

***Plasmodium malariae* and *ovale* genomes provide insights into malaria parasite evolution**

Gavin G. Rutledge¹, Ulrike Böhme¹, Mandy Sanders¹, Adam J. Reid¹, James A. Cotton¹, Oumou Maiga-Ascofare^{2,3}, Abdoulaye A. Djimdé^{1,2}, Tobias O. Apinjoh⁴, Lucas Amenga-Etego⁵, Magnus Manske¹, John W. Barnwell⁶, François Renaud⁷, Benjamin Ollomo⁸, Franck Prugnolle^{7,8}, Nicholas M. Anstey⁹, Sarah Auburn⁹, Ric N. Price^{9,10}, James S. McCarthy¹¹, Dominic P. Kwiatkowski^{1,12}, Chris I. Newbold^{1,13}, Matthew Berriman^{1*} & Thomas D. Otto^{1*}

¹ Malaria Programme, Wellcome Trust Sanger Institute, Hinxton, Cambridge, United Kingdom

² Malaria Research and Training Center, University of Science, Techniques, and Technologies of Bamako, Bamako, Mali

³ Bernhard-Nocht Institute for Tropical Medicine, Hamburg, Germany

⁴ University of Buea, Buea, Cameroon

⁵ Navrongo Health Research Centre, Navrongo, Ghana

⁶ Centers for Disease Control and Prevention, Atlanta, Georgia, United States

⁷ Laboratoire MIVEGEC (UM1-CNRS-IRD), Montpellier, France

⁸ Centre International de Recherches médicales de Franceville, Gabon

⁹ Global and Tropical Health Division, Menzies School of Health Research and Charles Darwin University, Darwin, NT, Australia

¹⁰ Centre for Tropical Medicine and Global Health, Nuffield Department of Clinical Medicine, University of Oxford, UK

¹¹ Clinical Tropical Medicine Laboratory, QIMR Berghofer Medical Research Institute, University of Queensland, Brisbane, Australia

¹² Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom

¹³ Weatherall Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Oxford, United Kingdom

Elucidating the evolutionary history and interrelatedness of human-infective *Plasmodium* species has been hampered by a lack of genetic information for three human-infective species: *P. malariae* and two *P. ovale* species (*P. o. curtisi* and *P. o. wallikeri*)¹. These species are prevalent across most malaria endemic regions^{2,3} (Figure 1) and are often undetectable by light microscopy⁴, making their study in human populations difficult⁵. The exact evolutionary relationship of these species to the other human-infective species has been contested^{6,7}. Using a new reference genome for *P. malariae* as well as a manually curated draft *P. o. curtisi* genome, we are now able to accurately place these species within the *Plasmodium* phylogeny. Sequencing of a chimpanzee-infective *P. malariae* relative reveals similar signatures of

selection in the *P. malariae* lineage to another *Plasmodium* lineage where colonisation of both human and chimpanzee hosts has occurred. Molecular dating suggests that these host adaptations occurred over similar evolutionary timescales. In addition to the core genome that is conserved between species, differences in gene content can be linked to their specific biology. The genome suggests that *P. malariae* expresses a novel family of heterodimeric proteins on its surface that have structural similarities to a protein crucial for red blood cell invasion. The data presented here provide an essential foundation into understanding the evolution of the *Plasmodium* genus as a whole.

A reference genome of *P. malariae* was produced from clinically isolated parasites and sequenced using long-read sequencing technology (Table 1; Supplementary Information). The assembly significantly surpasses available draft genome data for *P. malariae*⁷, especially in terms of contiguity (63 vs 7270 scaffolds, N50 = 2.3 Mb vs. 6.4 kb) (Supplementary Information; Extended Data Table 1) allowing large-scale structural changes to be accurately determined. Against a background of near-complete collinearity with *P. vivax*, we found a previously undescribed large reciprocal translocation between chromosomes 6 and 10 and a pericentric inversion of chromosome 5 (Extended Data Fig 1a,b). Additional draft genomes for both species of *P. ovale* were assembled from *P. falciparum* co-infections and the genome of a parasite that we call *P. malariae-like* was assembled from a chimpanzee co-infected with *P. reichenowi* (Figure 1; Table 1; Extended Data Table 2; Supplementary Information).

To investigate host-specific adaptation of parasites to human and chimpanzee hosts, we compared *P. malariae* to *P. malariae-like*. We find lower levels of nucleotide diversity in the human-infective than in the chimpanzee-infective species (Table 1; Extended Data Fig 2a). This mirrors the situation where the human parasite *P. falciparum* has lower levels of nucleotide diversity than its chimpanzee-infective relative *P. reichenowi*⁸. In both cases, the lack of diversity in human infective species suggests recent population expansions. However, we find a New World primate-infective species termed *P. brasilianum* to be indistinguishable from *P. malariae* (Extended Data Fig 2a), as previously suggested⁹. Thus host adaptation in the *P. malariae* lineage appears to be less restricted than in *P. falciparum*.

Using additional samples to calculate standard measures of molecular evolution (Methods; Supplementary Information), we identified a subset of genes under selection in both *P. malariae/P. malariae-like* and in an earlier *P. falciparum/P. reichenowi* study¹⁰ (Extended Data Fig 3a; Extended Data Table 3), showing some conservation of selection pressures in *Plasmodium* lineages and suggesting host-specific adaptation of parasites to human and chimpanzee hosts. There is evidence that five genes are under diversifying selection in both lineages (Extended Data Fig 3b), including one encoding *merozoite surface protein 1* (*msp1*). Two genes expressed in blood stages are under significant balancing selection in both comparisons (Extended Data Fig 3c): the genes encoding *apical membrane antigen 1* (*ama1*) and an uncharacterized conserved protein. Genes under significant selection in both comparisons are enriched for 'pathogenesis' and 'entry into/exit from host cell' Gene Ontology terms. The increased signatures of selection in the *P. malariae*

lineage reflect a genome-wide increase in fixed nonsynonymous mutations in this species (Extended Data Fig 3d). One of the genes with the highest ratio of nonsynonymous to synonymous mutations in the *P. malariae*/*P. malariae-like* comparison is *Reticulocyte Binding Protein 1a* (*RBP1a*), prompting us to hypothesize that human genes encoding transmembrane proteins acting as potential *RBP1a* receptors would have to be conserved in red blood cells between humans and New World primates but not with chimpanzees (Supplementary Information; Methods; Extended Data Table 4). As expected from the recent divergence of *P. malariae* from *P. malariae-like*, significantly more genes were found with signatures of selection in *P. malariae* than between the two *P. ovale* species (Table 1; Extended Data Fig 3a; Supplementary Information).

Using the added accuracy of manually curated gene sets for both *P. malariae* and *P. o. curtisi* (Table 1; Methods), maximum likelihood trees were constructed using 1000 conserved single-copy core genes that are present in 12 selected *Plasmodium* species (Figure 2; Supplementary Information). There is strong evidence that *P. malariae* forms an outgroup both to rodent infective species and to a primate-infective clade that includes *P. vivax*. This phylogenetic arrangement is identical to that found using apicoplast data⁶ but contrasts with other recent studies^{1,7,11} (Supplementary Information). However, our phylogenetic analysis is based on the most comprehensive amino acid alignment to date, which is enriched for neutrally evolving sites through stringent filtering and has been subjected to a number of different sensitivity tests (Supplementary Information). Assuming consistent mutation rates and generation times across the branches (Supplementary

Information), we find that the relative split between the two *P. ovale* species is about 5 times earlier than the split between *P. falciparum* and *P. reichenowi*, whereas *P. malariae* and *P. malariae-like* seem to have split at a similar time to the latter two (Figure 2).

Manual curation of about 5000 gene models of both *P. malariae* and *P. o. curtisi* enabled a detailed exploration of lineage-specific differences in gene content (Table 1), with some paralogous expansions being particularly striking (Extended Data Fig 1c,d). Genes potentially involved in formation of hypnozoites in *P. ovale*- the lifecycle stage responsible for relapse infections- were also identified (Supplementary Information). The manual curation enabled pseudogenes differentially distributed between the two genomes and other human-infective *Plasmodium* species (Extended Data Table 5) to be analysed in ways not possible using computer annotated draft genome data⁷. For instance, pseudogenes are found amongst a paralogously-expanded family of invasion-associated RBPs (Extended Data Fig 4a-e), and a homolog of a *P. falciparum* cyclin (PF3D7_1227500) is only pseudogenized in *P. o. wallikeri* and may be linked with the difference in relapse times between the two *P. ovale* species¹².

Large multigene families are a defining feature that distinguish the genomes of malaria species but are refractory to detailed analysis in non-curated draft genome data⁷. In *P. malariae* and *P. ovale*, approximately 40% of the total genome is subtelomeric. However, the gene content of the subtelomeres differs substantially between the two species (Figure 3a; Table 1). The breadth and sequence types of the *pir* gene repertoires of the *P. ovale* species are similar to *P. vivax*, whereas *P.*

131 *malariae* contains only a restricted subset (Extended Data Fig 5a). The ancient
132 divergence of the two *P. ovale* is supported by their *pir* repertoires being readily
133 distinguishable (Extended Data Fig 5b). Despite being sister taxa to *P. ovale*, the *pir*
134 repertoire of rodent malaria parasites however appears to be completely different
135 (Extended Data Fig 5a). Moreover, almost 50% of *pir* genes in *P. malariae* are
136 pseudogenes (c.f. 25% in *P. o. curtisi* and 9% in *P. vivax*), suggesting an even smaller
137 functional repertoire.

138 The most striking difference in the subtelomeres of *P. malariae* is the presence of
139 two large novel gene families that were not apparent in earlier partial genome data⁷,
140 and which we have termed *fam-l* and *fam-m* (Figure 3a). *fam-l* and *-m* encoded
141 proteins show characteristics of proteins likely to be exported from the parasite to
142 the infected red blood cell surface (PEXEL export signal, signal peptide,
143 transmembrane domains, and a variable region). Most *fam-l* and *-m* genes face the
144 telomeres and occur as doublets (Figure 3b), and we find some evidence that they
145 are co-evolving (Extended Data Fig 5c,d). Proteins encoded by *fam-l* and *-m* genes
146 may therefore form heterodimers, a feature not previously seen among
147 subtelomeric gene families in other *Plasmodium* species. Finally, 3D structures of
148 *fam-l* and *-m* proteins, predicted with high-confidence (TM score > 0.5)¹³, overlap
149 the crystal structure of the *P. falciparum* *Rh5* protein (TM score > 0.8)¹⁴, with 100%
150 of the *Rh5* structure covered despite having only 10% sequence similarity (Figure
151 3c). *Rh5* is the only known *P. falciparum* protein that is essential for erythrocyte
152 invasion, via its binding to basigin on the erythrocyte surface¹⁵. The *Rh5* kite-shaped
153 fold is known to be present in *RBP2a* in *P. vivax*¹⁶, and may be a conserved structure

necessary for the binding capabilities of all *Rh* and *RBP* genes. This suggests that *fam-l* and *fam-m* genes also have an adhesion role, possibly binding host receptors. The present study highlights features of host-specific adaptations at several levels: ape to human, primate to rodent and within human hosts. As noted in previous comparative genomics studies involving host switches in the *Plasmodium* genus, invasion related genes are consistently found to be rapidly evolving. The RBP family is highly expanded (Extended Data Fig 4a) but its differential distribution across species suggests that RBP3 may be essential for normocyte invasion (Extended Data Fig 4c). In contrast to other studies^{1,7,11}, we place the rodent malaria parasites as an outgroup to *P. ovale* but rooted by *P. malariae*. The rodent malaria parasites could therefore more closely model the biology of *P. ovale* than other human infective species and indicates that there must have been an ancestral host switch from primates to rodents. The relative dating of speciation events suggests that the move between non-human primates and humans occurred at approximately the same time in two well separated lineages, suggesting that a common historical event may have promoted host switching and speciation in *Plasmodium* at the time. The much older speciation of the two *P. ovale* parasites and the fact that they have been considered identical until recently shows the limitation of morphology alone in species determination.

Due to the importance of rapidly evolving multigene families and genome structure, high quality genomes for all human infective species of *Plasmodium* are desperately needed. While *P. malariae* and *P. ovale* are known to be widespread and common in co-infections with *P. falciparum* (Figure 1), their low parasitaemia levels have

complicated their study and consequently much is still unknown about them. The present study provides a high-quality reference genome for *P. malariae*, thereby providing a crucial step forward into better understanding this elusive species.

References

- 1 Sutherland, C. J. *et al.* Two nonrecombining sympatric forms of the human malaria parasite *Plasmodium ovale* occur globally. *J Infect Dis* **201**, 1544-1550, doi:10.1086/652240 (2010).
- 2 Collins, W. E. & Jeffery, G. M. *Plasmodium malariae*: parasite and disease. *Clin Microbiol Rev* **20**, 579-592, doi:10.1128/CMR.00027-07 (2007).
- 3 Collins, W. E. & Jeffery, G. M. *Plasmodium ovale*: parasite and disease. *Clin Microbiol Rev* **18**, 570-581, doi:10.1128/CMR.18.3.570-581.2005 (2005).
- 4 Doderer-Lang, C. *et al.* The ears of the African elephant: unexpected high seroprevalence of *Plasmodium ovale* and *Plasmodium malariae* in healthy populations in Western Africa. *Malar J* **13**, 240, doi:10.1186/1475-2875-13-240 (2014).
- 5 Bousema, T., Okell, L., Felger, I. & Drakeley, C. Asymptomatic malaria infections: detectability, transmissibility and public health relevance. *Nat Rev Microbiol* **12**, 833-840, doi:10.1038/nrmicro3364 (2014).
- 6 Arisue, N. *et al.* The *Plasmodium* apicoplast genome: conserved structure and close relationship of *P. ovale* to rodent malaria parasites. *Mol Biol Evol* **29**, 2095-2099, doi:10.1093/molbev/mss082 (2012).
- 7 Ansari, H. R. *et al.* Genome-scale comparison of expanded gene families in *Plasmodium ovale wallikeri* and *Plasmodium ovale curtisi* with *Plasmodium malariae* and with other *Plasmodium* species. *Int J Parasitol*, doi:10.1016/j.ijpara.2016.05.009 (2016).
- 8 Sundararaman, S. A. *et al.* Genomes of cryptic chimpanzee *Plasmodium* species reveal key evolutionary events leading to human malaria. *Nat Commun* **7**, 11078, doi:10.1038/ncomms11078 (2016).
- 9 Lalremruata, A. *et al.* Natural infection of *Plasmodium brasilianum* in humans: Man and monkey share quartan malaria parasites in the Venezuelan Amazon. *EBioMedicine* **2**, 1186-1192, doi:10.1016/j.ebiom.2015.07.033 (2015).
- 10 Otto, T. D. *et al.* Genome sequencing of chimpanzee malaria parasites reveals possible pathways of adaptation to human hosts. *Nat Commun* **5**, 4754, doi:10.1038/ncomms5754 (2014).
- 11 Schaer, J. *et al.* High diversity of West African bat malaria parasites and a tight link with rodent *Plasmodium* taxa. *P Natl Acad Sci USA* **110**, 17415-17419, doi:10.1073/pnas.1311016110 (2013).

- 12 Nolder, D. *et al.* An observational study of malaria in British travellers:
Plasmodium ovale wallikeri and Plasmodium ovale curtisi differ significantly
in the duration of latency. *BMJ Open* **3**, doi:10.1136/bmjopen-2013-002711
(2013).
- 13 Yang, J. *et al.* The I-TASSER Suite: protein structure and function prediction.
Nat Methods **12**, 7-8, doi:10.1038/nmeth.3213 (2015).
- 14 Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm
based on the TM-score. *Nucleic Acids Res* **33**, 2302-2309,
doi:10.1093/nar/gki524 (2005).
- 15 Crosnier, C. *et al.* Basigin is a receptor essential for erythrocyte invasion by
Plasmodium falciparum. *Nature* **480**, 534-537, doi:10.1038/nature10606
(2011).
- 16 Gruszczyk, J. *et al.* Structurally conserved erythrocyte-binding domain in
Plasmodium provides a versatile scaffold for alternate receptor engagement.
Proc Natl Acad Sci U S A **113**, E191-200, doi:10.1073/pnas.1516512113
(2016).

Supplementary Information is linked to the online version of the paper at
www.nature.com/nature.

Acknowledgements

This work was supported by the Medical Research Council [MR/J004111/1] [MR/L008661/1] and the Wellcome Trust [098051]. S.A. and R.N.P. are funded by the Wellcome Trust (Senior Fellowship in Clinical Science awarded to RNP, 091625). F.R., B.O., and F.P. are financed by the ANR JCJC 2012 ORIGIN, the LMI Zofac, as well as by CNRS, IRD, and CIRMF. C.I.N. is funded by the Wellcome Trust [104792]. A.A.D. is funded as a Sanger International Fellow. The authors thank Eric Willaume from the Park of La Lékédi, and the different people involved in the sanitary controls of the chimpanzees. J.S.M. and N.M.A. are supported by NHMRC Practitioner Fellowships (#104072).

Author Contributions

G.G.R. carried out the sequence assembly, genome annotation and all the data analysis; U.C.B. performed manual gene curation; M.S. coordinated sequencing; A.J.R., J.A.C., M. M., and F.P. performed data analysis; G.G.R., T.O.A., L.A.E., J.W.B., D.P.K., C.I.N., M.B., and T.D.O. designed the *P. ovale* project; G.G.R., F.R., B.O., F.P., C.I.N., M.B., and T.D.O. designed the *P. malariae-like* project; G.G.R., A.A.D., O.M.A., N.M.A., S.A., R.N.P., J.S.M., C.I.N., M.B., and T.D.O. designed the *P. malariae* project; G.G.R., C.I.N., M.B., T.D.O. wrote the manuscript; All authors read and critically revised the manuscript; and C.I.N., M.B., T.D.O. directed the overall study.

Data deposition statement: All raw data has been deposited as described in Supplementary Information. Assembled genome sequences can be found under the study PRJEB14392 (<http://www.ebi.ac.uk/ena/data/view/PRJEB14392>). The individual accession numbers are as follows for PmUG01 (Contig accession: FLRL01000001-FLRL01000047, Chromosome accession: LT594622-LT594637), PocGH01 (Contig accession: FLRI01000001-FLRI01000638, Chromosome accession: LT594582-LT594597), PowCR01 (Contig accession: FLRJ01000001-FLRJ01000771, Chromosome accession: LT594505-LT594520) and PmlGA01 (Contig accession: FLRK01000001-FLRK01000035, Chromosome accession: LT594489-LT594503).

Reprints and permissions information is available at www.nature.com/reprints

Competing interest declaration: The authors declare no competing financial interests.

Correspondence and requests for materials should be addressed to mb4@sanger.ac.uk or tdo@sanger.ac.uk.

Table 1. Comparison of Genome Features of all Human-Infective *Plasmodium* species and *P. malariae-like*

Feature	<i>P. malariae</i>	<i>P. malariae-like</i>	<i>P. ovale curtisi</i>	<i>P. ovale wallikeri</i>	<i>P. falciparum</i>	<i>P. knowlesi</i>	<i>P. vivax</i>
Genome Size (Mb)	33.6	23.7	33.5	33.5	23.3	24.4	29.1
Scaffolds ^a	14(47)	14 (36)	14(638)	14 (771)	14(0)	14(297)	14(226)
Gaps	0	3697	894	1264	0	98	560
GC content	0.24	0.30	0.29	0.29	0.19	0.39	0.40
Gene Number*	6540	4764**	7132	7052**	5429	5291	6642
Pseudogenes	623	N/A	494	N/A	153	7	154
<i>pir</i>	255	4	1949	1375	227	70	1212
<i>var</i>	0	0	0	0	103	0	0
<i>SICAvar</i>	0	0	0	0	0	237	0
<i>STP1</i>	166	2	70	94	0	0	9
<i>tryptophan-rich antigen</i>	42	7	41	33	3	29	40
<i>ETRAMP</i>	8	4	7	11	15	10	10
<i>PHIST</i>	30	3	54	21	81	44	82
<i>fam-l^b</i>	396	0	0	0	0	0	0
<i>fam-m^b</i>	283	1	0	0	0	0	0
Nucleotide Diversity	3.2 x 10 ⁻⁴	6.5 x 10 ⁻³	1.9 x 10 ⁻⁴	3.7 x 10 ⁻⁴	5.7 x 10 ⁻⁴	N/A	9.9 x 10 ⁻⁴

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

** Non-curated gene-models

^a Unassigned contigs indicated in parentheses

^b Previously included in the *Pm-fam-a* family⁷, which consisted of all unannotated transmembrane containing genes

Figure 1. Prevalence of *P. malariae* and *P. ovale* with sample origins. Presence and absence of *P. malariae* (Red), *P. ovale* (Blue) or both (Purple) by country based on a literature review. Barplots show proportion of *P. falciparum* infections with co-

infections of *P. malariae* (Red), *P. ovale* (Blue), *P. vivax* (Green), or two species (Purple) based on the Pf3K dataset (Supplementary Information; Methods). Stars indicate origin of sample used for reference genome assembly and points show additional samples used.

Figure 2. Phylogenetic tree of the *Plasmodium* genus Maximum likelihood phylogenetic tree of the *Plasmodium* genus, showing the *P. malariae* clade (Red) and the *P. ovale* clade (Blue) together with the divergence levels of the species as calibrated to the *P. falciparum*/*P. reichenowi* split (x). Using a previously published date of 3.0-5.5 million year ago (MYA) for the *P. falciparum* and *P. reichenowi* split⁵², we thereby date the *P. ovale* split to 22.8MYA and the *P. malariae* split to 3.9MYA. Silhouettes show host-specificity of the different species. Values at branching points are bootstrap values (Methods).

Figure 3. Subtelomeric Gene Family Expansions in *P. malariae* and *P. ovale*. **a)** Gene network based on sequence similarity of all genes in *P. malariae* (Red), *P. ovale* (Blue), and *P. vivax* (Green). Cluster 1 contains *fam-l* genes, Cluster 2 contains *fam-m* genes, and Cluster 3 contains *surfins* and *STP1* genes. **b)** Chromosome 5 subtelomeric localization of *fam-l* and *fam-m* genes in doublets (Blue brackets) on the telomere-facing strand. Also showing pseudogenes (Grey) and hypothetical gene (Blue). **c)** Predicted 3D-structure of *fam-l* (above) and *fam-m* (below) overlaid with the *Rh5* crystal structure (Purple). Right-hand images show the protein rotated to the right.

Methods

Co-infection Mining

We aligned the *P. malariae* (AB354570) and *P. ovale* (AB354571) mitochondrial genome sequences against those of *P. falciparum*¹⁷, *P. vivax*¹⁸, and *P. knowlesi*¹⁹ using MUSCLE²⁰. For each species, we identified three 15bp stretches within the *Cox1* gene that contained two or more species-specific SNPs. We searched for these 15bp species-specific barcodes within the sequencing reads of all 2,512 samples from the Pf3K global collection that mapped to the *P. falciparum* mitochondrial genome (www.malariagen.net). Samples that contained at least two sequencing reads matching one or more of the 15bp barcodes for a specific species were considered to be positive for that species (Extended Data Table 2). We found good correspondence between the three different barcodes for each species, with over 80% of positive samples being positive for all three barcodes. We generated pseudo-barcodes by changing two randomly selected nucleotide bases at a time for 10 randomly selected 15bp region in the *P. vivax*¹⁸ mitochondrial genome. We did not detect any positive hits using these pseudo-barcodes. As an additional negative control, we searched for *P. knowlesi* co-infections, but did not find any samples positive for this species as expected from the limited geographical range of this species. Two samples (PocGH01, PocGH02) had high numbers of sequencing reads for all three *P. ovale* barcodes and were used for reference genome assembly and SNP calling respectively.

Parasite Material

All *P. ovale* samples were obtained from symptomatic patients diagnosed with a *P. falciparum* infection. The two *P. o. curtisi* samples (PocGH01, PocGH02) identified through co-infection mining (see above), were from two patients testing positive on a CareStart® (HRP2 based) rapid malaria diagnostic test (RDT) kit at the Navrongo War Memorial hospital, Ghana. One *P. o. wallikeri* sample (PowCR01) and one *P. o. curtisi* sample were from uncomplicated malaria patients testing positive by light microscopy at the Mile 16 - Bolifamba Health Centre, Buea, Cameroon. The other *P. o. wallikeri* sample (PowCR02) was obtained from an individual with asymptomatic parasitaemia enrolled through a community survey in Mutengene, Cameroon. For all samples, following consent obtainment, about 2-5mls of venous blood was obtained and then diluted with one volume of PBS. This was passed through CF11 cellulose powder columns to remove leucocytes prior to parasite DNA extraction.

The two *P. malariae*-like samples, PmlGA01 and PmlGA02, were extracted from Chimpanzee blood obtained during routine sanitary controls of animals living in a Gabonese sanctuary (Park of La Lékédi, Gabon). Blood collection was performed following international rules for animal health. Within six hours after collection, host white blood cell depletion was performed on fresh blood samples using the CF11 method²¹. After DNA extraction using the Qiagen blood and Tissue Kit and detection of *P. malariae* infections by *Cytb* PCR and sequencing²², the samples went through a whole genome amplification step²³.

One *P. malariae* sample, PmMA01, was collected from a patient with uncomplicated malaria in Faladje, Mali. Venous blood (2–5mL) was depleted of leukocytes within 6 hours of collection through CF11 cellulose powder columns as previously described²⁴. The study protocol was approved by the Ethics Committee of Faculty of Medicine and Odontomatology and Faculty of Pharmacy, Bamako, Mali.

Four samples of *P. malariae* were obtained from travellers returning to Australia with malaria. PmUG01 and PmID01 were sourced from patients returning from Uganda and Papua Indonesia respectively, who presented at the Royal Darwin Hospital, Darwin, with microscopy-positive *P. malariae* infection. PmMY01 was sourced from a patient presenting at the Queen Elizabeth Hospital, Sabah, Malaysia, with microscopy-positive *P. malariae* infection. Patient sample PmGN01 was collected from a patient who presented to Royal Brisbane and Womens Hospital in 2013 on return from Guinea. Venous blood samples were subject to leukodepletion within 6 hours of collection. PmUG01 was leukodepleted using a commercial Plasmodipur filter (EuroProxima, The Netherlands); home-made cellulose-based filters were used for PmID01 and PmMY01, while PmGN01 was leukodepleted using an inline leukodepletion filter present in the venesection pack (Pall Leukotrap; WBT436CEA). DNA extraction was undertaken on filtered blood using commercial kits (QIAamp DNA Blood Midi kit, Qiagen Australia). For samples PmUG01, PmID01 and PmMY01, ethical approval for the sample collection was obtained from the Human Research Ethics Committee of NT Department of Health and Families and Menzies School of Health Research (HREC-2010-1396 and HREC-2010-1431) and

the Medical Research Ethics Committee, Ministry of Health Malaysia (NMRR-10-754-6684). For sample PmGN02, ethical approval was obtained from the Royal Brisbane and Womens Hospital Human Research Ethics Committee (HREC/10/QRBW/379) and the Human Research Ethics Committee of the Queensland Institute of Medical Research (p1478).

Sample Preparation and Sequencing

One *P. malariae* sample, PmUG01, was selected for long read sequencing, using Pacific Biosciences (PacBio), due to its low host contamination and abundant DNA. Passing through a 25mm blunt-ended needle, 6ug of DNA was sheared to 20-25kb. SMRT bell template libraries were generated using the PacBio issued protocol (20kb Template Preparation using the BluePippin™ Size-Selection System). After a greater than 7kb size-selection using the BluePippin™ Size-Selection System (Sage Science, Beverly, MA), the library was sequenced using P6 polymerase and chemistry version 4 (P6/C4) in 20 SMRT cells (Supplementary Information).

The remaining isolates were sequenced with Illumina Standard libraries of 200-300bp fragments and amplification-free libraries of 400-600bp fragments were prepared²⁵ and sequenced on the Illumina HiSeq 2000 v3 or v4 and the MiSeq v2 according to the manufacturer's standard protocol (Supplementary Information). Raw sequence data was deposited in the European Nucleotide Archive (Supplementary Information).

Genome Assembly

The PacBio sequenced *P. malariae* sample, PmUG01, was assembled using HGAP²⁶ with an estimated genome size of 100Mb to account for the host contamination (~85% Human). The resulting assembly was corrected initially using Quiver²⁶, followed by iCORN²⁷. PmUG01 consisted of multiple haplotypes, with the majority haplotype being used for the iCORN²⁷, and a coverage analysis was performed to remove duplicate contigs. Additional duplicated contigs were identified using a BLASTN²⁸ search, with the shorter contigs being removed if they were fully contained within the longer contigs or merged with the longer contig if their contig ends overlapped. Host contamination was removed by manually filtering on GC, coverage, and BLASTN hits to the non-redundant nucleotide database²⁸.

The Illumina based genome assemblies for *P. o. curtisi*, *P. o. wallikeri*, and *P. malariae-like* were performed using MaSURCA²⁹ for samples PocGH01, PowCR01, and PmlGA01 respectively. To confirm that the assemblies were indeed *P. ovale*, we mapped existing *P. ovale* capillary reads to the assemblies (www.ncbi.nlm.nih.gov/Traces/trace.cgi?view=search, using as search terms: CENTER_PROJECT = "PO" and CENTER_NAME = "SC"). Prior to applying MaSURCA²⁹, the samples were mapped to the *P. falciparum* 3D7 reference genome¹⁷ to remove contaminating reads. The draft assemblies were further improved by iterative uses of SSPACE³⁰, GapFiller³¹ and IMAGE³². The resulting scaffolds were ordered using ABACAS³³ against the *P. vivax* PVP01 (<http://www.genedb.org/Homepage/PvivaxP01>) assembly (both *P. ovale*) or

against the *P. malariae* PacBio assembly (*P. malariae-like*). The assemblies were manually filtered on GC, coverage, and BLASTN hits to the non-redundant nucleotide database²⁸. iCORN²⁷ was used to correct frameshifts. Finally, contigs shorter than 1 kilobase (kb) were removed.

Using two more samples, PocGH02 and PowCR02, additional draft assemblies of both *P. ovale* species were produced using MaSURCA²⁹ followed by RATT³⁴ to transfer the gene models from the high-quality assemblies.

The genome sequences and annotations are currently available on GeneDB for *P. malariae* (<http://www.genedb.org/Homepage/Pmalariae>) and for *P. ovale curtisi* (<http://www.genedb.org/Homepage/Povale>). The genome sequences have been deposited into the European Nucleotide Archive (www.ebi.ac.uk/ena). Accession numbers for all reads generated for this study can be found in Supplementary Information. Accession IDs for the assembled genome sequences can be found in Supplementary Information.

Gene Annotation

RATT³⁴ was used to transfer gene models based on synteny conserved with other sequenced *Plasmodium* species (*P. falciparum*¹⁷, *P. vivax*¹⁸, *P. berghei*³⁵, and *P. gallinaceum* ([genedb.org/Homepage/Pgallinaceum](http://www.genedb.org/Homepage/Pgallinaceum))). In addition, genes were predicted *ab initio* using AUGUSTUS³⁶, trained on a set consisting of manually curated *P. malariae* and *P. ovale* genes respectively. Non-coding RNAs and tRNAs

were identified using Rfam 12.0³⁷. Gene models were then manually curated for both the *P. malariae* and *P. o. curtisi* reference genomes, using Artemis³⁸ and the Artemis Comparison Tool (ACT)³⁹. These tools were also used to manually identify deleted and disrupted genes (Extended Data Table 5). The *P. malariae-like* and *P. o. wallikeri* genomes were both annotated using the Companion tool⁴⁰.

Phylogenetics

Following ortholog assignment using BLASTP²⁸ and OrthoMCL⁴¹, amino acid sequences of 1000 core genes from 12 *Plasmodium* species (*P. gallinaceum* (genedb.org/Homepage/Pgallinaceum), *P. falciparum*¹⁷, *P. reichenowi*¹⁰, *P. knowlesi*¹⁹, *P. vivax* P01 (genedb.org/Homepage/PvivaxP01), *P. cynomolgi*⁴², *P. chabaudi*³⁵, *P. berghei*³⁵, and the four assemblies produced in this study) were aligned using MUSCLE²⁰. The alignments were cleaned using GBlocks⁴³ with default parameters to remove non-informative and gapped sites. The cleaned non-zero length alignments were then concatenated. This resulted in an alignment of 421,988 amino acid sites per species. The optimal substitution model for each gene partition was determined by running RAxML⁴⁴ for each gene separately using all implemented substitution models. The substitution models with the minimum AIC on a guide maximum parsimony tree were used for each gene partition. A maximum likelihood phylogenetic tree was constructed using RAxML⁴⁴ version 8.2.4. with 100 bootstraps⁴⁵ (Figure 1). To confirm this tree, we utilized different phylogenetic tools including PhyloBayes⁴⁶ and PhyML⁴⁷, a number of different substitution models within RAxML, starting the tree search from the commonly accepted phylogenetic

tree, and removing sites in the alignment which supported significantly different trees (Supplementary Information). Figtree was used to display and colour the tree (<http://tree.bio.ed.ac.uk/software/figtree/>).

A phylogenetic tree of four *P. malariae* (PmID01, PmGN01, PmGN02, PmMY01) and all *P. malariae-like* samples (PmlGA01, PmlGA02) was generated using PhyML⁴⁷ based on all *P. malariae* genes. For each sample, the raw SNPs, as called using the SNP calling pipeline (see below), were mapped onto all genes to morph them into sample specific gene copies using BCFtools⁴⁸. Amino acid alignments for all genes were concatenated and cleaned using GBlocks⁴³ with default parameters.

Divergence Dating

Species divergence times were estimated using the Bayesian inference tool G-PhoCS⁴⁹, a software which uses unlinked neutrally evolving loci and a given phylogeny to estimate demographic parameters. One additional sample per assembly (PmGN01 for *P. malariae*, PocGH02 for *P. o. curtisi*, PowCR02 for *P. o. wallikeri*, and PmlGA02 for *P. malariae-like*) was used to morph the respective assembly using iCORN²⁷. Regions in the genomes without mapping were masked, as iCORN²⁷ would not have morphed them. Unassigned contigs and subtelomeric regions were removed for this analysis due to the difficulty of alignment. Repetitive regions in the chromosomes of the four assemblies and the four morphed samples were masked using Dustmasker⁵⁰ and then the chromosomes were aligned using FSA⁵¹. The *P. o. wallikeri* and the *P. o. curtisi* chromosomes were aligned against each other, as were

the *P. malariae* and *P. malariae-like* chromosomes. The alignments were split into 1kb loci, removing those that contained gaps, masked regions, and coding regions to conform with the neutral loci assumption of G-PhoCS⁴⁹. G-PhoCS⁴⁹ was run for one million MCMC-iterations with a sample-skip of 1,000 and a burn-in of 10,000 for each of the two species pairs. Follow-up analyses using Tracer (<http://beast.bio.ed.ac.uk/Tracer>) confirmed that this was sufficient for convergence of the MCMC chain in all cases. In the model, we assumed a variable mutation rate across loci and allowed for on-going gene flow between the populations. The tau values obtained from this were 0.0049 for *P. malariae* and 0.0434 for *P. ovale*.

The tau values were used to calculate the date of the split, using the formula $(\tau \times G)/\mu$, where G is the generation time in years and μ is the mutation rate. Testing a number of different generation time and mutation rate estimates in order to optimize the *P. falciparum*/*P. reichenowi* split to 4 million years ago as estimated previously⁵², we found a mutation rate of 3.8×10^{-10} SNPs/site/lifecycle⁵³ and a generation time of 65 days⁵⁴ to generate this previously published date⁵². For *P. malariae*, a generation time of 100 days was used due to the longer intra-erythrocytic cycle.

3D Structure Prediction

The I-TASSER⁵⁵ Version 4.4 online web server⁵⁶ (zhanglab.ccmb.med.umich.edu/I-TASSER) was used for 3D protein structure prediction. Predicted structures with a

TM-score of over 0.5 were considered reliable as suggested in the I-TASSER user guidelines¹³. TM-align¹⁴, as implemented in I-TASSER⁵⁶, was used to overlay the predicted protein structure with existing published protein structures.

Hypnozoite Gene Search

Using the OrthoMCL⁴¹ clustering between all sequenced *Plasmodium* species used for the phylogenetic analysis (see above), we examined clusters containing only genes of hypnozoite-forming species: *P. vivax* P01 genes, *P. cynomolgi*⁴² genes and genes of both of the *P. ovale* species.

Additionally, we examined *P. o. curtisi* orthologs of previously published hypnozoite gene candidates⁴², looking in the 1kb 5' upstream region for any of the four ApiAP2 motifs⁵⁷ involved in sporozoite regulation and expression: GCATGC (PF3D7_1466400), GCGCCG (PF3D7_1342900), TAAGCC (PF3D7_1342900), and TGTTAC (PF3D7_0420300) (Supplementary Information).

Gene Family Analysis

All *P. malariae*, *P. ovale*, and *P. vivax* P01 genes were compared pairwise using BLASTP²⁸, with genes having a minimum local BLAST hit of 50% identity over 150 amino acids or more being considered connected. These gene connections were visualized in Gephi⁵⁸ using a Fruchterman-Reingold⁵⁹ layout and with unconnected genes removed.

P. malariae, *P. o. curtisi* and *P. o. wallikeri* protein sequences for *Plasmodium* interspersed repeat (*pir*) genes, excluding pseudogenes, were combined with those from *P. vivax* P01 (genedb.org/Homepage/PvivaxP01), *P. knowlesi*¹⁹, *P. chabaudi* AS v3 (genedb.org/Homepage/Pchabaudi), *P. yoelii yoelii* 17X v2³⁵, and *P. berghei* v3 (genedb.org/Homepage/Pberghei). Sequences were clustered using tribeMCL⁶⁰ with blast E-value 0.01 and inflation 2. This resulted in 152 subfamilies. We then excluded clusters with one member. The number of genes per species in each subfamily were plotted in a heatmap using the heatmap.2 function in ggplots in R-3.1.2.

The *pir* genes from two *P. o. curtisi* and two *P. o. wallikeri* assemblies (two high-quality and two draft genome assemblies) were compared pairwise using BLASTP²⁸ with a 99% identity over a minimum of 150 amino acids cutoff. The gene-gene connections were visualized in Gephi⁵⁸ using a Fruchterman-Reingold⁵⁹ layout after removing unconnected genes.

Mirror Tree Analysis

Using Artemis³⁸, 79 *fam-m* and *fam-l* doublets that were confidently predicted as being paired-up were manually selected based on their dispersal throughout the subtelomeres of different chromosomes. The Mirrortree⁶¹ web server (<http://csbg.cnb.csic.es/mtserver/>) was used to construct mirror trees for these 79 doublets. 35 doublets with recent branching from another doublet were manually selected to enrich for genes under recent selection (Extended Data Fig 5c). To

control for chance signals of co-evolution based on their subtelomeric location, the same methodology was repeated by choosing 79 *pir* genes in close proximity of *fam-m* genes as ‘pseudo-doublets’ for analysis and paired up in the Mirrortree⁶¹ web server (Extended Data Fig 5d).

Reticulocyte Binding Protein (RBP) Phylogenetic Plot

Full-length RBP genes were manually inspected using ACT³⁹ and verified to either be functional or pseudogenized by looking for sequencing reads in other samples that confirm mutations inducing premature stop codons or frameshifts. All functional RBPs were aligned using MUSCLE²⁰ and cleaned using GBlocks⁴³ with default parameters. PhyML⁴⁷ was used to construct a phylogenetic tree of the different RBPs (Extended Data Fig 4a). Figtree was used to colour the tree (<http://tree.bio.ed.ac.uk/software/figtree/>).

SNP Calling

Additional *P. malariae* (PmMY01, PmID01, PmMA01, PmGN01) and *P. o. curtisi* (PocGH01, PocGH02, PocCR01) samples were mapped back against the reference genomes using SMALT (-y 0.8, -i 300) (Supplementary Information). As outgroups, *P. malariae-like* (PmlGA01, PmlGA02) and *P. o. wallikeri* (PowCR01, PowCR02) were also mapped against the *P. malariae* and *P. o. curtisi* genomes respectively. The resulting ‘bam’ formatted files were merged for either of the two genomes, and GATK’s⁶² Unified Genotyper was used to call SNPs from the merged bam files (Supplementary Information). Per GATK’s⁶² best practices, SNPs were filtered by

quality of depth (QD > 2), depth of coverage (DP > 10), mapping quality (MQ > 20), and strand bias (FS < 60). Additionally, all sites for which we had missing data for any of the samples or where we had heterozygous calls were filtered. Finally, we filtered sites that were masked using Dustmasker⁵⁰ to remove repetitive and difficult to map regions. The same methodology was also applied to two *P. vivax* samples (SRR3400910 & SRR332566) and two *P. falciparum* Pf3K field samples (PF0066-C & PF0038-C) for comparative purposes.

Molecular Evolution Analysis

To calculate the nucleotide diversity for the different species, we extracted all filtered SNPs in the genomes excluding the subtelomeres. We then counted the number of pairwise differences between the different samples divided by the resulting genome size, comprising three comparisons for species with three samples (*P. malariae*, *P. o. curtisi*, *P. vivax*, *P. falciparum*) and one comparison for species with two samples (*P. o. wallikeri*, *P. malariae-like*). These estimates were then averaged by species (Table 1).

The filtered SNPs were used to morph the reference genomes using BCFtools⁴⁸ for each sample, from which sample-specific gene models were obtained. Nucleotide alignments of each gene were then generated. Codons with alignment positions that were masked using Dustmasker⁵⁰ were excluded. For each alignment (*ie.* gene), we calculated HKA⁶³, MK⁶⁴, and Ka/Ks⁶⁵ values (see below). Subtelomeric gene families

and pseudogenes were excluded from the analysis. The results were analysed and plotted in RStudio (<http://www.rstudio.com/>).

For the HKA⁶³, we counted the proportion of pairwise nucleotide differences intra-specifically (*ie.* within *P. malariae* and within *P. o. curtisi*) and inter-specifically (*ie.* between *P. malariae* and *P. malariae-like*, between *P. o. curtisi* and *P. o. wallikeri*). The intraspecific comparisons were averaged to get the genes' nucleotide diversity 'pi' and these were divided by the average interspecific comparisons, the nucleotide divergence, to get the HKA ratio (HKAr) for each gene.

The MK test⁶⁴ was performed for each gene by obtaining the number of fixed and polymorphic changes, as well as a p-value, as previously described⁶⁶ and then calculating the skew as $\log_2(((N_{\text{poly}}+1)/(S_{\text{poly}}+1))/((N_{\text{fix}}+1)/(S_{\text{fix}}+1)))$ where N_{poly} and N_{fix} are polymorphic and fixed non-synonymous substitutions respectively, while S_{poly} and S_{fix} refer to the synonymous substitutions.

To calculate the average Ka/Ks ratio⁶⁵, we took the cleaned alignments of the MK test, extracting the pairwise sequences of *P. malariae* and *P. malariae-like* (and of *P. o. curtisi* and *P. o. wallikeri*). The Bio::Align::DNASStatistics module was used to calculate the Ka/Ks values for each pair⁶⁷, averaging across samples within a species.

Using existing RNA-Seq data from seven different life-cycle stages in *P. falciparum*⁶⁸, reads were mapped against spliced gene sequences (exons, but not UTRs) from the *P. falciparum* 3D7 reference genome¹⁷ using Bowtie2⁶⁹ v2.1.0 (-a -X 800 -x). Read counts per transcript were estimated using eXpress v1.3.0⁷⁰. Genes with an effective length cutoff below 10 in any sample were removed. Summing over transcripts generated read counts per gene. Numbers were averaged for all gametocyte stages and for all blood stages. Genes with no stage having 10 or more reads were classified as being expressed elsewhere. Genes in *P. malariae* and *P. ovale* were classified by their *P. falciparum* ortholog's maximum expression stage if the difference between the maximum expression stage and the second highest stage was larger than the difference between the second and third highest stage, otherwise the gene was classified as having no peak expression.

The GO term enrichment analysis was performed in R, using TopGO⁷¹. As a GO-database, the predicted GO terms from the *P. falciparum* 3D7 genes orthologous to the *P. malariae* and *P. o. curtisi* genes included in the analysis were used. Collated tables for all molecular evolution measures for all genes can be found in Extended Data File 1 for *P. malariae* and *P. o. curtisi*.

RBP1a Receptor Search

To find the putative RBP1a receptor, we performed an OrthoMCL⁴¹ clustering between Human, chimpanzee⁷², and common marmoset⁷³ genes. *P. brasilianum* (*P. malariae*) results in chronic infection in the common marmoset⁷⁴. Genes without

transmembrane domains as well as those annotated as 'predicted' were removed.
To limit false positives, all remaining genes were searched against the chimpanzee
genes using BLASTP²⁸ with a threshold of 1e-10.

References

- 17 Gardner, M. J. *et al.* Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**, 498-511, doi:10.1038/nature01097 (2002).
- 18 Carlton, J. M. *et al.* Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. *Nature* **455**, 757-763, doi:10.1038/nature07327 (2008).
- 19 Pain, A. *et al.* The genome of the simian and human malaria parasite *Plasmodium knowlesi*. *Nature* **455**, 799-803, doi:10.1038/nature07306 (2008).
- 20 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-1797, doi:10.1093/nar/gkh340 (2004).
- 21 Auburn, S. *et al.* An effective method to purify *Plasmodium falciparum* DNA directly from clinical blood samples for whole genome high-throughput sequencing. *PLoS One* **6**, e22213, doi:10.1371/journal.pone.0022213 (2011).
- 22 Ollomo, B. *et al.* A new malaria agent in African hominids. *PLoS Pathog* **5**, e1000446, doi:10.1371/journal.ppat.1000446 (2009).
- 23 Oyola, S. O. *et al.* Optimized whole-genome amplification strategy for extremely AT-biased template. *DNA Res* **21**, 661-671, doi:10.1093/dnares/dsu028 (2014).
- 24 Kamau, E. *et al.* K13-propeller polymorphisms in *Plasmodium falciparum* parasites from sub-Saharan Africa. *J Infect Dis* **211**, 1352-1355, doi:10.1093/infdis/jiu608 (2015).
- 25 Bronner, I. F., Quail, M. A., Turner, D. J. & Swerdlow, H. Improved Protocols for Illumina Sequencing. *Curr Protoc Hum Genet* **80**, 18.12.11-42, doi:10.1002/0471142905.hg1802s80 (2014).
- 26 Chin, C. S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* **10**, 563-569, doi:10.1038/nmeth.2474 (2013).
- 27 Otto, T. D., Sanders, M., Berriman, M. & Newbold, C. Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing

746 technology. *Bioinformatics* **26**, 1704-1707,
 747 doi:10.1093/bioinformatics/btq269 (2010).
 748 28 Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of
 749 protein database search programs. *Nucleic Acids Res* **25**, 3389-3402 (1997).
 750 29 Zimin, A. V. *et al.* The MaSuRCA genome assembler. *Bioinformatics* **29**, 2669-
 751 2677, doi:10.1093/bioinformatics/btt476 (2013).
 752 30 Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding
 753 pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578-579,
 754 doi:10.1093/bioinformatics/btq683 (2011).
 755 31 Nadalin, F., Vezzi, F. & Policriti, A. GapFiller: a de novo assembly approach to
 756 fill the gap within paired reads. *BMC Bioinformatics* **13 Suppl 14**, S8,
 757 doi:10.1186/1471-2105-13-S14-S8 (2012).
 758 32 Tsai, I. J., Otto, T. D. & Berriman, M. Improving draft assemblies by iterative
 759 mapping and assembly of short reads to eliminate gaps. *Genome Biol* **11**, R41,
 760 doi:10.1186/gb-2010-11-4-r41 (2010).
 761 33 Assefa, S., Keane, T. M., Otto, T. D., Newbold, C. & Berriman, M. ABACAS:
 762 algorithm-based automatic contiguation of assembled sequences.
 763 *Bioinformatics* **25**, 1968-1969, doi:10.1093/bioinformatics/btp347 (2009).
 764 34 Otto, T. D., Dillon, G. P., Degraeve, W. S. & Berriman, M. RATT: Rapid
 765 Annotation Transfer Tool. *Nucleic Acids Res* **39**, e57,
 766 doi:10.1093/nar/gkq1268 (2011).
 767 35 Otto, T. D. *et al.* A comprehensive evaluation of rodent malaria parasite
 768 genomes and gene expression. *BMC Biol* **12**, 86, doi:10.1186/s12915-014-
 769 0086-0 (2014).
 770 36 Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts.
 771 *Nucleic Acids Res* **34**, W435-439, doi:10.1093/nar/gkl200 (2006).
 772 37 Nawrocki, E. P. *et al.* Rfam 12.0: updates to the RNA families database. *Nucleic*
 773 *Acids Res* **43**, D130-137, doi:10.1093/nar/gku1063 (2015).
 774 38 Rutherford, K. *et al.* Artemis: sequence visualization and annotation.
 775 *Bioinformatics* **16**, 944-945 (2000).
 776 39 Carver, T. J. *et al.* ACT: the Artemis Comparison Tool. *Bioinformatics* **21**,
 777 3422-3423, doi:10.1093/bioinformatics/bti553 (2005).
 778 40 Steinbiss, S. *et al.* Companion: a web server for annotation and analysis of
 779 parasite genomes. *Nucleic Acids Res* **44**, W29-34, doi:10.1093/nar/gkw292
 780 (2016).
 781 41 Li, L., Stoeckert, C. J., Jr. & Roos, D. S. OrthoMCL: identification of ortholog
 782 groups for eukaryotic genomes. *Genome Res* **13**, 2178-2189,
 783 doi:10.1101/gr.1224503 (2003).
 784 42 Tachibana, S. *et al.* Plasmodium cynomolgi genome sequences provide insight
 785 into Plasmodium vivax and the monkey malaria clade. *Nat Genet* **44**, 1051-
 786 1055, doi:10.1038/ng.2375 (2012).
 787 43 Talavera, G. & Castresana, J. Improvement of phylogenies after removing
 788 divergent and ambiguously aligned blocks from protein sequence
 789 alignments. *Syst Biol* **56**, 564-577, doi:10.1080/10635150701472164
 790 (2007).

791 44 Stamatakis, A., Ludwig, T. & Meier, H. RAxML-III: a fast program for
792 maximum likelihood-based inference of large phylogenetic trees.
793 *Bioinformatics* **21**, 456-463, doi:10.1093/bioinformatics/bti191 (2005).

794 45 Stamatakis, A., Hoover, P. & Rougemont, J. A rapid bootstrap algorithm for the
795 RAxML Web servers. *Syst Biol* **57**, 758-771,
796 doi:10.1080/10635150802429642 (2008).

797 46 Lartillot, N., Lepage, T. & Blanquart, S. PhyloBayes 3: a Bayesian software
798 package for phylogenetic reconstruction and molecular dating.
799 *Bioinformatics* **25**, 2286-2288, doi:10.1093/bioinformatics/btp368 (2009).

800 47 Guindon, S. *et al.* New algorithms and methods to estimate maximum-
801 likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**,
802 307-321, doi:10.1093/sysbio/syq010 (2010).

803 48 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools.
804 *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

805 49 Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G. & Siepel, A. Bayesian inference
806 of ancient human demography from individual genome sequences. *Nat Genet*
807 **43**, 1031-1034, doi:10.1038/ng.937 (2011).

808 50 Morgulis, A., Gertz, E. M., Schaffer, A. A. & Agarwala, R. A fast and symmetric
809 DUST implementation to mask low-complexity DNA sequences. *J Comput Biol*
810 **13**, 1028-1040, doi:10.1089/cmb.2006.13.1028 (2006).

811 51 Bradley, R. K. *et al.* Fast statistical alignment. *PLoS Comput Biol* **5**, e1000392,
812 doi:10.1371/journal.pcbi.1000392 (2009).

813 52 Silva, J. C., Egan, A., Arze, C., Spouge, J. L. & Harris, D. G. A new method for
814 estimating species age supports the coexistence of malaria parasites and
815 their Mammalian hosts. *Mol Biol Evol* **32**, 1354-1364,
816 doi:10.1093/molbev/msv005 (2015).

817 53 Claessens, A. *et al.* Generation of antigenic diversity in Plasmodium
818 falciparum by structured rearrangement of Var genes during mitosis. *PLoS*
819 *Genet* **10**, e1004812, doi:10.1371/journal.pgen.1004812 (2014).

820 54 Daniels, R. F. *et al.* Modeling malaria genomics reveals transmission decline
821 and rebound in Senegal. *Proc Natl Acad Sci U S A* **112**, 7067-7072,
822 doi:10.1073/pnas.1505691112 (2015).

823 55 Zhang, Y. Template-based modeling and free modeling by I-TASSER in CASP7.
824 *Proteins* **69 Suppl 8**, 108-117, doi:10.1002/prot.21702 (2007).

825 56 Zhang, Y. I-TASSER server for protein 3D structure prediction. *BMC*
826 *Bioinformatics* **9**, 40, doi:10.1186/1471-2105-9-40 (2008).

827 57 Campbell, T. L., De Silva, E. K., Olszewski, K. L., Elemento, O. & Llinas, M.
828 Identification and genome-wide prediction of DNA binding specificities for
829 the ApiAP2 family of regulators from the malaria parasite. *PLoS Pathog* **6**,
830 e1001165, doi:10.1371/journal.ppat.1001165 (2010).

831 58 Bastian M., H. S., & Jacomy M. . Gephi: an open source software for exporing
832 and manipulating networks. . *International AAAI Conference on Weblogs and*
833 *Social Media* (2009).

834 59 Fruchterman, T. M. J., & Reingold, E.M. Graph drawing by force-directed
835 placement. *Software: Practice and Experience* **21**, 1129-1164 (1991).

- 60 Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**, 1575-1584 (2002).
- 61 Ochoa, D. & Pazos, F. Studying the co-evolution of protein families with the Mirrortree web server. *Bioinformatics* **26**, 1370-1371, doi:10.1093/bioinformatics/btq137 (2010).
- 62 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).
- 63 Innan, H. Modified Hudson-Kreitman-Aguade test and two-dimensional evaluation of neutrality tests. *Genetics* **173**, 1725-1733, doi:10.1534/genetics.106.056242 (2006).
- 64 Kreitman, M. & Hudson, R. R. Inferring the evolutionary histories of the Adh and Adh-dup loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. *Genetics* **127**, 565-582 (1991).
- 65 Nekrutenko, A., Makova, K. D. & Li, W. H. The K(A)/K(S) ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. *Genome Res* **12**, 198-202, doi:10.1101/gr.200901 (2002).
- 66 Holloway, A. K., Lawniczak, M. K., Mezey, J. G., Begun, D. J. & Jones, C. D. Adaptive gene expression divergence inferred from population genomics. *PLoS Genet* **3**, 2007-2013, doi:10.1371/journal.pgen.0030187 (2007).
- 67 Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* **3**, 418-426 (1986).
- 68 Lopez-Barragan, M. J. *et al.* Directional gene expression and antisense transcripts in sexual and asexual stages of *Plasmodium falciparum*. *BMC Genomics* **12**, 587, doi:10.1186/1471-2164-12-587 (2011).
- 69 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).
- 70 Roberts, A. & Pachter, L. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat Methods* **10**, 71-73, doi:10.1038/nmeth.2251 (2013).
- 71 Alexa, A. R., J. topGO: Enrichment Analysis for Gene Ontology. *R package version 2.24.0* (2016).
- 72 Chimpanzee, S. & Analysis, C. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69-87, doi:10.1038/nature04072 (2005).
- 73 Marmoset Genome, S. & Analysis, C. The common marmoset genome provides insight into primate biology and evolution. *Nat Genet* **46**, 850-857, doi:10.1038/ng.3042 (2014).
- 74 Wedderburn, N., Mitchell, G. H. & Davies, D. R. *Plasmodium brasilianum* in the common marmoset *Callithrix jacchus*. *Parasitology* **90 (Pt 3)**, 573-578 (1985).

Extended Data Table 1.

Assembly and annotation statistics for the recently described assemblies compared to the present assemblies

^aless than 100 amino acids

^bannotated as either 'hypothetical protein' or 'conserved *Plasmodium* protein'

****Core defined as genes that have 1-1 orthologs between *P. falciparum* 3D7 and *P. vivax* P01**

Grey columns indicate genome assemblies by Ansari, et al⁷

Extended Data Table 2.

Samples positive for different *Plasmodium* species in the Pf3K dataset

The first column shows country of origin for the different samples, with the second column showing the total number of samples collected in that country. The following five columns show the number of these samples that are positive for the different *Plasmodium* species. All samples are positive for *P. falciparum*, which is expected because all the samples were initially identified as *P. falciparum* infections. We do not see any samples positive for *P. knowlesi*, because it has a very limited geographic range and isn't found in any of the sampled countries to our knowledge.

Extended Data Table 3.

Genes with significant scores in same test for both *P. falciparum*/*P. reichenowi* and either *P. o. curtisi*/*P. o. wallikeri* or *P. malariae*/*P. malariae-like*

For the three population genetics measures (HKAr, Ka/Ks, and MK Skew), the table shows the genes that have significant values in both the *P. falciparum*/*P. reichenowi* comparison and either the *P. o. curtisi*/*P. o. wallikeri* or the *P. malariae*/*P. malariae-like* comparison.

Extended Data Table 4.

List of Potential RBP1a receptors

The first column shows the 19 transmembrane-containing human genes that are shared between humans and the common marmoset, but not with chimpanzees. As RBP1a is the RBP with the largest differences between *P. malariae* and *P. malariae-like*, these genes may represent interesting candidates for the RBP1a receptor.

Extended Data Table 5.

Pseudogenized and deleted core genes in the two reference genomes

The first column shows the gene identifier of the *P. vivax* P01 homolog of the gene pseudogenized/deleted in one or more of the two reference genome assemblies. The second column is the *P. vivax* P01 annotation of that gene. The following two columns show whether the gene is functional (blank), pseudogenized (Pseudo) or deleted (Deleted).

Extended Data Fig 1. Large genomic changes in the *P. malariae* and *P. ovale* genome sequences. **a)** Artemis Comparison Tool (ACT)³⁹ view showing reciprocal translocation of chromosomes 6 and 10 in *P. malariae*. The red lines indicate blast similarities, chromosome 6 in orange and chromosome 10 in brown. **b)** ACT³⁹ view showing a pericentric inversion in chromosome 5 of *P. malariae*. Red lines indicate blast similarities and blue lines indicate inverted blast hits. **c)** Expansion of 22 copies (20 functional) of Pfg27 in *P. malariae* (top) compared to a single copy in *P. falciparum* (bottom) with red lines indicating blast similarities. Functional genes are in red and pseudogenes in grey. This compares to only 17 Pfg27 copies described previously⁷, which were found on six separate contigs, while all Pfg27 copies described here are found on the same contig. **d)** Expansion of PVP01_1270800 (PF3D7_1475900 in *P. falciparum*), a gene with no known function, in *P. o. curtisi* and *P. o. wallikeri*, with different copy numbers in each, compared to the one copy in *P. vivax*. Functional genes shown in orange and pseudogenes shown in grey. This gene family was recently named KELT⁷, and we confirm the 8 copies present in *P. o. wallikeri*, but show that *P. o. curtisi* has 9 copies of which one is a pseudogene.

Extended Data Fig 2. *P. malariae*-like has significantly longer branch lengths than *P. malariae*, and *P. brasilianum* is identical to *P. malariae*. **a)** A PhyML⁴⁷ phylogenetic tree of all *P. brasilianum* 18S rRNA sequences⁹, indicated by a red bar and red tip branches, and the corresponding 18S rRNA sequences from the *P. malariae* and *P. malariae*-like assemblies, labeled as such. Silhouettes represent the host origin for each sample. **b)** A phylogenetic tree of all *P. malariae* and *P. malariae*-like samples generated using PhyML⁴⁷ based on all *P. malariae* genes. *P. malariae* samples are indicated by a green bar and *P. malariae*-like samples are indicated by a purple bar. Silhouettes represent host infectivity.

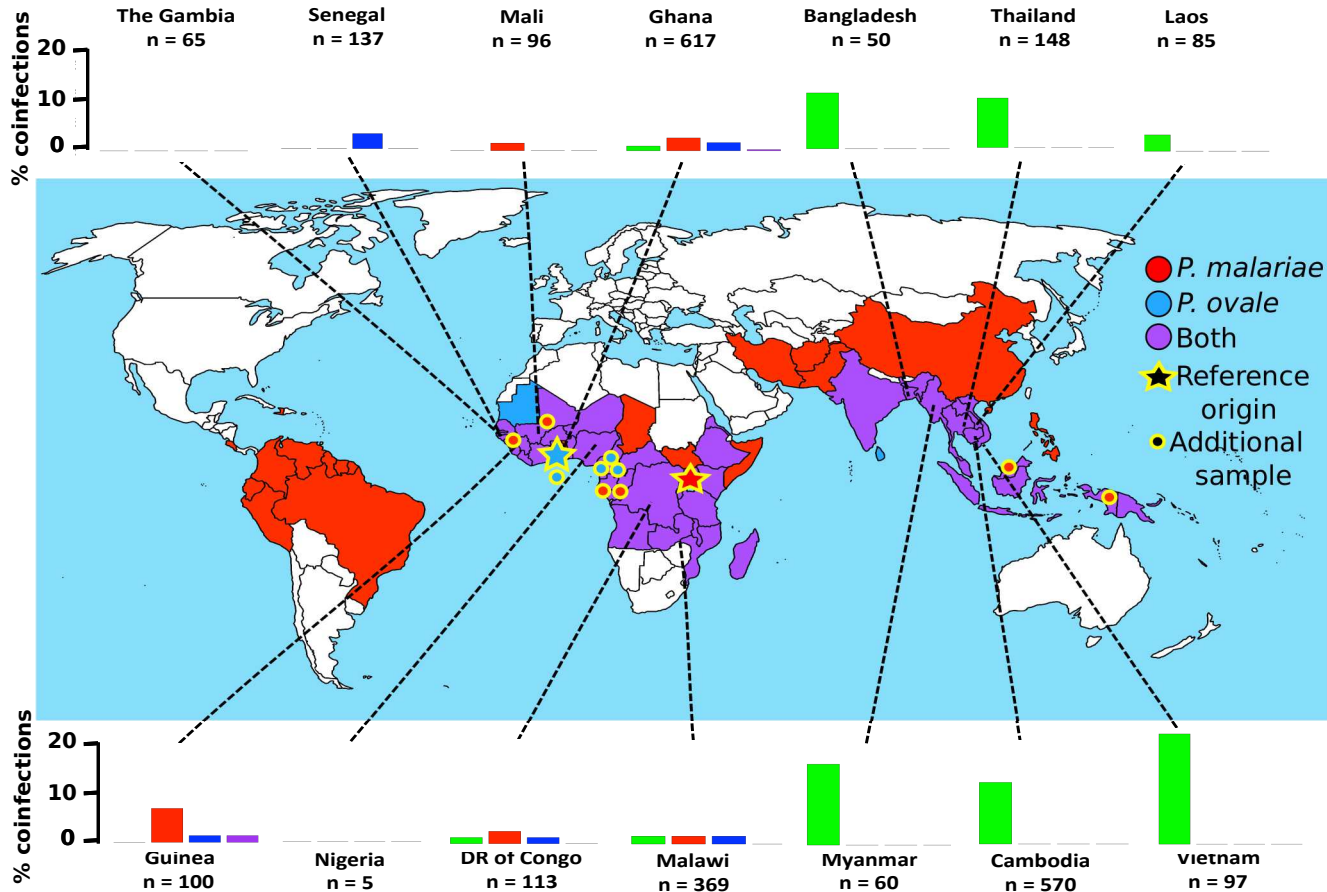
Extended Data Fig 3. Different population genetics in *P. malariae* and *P. o. curtisi*. **a)** HKA ratio and MK Skew for both *P. malariae* versus *P. malariae*-like (left) and *P. o. curtisi* versus *P. o. wallikeri* (right). Genes with high HKAr values (>0.15, vertical line) are coloured by the peak expression of their ortholog in *P. falciparum*⁶⁸ (Red, blood stage; Green, gametocyte; Blue, ookinete; Yellow, other; Grey, no peak expression) (Methods). Genes with high HKAr and a significant MK skew (square symbols): (1) merozoite surface protein 9, PF3D7_1228600; (2) rRNA (adenosine-2'-O-)-methyltransferase, PF3D7_1429400; (3) merozoite surface protein 1, PF3D7_0930300; (4) formin 1, PF3D7_0530900. **b)** Nonsynonymous versus synonymous fixed mutations per gene for both the *P. o. curtisi*/*P. o. wallikeri* (Blue) and the *P. malariae*/*P. malariae*-like (Red) comparisons. While the former has most genes centred around the x=y line, the latter has most genes below this line with more nonsynonymous than synonymous mutations, indicative of an ancestral bottleneck. **c)** Gene-wide HKAr values for the *P. falciparum*/*P. reichenowi* comparison, described earlier¹⁰, versus HKAr for the *P. o. curtisi*/*P. o. wallikeri* (Blue) and the *P. malariae*/*P. malariae*-like (Red) comparisons. Five genes show significant HKAr (>0.15) values for both comparisons: 1) ferredoxin reductase-like protein (PF3D7_0720400) 2) EGF-like membrane protein (PF3D7_0623300) 3)

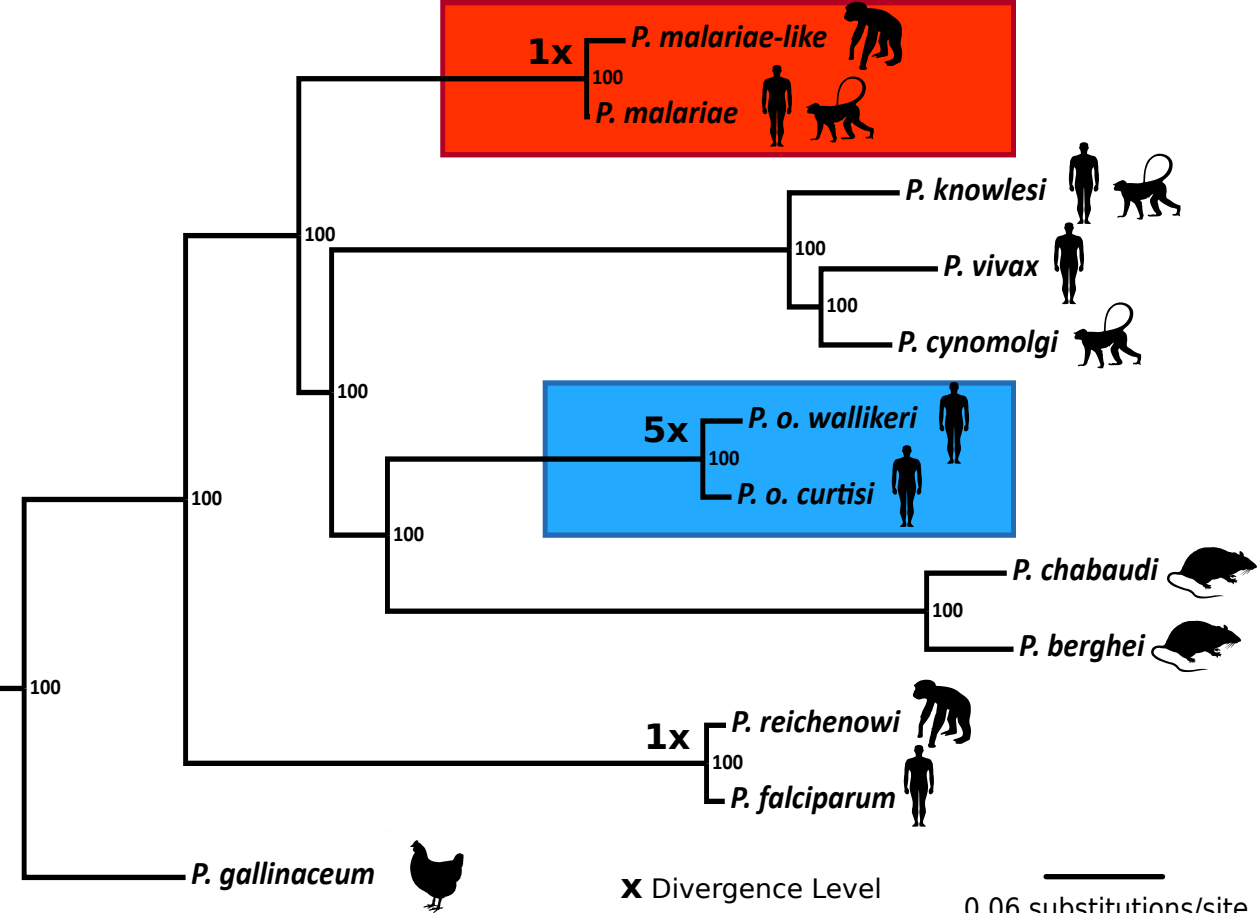
ADP/ATP carrier protein (PF3D7_1004800) 4) merozoite surface protein 1 (PF3D7_0930300) 5) conserved Plasmodium protein (PF3D7_0311000). **d)** Log2 of p-values of gene-wide MK tests for the *P. falciparum*/*P. reichenowi* comparison¹⁰ by *P. o. curtisi*/*P. o. wallikeri* (Blue) and *P. malariae*/*P. malariae-like* (Red) comparisons. Three genes have significant MK skews ($\log_2(p) < -3$) for both comparisons: 1) conserved Plasmodium protein (PF3D7_1361800) 2) apical membrane antigen 1 (PF3D7_1133400) 3) NAD(P)H-dependent glutamate synthase (PF3D7_1435300). **e)** Barplot of proportion of *P. malariae* (above) and *P. o. curtisi* (below) genes expressed at different stages (No peak expression (Grey), ookinete (Blue), gametocyte (Green), intraerythrocytic (Red), and other stage (Yellow)) binned by Ka/Ks, with the number of genes in each bin displayed (n =). *P. o. curtisi* genes with very high Ka/Ks values (> 2.5) are enriched for genes with peak expression in gametocyte.

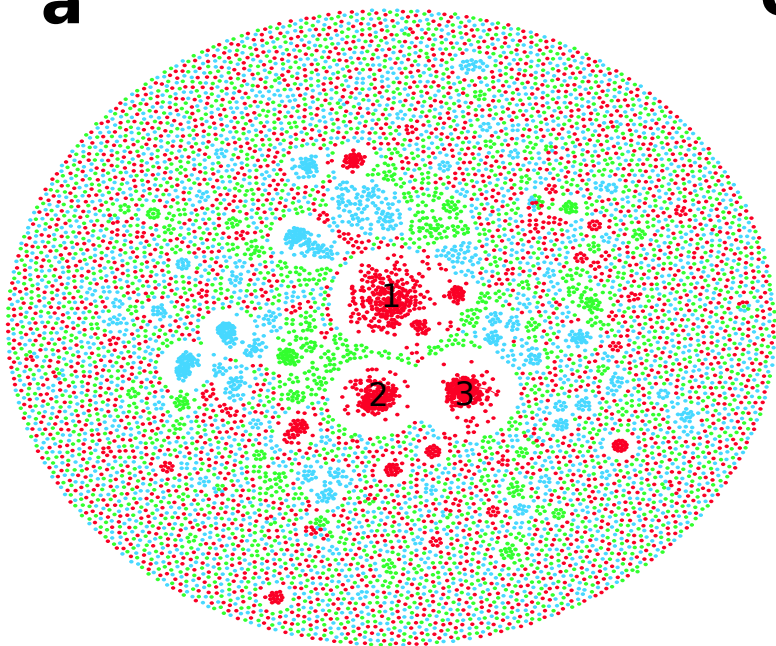
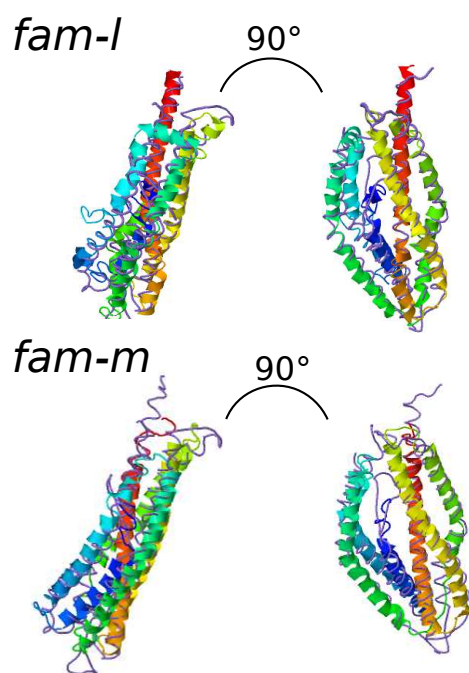
Extended Data Fig 4. Reticulocyte Binding Protein Changes in *P. malariae* and *P. ovale*. **a)** Phylogenetic tree of all full-length functional RBPs in *P. malariae* (Red branches), *P. o. curtisi* (Blue branches without stars), *P. o. wallikeri* (Blue branches with stars), and *P. vivax* (Green branches). Brackets indicate the different subclasses of RBPs: RBP1a, RBP1b, RBP2, and RBP3. **b)** ACT³⁹ view of functional (Orange) and pseudogenized (Grey) RBP1a and RBP1b in five species (*P. vivax*, *P. o. curtisi*, *P. o. wallikeri*, *P. malariae*, *P. malariae-like*). Blue indicates assembly gaps. Red bars between species indicate level of sequence similarity, with darker colour indicating higher similarity. **c)** Number of RBP genes in each of the three RBP classes (RBP1, RBP2, RBP3) by species (*P. vivax*, *P. o. curtisi*, *P. o. wallikeri*, *P. cynomolgi*, *P. malariae*, *P. knowlesi*) grouped by erythrocyte invasion preference (reticulocyte versus normocyte). **d)** PhyML⁴⁷ generated phylogenetic tree of all RBP genes over 1kb long in *P. o. curtisi* (Light blue) and *P. o. wallikeri* (Light red). Pseudogenes are denoted with **P**. Multiple functional RBP2 genes match up with pseudogenized copies in the other genome. **e)** ACT³⁹ view of RBP1b in red for *P. o. curtisi* (bottom) and the corresponding disrupted open reading frame in *P. o. wallikeri* (top), with black ticks indicating stop codons. Reads (in blue) from an additional *P. o. wallikeri* sample (PowCR02) confirm the bases introducing the frameshift (Green box) and premature stop codon (Yellow box) in RBP1b.

Extended Data Fig 5. Subtelomeric gene family dynamics in *P. ovale* and *P. malariae*. **a)** Heatmap showing the sharing of *pir* subfamilies between different species based on tribeMCL⁶⁰. Columns show *pir* subfamilies and rows show species. Colours indicate the number of genes classified into each subfamily for each species. Subfamilies were ordered by size, species were ordered for clarity. *pir* genes in rodent-infecting species fall into a small number of well-defined families. Those in *P. vivax*, *P. malariae* and *P. ovale* are however much more diverse. There is little overlap between rodent subfamilies and human-infecting subfamilies, despite *P. ovale* being a sister taxa to the rodent-infecting species. *P. knowlesi* has some sharing with other species, but its largest families are species-specific, suggesting it has undergone specialization of its *pir* repertoire. **b)** Gene network of *pir* genes for both high-quality assemblies of *P. o. curtisi* (Dark red) and *P. o. wallikeri* (Dark blue) and

draft assemblies of each (Light red and light blue respectively). *pir* genes with BLASTP²⁸ identity hits of 99%+ over 150 amino acids become connected in the graph. Genes without connections were removed. There is one connection between the two species (circled in black and with a zoomed in version), 801 between the *P. o. curtisi* assemblies, 524 between the *P. o. wallikeri* assemblies, 527 on average within each *P. o. curtisi* assembly, and 423 on average within each *P. o. wallikeri* assembly. This indicates that there is considerably less sharing of *pir* genes between the two *P. ovale* species than within each species, as expected if the two do not recombine with each other. **c)** Mirror tree⁶¹ for 79 *fam-m* and *fam-l* doublets, where the two phylogenetic trees correspond to either of the families with lines connecting branch tips of the same doublet. 35 branches (Red) were manually selected due to exhibiting recent branching. Inset below shows the correlation between the two trees for all branches (above, $r^2=0.19$, $p < 0.001$) and red branches (below, $r^2=0.53$, $p < 0.001$). This shows that the two families are co-evolving, especially when doublets that recently branched are selected, suggesting that the co-evolution may be disrupted over longer periods of time, potentially through recombination. **d)** Mirror tree⁶¹ for 79 *pir* and *fam-m* pseudo-doublets (Methods), where the two phylogenetic trees correspond to either of the families with lines connecting branch tips of the same doublet. 35 branches (Red) were manually selected due to exhibiting recent branching. Inset shows the correlation between the two trees for all branches (above, $r^2=-0.09$, $p > 0.05$) and red branches (below, $r^2=-0.10$, $p > 0.05$). This shows that the two families are not co-evolving, and that subtelomeric location does not produce sporadic signals of co-evolution.





a**c****b**