plasmodium chabaudi cimultigene family. Malar J 2011, 10:272.

Lawton J, Brugat T, Yarn YX, Reid AJ, Boehme U, Otto TD, Pain A, Jackson A, Berrien M, Cunningham D, Preiser P, Langhorne J: Characterization and gene expression analysis of the multi-gene family of Plasmodium chabaudi chabaudi (AS). BMC Genomics 2012, 13:125.

Ebbinghaus P, Krucken J: Characterization and tissue-specific expression patterns of the Plasmodium chabaudi cimultigene family. Malar J 2011, 10:272.
86. Janse CJ, Mons B, Rouwenhorst RJ, Van der Klooster PF, Overduivel J, Van der Kaay HJ: In vitro formation of ookinetes and functional maturity of Plasmodium berghei gametocytes. Parasitology 1985, 91:19–29.
87. Sebastian S, Brochet M, Collins MO, Schwach F, Jones ML, Goulding D, Rayner JC, Choudhary JS, Billker O: A Plasmodium calcium-dependent protein kinase controls zygote development and transmission by translationally activating repressed mRNAs. Cell Host Microbe 2012, 12:9–19.
88. Hoeijmakers WA, Bartfai R, Francois KJ, Stunnenberg HG: Linear amplification for deep sequencing. Nat Protoc 2011, 6:1026–1036.
89. Trapnell C, Pachter L, Salzberg SL: TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 2009, 25:1105–1111.
90. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L: Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol 2013, 31:46–53.
91. The R Project for Statistical Computing. [www.r-project.org]
92. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins GD: Clustal W and Clustal X version 2.0. Bioinformatics 2007, 23:2947–2948.
93. Stamatakis A, Ludwig T, Meier H: RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 2005, 21:456–463.
94. Stamatakis A, Hoover P, Rougemont J: A rapid bootstrap algorithm for the RAxML Web servers. Syst Biol 2008, 57:758–771.
95. Ronquist F, Huelsenbeck JP: MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 2003, 19:1572–1574.
96. The European Malaria Reagent Repository. [www.malariaresearch.eu]

Cite this article as: Otto et al.: A comprehensive evaluation of rodent malaria parasite genomes and gene expression. BMC Biology 2014 12:86.