Complete avian malaria parasite genomes reveal host-specific parasite evolution in birds and mammals

Ulrike Böhme[1]*, Thomas D. Otto[1]*, James Cotton[1], Sascha Steinbiss[1], Mandy Sanders[1], Samuel O. Oyola[1,2], Antoine Nicot[3], Sylvain Gandon[3], Kailash P. Patra[4], Colin Herd[1], Ellen Bushell[1], Katarzyna K. Modrzynska[1], Oliver Billker[1], Joseph M. Vinetz[4], Ana Rivero[5], Chris I. Newbold[1,6], Matthew Berriman[1]

[1] Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SA, UK
[2] International Livestock Research Institute, Box 30709, Nairobi, Kenya
[3] CEFE UMR 5175, CNRS – Université de Montpellier –Université Paul-Valéry Montpellier – EPHE, 1919, route de Mende, 34293 Montpellier Cedex 5, France
[4] Department of Medicine, Division of Infectious Diseases, University of California San Diego, School of Medicine, La Jolla, CA, 92093, USA
[5] MIVEGEC (CNRS UMR 5290), Montpellier, France
[6] University of Oxford, United Kingdom, Weatherall Institute of Molecular Medicine, John Radcliffe Hospital, UK
* These authors contributed equally to this work.

Abstract

Avian malaria parasites are prevalent around the world, and infect a wide diversity of bird species. Here we report the sequencing and analysis of high quality draft genome sequences for two avian malaria species, *Plasmodium relictum* and *Plasmodium gallinaceum*. We identify 50 genes that are specific to avian malaria, located in an otherwise conserved core of the genome that shares gene synteny with all other sequenced malaria genomes. Phylogenetic analysis suggests that the avian malaria species form an outgroup to the mammalian *Plasmodium* species**.** Consistent with their phylogenetic position, we identify orthologs of genes that had previously appeared to be restricted to the clades of parasites containing *P. falciparum* and *P. vivax* – the species with the greatest impact on human health. The subtelomeres of *P. relictum* and *P. gallinaceum* contain several novel gene families, including an expanded surf multigene family. We also identify an expansion of reticulocyte binding protein homologs in *P. relictum* and within these proteins proteins, we detect distinct regions that are specific to non-human primate, humans, rodent and avian hosts. For the first time in the *Plasmodium* lineage we find evidence of transposable elements, including several hundred fragments of LTR-retrotransposons in both species and an apparently complete LTR-retrotransposon in the genome of *P. gallinaceum.*

Introduction

Malaria parasites of birds are more widespread, prevalent and genetically diverse than those infecting other vertebrates (Bensch et al. 2009). They are present in all continents except Antarctica, and in some populations up to 98% of birds within a species may be infected (Glaizot et al. 2012).

The first avian malaria parasites were discovered in the late 19th century, shortly after the discovery of human malaria parasites. In the early 1900's avian malaria became a prominent experimental system to elucidate key aspects of the biology and transmission of malaria parasites (Huff and Bloom 1935; Raffaela and Marchiafava 1944) as well as for the routine testing and development of the first antimalarial drugs (Marshall 1942). Recently, avian malaria has returned as a focus of research as a unique model to understand the ecology and evolution of the parasite, both in the field and in the laboratory (Pigeault et al 2015).

Currently around 600 mitochondrial cytochrome *b* sequence-based lineages of avian *Plasmodium* have been described, grouped into *ca.* 40 different clades (Bensch et al. 2009). Some of these lineages are rare and have been found in a single host species in a restricted geographical area, while others are found in dozens of bird species within several orders, and across several different continents. The consequences of *Plasmodium* infections on host fitness are usually relatively mild but virulence depends on the sensitivity of the host and the parasite lineage. For instance, the accidental introduction of avian malaria and its mosquito vectors into Hawaii played a major role in the decline and extinction of several species of honeycreepers (Atkinson et al. 2000) and still poses a threat to geographically isolated bird species (Lapointe et al. 2012; Levin et al. 2013). Work on wild European bird populations has also revealed strong associations between endemic malaria infection and bird survival and recapture rates (Lachish et al. 2011). More recently, malaria infections have been found to accelerate bird senescence through telomere degradation (Asghar et al. 2015). In addition, some species of avian malaria pose a significant problem to the poultry industry, where mortality rates of up to 90% have been observed in domestic chickens (Springer 1996).

The biology and life cycle of avian-malaria parasites in both the vertebrate and vector hosts is similar to that of their mammalian counterparts, but with a few important differences. First, while mammalian parasites have a single exoerythrocytic (EE) cycle in hepatocytes (Huff 1969), avian *Plasmodium* have two obligate exoerythrocytic (EE) cycles, one occurring in the reticuloendothelial system of certain organs, and the other with a much wider tissue distribution (Valkiunas 2004). Second, while certain mammalian parasites (e.g. *P. vivax*) produce dormant forms exclusively during the EE cycle, avian malaria species can also produce dormant forms from the parasite blood stages (Valkiunas 2004). Finally, avian red blood cells are nucleated. Since it could be argued that the invasion and growth in nucleated cells – with their more complete metabolism and transport – is easier to evolve than development in enucleated mammalian erythrocytes, it is tempting to speculate that these parasites more closely resemble the ancestral state.

To date, several species of human, primate and rodent malaria have been sequenced, revealing some key common features. In particular, the core genome, which excludes the subtelomeres, shows extensive conservation of synteny that extends across most of the length of each chromosome. Even between distantly related *Plasmodium* species, long blocks of conserved sequence exist that are disrupted by only a small number of ancient recombination events. The subtelomeres, however, are populated by polymorphic multigene families that are frequently species specific. One exception is the *Plasmodium* interspersed repeat (*pir*) family which is both ubiquitous and the most numerous in currently sequenced malaria genomes. Malaria parasite genomes are also characterised by their unusual base composition, the most extreme of which is exhibited by the *Laverania* sub-genus that infects apes and humans with a GC content of ~20% (Gardner et al. 2002; Otto et al. 2014b; Sundararaman et al. 2016). In rodent parasites, GC content increases to ~23% and is as high as 40-45% in the *vivax/knowlesi* clade. An early surprise from malaria genomes was their relative lack of transcription factors. It was subsequently shown that each genome contained a similar set of plant-like AP2 transcription factors that orchestrate a unique wave like pattern of gene expression during the asexual blood stages (Campbell et al. 2010). The other important common factor among these genomes is that they contain gene family members that are important in red blood cell invasion. These include a number of merozoite surface proteins (MSPs) and a family of genes encoding so-called reticulocyte binding proteins (RBP).

Previous *Plasmodium* genome sequencing has been confined to mammalian clades. The genomes of avian malaria parasites should provide insights into the evolution of unique features of mammalian infective species and allow an exploration on how far the apparently shared features extend across the entire *Plasmodium* genus. In this study we describe the sequence, annotation and comparative genomics of two avian malaria genomes, *Plasmodium relictum* SGS1 and *Plasmodium gallinaceum* 8A. Our analysis reveals surprising features involving gene content, gene family expansion and for the first time in *Plasmodium*, the presence of transposable elements.

Results

Generation of two avian malaria genomes

Separating parasite DNA from that of its host has been a major obstacle to sequencing avian malaria parasite genomes because avian red blood cells are nucleated. In the present study, we obtained parasite DNA using two independent strategies (see methods) involving depleting host DNA based on methylation (Oyola et al. 2013), and using whole genome amplification (*P. gallinac*eum) or sequencing from oocysts from the dissected guts of infected *Culex* mosquitos (*P. relictum*). Using Illumina-sequencing, a 23.8-megabase (Mb) and a 22.6-Mb high quality draft of the *P. gallinaceum* and *P. relictum* genomes were produced and assembled into 152 and 498 scaffolds, respectively (Table 1). Both avian malaria *Plasmodium* genomes show very low GC-content. In fact with a GC-content of 17.8 %, *P. gallinaceum* has the lowest GC-content observed in any *Plasmodium* genome sequenced to date.

With 5273 and 5146 genes (Table 1), respectively, the *P. gallinaceum* and *P. relictum* genomes contain similar numbers of genes to other *Plasmodium* reference genomes (Figure 1). Those in *P. gallinaceum* were predicted *ab initio* and manually curated using the recently published *P. gallinaceum* blood-stage transcriptome (Lauron et al. 2014) as a guide. These annotated genes were projected onto the *P. relictum* genome, using RATT (Otto et al. 2011) and manually refined. The most up-to-date annotations for both species can be found on GeneDB (http://www.genedb.org) (Logan-Klumpler et al. 2012). A detailed table (Supplemental Table S1) provides information on gene expression of all the *P. gallinaceum* genes in comparison with *P. falciparum* 3D7 gene expression.

The chromosomes are similar in size and in number of genes to those found in other *Plasmodium* species, and have positionally conserved centromeres within regions sharing synteny with *P. falciparum* and other *Plasmodium* species (Supplemental Fig. S1). Likewise, the mitochondrial and apicoplast genomes have been sequenced and show similar size, GC-content and numbers of genes to those previously sequenced from other *Plasmodium* species (Table 1).

Relationship between *Plasmodium* species

The phylogeny of *Plasmodium* spp. has been controversial. There is broad agreement that three groups of well-studied parasites that infect mammals are monophyletic - the subgenus *Laverania* containing parasites of great apes including *P. falciparum* and close relatives, the subgenus *Plasmodium* containing all the other human (*P vivax, P. malariae, P. ovale*) and primate parasites, and the laboratory model rodent malaria species. However, almost every possible arrangement has been proposed at some point for these three mammal-infecting groups relative to those that infect birds and reptiles (Blanquart and Gascuel 2011; Perkins 2014) and the relationships of other *Plasmodium* species is unclear. Recently, an extensive multi-locus molecular dataset (26 loci from up to 103 taxa; Borner et al., 2016) recovered the great ape parasites as a basal branch within the mammal clade, disagreeing with the earlier phylogenomic analyses (Pick et al., 2011) that had supported the hypothesis (Waters et al. 1991) that mammalian *Plasmodium* are polyphyletic, and that *P. falciparum* and its relatives evolved recently from an avian ancestor, perhaps explaining its high virulence.

The completion of genome sequences for two non-mammalian *Plasmodium* species enabled us to re-evaluate the phylogeny of the mammalian groups using genome-wide data. Using both Bayesian and maximum-likelihood phylogenetic models we find support for *P. gallinaceum* and *P. relictum* forming an outgroup to the other *Plasmodium* species, and the *Laverania* appear as the most basal group of mammalian *Plasmodium* (Figure 1; Supplemental Fig. S2). We also find that subgenus

Plasmodium is paraphyletic, with an unexpected sister-group relationship between *P. ovale* and the rodent-infective species and *P. malariae* branching as the deepest lineage outside the avian species and the *Laverania*.

This result is robust to changes in the substitution model used for phylogenetic inference. The same phylogeny is from a maximum-likelihood tree under both partitioned and non-partitioned models (Supplemental Fig. S3A) and maximum parsimony analysis also agrees. Our result is also not just a result of trimming the alignment to remove poorly-aligned regions, as this tree also maximises the likelihood of a non-trimmed concatenated alignment under a simple substitution model. Simpler, non-model based approaches produce different trees, with neighbour-joining and simple amino acid distance producing a phylogeny matching that shown by Pick et al (2011), with a clade of Laverania and avian malaria species, and the primate-infective species outside Laverania forming a clade related to a clade of rodent malaria (Supplemental Fig. S3B).

Our data give strong and consistent support for the relationships within *Plasmodium* and the root of *Plasmodium* when the data for *Haemoproteus* is used as an outgroup to the species we include (Supplemental Fig. S3C), it seems to be somewhat more difficult to correctly place the wider outgroup of more distantly related apicomplexa. When the non-haemosporidia outgroups are constrained to be monophyletic, a Bayesian analysis shows a paraphyletic *Plasmodium* (with *Haemoproteus* joining the avian *Plasmodium spp.*) and has lower support for the basal splits within *Plasmodium* due to uncertainty in placement of the root of this subtree (Supplemental Fig. S3D). Attempts to fit more complex, and so potentially more realistic phylogenetic models to resolve this discrepancy between Bayesian and maximum likelihood trees were unsuccessful, as MCMC runs failed to converge under these models.

Our results confirm that, as has been widely recognised (Perkins 2014; Borner et al. 2016), the key difficulty for previous analyses has been a lack of comprehensive genomic data for any close outgroup to *Plasmodium*. Previous phylogenomic analyses were effectively trying to root the tree of *Plasmodium* using data from only very distant outgroups (Pick et al. 2011; Borner et al. 2016). The availability of the *Haemoproteus tartakovskyi* genome (Bensch et al. 2016) and our complete, curated genome data for two avian *Plasmodium* species has allowed us to resolve the phylogeny of *Plasmodium*. However, little molecular data is available for many lineages of the genus Plasmodium (Perkins and Schaer 2016). Molecular data from these, and from additional non-*Plasmodium* lineages of *haemosporidia*, in particular other groups proposed to be most closely related to mammal and bird *Plasmodium,* such as *Hepatocystis, Nycteria* and *Parahaemoproteus* (Perkins, 2014), will be key to fully resolving the evolution of the human pathogens.

Synteny between *Plasmodium* species

The two avian malaria genomes contain a conserved core with extensive synteny shared across the *Plasmodium* genus. For 83% (4395) of predicted genes in *P. gallinaceum* and *P. relictum*, orthologs could be identified in the genomes of rodent malaria species, the *Laverania* subgenus and the clade of species that includes *P. vivax*. We found 50 avian malaria-specific core genes with orthologs in both species (Supplemental Table S2). Using the available *P. gallinaceum* transcriptome data (Lauron et al. 2014), we found that only two of these 50 genes showed evidence of expression in the blood stage indicating that the remaining 48 genes are likely to play a role elsewhere in the life cycle. For the majority (52%) of these genes, putative functions could not be ascribed (Supplemental Table S2) but, a possible new member of the AP2 family of transcription factors (PRELSG_1134000, PGAL8A_00142800) (Supplemental Fig. S4A) was found in both species. In addition, we found orthologs of the 28 AP2 transcription factors already known in other *Plasmodium* spp., an additional 6-cysteine protein, a protein phosphatase and an AMP-specific ABC transporter. To date, analyses have failed to identify homologs in *Plasmodium* of the non-homologous end joining (NHEJ) pathway, a pathway that repairs double-strand breaks in DNA. Ku70 is a member of this pathway that has apparent orthologs in both *P. gallinaceum* and *P. relictum* (PGAL8A_00014200, PRELSG_0411800), supported by a three-dimensional model created using I-TASSER (Yang et al. 2015) (Supplemental Fig. S5). However, an ortholog of Ku80 (the obligate partner of Ku70 in NHEJ activity) is not present (Fell and Schild-Poulter 2015).

Unique to the avian malaria genomes is the AMP-activated protein kinase that has so far not been identified in other mammalian *Plasmodium* species. This enzyme plays a key role in cellular energy metabolism and consists of three subunits. In *P. gallinaceum* and *P. relictum*, just the alpha (PGAL8A_00159300, PRELSG_1117500) and beta subunits (PGAL8A_00165250, PRELSG_1111850) can be clearly identified based on domain analysis but the gamma subunit is not immediately obvious. However, based on an analysis of Pfam domains and structural prediction using I-TASSER, we were able to find a possible candidate gamma subunit in each of the two avian malaria species (PGAL8A_00033950, PRELSG_1019550).

In the core regions of *P. relictum* and *P. gallinaceum* chromosomes there are only minor differences in gene content. An additional hypothetical gene is present in *P. relictum* (PRELSG_0909800) (Supplemental Table S2, Supplemental Fig. S4B) and five avian malaria-specific core genes are pseudogenes in *P. relictum* but not in *P. gallinaceum* (Supplemental Table S2). Examples of the latter include a protein kinase (PRELSG_0314000) (Supplemental Fig. S6A) and an ABC transporter (MRP1) pseudogene in *P. relictum* (Supplemental Fig. S6B). There is only one case in the core genome, a pseudogene in *P. gallinaceum* (protein phosphatase 2C; PGAL8A_00276550) but not in *P. relictum* and we identified more pseudogenes on a genomic scale in *P. relictum* (37) than in *P. gallinaceum* (25).

We found 15 genes present in *P. gallinaceum* and *P. relictum* that were previously defined as *Laverania* specific (Supplemental Table S3). Apart from hypothetical proteins this includes ATPase1 (Supplemental Fig. S7A), apyrase and a sugar transporter. We also found 12 genes that so far have not previously been identified outside the *vivax, ovale or malariae* clades (Supplemental Table S4). Among these are the merozoite surface protein 1 paralog (MSP1P) and an ApiAP2 transcription factor (Supplemental Fig. S7B).

When comparing the predicted metabolic pathways of avian malaria parasites with those of previously sequenced *Plasmodium* species, one pathway, the shikimate pathway that provides precursors for folate biosynthesis is remarkably different. Genes encoding two enzymes in the pathway have become pseudogenes (Supplemental Fig. S8, Supplemental Table S5) and the gene encoding a key enzyme complex, the pentafunctional AROM polypeptide, is completely missing. Thus avian malaria parasites are not able to synthesize folate *de novo*. One explanation for this could be the fact that the host cells are nucleated and therefore provide a richer nutrient environment.

Multigene families

In addition to the multigene families present in previously sequenced *Plasmodium* genomes e.g. ETRAMPS, *pir* and reticulocyte binding proteins (Gardner et al. 2002; Pain et al. 2008), (Table 2, Supplemental Table S6), we identified four novel gene families in the avian malaria genomes (Supplemental Fig. S9) and found that a *Plasmodium* specific low copy number gene is expanded in the avian species. To maintain consistency with the gene family naming scheme established for other species (Otto et al., 2014a) the families are named *fam-e* to *fam-i* (Supplemental Fig. S9).

To explore the relationship between *Plasmodium* subtelomeric gene families across the genus we used two different clustering approaches, based either on global similarity or on conservation of short motifs. First, we compared all genes with BLASTP and created a gene network, where the genes (nodes) were connected if they shared a global similarity above a threshold (Figure 2A). Although the topology of the network changed with different sequence identity thresholds (Supplemental Fig S10A), at a threshold of 31% the STP1 and the Surface-associated interspersed proteins (SURFINs) of different species are connected. The SURFINs are encoded by a family of 10 genes in each of *P. falciparum* and *P. reichenowi*. We found a relatively high number of SURFINs in avian malaria genomes, 40 in *P. gallinaceum* and 14 in *P. relictum* (Table 2). As shown in previous studies, SURFINs show some sequence similarity to PIR proteins of *P. vivax* (Merino et al. 2006; Winter et al. 2005) and some SURFINs share a domain with the SICAvars of *P. knowlesi*. To examine this relationship more closely by highlighting similarity that could be missed by BLASTP, we used MEME (Bailey et al. 2009) to generate 96 sequence motifs from the STP1 and SURFIN families, respectively. Next we searched for those predicted motifs in all predicted proteins (excluding low complexity regions) of the 11 sequenced *Plasmodium* species. We visualized those results as a binary occurrence matrix (Supplemental Fig. S10B). Although some proteins share a limited repertoire of the STP1 or SURFIN motifs – namely DBL containing protein, antigen 332 and three putative proteins of unknown function (PRELSG_1445700, PmUG01_00032900, PmUG01_10034200), there is extensive motif-sharing amongst the STP1 and SURFIN proteins (Figure 2b) but only a single motif is shared between STP1*,* SURFIN and SICAvar (Supplemental Fig. S10B,C). This suggests that STP1 and SURFIN compose a superfamily. The SURFINs cluster into two groups. Group II is unique to *P. gallinaceum* but group I includes both homologs from the avian *Plasmodium* and the hominoid *Laverania* subgenus. The STP1 proteins are not found in the avian malaria parasites but form two *P. ovale* and one *P. malariae* specific clusters. Whether these poorly characterized families have functional similarities remains to be determined.

The *pir* (*Plasmodium* interspersed repeat) genes are the largest multigene family in *Plasmodium* spp. and have been found in high numbers in all malaria species sequenced to date (Janssen et al. 2004). In the avian malaria genomes we found only a small number of distantly related genes that are possibly members of the family: 20 in *P. gallinaceum* and 4 in *P. relictum* (Table 2). They follow the canonical 3-exon structure, with the second exon encoding a cysteine-rich low-complexity sequence, a transmembrane domain and a highly conserved third exon. However, the avian *pir* genes have only remote sequence similarity to those of other *Plasmodium* species (and have therefore been annotated as *pir*-like); the highest sequence similarity (41% over 60 bases) was found between a *pir* from *P. vivax* (PVP01_0800600) and *P. gallinaceum.*

We identified 290 genes in *P. gallinaceum*, and 203 in *P. relictum*, encoding the PEXEL motif that is frequently present in *Plasmodium* sub-telomeric multi-gene families and is important for trafficking proteins into and through the host cell (Marti et al. 2004). Three families in particular appear to be important in the avian malaria lineage. The *fam-f* family has only a single member in each species of the *Laverania* subgenus (PF3D7_1352900, PRCDC_1351900) and two members in the *vivax* clade (PVP01_1201900, PVP01_1147000, PKNH_1248100, PKNH_1148800) but has 16 members in *P. gallinaceum* and 14 members in *P. relictum* (Supplemental Fig. S9, Supplemental Table S6). Using I-TASSER to predict the structure, *fam-f* has some similarity (a modest C-score of -2.33) with a human protein (PDB:4IGG) involved in cell adhesion. *Fam-g* and *fam-h* to date appear to be specific to avian malaria. *Fam-g* has 107 members in *P. gallinaceum* and none in *P. relictum,* whereas *fam-h* has 49 copies in *P. relictum* and only 2 in *P. gallinaceum* (Supplemental Table S6). It is possible that the relative absence of *pir* genes might be in some way compensated by the expansion of these families.

There are two additional novel gene families in the avian malaria genomes, *fam-e* and *fam-i. Fam-e* is present in both avian malaria species and is a 2-exon gene with an average length of 350 amino acids and a transmembrane domain. There are 38 copies in *P. gallinaceum* and only 4 in *P. relictum* (Table 2, Supplemental Table S6). *Fam-i* is a 2-exon gene family only present in *P. gallinaceum* with 23 members (Supplemental Fig. S9). Both gene families lack a putative PEXEL motif. One aspect of these gene families is that the expression of a few individual members seems to dominate within the asexual blood stages, see Supplemental Table S1.

Expansion of the reticulocyte-binding protein (RBP) family in avian malaria parasites

Homologs of reticulocyte-binding protein (RBP) are known to be important in red cell invasion yet a recent publication (Lauron et al. 2015) indicated, based on a transcriptome assembly, a lack of RBPs in *P. gallinaceum*. In contrast we found an expansion of this family in both avian malaria parasites with 8 copies in *P. gallinaceum* and at least 29 in *P. relictum* (33 if fragments are included; see methods). Because these genes, are long (> 7.5 kb) and with large blocks of high sequence similarity they are difficult to assemble in *P. relictum* and the copy number in this species could be underestimated. Though the genes are diverse, we can see in a maximum likelihood tree of avian RBP ≥ 4.5kb that the separation into different sub-families predates the speciation of the two avian malaria lineages (Supplemental Fig. S11A). As with the STP1 and SURFIN families, we analysed sequence motifs produced by MEME to investigate the structure and evolution of the RBPs in 9 *Plasmodium* spp (Figure 3A, Supplemental Fig. S11B, C) in more detail. Conserved sequences were used, corresponding to two sets of shared motifs (black dashed box and stars in Figure 3A), to draw maximum likelihood trees for the 9 *Plasmodium* spp (Figure 3B). The phylogenetic trees confirm the general species diversity separating the avian species and the *Laverania*. The RBPs of *P. ovale*, *P. malariae*, *P. vivax* and *P. knowlesi* form three different clades. Interestingly the *P. berghei* RBP seem to be very similar to each other, but very different to the other species. This general classification of the species can also be seen in the motif occurrence matrix (Figure 3C). Some of the motifs are shared in all RBP. We also find host specific motifs, splitting the rodents, avian, and human and primate hosts.

A family of Long Terminal Repeat (LTR)- Retrotransposons in avian malaria genomes

Despite the presence of retrotransposons in the majority of eukaryotic genomes, none have been identified from *Plasmodium* species. A few transposable elements were annotated in the genome of *Plasmodium yoelii yoelii* 17XNL (eg Q7R980 and Q7R848, Carlton et al. 2002), but in all cases their subsequent identity as contaminants from the mouse genome was confirmed based on sequence similarity.

Surprisingly, in the avian malaria genomes we identified a large number of transposon fragments (Supplemental Fig. S12). Altogether there are 1244 transposon fragments in *P. gallinaceum* and 344 in *P. relictum* (Table 1). Except for one transposon fragment in the core region of a *P. gallinaceum* chromosome and two in core regions of *P. relictum* (Supplemental Fig. S13, Supplemental Fig S14A), all other fragments are in the subtelomeres (Supplemental Fig. S12, Figure 4C). A single complete, and potentially active, 5.7 kb retrotransposon is present in *P. gallinaceum* (PGAL8A_00410600) (Figure 4A). The retrotransposon contains a 4.5 kb open reading frame encoding a gag-pol polyprotein, including the following domains: a retroviral aspartyl protease (Pfam:PF00077), reverse transcriptase (Pfam:PF00078), RNAse H (Interpro:IPR012337) and integrase (Pfam:PF00665). It is bounded by long terminal repeats of 459 nucleotides (5'LTR) and 469 nucleotides (3'LTR) respectively. A primer binding site and polypurine tract were also identified (Figure 4B). Based on the order of encoded HMM domains the *P. gallinaceum* retrotransposon can be classified as ty3-gypsy retrotransposon (Steinbiss et al. 2009). In addition to the complete transposon, we found four nearly full-length copies, also bounded by long terminal repeats (PGAL8A_00328600, PGAL8A_00325400, PGAL8A_00189500, PGAL8A_00270200) (Figure 4A). *P. relictum* did not contain a complete retrotransposon but based on the programs LTRharvest/LTRdigest (Ellinghaus et al. 2008; Steinbiss et al. 2009) we found 7 near full-length copies with all the required domains. The most complete one is localized in the core area on chromosome 1 (Supplemental Fig S14B). It has a length of 5.3 kb, contains all the required HMM domains and is bounded by long terminal repeats that are shorter than those observed in *P. gallinaceum*, the 5'LTR and 3'LTR are only 253 and 257 bp long. A comparison between the complete retrotransposon from *P. gallinaceum* and the most complete fragment in *P. relictum* showed a similarity of 60% at the DNA level.

A BLAST comparison with other retrotransposons showed the highest similarity (28%) to a retrotransposon described in *Ascogregarina taiwanensis*, a gregarine that infects mosquito larvae (Templeton et al. 2010). This is also reflected in the phylogenetic tree (Supplemental Fig. S15).

To see if the complete retrotransposon was still active, we attempted to introduce a gag-pol expression cassette into the rodent malaria parasite *P. berghei*. Transfection was attempted at four independent occasions, without integration of the *P. gallinaceum* gag-pol expression cassette ever being detected. Parallel transfection of a second vector containing an unrelated insert acted as a positive control, ruling out technical difficulties. Failure to introduce the *P. gallinaceum* gag-pol transposase expression cassette is interpreted as potential toxicity associated with expression of the *P. gallinaceum* gag-pol under the very strong *pbhsp70* promoter and attempts to swap the promoter for the weaker *pbeef1a* promoter, or an inducible promoter are ongoing.

Discussion

Until now, all high quality and manually curated genomes available for the research community to understand malaria have been from mammalian parasites. The high-quality genome assemblies of two avian malaria genomes, *P. gallinaceum* and *P. relictum*, in the present study occupy a basal position relative to the phylogeny of the mammal-infecting *Plasmodium* species. To confirm this basal position was challenging, as the outgroups are diverse in sequence similarity and the avian parasites share the same extreme GC bias as those from the *Laverania* sub-genus (19% GC content).

All other mammalian genomes have a less extreme GC content, between 23-44 %. It has been suggested that *P. falciparum* lacks efficient base excision repair (BER) (Haltiwanger et al. 2000) and this drives the genome towards lower %GC content. If this is the case, it is likely that BER has been lost in both the *Laverania* and avian malaria lineages or that improvements to BER occurred after the evolution of the *Laverania* branch.

Though we have analysed 11 parasite species from diverse hosts, nearly all genes in regions previously defined as a conserved core occurred with 1:1 orthologs present in all 11 species. The roles of most genes are therefore probably shared between the species and transcend host differences. Our analysis focused on genes that are not shared with other species to provide insights to species-specific malaria biology. For example, we only find 50 core genes (1:1 orthologs) that are unique to the avian malaria parasites. As these include a novel AP2 transcription factor, a class of transcription factor known for its importance in developmental regulation and the majority (48) of the unique genes are not expressed in blood stages, one could speculate that they play a role in the second EE cycle unique to avian species (Garnham 1966). The differences in the folate and heme pathways of the avian species could also be attributed to their colonization of the more metabolically competent nucleated red blood cells of birds.

As expected we find variability in the subtelomeric gene families. We discovered four new gene families (*fam-e, fam-g, fam-h* and *fam-i*), a newly expanded family fam-f, expansion of SURFIN types, but also a reduction in the number of *pir* genes compared to other species. *Plasmodium interspersed* repeats (*pir*) genes are present in all *Plasmodium* species sequenced to date and are the largest multi-gene family within the genus. Their function is unclear, but their protein products are present at the host parasite interface, likely to be involved in parasite-host interactions and have been associated with immune evasion (Cunningham et al. 2010; Fernandez-Becerra et al. 2009). In the *P. chabaudi* rodent model, differential *pir* gene expression is associated with parasite virulence (Spence et al. 2013). It is not known whether the avian parasites employ a strategy of antigenic variation to evade their host's immune system, similar to that seen with *var* and SICAvar genes in *P. falciparum* and *P. knowlesi* respectively. Although avian malaria lack *var* genes, one of the other various avian malaria specific subtelomeric gene families may play a similar role. In particular, *fam-f* is single copy in the mammalian malaria genomes but is significantly expanded in both avian genomes and shows a distant homology to a human protein that is involved in cell-adhesion. SURFINs are also found on the surface of infected red blood cells, are expanded in *P. gallinaceum* and show substantial similarity to the STP1 family which leads us to hypothesise that these two genes families may have shared a common ancestor and their sequences evolved in a host-dependent manner.

Another polymorphic gene family involved in host cell invasion, and recently attributed to host specificity (Otto et al., 2014b) and red blood cell preference, are reticulocyte binding proteins (RBPs). This family is significantly expanded in *P. relictum*, which could explain the ability of this parasite to infect a wide range of tissues and avian species. The phylogenetic analysis shows that the genes predate the speciation of *Plasmodium* genus but we see a strong host-specific diversification. The rodent RBP cluster together and are different to the other clades but still share certain motifs across the genus (Figure 4A). Interestingly, the more diverse motifs are found at the N-terminus of the RBP, Figure 3A, (blue dashed boxes). The differences between the N-termini of the RBP, is intriguing as these regions mediate binding to host receptors. It is tempting to speculate that the conserved motifs are important for the general structure of the RBP, but the more variable N-terminal regions evolved to bind to specific host receptors.

The most striking difference between the avian parasites and their mammalian infecting relatives however is the presence of long terminal repeat (LTR) retrotransposons. Why were transposable elements (TEs) not found in any other *Plasmodium* species sequenced to date? The only other retrotransposon found so far in *Apicomplexa* are those from *Eimeria* (Ling et al. 2007; Reid et al. 2014). Both the retrotransposons found in *Eimeria* and the one we found in the avian malaria parasites belong to the ty3-gypsy family. The transposable element found in *Eimeria* is similar to chromoviruses, a subgroup of ty3-gypsy retrotransposons, whereas the transposon from *P. gallinaceum* does not contain chromodomains. Both transposons therefore seem to be from different lineages. We were able to identify several unreported fragments of ty3-gypsy retrotransposons in the recently published genome of the bird parasite *Haemoproteus tartakovskyi* (Bensch et al. 2016), a sister genus of *Plasmodium*. A complete transposon, as in *P. gallinaceum*, could not be identified. One reason for this could be the fragmented nature of the assembly which consists of 2,243 scaffolds. The TE was found in three avian parasite species, and since all appear to be from the same source we can assume that the TE was acquired in the ancestor of all the avian malaria species. The similarity of the avian *Plasmodium* TE to a sequence from *Ascogregarina taiwanensis* suggests that the TE in fact appeared long before the *Plasmodium* lineage. If this is the case, what has caused the contemporary loss of TEs from mammalian *Plasmodium*? It is tempting to speculate that the loss is linked to the absence of Ku70 in the mammalian parasites. Ku70 is a DNA binding protein and perhaps it is involved in repairing the otherwise-lethal lesions in the genome that result from retrotransposon-encoded integrase. However the absence of Ku80 rules out the NHEJ pathway and a novel function of the Ku70 monomer would therefore need to be elucidated. To date, *Piggybac* is the only TE to have been successfully mobilised in *Plasmodium* under experimental conditions (Balu et al. 2009) We have made multiple attempts to express *P. galllinaceum* gag-pol in *P. berghei* but these have been unsuccessful, perhaps due to its toxicity. Understanding the mechanism of action for this novel TE could open up exciting new possibilities for transposon based insertional mutagenesis within *Plasmodium* species.

Methods

Parasites used for the sequencing of P. gallinaceum strain 8A and collection of genomic DNA

The Institutional Animal Care and Use Committee (IACUC) of the University of California San Diego (UCSD) approved the animal protocol for the production of blood stages of *P. gallinaceum*. The 8A strain (catlog No MRA-310, ATCC, American Type Culture Collection, Manassas, VA, USA) used in these experiments was originally isolated in 1936 from chickens in Sri Lanka (Brumpt 1937) and has been since kept in laboratories across the world (largely through intraperitoneal passage between chickens and with occasional transmission via infected mosquitoes) as a model species for malaria research in the laboratory (Williams 2005). The *P. gallinaceum* 8A strain was cycled through White Leghorn chickens and *Aedes aegypti* mosquitoes, passage one (P1) parasites (10^5parasites/chick) were used to infect twenty chickens, and blood was collected as previously described (Patra and Vinetz 2012). Approximately, a total of 100 ml of infected blood sample (>10% parasitemia) was collected, centrifuged, the buffy coat was removed, and the RBC pellet was washed four times with cold phosphate buffered saline (PBS) pH 7.40. Washed RBCs were lysed by saponin (0.05% in PBS) and genomic DNA (gDNA) was extracted using a standard phenol-chloroform method. Because chicken RBC are nucleated only a small proportion of isolated DNA was that of *P. gallinaceum.* Hence Hoechst 33258-Cesium chloride (Cs-Cl) ultracentrifugation was used to separate AT rich *Plasmodium* DNA from the chicken DNA (Dame and Mccutchan 1987). Isolated *P. gallinaceum* DNA was extensively dialyzed against autoclaved Milli-Q water, precipitated with isopropanol and the DNA pellet washed with 70% ethanol. The DNA pellet was suspended in TE (10 mM Tris-HCl, 1 mM EDTA, pH 8) buffer and visualized in 0.7% agarose gel electrophoresis to confirm the quality of the DNA preparation. The DNA was stored at -80^0C or on dry ice prior to use.

Host DNA depletion and whole genome sequencing of P. gallinaceum

Purified *P. gallinaceum* genomic DNA from a batch prepared in 2003 was used to produce an amplification-free Illumina library of 400-600 base pairs (bp) (Quail et al. 2012) and 100bp paired end reads were generated on an Illumina HiSeq 2000 according to the manufacturer's standard sequencing protocol. To reduce host contamination and enrich for *P. gallinaceum* DNA, 2 µg of the DNA sample was mixed with 320 µl of methyl binding domain-Fc protein A beads complex (Feehery et al. 2013). The mixture was incubated at room temperature for 15 minutes with gentle rotation. Incubated mixture was placed on a magnetic rack for 3 minutes to separate the beads and the supernatant. A clear supernatant containing enriched *P. gallinaceum* DNA was pipetted in to a clean tube without disturbing the beads. The supernatant was purified using 1.8x volume of Agencourt AMPure® XP beads (Beckman Coulter #A63880) following manufacturer's instructions. The DNA was eluted in 80 µl of 1x TE buffer (pH 7.5).

An amplification-free Illumina library of 400-600bp was prepared from the enriched genomic DNA (Quail et al. 2012) and 150bp paired end reads were generated on an Illumina MiSeq using v2 chemistry according to the manufacturer's standard sequencing protocol.

From 20ng of the enriched genomic DNA whole genome amplification (WGA) was performed with REPLI-g Mini Kit (Qiagen) following a modified protocol (Oyola et al. 2014). Nuclease-free water and all tubes were UV-treated before use. WGA reactions were performed in 0.2 ml PCR tubes. Buffer D1 stock solution (Qiagen) was reconstituted by adding 500 µl of nuclease-free water and a working solution was prepared by mixing the stock solution and nuclease-free water in the ratio of 1:3.5

respectively. Buffer N1 was modified to include Tetramethylammonium chloride (TMAC) at a concentration of 300 mM. To denature DNA templates, 5 µl of the DNA solution was mixed with 5ul of buffer D1 (working solution prepared as described above). The mixture was vortexed and centrifuged briefly before incubating at room for 3 min. Denatured DNA was neutralized by adding 10 µl of the modified buffer N1. Neutralized DNA was mixed by vortexing and centrifuged briefly. To amplify the DNA template, denatured and neutralized sample was mixed with 29 µl of REPLI-g Mini Reaction Buffer and 1ul of REPLI-g Mini DNA polymerase to obtain a final reaction volume of 50 µl. The reaction mixture was incubated at 30°C for 16 hr using an MJ thermocycler with the heating lid set to track at +5^{0}C. Amplified DNA was cleaned using Agencourt Ampure XP beads (Beckman Coulter) using sample to beads ration of 1:1 and eluted with 50 µl of EB (Qiagen).

This material was then used to prepare a 3-4kb Illumina mate-paired library using an improved (Sanger) mate-paired protocol (Park et al. 2013) and 100bp paired end reads were generated on an Illumina HiSeq 2500 according to the manufacturer's standard sequencing protocol.

Parasites used for the sequencing of P. relictum and collection of genomic DNA

Experimental procedures were approved by the Ethical Committee for Animal Experimentation established by the CNRS under the auspices of the French Ministry of Education and Research (permit number CEEA- LR-1051). *Plasmodium relictum* (lineage SGS1-like, recently renamed DONANA05 (Bensch et al. 2009), GenBank KJ579152) was originally isolated by G. Sorci from wild sparrows (*Passer domesticus*) caught in 2009 in the region of Dijon (France) and subsequently passaged to naïve canaries (*Serinus canaria)* by intra peritoneal injection. The strain was maintained in an animal house by carrying out regular passages between our stock canaries, and occasionally through *Cx pipiens* mosquitoes every ca. 3 weeks (for details see Pigeault et al. 2015).

Heavily infected mosquito midguts were obtained in the following way. Mosquitoes from a laboratory line of *Cx. pipiens quinquefasciatus* (SLAB) were placed in a cage and allowed to blood feed from a heavily infected canary following standard laboratory protocols (Cornet et al. 2013). Two such cages, each with 70 mosquitoes, were set up in this way (bird parasitaemias were 4.45% and 7.89%). After the blood meal, mosquitoes were kept at 25°C and 80% relative humidity and dissected 7 days later to coincide with the midgut (oocyst) stage of the *Plasmodium* infection. Midguts were dissected and oocyst numbers assessed using standard laboratory procedures (Zélé et al. 2014). Total DNA was extracted from a single pool of 50 heavily infected midguts (>100 oocysts) using the QIAGEN protocol and materials (DNeasy 96 Tissue Kit, Qiagen NV, Venlo, The Netherlands) and total DNA was eluted in the final step with 100µL RNase free water (Qiagen).

Whole genome sequencing of P. relictum

An amplification-free Illumina library of 400-600 base pair (bp) was prepared from the genomic DNA of infected mosquito midguts and 150bp paired end reads were generated on an Illumina MiSeq using v2 chemistry according to the manufacturer's standard sequencing protocol.

Genome assembly and annotation of P. gallinaceum and *P. relictum*

Due to the better ratio of parasite versus host, the *P. relictum* assembly generated better contig results. First, low quality regions for the reads were clipped with SGA version 0.9.1 (Simpson and Durbin 2012); parameters: -m 51 --permute-ambiguous -f 3 -q 3). Those reads were assembled with Velvet (version 1.2.07) (Zerbino and Birney 2008) iterating through the following k-mers: 85, 81, 71 and 55. The other parameters were: -exp_cov 17 -max_coverage 30 -ins_length 450 -ins_length_sd 30 -cov_cutoff 9 -min_contig_lgth 200 -min_pair_count 10. The best k-mer was 81. Next we scaffolded the contigs further with SSPACE (Boetzer et al. 2011). To improve the assembly several tools were used, as described in PAGIT (Post Assembly Genome Improvement Toolkit (Swain et al. 2012)). First we ordered the contigs with ABACAS (Assefa et al. 2009) against *P. knowlesi*. Several rounds of iCORN2 (Otto et al. 2010) corrected single base pair errors and small indels. Assembly errors were detected with REAPR (Hunt et al. 2013), breaking the contig at each Fragment Coverage Distribution error (Parameter -l to also break contig errors). Those corrected contigs were ordered again with respect to described reference genomes. Next sequencing gaps were further closed with GapFiller (Boetzer and Pirovano 2012) and six iterations with IMAGE (Tsai et al. 2010), with two iterations each using of decreasing k-mer lengths of 71, 55 and 41.

The *P. gallinaceum* assembly was performed similarly, but needed more iterative steps of PAGIT and scaffolding with SSPACE and breaking with REAPR due to the mate pair 3kb library. Due to the library the assembly is more continuous, but due to the WGA and the higher host contamination and biases we have more contigs.

Annotation was performed using the Artemis and ACT software (Carver et al. 2013). Gene model structures were corrected based on orthology and transcriptome data (Lauron et al. 2014). Based on the transcriptome data we were able to correct 326 gene models and the transcriptome provided evidence for the addition of eight new genes. Functional assignments were extracted from literature or based on assessment of BLAST and FASTA similarity searches against public databases and searches in protein domain databases such as InterPro and Pfam (Finn et al. 2014). In addition TMHMMv2.0 (Krogh et al. 2001) and Rfamscan (Nawrocki et al. 2015) were used to identify transmembrane domains and non-coding RNA genes. To define orthologous and paralogous relationships between the predicted proteins of *P. gallinaceum* and of *P. falciparum, P. reichenowi*, *P. knowlesi*, *P. vivax* and the rodent malaria genomes *P. berghei*, *P. chabaudi* and *P. yoelii* the OrthoMCL protein clustering algorithm 38 was used (Li et al. 2003). The genomes of *P. gallinaceum* and *P. relictum* can be found on the ftp site (ftp://ftp.sanger.ac.uk/pub/project/pathogens/gff3/CURRENT) and on GeneDB (http://www.genedb.org) (Logan-Klumpler et al. 2012).

Phylogenetic analysis

OrthoMCL v2.0 was used to cluster predicted proteins from 19 species of apicomplexan parasite, including 11 previously published *Plasmodium* species (**P. berghei*, **P. chabaudi*, **P. yoelli* (Otto et al. 2014a)*, *P. cynomolgi* (Tachibana et al. 2012)*, *P. falciparum*, **P. knowlesi* (Pain et al. 2008), **P. reichenowi* (Otto et al. 2014b), **P. vivax, †P. ovale and †P. malariae* (Rutledge et al. 2016), the published *Haemoproteus tartakovskysi* genome (Bensch et al. 2016) and the two new *Plasmodium* genomes described here, together with *Toxoplasma gondii* and the piroplasms *Babesia microti, Babesia bovis, Theileria parvum* and *Theileria annulata*. Data for published *Plasmodium* genomes were downloaded from GeneDB (http://www.genedb.org (Logan-Klumpler et al. 2012)) those marked * on 17/7/2013 and those marked † on 01/06/2016, that for non-*Plasmodium* species from apiDB (http://www.apidb.org) on the same date. OrthoMCL was run with default parameters and an inflation parameter of 1.5; and the output parsed to identify a total of 881 clusters that were single-copy and present in all 19 species. Amino acid sequences for all of these clusters were aligned using mafft v7.205 (Katoh and Standley 2013) with the '--auto' flag and other parameters left as defaults, and then these alignments trimmed using GBlocks v0.91b (Castresana 2000) to keep well-aligned blocks of at least 4 consecutive well-aligned columns separated by no more than 4 less conserved columns and to discard columns with more than half gap characters. All trimmed gene cluster alignments with more than 10 amino acid residues (879 out of 881) were kept for subsequent analysis. Subsequent phylogenetic analyses were all based on this alignment of 289,315 amino acid residues, from 879 single-copy gene clusters.

Bayesian phylogenetic inference was performed using PhyloBayes 3.3f (Lartillot et al. 2009) under a CAT mixture model, allowing the rate of substitutions to vary between sites according to a discretised gamma distribution and the substitution process at each site to come from a mixture of amino acid composition matrices but with a single underlying Poisson process for the substitution process. We ran 8 independent MCMC chains of at least 60,000 steps each. The final 1500 trees from each chain were concatenated for inference (discarding approximately 20,000 steps per chain as burn-in). While model parameter estimates had not all converged across all chains, tree topologies appeared to be following visualisation with "R We There Yet?" (https://github.com/danlwarren/RWTY). Maximum-likelihood phylogenetic analysis using RAxML v.8.0.24 (Stamatakis 2014) was performed using a partitioned model where the alignment for each locus was assigned the best-fitting model under BIC from the set of empirical amino acid substitution matrices available in that version of RAxML and using observed amino acid composition, and under a single LG4X model for the whole alignment with maximum-likelihood estimates of amino acid composition. Additional analyses used PAUP v4.0b10 and Phylip v3.6.9 (Felsenstein 2005) for parsimony and neighbour-joining analysis of standard AA pairwise distances (under the JTT model) and Log-Det distances calculated using LDDist v1.3.2 (Thollesson 2004).

Transposon analysis

LTRharvest (from GenomeTools v1.5.2) (Ellinghaus et al. 2008) was used to search for putative LTR retrotransposon insertions in the sequence scaffold on which the ORF (4455 bp) in question was located. It successfully identified two flanking LTR sequences of 459 bp (5' LTR) and 469 bp (3' LTR) length and 90% similarity. Subsequent annotation of this element using LTRdigest (Steinbiss et al. 2009) revealed the presence of several profile HMM matches to retrotransposon-associated domains (Gag, protease, reverse transcriptase, RNAse H, integrase). Profiles used in this search were collected from the Pfam (Finn et al. 2014) (PF00075, PF00077, PF00078, PF00098, PF00385, PF00552, PF00665, PF00692, PF01021, PF01393, PF02022, PF03732, PF04094, PF04195, PF05380, PF06815, PF06817, PF07253, PF07727, PF08284) and GyDB databases (Llorens et al. 2011). The LTRdigest run also detected a primer binding site of length 15, complementary to a tRNA^{Ser} (anticodon GCT). For this purpose *P. gallinaceum* tRNA sequences were predicted *ab initio* using ARAGORN v1.2.36 (Laslett and Canback 2004). Moreover a polypurine tract of length 27 (AAAAAAAAAAAAAAAAAAAAAAAAAGA) was identified manually by examination of the area upstream of the 3' LTR. Filtering and manual inspection of the results of genome-wide LTRharvest/LTRdigest runs discovered at least four more potential near-full-length copies. However, none of these retains a complete ORF.

RepeatMasker (version open-4.0.2, with ABBlast/WUBlast 2.0MP-WashU, -nolow, default sensitivity) was used to identify fragmented insertions of the element in the genome DNA sequence using the DNA sequence of the full-length element as a custom library. All hits of length less than 400 bp were disregarded.

PEXEL-motif

All genes of the reference genomes were analysed for the presence of a PEXEL-motif using the updated HMM algorithm ExportPred v2.0 (Boddey et al. 2013). As a cutoff value 1.5 was used as in (Boddey et al. 2013). To compare genes with PEXEL-motifs between the species we used only orthologous genes with a one-to-one relationship in the 11 reference species.

Meme-Motif analysis

To predict new motifs we used meme version 4.9.1. For the STP1 and *SURFIN* analysis we searched for 96 motifs of the length between 10-150 amino acids on all the existing STP1 and SURFIN sequences of the used 9 genomes. Next, the conceptual proteomes of 9 *Plasmodium* spp were searched for the presence of those STP1/SURFIN meme motifs using fimo, a tool from the meme suit that finds predicted meme motifs in new sequences (cut-off 1.0E-6; seg used to exclude low complexity amino acid regions). Genes with less than 5 hits were excluded. The output was parsed with a PERL script into a matrix and visualized in R, using the heatmap.2 function and the ward clustering. The phylogenetic trees in Figure 3 were built with PhyML (Guindon et al. 2009). The alignments for those trees are based on the three meme motifs each. We tried to maximize the occurrence of number of sequences and species for the tree.

For the RBP analysis we took 15 RBP from each species. We chose 15 to have the same number of sequences per species. We joined the two Laverania samples and down sampled randomly the amount of sequences if needed. Motifs were predicted with the parameters -nmotifs 96 -minw 10 -maxw 150.

All structural predictions were performed on the I-Tasser web server (Yang et al. 2015) using default parameter. To determine Pfam domain enrichment we ran InterProScan (Mitchell et al. 2015), and parsed the output in a table for further analysis.

Determining the number of RBP copies in P. relictum

The long coding sequences (> 7.5 kb) with large blocks of highly similar sequence between RBPs confounded the assembly process and made determining the true number of RBPs a challenge. In the *P. relictum* assembly, 18 full-length RBP genes were annotated, of which five were pseudogenes. In addition, the *P. relictum* assembly contained numerous gene fragments that were truncated by the assembly process. Despite the high degree of polymorphism, the order of meme motifs in the full length RBP genes was conserved and this information was used to classify and count RBP fragments corresponding to the N and C termini. Of fifteen incomplete genes 8 and 6 were classified as unambiguously representing N- and C-terminal fragments, respectively. To identify copy number variants (CNVs) amongst near-identical RBPs, Illumina reads were mapped back against the genome using BWA (default parameters) and Bamview in Artemis. A CNV was assumed where read depth increased by a factor of 2, 3 or 4, relative to the median coverage depth for the whole genome. The occurrence of heterozygous SNPs within a CNV provided additional supporting evidence. Four almost base-perfect copies of one full length RBP had been collapsed into a single copy in the assembly process.

Expression of P. gallinaceum transposase in P. berghei

A 6.4 kb fragment carrying the *P. gallinaceum* gag-pol (without the LTR repeats) was synthesised by GeneArt Gene Synthesis service (Thermo Fischer Scientific). The *P. gallinaceum* gag-pol was sub-cloned into pL1694 (obtained from the Leiden Malaria Group) using BamHI and NotI, placing its expression under control of the constitutive *pbhsp70* promoter. In the resulting vector, the expression cassette *pbhsp70 5'utr:pgtransp:pbhsp70 3'utr* is flanked by homology arms to the P. *berghei* p230p gene. The vector was linearised by SacII prior to transfection into the 1596 *P. berghei* GIMO (gene in marker out) mother line (Leiden Malaria Group), with transfectant parasites injected intravenously into Balb/c mice followed by administration of 5-fluorocytosine in the drinking water (Janse et al. 2006; Lin et al. 2011; Braks et al. 2006).

Parasites typically appeared 10 day post-transfection, and were genotyped for integration of the *pbhsp70 5'utr:pgtransp:pbhsp70 3'utr* expression cassette (L1694-GT-F AGCGATAAAAATGATAAACCA, L1694-Pgtransp.-GT-R CGATTGACGCTAAATCATTCGG). Transfection was attempted at four independent occasions, in the presence of a positive control without integration of the Pg transposase expression cassette ever being detected.

Data Access

The *P. gallinaceum* 8A reads were deposited in the European Nucleotide Archive with the accession numbers ERS026186, ERS250112 and ERS328512. The *P. relictum* reads were deposited in the European Nucleotide Archive with the accession number ERS412008. Chromosome accession numbers for *P. relictum* are LN835296-LN835311 and scaffolds CVMU01000001-CVMU01000498. Accession numbers for *P. gallinaceum* are LN835293-LN835294 and scaffolds CVMV01000001-

CVMV01000152. The data is also available from GeneDB (http://www.genedb.org).

Acknowledgements

This work was supported by the Wellcome Trust (grant number 098051). Sascha Steinbiss was funded by the Wellcome Trust grant number WT099198MA.

References

Asghar M, Hasselquist D, Hansson B, Zehtindjiev P, Westerdahl H, Bensch S. 2015. Chronic infection. Hidden costs of infection: chronic malaria accelerates telomere degradation and senescence in wild birds. *Science* **347**: 436–438.

Assefa S, Keane TM, Otto TD, Newbold C, Berriman M. 2009. ABACAS: algorithm-based automatic contiguation of assembled sequences. *Bioinforma Oxf Engl* **25**: 1968–1969.

Atkinson CT, Dusek RJ, Woods KL, Iko WM. 2000. Pathogenicity of avian malaria in experimentally-infected Hawaii Amakihi. *J Wildl Dis* **36**: 197–204.

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**: W202-208.

Balu B, Chauhan C, Maher SP, Shoue DA, Kissinger JC, Fraser MJ, Adams JH. 2009. piggyBac is an effective tool for functional analysis of the Plasmodium falciparum genome. *BMC Microbiol* **9**: 83.

Bensch S, Canbäck B, DeBarry JD, Johansson T, Hellgren O, Kissinger JC, Palinauskas V, Videvall E, Valkiūnas G. 2016. The Genome of Haemoproteus tartakovskyi and Its Relationship to Human Malaria Parasites. *Genome Biol Evol* **8**: 1361–1373.

Bensch S, Hellgren O, Pérez-Tris J. 2009. MalAvi: a public database of malaria parasites and related haemosporidians in avian hosts based on mitochondrial cytochrome b lineages. *Mol Ecol Resour* **9**: 1353–1358.

Blanquart S, Gascuel O. 2011. Mitochondrial genes support a common origin of rodent malaria parasites and Plasmodium falciparum's relatives infecting great apes. *BMC Evol Biol* **11**: 70.

Boddey JA, Carvalho TG, Hodder AN, Sargeant TJ, Sleebs BE, Marapana D, Lopaticki S, Nebl T, Cowman AF. 2013. Role of plasmepsin V in export of diverse protein families from the Plasmodium falciparum exportome. *Traffic Cph Den* **14**: 532–550.

Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. 2011. Scaffolding pre-assembled contigs using SSPACE. *Bioinforma Oxf Engl* **27**: 578–579.

Boetzer M, Pirovano W. 2012. Toward almost closed genomes with GapFiller. *Genome Biol* **13**: R56.

Borner J, Pick C, Thiede J, Kolawole OM, Kingsley MT, Schulze J, Cottontail VM, Wellinghausen N, Schmidt-Chanasit J, Bruchhaus I, et al. 2016. Phylogeny of haemosporidian blood parasites revealed by a multi-gene approach. *Mol Phylogenet Evol* **94**: 221–231.

Braks JAM, Franke-Fayard B, Kroeze H, Janse CJ, Waters AP. 2006. Development and application of a positive-negative selectable marker system for use in reverse genetics in Plasmodium. *Nucleic Acids Res* **34**: e39.

Brumpt É. 1937. Schizogonie parfois intense du Plasmodium gallinaceum dans les cellules endotheliales des poules. *C R Soc Biol Paris* 810–813.

Campbell TL, De Silva EK, Olszewski KL, Elemento O, Llinás M. 2010. Identification and genome-wide prediction of DNA binding specificities for the ApiAP2 family of regulators from the malaria parasite. *PLoS Pathog* **6**: e1001165.

Carlton JM, Angiuoli SV, Suh BB, Kooij TW, Pertea M, Silva JC, Ermolaeva MD, Selengut JD, Koo HL, Peterson JD, et al. 2002. Genome sequence and comparative analysis of the model rodent malaria parasite Plasmodium yoelii yoelii. *Nature* **419**: 512–519.

Carver T, Harris SR, Otto TD, Berriman M, Parkhill J, McQuillan JA. 2013. BamView: visualizing and interpretation of next-generation sequencing read alignments. *Brief Bioinform* **14**: 203–212.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**: 540–552.

Cornet S, Nicot A, Rivero A, Gandon S. 2013. Malaria infection increases bird attractiveness to uninfected mosquitoes. *Ecol Lett* **16**: 323–329.

Cunningham D, Lawton J, Jarra W, Preiser P, Langhorne J. 2010. The pir multigene family of Plasmodium: antigenic variation and beyond. *Mol Biochem Parasitol* **170**: 65–73.

Dame JB, Mccutchan TF. 1987. Plasmodium falciparum: Hoechst Dye 33258-CsCl ultracentrifugation for separating parasite and host DNAs. *Exp Parasitol* **64**: 264–266.

Ellinghaus D, Kurtz S, Willhoeft U. 2008. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**: 18.

Feehery GR, Yigit E, Oyola SO, Langhorst BW, Schmidt VT, Stewart FJ, Dimalanta ET, Amaral-Zettler LA, Davis T, Quail MA, et al. 2013. A method for selectively enriching microbial DNA from contaminating vertebrate host DNA. *PloS One* **8**: e76096.

Fell VL, Schild-Poulter C. 2015. The Ku heterodimer: function in DNA repair and beyond. *Mutat Res Rev Mutat Res* **763**: 15–29.

Felsenstein J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. *Distrib Author*.

Fernandez-Becerra C, Yamamoto MM, Vêncio RZN, Lacerda M, Rosanas-Urgell A, del Portillo HA. 2009. Plasmodium vivax and the importance of the subtelomeric multigene vir superfamily. *Trends Parasitol* **25**: 44–51.

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, et al. 2014. Pfam: the protein families database. *Nucleic Acids Res* **42**: D222-230.

Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, et al. 2002. Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* **419**: 498–511.

Garnham PCC. 1966. *Malaria parasites and other haemosporidia*. Blackwell Scientific Publishers, Oxford, UK.

Glaizot O, Fumagalli L, Iritano K, Lalubin F, Van Rooyen J, Christe P. 2012. High prevalence and lineage diversity of avian malaria in wild populations of great tits (Parus major) and mosquitoes (Culex pipiens). *PloS One* **7**: e34964.

Guindon S, Delsuc F, Dufayard J-F, Gascuel O. 2009. Estimating maximum likelihood phylogenies with PhyML. *Methods Mol Biol Clifton NJ* **537**: 113–137.

Haltiwanger BM, Matsumoto Y, Nicolas E, Dianov GL, Bohr VA, Taraschi TF. 2000. DNA base excision repair in human malaria parasites is predominantly by a long-patch pathway. *Biochemistry (Mosc)* **39**: 763–772.

Huff CG. 1969. Exoerythrocytic stages of avian and reptilian malarial parasites. *Exp Parasitol* **24**: 383–421.

Huff CG, Bloom W. 1935. A malarial parasite infecting all blood and blood-forming cells in birds. *J Infect Dis* **57**: 315–336.

Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, Otto TD. 2013. REAPR: a universal tool for genome assembly evaluation. *Genome Biol* **14**: R47.

Janse CJ, Ramesar J, Waters AP. 2006. High-efficiency transfection and drug selection of genetically transformed blood stages of the rodent malaria parasite Plasmodium berghei. *Nat Protoc* **1**: 346–356.

Janssen CS, Phillips RS, Turner CMR, Barrett MP. 2004. Plasmodium interspersed repeats: the major multigene superfamily of malaria parasites. *Nucleic Acids Res* **32**: 5712–5720.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**: 772–780.

Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305**: 567–580.

Lachish S, Knowles SCL, Alves R, Wood MJ, Sheldon BC. 2011. Fitness effects of endemic malaria infections in a wild bird population: the importance of ecological structure. *J Anim Ecol* **80**: 1196–1206.

Lapointe DA, Atkinson CT, Samuel MD. 2012. Ecology and conservation biology of avian malaria. *Ann N Y Acad Sci* **1249**: 211–226.

Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinforma Oxf Engl* **25**: 2286–2288.

Laslett D, Canback B. 2004. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res* **32**: 11–16.

Lauron EJ, Aw Yeang HX, Taffner SM, Sehgal RNM. 2015. De novo assembly and transcriptome analysis of Plasmodium gallinaceum identifies the Rh5 interacting protein (ripr), and reveals a lack of EBL and RH gene family diversification. *Malar J* **14**: 296.

Lauron EJ, Oakgrove KS, Tell LA, Biskar K, Roy SW, Sehgal RN. 2014. Transcriptome sequencing and analysis of Plasmodium gallinaceum reveals polymorphisms and selection on the apical membrane antigen-1. *Malar J* **13**: 382.

Levin II, Zwiers P, Deem SL, Geest EA, Higashiguchi JM, Iezhova TA, Jiménez-Uzcátegui G, Kim DH, Morton JP, Perlut NG, et al. 2013. Multiple lineages of Avian malaria parasites (Plasmodium) in the Galapagos Islands and evidence for arrival via migratory birds. *Conserv Biol J Soc Conserv Biol* **27**: 1366–1377.

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**: 2178–2189.

Lin J, Annoura T, Sajid M, Chevalley-Maurel S, Ramesar J, Klop O, Franke-Fayard BMD, Janse CJ, Khan SM. 2011. A novel “gene insertion/marker out” (GIMO) method for transgene expression and gene complementation in rodent malaria parasites. *PloS One* **6**: e29289.

Ling K-H, Rajandream M-A, Rivailler P, Ivens A, Yap S-J, Madeira AMBN, Mungall K, Billington K, Yee W-Y, Bankier AT, et al. 2007. Sequencing and analysis of chromosome 1 of Eimeria tenella reveals a unique segmental organization. *Genome Res* **17**: 311–319.

Llorens C, Futami R, Covelli L, Domínguez-Escribá L, Viu JM, Tamarit D, Aguilar-Rodríguez J, Vicente-Ripolles M, Fuster G, Bernet GP, et al. 2011. The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. *Nucleic Acids Res* **39**: D70-74.

Logan-Klumpler FJ, De Silva N, Boehme U, Rogers MB, Velarde G, McQuillan JA, Carver T, Aslett M, Olsen C, Subramanian S, et al. 2012. GeneDB--an annotation database for pathogens. *Nucleic Acids Res* **40**: D98-108.

Marshall EK. 1942. Chemotherapy of avian malaria. *Physiol Rev* **22**: 190–204.

Marti M, Good RT, Rug M, Knuepfer E, Cowman AF. 2004. Targeting malaria virulence and remodeling proteins to the host erythrocyte. *Science* **306**: 1930–1933.

Merino EF, Fernandez-Becerra C, Durham AM, Ferreira JE, Tumilasci VF, d'Arc-Neves J, da Silva-Nunes M, Ferreira MU, Wickramarachchi T, Udagama-Randeniya P, et al. 2006. Multi-character population study of the vir subtelomeric multigene superfamily of Plasmodium vivax, a major human malaria parasite. *Mol Biochem Parasitol* **149**: 10–16.

Mitchell A, Chang H-Y, Daugherty L, Fraser M, Hunter S, Lopez R, McAnulla C, McMenamin C, Nuka G, Pesseat S, et al. 2015. The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res* **43**: D213-221.

Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, Floden EW, Gardner PP, Jones TA, Tate J, et al. 2015. Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res* **43**: D130-137.

Otto TD, Böhme U, Jackson AP, Hunt M, Franke-Fayard B, Hoeijmakers WAM, Religa AA, Robertson L, Sanders M, Ogun SA, et al. 2014a. A comprehensive evaluation of rodent malaria parasite genomes and gene expression. *BMC Biol* **12**: 86.

Otto TD, Dillon GP, Degrave WS, Berriman M. 2011. RATT: Rapid Annotation Transfer Tool. *Nucleic Acids Res* **39**: e57.

Otto TD, Rayner JC, Böhme U, Pain A, Spottiswoode N, Sanders M, Quail M, Ollomo B, Renaud F, Thomas AW, et al. 2014b. Genome sequencing of chimpanzee malaria parasites reveals possible pathways of adaptation to human hosts. *Nat Commun* **5**: 4754.

Otto TD, Sanders M, Berriman M, Newbold C. 2010. Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology. *Bioinforma Oxf Engl* **26**: 1704–1707.

Oyola SO, Gu Y, Manske M, Otto TD, O'Brien J, Alcock D, Macinnis B, Berriman M, Newbold CI, Kwiatkowski DP, et al. 2013. Efficient depletion of host DNA contamination in malaria clinical sequencing. *J Clin Microbiol* **51**: 745–751.

Oyola SO, Manske M, Campino S, Claessens A, Hamilton WL, Kekre M, Drury E, Mead D, Gu Y, Miles A, et al. 2014. Optimized whole-genome amplification strategy for extremely AT-biased template. *DNA Res Int J Rapid Publ Rep Genes Genomes* **21**: 661–671.

Pain A, Böhme U, Berry AE, Mungall K, Finn RD, Jackson AP, Mourier T, Mistry J, Pasini EM, Aslett MA, et al. 2008. The genome of the simian and human malaria parasite Plasmodium knowlesi. *Nature* **455**: 799–803.

Park N, Shirley L, Gu Y, Keane TM, Swerdlow H, Quail MA. 2013. An improved approach to mate-paired library preparation for Illumina sequencing. *Methods Gener Seq* **1**.

Patra KP, Vinetz JM. 2012. New Ultrastructural Analysis of the Invasive Apparatus of the Plasmodium Ookinete. *Am J Trop Med Hyg* **87**: 412–417.

Perkins SL. 2014. Malaria's many mates: past, present, and future of the systematics of the order Haemosporida. *J Parasitol* **100**: 11–25.

Perkins SL, Schaer J. 2016. A Modern Menagerie of Mammalian Malaria. *Trends Parasitol* **32**: 772–782.

Pick C, Ebersberger I, Spielmann T, Bruchhaus I, Burmester T. 2011. Phylogenomic analyses of malaria parasites and evolution of their exported proteins. *BMC Evol Biol* **11**: 167.

Pigeault R, Nicot A, Gandon S, Rivero A. 2015. Mosquito age and avian malaria infection. *Malar J* **14**: 383.

Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13**: 341.

Raffaela G, Marchiafava E. 1944. Avian malaria: a new lease of life for an old experimental model to study the evolutionary ecology of Plasmodium. *Ann Soc Belg Med Trop* 323–330.

Reid AJ, Blake DP, Ansari HR, Billington K, Browne HP, Bryant JM, Dunn M, Hung SS, Kawahara F, Miranda-Saavedra D, et al. 2014. Genomic analysis of the causative agents of coccidiosis in domestic chickens. *Genome Res*.

Rutledge GG, Boehme U, Sanders M, Reid AJ, Maiga-Ascofare O, Djimde AA, Apinjoh TO, Amenga-Etego L, Manske M, Barnwell JW, et al. 2016. *Elusive Plasmodium Species Complete the Human Malaria Genome Set*. bioRxiv doi: 10.1101/052696.

Simpson JT, Durbin R. 2012. Efficient de novo assembly of large genomes using compressed data structures. *Genome Res* **22**: 549–556.

Spence PJ, Jarra W, Lévy P, Reid AJ, Chappell L, Brugat T, Sanders M, Berriman M, Langhorne J. 2013. Vector transmission regulates immune control of Plasmodium virulence. *Nature* **498**: 228–231.

Springer WT. 1996. Other blood and tissue protozoa. In *Calnek, B.W, Beard, C.W., McDougald, L.R., Saif, Y.M. (Eds.). Diseases of Poultry*, pp. 900–911, Ames, IA: Iowa State University Press.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinforma Oxf Engl* **30**: 1312–1313.

Steinbiss S, Willhoeft U, Gremme G, Kurtz S. 2009. Fine-grained annotation and classification of de novo predicted LTR retrotransposons. *Nucleic Acids Res* **37**: 7002–7013.

Sundararaman SA, Plenderleith LJ, Liu W, Loy DE, Learn GH, Li Y, Shaw KS, Ayouba A, Peeters M, Speede S, et al. 2016. Genomes of cryptic chimpanzee Plasmodium species reveal key evolutionary events leading to human malaria. *Nat Commun* **7**: 11078.

Swain MT, Tsai IJ, Assefa SA, Newbold C, Berriman M, Otto TD. 2012. A post-assembly genome-improvement toolkit (PAGIT) to obtain annotated genomes from contigs. *Nat Protoc* **7**: 1260–1284.

Tachibana S-I, Sullivan SA, Kawai S, Nakamura S, Kim HR, Goto N, Arisue N, Palacpac NMQ, Honma H, Yagi M, et al. 2012. Plasmodium cynomolgi genome sequences provide insight into Plasmodium vivax and the monkey malaria clade. *Nat Genet* **44**: 1051–1055.

Templeton TJ, Enomoto S, Chen W-J, Huang C-G, Lancto CA, Abrahamsen MS, Zhu G. 2010. A genome-sequence survey for Ascogregarina taiwanensis supports evolutionary affiliation but metabolic diversity between a Gregarine and Cryptosporidium. *Mol Biol Evol* **27**: 235–248.

Thollesson M. 2004. LDDist: a Perl module for calculating LogDet pair-wise distances for protein and nucleotide sequences. *Bioinforma Oxf Engl* **20**: 416–418.

Tsai IJ, Otto TD, Berriman M. 2010. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome Biol* **11**: R41.

Valkiunas G. 2004. *Avian Malaria Parasites and other Haemosporidia*. CRC Press.

Ward JH. 1963. Hierarchical Grouping to Optimize an Objective Function. *J Am Stat Assoc* **58**: 236–244.

Waters AP, Higgins DG, McCutchan TF. 1991. Plasmodium falciparum appears to have arisen as a result of lateral transfer between avian and human hosts. *Proc Natl Acad Sci U S A* **88**: 3140–3144.

Williams RB. 2005. Avian malaria: clinical and chemical pathology of Plasmodium gallinaceum in the domesticated fowl Gallus gallus. *Avian Pathol J WVPA* **34**: 29–47.

Winter G, Kawai S, Haeggström M, Kaneko O, von Euler A, Kawazu S, Palm D, Fernandez V, Wahlgren M. 2005. SURFIN is a polymorphic antigen expressed on Plasmodium falciparum merozoites and infected erythrocytes. *J Exp Med* **201**: 1853–1863.

Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. 2015. The I-TASSER Suite: protein structure and function prediction. *Nat Methods* **12**: 7–8.

Zélé F, Nicot A, Berthomieu A, Weill M, Duron O, Rivero A. 2014. Wolbachia increases susceptibility to Plasmodium infection in a natural system. *Proc Biol Sci* **281**: 20132837.

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**: 821–829.

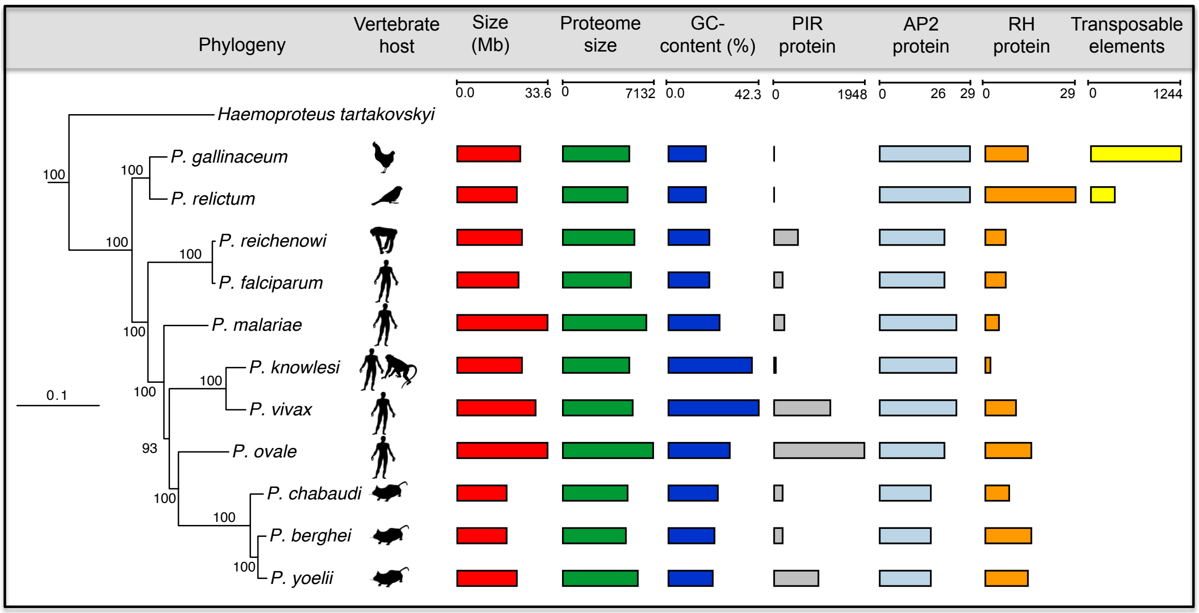

Figure 1. Phylogeny and key features of *Plasmodium* species.

The figure shows maximum likelihood phylogeny of *Plasmodium* species based on a concatenated alignment of 289,315 amino acid residues from 879 single-copy orthologs. Branch lengths are expected substitutions per amino acid site and values on nodes are number of bootstrap replicates (out of 100) displaying the partition induced by the node.The tree was rooted with sequences from *Toxoplasma* and four piroplasm species, with the full tree shown as Supplementary Figure S2. The phylogenetic tree is combined with a graphical overview of key features of all reference genomes (genome versions from 1.5.2016). Due to the fragmented nature of the *Haemoproteus tartakovskyi* (Bensch et al. 2016) genome, counts for its key features have not been included.

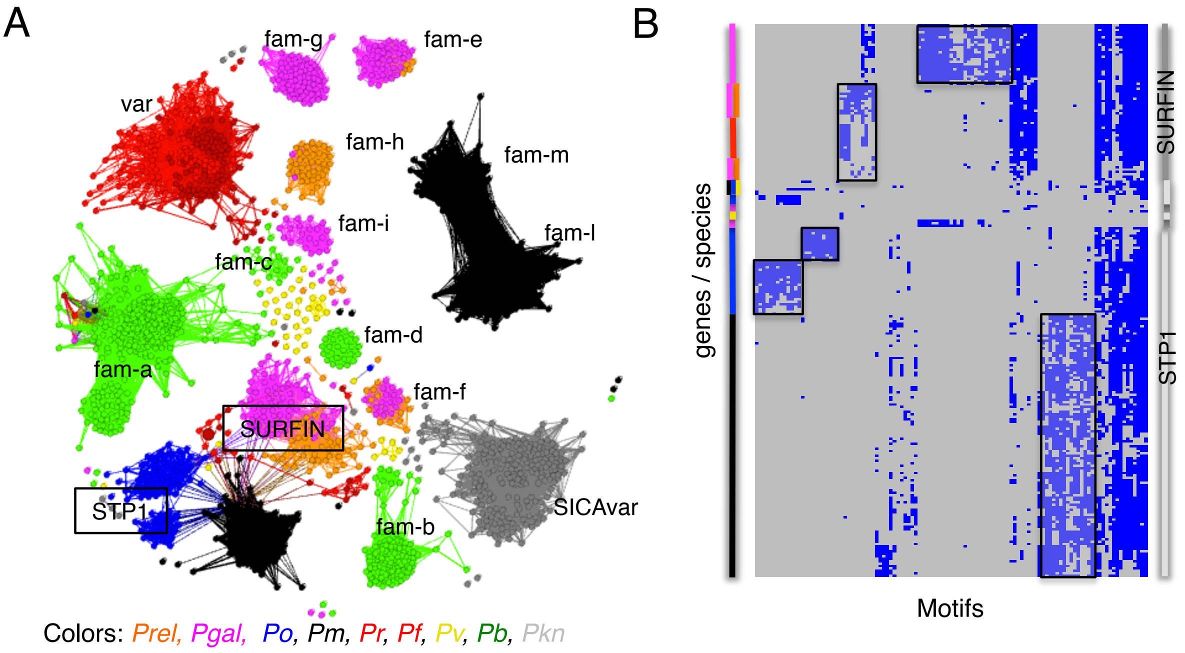

Figure 2. Similarity of gene families within *Plasmodium*

(A) A network of BLASTp similarity between genes (nodes) sharing at least 31% global identity. Genes are coloured by species. The *pir* genes were excluded due to their large numbers across the *Plasmodium* genus. *Fam-m* and *Fam-l* are *P. malariae* specific gene families (Rutledge et al. 2016) (B) Clustering of STP1 and SURFIN genes based on the occurrence motifs identified using MEME. Where a gene (row) has a specific motif (column) the value is set to one. The matrix is clustered through a hierarchical clustering algorithm (Ward 1963), to visualize similar patterns of motif-sharing. The x - axis represents motifs that occur in at least 10 genes, and individual genes are displaye don the y-axis (rows). Coloured bars on the left identify species, the bar on the right the gene annotation. Boxed areas indicate possible gene family sub-types.

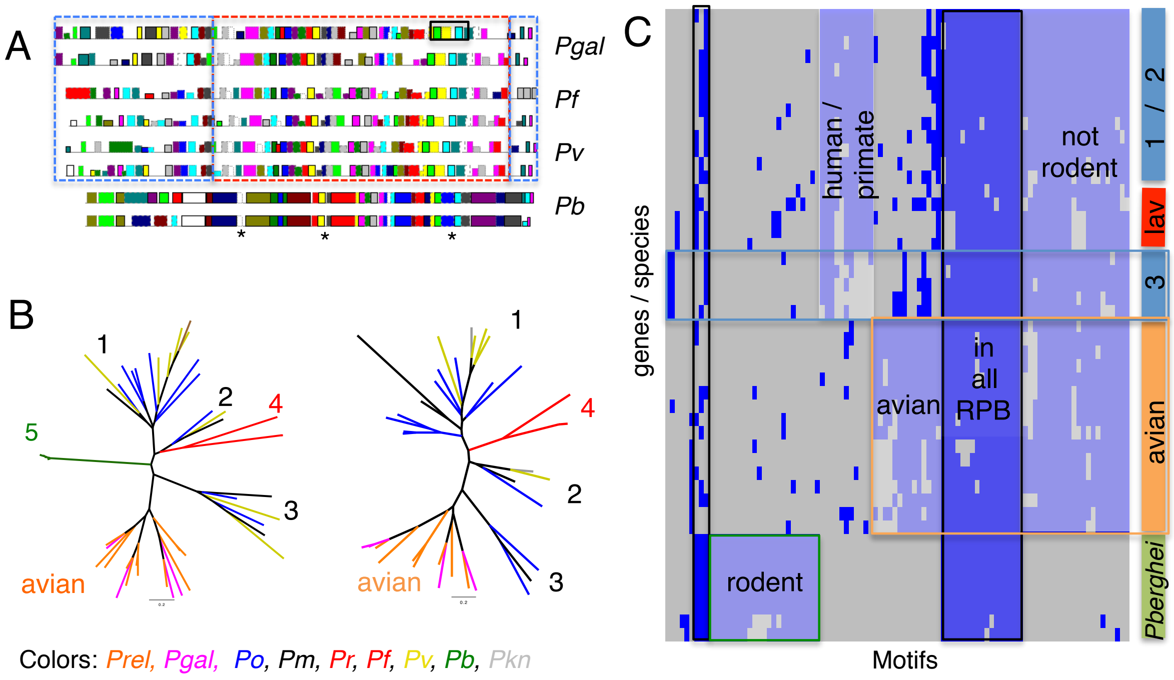

Figure 3: RBP Meme motifs comparison

Analysis of 96 meme motifs obtained from reticulocyte binding proteins (RBP) of nine species. (A) Example of motifs predicted on two RBP from each of four species. Each colored rectangle represents a different one of the 96 motifs. Its height represents the e-value of the individual meme motif at that position. The red dashed square highlights a similar order of motifs between *P. gallinaceum*, *P. falciparum* and *P. vivax*. The blue dashed boxes on either side highlight differences in motif content. The black box and the three stars are motifs used to build the tree in (B). (B): Two maximum likelihood phylogenetic trees based on two motifs set. The left tree was generated using the three motifs (indicated with an asterisk * in Panel A) and the second tree was generated using the motifs from the black dashed box in panel A. 1, 2 and 3 are the distinct clusters of the *P. malariae*, *P. ovale* and *P. vivax* RBP, as reported before (Rutledge et al. 2016), 4 *P. falciparum* and *P. reichenowi* and 5 *P. berghei*. (C) Clustering of the binary occurrence of meme motifs for each RBP, similar to figure (2B). The bar on the right represents either species (lav-*Laverania*, avian, *P.berghei*) or the groups 1,2 and 3 from (B) This analysis does not split group one and two of *P. malariae*, *P. ovale* and *P. vivax* RBPs. The x-axis represents the 96 motifs. Blue represents at least one occurrence of that motif for that gene. Shared patterns are highlighted with coloured boxes.

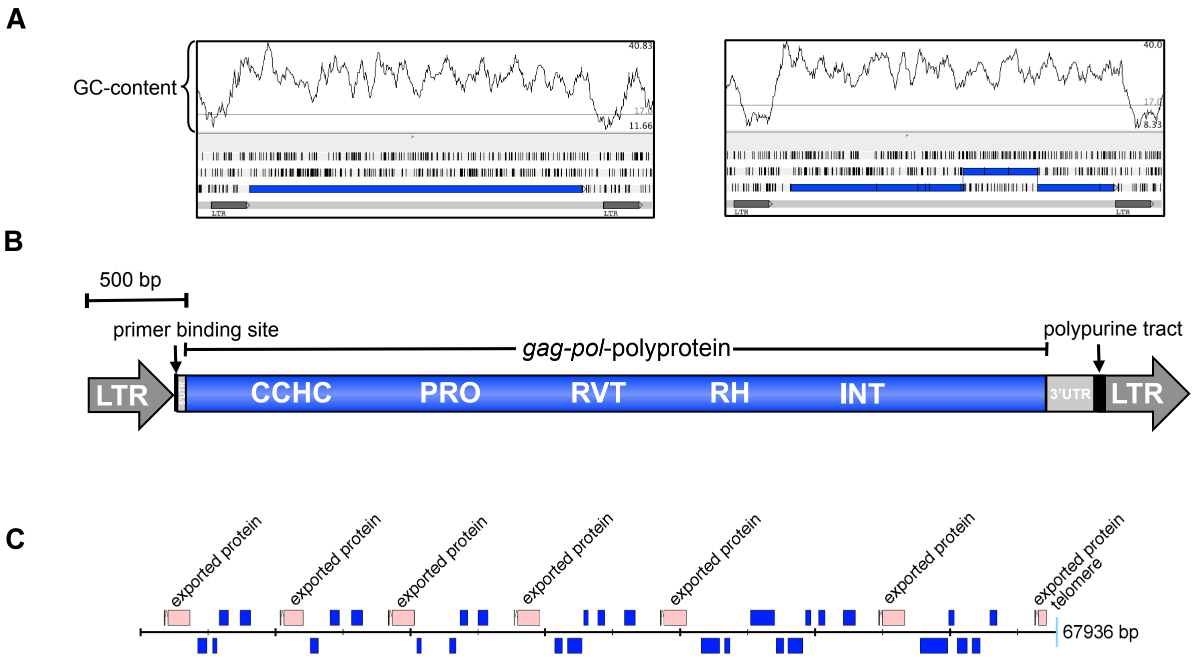
A
GC-content
40.83
11.66
40.0
8.33
LTR
LTR
LTR
LTR
B
500 bp
primer binding site
gag-pol-polyprotein
polypurine tract
LTR
CCHC
PRO
RVT
RH
INT
3'UTR
LTR
C
exported protein
exported protein
exported protein
exported protein
exported protein
exported protein
exported protein
telomere
67936 bp

Figure 4. Retotransposable elements in *P. gallinaceum*

(A) Artemis screenshot showing a complete retrotransposon of *P. gallinaceum* (PGAL8A_00410600) and a copy where the open reading frame encoding gag-pol-polyprotein is frame-shifted.

(B) Diagram of the *P. gallinaceum* retrotransposon (PGAL8A_00410600). The ty3/gypsy retro-element contains a continuous open reading frame including a CCHC-type zinc finger domain (CCHC), aspartic protease domain (PRO), reverse transcriptase domain (RVT), RNAse H domain (RH) and an integrase domain (INT). The element is bounded by long terminal repeats (LTR).

(C) A single subtelomeric region from *P. gallinaceum*. Retrotransposon elements are shown in blue.

Table 1: Genome and gene statistics for *P. gallinaceum* 8A, *P. relictum* SGS1, *P. falciparum* 3D7 and *P. knowlesi* H. Statistics of *P. falciparum* 3D7 (version 3.1) and *P. knowlesi* H (v2, from 01.05.2016) are shown for comparison.

	***P. gallinaceum* 8A**	***P. relictum* SGS1**	***P. falciparum* 3D7**	***P. knowlesi* H**
Nuclear genome				
genome size (Mb)	23.8	22.6	23.3	24.3
G+C content (%)	17.83	18.34	19.34	40.2
gaps	1840	236	0	83
no. of scaffolds	152	498	14	162
no. of chromosomes	ND	14	14	14
no. of genes*	5273	5146	5429	5291
no. of transposable elements (>400bp)	1244	344	0	0
no. of tRNAs	46	46	45	45
Mitochondrial genome				
genome size (bp)	6,747	6,092	5,967	5,957
G+C content (%)	32.58	31.68	31.6	30.52
no. of genes	3	3	3	3
Apicoplast genome				
genome size (kb)	29.4	29.4	34.3	30.6
G+C content (%)	12.9	13.06	14.22	14.03
no. of genes	30	30	30	30

* including pseudogenes and partial genes, excluding non-coding RNA genes

Table 2: Number of gene members of different subtelomeric multigene families in the genomes of *P. gallinaceum* 8A, *P. relictum* SGS1

Gene family	**No. of gene members**	
	P. gallinaceum	***P. relictum***
PIR-like protein	20	4
PIR-like, pseudogene	1	0
PIR-like, fragment	1	1
surface-associated interspersed protein (SURFIN)	40	14
surface-associated interspersed protein (SURFIN), pseudogene	4	16
surface-associated interspersed protein (SURFIN), fragment	35	20
early transcribed membrane protein	12	12
early transcribed membrane protein, pseudogene	0	1
early transcribed membrane protein, fragment	0	1
reticulocyte binding protein, putative	8	13
reticulocyte binding protein, pseudogene	4	5
reticulocyte binding protein, fragment	2	15
fam-e protein	38	4
fam-e protein, pseudogene	0	0
fam-e protein, fragment	11	0
fam-f protein	16	14
fam-f protein, pseudogene	2	0
fam-f protein, fragment	0	1
fam-g protein	107	0
fam-g protein, pseudogene	2	0
fam-g protein, fragment	12	0
fam-h protein	2	49
fam-h protein, pseudogene	0	0
fam-h protein, fragment	0	0
fam-i protein	23	0
fam-i, pseudogene	0	0
fam-i, fragment	3	0