Malaria.tools: comparative genomic and transcriptomic database for *Plasmodium* species

Qiao Wen Tan[1], Marek Mutwil[1*]

[1]School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore 637551, Singapore

* To whom correspondence should be addressed. Tel: +65 6904 7503; Email: mutwil@ntu.edu.sg

ABSTRACT

Malaria is a tropical parasitic disease caused by the *Plasmodium* genus, which resulted in an estimated 219 million cases of malaria and 435,000 malaria-related deaths in 2017. Despite the availability of the *P. falciparum* genome since 2002, almost 50% of the genes remain unannotated. To remedy this paucity of functional information, we used transcriptomic data to build gene co-expression networks for two *Plasmodium* species (*P. falciparum and P. berghei*), and included genomic data of four other *Plasmodium* species, *P. yoleii, P. knowlesi, P. vivax* and *P. cynomolgi*, as well as two non-*Plasmodium* species from the Apicomplexa, *Toxoplasma gondii* and *Theileria parva.* The database is preloaded with tools that allow the identification and cross-species comparison of co-expressed gene neighborhoods, clusters, and life stage-specific expression, thus providing sophisticated tools to predict gene function. Moreover, we exemplify how the tools can be used to easily identify genes relevant for pathogenicity and various life stages of the malaria parasite. The database is freely available at https://malaria.tools.

INTRODUCTION

Malaria is a widespread infectious disease transmitted by the *Anopheles* mosquito which caused an estimated 219 million cases and 435,000 malaria-related deaths in the year 2017 (1). It is caused by various *Plasmodium* species with *P. falciparum* and *P. vivax* being the two most widespread and deadly for humans (2). Over the years, varying degrees of resistance have emerged in *Plasmodium* against all drugs used for treating malaria. Hence, it is a race against time to find new treatments, and this requires an understanding of malaria biology, and more specifically, the characterisation of genes and their functions in order to develop drugs that target genes responsible for pathogenicity.

Despite the availability of genomes of the various *Plasmodium* parasites for already more than a decade, many of the genes have unknown functions. These genes are often parasite-specific (3), thus making conventional gene function prediction methods which are based on homology to genes from other organisms less efficient. To this end, alternative methods for gene function prediction such as gene co-expression networks have been developed. The rationale behind using gene co-expression networks stems from the observation that functionally related genes have similar gene expression profiles (4). Thus, by identifying clusters of genes that have highly similar expression profiles, we attain groups of functionally related genes where the function of uncharacterised genes can be inferred from its neighbours.

We have built the database Malaria.tools (https://malaria.tools) based on the CoNekT database framework (5) using publicly available RNA sequencing data of two model *Plasmodium* species. This database calculates gene co-expression networks from gene expression data obtained from more than 800 experiments and contains multiple tools to predict gene function from co-expression network neighbours, co-expressed clusters and specific expression profiles. Additionally, an in-built phylogenetic tree function combined with gene expression comparison allows *Plasmodium* researchers to identify

orthologs of *P. falciparum* genes. Malaria.tools provides the malaria research community a comprehensive and highly valuable resource for an efficient characterisation of genes based on their expression profile, and will aid in the identification of potential drug targets *in silico* in the race for new antimalarial drugs.

MATERIALS AND METHODS

RNA sequencing experiments for *P. falciparum* and *P. berghei* were downloaded as fastq files from European Nucleotide Archive (ENA) (6) via aspera v3.8.1.160447. For paired end experiments, only the file containing the first read, designated with "_1" was downloaded. To remove possible RNA contaminants from the parasite's hosts, the reads were mapped first against human and mouse for *P. falciparum* and *P. berghei*, respectively. The unmapped reads were then mapped against the mosquito vector. Human mosquito vector was used for both *Plasmodium* species as the CDS of the mouse mosquito vector was unavailable. Unmapped reads were mapped against the respective *Plasmodium* species, and experiments with at least 1 million reads and 50% of reads mapped to the *Plasmodium* species were used to construct the database.

The mapping was done using kallisto v0.44.0 (7). Kallisto index files were generated for *Homo sapiens* (GCF_000001405.38), *Anopheles gambiae* (GCF_000005575.2), *P. falciparum* 3D7 and *P. berghei* ANKA CDS sequences with default parameters. Both paired and single libraries were treated as single end libraries for mapping using kallisto quant for single end library with estimated fragment length of 200bp, estimated standard deviation of 20 and pseudobam option. Unmapped reads from the output BAM files were written into a new fastq file using samtools v1.9-52-g651bf14 (8).

In total, 206 experiments from *P. berghei* (Figure S1A) and 620 experiments from *P. falciparum* (Figure S1B) were included in database and annotated based on the information available from NCBI Sequence Read Archive, ENA and literature (Supplementary Table S3 (*P. berghei)* and Supplementary Table S4 (*P. falciparum*)). The Transcripts Per Kilobase Million (TPM) values from kallisto output of the selected experiments were represented as an expression matrix where genes are arranged in rows and experiments in columns. The expression matrices are available in Table S1 for *P. berghei* and Table S2 for *P. falciparum*. Highest Reciprocal Rank (HRR) co-expression networks were then constructed (9).

CDS sequences, description and associated GO terms for the 8 species in the database were obtained from the various sources described in Table 1. Superfamily and Pfam domain annotations from interproscan v5.32-71.0 (10) were used as sequence description when descriptions were not available. Orthologous groups of genes (Supplementary Table S5) and phylogenetic trees were obtained from Orthofinder v1.1.8 (11) using BLAST for sequence similarity inference with default settings. The database was based on the CoNekt database framework (5) with default settings, where Heuristic Cluster Chiseling Algorithm (HCCA) (9) cluster size was limited to 100 genes. Experiments involving wild type *Plasmodium* were further binned according to its various life stages (ring, trophozoite, schizont, male

gametocyte and female gametocyte) and used to calculate tissue specificity. Nucleotide and protein blast databases for blastn and blastp were created using makeblastdb v2.9.0+ (12).

RESULTS

Malaria.tools offers a wide selection of tools to query the database. For example, the user can find the genes of interest by using BLAST, gene IDs (e.g., *PF3D7_1223100*) and keywords (e.g., rhoptry). Genes that work together in a specific biological process or contain a particular domain can be identified by querying the database with Gene Ontology terms (e.g., GO:0009405) or a Pfam domain (e.g., VSA_Rifin), respectively. The database offers multiple comparative genomic and transcriptomic tools that allow the user to view and compare expression profiles within and across species, and to investigate the phylogenetic and expression relationships of gene families. A full description of the features is found at https://malaria.tools/features. To exemplify some of the features of malaria.tools, we provide three analyses showing typical case studies.

Identification of a co-expression neighborhood important for erythrocyte invasion

A co-expression neighbourhood consists of a gene of interest and its co-expressed genes (neighbors) calculated based on Highest Reciprocal Rank (HRR). To identify co-expression neighborhoods of interest to malaria researchers, we calculated which genes are network neighbors to already functionally characterized genes. We identified gene *PF3D7_1223100*, which was co-expressed with 91 other genes of which 48 (53%) are annotated with specific GO terms indicating that experimental evidence for the gene function already exists (e.g. evidence codes EXP, IDA, IPI). The high percentage of functionally characterized genes in this neighborhood indicates that the corresponding biological process has received special attention from the malaria researchers, likely due to its involvement in pathogenicity.

To gain insight into the function of this neighborhood, we first studied the expression profile of *PF3D7_1223100* (https://malaria.tools/sequence/view/16011). The expression profiles in malaria.tools are available in a detailed format that showcase expression in all annotated samples (Figure 1A), and as an average expression in the major life stages of malaria (Figure 1B, https://malaria.tools/profile/view/8699). In both expression profiles, we observed that *PF3D7_1223100* and its co-expressed genes are expressed in all major life stages with particularly high expression in the schizont stage (Figure 1A,B). Next, we retrieved publications on the functional characterization of the genes and we found that the highly studied genes are clearly associated with functions important for erythrocyte invasion. The genes can be classified into 3 major groups relating to motility, cytoadherence and erythrocyte invasion. Genes relating to the glideosome complex [*PF3D7_0918000* (GAP50) (13), *PF3D7_1323700* (GAPM1) (14)], the inner membrane complex (15) [*PF3D7_0109000* (PHIL1) and *PF3D7_1003600* (IMC1c)] and cytoskeleton (16, 17) [*PF3D7_1251200* (coronin), *PF3D7_0932200* (Profilin) and *PF3D7_1246200* (Actin 1)] are essential for the parasite to move towards a new red blood cell through gliding motility. Upon reaching the red blood cell, merozoite surface proteins such as *PF3D7_1035400* (MSP3), *PF3D7_1335100* (MSP7), *PF3D7_1035500* (MSP6) (18) and *PF3D7_1035700* (DBLMSP) facilitate the

binding of the parasite to the red blood cell. Finally, erythrocyte invasion is enabled by various genes such as enzymes [*PF3D7_0507500* (SUB1) (19), *PF3D7_1136500* (casein kinase 1) (20) and *PF3D7_0404700* (DPAP3) (21), signalling mediators (22) [*PF3D7_0934800* (PKAc), *PF3D7_1223100* (PKAr)], rhoptry proteins (23, 24) (*PF3D7_0929400* (RhopH2), *PF3D7_0905400* (RhopH3), *PF3D7_1410400* (RAP1), *PF3D7_0414900* (ARO), *PF3D7_1017100* (RON12), *PF3D7_0501600* (RAP2) and *PF3D7_0817700* (RON5)] and others [*PF3D7_0423800* (CyRPA) (25), *PF3D7_0935800* (CLAG9) (26), *PF3D7_0612700* (P12), *PF3D7_0404900* (P41) (27)]. The schizont is a non-infective life stage during the erythrocytic cycle. However, a mature schizont contains multiple merozoites, which upon rupture of the schizont moves and invades fresh erythrocytes. In the co-expression neighbourhood of *PF3D7_1223100*, we observe an upregulation of merozoite and erythrocyte invasion-related genes during the schizont stage. In conclusion, the remaining 47% of genes in this cluster that are not yet functionally characterized are prime candidates for further studies on parasite motility, cytoadherence and erythrocyte invasion.

Comparative transcriptomic analysis of gene modules involved in male gametocyte-specific motility

Comparative transcriptomic analyses can reveal which gene modules are conserved across species (28, 29), thus enabling the identification of the core genetic components of specific biological processes (29, 30). Malaria.tools provides two methods to extract these conserved transcriptional programs by (i) identifying common gene families that are specifically expressed in a particular life stage of two malaria species or by (ii) identifying conserved clusters of co-expressed genes.

Using the first method to identify conserved transcriptional programs, we navigated to 'Tools\Compare specificities', selected species *P. falciparum* and *P. berghei*, set condition 'Gametocyte (male) for both species and clicked 'Compare specificity'. The database first identified genes that are preferentially expressed in male gametocytes in both species (specificity measure (SPM) > 0.85) (31), which revealed that 187 gene families are expressed at this life stage in the two parasites (Figure 2A, Table S6). The table below the Venn diagram shows the identity and links to the 187 gene families, and clicking on the tree links of a gene family depicted the phylogenetic and expression relationships of the genes in the family. Not surprisingly, many of the gene families show conserved male gametocyte-specific expression in the two *Plasmodium* species, as exemplified by the phylogenetic tree of gene family OG0000502 (Figure 2B, https://malaria.tools/tree/view/503), indicating that this family is male gametocyte-specific. Interestingly, we also observed cases where only one of the clades of the phylogenetic tree showed a male gametocyte-specific expression (Figure 2C, OG0000055, https://malaria.tools/tree/view/56), suggesting that for this particular gene family an ancient gene duplication took place in the ancestor of the *Plasmodium* species followed by either a sub-functionalization or neo-functionalization of the genes. The list of the conserved male-specific genes and gene families provides a good starting point to dissect the genetic basis of male gametocyte-specific biological processes.

The second method to identify conserved gene modules using malaria.tools is based on co-expression network clusters. The clusters are used to identify functionally related genes based on the topology of the networks (9), i.e. similar clusters are found by identifying which cluster pairs contain significantly similar (P<0.05, hypergeometric test) number of gene families (expressed as Jaccard index,(5)). To exemplify this feature, we clicked first on one of the *P. berghei* genes found in the table (*PBANKA_0102700*, https://malaria.tools/sequence/view/17914) and then on the co-expression cluster 13 containing this gene (https://malaria.tools/cluster/view/40). The 'Similar clusters' table found on this page identified *P. falciparum* cluster 15 as being significantly similar to cluster 13 (Jaccard index = 0.246). Clicking on the 'Compare' button revealed the co-expression networks of the two modules (Figure 2D). As expected, the two conserved clusters show male gametocyte-specific expression profiles (https://malaria.tools/cluster/view/40, https://malaria.tools/cluster/view/124), further reinforcing that the two clusters represent a *bona fide* conserved transcriptional program for male gametocyte-specific motility.

In summary, the analysis of conserved gene modules resulted in 132 genes present in the co-expression networks whereof 123 genes showed male gametocyte-specific expression. However, only 60 of these homologs were annotated and the remaining ones were conserved proteins of unknown function (4 genes) or conserved *Plasmodium* proteins of unknown function (59 genes, Table S7). A closer inspection of the homologs present in the clusters revealed motor proteins such as kinesin (*PBANKA_0202700*, *PF3D7_0111000*, *PBANKA_0902400*, *PF3D7_1146700*, *PBANKA_1458800* and *PF3D7_1245600*) and dynein (*PBANKA_1022400* and *PF3D7_1420800*) (32), as well as flagella-related proteins such as the radial spoke protein 3 (*PBANKA_1039000*), growth arrest protein (33) (*PBANKA_1455800* and *PF3D7_1242400*) and MORN repeat containing protein which localises near the flagellar basal body in male gametocytes (34) (*PBANKA_1018200* and *PF3D7_1426400*). Taken together, the functions of these genes suggest an overall motility and flagella-related function associated with the clusters. Hence, the unannotated genes in these homologous networks will be of prime interest for researchers interested in male gametocyte motility.

Identification of microneme-specific co-expressed gene clusters

Micronemes are protein rich, secretory organelles important for host-cell invasion and gliding motility in parasitic Apicomplexans (35). Proteins are being discharged to facilitate entry of the parasites into red blood cells.

To gain insight into microneme biogenesis and function using malaria.tools, we entered GO:0020009 (GO term for microneme) to arrive at the page dedicated to microneme cellular component (https://malaria.tools/go/view/12866). The page revealed 270 and 299 annotated microneme-associated genes in *P. berghei* and *P. falciparum*, respectively. Furthermore, the page contains information about Pfam domains (Prot_kinase_dom, VWF_A, MORN and others) and gene families (OG_01_0000012, OG_01_0000038 and others), which may also play a role in the microneme function. Furthermore, the database identified cluster 2 (https://malaria.tools/cluster/view/31) and cluster 7

(https://malaria.tools/cluster/view/33) from *P. berghei* and cluster 2 (https://malaria.tools/cluster/view/74) and cluster 12 (https://malaria.tools/cluster/view/146) from *P. falciparum* as being significantly similar (P<0.05, Benjamini-Hochenberg corrected p-value, ref), implicating these clusters in a microneme-specific process.

To learn more about the function of these four clusters, we investigated their expression profiles. While cluster 2 from *P. berghei* and cluster 12 from *P. falciparum* show ubiquitous expression at all life stages of malaria (Figure 3A), cluster 7 from *P.berghei* and cluster 2 from *P. falciparum* show ookinete- and sporozoite-specific expression respectively. Since ookinetes and sporozoites are mosquito stage-specific (ref), we speculate that the two *Plasmodium* species have at least two types of micronemes, one being ubiquitously expressed (clusters 2 and 12) and another being mosquito-specific (clusters 7 and 2).

We further investigated the putative function of the ookinete-specific cluster 7 from *P. berghei* (https://malaria.tools/cluster/view/33), and found that the cluster is significantly enriched for GO terms such as entry into host cell (GO:0030260, Table S8), which is in line with the microneme being involved in invasion of blood cells. A closer look at the genes found in cluster 7 from *P. berghei* revealed that it contains genes that are essential for the infectivity and maturation of ookinetes (Figure 3B). Specifically, a group of genes relating to the inner membrane complex, surface ookinete protein and secreted ookinete protein is required for efficient gliding motility and midgut traversal [*PBANKA_1354600* (ISC1), *PBANKA_1025700* (IMC1l), *PBANKA_0513000* (IMC1m) (36), *PBANKA_1106900* (PIMMS2) (37), *PBANKA_0714300* (HSP20) (38) and *PBANKA_1432300* (CelTOS) (39)]. Another group of co-expressed genes contains perforins and a secreted protein, which are known to be important in midgut invasion where the parasite disrupts the membrane of the endothelial cell [*PBANKA_0824200* (PLP3), *PBANKA_0711400* (PLP4), *PBANKA_0711600* (PLP5) (40) and *PBANKA_1037800* (SOAP) (41)]. Last but not least, genes important for the transition from the ookinete to oocyst stage are also found enriched in this cluster [*PBANKA_0701900* (GAMA/PSOP9) (42), *PBANKA_1119200* (PSOP25) (43) and *PBANKA_0412900* (CTRP) (44)]. Taken together, cluster 7 from *P. berghei* contain genes that are most likely important for microneme function and host cell invasion.

CONCLUSIONS

The lack of comparative genomic and transcriptomic resources for malaria prompted us to construct malaria.tools, a state-of-the-art database containing a wide range of user-friendly features. The database can be queried by BLAST, gene IDs, keywords and Pfam domains and Gene Ontology searches. To identify novel genes relevant for a biological process of interest, the co-expression neighborhoods and clusters can be mined for uncharacterized candidates that are connected to well-studied genes. Alternatively, the database allows an easy identification of genes that are expressed during a specific life stage of the malaria parasite, thus allowing researchers to dissect the transcriptome critical for pathogenicity and other life stages. Finally, the database can compare the clusters and stage-specific expression profiles to identify the conserved core components of various biological processes. We

envision that malaria.tools will aid malaria researchers in selecting relevant genes for experimental functional characterisation and potential drug development for successful combating the emerging drug resistances.

DATA AVAILABILITY

The expression matrices, RNAseq sample annotation and gene families are available from the supplementary material. The co-expression networks, coding and protein sequences can be downloaded from malaria.tools.

SUPPLEMENTARY DATA

The Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Malaria.tools is hosted at Nanyang Technological University Singapore and we would like to thank Ryan Chee Kiang Ng for excellent tech support. Furthermore, we would like to thank Dr. Daniela Mutwil-Anderwald for proofreading the manuscript.

Author Contributions: Malaria.tools was implemented by Q.W.T. who also prepared the data and built malaria.tools with input from M.M. Both Q.W.T. and M.M. wrote the manuscript.

FUNDING

We would like to thank Nanyang Technological University Start-Up Grant for funding.

Conflict of interest statement. None declared.

REFERENCES

1. World Health Organisation (2018) World Malaria Report 2018 Geneva.
2. Thu,A.M., Phyo,A.P., Landier,J., Parker,D.M. and Nosten,F.H. (2017) Combating multidrug-resistant Plasmodium falciparum malaria. *FEBS J.*, **284**, 2569–2578.
3. Florent,I., Maréchal,E., Gascuel,O. and Bréhélin,L. (2010) Bioinformatic strategies to provide functional clues to the unknown genes in Plasmodium falciparum genome.

4. Zhou,X., Kao,M.-C.J. and Wong,W.H. (2002) Transitive functional annotation by shortest-path analysis of gene expression data. *Proc. Natl. Acad. Sci. U. S. A.*, **99**, 12783–12788.

5. Proost,S. and Mutwil,M. (2018) CoNekT: an open-source framework for comparative genomic and transcriptomic network analyses. *bioRxiv*, 10.1101/255075.

6. Silvester,N., Alako,B., Amid,C., Cerdeño-Tarrága,A., Clarke,L., Cleland,I., Harrison,P.W., Jayathilaka,S., Kay,S., Keane,T., *et al.* (2017) The European Nucleotide Archive in 2017. *Nucleic Acids Res.*, **46**, D36–D40.

7. Bray,N.L., Pimentel,H., Melsted,P. and Pachter,L. (2016) Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, **34**, 525–527.

8. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

9. Mutwil,M., Usadel,B., Schütte,M., Loraine,A., Ebenhöh,O. and Persson,S. (2010) Assembly of an interactive correlation network for the Arabidopsis genome using a novel heuristic clustering algorithm. *Plant Physiol.*, **152**, 29–43.

10. Jones,P., Binns,D., Chang,H.-Y., Fraser,M., Li,W., McAnulla,C., McWilliam,H., Maslen,J., Mitchell,A., Nuka,G., *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.

11. Emms,D.M. and Kelly,S. (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.*, **16**, 157.

12. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

13. Yeoman,J.A., Hanssen,E., Maier,A.G., Klonis,N., Maco,B., Baum,J., Turnbull,L., Whitchurch,C.B., Dixon,M.W.A. and Tilley,L. (2011) Tracking Glideosome-Associated Protein 50 Reveals the Development and Organization of the Inner Membrane Complex of Plasmodium falciparum. *Eukaryot. Cell*, **10**, 556–564.

14. Bullen,H.E., Tonkin,C.J., O'Donnell,R.A., Tham,W.-H., Papenfuss,A.T., Gould,S., Cowman,A.F., Crabb,B.S. and Gilson,P.R. (2009) A Novel Family of Apicomplexan Glideosome-associated Proteins with an Inner Membrane-anchoring Role. *J. Biol. Chem.* , **284**, 25353–25363.

15. Saini,E., Zeeshan,M., Brady,D., Pandey,R., Kaiser,G., Koreny,L., Kumar,P., Thakur,V., Tatiya,S., Katris,N.J., *et al.* (2017) Photosensitized INA-Labelled protein 1 (PhIL1) is novel component of the inner membrane complex and is required for Plasmodium parasite development. *Sci. Rep.*, **7**, 15577.

16. Olshina,M.A., Angrisano,F., Marapana,D.S., Riglar,D.T., Bane,K., Wong,W., Catimel,B., Yin,M.-X., Holmes,A.B., Frischknecht,F., *et al.* (2015) Plasmodium falciparum coronin organizes arrays of parallel actin filaments potentially guiding directional motility in invasive malaria parasites. *Malar. J.*, **14**, 280.

17. Moreau,C.A., Bhargav,S.P., Kumar,H., Quadt,K.A., Piirainen,H., Strauss,L., Kehrer,J., Streichfuss,M., Spatz,J.P., Wade,R.C., *et al.* (2017) A unique profilin-actin interface is important for malaria parasite motility. *PLoS Pathog.*, **13**, e1006412–e1006412.

18. Beeson,J.G., Drew,D.R., Boyle,M.J., Feng,G., Fowkes,F.J.I. and Richards,J.S. (2016) Merozoite surface proteins in red blood cell invasion, immunity and vaccines against malaria. *FEMS Microbiol. Rev.*, **40**, 343–372.
19. Withers-Martinez,C., Suarez,C., Fulle,S., Kher,S., Penzo,M., Ebejer,J.-P., Koussis,K., Hackett,F., Jirgensons,A., Finn,P., *et al.* (2012) Plasmodium subtilisin-like protease 1 (SUB1): insights into the active-site structure, specificity and function of a pan-malaria drug target. *Int. J. Parasitol.*, **42**, 597–612.
20. Dorin-Semblat,D., Demarta-Gatsi,C., Hamelin,R., Armand,F., Carvalho,T.G., Moniatte,M. and Doerig,C. (2015) Malaria Parasite-Infected Erythrocytes Secrete PfCK1, the Plasmodium Homologue of the Pleiotropic Protein Kinase Casein Kinase 1. *PLoS One*, **10**, e0139591.
21. Lehmann,C., Tan,M.S.Y., de Vries,L.E., Russo,I., Sanchez,M.I., Goldberg,D.E. and Deu,E. (2018) Plasmodium falciparum dipeptidyl aminopeptidase 3 activity is important for efficient erythrocyte invasion by the malaria parasite. *PLoS Pathog.*, **14**, e1007031.
22. Baker,D.A., Drought,L.G., Flueck,C., Nofal,S.D., Patel,A., Penzo,M. and Walker,E.M. (2017) Cyclic nucleotide signalling in malaria parasites. *Open Biol.*, **7**, 170213.
23. Counihan,N.A., Kalanon,M., Coppel,R.L. and De Koning-Ward,T.F. (2013) Plasmodium rhoptry proteins: Why order is important. *Trends Parasitol.*, **29**, 228–236.
24. Counihan,N.A., Chisholm,S.A., Bullen,H.E., Srivastava,A., Sanders,P.R., Jonsdottir,T.K., Weiss,G.E., Ghosh,S., Crabb,B.S., Creek,D.J., *et al.* (2017) Plasmodium falciparum parasites deploy RhopH2 into the host erythrocyte to obtain nutrients, grow and replicate. *Elife*, **6**.
25. Favuzza,P., Guffart,E., Tamborrini,M., Scherer,B., Dreyer,A.M., Rufer,A.C., Erny,J., Hoernschemeyer,J., Thoma,R., Schmid,G., *et al.* (2017) Structure of the malaria vaccine candidate antigen CyRPA and its complex with a parasite invasion inhibitory antibody. *Elife*, **6**.
26. Ling,I.T., Florens,L., Dluzewski,A.R., Kaneko,O., Grainger,M., Yim Lim,B.Y.S., Tsuboi,T., Hopkins,J.M., Johnson,J.R., Torii,M., *et al.* (2004) The Plasmodium falciparum clag9 gene encodes a rhoptry protein that is transferred to the host erythrocyte upon invasion. *Mol. Microbiol.*, **52**, 107–118.
27. Tonkin,M.L., Arredondo,S.A., Loveless,B.C., Serpa,J.J., Makepeace,K.A.T., Sundar,N., Petrotchenko,E. V, Miller,L.H., Grigg,M.E. and Boulanger,M.J. (2013) Structural and biochemical characterization of Plasmodium falciparum 12 (Pf12) reveals a unique interdomain organization and the potential for an antiparallel arrangement with Pf41. *J. Biol. Chem.*, **288**, 12805–12817.
28. Mutwil,M., Klie,S., Tohge,T., Giorgi,F.M., Wilkins,O., Campbell,M.M., Fernie,A.R., Usadel,B., Nikoloski,Z. and Persson,S. (2011) PlaNet: Combined Sequence and Expression Comparisons across Plant Networks Derived from Seven Species. *Plant Cell*, **23**, 895–910.
29. Movahedi,S., Van Bel,M., Heyndrickx,K.S. and Vandepoele,K. (2012) Comparative co-expression analysis in plant biology. *Plant, Cell Environ.*, **35**, 1787–1798.
30. Hansen,B.O., Vaid,N., Musialak-Lange,M., Janowski,M. and Mutwil,M. (2014) Elucidating gene function and function evolution through comparison of co-expression networks of plants. *Front. Plant Sci.*, **5**, 1–9.

31. Xiao,S.J., Zhang,C., Zou,Q. and Ji,Z.L. (2010) TiSGeD: A database for tissue-specific genes. *Bioinformatics*, **26**, 1273–1275.

32. Talman,A.M., Prieto,J.H., Marques,S., Ubaida-Mohien,C., Lawniczak,M., Wass,M.N., Xu,T., Frank,R., Ecker,A., Stanway,R.S., *et al.* (2014) Proteomic analysis of the Plasmodium male gamete reveals the key role for glycolysis in flagellar motility. *Malar. J.*, **13**, 315.

33. Yeh,S.-D., Chen,Y.-J., Chang,A.C.Y., Ray,R., She,B.-R., Lee,W.-S., Chiang,H.-S., Cohen,S.N. and Lin-Chao,S. (2002) Isolation and properties of Gas8, a growth arrest-specific gene regulated during male gametogenesis to produce a protein associated with the sperm motility apparatus. *J. Biol. Chem.*, **277**, 6311–7.

34. Ferguson,D.J.P., Sahoo,N., Pinches,R.A., Bumstead,J.M., Tomley,F.M. and Gubbels,M.-J. (2008) MORN1 Has a Conserved Role in Asexual and Sexual Development across the Apicomplexa. *Eukaryot. Cell*, **7**, 698–711.

35. Black,M.W. and Boothroyd,J.C. (2000) Lytic cycle of Toxoplasma gondii. *Microbiol. Mol. Biol. Rev.*, **64**, 607–623.

36. Harding,C.R. and Meissner,M. (2014) The inner membrane complex through development of Toxoplasma gondii and Plasmodium. *Cell. Microbiol.*, **16**, 632–641.

37. Ukegbu,C. V, Akinosoglou,K.A., Christophides,G.K. and Vlachou,D. (2017) Plasmodium berghei PIMMS2 Promotes Ookinete Invasion of the Anopheles gambiae Mosquito Midgut. *Infect. Immun.*, **85**, e00139-17.

38. Montagna,G.N., Buscaglia,C.A., Münter,S., Goosmann,C., Frischknecht,F., Brinkmann,V. and Matuschewski,K. (2012) Critical Role for Heat Shock Protein 20 (HSP20) in Migration of Malarial Sporozoites. *J. Biol. Chem.*, **287**, 2410–2422.

39. Kariu,T., Ishino,T., Yano,K., Chinzei,Y. and Yuda,M. (2006) CelTOS, a novel malarial protein that mediates transmission to mosquito and vertebrate hosts. *Mol. Microbiol.*, **59**, 1369–1379.

40. Deligianni,E., Silmon de Monerri,N.C., McMillan,P.J., Bertuccini,L., Superti,F., Manola,M., Spanos,L., Louis,C., Blackman,M.J., Tilley,L., *et al.* (2018) Essential role of Plasmodium perforin-like protein 4 in ookinete midgut passage. *PLoS One*, **13**, e0201651.

41. Dessens,J.T., Siden-Kiamos,I., Mendoza,J., Mahairaki,V., Khater,E., Vlachou,D., Xu,X.-J., Kafatos,F.C., Louis,C., Dimopoulos,G., *et al.* (2003) SOAP, a novel malaria ookinete protein involved in mosquito midgut invasion and oocyst development. *Mol. Microbiol.*, **49**, 319–329.

42. Ecker,A., Bushell,E.S.C., Tewari,R. and Sinden,R.E. (2008) Reverse genetics screen identifies six proteins important for malaria development in the mosquito. *Mol. Microbiol.*, **70**, 209–220.

43. Zheng,W., Liu,F., He,Y., Liu,Q., Humphreys,G.B., Tsuboi,T., Fan,Q., Luo,E., Cao,Y. and Cui,L. (2017) Functional characterization of Plasmodium berghei PSOP25 during ookinete development and as a malaria transmission-blocking vaccine candidate. *Parasit. Vectors*, **10**, 8.

44. Ramakrishnan,C., Dessens,J.T., Armson,R., Pinto,S.B., Talman,A.M., Blagborough,A.M. and Sinden,R.E. (2011) Vital functions of the malarial ookinete protein, CTRP, reside in the A domains. *Int. J.*

Parasitol., **41**, 1029–1039.

TABLE AND FIGURES LEGENDS

Table 1. Source of data for species in database.

Species	RNA-seq samples	CDS	CDS description	GO annotation
Theilera parva Muguga	N/A	NCBI Ref-seq	NCBI Ref-seq	Interproscan
Toxoplasma gondii ME49	N/A	NCBI Ref-seq	NCBI Ref-seq	Interproscan
Plasmodium berghei ANKA	206	PlasmoDB	GeneDB	PlasmoDB
Plasmodium cynomolgi B	N/A	PlasmoDB	Interproscan	Interproscan
Plasmodium falciparum 3D7	620	PlasmoDB	GeneDB	PlasmoDB
Plasmodium knowlesi Malayan Pk1A	N/A	PlasmoDB	Interproscan	Interproscan
Plasmodium vivax P01	N/A	PlasmoDB	GeneDB	Interproscan
Plasmodium yoelii YM	N/A	PlasmoDB	Interproscan	Interproscan

Figure 1. Expression profiles and co-expression neighborhood of gene *PF3D7_1223100*. A) Full expression profile of the gene. The x-axis represents the different RNA-seq experiments capturing the life stages and genetic perturbations of *P. falciparum*, while the y-axis indicates the expression level (Transcripts Per Million, TPM). The different life stages are color-coded by blue (ring), purple (trophozoite), green (schizont), light blue (male gametocyte), pink (female gametocyte) and brown (sporozoite). The bars indicate the mean expression value, while the dots show the expression values of the individual samples. For brevity, only the general descriptions of the samples are shown. B) Simplified expression profile of *PF3D7_1223100*, showing the average expression in the five major life stages. C) Co-expression network neighborhood containing functionally characterized genes. Nodes represent genes, edges (lines) connect co-expressed genes, while colored shapes indicate orthogroups. For brevity, only genes with experimentally verified function supported by at least two publications are shown in the figure.

Figure 2. Comparative analysis of male gametocyte-specific gene expression in *P. berghei* and *P. falciparum*. A) Venn diagram showing the overlap of the male gametophyte-specific gene families in the two malaria species. B) Phylogenetic gene tree of gene family OG0109400. The different species are color coded and represent *Toxoplasma gondii* in red (gene IDs XP_NNNNNNNNN, NP_XXXXXX), *Theilera parva* in orange (gene IDs XP_NNNNNN), *P. falciparum* in olive (gene IDs PF3D7_XXXXXXX),

P. berghei in dark green (gene IDs PBANKA_XXXXXXX), *P. yoleii* in light green (gene IDs PYYM_XXXXXXX), *P. knowlesi* in blue (gene IDs PKNOH_SXXXXXXXXX-t35_1), *P. vivax* in dark blue (gene IDs PVP01_XXXXXXX) and *P. cynomolgi* in purple (gene IDs PCYB_XXXXXX-t26_1). The colored boxes to the right of the gene IDs show the average gene expression in five major life stages of the malaria parasite, where yellow and blue color indicates low and high gene expression. C) Phylogenetic gene tree of gene family OG0000055. D) Comparison of cluster 13 (left blue box) and cluster 15 (right green box) from *P. berghei* and *P. falciparum*, respectively. Nodes represent, genes, solid edges connect co-expressed genes, dashed edges connect orthologs, while colored shapes indicate genes belonging to the same gene families. E) Average expression profiles of the genes found in cluster 13 (left) and 15 (right).

Figure 3. Expression profiles and co-expression network of the ookinete-enriched clusters in *P. falciparum* and *P. berghei*. A) Expression profiles of clusters 2 (first) and 7 (second) from *P. berghei* and clusters 2 (third) and 12 (fourth) from *P. falciparum*. The different life stages are color-coded. For brevity, the sample annotations are abbreviated to the major life stages. B) Co-expression cluster 7 from *P. berghei.* Nodes represent genes, co-expressed genes are connected by gray edges, while colored shapes indicate orthogroups. For brevity, only the discussed genes are highlighted.

Table S1. TPM expression matrix for *Plasmodium berghei*. Genes are found in rows, while samples are found in columns.

Table S2. TPM expression matrix for *Plasmodium falciparum*. Genes are found in rows, while samples are found in columns.

Table S3. Annotation of RNAseq samples for *Plasmodium berghei*.

Table S4. Annotation of RNAseq samples for *Plasmodium falciparum*.

Table S5. Orthogroups of the eight Apicomplexa species included in this study. Each row represents one orthogroup and the genes belonging to the orthogroup.

Table S6. Male gametocyte specific genes of *Plasmodium falciparum* and *Plasmodium berghei*.

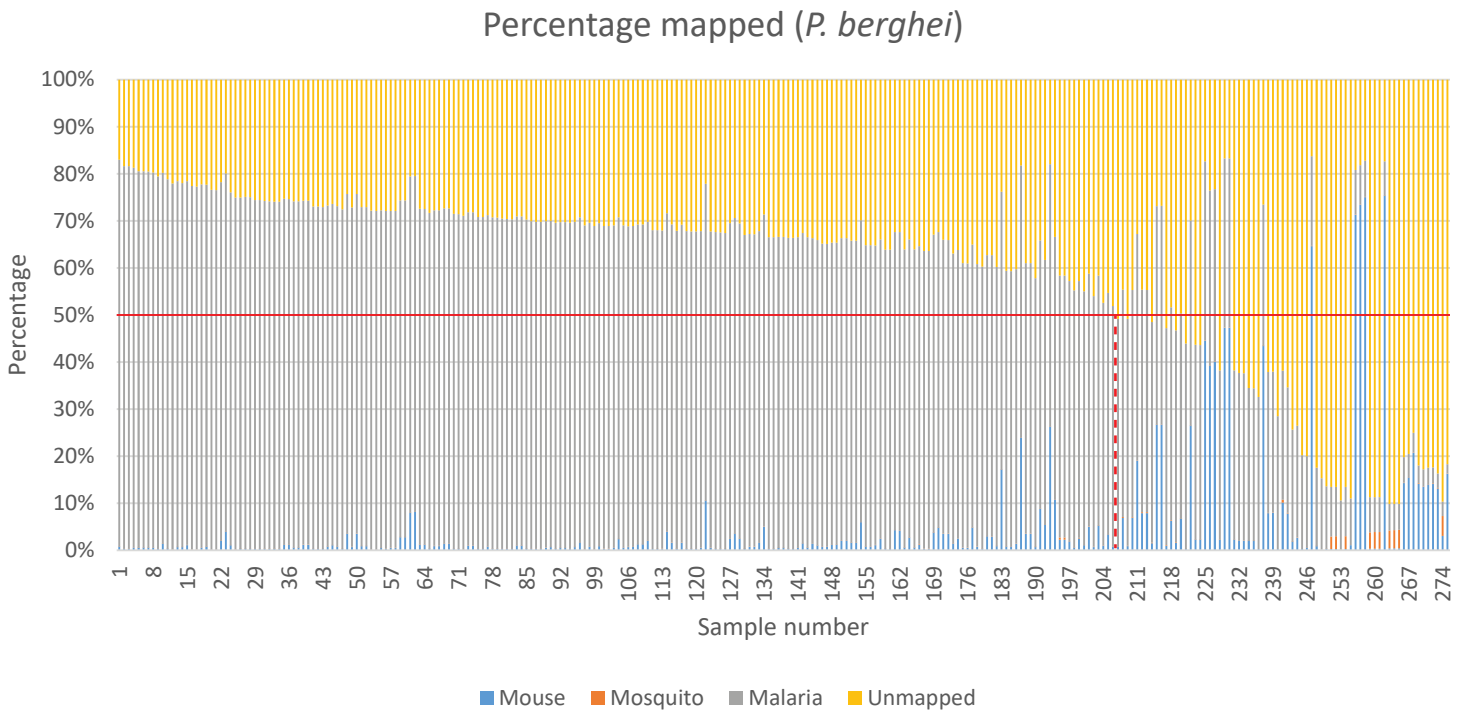
Table S7. Annotation of comparative clusters *Plasmodium berghei* cluster 13 and *Plasmodium falciparum* cluster 15

Table S8. Enriched GO terms of *P. berghei* HCCA cluster 7 and genes associated with it.

Figure S1. Overview of the sample quality of the RNA-seq samples that contain at least 1 million reads. A) *Plasmodium berghei*. For each sample, the percentage of reads mapping to the parasite (grey), mouse or human (blue), mosquito (orange) and not mapping to either (yellow) is shown. The sample cutoff is indicated by the red line. B) Overview of the sample quality of *Plasmodium falciparum*.

Figure S1

A


Percentage mapped (*P. berghei*)
Percentage
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1 8 15 22 29 36 43 50 57 64 71 78 85 92 99 106 113 120 127 134 141 148 155 162 169 176 183 190 197 204 211 218 225 232 239 246 253 260 267 274
Sample number
Mouse
Mosquito
Malaria
Unmapped

B

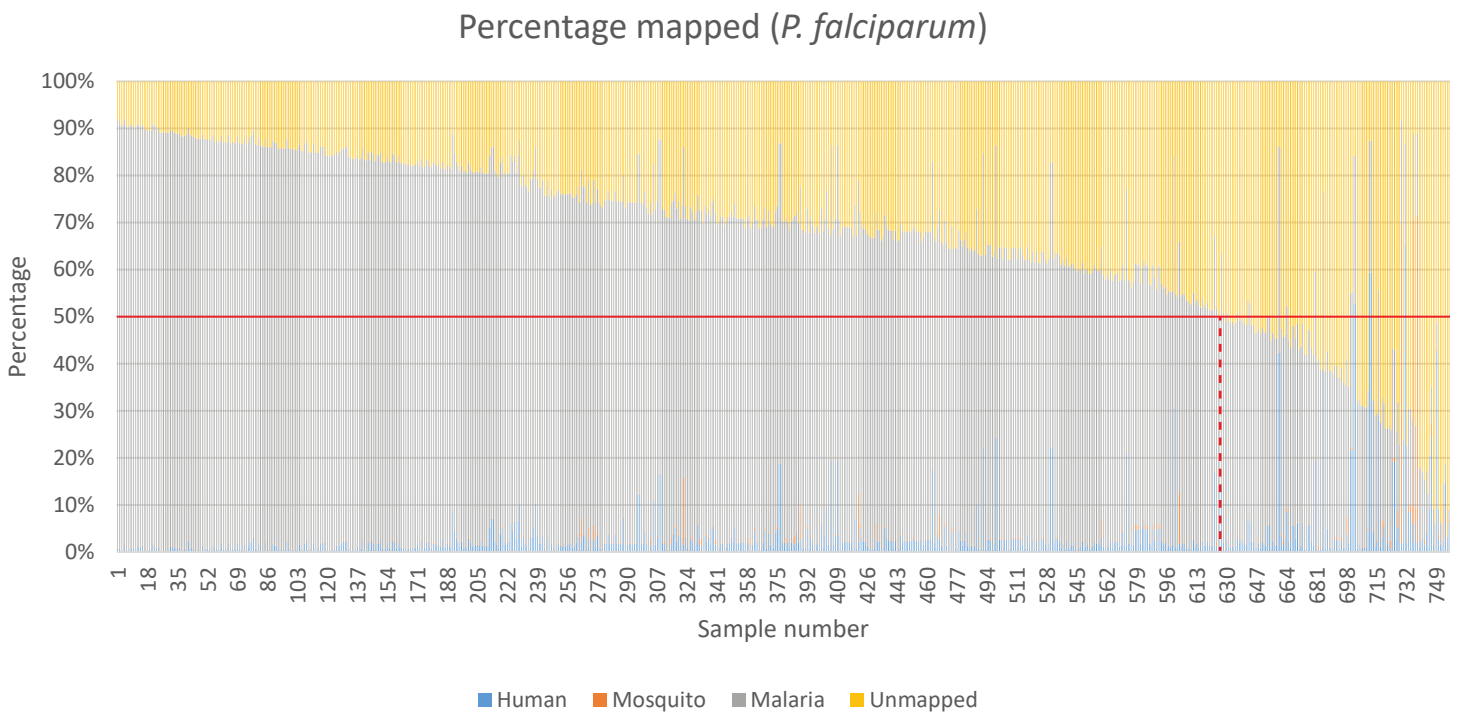

Percentage mapped (*P. falciparum*)
Percentage
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1 18 35 52 69 86 103 120 137 154 171 188 205 222 239 256 273 290 307 324 341 358 375 392 409 426 443 460 477 494 511 528 545 562 579 596 613 630 647 664 681 698 715 732 749
Sample number
Human
Mosquito
Malaria
Unmapped

Figure 1

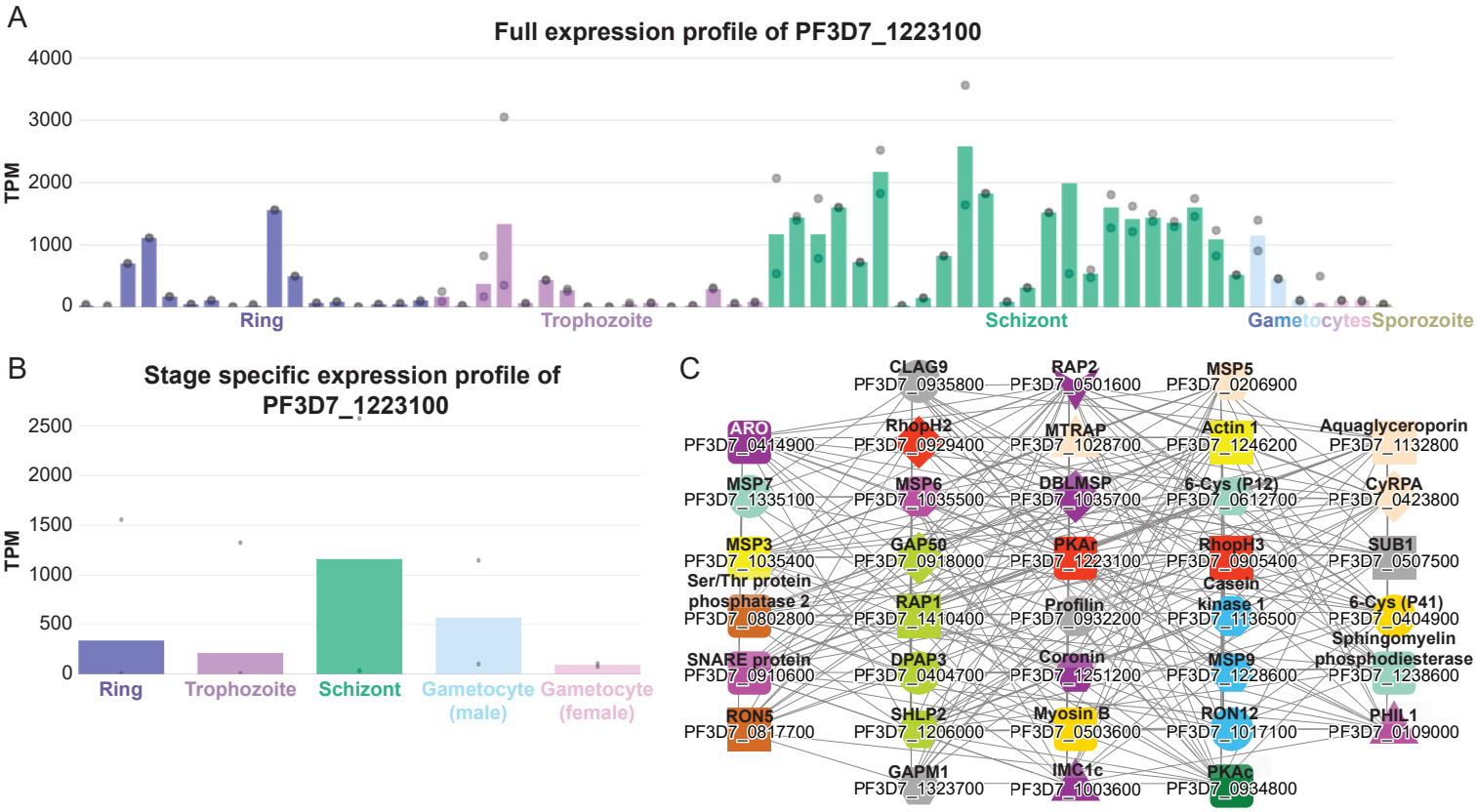

A
Full expression profile of PF3D7_1223100
TPM
4000
3000
2000
1000
0
Ring
Trophozoite
Schizont
Gametocytes
Sporozoite
B
Stage specific expression profile of PF3D7_1223100
TPM
2500
2000
1500
1000
500
0
Ring
Trophozoite
Schizont
Gametocyte (male)
Gametocyte (female)
C
CLAG9 PF3D7_0935800
RAP2 PF3D7_0501600
MSP5 PF3D7_0206900
ARO PF3D7_0414900
RhopH2 PF3D7_0929400
MTRAP PF3D7_1028700
Actin 1 PF3D7_1246200
Aquaglyceroporin PF3D7_1132800
MSP7 PF3D7_1335100
MSP6 PF3D7_1035500
DBLMSP PF3D7_1035700
6-Cys (P12) PF3D7_0612700
CyRPA PF3D7_0423800
MSP3 PF3D7_1035400
GAP50 PF3D7_0918000
PKAr PF3D7_1223100
RhopH3 PF3D7_0905400
SUB1 PF3D7_0507500
Ser/Thr protein phosphatase 2 PF3D7_0802800
RAP1 PF3D7_1410400
Profilin PF3D7_0932200
Casein kinase 1 PF3D7_1136500
6-Cys (P41) PF3D7_0404900
SNARE protein PF3D7_0910600
DPAP3 PF3D7_0404700
Coronin PF3D7_1251200
MSP9 PF3D7_1228600
Sphingomyelin phosphodiesterase PF3D7_1238600
RON5 PF3D7_0817700
SHLP2 PF3D7_1206000
Myosin B PF3D7_0503600
RON12 PF3D7_1017100
PHIL1 PF3D7_0109000
GAPM1 PF3D7_1323700
IMC1c PF3D7_1003600
PKAc PF3D7_0934800

Figure 2

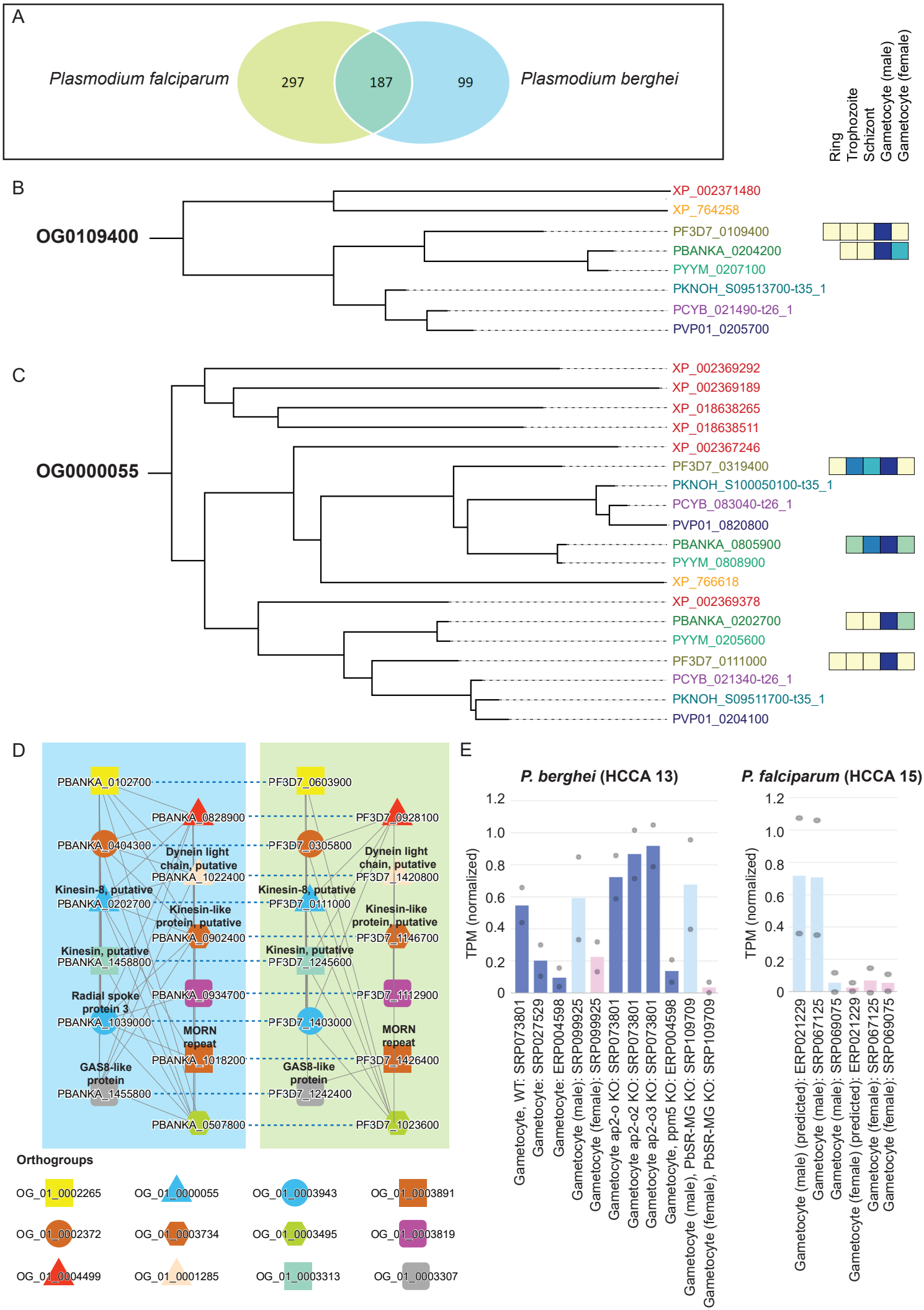
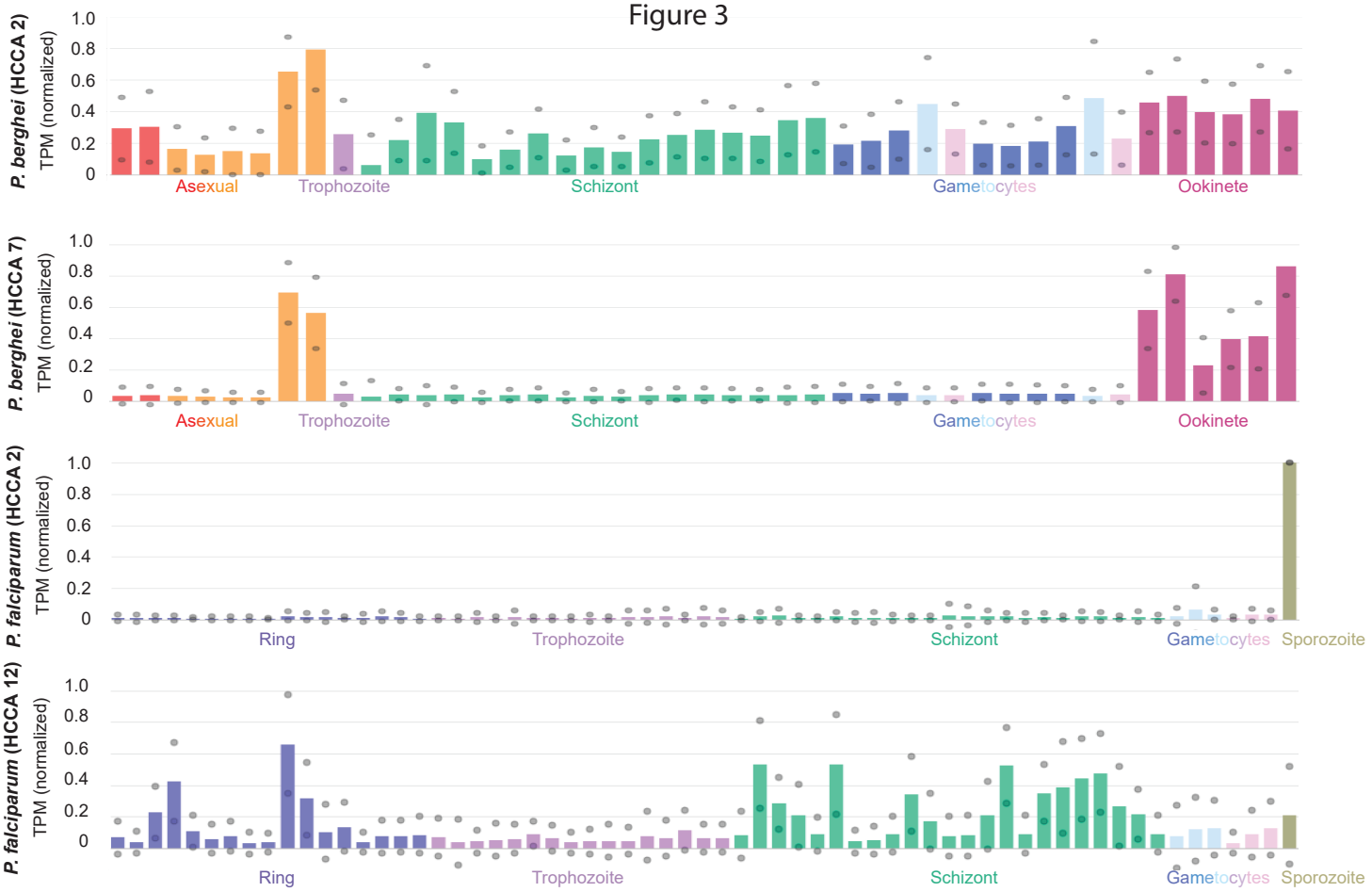
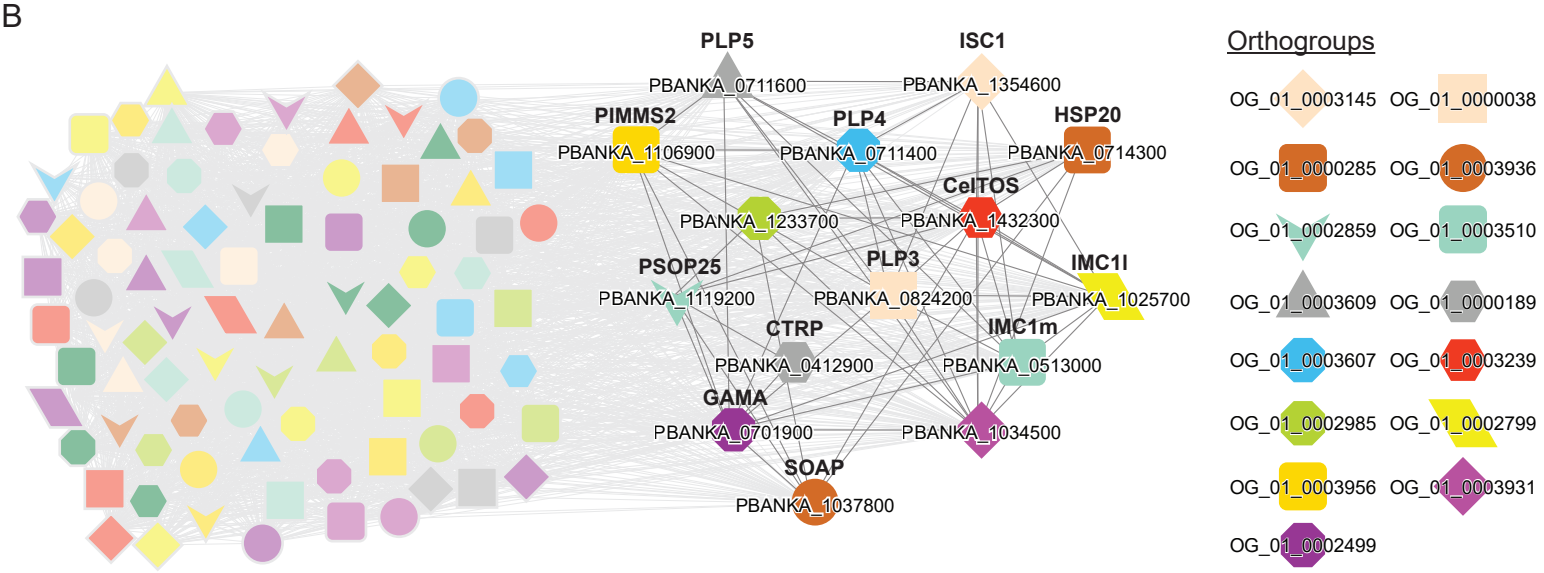

A
Plasmodium falciparum
297
187
99
Plasmodium berghei
Ring
Trophozoite
Schizont
Gametocyte (male)
Gametocyte (female)
B
OG0109400
XP_002371480
XP_764258
PF3D7_0109400
PBANKA_0204200
PYYM_0207100
PKNOH_S09513700-t35_1
PCYB_021490-t26_1
PVP01_0205700
C
OG0000055
XP_002369292
XP_002369189
XP_018638265
XP_018638511
XP_002367246
PF3D7_0319400
PKNOH_S100050100-t35_1
PCYB_083040-t26_1
PVP01_0820800
PBANKA_0805900
PYYM_0808900
XP_766618
XP_002369378
PBANKA_0202700
PYYM_0205600
PF3D7_0111000
PCYB_021340-t26_1
PKNOH_S09511700-t35_1
PVP01_0204100
D
PBANKA_0102700
PF3D7_0603900
PBANKA_0828900
PF3D7_0928100
PBANKA_0404300
PF3D7_0305800
Dynein light chain, putative
PBANKA_1022400
PF3D7_1420800
Kinesin-8, putative
PBANKA_0202700
PF3D7_0111000
Kinesin-like protein, putative
PBANKA_0902400
PF3D7_1146700
Kinesin, putative
PBANKA_1458800
PF3D7_1245600
Radial spoke protein 3
PBANKA_0934700
PF3D7_1112900
PBANKA_1039000
PF3D7_1403000
MORN repeat
PBANKA_1018200
PF3D7_1426400
GAS8-like protein
PBANKA_1455800
PF3D7_1242400
PBANKA_0507800
PF3D7_1023600
Orthogroups
OG_01_0002265
OG_01_0000055
OG_01_0003943
OG_01_0003891
OG_01_0002372
OG_01_0003734
OG_01_0003495
OG_01_0003819
OG_01_0004499
OG_01_0001285
OG_01_0003313
OG_01_0003307
E
P. berghei (HCCA 13)
P. falciparum (HCCA 15)
TPM (normalized)
1.2
1.0
0.8
0.6
0.4
0.2
0
Gametocyte, WT: SRP073801
Gametocyte: SRP027529
Gametocyte: ERP004598
Gametocyte (male): SRP099925
Gametocyte (female): SRP099925
Gametocyte ap2-o KO: SRP073801
Gametocyte ap2-o2 KO: SRP073801
Gametocyte ap2-o3 KO: SRP073801
Gametocyte, ppm5 KO: ERP004598
Gametocyte (male), PbSR-MG KO: SRP109709
Gametocyte (female), PbSR-MG KO: SRP109709
Gametocyte (male) (predicted): ERP021229
Gametocyte (male): SRP067125
Gametocyte (male): SRP069075
Gametocyte (female) (predicted): ERP021229
Gametocyte (female): SRP067125
Gametocyte (female): SRP069075

Figure 3

A


P. berghei (HCCA 2)
TPM (normalized)
1.0
0.8
0.6
0.4
0.2
0
Asexual
Trophozoite
Schizont
Gametocytes
Ookinete
P. berghei (HCCA 7)
TPM (normalized)
Asexual
Trophozoite
Schizont
Gametocytes
Ookinete
P. falciparum (HCCA 2)
TPM (normalized)
Ring
Trophozoite
Schizont
Gametocytes
Sporozoite
P. falciparum (HCCA 12)
TPM (normalized)
Ring
Trophozoite
Schizont
Gametocytes
Sporozoite

B


PLP5
PBANKA_0711600
ISC1
PBANKA_1354600
PIMMS2
PBANKA_1106900
PLP4
PBANKA_0711400
HSP20
PBANKA_0714300
CelTOS
PBANKA_1233700
PBANKA_1432300
PSOP25
PBANKA_1119200
PLP3
PBANKA_0824200
IMC1l
PBANKA_1025700
CTRP
PBANKA_0412900
IMC1m
PBANKA_0513000
GAMA
PBANKA_0701900
PBANKA_1034500
SOAP
PBANKA_1037800
Orthogroups
OG_01_0003145
OG_01_0000038
OG_01_0000285
OG_01_0003936
OG_01_0002859
OG_01_0003510
OG_01_0003609
OG_01_0000189
OG_01_0003607
OG_01_0003239
OG_01_0002985
OG_01_0002799
OG_01_0003956
OG_01_0003931
OG_01_0002499