Genome Sequence, Transcriptome, and Annotation of Rodent Malaria Parasite Plasmodium Yoelii Nigeriensis N67

Cui Zhang
National Institutes of Health
Cihan Oguz
National Institutes of Health
Sue Huse
National Institutes of Health
Lu Xia
National Institutes of Health
Jian Wu
National Institutes of Health
Yu-Chih Peng
National Institutes of Health
Margaret Smith
National Institutes of Health
Jack Chen
National Institutes of Health
Carole A. Long
National Institutes of Health
Justin Lack
National Institutes of Health
Xin-zhuan Su (xsu@niaid.nih.gov)
National Institutes of Health

Research Article

Keywords: Plasmodium, mouse, DNA sequence, transcript, proteome, polymorphism

Posted Date: February 4th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-162427/v1

Version of Record: A version of this preprint was published at BMC Genomics on April 26th, 2021. See the published version at https://doi.org/10.1186/s12864-021-07555-9.

Genome sequence, transcriptome, and annotation of rodent malaria parasite *Plasmodium yoelii nigeriensis* N67

Cui Zhang[1*], Cihan Oguz[2*], Sue Huse[2], Lu Xia[1,3], Jian Wu[1], Yu-Chih Peng[1], Margaret Smith[1], Jack Chen[4], Carole A. Long[1], Justin Lack[2], and Xin-zhuan Su[1#]

[1]*Malaria Functional Genomics Section, Laboratory of Malaria and Vector Research, National Institute of Allergy and Infectious Disease, National Institutes of Health, Bethesda, MD 20892-8132, USA;*

[2]*NIAID Collaborative Bioinformatics Resource (NCBR), Frederick National Laboratory for Cancer Research, Leidos Biomedical Research, Inc., Frederick, MD 21702, USA*

[3]*State Key Laboratory of Medical Genetics, Xiangya School of Medicine, Central South University, Changsha, Hunan 410078, The People's Republic of China.*

[4] *The NCI sequencing facility, 8560 Progress Drive, Room 3007, Frederick Md 21701, USA*

[*]Co-first authors

[#]Correspondence address to: Xin-zhuan Su at xsu@niaid.nih.gov

Short title: *P. yoelii nigeriensis* genome and transcriptome

Abstract

Background: Rodent malaria parasites are important models for studying host-malaria parasite interactions such as host immune response, mechanisms of parasite evasion of host killing, and vaccine development. One of the rodent malaria parasites is *Plasmodium yoelii*, and multiple *P. yoelii* strains or subspecies that cause different disease phenotypes have been widely employed in various studies. The genomes and transcriptomes of several *P. yoelii* strains have been analyzed and annotated, including the lethal strains of *Plasmodium y. yoelii* YM (or 17XL) and non-lethal strains of *Plasmodium y. yoelii* 17XNL/17X. Genomic DNA sequences and cDNA reads from another subspecies *P. y. nigeriensis* N67 have been reported for studies of genetic polymorphisms and parasite response to drugs, but its genome has not been assembled and annotated.

Results: We performed genome sequencing of the N67 parasite using the PacBio long-read sequencing technology, *de novo* assembled its genome and transcriptome, and predicted 5,383 genes with high overall annotation quality. Comparison of the annotated genome of the N67 parasite with those of YM and 17X parasites revealed a set of genes with N67-specific orthology, expansion of gene families, particularly the homologs of the *Plasmodium chabaudi* erythrocyte membrane antigen, large numbers of SNPs and indels, and proteins predicted to interact with host immune responses based on their functional domains.

Conclusions: The genomes of N67 and 17X parasites are highly diverse, having approximately one polymorphic site per 50 base pairs of DNA. The annotated N67 genome and transcriptome provide searchable databases for fast retrieval of genes and proteins, which will greatly facilitate our efforts in studying the parasite biology and gene function and in developing effective control measures against malaria.

Key words: *Plasmodium*, mouse, DNA sequence, transcript, proteome, polymorphism

Background

Malaria is one of the deadly tropical infectious diseases that impacts the health of hundreds of millions of people [1]. The lack of an effective vaccine, emergence of drug resistant parasites and insecticide resistant mosquitoes, and incomplete understanding of the disease mechanisms are the major factors that impede disease control and elimination. Vaccine development and in depth studies of disease molecular mechanisms using human populations are limited by ethical regulations and relatively high costs. Animal disease models such as parasites infecting rodents and non-human primates are important systems for studying malaria and have been widely used for vaccine development and for studying the molecular mechanisms of host-parasite interaction [2, 3]. Of course, results obtained from animal models need to be verified in human infection because there are differences in disease mechanism due to variation in genetic backgrounds of both the parasites and the hosts.

Plasmodium yoelii is one of the rodent malaria species that includes several parasite strains or subspecies well-characterized genetically and phenotypically [4, 5]. Some of the *P. yoelii* strains are genetically diverse, whereas others are closely related or derived from a common ancestor during laboratory passages in mice [4, 6, 7]. Mice infected with these *P. yoelii* strains generally have dramatic differences in parasitemia, disease severity, pathology, and host immune response [8]. For example, *P. y. yoelii* 17X (or 17XNL) and *P. y. yoelii* 17XL (or YM) are closely related parasites genetically. Indeed, these parasites were derived from a parasite isolated from a wild thicket rat in the Central African Republic [7]. 17XL and YM lines became fast growing and

lethal during passages of 17X parasites in mice in two separate laboratories, whereas 17X or 17XNL remained slow growing and non-lethal [7]. These parasites also stimulate different host responses and pathology [9-12]. Another example of parasites having closely related genomes but with different virulence is the *P. y. nigeriensis* N67 and *P. y. nigeriensis* N67C parasite pair. The genomes of N67 and N67C are very similar [5, 6]; however, they produce quite different disease phenotypes in C57BL/6 mice. Infection of N67 stimulates a strong early type I interferon (IFN-I) response, leading to a decline of parasitemia to below 5% day 7 post infection (pi). The parasitemia rebounds to 50-60%, and the host dies at day 20 pi [13]. In contrast, mice infected with N67C produce a strong T cell and INF-γ mediated inflammatory responses and die day 7 pi [14]. A C741Y amino acid substitution in the *P. yoelii* erythrocyte binding-like protein (PyEBL) contributes to the differences in virulence and immune response, but other parasite genes also play a role in the differences in disease phenotypes [15]. Identification of the genes or genetic differences between N67 and N67C parasite will facilitate our understanding of the molecular mechanisms of virulence and disease phenotypes in these infections.

With the advance of DNA sequencing technologies, the genomes and transcriptomes of many rodent malaria parasites, including those of YM, 17X, and 17XNL strains, have been sequenced and annotated [16-20]. The genomes of *Plasmodium berghei* and *Plasmodium chabaudi* parasites are approximately 18.5-19 Mb, whereas the *P. yoelii* YM genome is 22.75 Mb containing 5,675 predicted genes [18]. There are only eight genes with single nucleotide polymorphisms (SNPs) detected between the genomes of the YM and 17X strains [18], supporting isogenic parasites recently derived from the same ancestor [7]. Although RNA and DNA sequencing studies using short Illumina reads from the N67 parasite have been previously carried out to investigate

genetic polymorphisms and parasite response to drugs [6, 19], the N67 genome has not been assembled and annotated, which impedes studies of the gene functions, parasite biology, and virulence of the parasite. In this study, we sequenced the genome of the N67 parasite using PacBio sequencing technology that produces long sequence reads, assembled, and annotated its genome based on *de novo* assembled genome sequences and multiple transcriptomes. Comparison of the N67 genome sequences with those of the YM and 17X parasites revealed a set of proteins with N67-specific orthology, protein families predicted to regulate host immune responses, expansion of critical gene families, and a large number of SNPs and indels that pass stringent filtering criteria. These results have the potential to greatly facilitate our efforts in studying the parasite biology and in developing effective control measures against malaria.

Results

Genome sequencing, read statistics, and de novo assembly of the parasite genomes

We prepared DNA samples for PacBio sequencing from the N67 parasite we obtained previously [5]. Genomic DNAs were fragmented and sequenced on a PacBio Sequel using PacBio SMRT cell long read technology [21]. The polymerase reads from sequencing machine were first filtered to remove barcodes and low-quality sequences using the Hierarchical Genome Assembly Process (HGAP) (**Fig. S1a**). We obtained 1,111,721 subreads consisting of 6,733,837,360 bp for the N67 parasite, providing 233 mean coverages with an averaged barcode quality of 72. The longest subread length was 195,628 bp and the mean read length was 70,695 bp. The subreads were then assembled into 61,130 circular consensus sequencing (CCS) reads with a mean CCS coverage of 13.5-fold for the parasite.

We next *de novo* assembled the N67 CCS reads into 121 contigs consisting of 21,277,183 bp, with the largest contigs being 979,279 bp (**Table 1**). For the assembled sequences, the N50 index was 300,848 bp with 95.8% of the N67 sequences in contigs > 50 kb (**Fig. S1b).** The GC content of the sequences for the parasites is ~22% for the nuclear genome and ~30% for mitochondrial and the plastid genomes, similar to those of the 17X parasite.

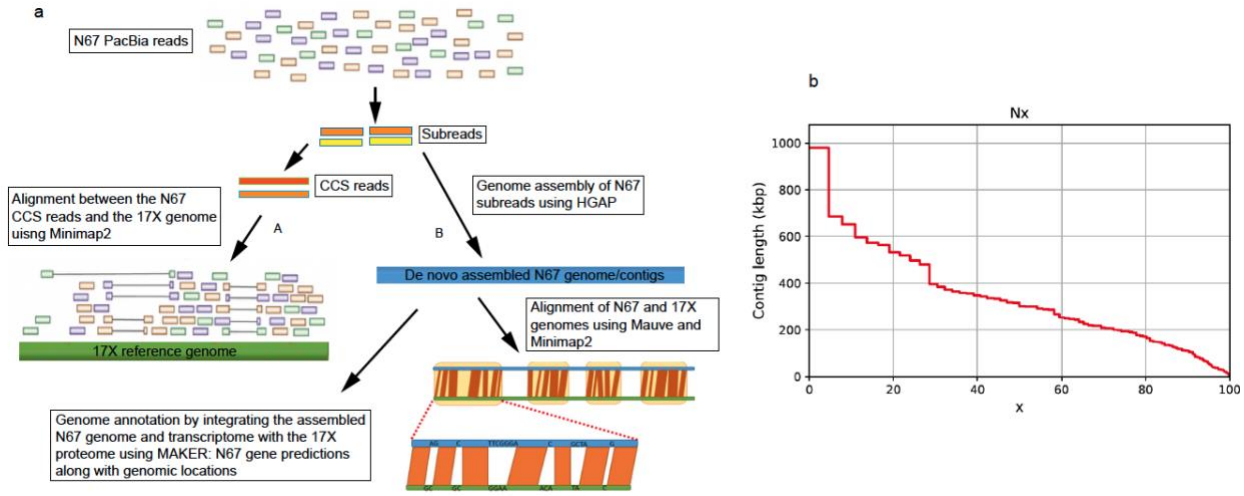


Fig. S1. Strageties of genome assembly and annotation, and plot of contig length distributions of the *Plasmodium y. yoelii* N67 parasite genome assembly. **a**, Diagram illustrating the processes of aligning N67 CCS reads (A) and contigs (B) to the 17X genomes. HGAP, Hierarchical genome assembly process; CCS read, circular consensus sequencing read. **b**, Plot of contig length distribution. The X-axis is percentage of the contigs with lengths (base pair) greater than the values indicated on the Y-axis.

Alignment of N67 sequences to the 17X assembled genome

Before investigating the diversity of the N67 genome and performing genome annotation, we aligned both the CCS reads and the assembled contigs to the updated 17X reference genome in PlasmoDB, version 46 (https://plasmodb.org/plasmo/) [18, 22] using Minimap2 [23] and the

progressiveMauve algorithm [24] that performs contig-by-contig alignment between the assembly and the 17X reference (**Fig. S1a**). A total of ~23 Mb from the N67 CCS reads were aligned to the 14 chromosomes of the 17X parasite, 34,324 bp to the plastid genome, and 6,083 bp to the mitochondrial genome, suggesting good genome coverages (**Table 2**). The mean CCS read coverages were 11.0-13.7 for the autosomes, 43.9 for the plastid genome, and 334.3 for the mitochondrial genome. In addition to the base-level alignment, we also aligned 101 of the 121 N67 contigs to the 17X reference genome using the progressiveMauve algorithm (**Fig. 1**) and 18.1 Mb of the 21.1 Mb (86%) *de novo* assembled N67 genome to the 17X genome using Minimap2. The low GC content of the parasite DNA and the abundance of low-complexity repeats in the genomes pose challenges to the assembly process and the alignment of the assembled N67 genome to the 17X reference. Therefore, approximately 14% of the N67 assembly did not align to the 17X genome at contig level.

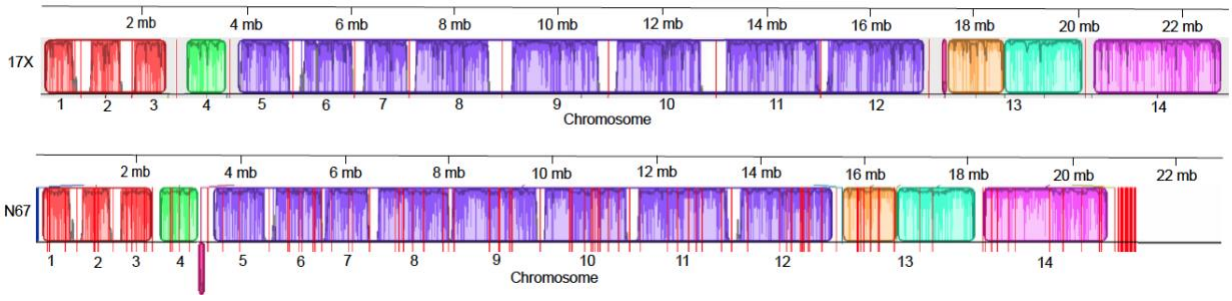


Fig. 1. Alignment of Hierarchical Genome Assembly Process (HGAP) assembled N67 contigs to the 17X chromosomes. The alignments were generated using progressiveMauve. Each color corresponds to a localized co-linear block (LCB) that is conserved across the two genomes. Inside each LCB, the jagged dark lines represent the similarity profile; with darker colors representing higher similarity regions. The vertical red lines indicate chromosome boundaries in

17X and the contig boundaries on the N67 sequences. Note a contig on N67 chromosome 4 that is inverted (presented under the chromosome line) in reference to that of 17X sequence.

RNA-Seq data and *de novo* transcriptome assembly of the N67 parasite

To facilitate genome annotation and gene prediction, we also sequenced mRNA of blood stages from eight mice infected with N67 Illumina sequencing method. Overall, 82.9% of the RNA-Seq reads from the samples uniquely mapped to the 17X genome and were retained for transcriptome assembly, The majority of the remaining reads were either uniquely mapped to the mouse genome (4.7%) or did not map to any genomes of human, mouse, bacteria, fungi and virus (9.3%) based on results from the FastQ Screen [25]. The rest of 3.1% of reads were mapped to human, fungi, bacteria or multiple genomes. We then used Trinity [26] to perform *de novo* transcriptome assembly and obtained 25,689 transcripts containing 39,856,633 bp with an average GC content of 23.5% (**Table S1**). The N50 was 1,952 bp with the largest transcript being 19,550 bp. The N67 Illumina reads were aligned to the *de novo* assembled *P. yoelii* transcriptome using Bowtie2 [27], resulting in 95.4% of the N67 read pairs concordantly aligned to the assembled transcriptome, showing a high level of overall read support for the assembly.

Gene predictions and functional annotation

We predicted 5,383 genes/proteins from the N67 genome, including all the sequences not aligned to the 17X genome, using the MAKER pipeline [28] as described in the Methods (**Table S2**). For a high quality and well-annotated assembly, at least 90% of the predicted proteins are required to have annotation edit distance (AED) values of less than 0.5 [28]. For the N67 proteome, 98% and 94% had AED (base pair level) and eAED (exon level) values less than 0.5, respectively.

Additionally, more than 50% of the proteome should ideally contain a recognizable protein domain for a well-annotated proteome [28]. Ninety-two percent of the predicted N67 proteins have recognizable domains and/or are assigned to protein families. Furthermore, the smallest predicted N67 protein has 16 amino acids (N67_005372, **Table S2**), similar to the smallest predicted protein of 15 amino acids in the 17X proteome. Search of N67_005372 protein sequence (MRVNKYVSVNMKMNYT) against the 17X and YM proteome did not return any hit; however, it has a 79% sequence identity to serine hydroxymethyltransferase of thermoacidophilic archaea *Thermoplasma volcanium*.

Search of the N67 proteome against InterPro database (https://www.ebi.ac.uk/interpro/search/sequence/) of protein families, domains and functional sites using InterProScan revealed that the largest five groups of proteins were the YIR antigens (750 members), P-loop containing nucleoside triphosphate hydrolases (272), subtelomeric PYST-A proteins (118), WD40-repeat-containing domain superfamily (89), and homologous proteins of *P. chabaudi* erythrocyte membrane protein 1 (PcEMA1) (83) (**Table S3**). Interestingly, the PcEMA1 was initially described from *P. chabaudi* parasites as an acidic phosphoprotein that might modulate the structure of the red cell membrane to the advantage of the parasite [29]. It has two tandem repats (16 × 8 AA and 2 × 9 AA) that may mediate genetic recombination and gene member expansion possibly through microhomology-mediated end joining (MMEJ) [30]. The expansion of this gene family in N67 parasites suggests that the PcEMA1 proteins may play a role in interaction with host immune system. Some other interesting groups included 43 proteins with DEAD/DEAH box helicase domain, 22 proteins with AP2/ERF domain, and 7 proteins with Rh5 coiled-coil domain.

198 We also searched the predicted N67 proteome for protein domains associated with pathways
199 within the Reactome pathway database. The top five largest Reactome groups were major
200 pathways of rRNA processing (122 proteins), regulation of expression of SLITs and ROBOs
201 (117), SRP-dependent cotranslational protein targeting to membrane (92), GTP hydrolysis and
202 joining of the 60S ribosomal subunit (91), and L13a-mediated translational silencing of
203 ceruloplasmin expression (89) (**Table S4**). Interestingly, there were also many proteins involved
204 in viral mRNA translation (78) and immune responses such as pathways of antigen processing
205 (64), neutrophil degranulation (36), NFκB activation in B cells (35), CLEC7A (Dectin-1)
206 signaling (35), downstream TCR signaling (35), FCERI mediated NFκB activation (35),
207 interleukin-1 signaling (35), NIK noncanonical NFκB signaling (35), and Vpu mediated
208 degradation of CD4 (35), TNFR2 non-canonical NFκB (34), and genes in MHC class II antigen
209 presentation (23) (**Table S5**). The molecules in the viral mRNA translation are mostly structural
210 constituents of ribosome proteins that are likely essential for the translation of parasite proteins.
211 *Toxoplasma* parasites secrete effector proteins into the host cell to co-opt host transcription
212 factors and modulate host immune responses [31]. Some of the proteins grouped with immune
213 response pathways could play important roles in regulating host immune response to the parasite
214 infection.
215
216 **Estimates of completeness of the N67 genome and transcriptome**
217 We next used Benchmarking Universal Single-Copy Orthologs (BUSCO 3.0.2) to assess the
218 completeness of the assembled N67 genome. Of the 3,642 *Plasmodium* and 446 *Apicomplexa*
219 BUSCO gene sets, 3,369 (92.5%) and 431 (96.6%) were present in the N67 genome assembly,
220 respectively (**Table S6**). We also evaluated the extent to which the assembled N67 transcriptome

matched the BUSCO gene sets across the *Apicomplexa* and *Plasmodium*. Approximately 92.8% of the BUSCO *Apicomplexa* gene set and 71.3% of the *Plasmodium* gene set were present in the assembled N67 transcriptome (**Table S6**). The N67 transcriptome genes matching the *Plasmodium* BUSCO gene set included 1,448 complete and single-copy genes (39.8%), 1,149 (31.5%) complete and duplicated genes, and 338 fragmented sequences (9.3%) (**Table S6**). There were also 707 genes (19.4%) missing from the *Plasmodium* BUSCO gene set; some of the missed genes might not be expressed in the blood stages. The genomic sequences from PacBio sequencing appear to provide more complete gene assembly than those from short Illumina RNA seq data.

P. yoelii common orthogroups and putative proteins with N67-specific orthology

The N67 and 17X (or YM) parasites belong to two subspecies of *P. yoelii*, and the genomes of these parasites are quite diverse [4, 6]. It is potentially interesting to identify genes common and unique (or highly diverse) in these parasite genomes. Therefore, we compared the 5,383 predicted proteins from N67 with 6,092 17X proteins and 5,685 YM proteins using OrthoFinder [32] and identified a core set of 4,539 orthogroups shared among the N67, 17X, and YM genomes (**Fig. 2a**). Out of a total 17,160 proteins from the three parasite strains, 17,035 (99.3%) were placed in 5,230 orthogroups based on searches of sequence similarity using DIAMOND within the latest OrthoFinder framework [33, 34]. Of the 5,383 N67 proteins, 5,294 were assigned to orthogroups, including 110 in 12 N67 specific orthogroups (**Table S7** and **Table S8)**. There were also 89 proteins that could not be assigned to any orthogroup, leading to a total of 199 proteins that appear to have N67-specific orthology. These proteins had slightly lower pairwise bit-scores than those assigned to the common orthogroups (**Fig. 2b**).

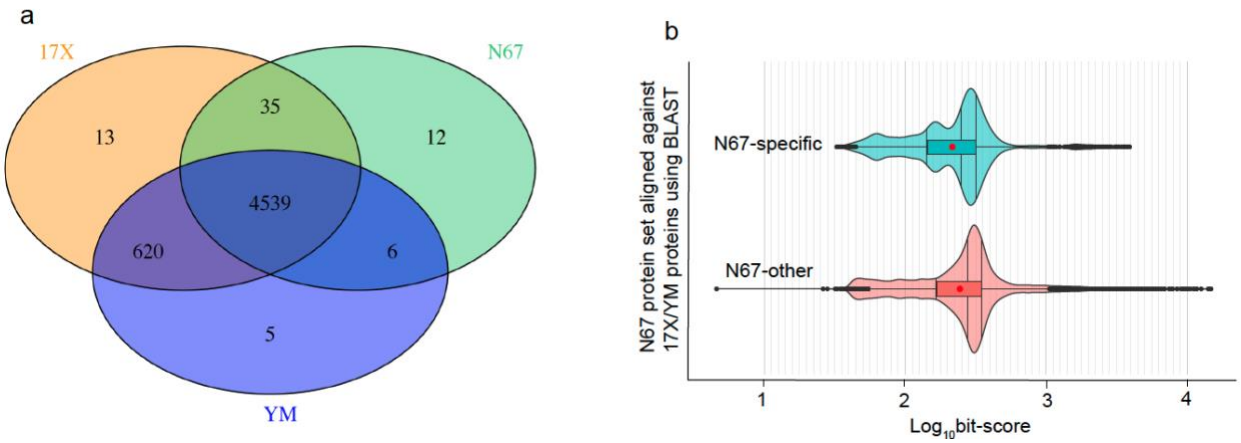


Fig. 2. Shared and strain-specific orthogroup counts identified from *Plasmodium y. yoelii* YM, *P. y. yoelii* 17X, and *P. y. nigeriensis* N67 parasites using OrthoFinder [32]. **a**, Venn diagram of shared and strain-specific orthogroups; **b**, $\log_{10}$-transformed bit-score distributions for N67 proteins that are not assigned to any orthogroup plus those in N67-specific orthogroups (N67-specific) and proteins assigned to orthogroups having at least one 17X or YM protein (N67-other). The bit-scores are derived from pairwise BLAST alignments within the Orthofinder framework, where all queries were N67 protein sequences that were aligned against the 17X and YM sequences. The red dots indicate the mean values of bit-score distributions, whereas the vertical lines within the violins indicate the median, upper and lower quartile values.

To further characterize N67-specific proteins, we used BLAST to align the 199 N67 proteins against the 17X proteome and showed that the majority proteins had motifs matching members of highly diverse gene families. Among the 199 proteins, 91 are hypothetical or uncharacterized proteins, 64 are PIR/YIR proteins, 22 are Fam-A/B proteins, and five are reticulocyte binding proteins (**Table S9**). Clustering the proteins based on sequence similarity generated three dendrograms, one consisting of Fam-A and Fam-B proteins (**Fig. S2a and Table S9**), another one consisting of 10 YIR proteins (**Fig. S2b**), and a third one of three subclusters of YIR proteins

(**Fig. S2c**). The YIR proteins in cluster B and C are quite different and could not be clustered together, suggesting potentially different origins.

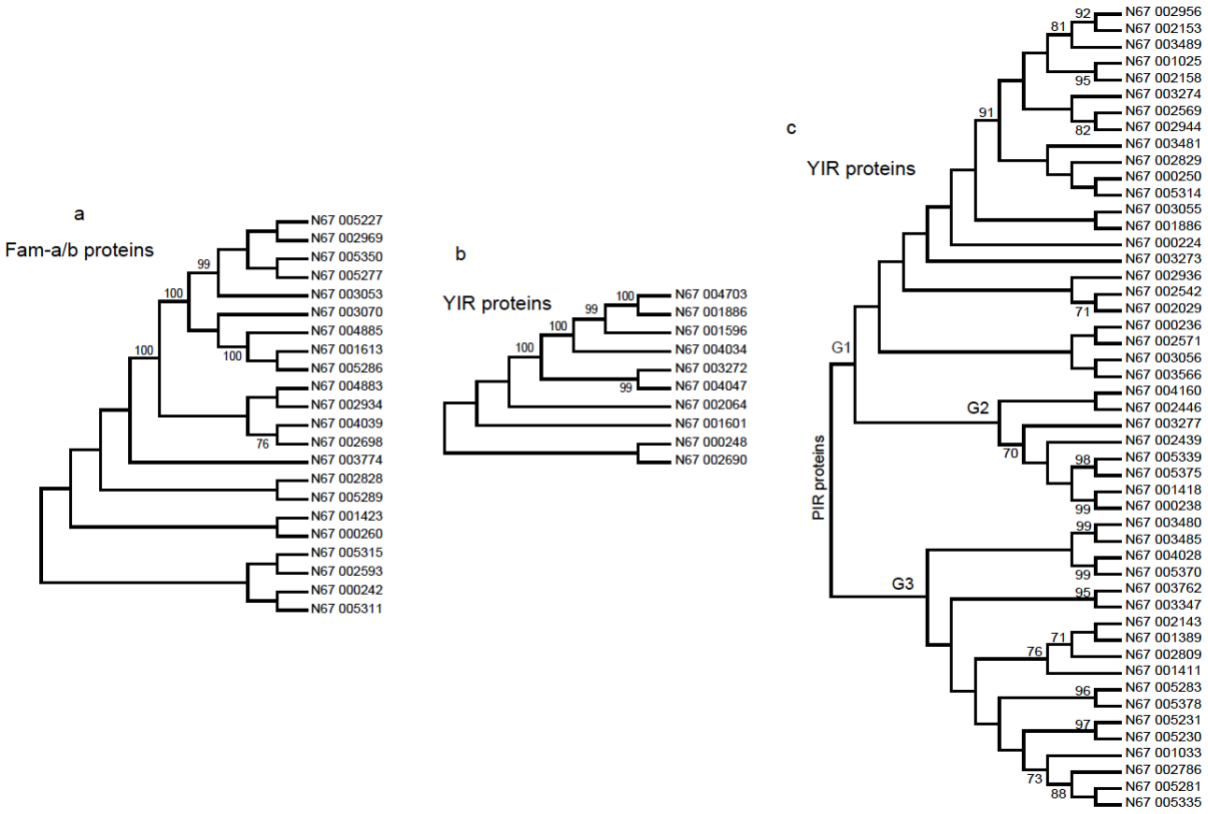
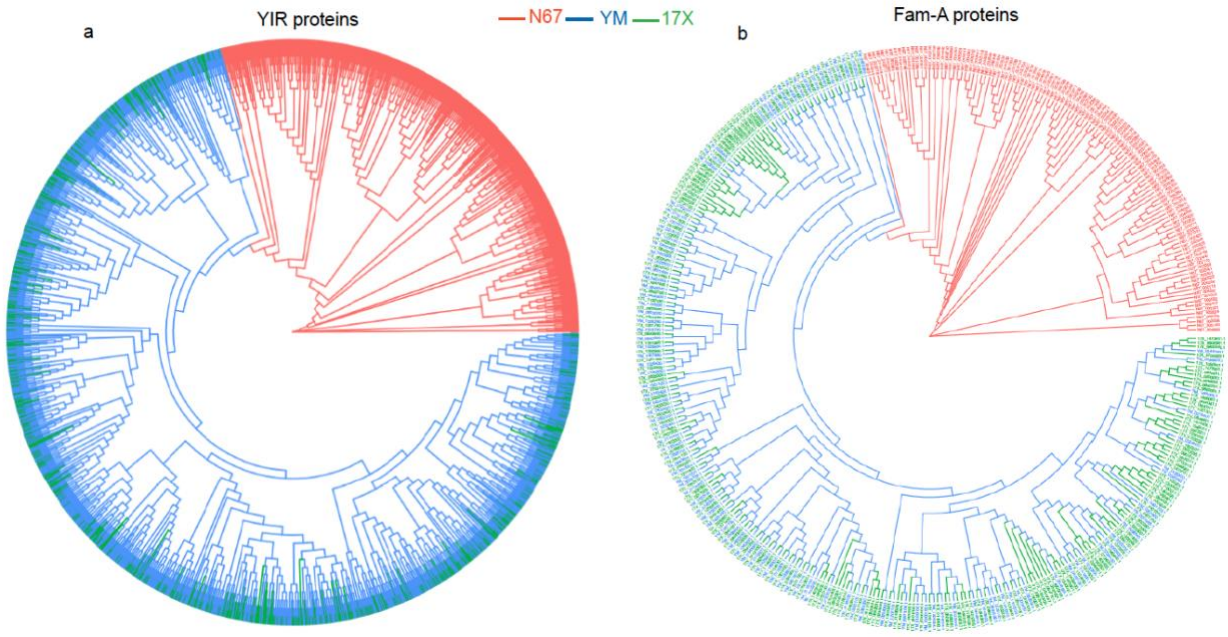


Fig. S2. Clustering of protein sequences from the *Plasmodium y. nigeriensis* N67-specific orthogroups and those that are not assigned to any orthogroup. The predicted protein sequences were aligned using ClustalW algorithm and clustered using procedures described in the Methods section. **a**, Fam-A/B proteins; **b**, YIR proteins (group 1); **c**, YIR proteins (group 2). Only bootstrap values higher than 70% are shown.

Gene families from three *P. yoelii* parasites

Among the predicted genes and proteins, we identified 22 gene families that have been previously found in *Plasmodium yoelii* [35] with at least one member detected in N67 (**Table S10**). The gene families consist of 1,475 genes (24% of the predicted genes) for 17X, 1,141

genes (21%) for N67, and 1,075 genes (19%) for YM parasite (**Table S10**). The largest gene families are the *yir* and *fam-a/b/c/d* families. There are 1,057 *yir* and 301 *fam-a/b/c/d* genes for 17X, 750 *yir* and 213 *fam-a/b/c/d* genes for N67, and 773 *yir* and 190 *fam-a/b/c/d* genes for YM parasite, respectively. As expected, clustering of YIR and Fam-A protein families showed that the proteins from N67 grouped separately from those of 17X and YM (**Fig. 3a** and **Fig. 3b**), consistent with N67 being a subspecies of *P. yoelii*. The true numbers of *yir* and *fam-a/b/c/d* genes for the N67 and YM parasites could be larger because some genes in these gene families are likely not assembled into the genome. Other important gene families include genes encoding early transcribed membrane proteins (ETRAMPs), lysophospholipase, erythrocyte membrane antigens, and reticulocyte binding proteins associated with host-parasite interactions and Cys6 (6-Cysteine) proteins. As noted above, there are 83 copies of the gene encoding PcEMA1 homologs [29] in the N67 parasite, compared with only one gene in the 17X and YM parasites, respectively (**Table S10**). Clustering of the PcEMA1 proteins from the three parasites showed that N67_000859 and N67_000245 were closely related to the two single copies from 17X (17X_10019001) and YM (YM_100119001) (**Fig. 4**). Similarly, there are nine genes encoding haloacid dehalogenase-like hydrolase in N67, but only 4 genes in both 17X and YM parasites. In contrast, the number of genes encoding reticulocyte binding proteins appears to be reduced in N67 parasites; 12 genes for N67 (including five from N67 specific orthogroups), whereas 17X and YM have 33 and 31, respectively. Although we cannot rule out that the differences in gene number are due to different degrees of genome completeness, the increases in gene number of PcEMA1 homolog and haloacid dehalogenase-like hydrolase in N67 could be associated with its unique growth characteristics and stimulation of strong early type I interferon response [13]. The expansion of the PcEMA1 homolog genes deserve additional investigation.



298

299 **Fig. 3**. Clustering of YIR and Fam-A proteins from the *Plasmodium y. nigeriensis* N67, *P. y.*

300 *yoelii* 17X, and *P. y. yoelii* YM parasites. The predicted protein sequences were aligned using

301 ClustalW algorithm in msa R package, and the dendrograms were inferred and visualized using

302 the ape, seqinr, and ggtree packages in R. **a**, YIR proteins from N67, 17X, and YM parasites; **b**,

303 Fam-A proteins. Proteins are colored based on their parasite origins.

304

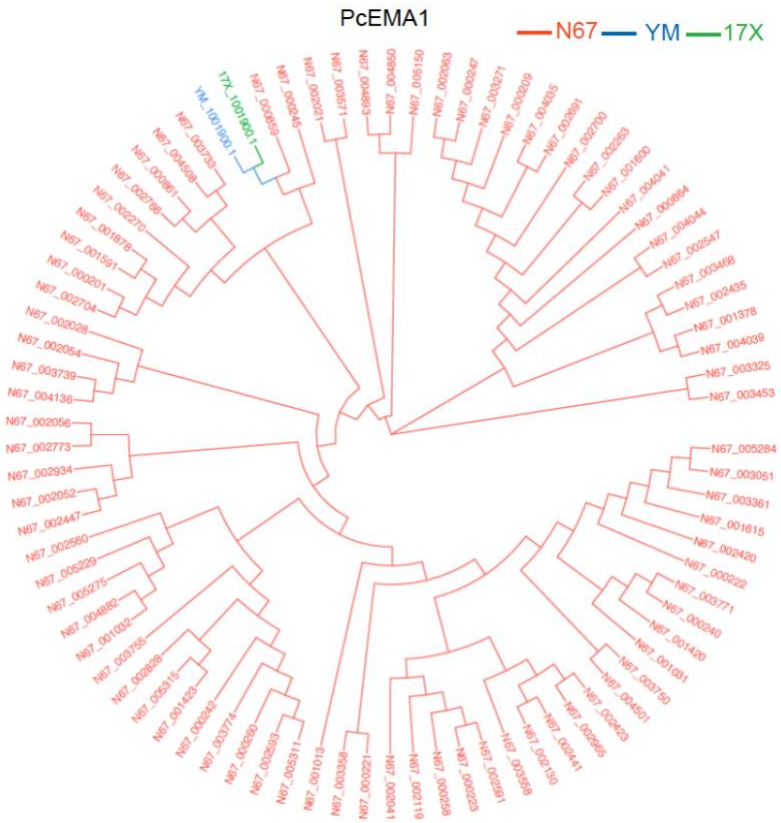


Fig. 4. Clustering of homologous *P. chabaudi* erythrocyte membrane protein 1 (PcEMA1) from *Plasmodium y. nigeriensis* N67, *P. y. yoelii* 17X and *P. y. yoelii* YM parasites. The gene family is expanded only in the N67 parasite. The predicted protein sequences were aligned using ClustalW algorithm in msa R package, and the dendrogram was inferred and visualized using the ape, seqinr, and ggtree packages in R. Proteins are colored based on their parasite origins.

Sequence polymorphisms between N67 and 17X

Initial alignment of the N67 sequences to those of 17X identified 486,102 SNPs and 41,317 indels, leading to approximately one SNP per 47.3 bp and one indel per 556.7 bp DNA between 17X and N67 assuming a genome size of 23 Mb [18]. The number of SNPs of this study are similar to the previous 458,922 SNPs from Illumina reads [6]. We further filtered the SNPs using Ensembl Variant Effect Predictor (VEP) based on the impact of the variants on the protein

function and the following criteria: coverage in both strains with a minimum depth of 5X and a dominant allele frequency of 75%. We identified 69,413 SNPs and 11,076 indels that passed the following criteria and had high or moderate impacts (**Table S11**). These variants represent approximately one SNP per 331.4 bp and one indel per 2076.6 bp of DNA. High impact variants are those assumed to be disruptive on the protein functions such as total loss of function caused by protein truncation. Moderate impact variants are non-disruptive variants that might change protein effectiveness such as missense mutations, in-frame indels, and splice-region variants outside the canonical splice site (http://uswest.ensembl.org/info/genome/variation/prediction/predicted_data.html). Given that there are 80,489 and 79,946 high or moderate impact variants between N67 vs 17X and YM, respectively (**Table S11**), it is quite interesting that with so many differences between the genomes, the YM (or 17XNL) and N67 parasites could be genetically crossed without apparent difficulty in terms of mating or genetic crossover [5, 36].

Discussion

This study reports an annotated genome for the N67 parasite, a subspecies of *P. yoelii nigeriensis*, using PacBio long sequence reads and RNA-Seq sequences from blood stage parasites. The genomes of several *P. yoelii* strains belonging to different subspecies (*P. y. yoelii*), including 17X, 17XNL, and YM strains, have been reported and well-characterized [16, 18, 20]. Although Illumina-based RNA and DNA sequencing of the N67 parasite have been performed for studying genetic polymorphisms and parasite response to drugs [6, 19], assembly of the N67 genome and prediction of gene functions have not been reported previously due to difficulties in assembling AT-rich short sequence reads. Our study predicted and annotated 5,383 genes for

N67 and identified a set of proteins with N67-specific orthology. Approximately 14% of the assembled N67 sequences did not align to the 17X genome. The unaligned sequences were likely due to high level of sequence diversity between the 17X and N67 genomes and the limitations of the assembling process. There are more than 80,000 SNPs and indels having medium to high functional imacts betwen the N67 and 17X genomes. Improvement of the N67 genome with increased overall alignment with the 17X reference would likely increase the numbers of SNPs and indels. Interestingly, despite having highly diverse genomes (the level of diversity is greater than human and chimpanzee), the N67 and YM or 17XNL can be genetically crossed to produce progenies for mapping parasite and host genes [5, 36, 37]. The genome and transcriptome sequences and annotations of the N67 parasite we present here will be valuable resources for future studies on gene functions of this important *P. yoelii* subspecies.

The majority of studies using *P. yoelii* parasites involved the lethal YM (or 17XL) and nonlethal 17X (17XNL) strains. However, *P. y. nigeriensis* parasites have also been used as models for studying unique disease phenotypes, transmission in mosquitoes, strain-specific host immune responses and drug resistances [13, 14, 38-41]. The N67 and N67C are two isogenic strains of *P. y. nigeriensis* subspecies that cause very unique and interesting disease phenotypes in C57BL/6 mice. Whereas mice infected with N67 parasites produce an early peak of IFN-I that has been linked to suppression of parasitemia day 6 pi, mice infected with N67C produce low levels of IFN-I and die approximately day 7 pi due to T cell and IFN-γ mediated inflammation [13, 14]. A mutation (C741Y) in PyEBL partially contributes to the difference in disease phenotypes [15]; however, other genes in the genome are likely involved, particularly the variant gene families such as the *yir* and *fam-a/b/c/d* genes. For example, a locus at one end of chromosome 13 was

shown to be significantly linked to many host genes functioning in IFN-I response pathways or interferon-stimulated genes (ISGs) [37]. Among the proteins in Reactome pathways that may play a role in viral mRNA translation and immune responses, several proteins can bind RNA and DNA or have endopeptidase activities. These proteins may function to activate NFκB signaling, ISG15 antiviral mechanism, and DDX58/IFIH1-mediated induction of IFN-α/β if they are secreted into the host cytoplasm, particularly in liver cells. Further experiments are required to demonstrate whether any of these proteins can influence host immune responses.

Approximately 20% of the annotated genes in the *P. yoelii* genomes are members of multi-copy gene families. The number of the *yir* genes in 17X is much larger than that of N67 or YM. Although the observation of more *yir* genes in the 17X genome could be due to more complete and better annotated 17X genome, it is not surprising to find variation in the number of *yir* genes. The *yir* genes belongs to the *Plasmodium* interspersed repeat (*pir*) gene families. The *pir* genes are mostly distributed in the subtelomeric regions of chromosomes with gene copies numbering from a few dozen to hundreds or even over a thousand [42], and up to 40% of the *cir* gene (gene family from *P. chanbaudi*) repertoire are expressed during the intraerythrocytic cycle [43]. Many of the *pir* genes are expressed on the surface of iRBCs and merozoites and play an important role in immune evasion [44, 45]. In addition to the *yir* genes, the variation in number of genes encoding PcEMA1 homologs, haloacid dehalogenase-like hydrolases, reticulocyte binding proteins between N67 and 17X/YM are interesting, particularly the expansion of the gene encoding PcEMA1 homologs. These proteins likely play some important roles in host-parasite interaction and parasite development.

Conclusions

The rodent parasite N67 is an important subspecies of *P. yoelii* for studying host immune responses and parasite biology. The lack of the assembled genome and genome annotation has impeded functional studies of the parasite, including genetic mapping determinants playing important roles in modulating host IFN-I responses. This study provides the first assembled and annotated N67 parasite geneme, including prediction of 5,383 genes, although there are still many gaps in the genome. Comparison of the annotated genome of the N67 parasite with those of YM and 17X parasites reveals a large numbers of SNPs and indels that may have functional impact on parasite development and biology. Additionally, unique N67 gene sets, expansion of gene families, and genes potentially regulating host immune responses are also identified. Although further efforts such as manual curation are necessary to completely assemble the genome sequences, the assembled genome with over 5,000 predicted genes from this study will greatly facilitate our investigations of the parasite biology and mechanisms of the disease.

Methods

Parasite and infection of mice

The N67 parasite was initially obtained from MR4-BEI (https://www.beiresources.org/About/MR4.aspx) and were described previously (26). Inbred female C57BL/6 mice, aged 6–8 weeks old, were obtained from NIAID/Taconic repository. The procedures for infecting mice with the parasites were as reported previously (26, 27). Parasitemia was monitored by microscopic examination of Giemsa-stained thin blood smears.

DNA preparation for PacBio sequencing

Mice were injected *ip* with an inoculum containing 1×10^6 infected red blood cells (iRBCs). Blood samples (200 μl) with approximately 30-40% parasitemia were collected on day 4 after injection. Infected red blood cells collected in 1 ml 0.15% sodium citrate/PBS buffer were pelleted at 2,000 rpm for 5 min in an Eppendorf centrifuge, re-suspended in 1 ml of PBS, and passed through two consecutive NWF filters (Zhixing Bio, Bengbu, China) to remove the host white blood cells [19]. The flow-through cell suspension was washed in 800 μl PBS 3X through centrifugation at 3,000 rpm for 3 min. The pellet was dissolved in 400 μl lysis buffer (100 mM NaCl, 10 mM Tris, 25 mM EDTA, pH 8.0, 0.5% SDS) containing 20 μl RNase (500 μg/ml) and 20 μl protease K (10 mg/ml) and incubated at 50°C overnight. DNA was extracted using 400 μl phenol, chloroform, and isopropanol at a ratio of 25:24:1 and precipitated by adding 2 volume 100% ethanol overnight at -20°C. The sample was centrifuged at 13,000 rpm for 15 mins at 4°C, and the DNA pellet was washed in 500 μl 70% ethanol twice before addition of 20 μl water. The quality of DNA was estimated on 1% agarose gel showing a typical high molecular weight band.

Fragmentation of DNA and PacBio sequencing

A SMRTbell library was constructed using standard PacBio library preparation procedure (Pacific Biosciences, Menlo Park, CA, USA). The genomic DNA was fragmented with the majority of DNA fragments above 20 kb, then the DNA was carried into the first enzymatic reaction to remove single-stranded overhangs and tailed with an A-overhang. Ligation with T-overhang SMRTbell adapters was performed and the SMRTbell library was purified. The size and concentration of the final library were assessed.

Sequencing primer and Sequel DNA Polymerase were annealed and bound, respectively, to the SMRTbell library. The library was loaded on PacBio Sequel using diffusion loading. SMRT sequencing was performed on the Sequel System with Sequel Sequencing Kit 3.0, 1200 min movies. Quality control (QC) for raw reads (subreads) generated from the sequencer were performed by the default smrtlink QC pipeline. Pass-filter reads were then used as input for the genome assembly.

Genome sequence assembly

The genome was assembled using HGAP v4.0, a standard assembler from PacBio SMRTLink software (Pacific Biosciences, Menlo Park, CA, USA), Subreads longer than 6 kb were designated as "seed reads" and used as template sequences for preassembly/error correction. After assembly, two rounds of polishing were performed to increase the consensus sequence quality of the assembly, including aligning the PacBio data to the contigs and computing consensus using the Arrow consensus caller (SMRTLink).

RNA-Seq, transcriptome assembly and gene predictions

To assemble the N67 transcriptome, we extracted RNA from mixed-stage iRBC samples of eight mice infected with N67 and performed Illumina sequencing as reported previously [15]. The resulting RNA-Seq reads were trimmed with Trimmomatic [46] to remove the adapter sequences, and the reads were mapped to the 17X genome using the STAR aligner [47] and disambiguate [48] tools. RNA-Seq reads from the samples uniquely mapped to the 17X genome were retained for the transcriptome assembly step that was performed using Trinity [26].

Gene predictions were generated using the MAKER pipeline [28]. Specifically, MAKER utilized BLAST to align the *de novo* assembled N67 transcriptome and 17X transcriptome to the *de novo* assembled N67 genome, polished these alignments using Exonerate in a splice-aware fashion, and implemented SNAP and Augustus hidden Markov models (HMMs) to generate *ab initio* gene models [49, 50]. Functional analysis and annotation was performed with InterProScan [51] after homology searches of over 15 databases including Pfam, ProSite, TIGRFAM, and PANTHER. Our final set of N67 genes/proteins only include those that are adequately supported by the assembled N67 genome/transcriptome and the 17X proteome. For each predicted N67 protein, we computed two AED values (AED/eAED: at the base pair and exon levels) to quantify how well each N67 protein is supported by these data sources [52].

Estimates of completeness

We quantified the completeness of the *de novo* assembled N67 transcriptome, genome, and proteome using BUSCO [53] against the single-copy orthologs conserved among *Apicomplexa* (446 BUSCOs) and *Plasmodium* (3642 BUSCOs) from the OrthoDB v10.1 database [54].

Identification of orthologs

We used the Orthofinder [34] framework for identifying the ortholog sets among the 17X, YM and N67 proteins. Orthofinder utilizes DIAMOND [33] for identifying sequence similarity and DendroBLAST [55] for gene tree inference.

Gene family clustering

476 Hierarchical clustering analyses were performed using MEGAX [56]. Protein sequences of YIR,
477 Fam-A as well as proteins from the N67-specific orthogroups and the ones not assigned to any
478 orthogroup were aligned using the ClustalW algorithm. The maximum likelihood method and
479 Jones-Taylor-Thornton (JTT) matrix-based model were used to construct cladograms from the
480 aligned sequences [57].
481

482 Availability of data and materials

483 This sequencing data and assembled genomes have been deposited to GenBank with Accession
484 number JAEVLW000000000.
485

486 Abbreviation

487 YM, *Plasmodium y. yoelii* YM
488 17X, *Plasmodium y. yoelii* 17X
489 17XNL, *Plasmodium y. yoelii* 17XNL
490 17XL, *Plasmodium y. yoelii* 17XL
491 N67, *Plasmodium y. nigeriensis* N67
492 IFN-I, type I interferon
493 PyEBL, *P. yoelii* erythrocyte binding-like protein
494 SNPs, single nucleotide polymorphisms
495 HGAP, hierarchical genome assembly process
496 CCS, circular consensus sequencing
497 AED, annotation edit distance
498 pi, post infection

PcEMA1, *P. chabaudi* erythrocyte membrane protein 1

MMEJ, microhomology-mediated end joining

BUSCO, benchmarking universal single-copy orthologs

ETRAMPs, early transcribed membrane proteins

VEP, variant effect predictor

ISGs, interferon-stimulated genes

pir, interspersed repeat

yir, *yoelii* interspersed repeats

iRBCs, red blood cells

QC, Quality control

HMMs, augustus hidden markov models

LCB, co-linear block

References

1. WHO: **World malaria report 2019**. *https://wwwwhoint/malaria/publications/world-malaria-report-2019/en/* 2019.
2. Craig AG, Grau GE, Janse C, Kazura JW, Milner D, Barnwell JW, Turner G, Langhorne J, participants of the Hinxton Retreat meeting on Animal Models for Research on Severe M: **The role of animal models for research on severe malaria**. *PLoS Pathog* 2012, **8**(2):e1002401.
3. Langhorne J, Buffet P, Galinski M, Good M, Harty J, Leroy D, Mota MM, Pasini E, Renia L, Riley E *et al*: **The relevance of non-human primate and rodent malaria models for humans**. *Malar J* 2011, **10**:23.
4. Li J, Zhang Y, Sullivan M, Hong L, Huang L, Lu F, McCutchan TF, Su XZ: **Typing *Plasmodium yoelii* microsatellites using a simple and affordable fluorescent labeling method**. *Mol Biochem Parasitol* 2007, **155**(2):94-102.

5. Li J, Pattaradilokrat S, Zhu F, Jiang H, Liu S, Hong L, Fu Y, Koo L, Xu W, Pan W *et al*: **Linkage maps from multiple genetic crosses and loci linked to growth-related virulent phenotype in *Plasmodium yoelii***. *Proc Natl Acad Sci U S A* 2011, **108**(31):E374-382.

6. Nair SC, Pattaradilokrat S, Zilversmit MM, Dommer J, Nagarajan V, Stephens MT, Xiao W, Tan JC, Su XZ: **Genome-wide polymorphisms and development of a microarray platform to detect genetic variations in *Plasmodium yoelii***. *Mol Biochem Parasitol* 2014, **194**(1-2):9-15.

7. Pattaradilokrat S, Cheesman SJ, Carter R: **Congenicity and genetic polymorphism in cloned lines derived from a single isolate of a rodent malaria parasite**. *Mol Biochem Parasitol* 2008, **157**(2):244-247.

8. Pattaradilokrat S, Li J, Wu J, Qi Y, Eastman RT, Zilversmit M, Nair SC, Huaman MC, Quinones M, Jiang H *et al*: **Plasmodium genetic loci linked to host cytokine and chemokine responses**. *Genes Immun* 2014, **15**(3):145-152.

9. Martin-Jaular L, Ferrer M, Calvo M, Rosanas-Urgell A, Kalko S, Graewe S, Soria G, Cortadellas N, Ordi J, Planas A *et al*: **Strain-specific spleen remodelling in *Plasmodium yoelii* infections in Balb/c mice facilitates adherence and spleen macrophage-clearance escape**. *Cell Microbiol* 2011, **13**(1):109-122.

10. Bakir HY, Sayed FG, Rahman SA, Hamza AI, Mahmoud AE, Galal LA, Attia RA: **Comparative study between non lethal and lethal strains of *Plasmodium yoelii* with reference to its immunological aspect**. *J Egypt Soc Parasitol* 2009, **39**(2):585-593.

11. Wykes MN, Liu XQ, Beattie L, Stanisic DI, Stacey KJ, Smyth MJ, Thomas R, Good MF: ***Plasmodium* strain determines dendritic cell function essential for survival from malaria**. *PLoS Pathog* 2007, **3**(7):e96.

12. Otsuki H, Kaneko O, Thongkukiatkul A, Tachibana M, Iriko H, Takeo S, Tsuboi T, Torii M: **Single amino acid substitution in *Plasmodium yoelii* erythrocyte ligand determines its localization and controls parasite virulence**. *Proc Natl Acad Sci U S A* 2009, **106**(17):7167-7172.

13. Wu J, Tian L, Yu X, Pattaradilokrat S, Li J, Wang M, Yu W, Qi Y, Zeituni AE, Nair SC *et al*: **Strain-specific innate immune signaling pathways determine malaria**

parasitemia dynamics and host mortality. *Proc Natl Acad Sci U S A* 2014, **111**(4):E511-520.

14. Lacerda-Queiroz N, Riteau N, Eastman RT, Bock KW, Orandle MS, Moore IN, Sher A, Long CA, Jankovic D, Su XZ: **Mechanism of splenic cell death and host mortality in a *Plasmodium yoelii* malaria model**. *Sci Rep* 2017, **7**(1):10438.

15. Peng YC, Qi Y, Zhang C, Yao X, Wu J, Pattaradilokrat S, Xia L, Tumas KC, He X, Ishizaki T *et al*: ***Plasmodium yoelii* Erythrocyte-Binding-like Protein Modulates Host Cell Membrane Structure, Immunity, and Disease Severity**. *mBio* 2020, **11**(1).

16. Carlton JM, Angiuoli SV, Suh BB, Kooij TW, Pertea M, Silva JC, Ermolaeva MD, Allen JE, Selengut JD, Koo HL *et al*: **Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii***. *Nature* 2002, **419**(6906):512-519.

17. Vaughan A, Chiu SY, Ramasamy G, Li L, Gardner MJ, Tarun AS, Kappe SH, Peng X: **Assessment and improvement of the *Plasmodium yoelii yoelii* genome annotation through comparative analysis**. *Bioinformatics* 2008, **24**(13):i383-389.

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

19. Li J, Cai B, Qi Y, Zhao W, Liu J, Xu R, Pang Q, Tao Z, Hong L, Liu S *et al*: **UTR introns, antisense RNA and differentially spliced transcripts between *Plasmodium yoelii* subspecies**. *Malar J* 2016, **15**:30.

20. Kooij TW, Carlton JM, Bidwell SL, Hall N, Ramesar J, Janse CJ, Waters AP: **A *Plasmodium* whole-genome synteny map: indels and synteny breakpoints as foci for species-specific genes**. *PLoS Pathog* 2005, **1**(4):e44.

21. Nakano K, Shiroma A, Shimoji M, Tamotsu H, Ashimine N, Ohki S, Shinzato M, Minami M, Nakanishi T, Teruya K *et al*: **Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area**. *Hum Cell* 2017, **30**(3):149-161.

22. Aurrecoechea C, Brestelli J, Brunk BP, Dommer J, Fischer S, Gajria B, Gao X, Gingle A, Grant G, Harb OS *et al*: **PlasmoDB: a functional genomic database for malaria parasites**. *Nucleic Acids Res* 2009, **37**(Database issue):D539-543.

23. Li H: **Minimap2: pairwise alignment for nucleotide sequences**. *Bioinformatics* 2018, **34**(18):3094-3100.
24. Darling AE, Mau B, Perna NT: **progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement**. *PLoS One* 2010, **5**(6):e11147.
25. Wingett SW, Andrews S: **FastQ Screen: A tool for multi-genome mapping and quality control**. *F1000Res* 2018, **7**:1338.
26. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q *et al*: **Full-length transcriptome assembly from RNA-Seq data without a reference genome**. *Nat Biotechnol* 2011, **29**(7):644-652.
27. Langdon WB: **Performance of genetic programming optimised Bowtie2 on genome comparison and analytic testing (GCAT) benchmarks**. *BioData Min* 2015, **8**(1):1.
28. Campbell MS, Holt C, Moore B, Yandell M: **Genome Annotation and Curation Using MAKER and MAKER-P**. *Curr Protoc Bioinformatics* 2014, **48**:4 11 11-39.
29. Deleersnijder W, Prasomsitti P, Tungpradubkul S, Hendrix D, Hamers-Casterman C, Hamers R: **Structure of a *Plasmodium chabaudi* acidic phosphoprotein that is associated with the host erythrocyte membrane**. *Mol Biochem Parasitol* 1992, **56**(1):59-68.
30. Xu R, Liu Y, Fan R, Liang R, Yue L, Liu S, Su XZ, Li J: **Generation and functional characterisation of Plasmodium yoelii csp deletion mutants using a microhomology-based CRISPR/Cas9 method**. *Int J Parasitol* 2019, **49**(9):705-714.
31. Hakimi MA, Bougdour A: **Toxoplasma's ways of manipulating the host transcriptome via secreted effectors**. *Curr Opin Microbiol* 2015, **26**:24-31.
32. Emms DM, Kelly S: **OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy**. *Genome Biol* 2015, **16**:157.
33. Buchfink B, Xie C, Huson DH: **Fast and sensitive protein alignment using DIAMOND**. *Nat Methods* 2015, **12**(1):59-60.
34. Emms DM, Kelly S: **OrthoFinder: phylogenetic orthology inference for comparative genomics**. *Genome Biol* 2019, **20**(1):238.

35. Frech C, Chen N: **Variant surface antigens of malaria parasites: functional and evolutionary insights from comparative gene family classification and analysis**. *BMC Genomics* 2013, **14**:427.
36. Nair SC, Xu R, Pattaradilokrat S, Wu J, Qi Y, Zilversmit M, Ganesan S, Nagarajan V, Eastman RT, Orandle MS *et al*: **A *Plasmodium yoelii* HECT-like E3 ubiquitin ligase regulates parasite growth and virulence**. *Nat Commun* 2017, **8**(1):223.
37. Wu J, Cai B, Sun W, Huang R, Liu X, Lin M, Pattaradilokrat S, Martin S, Qi Y, Nair SC *et al*: **Genome-wide Analysis of Host-*Plasmodium yoelii* Interactions Reveals Regulators of the Type I Interferon Response**. *Cell Rep* 2015, **12**(4):661-672.
38. Dutta GP, Bajpai R, Vishwakarma RA: **Antimalarial efficacy of arteether against multiple drug resistant strain of *Plasmodium yoelii nigeriensis***. *Pharmacol Res* 1989, **21**(4):415-419.
39. Orfano AS, Duarte AP, Molina-Cruz A, Pimenta PF, Barillas-Mury C: ***Plasmodium yoelii nigeriensis* (N67) Is a Robust Animal Model to Study Malaria Transmission by South American Anopheline Mosquitoes**. *PLoS One* 2016, **11**(12):e0167178.
40. Beaute-Lafitte A, Altemayer-Caillard V, Chabaud AG, Landau I: **Plasmodium yoelii nigeriensis: biological mechanisms of resistance to chloroquine**. *Parasite* 1994, **1**(3):227-233.
41. Graves PM, Curtis CF: **Susceptibility of Anopheles gambiae to *Plasmodium yoelii nigeriensis* and *Plasmodium falciparum***. *Ann Trop Med Parasitol* 1982, **76**(6):633-639.
42. Cunningham D, Lawton J, Jarra W, Preiser P, Langhorne J: **The pir multigene family of Plasmodium: antigenic variation and beyond**. *Mol Biochem Parasitol* 2010, **170**(2):65-73.
43. Lawton J, Brugat T, Yan YX, Reid AJ, Bohme U, Otto TD, Pain A, Jackson A, Berriman M, Cunningham D *et al*: **Characterization and gene expression analysis of the cir multi-gene family of Plasmodium chabaudi chabaudi (AS)**. *BMC Genomics* 2012, **13**:125.
44. Yam XY, Brugat T, Siau A, Lawton J, Wong DS, Farah A, Twang JS, Gao X, Langhorne J, Preiser PR: **Characterization of the Plasmodium Interspersed Repeats (PIR) proteins of Plasmodium chabaudi indicates functional diversity**. *Sci Rep* 2016, **6**:23449.

45. Yam XY, Preiser PR: **Host immune evasion strategies of malaria blood stage parasite**. *Mol Biosyst* 2017, **13**(12):2498-2508.
46. Bolger AM, Lohse M, Usadel B: **Trimmomatic: a flexible trimmer for Illumina sequence data**. *Bioinformatics* 2014, **30**(15):2114-2120.
47. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: **STAR: ultrafast universal RNA-seq aligner**. *Bioinformatics* 2013, **29**(1):15-21.
48. Ahdesmaki MJ, Gray SR, Johnson JH, Lai Z: **Disambiguate: An open-source application for disambiguating two species in next generation sequencing data from grafted samples**. *F1000Res* 2016, **5**:2741.
49. Korf I: **Gene finding in novel genomes**. *BMC Bioinformatics* 2004, **5**:59.
50. Stanke M, Steinkamp R, Waack S, Morgenstern B: **AUGUSTUS: a web server for gene finding in eukaryotes**. *Nucleic Acids Res* 2004, **32**(Web Server issue):W309-312.
51. Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, Chang HY, Dosztanyi Z, El-Gebali S, Fraser M *et al*: **InterPro in 2017-beyond protein family and domain annotations**. *Nucleic Acids Res* 2017, **45**(D1):D190-D199.
52. Eilbeck K, Moore B, Holt C, Yandell M: **Quantitative measures for the management and comparison of annotated genomes**. *BMC Bioinformatics* 2009, **10**:67.
53. Waterhouse RM, Seppey M, Simao FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM: **BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics**. *Mol Biol Evol* 2018, **35**(3):543-548.
54. Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simao FA, Zdobnov EM: **OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs**. *Nucleic Acids Res* 2019, **47**(D1):D807-D811.
55. Kelly S, Maini PK: **DendroBLAST: approximate phylogenetic trees in the absence of multiple sequence alignments**. *PLoS One* 2013, **8**(3):e58537.
56. Kumar S, Stecher G, Li M, Knyaz C, Tamura K: **MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms**. *Mol Biol Evol* 2018, **35**(6):1547-1549.
57. Jones DT, Taylor WR, Thornton JM: **The rapid generation of mutation data matrices from protein sequences**. *Comput Appl Biosci* 1992, **8**(3):275-282.

Acknowledements

The authors thank Brigit Shea Sullivan, NIH Library Editing Service, for manuscript editing assistance. We alos thank Dr. Jian Li, Xiamen University of China, for providing the NWF filters.

Funding

This work was supported by the Division of Intramural Research, National Institute of Allergy and Infectious Diseases (NIAID), National Institutes of Health (NIH), USA.

Author information

Affiliations

Malaria Functional Genomics Section, Laboratory of Malaria and Vector Research, National Institute of Allergy and Infectious Disease, National Institutes of Health, Bethesda, MD 20892-8132, USA

Cui Zhang, Lu Xia, Jian Wu, Yu-Chih Peng, Margaret Smith, Carole A. Long & Xin-zhuan Su

NIAID Collaborative Bioinformatics Resource (NCBR), Frederick National Laboratory for Cancer Research, Leidos Biomedical Research, Inc., Frederick, MD 21702, USA

Cihan Oguz, Sue Huse & Justin Lack

Curret address: State Key Laboratory of Medical Genetics, Xiangya School of Medicine, Central South University, Changsha, Hunan 410078, The People's Republic of China.

Lu Xia

The NCI sequencing facility, 8560 Progress Drive, Room 3007, Frederick Md 21701, USA

Jack Chen

Contributions

CZ, LX, Y-CP, JW, MS, parasite infection of mice, DNA preparation, data analysis; JC, PacBio sequencing, CO, SH, JL, data analysis and writing; CAL and X-zS supervision and writing; Xz-S project conception.

Corresponding author

Correspondence to Xinzhuan Su.

Ethics declarations

Ethics approval and consent to participate

Plasmdoium yoelii nigeriensis N67 is a malaria parasite that infects red blood cells of rodents, including mice. This study involves infection of C57BL/6 mice with the parasite to obtain parasite genomic DNA for sequencing. The experiments were performed in accordance with the protocol approved (approval #LMVR11E) by the Institutional Animal Care and Use Committee (IACUC) at the National Institute of Allergy and Infectious Diseases (NIAID) following the guidelines of the Public Health Service Policy on Humane Care and Use of Laboratory Animals

and The Association for Assessment and Accreditation of Laboratory Animal Care International (AAALAC). The study was carried out in compliance with the ARRIVE guidelines.

Consent for publication

Not applicable.

Conflict of interests

The authors declare that they have no competing interests.

Supplementary information

Additional file 1. Fig. S1-2.

Additional file 2. Table S1.

Additional file 3. Table S2.

Additional file 4. Table S3.

Additional file 5. Table S4.

Additional file 6. Table S5.

Additional file 7. Table S6.

Additional file 8. Table S7.

Additional file 9. Table S8.

Additional file 10. Table S9.

Additional file 11. Table S10.

Additional file 12. Table S11.

Figures

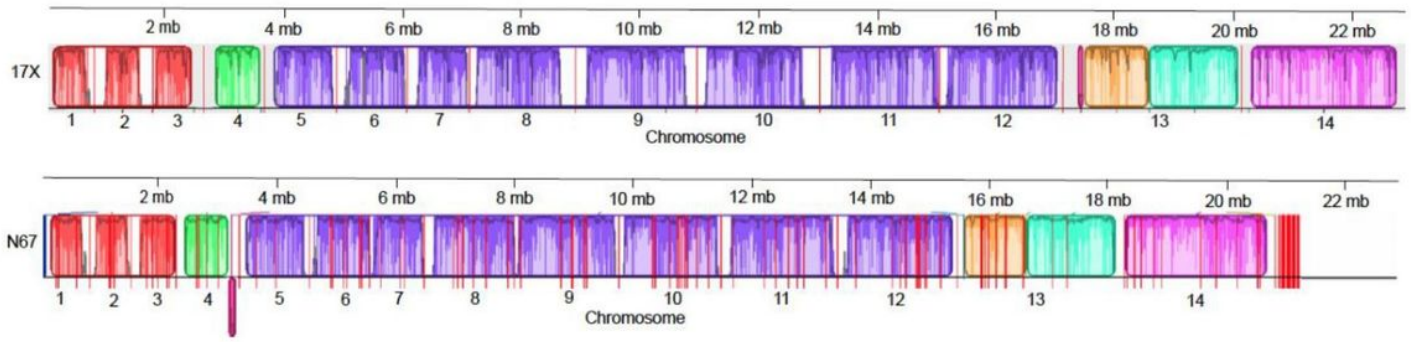


Figure 1

Alignment of Hierarchical Genome Assembly Process (HGAP) assembled N67 contigs to the 17X chromosomes. The alignments were generated using progressiveMauve. Each color corresponds to a localized co-linear block (LCB) that is conserved across the two genomes. Inside each LCB, the jagged dark lines represent the similarity profile; with darker colors representing higher similarity regions. The vertical red lines indicate chromosome boundaries in 17X and the contig boundaries on the N67 sequences. Note a contig on N67 chromosome 4 that is inverted (presented under the chromosome line) in reference to that of 17X sequence.

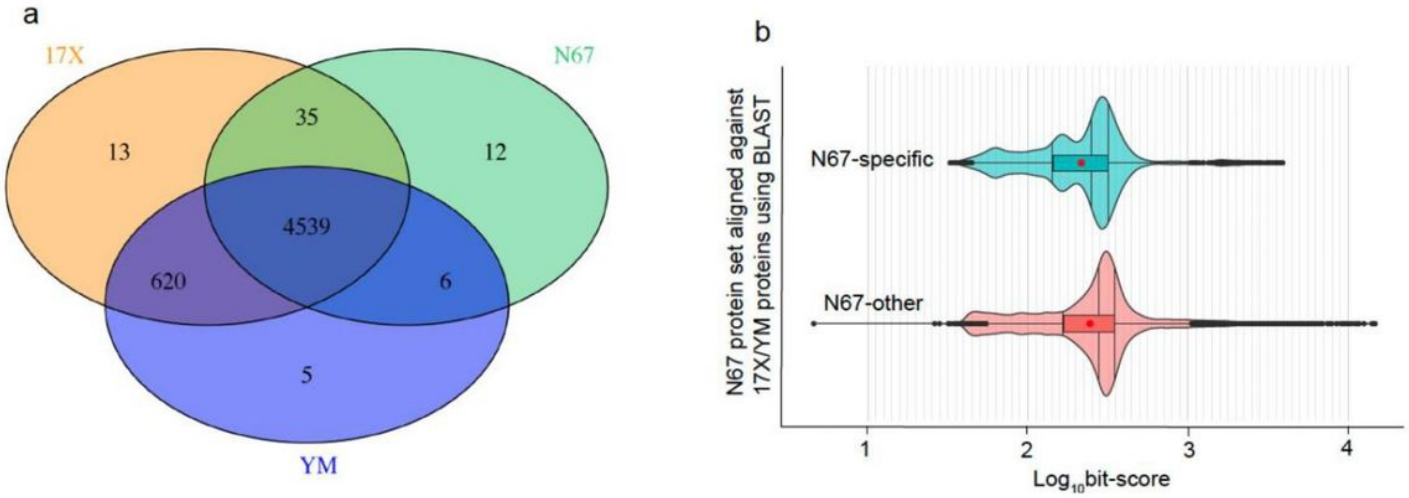


Figure 2

Shared and strain-specific orthogroup counts identified from Plasmodium y. yoelii YM, P. y. yoelii 17X, and P. y. nigeriensis N67 parasites using OrthoFinder [32]. a, Venn diagram of shared and strain-specific orthogroups; b, log10-transformed bit-score distributions for N67 proteins that are not assigned to any orthogroup plus those in N67-specific orthogroups (N67- specific) and proteins assigned to orthogroups having at least one 17X or YM protein (N67- other). The bit-scores are derived from pairwise BLAST

alignments within the Orthofinder framework, where all queries were N67 protein sequences that were aligned against the 17X and YM sequences. The red dots indicate the mean values of bit-score distributions, whereas the vertical lines within the violins indicate the median, upper and lower quartile values.

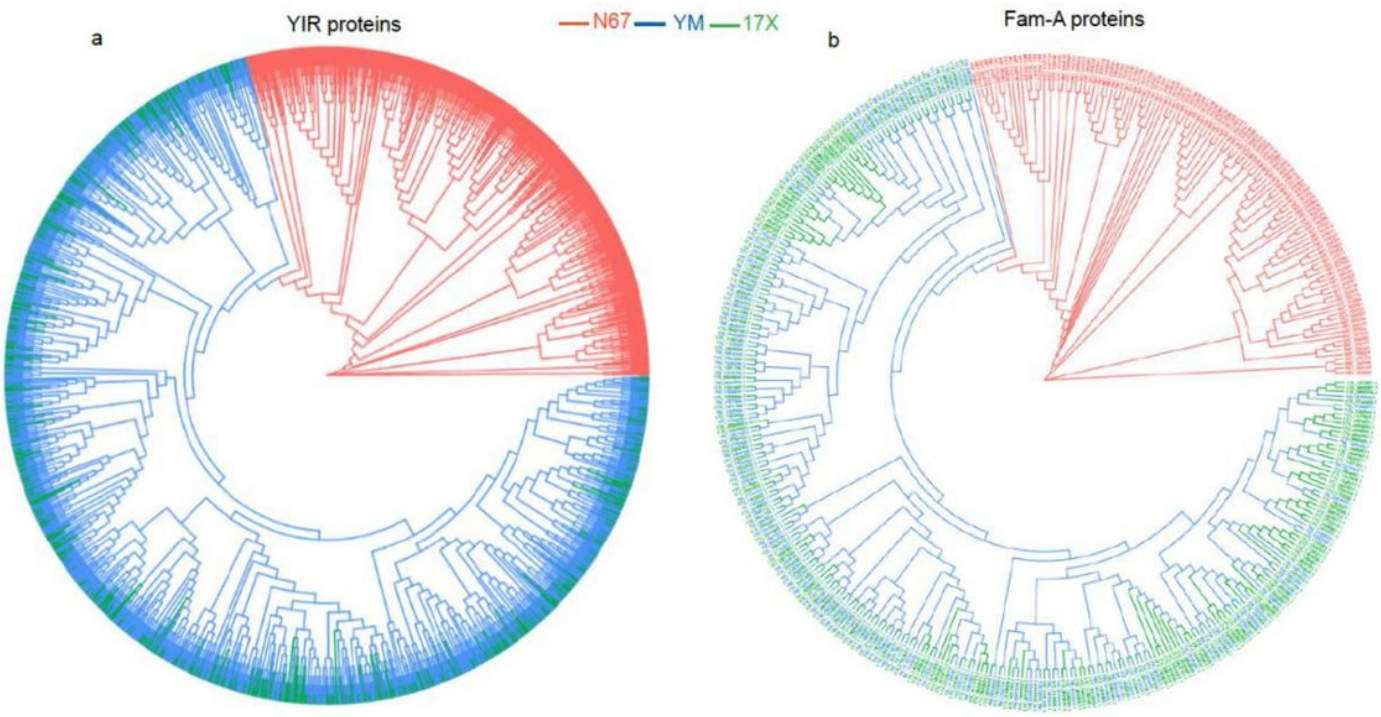


Figure 3

Clustering of YIR and Fam-A proteins from the Plasmodium y. nigeriensis N67, P. y. yoelii 17X, and P. y. yoelii YM parasites. The predicted protein sequences were aligned using ClustalW algorithm in msa R package, and the dendrograms were inferred and visualized using the ape, seqinr, and ggtree packages in R. a, YIR proteins from N67, 17X, and YM parasites; b, Fam-A proteins. Proteins are colored based on their parasite origins.

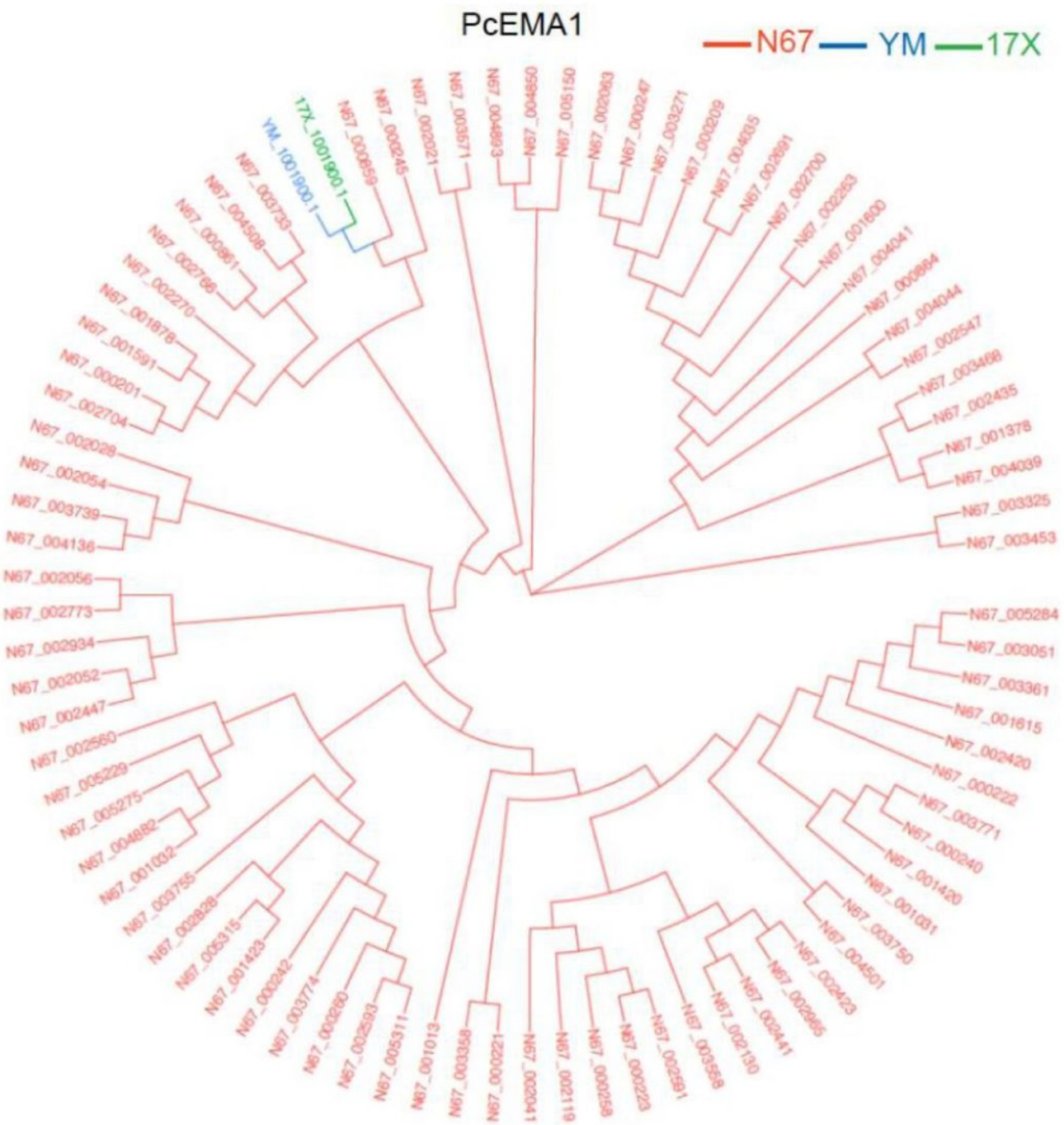


Figure 4

Clustering of homologous P. chabaudi erythrocyte membrane protein 1 (PcEMA1) from Plasmodium y. nigeriensis N67, P. y. yoelii 17X and P. y. yoelii YM parasites. The gene family is expanded only in the N67 parasite. The predicted protein sequences were aligned using ClustalW algorithm in msa R package, and the dendrogram was inferred and visualized using the ape, seqinr, and ggtree packages in R. Proteins are colored based on their parasite origins.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Table1.xlsx
- Table2.xlsx
- BNCN67Supplementalfiles.pdf
- TableS1.xlsx
- TableS10.xlsx
- TableS11..xlsx
- TableS2.xlsx
- TableS3.xlsx
- TableS4.xlsx
- TableS5.xlsx
- TableS6.xlsx
- TableS7.xlsx
- TableS8.xlsx
- TableS9.xlsx