Plasmodium malariae and P. ovale genomes provide insights into malaria parasite evolution

Gavin G. Rutledge1, Ulrike Böhme1, Mandy Sanders1, Adam J. Reid1, James A. Cotton1, Oumou Maiga–Ascofare2,3, Abdoulaye A. Djimdé1,2, Tobias O. Apinjoh4, Lucas Amenga–Etego5, Magnus Manske1, John W. Barnwell6, François Renaud7, Benjamin Ollomo8, Franck Prugnolle7,8, Nicholas M. Anstey9, Sarah Auburn9, Ric N. Price9,10, James S. McCarthy11, Dominic P. Kwiatkowski1,12, Chris I. Newbold1,13, Matthew Berriman1 & Thomas D. Otto1

Elucidation of the evolutionary history and interrelatedness of Plasmodium species that infect humans 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 regions in which malaria is endemic2,3 and are often undetectable by light microscopy4, rendering their study in human populations difficult5. The exact evolutionary relationship of these species to the other human-infective species has been contested6,7. Using a new reference genome for P. malariae and a manually curated draft P. o. curtisi genome, we are now able to accurately place these species within the Plasmodium phylogeny. Sequencing a P. malariae relative that infects chimpanzees reveals similar signatures of selection in the P. malariae lineage to another Plasmodium lineage shown to be capable of colonization of both human and chimpanzee hosts. 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 family of heterodimeric proteins on its surface that have structural similarities to a protein crucial for invasion of red blood cells. The data presented here provide insight into 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 surpasses available draft genome data for P. malariae5, especially in terms of contiguity (63 versus 7,270 scaffolds, N50 = 2.3 Mb versus 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 (Fig. 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 found lower levels of nucleotide diversity in the human-infective species than in the chimpanzee-infective species (Table 1; Extended Data Fig. 2a). This mirrors the lower levels of nucleotide diversity in the human parasite P. falciparum than in its chimpanzee-infective relative P. reichenowi8. In both cases, the lack of diversity in human-infective species suggests recent population expansions. However, we found that a species that infects New World primates termed P. brasilianum was indistinguishable from P. malariae (Extended Data Fig. 2b), 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 and P. malariae-like and in an earlier study of P. falciparum and P. reichenowi10 (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 protein antigen 1 (aman) 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 and P. malariae-like comparison is reticulocyte-binding protein 1a (RBP1a), prompting us to hypothesize that human genes encoding transmembrane proteins that act 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, 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 1,000 conserved single-copy core genes that are present in 12 selected Plasmodium species (Fig. 2; Supplementary Information). There is strong evidence that P. malariae forms an out-group 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 data but

1Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK. 2Malaria Research and Training Centre, University of Science, Techniques, and Technologies of Bamako, Bamako BP E.2528, Mali. 3German Center for Infection Research, 20359 Hamburg, Germany. 4University of Buea, Post Office Box 63, Buea, South West Region, Republic of Cameroon. 5Navrongo Health Research Centre, Post Office Box 114, Navrongo, Upper East Region, Ghana. 6Centers for Disease Control and Prevention, Atlanta, Georgia 30333, USA. 7Laboratoire MIVEGEC (UMI-CNRS-IRD), 34394 Montpellier, France. 8Centre International de Recherches Médicales de Franceville, BP E.2528, Mali. 9Global and Tropical Health Division, Menzies School of Health Research and Charles Darwin University, Darwin, Northern Territory 0810, Australia. 10Centre for Tropical Medicine and Global Health, Nuffield Department of Clinical Medicine, University of Oxford, Oxford OX3 7LJ, UK. 11Clinical Tropical Medicine Laboratory, QIMR Berghofer Medical Research Institute, University of Queensland, Brisbane, Queensland 4006, Australia. 12Welcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK. 13Weatherall Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Oxford OX3 9DS, UK.

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.
contrasts with other previous studies1,7,11 (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 five 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 (Fig. 2).

Manual curation of about 5,000 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 notable (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 that are differentially distributed between the two genomes and other human-infective \( Plasmodium \) species (Extended Data Table 5) to be analyzed in ways not possible using computer-annotated draft genome data7. For instance, pseudogenes are found among a paralogously expanded family of invasion-associated RBPs (Extended Data Fig. 4a–e), and a homologue of a \( P. falciparum \) cyclin (PF3D7_1227500) is only pseudogenized in \( P. o. wallikeri \) and may

| Feature | \( P. malariae \) | \( P. malariae \)-like | \( P. o. curtisi \) | \( P. o. wallikeri \) | \( P. falciparum \) | \( P. knowlesi \) | \( P. vivax \) |
|---------|------------------|------------------|-----------------|-----------------|-----------------|----------------|----------------|
| Genome size (Mb) | 33.6 | 23.7 | 33.5 | 33.5 | 23.3 | 24.4 | 29.1 |
| Scaffolds* | 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* | 6,540 | 4,764** | 7,132 | 7,052** | 5,429 | 5,291 | 6,642 |
| Pseudogenes | 623 | N/A | 494 | N/A | 153 | 7 | 154 |
| pir | 255 | 4 | 1,949 | 1,375 | 227 | 70 | 1,212 |
| var | 0 | 0 | 0 | 103 | 0 | 0 | 0 |
| SICAvR | 0 | 0 | 0 | 237 | 0 | 9 | 0 |
| STP1 | 166 | 2 | 70 | 94 | 0 | 0 | 0 |
| tryptophan-rich antigen | 42 | 7 | 41 | 33 | 3 | 29 | 40 |
| ETRAMP | 8 | 4 | 7 | 11 | 15 | 10 | 10 |
| PHIST | 30 | 3 | 54 | 21 | 81 | 44 | 82 |
| fam-P | 396 | 0 | 0 | 0 | 0 | 0 | 0 |
| fam-m* | 283 | 1 | 0 | 0 | 0 | 0 | 0 |
| Nucleotide Diversity | \( 3.2 \times 10^{-4} \) | \( 6.5 \times 10^{-3} \) | \( 1.9 \times 10^{-4} \) | \( 3.7 \times 10^{-4} \) | \( 5.7 \times 10^{-4} \) | N/A | \( 9.9 \times 10^{-4} \) |

*Including pseudogenes and partial genes, excluding non-coding RNA genes.
**Non-curated gene-models.
*Unassigned contigs indicated in parentheses.
*Previously included in the \( Pm-fam-a \) family, which consisted of all unannotated transmembrane-containing genes.
be linked to the difference in relapse times between the two *P. ovale* species\textsuperscript{12}.

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 data. In *P. malariae* and *P. ovale*, approximately 40% of the total genome is subtelomeric. However, the gene content of the subtelomers differs substantially between the two species (Fig. 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. malariae* contains only a restricted subset (Extended Data Fig. 5a). The ancient divergence of the two *P. ovale* species is supported by their *pir* repertoires being readily distinguishable (Extended Data Fig. 5b). Despite being sister taxa to *P. ovale*, the *pir* repertoire of rodent malaria parasites, however, appears to be completely different (Extended Data Fig. 5a). Moreover, almost 50% of *pir* genes in *P. malariae* are pseudogenes (compared to 25% in *P. o. curtisi* and 9% in *P. vivax*), suggesting an even smaller functional repertoire.

The most notable difference in the subtelomers of *P. malariae* is the presence of two large gene families that were not apparent in earlier partial genome data\textsuperscript{13}, and which we have termed *fam-l* and *fam-m* (Fig. 3a). Proteins encoded by *fam-l* and *fam-m* show characteristics of proteins that are probably exported from the parasite to the infected red blood cell surface (PEXEL export signal, signal peptide, transmembrane domains, and a variable region). Most *fam-l* and *fam-m* genes face the telomeres and occur as doublets (Fig. 3b), and we found some evidence that they are co-evolving (Extended Data Fig. 5c, d). Proteins encoded by *fam-l* and *fam-m* genes may therefore form heterodimers, a feature not previously seen among subtelomeric gene families in other *Plasmodium* species. Finally, 3D structures of *fam-l* and *fam-m* proteins, predicted with high confidence (template modelling (TM)-score > 0.5)\textsuperscript{13}, overlap the crystal structure of the *P. falciparum* RH5 protein (TM-score > 0.8)\textsuperscript{14}, with 100% of the RH5 structure covered despite having only 10% sequence similarity (Fig. 3c). RH5 is the only known *P. falciparum* protein that is essential for erythrocyte invasion, through binding to basigin on the erythrocyte surface\textsuperscript{15}. The RH5 kite-shaped fold is known to be present in RBP2a in *P. vivax*\textsuperscript{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 invasion of normocytes (Extended Data Fig. 4c). In contrast to other studies\textsuperscript{1,3,11}, 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-infected species; this 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.

Owing to the importance of rapidly evolving multigene families and genome structure, high quality genomes for all human infective species of *Plasmodium* are desperately needed. Although *P. malariae* and *P. ovale* are known to be widespread and common in co-infections with *P. falciparum* (Fig. 1), their low parasitaemia levels have complicated their study and consequently little is known about them. The present study provides a high-quality reference genome for *P. malariae*, thereby providing a step forward in better understanding this elusive species.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 15 April; accepted 4 December 2016.**

**Published online 25 January 2017.**

1. Sutherland, C. J. et al. Two nonrecombining syntenic forms of the human malaria parasite *Plasmodium ovale* occur globally. *J. Infect. Dis.* 201, 1544–1550 (2010).

2. Collins, W. E. & Jeffery, G. M. *Plasmodium malariae*: parasite and disease. *Clin. Microbiol. Rev.* 20, 579–592 (2007).
3. Collins, W. E. & Jeffery, G. M. *Plasmodium ovale*: parasite and disease. *Clin. Microbiol. Rev.* **18**, 570–581 (2005).

4. Doderer-Lang, C. et al. The ears of the African elephant: unexpected high sero-prevalence of *Plasmodium ovale* and *Plasmodium malariae* in healthy populations in Western Africa. *Malar. J.* **13**, 240 (2014).

5. Bousesma, T., Okeel, L., Feiger, I. & Drakeley, C. Asymptomatic malaria infections: detectability, transmissibility and public health relevance. *Nat. Rev. Microbiol.* **12**, 833–840 (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 (2012).

7. Ansari, H. R. et al. Geno-scale comparison of expanded gene families in *Plasmodium ovale* walkeri and *Plasmodium ovale curtisi* with *Plasmodium malariae* and with other *Plasmodium* species. *Int. J. Parasitol.* **46**, 685–696 (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 (2016).

9. Lairemruata, 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 (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 (2014).

11. Schaeberle, J. et al. High diversity of West African bat malaria parasites and a tight link with rodent *Plasmodium* taxa. *Proc. Natl Acad. Sci. USA* **110**, 17415–17419 (2013).

12. Nolder, D. et al. An observational study of malaria in British travellers: *Plasmodium ovale* walkeri and *Plasmodium ovale curtisi* differ significantly in the duration of latency. *BMJ Open* **3**, e002711 (2013).

13. Yang, J. et al. The i-TASSER suite: protein structure and function prediction. *Nat. Methods* **12**, 7–8 (2015).

14. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).

15. Crozier, C. et al. Basigin is a receptor essential for erythrocyte invasion by *Plasmodium falciparum*. *Nature* **480**, 534–537 (2011). 10.1038/nature10606.

16. Gruszczyk, J. et al. Structurally conserved erythrocyte-binding domain in *Plasmodium* provides a versatile scaffold for alternate receptor engagement. *Proc. Natl Acad. Sci. USA* **113**, E191–E200 (2016).

17. Silva, J. C., Egan, A., Arze, C., Spouge, J. L. & Harris, D. G. A new method for estimating species age supports the coexistence of malaria parasites and their Mammalian hosts. *Mol. Biol. Evol.* **32**, 1354–1364 (2015).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** This work was supported by the Medical Research Council (MR/R004111/1; MR/L008661/1) and the Wellcome Trust (096051). S.A. and R.N.P. are funded by the Wellcome Trust (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 a Wellcome Trust Investigator Award (104792/Z/14/Z). A.A.D. is funded as a Sanger International Fellow. J.S.M. and N.M.A. are supported by NHMRC Practitioner Fellowships (1041802; 1042072). The authors thank E. Willaume from the Park of La Lékédi and the people involved in the sanitary controls of the chimpanzees. The authors also thank P. Griffin and S. Sekuloski for assistance in collection of *P. malariae* isolate PmGNO1.

**Author Contributions** G.G.R. carried out the sequence assembly, genome annotation and all of 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. malarialae* 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.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.B. (mb4@sanger.ac.uk) or T.D.O. (tdo@sanger.ac.uk).

**Reviewer Information** Nature thanks D. Neafsey, D. Wirth and the other anonymous reviewer(s) for their contribution to the peer review of this work.

This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons licence, users will need to obtain permission from the licence holder to reproduce the material. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.
for each gene separately using all implemented substitution models. The substitution models with the minimum Akaika Information Criterion on a guide maximum parsimony tree were used for each gene partition. A maximum likelihood phylogenetic tree was constructed using RAxMLv7 8 version 8.2.4 with 100 bootstraps89 (Fig. 2). To confirm this tree, we used different phylogenetic tools including PhyloBayes83 and PhyML83, 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* (PmD101, PmgN01, PmgN02, PmMY01) and all *P. malariae*-like samples (PmgA01, PmgA02) was generated using PhyML83 on the basis of 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 BCFTools83. Amino acid alignments for all genes were concatenated and cleaned using GBlocks48 with default parameters.

**Divergence dating.** Species divergence times were estimated using the Bayesian inference tool G-PhoCS52, a software which uses unlabeled neutrally evolving loci and a given phylogeny to estimate demographic parameters. One additional sample per assembly (PmgN01 for *P. malariae*, PocGH01 for *P. o. curtisi*, PowellR02 for *P. o. wallikeri*, and PmgA02 for *P. malariae*-like) was used to morph the respective assembly using iCORN28. Regions in the genomes without mapping were masked, as iCORN28 would not have morphed them. Unassigned contigs and subtelomeric regions were removed for this analysis owing to the difficulty of alignment. Repetitive regions in the chromosomes of the four assemblies and the four morphed samples were masked using Dustmasker83 and then the chromosomes were aligned using FSA54. The *P. o. wallikeri* and the *P. curtisi* chromosomes were aligned against each other, as were the *P. malariae* and *P. malariae*-like chromosomes. The alignments were split into 1 kb loci, removing those that contained gaps, masked regions, and coding regions to conform with the neutral loci prediction of G-PhoCS52. G-PhoCS52 was run for one million Markov Chain Monte Carlo (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 × G) / m, where G is the generation time in years and m is the mutation rate. Testing a number of different generation time and mutation rate estimates in order to optimize the *P. falciparum* and *P. reichenowi* split to 4 million years ago as estimated previously77, we found a mutation rate of 3.8 × 10−10 SNPs per site per lifetime83 and a generation time of 65 days20 to generate this previously published date77. For *P. malariae*, a generation time of 100 days was used owing to the longer intra-erythrocytic cycle.

**3D structure prediction.** The I-TASSER87 version 4.4 online web server28 (http://www.ia.ac.cn/I-TASSER) was used for 3D protein structure prediction. Predicted structures with a TM-score of over 0.5 were considered to be reproducible. The same methodology was also applied to the reference genomes using SMALT (-y 0.8, -i 300) (Supplementary Information). As outgroups, *P. malariae*-like (PmgA01, PmgA02) and *P. wallikeri* (PowCR01, PowCR02) were also mapped against the *P. malariae* and *P. curtisi* genomes respectively. The resulting ‘bam’ formatted files were merged for either of the two genomes, and GATK65 Unified Genotyper was used to call SNPs from the merged bam files (Supplementary Information). As per GATK65 best practices, SNPs were filtered by quality of depth (DP > 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 for which we had heterozygous calls were filtered. Finally, we filtered sites that were masked using Dustmasker83 to remove repetitive and difficult to map regions. The same methodology was also applied to two *P. vivax* samples (SRR3409910 and SRR332566) and two *P. falciparum* *Pb*3k field samples (PF0066-C and PF0038-C, see https://www.malariaigen.net/projects/pb3k) 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 across species (Table 1).

The filtered SNPs were used to map the reference genomes using BCFTools83 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 Dustmasker83 were excluded. For each alignment (that is, gene), we calculated Hudson–Kreitman–Aguadé (HKA)66, McDonald–Kreitman (MK)67, and K8/K8 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 HKA66, we counted the proportion of pairwise nucleotide differences intra-species, looking into and within *P. o. curtisi* and inter-specifically (that is, between *P. malariae* and *P. malariae*-like, between *P. o. wallikeri* and *P. o. curtisi*). The invariable comparisons were averaged to obtain the nucleotide diversity π of the genes and these were divided by the average inter-specific comparisons, the nucleotide divergence, to get the HKA ratio (HKAπ) for each gene.

The MK test67 was performed for each gene by obtaining the number of fixed and polymorphic changes, as well as a P value, as previously described67 and then calculating the skew as log((Npol + 1) / (Nfix + 1)) / (Npol + 1) / (Nfix + 1)), where Npol and Nfix are polymorphic and fixed non-synonymous substitutions respectively, and Npol and Nfix refer to the synonymous substitutions.

To calculate the average K8/K8 ratio68, 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::DNAStatistics’ module was used to calculate the K8/K8 values for each pair68, averaging across samples within a species.
Using existing RNA sequencing data from seven different life-cycle stages in *P. falciparum*1, reads were mapped against spliced gene sequences (exons, but not UTRs) from the *P. falciparum* 3D7 reference genome22 using Bowtie2 (ref. 72) v2.1.0 (-a -X 800 –x). Read counts per transcript were estimated using eXpress v1.3.0 (ref. 73). 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 the maximum expression stage if their *P. falciparum* orthologue 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 Gene Ontology term enrichment analysis was performed in R, using TopGO4. As a Gene Ontology database, the predicted Gene Ontology 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 Source Data File 1 for *P. malariae* and *P. o. curtisi*.

**RBP1a receptor search.** To find the putative RBP1a receptor, we performed an OrthoMCL43 clustering between human, chimpanzee75, and common marmoset76 genes. *P. brasilianum* (*P. malariae*) results in chronic infection in the common marmoset77. 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 BLASTP29 with a threshold of 1 × 10–10.

**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: FLR01000001–FLR01000038; chromosome accession: LT594582–LT594597), PocGH01 (contig accession: FLR01000001–FLR010000638; chromosome accession: LT594562–LT594597), PovCR01 (contig accession: FLR01000101–FLR01000771; chromosome accession: LT594505–LT594520) and PmlGA01 (contig accession: FLR01000001–FLR01000335; chromosome accession: LT594489–LT594503).

18. Gardner, M. J. et al. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 419, 498–511 (2002).

19. Carlton, J. M. et al. Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. *Nature* 455, 757–763 (2008).

20. Pain, A. et al. The genome of the simian and human malaria parasite *Plasmodium knowlesi*. *Nature* 455, 799–803 (2008).

21. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32, 1792–1797 (2004).

22. Aubum, 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 (2011).

23. Oilomo, B. et al. A new malaria agent in African hominids. *PLoS Pathog.* 6, e1001466 (2009).

24. Øyola, S. O. et al. Optimized whole-genome amplification strategy for extremely AT-biased template. *DNA Res.* 21, 661–671 (2014).

25. Kamau, E. et al. K3-propeller polymorphisms in *Plasmodium falciparum* parasites from sub-Saharan Africa. *J. Infect. Dis.* 211, 1352–1355 (2015).

26. Bronner, I. F., Quail, M. A., Turner, D. J. & Swerdlow, H. Improved protocols for optimized whole-genome amplification strategy for extremely AT-biased template. *DNA Res.* 21, 661–671 (2014).

27. Chin, C. S. et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* 10, 563–569 (2013).

28. Otto, T. D., Dillon, G. P., Degrave, W. S. & Berriman, M. RATT: Rapid Annotation Transfer Tool. *Nucleic Acids Res.* 39, e57 (2011).

29. Otto, T. D., Dillon, G. P., Degrave, W. S. & Berriman, M. RATT: Rapid Annotation Transfer Tool. *Nucleic Acids Res.* 39, e57 (2011).

30. Tsai, I. J., Otto, T. D. & Berriman, M. Improving draft assemblies by iterative correction of *P. falciparum* orthologue 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.

31. The Gene Ontology term enrichment analysis was performed in R, using TopGO4. As a Gene Ontology database, the predicted Gene Ontology 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 Source Data File 1 for *P. malariae* and *P. o. curtisi*.

32. RBP1a receptor search. To find the putative RBP1a receptor, we performed an OrthoMCL43 clustering between human, chimpanzee75, and common marmoset76 genes. *P. brasilianum* (*P. malariae*) results in chronic infection in the common marmoset77. 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 BLASTP29 with a threshold of 1 × 10–10.

33. 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: FLR01000001–FLR01000038; chromosome accession: LT594582–LT594597), PocGH01 (contig accession: FLR01000001–FLR010000638; chromosome accession: LT594562–LT594597), PovCR01 (contig accession: FLR01000101–FLR01000771; chromosome accession: LT594505–LT594520) and PmlGA01 (contig accession: FLR01000001–FLR01000335; chromosome accession: LT594489–LT594503).

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.
70. Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**, 418–426 (1986).

71. López-Barragán, M. J. et al. Directional gene expression and antisense transcripts in sexual and asexual stages of *Plasmodium falciparum*. *BMC Genomics* **12**, 587 (2011).

72. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

73. Roberts, A. & Pachter, L. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat. Methods* **10**, 71–73 (2013).

74. Alexa, A. R., *topGO*: Enrichment Analysis for Gene Ontology. *R package* version 2.24.0 (2016).

75. Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87 (2005).

76. Marmoset Genome Sequencing and Analysis Consortium. The common marmoset genome provides insight into primate biology and evolution. *Nat. Genet.* **46**, 850–857 (2014).

77. Wedderburn, N., Mitchell, G. H. & Davies, D. R. *Plasmodium brasilianum* in the common marmoset *Callithrix jacchus*. *Parasitology* **90**, 573–578 (1985).
Extended Data Figure 1 | Large genomic changes in the *P. malariae* and *P. ovale* genome sequences. **a**, Artemis Comparison Tool (ACT) 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**, ACT 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 previously, 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 KELT, and we confirm the 8 copies present in *P. o. wallikeri*, but show that *P. o. curtisi* has 9 copies, two of which are pseudogenes.
Extended Data Figure 2 | *P. malariae*-like has significantly longer branch lengths than *P. malariae*, and *P. brasilianum* is identical to *P. malariae*. a, A phylogenetic tree of all *P. malariae* and *P. malariae*-like samples generated using PhyML on the basis of 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. b, A PhyML phylogenetic tree of all *P. brasilianum* 18S rRNA sequences, indicated by a red bar and red tip branches, and the corresponding 18S rRNA sequences from the *P. malariae* and *P. malariae*-like assemblies, labelled as such. Silhouettes represent the host origin for each sample.
Extended Data Figure 3 | See next page for caption.
Extended Data Figure 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 orthologue in *P. falciparum* (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, Gene-wide HKAr values for the *P. falciparum* to *P. reichenowi* comparison, described earlier, versus HKAr for the *P. o. curtisi* to *P. o. wallikeri* (blue) and the *P. malariae* to *P. malariae*-like (red) comparisons. Five genes show significant HKAr (>0.15) values for both comparisons: (1) ferrodoxin 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). c, log2 of P values of gene-wide MK tests for the *P. falciparum* to *P. reichenowi* comparison10 by *P. o. curtisi* to *P. o. wallikeri* (blue) and *P. malariae* to *P. malariae*-like (red) comparisons. Three genes have significant MK skews (log2(P) < −3) for both comparisons: (1) glideosome-associated connectyor (PF3D7_1361800); (2) apical membrane antigen 1 (PF3D7_1133400); (3) NAD(P)H-dependent glutamate synthase (PF3D7_1435300). d, Nonsynonymous versus synonymous fixed mutations per gene for both the *P. o. curtisi* to *P. o. wallikeri* (blue) and the *P. malariae* to *P. malariae*-like (red) comparisons. Whereas 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. e, Bar plot 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 \(K_a/K_s\), with the number of genes in each bin displayed (n). *P. o. curtisi* genes with very high \(K_a/K_s\) values (>2.5) are enriched for genes with peak expression in gametocyte.
Extended Data Figure 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, ACT41 view of functional (orange) and pseudogenized (grey) RBP1a and RBP1b in five species (*P. vivax*, *P. o. curtisi*, *P. o. wallikeri*, *P. malariae* and *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, PhyML50 generated phylogenetic tree of all RBP genes over 1 kb long in *P. o. curtisi* and *P. o. wallikeri*. Pseudogenes are denoted with (P). Multiple functional RBP2 genes match up with pseudogenized copies in the other genome. e, ACT41 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 Figure 5 | See next page for caption.
Extended Data Figure 5 | Subtelomeric gene family dynamics in P. ovale and P. malariae. a, Heat map showing the sharing of pir subfamilies between different species based on tribeMCL\textsuperscript{63}. 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-infective species fall into a small number of well-defined families. Those in P. vivax, P. malariae and P. ovale, however, are much more diverse. There is little overlap between rodent subfamilies and human-infective 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 BLASTP\textsuperscript{29} identity hits of 99% and 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 tree\textsuperscript{64} 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 owing to exhibiting recent branching. Inset below shows the correlations as calculated by the Mirrortree webserver\textsuperscript{64} between the two trees for all branches (above, correlation = 0.19, $P < 0.001$) and red branches (below, correlation = 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 tree\textsuperscript{64} 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. We manually selected 35 branches (red) as they exhibited recent branching. Inset shows the correlations as calculated by the Mirrortree webserver\textsuperscript{64} between the two trees for all branches (above, correlation = −0.09, $P > 0.05$) and red branches (below, correlation = −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.
Extended Data Table 1 | Assembly and annotation statistics for the recently described assemblies compared to the present assemblies

|                        | PmUG01 | PmAL | PmGA31 | PocGH01 | Poc1 | Poc2 | PowCR01 | Pow1 | Pow2 |
|------------------------|--------|------|--------|---------|------|------|---------|------|------|
| Size (kb)              | 33,618 | 31,925 | 23,663 | 33,485  | 34,519 | 38,010 | 33,579  | 35,285 | 35,192 |
| Largest (kb)           | 3,664  | 56    | 3,177  | 2,946   | 94    | 491   | 3,061   | 569   | 657   |
| Average (kb)           | 534    | 4     | 474    | 22      | 9     | 17    | 43      | 26    | 22    |
| Gaps                   | 0      | 2,236 | 3,697  | 894     | 1,224 | 2,049 | 1,264   | 62    | 76    |
| Scaffolds              | 63     | 7,270 | 50     | 654     | 4,025 | 2,227 | 767     | 1,362 | 1,811 |
| Scaffold N50 (kb)      | 2,312  | 6     | 2,076  | 1,039   | 18    | 46    | 990     | 174   | 137   |
| Contigs                | 63     | 9,506 | 3,717  | 1,548   | 5,249 | 4,276 | 2,047   | 1,424 | 1,687 |
| Contig N50 (kb)        | 2,312  | 5     | 14     | 39      | 12    | 17    | 30      | 140   | 114   |
| Genes                  | 6,591  | 6,343 | 4,764  | 7,198   | 7,776 | 8,625 | 7,052   | 8,421 | 8,646 |
| 1:1 Orthologs          | 4291   | 3783  | 3637   | 4296    | 3956  | 3874  | 4174    | 3950  | 3906  |

Core***

|                        |        |      |        |        |      |      |        |      |      |
|------------------------|--------|------|--------|--------|------|------|--------|------|------|
| Short Genes*           | 102    | 104  | 109    | 99     | 69   | 63   | 88     | 89   | 85   |
| Partial                | 2      | 551  | 90     | 18     | 252  | 201  | 7      | 4    | 4    |
| Pseudo                 | 20     | 0    | 245    | 10     | 0    | 0    | 322    | 0    | 0    |
| Unknown function†      | 1,713  | 1,586 | 1,508  | 1,781  | 1,833 | 1,894 | 1,562   | 1,780 | 1,778 |
| >7 exon orthologs      | 281    | 204  | 190    | 290    | 241  | 251  | 260    | 252  | 253  |
| Median length (>7 exon) (aa) | 477 | 368  | 340    | 478    | 500  | 495  | 462    | 455  | 443  |

Subgenomes**

|                        |        |      |        |        |      |      |        |      |      |
|------------------------|--------|------|--------|--------|------|------|--------|------|------|
| Short Genes*           | 46     | 270  | 117    | 71     | 536  | 531  | 131    | 857  | 997  |
| Partial                | 8      | 621  | 246    | 262    | 547  | 676  | 156    | 2    | 6    |
| Pseudogenes            | 1236   | 3    | 21     | 978    | 4    | 6    | 393    | 11   | 10   |
| Unknown function†      | 765    | 1328 | 447    | 437    | 1176 | 1330 | 734    | 1824 | 2122 |

*Less than 100 amino acids.
†Annotated as either ‘hypothetical protein’ or ‘conserved Plasmodium protein’.
**Core defined as genes that have 1:1 orthologues between P. falciparum 3D7 and P. vivax P01.
Grey columns indicate genome assemblies from ref. 7.
### Extended Data Table 2 | Samples positive for different *Plasmodium* species in the Pf3K data set

| Country      | Total Samples | *P. falciparum* Positive | *P. vivax* Positive | *P. malariae* Positive | *P. ovale* Positive | *P. knowlesi* Positive |
|--------------|---------------|--------------------------|---------------------|------------------------|---------------------|------------------------|
| The Gambia   | 65            | 65                       | 0                   | 0                      | 0                   | 0                      |
| Guinea       | 100           | 100                      | 0                   | 7                      | 3                   | 0                      |
| Thailand     | 148           | 148                      | 11                  | 0                      | 0                   | 0                      |
| Ghana        | 617           | 617                      | 5                   | 12                     | 9                   | 0                      |
| Cambodia     | 570           | 570                      | 50                  | 0                      | 0                   | 0                      |
| Mali         | 96            | 96                       | 0                   | 1                      | 0                   | 0                      |
| Bangladesh   | 50            | 50                       | 4                   | 0                      | 0                   | 0                      |
| Malawi       | 369           | 369                      | 4                   | 4                      | 4                   | 0                      |
| Vietnam      | 97            | 97                       | 16                  | 0                      | 0                   | 0                      |
| Myanmar      | 60            | 60                       | 7                   | 0                      | 0                   | 0                      |
| Laos         | 85            | 85                       | 4                   | 0                      | 2                   | 0                      |
| DR Congo     | 113           | 113                      | 1                   | 2                      | 1                   | 0                      |
| Nigeria      | 5             | 5                        | 0                   | 0                      | 0                   | 0                      |
| Senegal      | 137           | 137                      | 0                   | 0                      | 1                   | 0                      |
| GLOBAL       | 2512          | 2512                     | 102                 | 26                     | 19                  | 0                      |

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* and *P. reichenowi* and either *P. o. curtisi* and *P. o. wallikeri* or *P. malariae* and *P. malariae*-like

| Species                  | Gene ID             | Gene Product                                             |
|--------------------------|---------------------|----------------------------------------------------------|
| *P. malariae*            | PmUG01_09042600     | apical membrane antigen 1                                |
| *P. malariae*            | PmUG01_11024300     | conserved Plasmodium protein                             |
| *P. malariae*            | PmUG01_03026800     | ferrodoxin reductase-like protein                         |
| *P. malariae*            | PmUG01_07042000     | merozoite surface protein 1                              |
| *P. malariae*            | PmUG01_08020600     | ADP/ATP carrier protein                                  |
| *P. malariae*            | PmUG01_08045200     | conserved Plasmodium protein                             |
| *P. malariae*            | PmUG01_11040300     | EGF-like membrane protein                                |
| *P. o. curtisi*          | PocGH01_13025000    | NAD(P)H-dependent glutamate synthase                     |

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

| P. vivax ID     | Annotation                                      | P. malariae | P. o. curtisi |
|-----------------|-------------------------------------------------|-------------|--------------|
| PVP01_0412100   | Multidrug efflux pump                           |             | Pseudo       |
| PVP01_0309300   | Erythrocyte vesicle protein 1                   | Pseudo      | Pseudo       |
| PVP01_1032500   | Conserved *Plasmodium* protein, unknown function|             | Pseudo       |
| PVP01_1344900   | Serine/Threonine protein phosphatase CPPED1     | Pseudo      |              |
| PVP01_1407400   | MORN repeat protein                             | Pseudo      |              |
| PVP01_1107900   | 6-cysteine protein (P92)                        | Deleted     |              |
| PVP01_1117100   | Conserved *Plasmodium* protein, unknown function| Pseudo      |              |
| PVP01_0906000   | WD repeat-containing protein WRAP73             | Deleted     |              |
| PVP01_0929100   | 6-phosphofructokinase                           | Pseudo      |              |
| PVP01_0940700   | Carbonic anhydrase                              | Deleted     | Pseudo       |
| PVP01_1445600   | Conserved *Plasmodium* protein, unknown function| Pseudo      |              |
| PVP01_1237400   | Nucleoside Transporter 3                        | Pseudo      |              |
| PVP01_1123700   | Conserved *Plasmodium* protein, unknown function| Pseudo      | Pseudo       |
| PVP01_1246900   | Biotin protein ligase                           | Pseudo      |              |

The first column shows the gene identifier of the *P. vivax* P01 homologue of the gene pseudogenized and/or 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').