bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Genomic and transcriptomic evidence for descent from *Plasmodium* and loss of blood schizogony in *Hepatocystis* parasites from naturally infected red colobus monkeys

Eerik Aunin[1], Ulrike Böhme[1], Theo Sanderson[2], Noah D Simons[3], Tony L Goldberg[4] , Nelson Ting[3], Colin A Chapman[5,6,7], Chris I Newbold[1,8], Matthew Berriman[1] & Adam J Reid[1]

[1]Parasite Genomics, Wellcome Sanger Institute, Hinxton, Cambridge, United Kingdom

[2]Malaria Biochemistry Laboratory, The Francis Crick Institute, London, United Kingdom

[3]Department of Anthropology and Institute of Ecology and Evolution, University of Oregon, Eugene, Oregon

[4]Department of Pathobiological Sciences, School of Veterinary Medicine, University of Wisconsin-Madison, Madison, WI 53706, USA

[5]Department of Anthropology, Center for the Advanced Study of Human Paleobiology, The George Washington University, Washington DC, USA, 20037

[6]Shaanxi Key Laboratory for Animal Conservation, Northwest University, Xi'an, China

[7]School of Life Sciences, University of KwaZulu-Natal, Scottsville, Pietermaritzburg, South Africa,

[8]Weatherall Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Oxford, United Kingdom

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

ABSTRACT

Hepatocystis is a genus of single-celled parasites infecting monkeys, bats and squirrels. Although thought to descend from malaria parasites (*Plasmodium spp.*), *Hepatocystis spp.* are thought not to undergo replication in the blood – the part of the *Plasmodium* life cycle which causes the symptoms of malaria. Furthermore, *Hepatocystis* is transmitted by midges, not mosquitoes. Comparative genomics of *Hepatocystis* and *Plasmodium* species therefore presents an opportunity to better understand some of the most important aspects of malaria parasite biology. We were able to generate a draft genome for *Hepatocystis* using DNA sequencing reads from the blood of a naturally infected red colobus monkey. We provide robust phylogenetic support for *Hepatocystis* as a sister group to *Plasmodium* parasites infecting rodents. We show transcriptomic support for a lack of replication in the blood and genomic support for a complete loss of a family of genes involved in red blood cell invasion. Our analyses highlight the rapid evolution of genes involved in parasite vector stages, revealing genes that may be critical for interactions between malaria parasites and mosquitoes.

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Introduction

Species of the genus *Hepatocystis* are single-celled eukaryotic parasites infecting Old World monkeys, fruit bats and squirrels (1). Phylogenetically, they are thought to reside within a clade containing *Plasmodium* species, including the parasites causing malaria in humans (2). They were originally considered distinct from *Plasmodium* and have remained in a different genus because they lack the defining feature of asexual development in the blood, known as erythrocytic schizogony (3). The presence of macroscopic exoerythrocytic schizonts (merocysts) in the liver of the vertebrate host is the most prominent feature of *Hepatocystis (3)*. Similar to *Plasmodium* parasites, *Hepatocystis* merocysts yield many single-celled merozoites. However, unlike *Plasmodium*, *Hepatocystis* merocysts appear to be the only replication phase in the vertebrate host (4). First generation merozoites of *Plasmodium* spp. are released from liver cells and invade red blood cells, where they multiply asexually, before erupting from red cells as secondary merozoites. These merozoites invade further red blood cells before some develop into stages that can be transmitted to the vector. In contrast, liver merozoites of *Hepatocystis* spp. are thought to commit to the development of transmission stages directly upon invading red blood cells. They are then vectored not by mosquitoes, but by biting midges of the genus *Culicoides* (5). After fertilisation, *Hepatocystis* ookinetes encyst in the head and thorax of the midge between muscle fibres, whereas *Plasmodium* ookinetes encyst in the midgut wall of mosquitoes. After maturation, oocysts of both *Plasmodium* and *Hepatocystis* rupture and release sporozoites that migrate to the salivary glands (1). These discrete biological differences, in the face of phylogenetic similarity and many shared biological features, make *Hepatocystis* a potentially powerful comparator for understanding important aspects of malaria parasite biology, such as transmission and host specificity.

A population of red colobus monkeys (*Piliocolobus tephrosceles*), from Kibale National Park, Uganda were previously shown to host *Hepatocystis* based on morphological identification of infected red blood cells and DNA sequencing of the *cytochrome b* gene (6). In this work we use *Hepatocystis* genome and transcriptome sequences derived from *P. tephrosceles* whole blood samples to generate a draft genome sequence and gain insights into *Hepatocystis* evolution. We go on to use these insights to explore key aspects of malaria parasite biology such as red blood cell invasion, gametocytogenesis and parasite-vector interactions.

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Results

Genome assembly and annotation

To produce a genome assembly for *Hepatocystis* we took advantage of a published genomic sequence from a red colobus monkey (*Piliocolobus tephrosceles*) in Kibale National Park, Uganda (NCBI assembly ASM277652v1), where *Hepatocystis* had previously been reported (6). We examined AT-content and sequence similarity to *Plasmodium spp.* and macaque (Fig 1A). This revealed a subset of contigs with high AT content and similarity to *Plasmodium* spp. Phylogenetic analysis, using an orthologue of *cytochrome b* from these contigs, suggested that they represent the first substantial genomic sequence from the genus *Hepatocystis* (Fig 1B), which is known to infect similar types of Old World monkeys (6). At least four species of *Hepatocystis* are known to infect African monkeys – *H. kochi, H. simiae, H. bouillezi* and *H. cercopitheci* (6) – but with little sequence data currently linked to morphological identification, it was not possible to determine the species. We have thus classified the parasite as *Hepatocystis* sp. ex *Piliocolobus tephrosceles* (*HexPt*; NCBI Taxonomy ID: 2600580). The extraction of the *Hepatocystis* sequences from the *P. tephrosceles* assembly yielded a set of 11,877 scaffolds with a total size of 26.26 Mb and an N50 of 2.4 kb. Automated genome annotation with Companion (7) identified 2,967 genes and 1,432 pseudogenes in these scaffolds. To improve upon this assembly, we isolated putative *Hepatocystis* reads from the original short read DNA sequencing data. These were assembled into a draft quality nuclear genome assembly of 19.94 Mb, comprising 2,435 contigs with an N50 of 18.4 kb (Table 1). The GC content (22.05%) was identical to that of *Plasmodium* spp. infecting rodents, and slightly higher than *P. falciparum* (19.34%). We identified 5,340 genes, compared to 5,441 in *P. falciparum* and 5,049 in *P. berghei*, suggesting a largely complete sequence (Table 1; S1 Table). Despite the fragmented nature of the assembly, we were able to identify synteny with *P. falciparum* around centromeres (S1 Fig) and evidence of clustering of contingency gene families (S2 Fig), as seen in most *Plasmodium* species.

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 1. Features of the *Hepatocystis* sp. ex. *Piliocolobus tephrosceles* assembly compared to *P. falciparum* 3D7, *P. vivax* P01 and *P. berghei* ANKA.

	Hepatocystis	*P. falciparum* 3D7	*P. vivax* P01	*P. berghei* ANKA
Nuclear genome				
Genome size (Mb)	19.94	23.3	29.0	18.7
G+C content (%)	22.05	19.34	39.8	22.05
Gaps within scaffolds	979	0	431	0
No. of scaffolds	2435	14	240	19
No. of chromosomes	ND	14	14	14
No. of genes*	5,340	5,441	6,650	5,049
No. of pseudogenes	25	158	158	129
No. of partial genes	1,475	0	196	8
No. of ncRNA	19	103	35	47
No. of tRNAs	41	45	45	45
No. of telomeres	0**	26	1	12
No. of centromeres	5	13	14	14
Mitochondrial genome				
Genome size (bp)	6,595	5,967	5,989	5,957
G+C content (%)	30.99	31.6	30.5	30.9
No. of genes	3	3	3	3
Apicoplast genome				
Genome size (kp)	27.0	34.3	29.6	34.3
G+C content (%)	13.29	14.22	13.3	15.1
No. of genes	28	30	30	30
Completeness				
CEGMA - complete	63.31%	69.35%	68.15%	70.16%
as least partial	69.35%	71.77%	71.77%	73.39%

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

** two small contigs have telomeric repeats (scaffold 2410 (5 telomeric repeats, scaffold 2364, 9 telomeric repeats))

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Phylogenetic position of *Hepatocystis* sp. ex. *Piliocolobus tephrosceles*

There is consensus that *Hepatocystis* spp. are nested within the *Plasmodium* genus (2,8,9), however their placement within the genus has not been robustly determined. Indeed, our *cytochrome b* phylogeny confirms that our assembled genome is that of *Hepatocystis spp.*, but it provides little support for the placement of this genus in relation to *Plasmodium* spp. A phylogeny generated using all mitochondrially encoded protein sequences also provided little support for key nodes (S3 Fig; S1 Dataset). The mitochondrial genome is therefore not reliable for determining the species phylogeny. A phylogeny based on 18 apicoplast proteins was more robust and placed *Hepatocystis* as an outgroup to the *Plasmodium* species infecting rodents (*Vinckeia;* S4 Fig; S1 Dataset). We wanted to reliably place *HexPt* relative to other *Hepatocystis* species. Limited sequence data are available for *Hepatocystis* outside of this study, however 11 genes from multiple organellar genomes have been sequenced for *H. epomophori*, a parasite of bats (2). Based on the sequence of these genes, we found that *HexPt* forms a sister group to *H. epomophori* (S5 Fig; S1 Dataset). Furthermore, the *Hepatocystis* genus again forms a sister group to the *Vinckei* subgenus of *Plasmodium,* although the tree contains some ambiguous branch points. To improve the robustness of the placement of *Hepatocystis* within *Plasmodium* we used 3271 orthologous nuclear genes from each of 12 species across the *Plasmodium* genus, which robustly places *Hepatocystis* as a sister clade to the *Plasmodium* species infecting rodents (subgenus *Vinckeia;* Fig 2; S1 Dataset). Interestingly, some *Vinckeia* species (*P. cyclopsi*) also infect bats, supporting an earlier suggestion that the ancestor of *Hepatocystis* and *Vinckeia* might have infected bats (10). However, whole-genome data also suggest that *Vinckeia* is derived from a group of monkey-infecting parasites (11). *Hepatocystis* has a long branch that could indicate rapid evolution after splitting from its ancestor with *Vinckeia*.

In vivo transcriptome data supports a lack of erythrocytic schizogony

Transcriptome sequencing of blood samples from 29 individuals was performed as part of the red colobus monkey genome sequencing project (12). We found evidence that each of these individuals was infected with the same species of *Hepatocystis* as found in the genomic reads, consistent with high prevalence of this parasite in Kibale red colobus monkeys as previously reported (6). The extremely low SNP density suggested parasites between hosts were highly related (Fig 3A). We identified an average of 1.37 SNPs (standard deviation = 0.46) and 0.41 indels (standard deviation 0.17) per 10 kb of genome when calling variants using RNA-seq reads. Although it is believed that *Hepatocystis* spp. do not undergo erythrocytic schizogony (3), this has been challenged by limited microscopic evidence for asexual stages in the blood (13). To determine whether there was transcriptomic evidence for schizonts in the blood we deconvoluted the *Hepatocystis* transcriptomes

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

using transcriptome profiles representing different *Plasmodium* life stages. We found no such evidence, but observed varying proportions of cells identified as early blood stages (rings/trophozoites) and mature gametocytes (Fig 3B; S6 Fig; S2 Table). It is thought that chronic infections (of up to 15 months) may be maintained from continual development in the liver (3). The presence of early blood stages in these individuals may therefore reflect this continual production of new blood forms, rather than recent infection. Proportions of rings and trophozoites were positively correlated and both these forms correlated negatively with female gametocytes (Fig 3C). Interestingly, the inferred proportions of male and female gametocytes were not strongly correlated suggesting there might be variation in commitment rates of gametocytes to male or female development.

Expanded and novel gene families

The largest gene family in *Hepatocystis* sp. was a novel family, which we have named *Hepatocystis*-specific family 1 (*hep1*; Table 2). These 12 single-exon genes (plus four pseudogenes) each encode proteins of ~250 amino acids, beginning with a predicted signal peptide (S7A Fig). We could find no significant sequence similarity to genes from any other sequenced genome (using HHblits (14)). However, they contain a repeat region with striking similarity to that in *Plasmodium kahrp*, a gene involved in presenting proteins on the red blood cell surface (15). Three members were highly expressed in *in vivo* blood stages, with one correlating well with the presence of early stages (S8A Fig; HEP_00211100, Pearson's r=0.80 with rings). Another novel family, *hep2*, contained N-terminal PEXEL motifs, suggesting it is exported (S7B Fig). Distinct parts of its sequence showed similarity (albeit with low significance) to exported proteins from *P. malariae* (PmUG01_00051800; probability: 77.88% HHblits) and *P. ovale curtisi* (PocGH01_00025800; 78.47%) as well as a gene in *P. gallinaceum* (PGAL8A_00461100; 69.52%). Three or four members were highly expressed in blood stages *in vivo*, with one member highly correlated with predicted proportions of early blood stages (S8B Fig; HEP_00165500, Pearson's r=0.83 with rings; HEP_00480100, Pearson's r=0.77 with rings).

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 2. Frequencies of members of known and novel gene families in *Hepatocystis*

Gene family	**Frequency**
ApiAP2 transcription factors	27
Hepatocystis-specific family 1 (*hep1*)	16 (includes 4 pseudogenes)
Hepatocystis-specifc family 2 (*hep2*)	10 (includes 4 pseudogenes)
cpw-wpc	8
6-cysteine proteins	7
lccl	6
Thrombospondin-Related Anonymous Protein (*trap*)	6
pir	5 (includes 1 pseudogene)
Serine repeat antigen (SERA)	5 (1 pseudogene)
early transcribed membrane protein (*etramp*)	4
exported protein 1 (*exp1*)	4 (includes 2 pseudogenes)
Tryptophan-rich antigen (*trap*)	4
Lysophospholipase	2-4
Phist domain-containing	2
fam-a	1

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The gene encoding Thrombospondin-Related Anonymous Protein (*trap*) is involved in infection of salivary glands and liver cells by *Plasmodium* sporozoites. It is found strictly in a single copy in all *Plasmodium* species sequenced to date. However, it is present in six copies in *Hepatocystis*, suggesting that *trap*-mediated aspect of sporozoite-host interactions may be more complex. None of these genes were highly expressed in blood stage transcriptomes, consistent with their known role in sporozoites. Exported protein 1 (e.g. PF3D7_1121600) is a single copy gene in all *Plasmodium* species. It encodes a Parasitophorous Vacuolar Membrane (PVM) protein and is important for host-parasite interactions in the liver (16). In *Hepatocystis* it is expanded to four copies.

Missing orthologues tend to be involved in erythrocytic schizogony

The genomes of *Plasmodium* spp. each contain large families of genes known or thought to be involved in host parasite interactions (17). These include, amongst others the *var*, *rif* and *stevor* genes in the *Laverania* subgenus, *SICAvar* genes in *P. knowlesi* and the *pir* genes across the genus. We find only four intact *pir* genes and a single *pir* pseudogene in *Hepatocystis*, compared to ~200-1000 in *Plasmodium* spp. infecting rodents. One was particularly highly expressed in the blood stages from most monkeys that were sampled (S8C Fig; HEP_00069900). All of these are most similar to the ancestral *pir* subfamily, present in single copy in *Vinckeia* species and in 19 copies in *P. vivax* P01 (Table 2; S9 Fig). The best described role for *pir* genes is in *Vinckeia* parasites, where they are involved in establishing chronic infections in mice (18). Given that asexual *Hepatocystis* are not thought to exist in the blood of monkeys or bats, there would be no need for this function of *pir* genes. However, in *Vinckeia*, *pir* genes are expressed in several other stages, including male gametocytes (19), which do feature in the *Hepatocystis* lifecycle. The function of the ancestral *pir* gene subfamily is unknown, although it is expressed at multiple stages of the lifecycle in *P. berghei* (20).

We wanted to determine, more generally, the types of genes that might have been lost in *HexPt* relative to *Plasmodium* spp. This is made difficult due to uncertainty in determining missingness in a draft genome. We overcame this problem using clusters of genes identified as having common expression patterns across the lifecycle of the close *Hepatocystis* relative *P. berghei* (20). Clusters 10 (late schizont expression; Fisher's Exact test odds ratio = 0.12, FDR = 0.0003) and 4 (mixed stages; Fisher's Exact test odds ratio = 1.95, FDR = 0.04) tended to contain orthologues shared by *P. berghei*, *P. ovale wallikeri* and *P. vivax*, but not *HexPt* (S3 Table; S10 Fig). When we considered an assembly of all HexPt RNA-seq reads, we found that the late schizont cluster was still significant (Fisher's Exact test odds ratio = 0.09, FDR = 0.00014), but cluster 4 was not (S3 Table). Eleven out of 25 orthologues

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

missing in the late schizont cluster (including two pseudogenes) encoded Reticulocyte Binding Proteins (RBPs). In fact, we could not identify any RBPs in the *Hepatocystis* genome or *Hepatocystis* RNA-seq assemblies. Also missing from this cluster was *Cdpk5,* a kinase that regulates parasite egress from red cells (21). The other principal gene families involved in erythrocyte binding and invasion by *Plasmodium* are the erythrocyte binding ligands (*eh*)/duffy-binding protein (*dbp*) and the merozoite surface protein (*msp*) families (22). These are largely conserved relative to *P. berghei*. Thus, orthologues missing relative to *Plasmodium* spp. tend to be involved in erythrocytic schizogony, the part of the life cycle also absent.

The most rapidly evolving genes are often involved in vector biology and control of gene expression

Our whole-genome phylogeny (Fig 2) suggested that some genes have changed extensively in *Hepatocystis* compared to *Plasmodium* spp. This might indicate functional changes important for the particular biology of *Hepatocystis*. We found previously that the ratio of synonymous mutations to synonymous sites (dS) saturates between *Plasmodium* clades (Böhme et al., 2018) and therefore considered the ratio of non-synonymous mutations (dN) rather than the more commonly used dN/dS. We first looked for enrichment of conserved protein domain families in genes with the highest 5% of dN values. There was an enrichment for the AP2 domain (Pfam:PF00847.20; Fisher test with BH correction; p-value = 0.0042). Most *Plasmodium* species possess 27 ApiAP2 transcription factors containing this domain, which are thought to be the key players in control of gene expression and parasite development across the life cycle. AP2-G plays an important role in exiting the cycle of schizogony and commitment to gametocytogenesis in *Plasmodium* spp. (23,24), whereas, as demonstrated, *HexPt* lacks erythrocytic schizogony. *Hepatocystis* spp. also form much larger cysts in the liver (giving the genus its name) and develop in different tissues within a different insect vector compared to *Plasmodium* spp. Our *Hepatocystis* assembly contained orthologues of all 27 ApiAP2 genes present in *P. falciparum* (Table 2). This suggests that life cycle differences between *Plasmodium* and *Hepatocystis* spp. are not reflected in gain or loss of these key transcription factors. However, the relatively high rate of non-synonymous mutations suggests there may have been significant adjustment in how these transcription factors act. To determine parts of the life cycle that were enriched for the most rapidly evolving genes, we looked at whether particular gene expression clusters from the Malaria Cell Atlas (19,20) were enriched for genes with high dN (top 5% of values; Table 3; S11 Fig; S4 Table). We found that three clusters (2, 4 and 6) had fewer genes with high dN than expected by chance (Fisher's exact test with Holm multiple hypothesis testing correction, p-value < 0.05) and that these contained genes expressed across much of the life cycle, especially

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

growth phases and female gametocytes. These clusters also tended to contain essential genes that might be expected to be highly conserved in *Hepatocystis*. Although there was not a significant trend for gametocyte-associated genes having higher than average dN, the top 5% of genes ranked by dN contained several putative gametocyte genes (Table 3). Two of these encode putative 6-cysteine proteins P47 and P38, the first required for female gamete fertility (25). Additionally, Merozoite TRAP-like protein (MTRAP), essential for *Plasmodium* gamete egress from erythrocytes (26) and two genes (HEP_00254800 and HEP_00195400) with orthologues involved in osmiophilic body formation (27,28) had high dN values. Overall, clusters involved in ookinete (15) and general mosquito stages (16) had significantly higher values than other clusters (Kolmogorov-Smirnov test: Cluster 15 vs all other clusters: D = 0.42, p-value = 1.08e-05. Cluster 16 vs all other clusters: D = 0.52, p-value = 4.68e-12). This is also reflected in the *Hepatocystis* genes with the highest dN values, which include oocyst rupture protein 2 (*orp2*; dN = 1.08), *ap2-o*, *ap2-sp2*, secreted ookinete protein (*psop7*) and osmiophilic body protein (*g377*). These genes provide clues about changes in the parasite that might relate to its adaptation to transmission by biting midges, rather than mosquitoes.

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 3. Top 15 genes with functional annotations ranked by *Hepatocystis* dN in comparison of *Hepatocystis*, *P. berghei* ANKA and *P. ovale curtisi*. Genes with completely unknown function and genes with very little information on their possible functions have been left out from this table. The rank column indicates the *Hepatocystis* dN rank of each gene in the complete table (with 4009 genes) that includes genes with unknown function (S4 Table)

Gene id	*Hepato-cystis* dN	*P. berghei* dN	*P. ovale* dN	Annotations	Rank	Putative function
HEP_00146800 PBANKA_1303400 PocGH01_12075300	1.08	0.21	0.42	Oocyst rupture protein 2 (ORP2)	3	Sporozoite egress from the oocyst (29)
HEP_00446500 PBANKA_1003000 PocGH01_03012800	0.71	0.19	0.34	liver specific protein 2 (LISP2)	19	Involved in liver stage development (30)
HEP_00295100 PBANKA_0905900 PocGH01_09049300	0.71	0.31	0.63	AP2-O	20	Essential for morphogenesis in ookinete stage in *Plasmodium* (31)
HEP_00035800 PBANKA_1107600 PocGH01_10033500	0.69	0.27	0.16	6-cysteine protein (p38)	24	*P. berghei* p38 is expressed in gametocytes and in asexual blood stages (25)
HEP_00213800 PBANKA_1001800 PocGH01_03011500	0.67	0.45	0.63	AP2 domain transcription factor AP2-SP2	28	Required for sporozoite production in *Plasmodium (32)*
HEP_00337100 PBANKA_0112100 PocGH01_11043100	0.62	0.36	0.42	AP2 domain transcription factor ApiAP2	40	Involved in blood stage replication (32,33)
HEP_00456700 PBANKA_0512800 PocGH01_06021700	0.62	0.61	0.38	Merozoite TRAP-like protein (MTRAP)	42	Essential for gamete egress from erythrocytes (26)
HEP_00304800 PBANKA_1353400 PocGH01_12023100	0.62	0.32	0.26	Secreted ookinete protein (PSOP7)	43	Secreted ookinete proteins are necessary for invasion of the mosquito midgut (34)
HEP_00254800 PBANKA_1449000 PocGH01_14060000	0.59	0.26	0.26	Microgamete surface protein MiGS	54	Plays a critical role in male gametocyte osmiophilic body formation and exflagellation (27)
HEP_00115700 PBANKA_0304400 PocGH01_04023000	0.55	0.41	0.15	Merozoite surface protein 4 (MSP4)	62	Merozoite surface proteins are involved in red blood cell invasion (35)
HEP_00166600 PBANKA_0301000 PocGH01_04026800	0.54	0.36	0.2	Repetitive organellar protein (ROPE)	64	Localized to the apical end of merozoites, possibly involved in red blood cell invasion (36)

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

HEP_00195400 PBANKA_1463000 PocGH01_14074600	0.54	0.24	0.26	Osmiophilic body protein (G377)	67	Female-specific protein, affects the size of the osmiophilic body and female gamete egress efficiency (28)
HEP_00130400 PBANKA_1358000 PocGH01_12018000	0.53	0.17	0.11	Thioredoxin 2 (TRX2)	75	Part of a protein complex in parasitophorous vacuolar membrane, required for pathogenic protein secretion into host (37), important for maintaining normal blood-stage growth (38)
HEP_00391300 PBANKA_0313400 PocGH01_04013500	0.52	0.49	0.27	Autophagy-related protein 11 (ATG11)	76	Predicted to be involved in cargo selection in selective autophagy (39)
HEP_00155500 PBANKA_1302300 PocGH01_12076400	0.52	0.18	0.17	Metacaspase-2	80	Protease with caspase-like activity (40)

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Discussion

We have assembled and annotated a draft quality genome sequence of *Hepatocystis* sp. ex *Piliocolobus tephrosceles* (*HexPt*). Our nuclear and apicoplast genome phylogenies confirm the recently proposed phylogenetic placement of this genus as an outgroup to the rodent-infecting *Vinckei* subgenus of *Plasmodium (2)*. However, a distinct branching pattern and low bootstrap support for many nodes in our mitochondrial genome phylogeny highlights why some previous analyses have come to different conclusions about the placement of *Hepatocystis*. Thus, the use of mitochondrial genes to infer phylogenetic relationships between species within the *Hemosporidia* should be approached with caution. We found a long branch leading to *HexPt*, suggesting a relatively deep split from the rodent-infecting species. In addition, we showed robustly that *HexPt* groups, as expected, with *Hepatocystis epomophori*, which infects bats. This finding supports the polyphyly of the *Plasmodium* parasites infecting apes, monkeys and rodents but the monophyly of *Hepatocystis* itself. A close relative of rodent-infecting *P. berghei* (*P. cyclopsi*) has been found in bats (10) and thus the *Hepatocystis*/*Vinckeia* group represents a relatively labile group with respect to host preference. Indeed, the possibility of cross-species transmission of *HexPt* was reported previously (6).

The paraphyly of *Plasmodium* with respect to *Hepatocystis* exists because *Hepatocystis* spp. lack a defining characteristic of the *Plasmodium* genus, namely erythrocytic schizogony - asexual development in the blood. Thus, the very part of the *Plasmodium* life cycle which causes the symptoms of malaria is thought to be absent in *Hepatocystis*. While multiple lines of enquiry have failed to identify these forms (Garnham, 1966) there has remained some doubt, with reports of cells with schizont-like morphology in the blood for some species (13,41). We were able to take advantage of bulk RNA-seq data collected from blood samples of a number of monkeys, all apparently infected with very closely related *Hepatocystis* parasites. By comparing to single-cell RNA-seq data from known cell types, we found no evidence for schizont stages in the blood. The apparent lack of schizonts could be due to sequestration of these stages away from the bloodstream. This is seen in some *Plasmodium* species, so we cannot rule out schizogony away from peripheral circulation. However, in its current state, this lack of evidence supports the idea that erythrocytic schizogony has been lost three times: in *Hepatocystis*, *Polychromophilus* and *Nycteria* (2). Ancestrally, gametocytogenesis must have been the default developmental pathway, being required for transmission. However, in *Plasmodium*, it seems that erythrocytic schizogony became the default developmental pathway with epigenetic control of the ApiAP2-G transcription factor, required for development into sexual stages (23,24). Perhaps the simplest possible explanation for a

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

loss of erythrocytic schizogony would be that ApiAP2-G is no longer under control, but is constitutively expressed in parasites leaving the liver. In line with this we find ApiAP2-G present and highly expressed in blood stage *HexPt*. Furthermore, all *Plasmodium* ApiAP2 transcription factors are conserved in *HexPt*, indicating that changes in its life cycle are not associated with their loss or gain.

The lack of erythrocytic schizogony is supported by a tendency for orthologues missing relative to *Plasmodium* spp. to be those expressed in blood schizonts. The most noticeable example being the complete absence of the Reticulocyte Binding Protein (RBP) family, found across all *Plasmodium* spp. examined so far, including those which infect birds (42). RBP proteins are known to function as essential red blood cell invasion ligands in *Plasmodium falciparum* (43) and multiple copies are thought to provide alternative invasion pathways (22). However, previous transcriptomic data have suggested that, while *rbp* genes in *P. berghei* are highly expressed in schizonts, they are less abundant or have a distinct repertoire in liver stages (20,44,45). This implies that RBPs are less important, or at least that distinct invasion pathways are used by first generation merozoites. Also missing were *cdpk5* (involved in schizont egress) and *msp9* (an invasion related gene enriched in blood vs. liver schizonts in Caldelari et al., 2019). Taken together, these results underscore the increasing realisation that first generation merozoites have distinct properties from merozoites that have developed in the blood and suggest the way in which first generation merozoites invade red blood cells may be distinct.

The genomes of *Plasmodium* spp. each contain large, rapidly evolving gene families that are known, or thought to be involved in host-parasite interactions, principally in asexual blood stages. The reason for their numbers may be due to a bet-hedging strategy, providing the diversity necessary for evading adaptive immune responses or dealing with unpredictable host variation. Although the *Hepatocystis* genome contains two novel multigene families, we identified only 10-15 copies of each. The largest gene families in the closely related rodent-infecting *Plasmodium* species (*pir* and *fam-a*) are present here only as their ancestral orthologues, also conserved in the monkey-infecting *Plasmodium* species. We should be cautious in noting a lack of expansion in such families in *Hepatocystis*, as previous draft *Plasmodium* genome sequences have been shown to under-represent these genes. However, it is perhaps not surprising that *pir* genes are poorly represented. They are thought to have a role in the maintenance of chronic infections mediated by asexual stages in the blood (18) and *Hepatocystis* infection does not involve this stage of development. Given that

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Hepatocystis can sustain long chronic infections (1), presumably from the liver, this parasite may help us to better understand how *Plasmodium* survives in the liver.

A striking feature of the *Hepatocystis* life cycle is its vector - a midge rather than a mosquito. We found evidence of rapid evolution amongst orthologues of *Plasmodium* genes involved in mosquito stages of development, suggesting that adaptation to a new insect vector was a major evolutionary force. The rapidly evolving genes provide clues to the players in parasite-vector interactions and may provide avenues for the development of interventions to prevent transmission of the malaria parasite.

We expect that high-quality genome sequences of additional *Plasmodium* relatives such as *Nycteria*, *Haemoproteus,* and *Polychromophilus* will provide more insights into the evolution and molecular biology of one of humanity's greatest enemies - the malaria parasite.

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Methods

Sample collection and data generation

The sequence data used in this study were part of different project originally designed to generate a reference genome for the red colobus monkey (genus *Piliocolobus*). Biomaterials used were from wild Ugandan (or Ashy) red colobus monkey (*Piliocolobus tephrosceles*) individuals from Kibale National Park, Uganda. These animals reside in a habituated group that has been a focus of long-term studies in health, ecology, and disease (12,46,47). Red colobus individuals were immobilized in the field as previously described (48). Whole blood was collected using a modified PreAnalytiX PAXgene Blood RNA System protocol as described in Simons et al. (12). Additionally, whole blood was collected into BD Vacutainer Plasma Preparation Tubes, blood plasma and cells were separated via centrifugation, and both were subsequently aliquoted into cryovials and stored in liquid nitrogen. Samples were transported to the United States in an IATA-approved liquid nitrogen dry shipper and then transferred to −80 °C for storage until further processing.

Methods for DNA extraction, library preparation, and whole genome sequencing are described in Simons (49). Briefly, high molecular weight DNA was extracted from the blood cells of one red colobus monkey individual and size selected for fragments larger than 50,000 base pairs. A 10X Genomics Chromium System library preparation was performed and subsequently sequenced on two lanes of a 150 bp paired-end Illumina HiSeqX run as well as two lanes of a 150 bp paired-end Illumina HiSeq 4000 run.

Methods for RNA extraction and library preparation are described in Simons et al. (2019). Briefly, RNA was extracted from 29 red colobus individuals using a modified protocol for the PreAnalytiX PAXgene Blood RNA Kit protocol. Total RNA extracts were concentrated, depleted of alpha and beta globin mRNA, and assessed for integrity (RIN mean: 8.1, range: 6.6–9.2). Sequencing libraries were prepared using the KAPA Biosystems Stranded mRNA-seq Kit and sequenced on four partial lanes of a 150 bp paired-end Illumina HiSeq 4000 run. These data were uploaded to NCBI as part of BioProject PRJNA413051.

Separation of *Hepatocystis* and *Piliocolobus* scaffolds

The *Piliocolobus tephrosceles* genome assembly (ASM277652v1) was downloaded from the NCBI database. Scaffolds were first sorted by their GC% and Diamond 0.9.22 (50) BLASTX hits against a database of representative apicomplexan and Old World monkey proteomes. The sorting was improved by examining mapping scores of the scaffolds mapped to *Plasmodium* species and *Macaca*

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

mulatta genomes (Mmul_8.0.1, GenBank assembly accession GCA_000772875.3) using Minimap2 2.12 (51). The separation of scaffolds was further verified and refined by running NCBI BLAST of 960 bp fragments of all scaffolds against the NCBI nt database (Jul 18 2017 version) (52). To predict genes in the apicomplexan scaffolds, Companion automatic annotation software (7) was run with these scaffolds as input and the *P. vivax* P01 genome as the reference.

Identification of Hepatocystis sequences in Piliocolobus RNA-seq data

Illumina HiSeq 4000 RNA-seq reads from the study PRJNA413051 were downloaded from the European Nucleotide Archive. In order to find out if the RNA-seq data contained apicomplexan sequences, mapping of these reads to apicomplexan scaffolds from *Piliocolobus tephrosceles* genome assembly (ASM277652v1) was done using HISAT2 2.1.0 (53).

Hepatocystis genome assembly

Filtering of reads for assembly

Minimap2 (51) and Kraken 2.0.8-beta (54) were used to identify the best matching species for each 10x Chromium genomic DNA read (from Illumina HiSeq X and HiSeq4000 platforms). Our Kraken database contained 17 Old World monkey genomes and 19 *Plasmodium* genomes downloaded from NCBI FTP in June 2018 (52) . The Kraken database also included the contigs of the *P. tephrosceles* assembly ASM277652v1, separated into *P. tephrosceles* and *Hepatocystis* sp. *Plasmodium malariae* UG01 (from PlasmoDB (55) version 39) and *Macaca mulatta* (Mmul_8.0.1) assemblies were used as reference genomes for the assignment of reads based on Minimap2 mapping scores. Reads that were unambiguously identified as monkey sequences using Kraken and Minimap2 were excluded from subsequent assemblies. The Supernova assembler manual (56) warns against exceeding 56x coverage in assemblies. Reads selected for Supernova assemblies were therefore divided into 34 batches, with ~10 million reads in each batch. Reads were ordered by their barcodes so that those with the same barcode would preferentially occur in the same batch.

Supernova and SPAdes assemblies

We generated 34 assemblies with Supernova v2.1.1 with default settings. In addition, two SPAdes v3.11.0 (57) assemblies with default settings were generated with *Hepatocystis* reads: one with HiSeq X reads and another with HiSeq 4000 reads. Chromium barcodes were removed from the reads before the SPAdes assemblies.

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Deriving the mitochondrial sequence

Hepatocystis Supernova and SPAdes assembly contigs were mapped to the *P. malariae* UG01 genome from PlasmoDB version 40 with Minimap2. The sequences of contigs that mapped to the *P. malariae* mitochondrion were extracted using SAMTools 0.1.19-44428cd (58) and BEDtools v2.17.0 (59). The contigs were oriented and then aligned using Clustal Omega 1.2.4 (60). Consensus sequence of aligned contigs was derived using Jalview 2.10.4b1 (61). The consensus sequence was circularised with Circlator minimus2 (62).

Canu assembly

Scaffolds from the Supernova assemblies were broken into contigs. All contigs from the Supernova and SPAdes assemblies were pooled and used as the input for Canu assembler 1.6 (63) in place of long reads. Canu assembly was done without read correction and trimming stages. The settings for Canu were as follows: -assemble genomeSize=23000k minReadLength=300 minOverlapLength=250 corMaxEvidenceErate=0.15 correctedErrorRate=0.16 stopOnReadQuality=false -nanopore-raw.

Processing of Canu unassembled sequences file

Selected contigs from the Canu unassembled sequences output file (*.unassembled.fasta) were recovered and pooled with assembled contigs (*.contigs.fasta). The first step in the filtering of the contigs of the unassembled sequences file was to exclude contigs that had a BLAST match in the assembled sequences output file (with E value cutoff 1e-10). Next, contigs where low complexity sequence content exceeded 50% (detected using Dustmasker 1.0.0 (64)) were removed. Contigs with GC content higher than 50% were also removed. Diamond BLASTX (against a database of *Macaca mulatta*, *P. malariae* UG01, *P. ovale wallikeri*, *P. falciparum* 3D7 and *P. vivax* P0 proteomes) and BLAST (using the nt database from Jul 18 2017 and nr database from Jul 19 2017) were then used to exclude all contigs where the top hits were not an apicomplexan species. In total, 0.34% of contigs from the unassembled sequences file were selected to be included in the assembly.

Deduplication of contigs

Initial deduplication of contigs was done using BBTools dedupe (65) (Nov 20, 2017 version) and GAP5 v1.2.14-r3753M (66) autojoin. In addition, BUSCO 3.0.1 (67) was used to detect duplicated core genes with the protists dataset. Two contigs flagged by BUSCO as containing duplicated genes were removed. All vs all BLAST of contigs (with E-value cutoff 1e-20, minimum overlap length 100 bp, minimum identity 85%) was used to find possible cases of remaining duplicated contigs. Contigs yielding BLAST hits were aligned with MAFFT v7.205 (68) and the alignments were manually inspected. Contained contigs were deleted and contigs that had unique overlaps with high identity were merged into consensus sequences using Jalview.

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Removal of contaminants after Canu assembly

All Canu assembly contigs were checked with Diamond against a database of *Macaca mulatta*, *P. malariae* UG01, *P. ovale wallikeri*, *P. falciparum* 3D7 and *P. vivax* P01 proteomes. The Diamond search did not detect any contaminants. Contigs not identified by Diamond were checked with BLAST against the nt database (Jul 18 2017 version). Contigs where the top BLAST hit was a human or monkey sequence were removed from the assembly.

A subset of contigs in the assembly was observed to consist of short sequences with low complexity, high GC% and low frequency of stop codons. These contigs did not match any sequences by BLAST search against nt and nr databases (with E-value cutoff 1e-10). Due to their difference from the rest of the contigs in the assembly, it was assumed that these contigs were contaminants rather than *Hepatocystis* sequences. In order to programmatically find these contigs, GC%, tandem repeats percentage, percentage of low complexity content and frequency of stop codons were recorded for all contigs in the assembly. Tandem Repeats Finder 4.04 (69) was used to assess tandem repeats percentage and Dustmasker 1.0.0 (64) was used to find low complexity sequence content. PCA and k-means clustering (using R version 3.5.1) showed that the assembly contigs separated into two groups based on these parameters. The group of contigs with low complexity (189 contigs) was removed from the assembly.

Scaffolding and polishing of Canu assembly contigs

Before scaffolding, contigs were filtered by size to remove sequences shorter than 200 bp. *Hepatocystis* RNA-seq reads were extracted from RNA-seq sample SAMN07757854 using Kraken 2. Canu assembly contigs were scaffolded with these reads using P_RNA_scaffolder (70). To correct scaffolding errors, the scaffolds were processed with REAPR 1.0.18 (71) using 197819014 unbarcoded *Hepatocystis* DNA read pairs. REAPR was run with the *perfectmap* option and -break b=1. Next, the assembly was scaffolded using Scaff10x (https://github.com/wtsi-hpag/Scaff10X) version 3.1, run for 4 iterations with the following settings: -matrix 4000 -edge 1000 -block 10000 -longread 0 -link 3 -reads 5. 197,819,014 *Hepatocystis* DNA read pairs were used for Scaff10x scaffolding. After this, P_RNA_scaffolder was run again as above. This was followed by running Tigmint 1.1.2 (72) with 419,652,376 *Hepatocystis* read pairs to correct misassemblies. fill_gaps_with_gapfiller (https://github.com/sanger-pathogens/assembly_improvement/blob/master/bin/fill_gaps_with_gapfiller) was used to fill gaps in scaffolds, using 197819014 unbarcoded *Hepatocystis* DNA read pairs. After this, ICORN v0.97 (73) was run for 5 iterations with 4608740 *Hepatocystis* read pairs. This was followed by polishing the assembly with Pilon 1.19 (74) using 21,794,613 *Hepatocystis* read pairs. Assembly completeness was assessed with BUSCO v3.0.1 (75) (run with the protists lineage and with -m geno -sp pfalciparum --

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

long flags) and CEGMA v2.5 (76). *P. berghei* ANKA, *P. ovale curtisi* and *P. falciparum* 3D7 genomes from PlasmoDB release 45 were also assessed with BUSCO and CEGMA with the same settings in order to compare the *Hepatocystis* assembly with *Plasmodium* assemblies.

Curation and Annotation

The assembly was annotated using Companion (7). The alignment of reference proteins to target sequence was enabled in the Companion run but all other parameters were left as default. A GTF file derived from mapping of *Hepatocystis* RNA-seq reads of three biological samples (SAMN07757854, SAMN07757861 and SAMN07757872) to the assembly was used as transcript evidence for Companion. To produce the GTF file, the RNA-seq reads were mapped to the assembly using 2-pass mapping with STAR RNA-seq aligner (77) (as described in the "Variant calling of RNA-seq samples" section) and the mapped reads were processed with Cufflinks (78). All *Plasmodium* genomes available in the web version of Companion were tested as the reference genome for annotating the *Hepatocystis* genome, in order to find out which reference genome yields the highest gene density. For the final Companion run the *P. falciparum* 3D7 reference genome (version from June 2015) was used. The Companion output was manually curated using Artemis (79) and ACT (80) version 18.0.2. Manual curation was carried out to correct the overprediction of coding sequences, add missing genes and correct exon-intron boundaries. Altogether 680 gene models were corrected, 546 genes added and 221 genes deleted. RNA-seq data was used as supporting evidence. Non-coding RNAs were predicted with Rfam (81).

All genes were analyzed for the presence of a PEXEL-motif using the updated HMM algorithm ExportPred v2.0 (82). Distant homology to *hep1* and *hep2* gene families was sought by using the HHblits webserver with default options (14).

The reference genomes used to produce statistics on features of *Plasmodium* genomes in Fig 2 and Table 1 were as follows: *P. relictum* SGS1, *P. gallinaceum* 8A (42), *P. malariae* UG01, *P. ovale wallikeri*, *P. ovale curtisi* GH01 (11), *P knowlesi* H (83), *P. vivax* P01 (84), *P. cynomolgi* M (85), *P. chabaudi* AS (18), *P. berghei* ANKA (86), *P. reichenowi* CDC (87), *P. falciparum* 3D7 (88).

For S1 Table, transmembrane domains of proteins were predicted using TMHMM 2.0 (89). Conserved domains were detected in proteins using HMMER i1.1rc3 (http://hmmer.org/) and Pfam-A database release 28.0 (90), with E-value cutoff 1e-5. Besides predicting exported proteins with ExportPred 2 (82), matches to PEXEL consensus sequence (RxLxE/Q/D) were counted in protein sequences using string search in Python. Signal peptides were detected using SignalP-5 (91).

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Phylogenetic trees

Haemosporidian sequences were downloaded from NCBI FTP and PlasmoDB (release 43). The phylogenetic tree of cytochrome B and the tree that included 11 *Hepatocystis epomophori* genes were based on DNA alignments. The cytochrome B tree also included cytochrome B sequences from *de novo* assemblies of *Hepatocystis* RNA-seq reads derived from *Piliocolobus tephrosceles* blood. The trees of mitochondrial, apicoplast and nuclear proteomes were based on protein alignments. For apicoplast proteome and nuclear proteome trees, orthologous proteins were identified using OrthoMCL 1.4 (92). All vs all BLAST for OrthoMCL was done using blastall 2.2.25 with E-value cutoff 1e-5. OrthoMCL was run with mode 3. Proteins with single copy orthologs across all the selected species were used for the protein phylogenetic trees. Sequences were aligned with MAFFT 7.205 (93) (with --auto flag) and the alignments were processed using Gblocks 0.91b (94) with default settings. Individual Gblocks-processed alignments were concatenated into one alignment. The phylogenetic trees were generated using IQ-TREE multicore version 1.6.5 (95) with default settings and plotted using FigTree 1.4.4 (https://github.com/rambaut/figtree/releases). Inkscape (https://inkscape.org) version 0.92 was used to edit text labels of the phylogenetic trees generated with FigTree.

Clustering of *pir* proteins into subfamilies

Sequences of *Plasmodium pir* family proteins (including *bir*, *cyir*, *kir*, *vir* and *yir* proteins) were downloaded from PlasmoDB (55) (release 39). The sequences were clustered using MCL (96), following the procedures described in the section "Clustering similarity graphs encoded in BLAST results" in clmprotocols (https://micans.org/mcl/man/clmprotocols.html). The BLAST E-value cutoff used for clustering was 0.01 and the MCL inflation value was 2. The *pir* protein counts per subfamily in each species were plotted as a heatmap using the heatmap.2 function in gplots package version 3.0.1.1 in R version 3.5.1.

RNA-seq mapping and assembly

To separate *Hepatocystis* reads from *Piliocolobus* reads, RNA-seq data from the ENA (study PRJNA413051) were mapped to a FASTA file containing genome assemblies of *Hepatocystis* and *M. mulatta* (NCBI assembly Mmul_8.0.1), using HISAT2 version 2.1.0 (53), with "--rna-strandness RF". BED files were generated from the mapped reads using BEDTools 2.17.0 (59). Reads from each

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

technical replicate were merged, resulting in a single set of read counts for each individual monkey. The BED files were filtered to remove multimapping reads and reads with mapping quality score lower than 10. Names of reads that specifically mapped to the *Hepatocystis* assembly were extracted from the BED file. SeqTK 1.0-r31 (https://github.com/lh3/seqtk) was used to isolate *Hepatocystis* FASTQ reads based on the list of reads from the previous step. The *Hepatocystis* reads were then mapped to the *Hepatocystis* genome assembly using HISAT2 2.1.0 with "--rna-strandness RF" flag. The SAM files with mapped reads were converted to sorted BAM files with SamTools 0.1.19-44428cd (58). The EMBL file of *Hepatocystis* genome annotations was converted to GFF format using Artemis 18.0.1 (79). Htseq-count 0.7.1 (97) was used to count mapped reads per gene in the GFF file with "-t mRNA -a 0 -s reverse". Htseq-count files of individual RNA-seq runs were merged into a single file.

In order to extract *Hepatocystis* cytochrome b sequences of each RNA-seq sample, *Hepatocystis* RNA-seq reads of each sample were isolated from *Piliocolobus* reads as described above and then assembled with the SPAdes assembler v3.11.0 (98) with the "--rna" flag. *Hepatocystis* cytochrome b contigs were identified in each of the 29 RNA-seq assemblies using BLAST against *Hepatocystis* cytochrome b from the DNA assembly (E-value cutoff 1e-10).

In addition to assemblies of individual RNA-seq samples, an assembly of all RNA-seq samples pooled was done. The reads for this assembly were sorted by competitive mapping to *P. ovale curtisi* GH01 (from PlasmoDB release 45) and *Macaca mulatta* (Mmul_8.0.1, GenBank assembly accession GCA_000772875.3) genomes with minimap2 (with the "-ax sr" flag). Reads mapping to the *Macaca mulatta* genome with minimum mapping score 20 were removed and the rest of the reads were assembled with the SPAdes assembler v3.13.1 (98) with the "--rna" flag. *Hepatocystis* contigs were identified by comparison of sequences with *Plasmodium* and *Macaca mulatta* reference genomes using Diamond, minimap2 and BLAST, similarly to what is described in the section "Separation of *Hepatocystis* and *Piliocolobus* scaffolds". Further decontamination was done using Diamond and BLAST searches against 19747 sequences from *Ascomycota* and 165860 bacterial sequences downloaded from UniProt (release 2019_10) (99) and 3 *Babesia* proteomes from PiroplasmaDB (release 46) (100). Selected contigs were also checked with BLAST against the NCBI nt database. The assembly was deduplicated using BBTools dedupe (Nov 20, 2017 version) and GAP5 v1.2.14-r3753M. Assembly completeness was assessed using CEGMA 2.5. In order to reduce the number of contigs so that they could be used as input for Companion, the assembly was scaffolded with RaGOO Version 1.1 (101), using the *Hepatocystis* DNA assembly as the reference. The assembly was then processed by the Companion annotation software (Glasgow server, November 2019 version, with *P. falciparum* 3D7 reference genome, with protein evidence enabled and the rest of the settings left as default). In

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

order to detect proteins missed by Companion, EMBOSS Transeq (version 6.3.1) was used to translate the transcriptome assembly in all 6 reading frames. The output of Transeq was then filtered to keep sequences between stop codons with minimum length of 240 amino acids. Protein BLAST with E-value cutoff 1e-20 was used to detect sequences in Transeq output that were not present in the proteins annotated by Companion. These selected Transeq output sequences were checked for contaminants with BLAST similarly to what was described before. The sequences that passed the contaminant check were combined with the set of *Hepatocystis* RNA-seq assembly proteins that were detected by Companion. OrthoMCL was run with proteins from *Hepatocystis* RNA-seq assembly (Companion and selected Transeq sequences combined), *Hepatocystis* DNA assembly proteins, 20 *Plasmodium* proteomes from PlasmoDB release 43 and *P. ovale wallikeri* proteome (GenBank GCA_900090025.2). The settings for OrthoMCL were as described in the "Phylogenetic trees" section.

dN analysis

P. berghei ANKA and *P. ovale curtisi* protein and transcript sequences were retrieved from PlasmoDB (55) (release 45). One-to-one orthologs between *Hepatocystis*, *P. berghei* ANKA and *P. ovale curtisi* were identified using OrthoMCL (92)and a Newick tree of the three species was generated with IQ-TREE (95). The settings for OrthoMCL and IQ-TREE were as described in the "Phylogenetic trees" section. Transcripts of one-to-one orthologs were aligned using command line version of TranslatorX (102) with "-p F -t T" flags, so that each alignment file contained sequences from three species. Gaps were removed from alignments while retaining the correct reading frame. Alignment regions where the nucleotide sequence surrounded by gaps was shorter than 42 bp were also removed. In addition, the script truncated alignments at the last whole codon if a sequence ended with a partial codon due to a contig break. The alignments and the Newick tree of the 3 species were then used as input for codeml (103) in order to determine the dN and dN/dS of each alignment. The codeml settings that differed from default settings were: seqtype = 1, model = 1. *P. berghei* RNA-seq cluster numbers from Malaria Cell Atlas (20) were assigned to each alignment based on the *P. berghei* gene in the alignment. Transcriptomics-based gametocyte specificity scores of *Plasmodium* genes were taken from an existing study on this topic (104) (transcripts table S2 of "Transcriptomics_all_studies" tab). The *P. falciparum* genes in the gametocyte specificity scores table were matched with equivalent *Hepatocystis* genes using OrthoMCL (run with the same settings as when used for phylogenetic trees). Statistical tests with the dN results (Kolmogorov Smirnov test, Fisher test and Spearman correlation) were performed using the *stats* library in R.

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Variant calling of RNA-seq samples

SNPs and indels were called in *Hepatocystis* RNA-seq reads that had been separated from *Piliocolobus* reads as described above. Four technical replicates of each RNA-seq sample were pooled. Variant calling followed the "Calling variants in RNAseq" workflow in GATK (105) user guide (https://gatkforums.broadinstitute.org/gatk/discussion/3891/calling-variants-in-rnaseq). First, the reads were mapped to the reference genome using 2-pass mapping with the STAR RNA-seq aligner (77) version 2.5.3a. 2-pass mapping consisted of indexing the genome with genomeGenerate command, aligning the reads with the genome, generating a new index based on splice junction information contained in the output of the first pass and then producing a final alignment using the new index. GATK (105) version 4.0.3.0 was used for the next steps. The mapped reads were processed with GATK MarkDuplicates and SplitNCigarReads commands. GATK HaplotypeCaller was then run with the following settings: --dont-use-soft-clipped-bases --emit-ref-confidence GVCF --sample-ploidy 1 --standard-min-confidence-threshold-for-calling 20.0. Joint genotyping of the samples was then done using GATK CombineGVCFs and GenotypeGVCFs commands. This was followed by running VariantFiltration with these settings: -window 35 -cluster 3 --filter-name FS -filter 'FS > 30.0' --filter-name QD -filter 'QD < 2.0'. SNPs were separated from indels using GATK SelectVariants. Samples SAMN07757853, SAMN07757863 and SAMN07757870 were excluded from further analysis due to their low expression of *Hepatocystis* genes. The average filtered SNP counts per 10 kb of reference genome for each sample were calculated as the number of filtered SNPs divided by (genome size in kb * 10).

RNA-seq deconvolution

Deconvolution of a bulk RNA-seq transcriptome sequence aims to determine the relative proportions of different cell types in the original sample. This requires a reference dataset of transcriptomes from “pure” cell types. To create this, we used single-cell *P. berghei* transcriptome sequences from the Malaria Cell Atlas (20). For each cell type, single-cell transcriptome sequences were combined by summing read counts per gene to generate a set of pseudobulk transcriptome sequences (see our GitHub repository). The aim of summing across cells is to reduce the number of dropouts which are common in individual single-cell transcriptome sequences. Bulk *Hepatocystis* RNA-seq transcriptome sequences, mapped and counted as above, were summed across replicates and filtered to exclude those with fewer than 100,000 reads. *Hepatocystis* and *P. berghei* pseudobulk read counts were converted to Counts Per Million (CPM) and *Hepatocystis* gene ids were

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

converted to those of *P. berghei* one-to-one orthologues. Genes without one-to-one orthologues (defined by orthoMCL analysis) were excluded. CIBERSORT v1.06 (106) was used to deconvolute the *Hepatocystis* transcriptomes with the MCA pseudobulk as the signature matrix file. To test the accuracy of this deconvolution process we generated mixtures of the pseudobulk resulting in e.g. equal representation of read counts from male gametocyte, female gametocyte, ring, trophozoite and schizont pseudobulk transcriptomes (see our GitHub repository). We also deconvoluted bulk RNA-seq transcriptomes from Otto et al. (86) processed as in Reid et al. (19).

Enrichment of missing genes in Malaria Cell Atlas gene clusters

We wanted to determine whether there were functional patterns common to orthologues missing from the *Hepatocystis* genome relative to *Plasmodium* species. To do this we looked for orthologous groups (orthoMCL as above) containing genes from *P. berghei*, *P. ovale wallikeri* and *P. vivax* P01, but not *Hepatocystis*. Genes from *P. berghei* have previously been assigned to 20 clusters based on their gene expression patterns across the whole life cycle (20). We looked to see whether missing orthologues tended to fall into particular clusters more often than expected by chance (see our GitHub repository). We used Fisher's exact test with Benjamini-Hochberg correction to control the false discovery rate. We reported clusters with FDR >= 0.05.

Data availability

The raw sequence data for *Hepatocystis* can be retrieved from the European Nucleotide Archive, project accession number ERP115621 and sample accession number ERS3649919. The assembly can be found under the study PRJEB32891. The individual accession numbers for the contigs are: CABPSV010000001-CABPSV010002439. Accession numbers for the apicoplast and the mitochondrion are LR699571-LR699572. Illumina HiSeq 4000 RNA-seq reads, containing a mix of *Piliocolobus tephrosceles* and *Hepatocystis sp.* sequences can be found in the European Nucleotide Archive under study accession PRJNA413051. Other data and code are available from our GitHub repository: https://github.com/adamjamesreid/hepatocystis-genome.

Acknowledgements

We would like to thank Alan Tracey for advice on genome assembly and J. Byaruhanga, P. Katurama, A. Mbabazi, A. Nyamwija, J. Rusoke, D. Hyeroba and G. Weny and the staff of Makerere University Biological Field Station for assistance in the field. Manoj Duraisingh provided helpful comments on the manuscript.

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Funding

This work was funded by National Institutes of Health (NIH), USA grant TW009237 as part of the joint NIH-NSF Ecology of Infectious Disease program and the UK Economic and Social Research Council (TLG, NT, CAC). National Science Foundation Grant BCS-1540459 (NT, NDS). The Wellcome Sanger Institute is funded by the Wellcome Trust (grant 206194/Z/17/Z) which supports EA, UB, MB, and AJR. AJR is also supported by funding from the UK Medical Research Council (MRC Programme grant #: MR/M003906/1). CIN is funded by a Wellcome Investigator Award (104792/Z/14/Z).

Author Contributions

Conceptualization AJR, TS

Data Curation EA, UB

Formal Analysis AJR, EA

Investigation AJR, EA, NDS, UB

Project Administration CAC, NT, TLG

Resources CAC, NT, TLG

Software AJR, EA

Supervision AJR, MB, NT

Visualization AJR, EA, UB

Writing – Original Draft Preparation AJR, EA, TS, UB

Writing – Review & Editing AJR, CAC, CIN, EA, MB, NDS, NT, TS, TLG, UB

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

1. Garnham PCC. Malaria parasites and other haemosporidia. Blackwell Scientific; 1966. 1114 p.

2. Galen SC, Borner J, Martinsen ES, Schaer J, Austin CC, West CJ, et al. The polyphyly of Plasmodium: comprehensive phylogenetic analyses of the malaria parasites (order Haemosporida) reveal widespread taxonomic conflict. R Soc Open Sci. 2018 May;5(5):171780.

3. Garnham PCC. The developmental cycle of Hepatocystes (Plasmodium) kochi in the monkey host. Trans R Soc Trop Med Hyg. 1948 Mar;41(5):601–16.

4. Perkins SL, Schaer J. A Modern Menagerie of Mammalian Malaria. Trends Parasitol. 2016 Oct;32(10):772–82.

5. Garnham PCC, Heisch RB, Minter DM, Others. The Vector of Hepatocystis (= Plasmodium) kocht; the Successful Conclusion of Observations in Many Parts of Tropical Africa. Trans R Soc Trop Med Hyg. 1961;55(6):497–502.

6. Thurber MI, Ghai RR, Hyeroba D, Weny G, Tumukunde A, Chapman CA, et al. Co-infection and cross-species transmission of divergent Hepatocystis lineages in a wild African primate community. Int J Parasitol. 2013 Jul;43(8):613–9.

7. Steinbiss S, Silva-Franco F, Brunk B, Foth B, Hertz-Fowler C, Berriman M, et al. Companion: a web server for annotation and analysis of parasite genomes. Nucleic Acids Res. 2016 Jul 8;44(W1):W29–34.

8. Boundenga L, Ngoubangoye B, Mombo IM, Tsoubmou TA, Renaud F, Rougeron V, et al. Extensive diversity of malaria parasites circulating in Central African bats and monkeys. Ecol Evol. 2018 Nov;8(21):10578–86.

9. Chang Q, Sun X, Wang J, Yin J, Song J, Peng S, et al. Identification of Hepatocystis species in a macaque monkey in northern Myanmar. Res Rep Trop Med. 2011 Nov 30;2:141–6.

10. Schaer J, Perkins SL, Decher J, Leendertz FH, Fahr J, Weber N, et al. High diversity of West African bat malaria parasites and a tight link with rodent Plasmodium taxa. Proc Natl Acad Sci U S A. 2013 Oct 22;110(43):17415–9.

11. Rutledge GG, Böhme U, Sanders M, Reid AJ, Cotton JA, Maiga-Ascofare O, et al. Plasmodium malariae and P. ovale genomes provide insights into malaria parasite evolution. Nature. 2017 Feb 2;542(7639):101–4.

12. Simons ND, Eick GN, Ruiz-Lopez MJ, Hyeroba D, Omeja PA, Weny G, et al. Genome-Wide Patterns of Gene Expression in a Wild Primate Indicate Species-Specific Mechanisms Associated with Tolerance to Natural Simian Immunodeficiency Virus Infection. Genome Biol Evol. 2019 Jun 1;11(6):1630–43.

13. Schaer J, Perkins SL, Ejotre I, Vodzak ME, Matuschewski K, Reeder DM. Epauletted fruit bats display exceptionally high infections with a Hepatocystis species complex in South Sudan. Sci Rep. 2017 Jul 31;7(1):6928.

14. Remmert M, Biegert A, Hauser A, Söding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods. 2011 Dec 25;9(2):173–5.

15. Looker O, Blanch AJ, Liu B, Nunez-Iglesias J, McMillan PJ, Tilley L, et al. The knob protein KAHRP assembles into a ring-shaped structure that underpins virulence complex assembly. PLoS Pathog. 2019 May;15(5):e1007761.

16. Sá E Cunha C, Nyboer B, Heiss K, Sanches-Vaz M, Fontinha D, Wiedtke E, et al. Plasmodium berghei EXP-1 interacts with host Apolipoprotein H during Plasmodium liver-stage development. Proc Natl Acad Sci U S A. 2017 Feb 14;114(7):E1138–47.

17. Reid AJ. Large, rapidly evolving gene families are at the forefront of host–parasite interactions in Apicomplexa. Parasitology. 2015 Feb;142(S1):S57–70.

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

18. Brugat T, Reid AJ, Lin J, Cunningham D, Tumwine I, Kushinga G, et al. Antibody-independent mechanisms regulate the establishment of chronic Plasmodium infection. Nat Microbiol. 2017 Feb 6;2:16276.

19. Reid AJ, Talman AM, Bennett HM, Gomes AR, Sanders MJ, Illingworth CJR, et al. Single-cell RNA-seq reveals hidden transcriptional variation in malaria parasites. Elife [Internet]. 2018 Mar 27;7. Available from: http://dx.doi.org/10.7554/eLife.33105

20. Howick VM, Russell AJC, Andrews T, Heaton H, Reid AJ, Natarajan K, et al. The Malaria Cell Atlas: Single parasite transcriptomes across the complete Plasmodium life cycle. Science [Internet]. 2019 Aug 23;365(6455). Available from: http://dx.doi.org/10.1126/science.aaw2619

21. Dvorin JD, Martyn DC, Patel SD, Grimley JS, Collins CR, Hopp CS, et al. A plant-like kinase in Plasmodium falciparum regulates parasite egress from erythrocytes. Science. 2010 May 14;328(5980):910–2.

22. Wright GJ, Rayner JC. Plasmodium falciparum erythrocyte invasion: combining function with immune evasion. PLoS Pathog. 2014 Mar;10(3):e1003943.

23. Sinha A, Hughes KR, Modrzynska KK, Otto TD, Pfander C, Dickens NJ, et al. A cascade of DNA-binding proteins for sexual commitment and development in Plasmodium. Nature. 2014 Mar 13;507(7491):253–7.

24. Kafsack BFC, Rovira-Graells N, Clark TG, Bancells C, Crowley VM, Campino SG, et al. A transcriptional switch underlies commitment to sexual development in malaria parasites. Nature. 2014 Mar 13;507(7491):248–52.

25. van Dijk MR, van Schaijk BCL, Khan SM, van Dooren MW, Ramesar J, Kaczanowski S, et al. Three members of the 6-cys protein family of Plasmodium play a role in gamete fertility. PLoS Pathog. 2010 Apr 8;6(4):e1000853.

26. Bargieri DY, Thiberge S, Tay CL, Carey AF, Rantz A, Hischen F, et al. Plasmodium Merozoite TRAP Family Protein Is Essential for Vacuole Membrane Disruption and Gamete Egress from Erythrocytes. Cell Host Microbe. 2016 Nov 9;20(5):618–30.

27. Tachibana M, Ishino T, Takashima E, Tsuboi T, Torii M. A male gametocyte osmiophilic body and microgamete surface protein of the rodent malaria parasite Plasmodium yoelii (PyMiGS) plays a critical role in male osmiophilic body formation and exflagellation. Cell Microbiol. 2018 May;20(5):e12821.

28. Olivieri A, Bertuccini L, Deligianni E, Franke-Fayard B, Currà C, Siden-Kiamos I, et al. Distinct properties of the egress-related osmiophilic bodies in male and female gametocytes of the rodent malaria parasite Plasmodium berghei. Cell Microbiol. 2015 Mar;17(3):355–68.

29. Siden-Kiamos I, Pace T, Klonizakis A, Nardini M, Garcia CRS, Currà C. Identification of Plasmodium berghei Oocyst Rupture Protein 2 (ORP2) domains involved in sporozoite egress from the oocyst. Int J Parasitol. 2018 Dec;48(14):1127–36.

30. Gupta DK, Dembele L, Voorberg-van der Wel A, Roma G, Yip A, Chuenchob V, et al. The Plasmodium liver-specific protein 2 (LISP2) is an early marker of liver stage development. Elife [Internet]. 2019 May 16;8. Available from: http://dx.doi.org/10.7554/eLife.43362

31. Yuda M, Iwanaga S, Shigenobu S, Mair GR, Janse CJ, Waters AP, et al. Identification of a transcription factor in the mosquito-invasive stage of malaria parasites. Mol Microbiol. 2009 Mar;71(6):1402–14.

32. Modrzynska K, Pfander C, Chappell L, Yu L, Suarez C, Dundas K, et al. A Knockout Screen of ApiAP2 Genes Reveals Networks of Interacting Transcriptional Regulators Controlling the Plasmodium Life Cycle. Cell Host Microbe. 2017 Jan 11;21(1):11–22.

33. Jeninga MD, Quinn JE, Petter M. ApiAP2 Transcription Factors in Apicomplexan Parasites. Pathogens [Internet]. 2019 Apr 7;8(2). Available from: http://dx.doi.org/10.3390/pathogens8020047

34. Langer RC, Li F, Vinetz JM. Identification of novel Plasmodium gallinaceum zygote- and ookinete-expressed proteins as targets for blocking malaria transmission. Infect Immun. 2002 Jan;70(1):102–6.

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

35. Beeson JG, Drew DR, Boyle MJ, Feng G, Fowkes FJI, Richards JS. Merozoite surface proteins in red blood cell invasion, immunity and vaccines against malaria. FEMS Microbiol Rev. 2016 May;40(3):343–72.

36. Werner EB, Taylor WR, Holder AA. A Plasmodium chabaudi protein contains a repetitive region with a predicted spectrin-like structure. Mol Biochem Parasitol. 1998 Aug 1;94(2):185–96.

37. Sharma A, Sharma A, Dixit S, Sharma A. Structural insights into thioredoxin-2: a component of malaria parasite protein secretion machinery. Sci Rep. 2011 Dec 1;1:179.

38. Matthews K, Kalanon M, Chisholm SA, Sturm A, Goodman CD, Dixon MWA, et al. The Plasmodium translocon of exported proteins (PTEX) component thioredoxin-2 is important for maintaining normal blood-stage growth. Mol Microbiol. 2013 Sep;89(6):1167–86.

39. Navale R, Atul, Allanki AD, Sijwali PS. Characterization of the autophagy marker protein Atg8 reveals atypical features of autophagy in Plasmodium falciparum. PLoS One. 2014 Nov 26;9(11):e113220.

40. Vandana, Singh AP, Singh J, Sharma R, Akhter M, Mishra PK, et al. Biochemical characterization of unusual cysteine protease of P. falciparum, metacaspase-2 (MCA-2). Mol Biochem Parasitol. 2018 Mar;220:28–41.

41. Rodhain J. Plasmodium epomophori n. sp. parasite commun des Roussettes epaulieres au Congo Belge. Bull Soc Pathol Exot Filiales. 1926;19:828–38.

42. Böhme U, Otto TD, Cotton JA, Steinbiss S, Sanders M, Oyola SO, et al. Complete avian malaria parasite genomes reveal features associated with lineage-specific evolution in birds and mammals. Genome Res. 2018 Apr;28(4):547–60.

43. Crosnier C, Bustamante LY, Bartholdson SJ, Bei AK, Theron M, Uchikawa M, et al. Basigin is a receptor essential for erythrocyte invasion by Plasmodium falciparum. Nature. 2011 Nov 9;480(7378):534–7.

44. Caldelari R, Dogga S, Schmid MW, Franke-Fayard B, Janse CJ, Soldati-Favre D, et al. Transcriptome analysis of Plasmodium berghei during exo-erythrocytic development. Malar J. 2019 Sep 24;18(1):330.

45. Preiser PR, Khan S, Costa FTM, Jarra W. Stage-specific transcription of distinct repertoires of a multigene family during Plasmodium life cycle. 2002; Available from: https://science.sciencemag.org/content/295/5553/342.short

46. Goldberg TL, Sintasath DM, Chapman CA, Cameron KM, Karesh WB, Tang S, et al. Coinfection of Ugandan red colobus (Procolobus [Piliocolobus] rufomitratus tephrosceles) with novel, divergent delta-, lenti-, and spumaretroviruses. J Virol. 2009 Nov;83(21):11318–29.

47. Lauck M, Sibley SD, Hyeroba D, Tumukunde A, Weny G, Chapman CA, et al. Exceptional simian hemorrhagic fever virus diversity in a wild African primate community. J Virol. 2013 Jan;87(1):688–91.

48. Lauck M, Hyeroba D, Tumukunde A, Weny G, Lank SM, Chapman CA, et al. Novel, divergent simian hemorrhagic fever viruses in a wild Ugandan red colobus monkey discovered using direct pyrosequencing. PLoS One. 2011 Apr 22;6(4):e19056.

49. Simons N. The Role of Gene Regulation in Infectious Disease in the Ugandan Red Colobus Monkey (Piliocolobus tephrosceles). 2018; Available from: https://scholarsbank.uoregon.edu/xmlui/handle/1794/23729

50. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015 Jan;12(1):59–60.

51. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018 Sep 15;34(18):3094–100.

52. Sayers EW, Agarwala R, Bolton EE, Brister JR, Canese K, Clark K, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2019 Jan 8;47(D1):D23–8.

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

53. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015 Apr;12(4):357–60.

54. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014 Mar 3;15(3):R46.

55. PlasmoDB: An integrative database of the Plasmodium falciparum genome. Tools for accessing and analyzing finished and unfinished sequence data. The Plasmodium Genome Database Collaborative. Nucleic Acids Res. 2001 Jan 1;29(1):66–9.

56. Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. Direct determination of diploid genome sequences. Genome Res. 2017 May;27(5):757–67.

57. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012 May;19(5):455–77.

58. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078–9.

59. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 Mar 15;26(6):841–2.

60. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011 Oct 11;7:539.

61. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics. 2009 May 1;25(9):1189–91.

62. Hunt M, Silva ND, Otto TD, Parkhill J, Keane JA, Harris SR. Circlator: automated circularization of genome assemblies using long sequencing reads. Genome Biol. 2015 Dec 29;16:294.

63. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017 May;27(5):722–36.

64. Morgulis A, Gertz EM, Schäffer AA, Agarwala R. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. J Comput Biol. 2006 Jun;13(5):1028–40.

65. Bushnell B, Rood J, Singer E. BBMerge - Accurate paired shotgun read merging via overlap. PLoS One. 2017 Oct 26;12(10):e0185056.

66. Bonfield JK, Whitwham A. Gap5--editing the billion fragment sequence assembly. Bioinformatics. 2010 Jul 15;26(14):1699–703.

67. Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, et al. BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol [Internet]. 2017 Dec 6; Available from: http://dx.doi.org/10.1093/molbev/msx319

68. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013 Apr;30(4):772–80.

69. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999 Jan 15;27(2):573–80.

70. Zhu B-H, Xiao J, Xue W, Xu G-C, Sun M-Y, Li J-T. P_RNA_scaffolder: a fast and accurate genome scaffolder using paired-end RNA-sequencing reads. BMC Genomics. 2018 Mar 2;19(1):175.

71. Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, Otto TD. REAPR: a universal tool for genome assembly evaluation. Genome Biol. 2013 May 27;14(5):R47.

72. Jackman SD, Coombe L, Chu J, Warren RL, Vandervalk BP, Yeo S, et al. Tigmint: correcting assembly errors

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

using linked reads from large molecules. BMC Bioinformatics. 2018 Oct 26;19(1):393.

73. Otto TD, Sanders M, Berriman M, Newbold C. Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology. Bioinformatics. 2010 Jul 15;26(14):1704–7.

74. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014 Nov 19;9(11):e112963.

75. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015 Oct 1;31(19):3210–2.

76. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 2007 May 1;23(9):1061–7.

77. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan 1;29(1):15–21.

78. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010 May;28(5):511–5.

79. Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. Bioinformatics. 2012 Feb 15;28(4):464–9.

80. Carver TJ, Rutherford KM, Berriman M, Rajandream M-A, Barrell BG, Parkhill J. ACT: the Artemis Comparison Tool. Bioinformatics. 2005 Aug 15;21(16):3422–3.

81. Kalvari I, Nawrocki EP, Argasinska J, Quinones-Olvera N, Finn RD, Bateman A, et al. Non-Coding RNA Analysis Using the Rfam Database. Curr Protoc Bioinformatics. 2018 Jun;62(1):e51.

82. Boddey JA, Carvalho TG, Hodder AN, Sargeant TJ, Sleebs BE, Marapana D, et al. Role of plasmepsin V in export of diverse protein families from the Plasmodium falciparum exportome. Traffic. 2013 May;14(5):532–50.

83. Pain A, Böhme U, Berry AE, Mungall K, Finn RD, Jackson AP, et al. The genome of the simian and human malaria parasite Plasmodium knowlesi. Nature. 2008 Oct 9;455(7214):799–803.

84. Auburn S, Böhme U, Steinbiss S, Trimarsanto H, Hostetler J, Sanders M, et al. A new Plasmodium vivax reference sequence with improved assembly of the subtelomeres reveals an abundance of pir genes. Wellcome Open Res. 2016 Nov 15;1:4.

85. Pasini EM, Böhme U, Rutledge GG, Voorberg-Van der Wel A, Sanders M, Berriman M, et al. An improved Plasmodium cynomolgi genome assembly reveals an unexpected methyltransferase gene expansion. Wellcome Open Res. 2017 Jun 16;2:42.

86. Otto TD, Böhme U, Jackson AP, Hunt M, Franke-Fayard B, Hoeijmakers WAM, et al. A comprehensive evaluation of rodent malaria parasite genomes and gene expression. BMC Biol. 2014 Oct 30;12:86.

87. Otto TD, Rayner JC, Böhme U, Pain A, Spottiswoode N, Sanders M, et al. Genome sequencing of chimpanzee malaria parasites reveals possible pathways of adaptation to human hosts. Nat Commun. 2014 Sep 9;5:4754.

88. Böhme U, Otto TD, Sanders M, Newbold CI, Berriman M. Progression of the canonical reference malaria parasite genome from 2002-2019. Wellcome Open Res. 2019 May 28;4:58.

89. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 2001 Jan 19;305(3):567–80.

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

90. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. Nucleic Acids Res. 2019 Jan 8;47(D1):D427–32.

91. Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat Biotechnol. 2019 Apr;37(4):420–3.

92. Li L, Stoeckert CJ Jr, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003 Sep;13(9):2178–89.

93. Katoh K, Misawa K, Kuma K-I, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002 Jul 15;30(14):3059–66.

94. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000 Apr;17(4):540–52.

95. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015 Jan;32(1):268–74.

96. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 2002 Apr 1;30(7):1575–84.

97. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015 Jan 15;31(2):166–9.

98. Nurk S, Bankevich A, Antipov D, Gurevich AA, Korobeynikov A, Lapidus A, et al. Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. J Comput Biol. 2013 Oct;20(10):714–37.

99. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019 Jan 8;47(D1):D506–15.

100. Warrenfeltz S, Basenko EY, Crouch K, Harb OS, Kissinger JC, Roos DS, et al. EuPathDB: The Eukaryotic Pathogen Genomics Database Resource. Methods Mol Biol. 2018;1757:69–113.

101. Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, et al. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. Genome Biol. 2019 Oct 28;20(1):224.

102. Abascal F, Zardoya R, Telford MJ. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W7–13.

103. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007 Aug;24(8):1586–91.

104. Meerstein-Kessel L, van der Lee R, Stone W, Lanke K, Baker DA, Alano P, et al. Probabilistic data integration identifies reliable gametocyte-specific proteins and transcripts in malaria parasites. Sci Rep. 2018 Jan 11;8(1):410.

105. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep;20(9):1297–303.

106. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015 May;12(5):453–7.

107. Tachibana S-I, Sullivan SA, Kawai S, Nakamura S, Kim HR, Goto N, et al. Plasmodium cynomolgi genome sequences provide insight into Plasmodium vivax and the monkey malaria clade. Nat Genet. 2012 Sep;44(9):1051–5.

108. Frech C, Chen N. Variant surface antigens of malaria parasites: functional and evolutionary insights from comparative gene family classification and analysis. BMC Genomics. 2013 Jun 27;14:427.

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure Legends

Fig 1. An assembly of genomic sequencing reads from a red colobus monkey blood sample contained significant amounts of sequence from the parasite *Hepatocystis* spp. (A) Contigs from the *Piliocolobus* assembly had a bimodal distribution of AT-content and sequence similarity to *Plasmodium* spp. (B). A phylogenetic tree of *cytochrome b* indicated that the closest match for the apicomplexan parasite sequenced from *Piliocolobus tephrosceles* blood is a *Hepatocystis* isolate from a monkey host. Parasite *cytochrome b* sequences derived from RNA-seq assemblies from *Piliocolobus tephrosceles* blood samples are almost entirely identical to the *cytochrome b* sequence assembled from *Hepatocystis* DNA reads from a single monkey. Branches of the tree have been coloured by bootstrap support values from 15 (red) to 100 (green). Some of the bootstrap support values have also been added next to the nodes as text. Red arrows highlight the *Hepatocystis* samples from the current study.

Fig 2. Whole genome phylogeny and key features of the *Hepatocystis* genome. A whole-genome phylogenetic tree is combined with a graphical overview of key features of *Hepatocystis* and *Plasmodium* species (genome versions from August 2019). The maximum likelihood phylogenetic tree of *Hepatocystis* and *Plasmodium* species is based on 1112706 amino acid residues from 3271 single copy orthologs encoded by the nuclear genome. Bootstrap support values of all nodes were 100, except for one node where the value was 82. The rooting of the tree at *P. gallinaceum* is based on previously published *Plasmodium* phylogenetic trees (11,42). TRAP - thrombospondin-related anonymous protein. RBP protein - reticulocyte binding protein. For the phylogenetic tree *P. cynomolgi* B was used (107).

Fig 3. Deconvolution of Hepatocystis *in vivo* RNA-seq data supports a lack of erythrocytic schizogony and a variable sex ratio. (A) Distributions of SNPs per 100 kb in each *Hepatocystis* RNA-seq sample highlight low and consistent genetic diversity. (B) Deconvolution of RNA-seq samples to identify parasite stage composition shows no evidence for blood schizonts. Ring and trophozoite cells are assumed to relate to early stages of gametocyte development, which are not distinguishable from asexual rings and trophozoites. (C) Proportions of early blood stages (ring and trophozoite) are negatively correlated with mature female gametocytes, however male and female gameocyte ratios are poorly correlated, suggesting that sex ratios vary between samples.

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figures

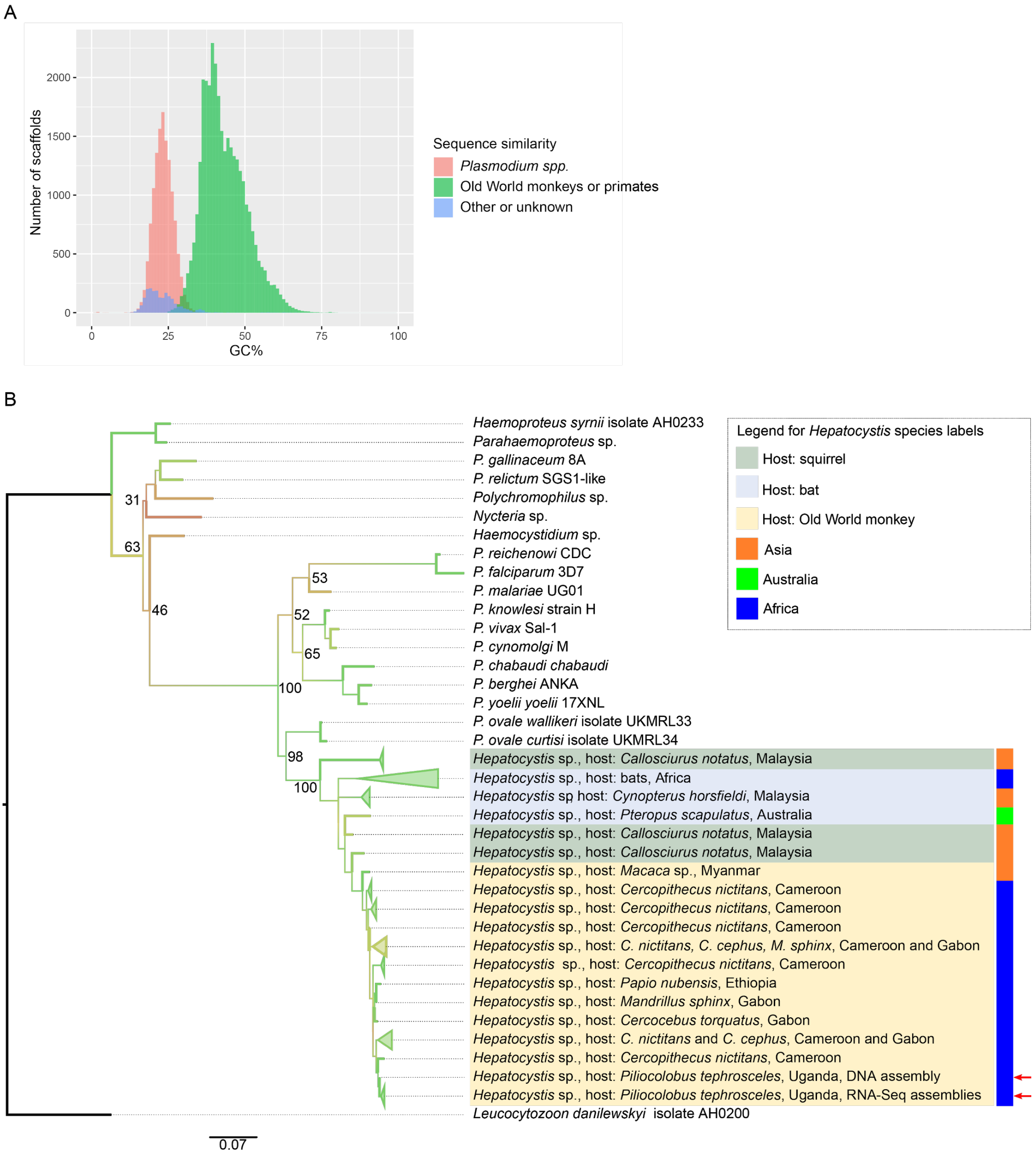


Fig 1

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

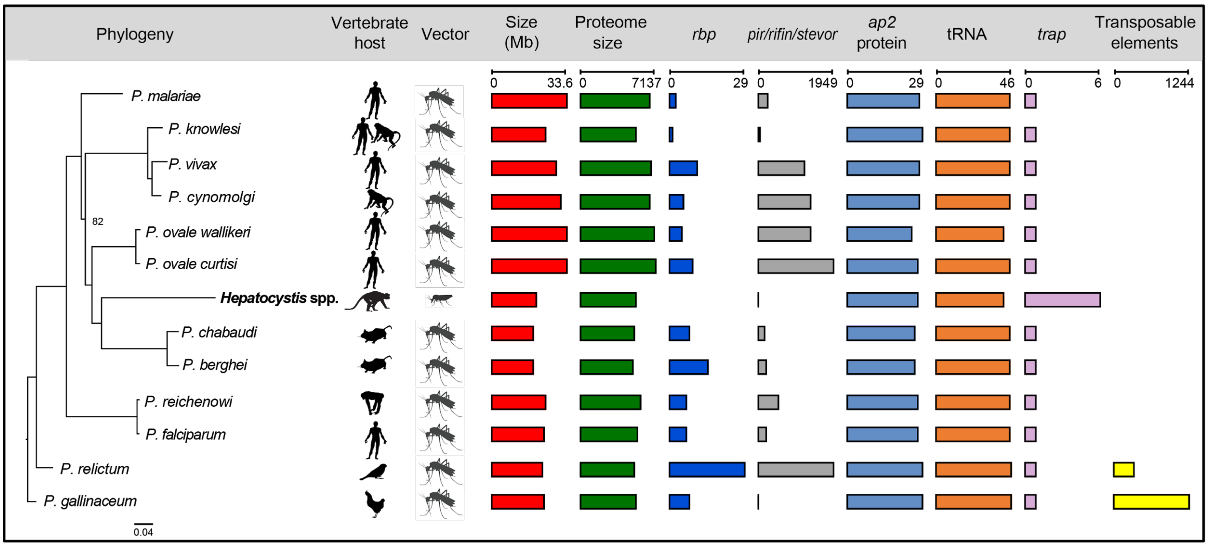


Fig 2

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

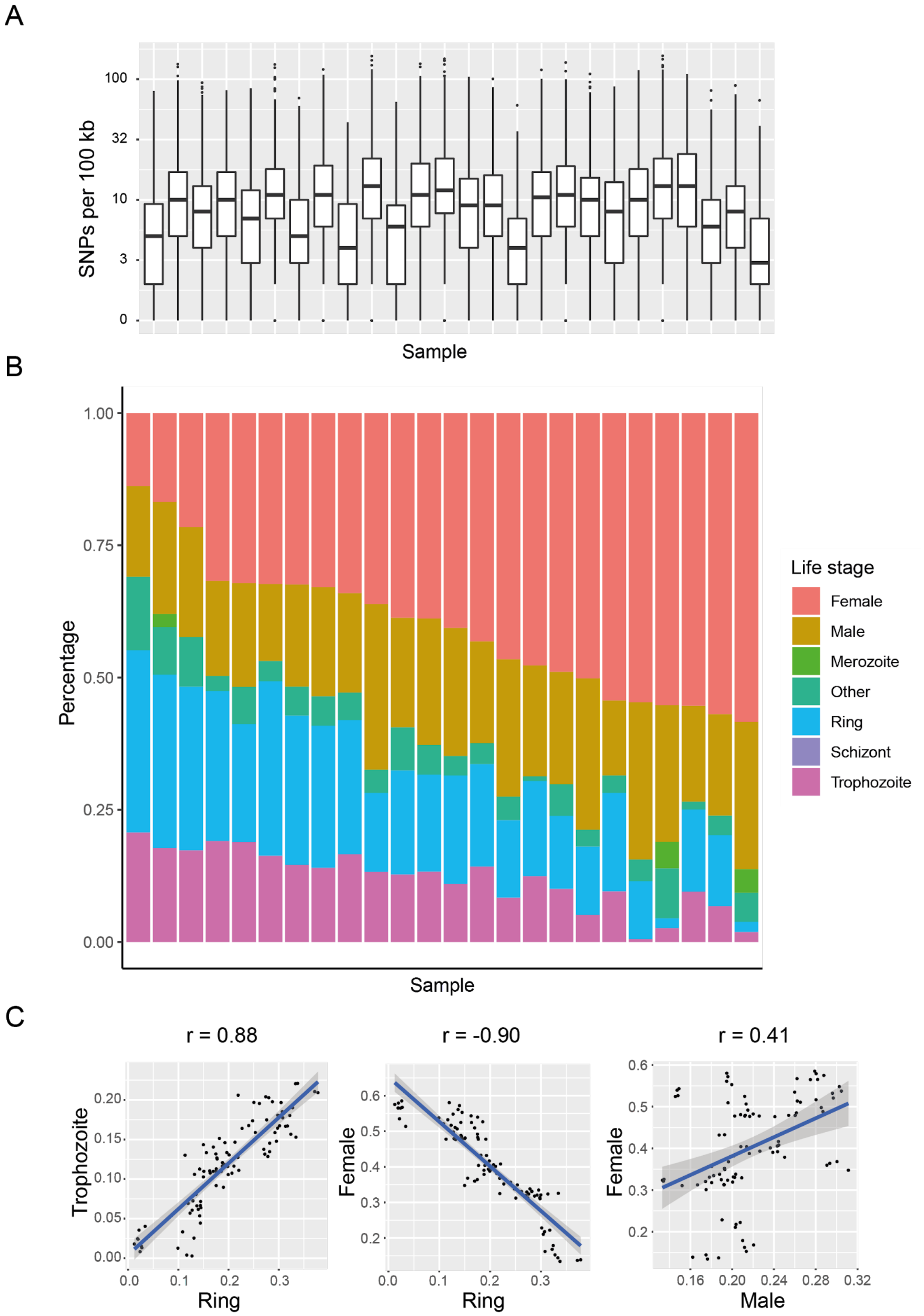


Fig 3

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supporting information

S1 Fig. Conservation of synteny in the core regions of the assembly. ACT (Artemis Comparison Tool) screenshot showing a comparison of centromere-proximal regions of *Hepatocystis* scaffold 132, *P. falciparum* 3D7 (Pf3D7) chromosome 4 and *P. vivax* (PvP01) chromosome 5. The red blocks represent sequence similarity (tBLASTx). The centromere is shown in green. Coloured boxes represent genes. The graph shows the GC-content.

S2 Fig: Organization of putative subtelomeric regions of *Hepatocystis* scaffold 67, scaffold 211, *P. knowlesi* H chromosome 4 and *P. falciparum* 3D7 chromosome 9. Exons are shown in coloured boxes with introns as linking lines. '//' represents a gap. The shaded/grey areas in *P. knowlesi* and *P. falciparum* mark the start of the conserved, syntenic regions to other *Plasmodium* species. The presence of genes that are subtelomeric in *Plasmodium* species, i.e. PHIST proteins, suggests that the *Hepatocystis* scaffolds are also subtelomeric. A complete subtelomere that includes telomeric repeats is missing in our *Hepatocystis* assembly. Thus, whether *Hepatocystis* chromosomes retain the organisation common to most *Plasmodium* species remains unclear.

S3 Fig. Phylogenetic tree of *Haemosporidian* mitochondrial proteins. *Hepatocystis* sp. ex. *Piliocolobus tephrosceles* (this work, marked with red arrow) appears next to a previously sequenced *Hepatocystis* sample from the flying fox *Pteropus hypomelanus* (NCBI accession FJ168565.1). Branches of the tree have been coloured by bootstrap support values from 45 (red) to 100 (green). Bootstrap values below 100 have also been added to the figure as text.

S4 Fig. Phylogenetic tree of 18 apicoplast protein sequences of *Plasmodium* spp. and *Hepatocystis*. Branches of the tree have been coloured by bootstrap support values from 66 (red) to 100 (green). Bootstrap values below 100 have also been added to the figure as text.

S5 Fig. Phylogenetic tree of 11 nuclear genes of *Hepatocystis* and *Plasmodium* species. Genes of *Hepatocystis* sp. ex *Piliocolobus tephrosceles* are highly similar to *Hepatocystis epomophori* genes sequenced in a different study (2). The tree is based on the following genes: splicing factor 3B subunit 1, tubulin gamma chain, DNA polymerase delta catalytic subunit, eukaryotic translation initiation factor 2 gamma subunit, T-complex protein 1 subunit alpha, pantothenate transporter, ribonucleoside-diphosphate reductase large subunit, aminophospholipid-transporting P-ATPase, GCN20, transport protein Sec24A and RuvB-like helicase 3. Branches of the tree have been coloured by bootstrap values from 73 (red) to 100 (green). Bootstrap values below 100 have also been added to the figure as text. The red arrow points to the *Hepatocystis* sample from the current study.

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

S6 Fig. Deconvolution using CIBERSORT and the Malaria Cell Atlas accurately determines the presence and absence of different *Plasmodium* life stages in bulk RNA-seq data. (A) Pre-defined mixtures of pseudobulk RNA-seq data were deconvoluted with very high accuracy. (B) Real samples of *P. berghei* bulk RNA-seq from Otto et al (2014) were deconvoluted showing almost pure mixtures of gametocyte, ookinete or asexual stages as expected. The low proportions of expected parts of the IDC in each asexual sample may result from differences between what the MCA defines as a ring/trophozoite/schizont and what would microscopically be defined as such.

S7 Fig. Multiple sequence alignments of two *Hepatocystis*-specific gene families. (A) Alignment of Hepatocystis-specific gene family 1 (*hep1*). Pseudogenes (HEP_00099300, HEP_00250500, HEP_00323900) were not included in the alignment. HEP_00353700 is 476 amino acids long and was truncated here. (B). Alignment of Hepatocystis-specific gene family 2 (Hep2). This gene family contains a PEXEL motif (marked with a black box). Pseudogenes (HEP_00165000, HEP_00165200, HEP_00324000, HEP_00489100) were not included in the alignment.

S8 Fig. Heatmaps of *Hepatocystis* gene family expression in the blood of its mammalian host. (A) Expression levels (log vst-normalised) of *hep1* genes across blood samples from multiple red colobus monkeys. The estimated proportions of early blood stages (rings/trophozoites) and mature gametocytes are highlighted above. (B) Expression levels of *hep2* genes (C) Expression levels of *pir* genes.

S9 Fig. Heat map of *pir* protein subfamilies in *Hepatocystis* and *Plasmodium* species. Rows correspond to species and columns correspond to *pir* subfamilies. The columns have been ordered by the number of sequences in each subfamily and the order of rows is approximately based on phylogeny. Colours represent the numbers of proteins belonging to each subfamily for each species. All *Hepatocystis pir* proteins belong to the only subfamily conserved across all these species (108) (indicated with red arrow).

S10 Fig. Some orthologues missing in *Hepatocystis sp.* relative to *Plasmodium* species show common gene expression patterns across the *Plasmodium* life cycle. (A) Malaria Cell Atlas (MCA) gene cluster 10 represents genes highly expressed in late schizonts. 25 genes from this cluster were conserved in *P. ovale wallikeri* and *P. vivax*, but were missing from our *Hepatocystis* genome assembly. Genes were clustered here by expression pattern and single-cells were ordered by pseudotime as in (20). (B) MCA cluster 4 represents genes highly expressed across much of the life cycle - liver stages, trophozoites, female gametocytes and ookinetes/oocysts. 27 genes from this cluster were conserved in *P. ovale wallikeri* and *P. vivax*, but were missing from our *Hepatocystis* genome assembly.

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

S11 Fig. Distributions of *Hepatocystis* dN values in Malaria Cell Atlas (MCA) clusters. *Hepatocystis* dN was calculated in 3-way comparison between *Hepatocystis*, *P. berghei* ANKA and *P. ovale curtisi* using codeml. The Malaria Cell Atlas clusters have been described in Figure 2B in the article on Malaria Cell Atlas (20). (A) *Hepatocystis* genes with dN in the top 5%: observed versus expected ratios for Malaria Cell Atlas clusters. *Hepatocystis* genes that correspond to Malaria Cell Atlas clusters 2, 4 and 6 have less genes with dN rank in the top 5% than expected by chance (Fisher exact test p-value < 0.05). None of the MCA clusters contain significantly more genes ranked in the top 5% of dN than expected by chance, although there is a trend towards clusters clusters 15 and 16 having higher dN. (B) Boxplot of all *Hepatocystis* dN values per each Malaria Cell Atlas cluster. Distribution of values in clusters 15 and 16 differs from the rest of the clusters. Kolmogorov-Smirnov test statistics are the following. Cluster 15 vs all other clusters: D = 0.42, p-value = 1.08e-05. Cluster 16 vs all other clusters: D = 0.52, p-value = 4.68e-12. Clusters 15 and 16 combined vs all other clusters: D = 0.46, p = 2.33e-15.

S1 Table. Summary of gene properties. For each gene in the assembly, the following is listed: annotation, number of exons, gene length (bp), the presence or absence of start and stop codons (reflecting the completeness of the assembly of the gene) and RNA-seq expression level (mean FPKM with standard deviation) in sample SAMN07757854 (RC106R). For the proteins encoded by the genes, the table shows the number of transmembrane segments predicted by TMHMM, ExportPred 2 score, 1 to 1 orthologs in *P. berghei* ANKA and *P. ovale curtisi* GH01 (based on OrthoMCL), PFAM domains, the number of matches to PEXEL motif (RxLxE/Q/D) and SignalP-5 signal peptide prediction.

S2 Table. Raw and normalised *Hepatocystis* gene expression data.

S3 Table. *Plasmodium* orthologues missing in the *Hepatocystis* genome assembly. *Plasmodium berghei* genes, which have an orthologue in *P. ovale curtisi or P. vivax*, but not the *HexPt* DNA (A) or RNA-seq (B) assemblies and are enriched in Malaria Cell Atlas gene clusters.

S4 Table. Genes with Hepatocystis dN rank in the top 5% in codeml 3-way comparison between *Hepatocystis*, *P. berghei* ANKA and *P. ovale curtisi* GH01. The total number of genes in dN analysis was 4009, out of which 200 correspond to 5%. The table includes Malaria Cell Atlas cluster numbers for each gene.

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

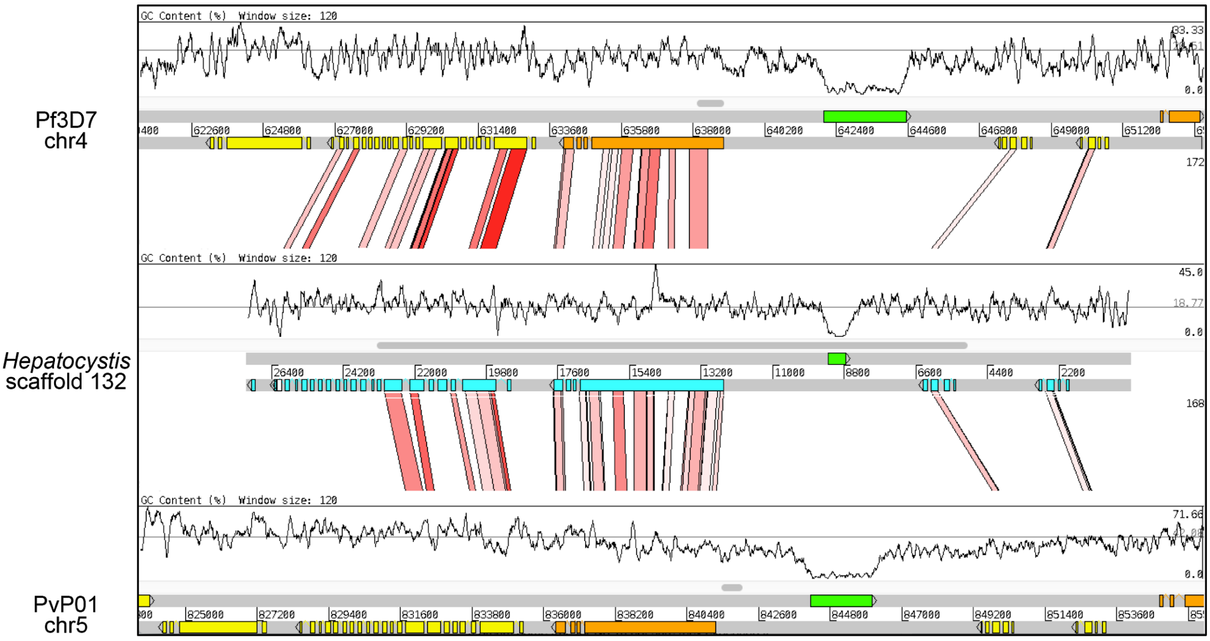


Fig S1

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

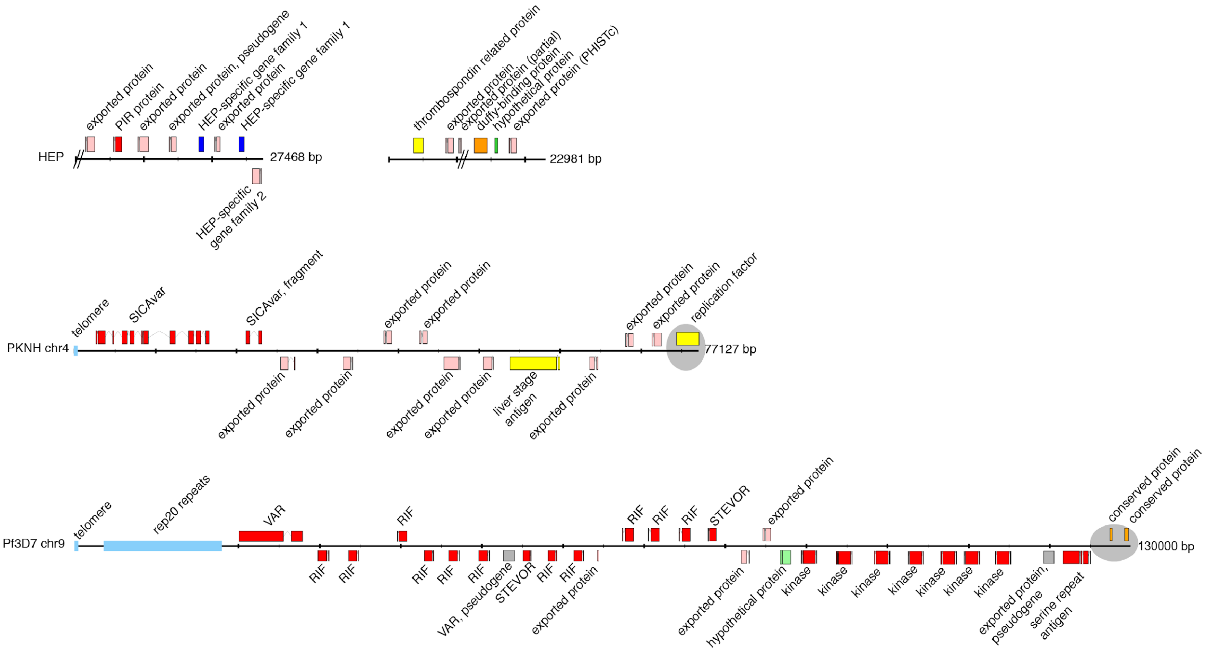


Fig S2

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

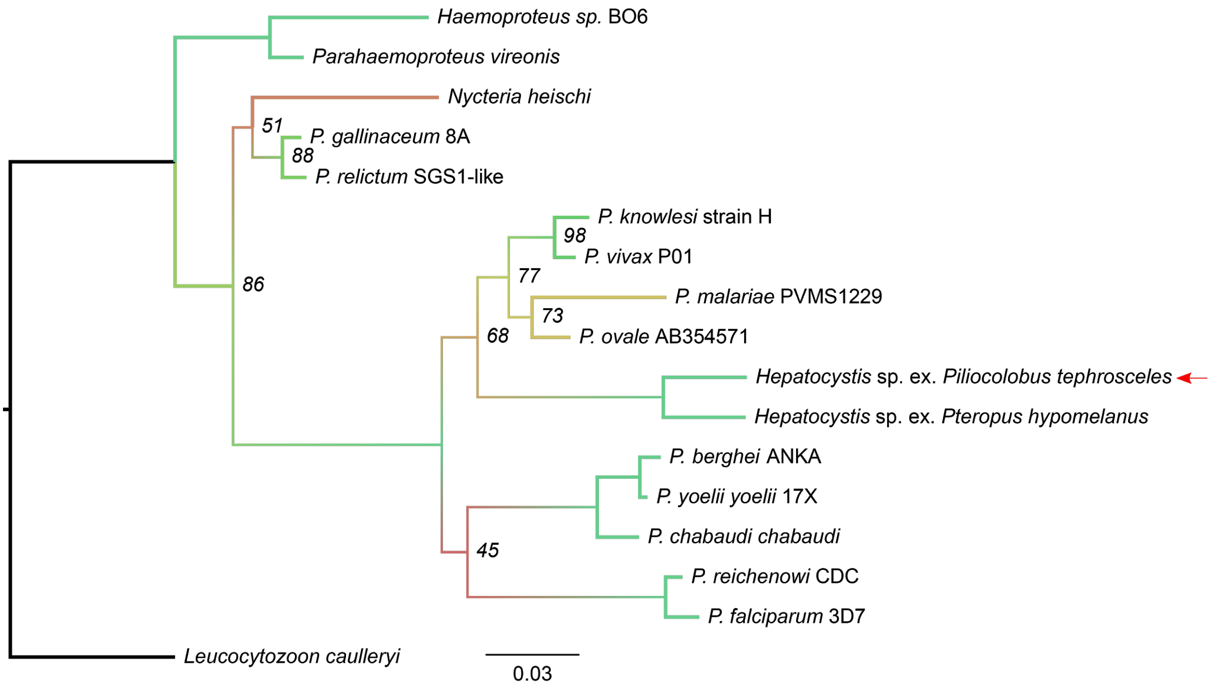
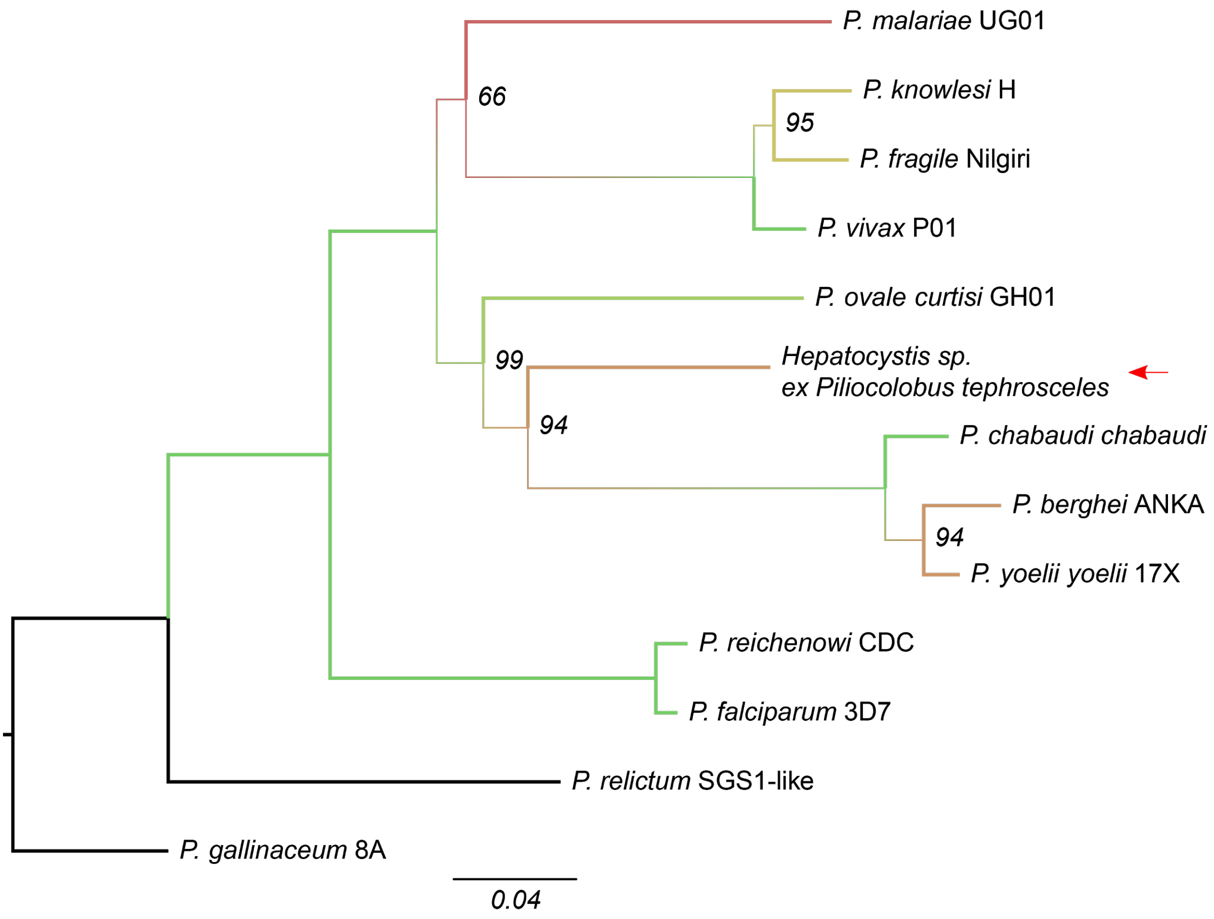


Fig S3



bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Fig S4

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

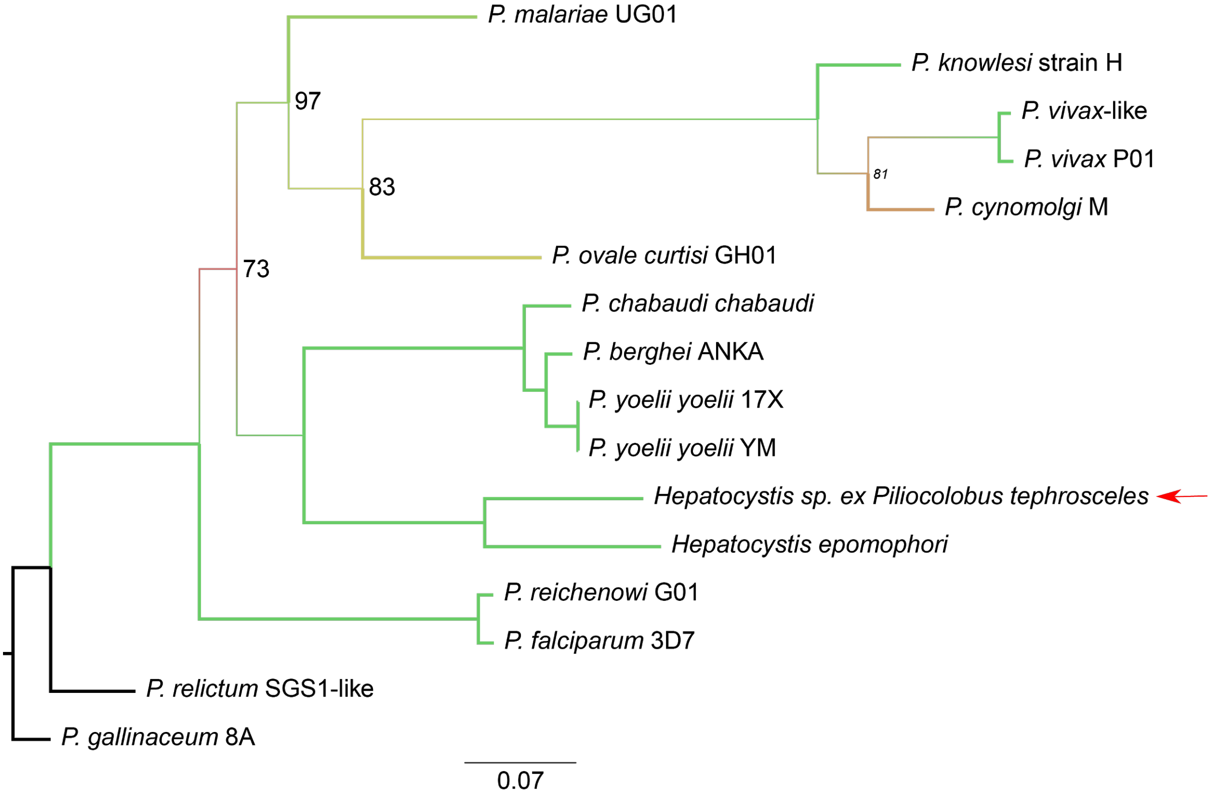


Fig S5

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

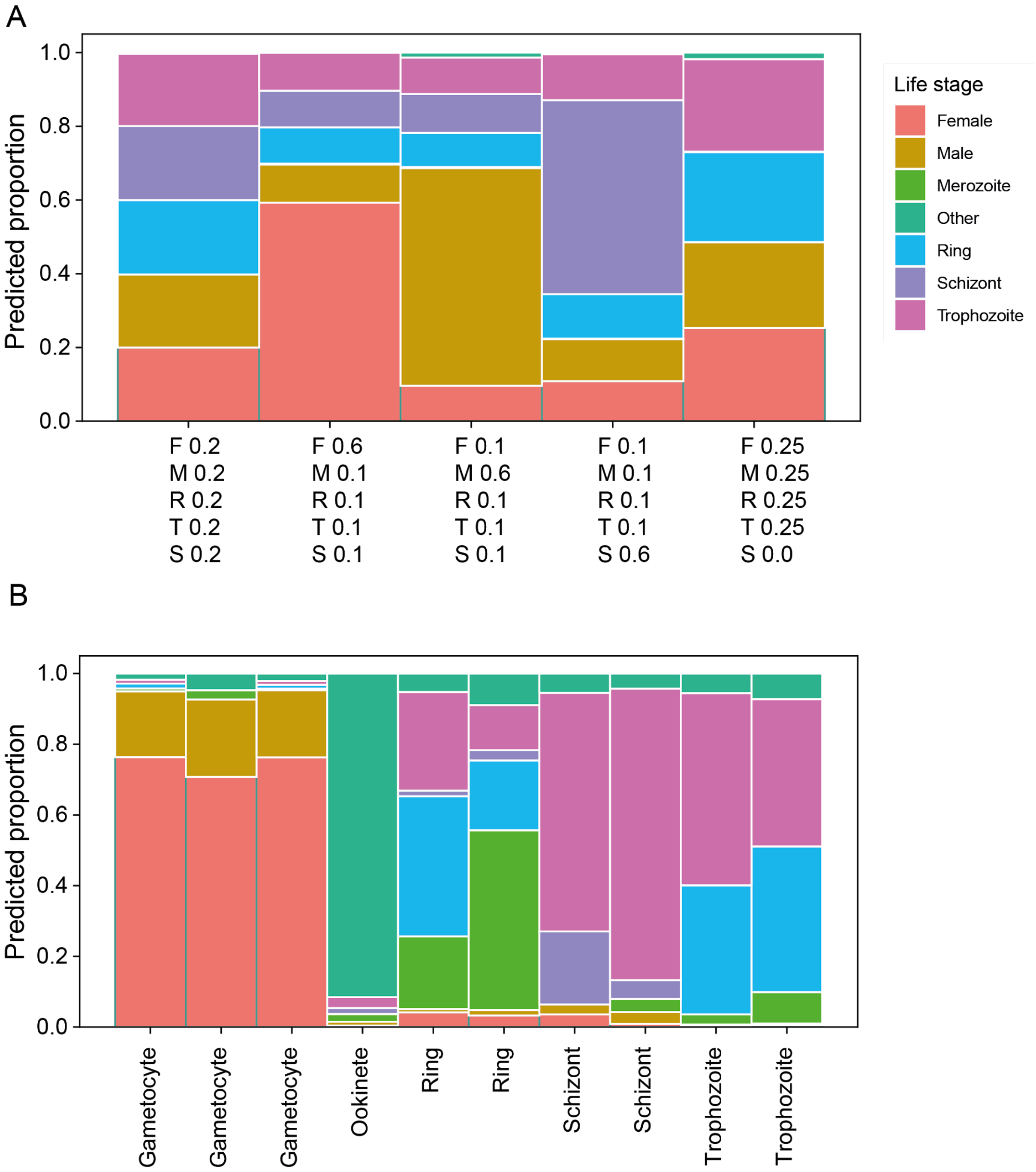


Fig S6

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

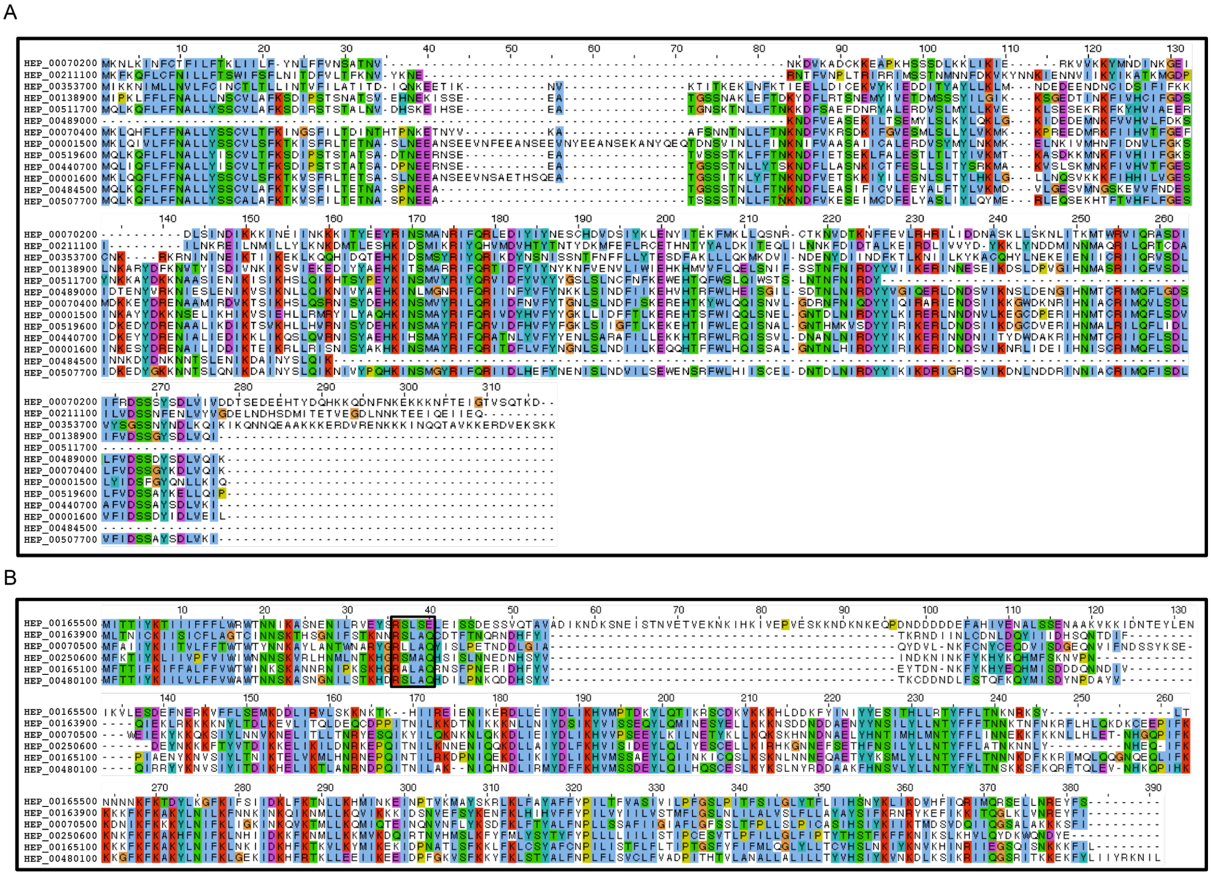


Fig S7

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

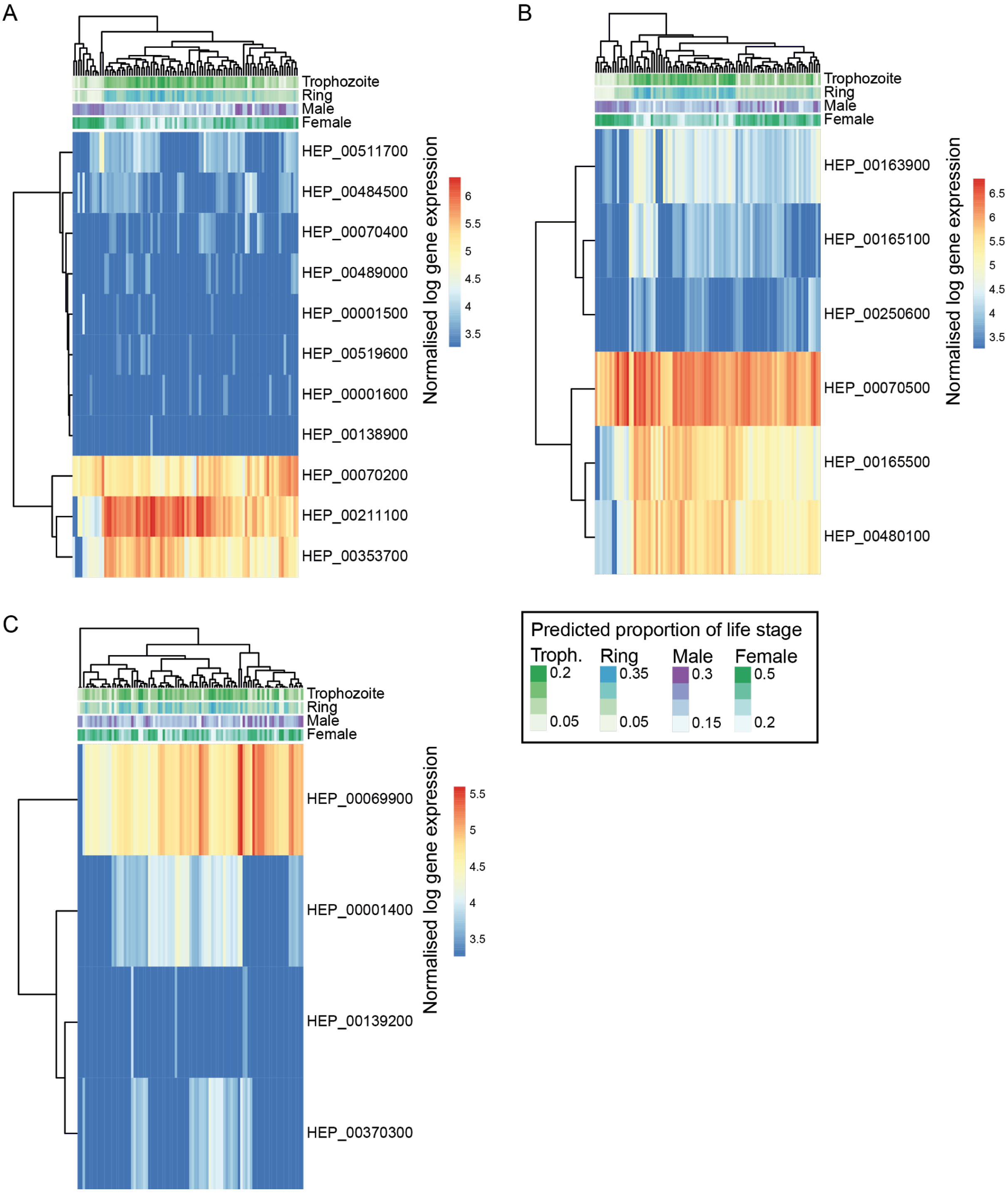


Fig S8

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

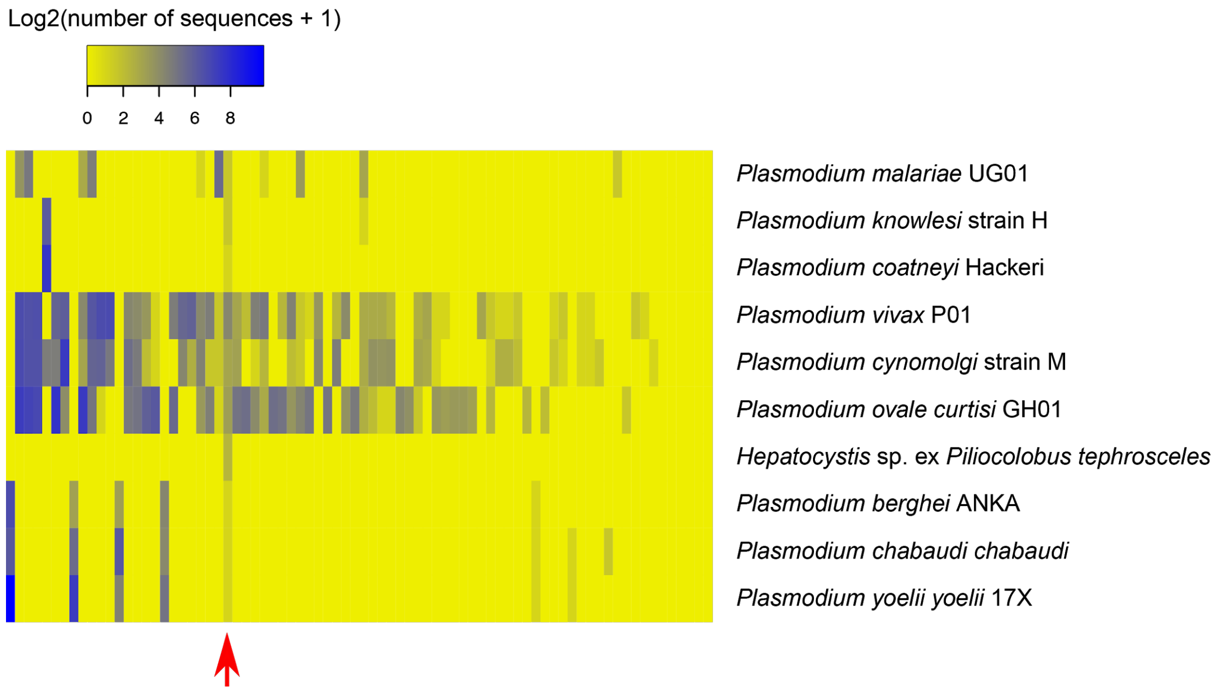


Fig S9

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

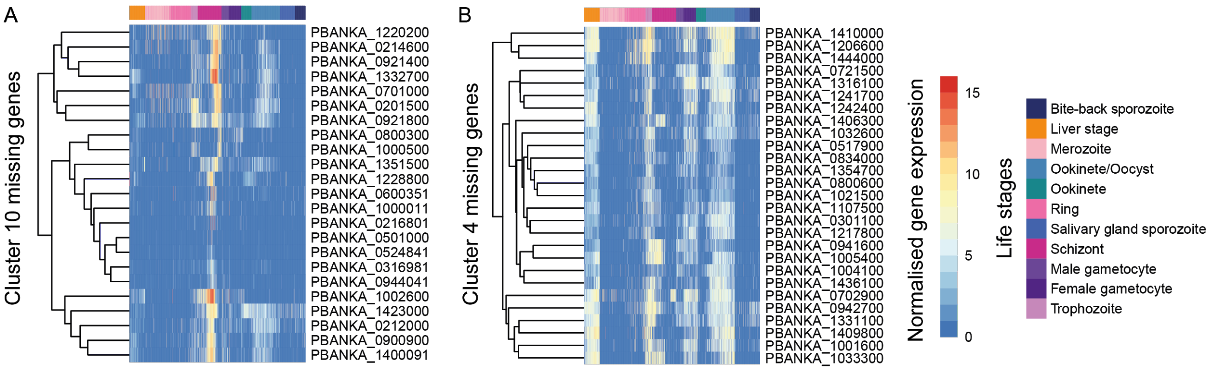


Fig S10

bioRxiv preprint doi: https://doi.org/10.1101/871327; this version posted December 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

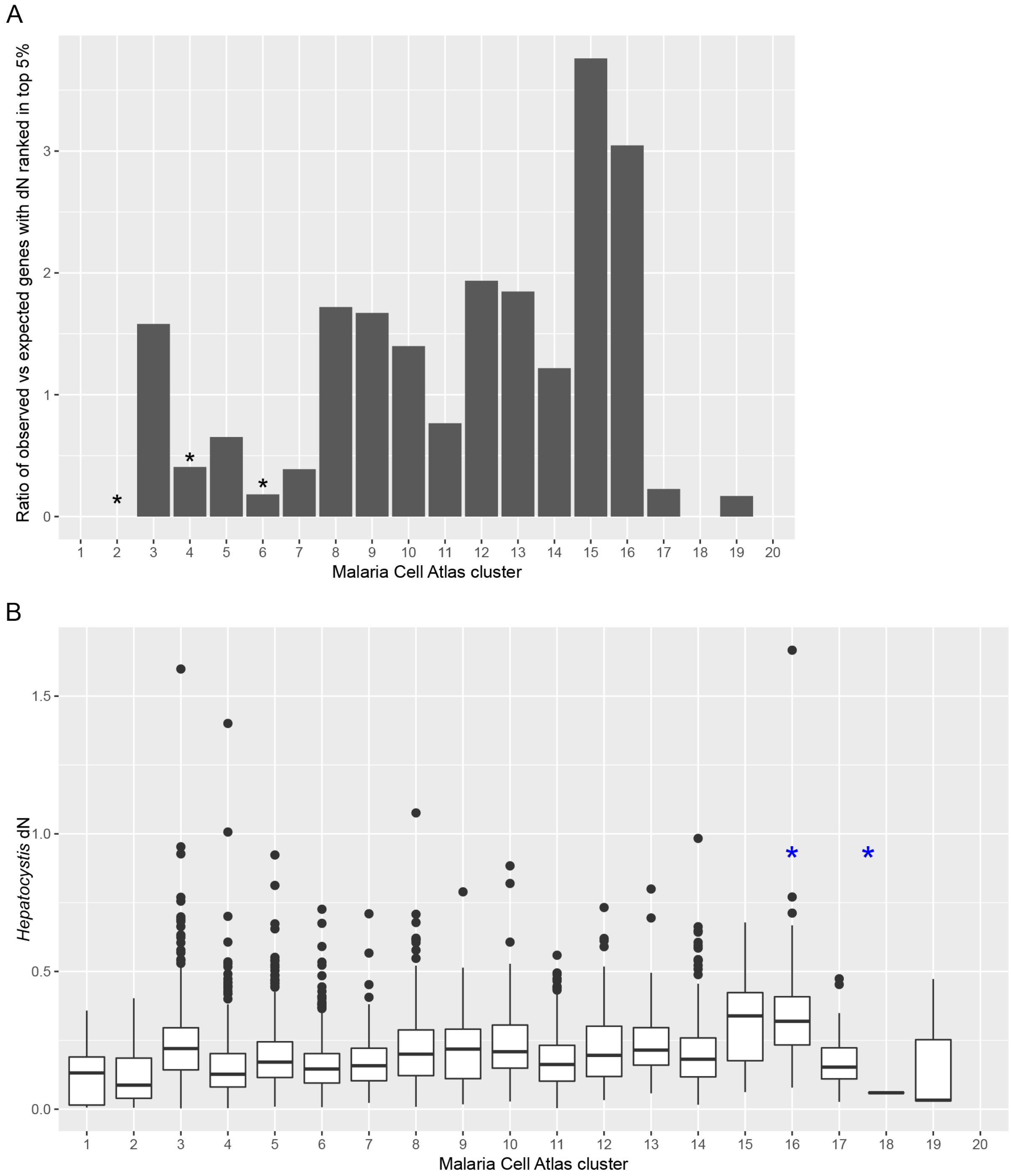


Fig S11