

1 **Genome sequence, transcriptome, and annotation of rodent malaria parasite**

2 ***Plasmodium yoelii nigeriensis* N67**

3

4 Cui Zhang^{1*}, Cihan Oguz^{2*}, Sue Huse², Lu Xia^{1,3}, Jian Wu¹, Yu-Chih Peng¹, Margaret Smith¹,

5 Jack Chen⁴, Carole A. Long¹, Justin Lack², and Xin-zhuan Su^{1#}

6

7 ¹*Malaria Functional Genomics Section, Laboratory of Malaria and Vector Research, National*

8 *Institute of Allergy and Infectious Disease, National Institutes of Health, Bethesda, MD 20892-*

9 *8132, USA;*

10 ²*NIAID Collaborative Bioinformatics Resource (NCBR), Frederick National Laboratory for*

11 *Cancer Research, Leidos Biomedical Research, Inc., Frederick, MD 21702, USA*

12 ³*State Key Laboratory of Medical Genetics, Xiangya School of Medicine, Central South*

13 *University, Changsha, Hunan 410078, The People's Republic of China.*

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

15

16 *Co-first authors

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

18

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

20

21

22

23

24 **Abstract**

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

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

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

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

48

49 **Background**

50 Malaria is one of the deadly tropical infectious diseases that impacts the health of hundreds of
51 millions of people [1]. The lack of an effective vaccine, emergence of drug resistant parasites
52 and insecticide resistant mosquitoes, and incomplete understanding of the disease mechanisms
53 are the major factors that impede disease control and elimination. Vaccine development and in
54 depth studies of disease molecular mechanisms using human populations are limited by ethical
55 regulations and relatively high costs. Animal disease models such as parasites infecting rodents
56 and non-human primates are important systems for studying malaria and have been widely used
57 for vaccine development and for studying the molecular mechanisms of host-parasite interaction
58 [2, 3]. Of course, results obtained from animal models need to be verified in human infection
59 because there are differences in disease mechanism due to variation in genetic backgrounds of
60 both the parasites and the hosts.

61

62 *Plasmodium yoelii* is one of the rodent malaria species that includes several parasite strains or
63 subspecies well-characterized genetically and phenotypically [4, 5]. Some of the *P. yoelii* strains
64 are genetically diverse, whereas others are closely related or derived from a common ancestor
65 during laboratory passages in mice [4, 6, 7]. Mice infected with these *P. yoelii* strains generally
66 have dramatic differences in parasitemia, disease severity, pathology, and host immune response
67 [8]. For example, *P. y. yoelii* 17X (or 17XNL) and *P. y. yoelii* 17XL (or YM) are closely related
68 parasites genetically. Indeed, these parasites were derived from a parasite isolated from a wild
69 thicket rat in the Central African Republic [7]. 17XL and YM lines became fast growing and

70 lethal during passages of 17X parasites in mice in two separate laboratories, whereas 17X or
71 17XNL remained slow growing and non-lethal [7]. These parasites also stimulate different host
72 responses and pathology [9-12]. Another example of parasites having closely related genomes
73 but with different virulence is the *P. y. nigeriensis* N67 and *P. y. nigeriensis* N67C parasite pair.
74 The genomes of N67 and N67C are very similar [5, 6]; however, they produce quite different
75 disease phenotypes in C57BL/6 mice. Infection of N67 stimulates a strong early type I interferon
76 (IFN-I) response, leading to a decline of parasitemia to below 5% day 7 post infection (pi). The
77 parasitemia rebounds to 50-60%, and the host dies at day 20 pi [13]. In contrast, mice infected
78 with N67C produce a strong T cell and INF- γ mediated inflammatory responses and die day 7 pi
79 [14]. A C741Y amino acid substitution in the *P. yoelii* erythrocyte binding-like protein (PyEBL)
80 contributes to the differences in virulence and immune response, but other parasite genes also
81 play a role in the differences in disease phenotypes [15]. Identification of the genes or genetic
82 differences between N67 and N67C parasite will facilitate our understanding of the molecular
83 mechanisms of virulence and disease phenotypes in these infections.

84

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

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

103

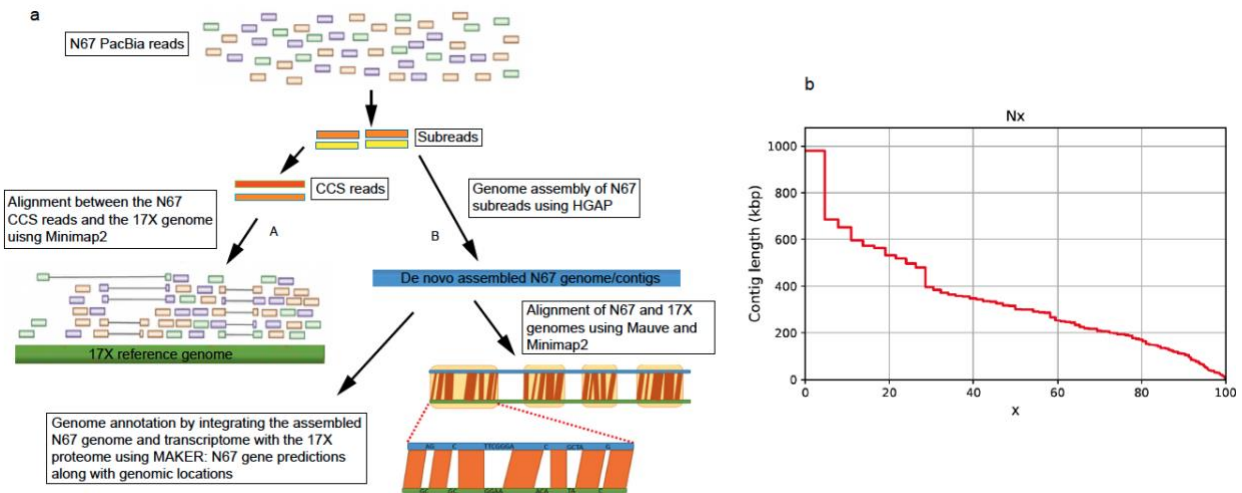
104 **Results**

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

106 We prepared DNA samples for PacBio sequencing from the N67 parasite we obtained previously
107 [5]. Genomic DNAs were fragmented and sequenced on a PacBio Sequel using PacBio SMRT
108 cell long read technology [21]. The polymerase reads from sequencing machine were first
109 filtered to remove barcodes and low-quality sequences using the Hierarchical Genome Assembly
110 Process (HGAP) (**Fig. S1a**). We obtained 1,111,721 subreads consisting of 6,733,837,360 bp for
111 the N67 parasite, providing 233 mean coverages with an averaged barcode quality of 72. The
112 longest subread length was 195,628 bp and the mean read length was 70,695 bp. The subreads
113 were then assembled into 61,130 circular consensus sequencing (CCS) reads with a mean CCS
114 coverage of 13.5-fold for the parasite.

115

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



121
 122 **Fig. S1.** Strategies of genome assembly and annotation, and plot of contig length distributions of
 123 the *Plasmodium y. yoelii* N67 parasite genome assembly. **a**, Diagram illustrating the processes of
 124 aligning N67 CCS reads (A) and contigs (B) to the 17X genomes. HGAP, Hierarchical genome
 125 assembly process; CCS read, circular consensus sequencing read. **b**, Plot of contig length
 126 distribution. The X-axis is percentage of the contigs with lengths (base pair) greater than the
 127 values indicated on the Y-axis.

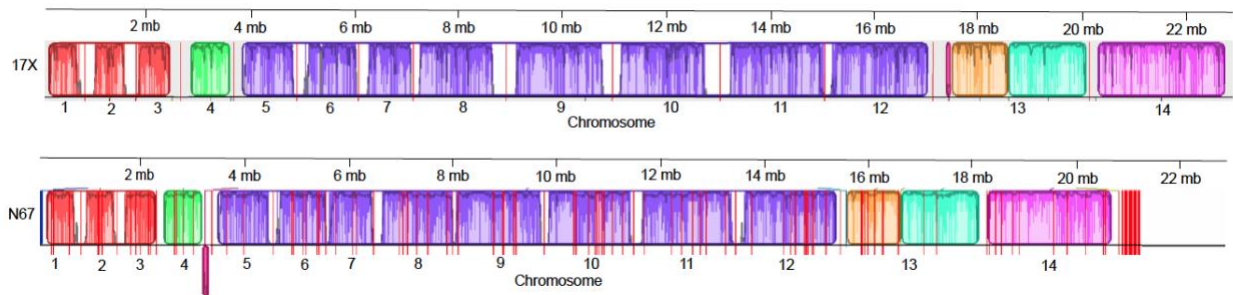
128

129 **Alignment of N67 sequences to the 17X assembled genome**

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

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

145



146

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

152 17X and the contig boundaries on the N67 sequences. Note a contig on N67 chromosome 4 that
153 is inverted (presented under the chromosome line) in reference to that of 17X sequence.

154

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

156 To facilitate genome annotation and gene prediction, we also sequenced mRNA of blood stages
157 from eight mice infected with N67 Illumina sequencing method. Overall, 82.9% of the RNA-Seq
158 reads from the samples uniquely mapped to the 17X genome and were retained for transcriptome
159 assembly, The majority of the remaining reads were either uniquely mapped to the mouse
160 genome (4.7%) or did not map to any genomes of human, mouse, bacteria, fungi and virus
161 (9.3%) based on results from the FastQ Screen [25]. The rest of 3.1% of reads were mapped to
162 human, fungi, bacteria or multiple genomes. We then used Trinity [26] to perform *de novo*
163 transcriptome assembly and obtained 25,689 transcripts containing 39,856,633 bp with an
164 average GC content of 23.5% (**Table S1**). The N50 was 1,952 bp with the largest transcript
165 being 19,550 bp. The N67 Illumina reads were aligned to the *de novo* assembled *P. yoelii*
166 transcriptome using Bowtie2 [27], resulting in 95.4% of the N67 read pairs concordantly aligned
167 to the assembled transcriptome, showing a high level of overall read support for the assembly.

168

169 **Gene predictions and functional annotation**

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

175 Additionally, more than 50% of the proteome should ideally contain a recognizable protein
176 domain for a well-annotated proteome [28]. Ninety-two percent of the predicted N67 proteins
177 have recognizable domains and/or are assigned to protein families. Furthermore, the smallest
178 predicted N67 protein has 16 amino acids (N67_005372, **Table S2**), similar to the smallest
179 predicted protein of 15 amino acids in the 17X proteome. Search of N67_005372 protein
180 sequence (MRVNKYVSVNMKMNYT) against the 17X and YM proteome did not return any
181 hit; however, it has a 79% sequence identity to serine hydroxymethyltransferase of
182 thermoacidophilic archaea *Thermoplasma volcanium*.
183
184 Search of the N67 proteome against InterPro database
185 (<https://www.ebi.ac.uk/interpro/search/sequence/>) of protein families, domains and functional
186 sites using InterProScan revealed that the largest five groups of proteins were the YIR antigens
187 (750 members), P-loop containing nucleoside triphosphate hydrolases (272), subtelomeric
188 PYST-A proteins (118), WD40-repeat-containing domain superfamily (89), and homologous
189 proteins of *P. chabaudi* erythrocyte membrane protein 1 (PcEMA1) (83) (**Table S3**).
190 Interestingly, the PcEMA1 was initially described from *P. chabaudi* parasites as an acidic
191 phosphoprotein that might modulate the structure of the red cell membrane to the advantage of
192 the parasite [29]. It has two tandem repeats (16×8 AA and 2×9 AA) that may mediate genetic
193 recombination and gene member expansion possibly through microhomology-mediated end
194 joining (MMEJ) [30]. The expansion of this gene family in N67 parasites suggests that the
195 PcEMA1 proteins may play a role in interaction with host immune system. Some other
196 interesting groups included 43 proteins with DEAD/DEAH box helicase domain, 22 proteins
197 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**

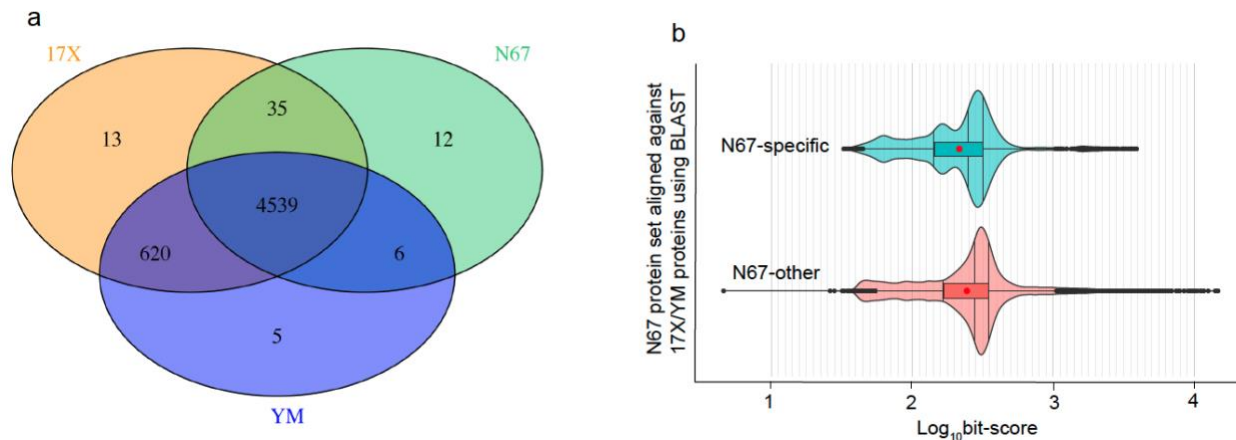
217 We next used Benchmarking Universal Single-Copy Orthologs (BUSCO 3.0.2) to assess the
218 completeness of the assembled N67 genome. Of the 3,642 *Plasmodium* and 446 *Apicomplexa*
219 BUSCO gene sets, 3,369 (92.5%) and 431 (96.6%) were present in the N67 genome assembly,
220 respectively (**Table S6**). We also evaluated the extent to which the assembled N67 transcriptome

221 matched the BUSCO gene sets across the *Apicomplexa* and *Plasmodium*. Approximately 92.8%
222 of the BUSCO *Apicomplexa* gene set and 71.3% of the *Plasmodium* gene set were present in the
223 assembled N67 transcriptome (**Table S6**). The N67 transcriptome genes matching the
224 *Plasmodium* BUSCO gene set included 1,448 complete and single-copy genes (39.8%), 1,149
225 (31.5%) complete and duplicated genes, and 338 fragmented sequences (9.3%) (**Table S6**).
226 There were also 707 genes (19.4%) missing from the *Plasmodium* BUSCO gene set; some of the
227 missed genes might not be expressed in the blood stages. The genomic sequences from PacBio
228 sequencing appear to provide more complete gene assembly than those from short Illumina RNA
229 seq data.

230

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

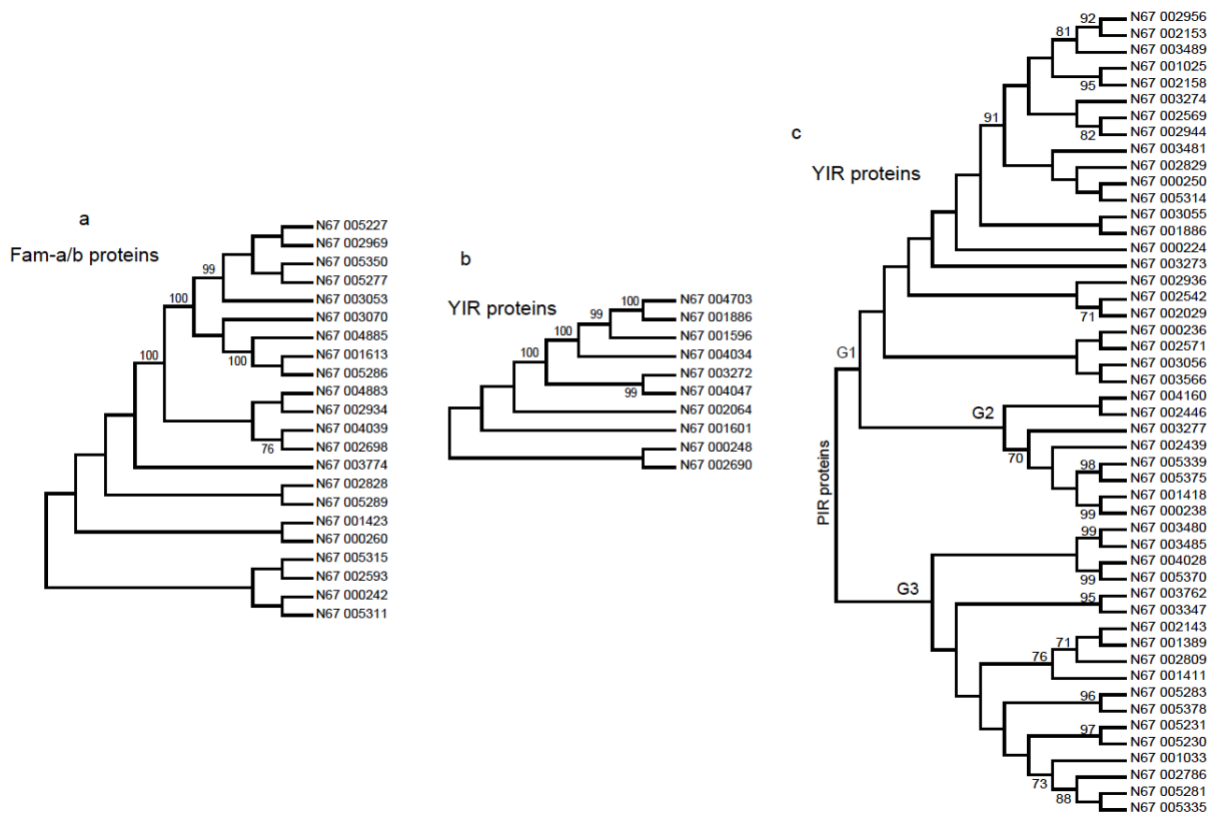
232 The N67 and 17X (or YM) parasites belong to two subspecies of *P. yoelii*, and the genomes of
233 these parasites are quite diverse [4, 6]. It is potentially interesting to identify genes common and
234 unique (or highly diverse) in these parasite genomes. Therefore, we compared the 5,383
235 predicted proteins from N67 with 6,092 17X proteins and 5,685 YM proteins using OrthoFinder
236 [32] and identified a core set of 4,539 orthogroups shared among the N67, 17X, and YM
237 genomes (**Fig. 2a**). Out of a total 17,160 proteins from the three parasite strains, 17,035 (99.3%)
238 were placed in 5,230 orthogroups based on searches of sequence similarity using DIAMOND
239 within the latest OrthoFinder framework [33, 34]. Of the 5,383 N67 proteins, 5,294 were
240 assigned to orthogroups, including 110 in 12 N67 specific orthogroups (**Table S7** and **Table S8**).
241 There were also 89 proteins that could not be assigned to any orthogroup, leading to a total of
242 199 proteins that appear to have N67-specific orthology. These proteins had slightly lower
243 pairwise bit-scores than those assigned to the common orthogroups (**Fig. 2b**).



244
 245 **Fig. 2.** Shared and strain-specific orthogroup counts identified from *Plasmodium y. yoelii* YM, *P.*
 246 *y. yoelii* 17X, and *P. y. nigeriensis* N67 parasites using OrthoFinder [32]. **a**, Venn diagram of
 247 shared and strain-specific orthogroups; **b**, \log_{10} -transformed bit-score distributions for N67
 248 proteins that are not assigned to any orthogroup plus those in N67-specific orthogroups (N67-
 249 specific) and proteins assigned to orthogroups having at least one 17X or YM protein (N67-
 250 other). The bit-scores are derived from pairwise BLAST alignments within the Orthofinder
 251 framework, where all queries were N67 protein sequences that were aligned against the 17X and
 252 YM sequences. The red dots indicate the mean values of bit-score distributions, whereas the
 253 vertical lines within the violins indicate the median, upper and lower quartile values.

254
 255 To further characterize N67-specific proteins, we used BLAST to align the 199 N67 proteins
 256 against the 17X proteome and showed that the majority proteins had motifs matching members
 257 of highly diverse gene families. Among the 199 proteins, 91 are hypothetical or uncharacterized
 258 proteins, 64 are PIR/YIR proteins, 22 are Fam-A/B proteins, and five are reticulocyte binding
 259 proteins (**Table S9**). Clustering the proteins based on sequence similarity generated three
 260 dendrograms, one consisting of Fam-A and Fam-B proteins (**Fig. S2a and Table S9**), another
 261 one consisting of 10 YIR proteins (**Fig. S2b**), and a third one of three subclusters of YIR proteins

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



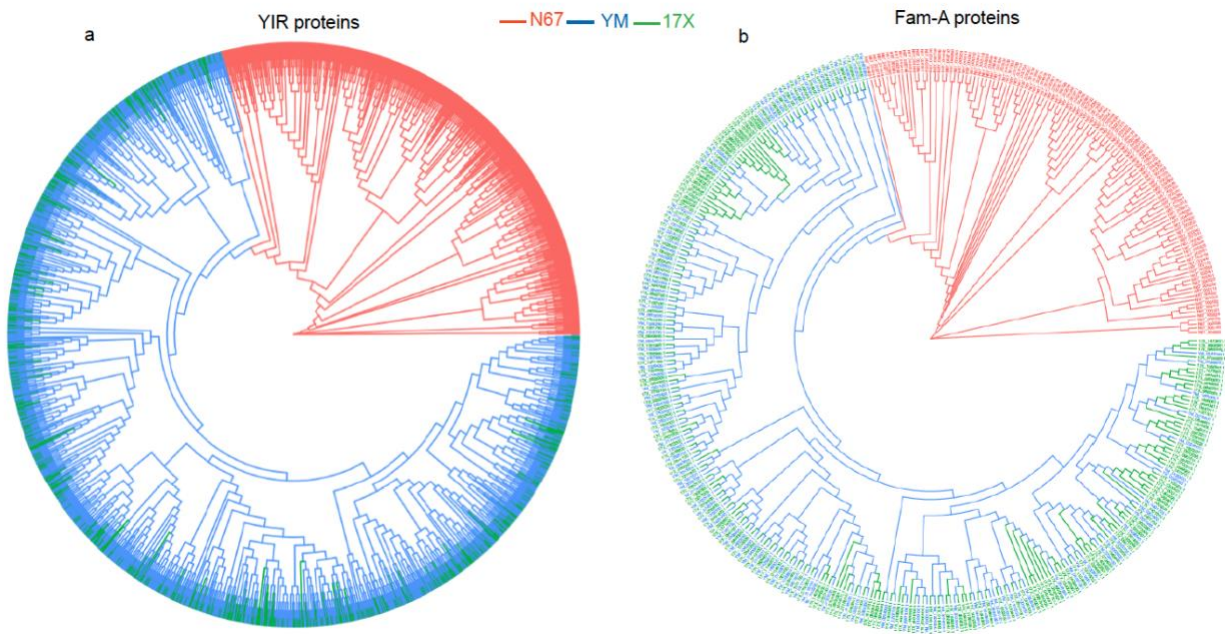
264
 265 **Fig. S2.** Clustering of protein sequences from the *Plasmodium y. nigeriensis* N67-specific
 266 orthogroups and those that are not assigned to any orthogroup. The predicted protein sequences
 267 were aligned using ClustalW algorithm and clustered using procedures described in the Methods
 268 section. **a**, Fam-A/B proteins; **b**, YIR proteins (group 1); **c**, YIR proteins (group 2). Only
 269 bootstrap values higher than 70% are shown.

270

271 **Gene families from three *P. yoelii* parasites**

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

275 genes (21%) for N67, and 1,075 genes (19%) for YM parasite (**Table S10**). The largest gene
276 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
277 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
278 parasite, respectively. As expected, clustering of YIR and Fam-A protein families showed that
279 the proteins from N67 grouped separately from those of 17X and YM (**Fig. 3a** and **Fig. 3b**),
280 consistent with N67 being a subspecies of *P. yoelii*. The true numbers of *yir* and *fam-a/b/c/d*
281 genes for the N67 and YM parasites could be larger because some genes in these gene families
282 are likely not assembled into the genome. Other important gene families include genes encoding
283 early transcribed membrane proteins (ETRAMPs), lysophospholipase, erythrocyte membrane
284 antigens, and reticulocyte binding proteins associated with host-parasite interactions and Cys6
285 (6-Cysteine) proteins. As noted above, there are 83 copies of the gene encoding PcEMA1
286 homologs [29] in the N67 parasite, compared with only one gene in the 17X and YM parasites,
287 respectively (**Table S10**). Clustering of the PcEMA1 proteins from the three parasites showed
288 that N67_000859 and N67_000245 were closely related to the two single copies from 17X
289 (17X_10019001) and YM (YM_100119001) (**Fig. 4**). Similarly, there are nine genes encoding
290 haloacid dehalogenase-like hydrolase in N67, but only 4 genes in both 17X and YM parasites. In
291 contrast, the number of genes encoding reticulocyte binding proteins appears to be reduced in
292 N67 parasites; 12 genes for N67 (including five from N67 specific orthogroups), whereas 17X
293 and YM have 33 and 31, respectively. Although we cannot rule out that the differences in gene
294 number are due to different degrees of genome completeness, the increases in gene number of
295 PcEMA1 homolog and haloacid dehalogenase-like hydrolase in N67 could be associated with its
296 unique growth characteristics and stimulation of strong early type I interferon response [13]. The
297 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

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

331

332 **Discussion**

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

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

352

353 The majority of studies using *P. yoelii* parasites involved the lethal YM (or 17XL) and nonlethal
354 17X (17XNL) strains. However, *P. y. nigeriensis* parasites have also been used as models for
355 studying unique disease phenotypes, transmission in mosquitoes, strain-specific host immune
356 responses and drug resistances [13, 14, 38-41]. The N67 and N67C are two isogenic strains of *P.*
357 *y. nigeriensis* subspecies that cause very unique and interesting disease phenotypes in C57BL/6
358 mice. Whereas mice infected with N67 parasites produce an early peak of IFN-I that has been
359 linked to suppression of parasitemia day 6 pi, mice infected with N67C produce low levels of
360 IFN-I and die approximately day 7 pi due to T cell and IFN- γ mediated inflammation [13, 14]. A
361 mutation (C741Y) in PyEBL partially contributes to the difference in disease phenotypes [15];
362 however, other genes in the genome are likely involved, particularly the variant gene families
363 such as the *yir* and *fam-a/b/c/d* genes. For example, a locus at one end of chromosome 13 was

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

371
372 Approximately 20% of the annotated genes in the *P. yoelii* genomes are members of multi-copy
373 gene families. The number of the *yir* genes in 17X is much larger than that of N67 or YM.
374 Although the observation of more *yir* genes in the 17X genome could be due to more complete
375 and better annotated 17X genome, it is not surprising to find variation in the number of *yir* genes.
376 The *yir* genes belongs to the *Plasmodium* interspersed repeat (*pir*) gene families. The *pir* genes
377 are mostly distributed in the subtelomeric regions of chromosomes with gene copies numbering
378 from a few dozen to hundreds or even over a thousand [42], and up to 40% of the *cir* gene (gene
379 family from *P. chanbaudi*) repertoire are expressed during the intraerythrocytic cycle [43]. Many
380 of the *pir* genes are expressed on the surface of iRBCs and merozoites and play an important role
381 in immune evasion [44, 45]. In addition to the *yir* genes, the variation in number of genes
382 encoding PcEMA1 homologs, haloacid dehalogenase-like hydrolases, reticulocyte binding
383 proteins between N67 and 17X/YM are interesting, particularly the expansion of the gene
384 encoding PcEMA1 homologs. These proteins likely play some important roles in host-parasite
385 interaction and parasite development.

386

387 **Conclusions**

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

400

401 **Methods**

402 **Parasite and infection of mice**

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

408

409 **DNA preparation for PacBio sequencing**

410 Mice were injected *ip* with an inoculum containing 1×10^6 infected red blood cells (iRBCs).
411 Blood samples (200 μ l) with approximately 30-40% parasitemia were collected on day 4 after
412 injection. Infected red blood cells collected in 1 ml 0.15% sodium citrate/PBS buffer were
413 pelleted at 2,000 rpm for 5 min in an Eppendorf centrifuge, re-suspended in 1 ml of PBS, and
414 passed through two consecutive NWF filters (Zhixing Bio, Bengbu, China) to remove the host
415 white blood cells [19]. The flow-through cell suspension was washed in 800 μ l PBS 3X through
416 centrifugation at 3,000 rpm for 3 min. The pellet was dissolved in 400 μ l lysis buffer (100 mM
417 NaCl, 10 mM Tris, 25 mM EDTA, pH 8.0, 0.5% SDS) containing 20 μ l RNase (500 μ g/ml) and
418 20 μ l protease K (10 mg/ml) and incubated at 50°C overnight. DNA was extracted using 400 μ l
419 phenol, chloroform, and isopropanol at a ratio of 25:24:1 and precipitated by adding 2 volume
420 100% ethanol overnight at -20°C. The sample was centrifuged at 13,000 rpm for 15 mins at 4°C,
421 and the DNA pellet was washed in 500 μ l 70% ethanol twice before addition of 20 μ l water. The
422 quality of DNA was estimated on 1% agarose gel showing a typical high molecular weight band.

423

424 **Fragmentation of DNA and PacBio sequencing**

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

431

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

438

439 **Genome sequence assembly**

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

446

447 **RNA-Seq, transcriptome assembly and gene predictions**

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

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

464

465 **Estimates of completeness**

466 We quantified the completeness of the *de novo* assembled N67 transcriptome, genome, and
467 proteome using BUSCO [53] against the single-copy orthologs conserved among *Apicomplexa*
468 (446 BUSCOs) and *Plasmodium* (3642 BUSCOs) from the OrthoDB v10.1 database [54].

469

470 **Identification of orthologs**

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

474

475 **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 JAEVLW0000000000.

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

499 PcEMA1, *P. chabaudi* erythrocyte membrane protein 1
500 MMEJ, microhomology-mediated end joining
501 BUSCO, benchmarking universal single-copy orthologs
502 ETRAMPs, early transcribed membrane proteins
503 VEP, variant effect predictor
504 ISGs, interferon-stimulated genes
505 *pir*, interspersed repeat
506 *yir*, *yoelii* interspersed repeats
507 iRBCs, red blood cells
508 QC, Quality control
509 HMMs, augustus hidden markov models
510 LCB, co-linear block

511

512 **References**

- 513 1. WHO: **World malaria report 2019**. [https://www.who.int/malaria/publications/world-](https://www.who.int/malaria/publications/world-malaria-report-2019/en/)
514 [malaria-report-2019/en/](https://www.who.int/malaria/publications/world-malaria-report-2019/en/) 2019.
- 515 2. Craig AG, Grau GE, Janse C, Kazura JW, Milner D, Barnwell JW, Turner G, Langhorne
516 J, participants of the Hinxton Retreat meeting on Animal Models for Research on Severe
517 M: **The role of animal models for research on severe malaria**. *PLoS Pathog* 2012,
518 **8(2):e1002401**.
- 519 3. Langhorne J, Buffet P, Galinski M, Good M, Harty J, Leroy D, Mota MM, Pasini E,
520 Renia L, Riley E *et al*: **The relevance of non-human primate and rodent malaria**
521 **models for humans**. *Malar J* 2011, **10:23**.
- 522 4. Li J, Zhang Y, Sullivan M, Hong L, Huang L, Lu F, McCutchan TF, Su XZ: **Typing**
523 ***Plasmodium yoelii* microsatellites using a simple and affordable fluorescent labeling**
524 **method**. *Mol Biochem Parasitol* 2007, **155(2):94-102**.

- 525 5. Li J, Pattaradilokrat S, Zhu F, Jiang H, Liu S, Hong L, Fu Y, Koo L, Xu W, Pan W *et al*:
526 **Linkage maps from multiple genetic crosses and loci linked to growth-related**
527 **virulent phenotype in *Plasmodium yoelii***. *Proc Natl Acad Sci U S A* 2011,
528 **108(31):E374-382**.
- 529 6. Nair SC, Pattaradilokrat S, Zilversmit MM, Dommer J, Nagarajan V, Stephens MT, Xiao
530 W, Tan JC, Su XZ: **Genome-wide polymorphisms and development of a microarray**
531 **platform to detect genetic variations in *Plasmodium yoelii***. *Mol Biochem Parasitol*
532 2014, **194(1-2):9-15**.
- 533 7. Pattaradilokrat S, Cheesman SJ, Carter R: **Congenicity and genetic polymorphism in**
534 **cloned lines derived from a single isolate of a rodent malaria parasite**. *Mol Biochem*
535 *Parasitol* 2008, **157(2):244-247**.
- 536 8. Pattaradilokrat S, Li J, Wu J, Qi Y, Eastman RT, Zilversmit M, Nair SC, Huaman MC,
537 Quinones M, Jiang H *et al*: **Plasmodium genetic loci linked to host cytokine and**
538 **chemokine responses**. *Genes Immun* 2014, **15(3):145-152**.
- 539 9. Martin-Jaular L, Ferrer M, Calvo M, Rosanas-Urgell A, Kalko S, Graewe S, Soria G,
540 Cortadellas N, Ordi J, Planas A *et al*: **Strain-specific spleen remodelling in**
541 ***Plasmodium yoelii* infections in Balb/c mice facilitates adherence and spleen**
542 **macrophage-clearance escape**. *Cell Microbiol* 2011, **13(1):109-122**.
- 543 10. Bakir HY, Sayed FG, Rahman SA, Hamza AI, Mahmoud AE, Galal LA, Attia RA:
544 **Comparative study between non lethal and lethal strains of *Plasmodium yoelii* with**
545 **reference to its immunological aspect**. *J Egypt Soc Parasitol* 2009, **39(2):585-593**.
- 546 11. Wykes MN, Liu XQ, Beattie L, Stanistic DI, Stacey KJ, Smyth MJ, Thomas R, Good MF:
547 ***Plasmodium* strain determines dendritic cell function essential for survival from**
548 **malaria**. *PLoS Pathog* 2007, **3(7):e96**.
- 549 12. Otsuki H, Kaneko O, Thongkukiatkul A, Tachibana M, Iriko H, Takeo S, Tsuboi T, Torii
550 M: **Single amino acid substitution in *Plasmodium yoelii* erythrocyte ligand**
551 **determines its localization and controls parasite virulence**. *Proc Natl Acad Sci U S A*
552 2009, **106(17):7167-7172**.
- 553 13. Wu J, Tian L, Yu X, Pattaradilokrat S, Li J, Wang M, Yu W, Qi Y, Zeituni AE, Nair SC
554 *et al*: **Strain-specific innate immune signaling pathways determine malaria**

- 555 **parasitemia dynamics and host mortality.** *Proc Natl Acad Sci U S A* 2014,
556 **111(4):E511-520.**
- 557 14. Lacerda-Queiroz N, Riteau N, Eastman RT, Bock KW, Orandle MS, Moore IN, Sher A,
558 Long CA, Jankovic D, Su XZ: **Mechanism of splenic cell death and host mortality in a**
559 ***Plasmodium yoelii* malaria model.** *Sci Rep* 2017, **7(1):10438.**
- 560 15. Peng YC, Qi Y, Zhang C, Yao X, Wu J, Pattaradilokrat S, Xia L, Tumas KC, He X,
561 Ishizaki T *et al*: ***Plasmodium yoelii* Erythrocyte-Binding-like Protein Modulates Host**
562 **Cell Membrane Structure, Immunity, and Disease Severity.** *mBio* 2020, **11(1).**
- 563 16. Carlton JM, Angiuoli SV, Suh BB, Kooij TW, Perteu M, Silva JC, Ermolaeva MD, Allen
564 JE, Selengut JD, Koo HL *et al*: **Genome sequence and comparative analysis of the**
565 **model rodent malaria parasite *Plasmodium yoelii yoelii*.** *Nature* 2002, **419(6906):512-**
566 **519.**
- 567 17. Vaughan A, Chiu SY, Ramasamy G, Li L, Gardner MJ, Tarun AS, Kappe SH, Peng X:
568 **Assessment and improvement of the *Plasmodium yoelii yoelii* genome annotation**
569 **through comparative analysis.** *Bioinformatics* 2008, **24(13):i383-389.**
- 570 18. Otto TD, Bohme U, Jackson AP, Hunt M, Franke-Fayard B, Hoeijmakers WA, Religa
571 AA, Robertson L, Sanders M, Ogun SA *et al*: **A comprehensive evaluation of rodent**
572 **malaria parasite genomes and gene expression.** *BMC Biol* 2014, **12:86.**
- 573 19. Li J, Cai B, Qi Y, Zhao W, Liu J, Xu R, Pang Q, Tao Z, Hong L, Liu S *et al*: **UTR**
574 **introns, antisense RNA and differentially spliced transcripts between *Plasmodium***
575 ***yoelii* subspecies.** *Malar J* 2016, **15:30.**
- 576 20. Kooij TW, Carlton JM, Bidwell SL, Hall N, Ramesar J, Janse CJ, Waters AP: **A**
577 ***Plasmodium* whole-genome synteny map: indels and synteny breakpoints as foci for**
578 **species-specific genes.** *PLoS Pathog* 2005, **1(4):e44.**
- 579 21. Nakano K, Shiroma A, Shimoji M, Tamotsu H, Ashimine N, Ohki S, Shinzato M,
580 Minami M, Nakanishi T, Teruya K *et al*: **Advantages of genome sequencing by long-**
581 **read sequencer using SMRT technology in medical area.** *Hum Cell* 2017, **30(3):149-**
582 **161.**
- 583 22. Aurrecoechea C, Brestelli J, Brunk BP, Dommer J, Fischer S, Gajria B, Gao X, Gingle A,
584 Grant G, Harb OS *et al*: **PlasmoDB: a functional genomic database for malaria**
585 **parasites.** *Nucleic Acids Res* 2009, **37(Database issue):D539-543.**

- 586 23. Li H: **Minimap2: pairwise alignment for nucleotide sequences.** *Bioinformatics* 2018,
587 **34(18):3094-3100.**
- 588 24. Darling AE, Mau B, Perna NT: **progressiveMauve: multiple genome alignment with**
589 **gene gain, loss and rearrangement.** *PLoS One* 2010, **5(6):e11147.**
- 590 25. Wingett SW, Andrews S: **FastQ Screen: A tool for multi-genome mapping and**
591 **quality control.** *F1000Res* 2018, **7:1338.**
- 592 26. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan
593 L, Raychowdhury R, Zeng Q *et al*: **Full-length transcriptome assembly from RNA-**
594 **Seq data without a reference genome.** *Nat Biotechnol* 2011, **29(7):644-652.**
- 595 27. Langdon WB: **Performance of genetic programming optimised Bowtie2 on genome**
596 **comparison and analytic testing (GCAT) benchmarks.** *BioData Min* 2015, **8(1):1.**
- 597 28. Campbell MS, Holt C, Moore B, Yandell M: **Genome Annotation and Curation Using**
598 **MAKER and MAKER-P.** *Curr Protoc Bioinformatics* 2014, **48:4 11 11-39.**
- 599 29. Deleersnijder W, Prasomsitti P, Tungpradubkul S, Hendrix D, Hamers-Casterman C,
600 Hamers R: **Structure of a Plasmodium chabaudi acidic phosphoprotein that is**
601 **associated with the host erythrocyte membrane.** *Mol Biochem Parasitol* 1992,
602 **56(1):59-68.**
- 603 30. Xu R, Liu Y, Fan R, Liang R, Yue L, Liu S, Su XZ, Li J: **Generation and functional**
604 **characterisation of Plasmodium yoelii csp deletion mutants using a microhomology-**
605 **based CRISPR/Cas9 method.** *Int J Parasitol* 2019, **49(9):705-714.**
- 606 31. Hakimi MA, Bougdour A: **Toxoplasma's ways of manipulating the host**
607 **transcriptome via secreted effectors.** *Curr Opin Microbiol* 2015, **26:24-31.**
- 608 32. Emms DM, Kelly S: **OrthoFinder: solving fundamental biases in whole genome**
609 **comparisons dramatically improves orthogroup inference accuracy.** *Genome Biol*
610 2015, **16:157.**
- 611 33. Buchfink B, Xie C, Huson DH: **Fast and sensitive protein alignment using**
612 **DIAMOND.** *Nat Methods* 2015, **12(1):59-60.**
- 613 34. Emms DM, Kelly S: **OrthoFinder: phylogenetic orthology inference for comparative**
614 **genomics.** *Genome Biol* 2019, **20(1):238.**

- 615 35. Frech C, Chen N: **Variant surface antigens of malaria parasites: functional and**
616 **evolutionary insights from comparative gene family classification and analysis.** *BMC*
617 *Genomics* 2013, **14**:427.
- 618 36. Nair SC, Xu R, Pattaradilokrat S, Wu J, Qi Y, Zilversmit M, Ganesan S, Nagarajan V,
619 Eastman RT, Orandle MS *et al*: **A *Plasmodium yoelii* HECT-like E3 ubiquitin ligase**
620 **regulates parasite growth and virulence.** *Nat Commun* 2017, **8**(1):223.
- 621 37. Wu J, Cai B, Sun W, Huang R, Liu X, Lin M, Pattaradilokrat S, Martin S, Qi Y, Nair SC
622 *et al*: **Genome-wide Analysis of Host-*Plasmodium yoelii* Interactions Reveals**
623 **Regulators of the Type I Interferon Response.** *Cell Rep* 2015, **12**(4):661-672.
- 624 38. Dutta GP, Bajpai R, Vishwakarma RA: **Antimalarial efficacy of arteether against**
625 **multiple drug resistant strain of *Plasmodium yoelii nigeriensis*.** *Pharmacol Res* 1989,
626 **21**(4):415-419.
- 627 39. Orfano AS, Duarte AP, Molina-Cruz A, Pimenta PF, Barillas-Mury C: ***Plasmodium***
628 ***yoelii nigeriensis* (N67) Is a Robust Animal Model to Study Malaria Transmission by**
629 **South American Anopheline Mosquitoes.** *PLoS One* 2016, **11**(12):e0167178.
- 630 40. Beaute-Lafitte A, Altemayer-Caillard V, Chabaud AG, Landau I: ***Plasmodium yoelii***
631 ***nigeriensis*: biological mechanisms of resistance to chloroquine.** *Parasite* 1994,
632 **1**(3):227-233.
- 633 41. Graves PM, Curtis CF: **Susceptibility of *Anopheles gambiae* to *Plasmodium yoelii***
634 ***nigeriensis* and *Plasmodium falciparum*.** *Ann Trop Med Parasitol* 1982, **76**(6):633-639.
- 635 42. Cunningham D, Lawton J, Jarra W, Preiser P, Langhorne J: **The *pir* multigene family of**
636 ***Plasmodium*: antigenic variation and beyond.** *Mol Biochem Parasitol* 2010,
637 **170**(2):65-73.
- 638 43. Lawton J, Brugat T, Yan YX, Reid AJ, Bohme U, Otto TD, Pain A, Jackson A, Berriman
639 M, Cunningham D *et al*: **Characterization and gene expression analysis of the *cir***
640 **multi-gene family of *Plasmodium chabaudi chabaudi* (AS).** *BMC Genomics* 2012,
641 **13**:125.
- 642 44. Yam XY, Brugat T, Siau A, Lawton J, Wong DS, Farah A, Twang JS, Gao X, Langhorne
643 J, Preiser PR: **Characterization of the *Plasmodium* Interspersed Repeats (PIR)**
644 **proteins of *Plasmodium chabaudi* indicates functional diversity.** *Sci Rep* 2016,
645 **6**:23449.

- 646 45. Yam XY, Preiser PR: **Host immune evasion strategies of malaria blood stage**
647 **parasite**. *Mol Biosyst* 2017, **13**(12):2498-2508.
- 648 46. Bolger AM, Lohse M, Usadel B: **Trimmomatic: a flexible trimmer for Illumina**
649 **sequence data**. *Bioinformatics* 2014, **30**(15):2114-2120.
- 650 47. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M,
651 Gingeras TR: **STAR: ultrafast universal RNA-seq aligner**. *Bioinformatics* 2013,
652 **29**(1):15-21.
- 653 48. Ahdesmaki MJ, Gray SR, Johnson JH, Lai Z: **Disambiguate: An open-source**
654 **application for disambiguating two species in next generation sequencing data from**
655 **grafted samples**. *F1000Res* 2016, **5**:2741.
- 656 49. Korf I: **Gene finding in novel genomes**. *BMC Bioinformatics* 2004, **5**:59.
- 657 50. Stanke M, Steinkamp R, Waack S, Morgenstern B: **AUGUSTUS: a web server for gene**
658 **finding in eukaryotes**. *Nucleic Acids Res* 2004, **32**(Web Server issue):W309-312.
- 659 51. Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, Chang HY,
660 Dosztanyi Z, El-Gebali S, Fraser M *et al*: **InterPro in 2017-beyond protein family and**
661 **domain annotations**. *Nucleic Acids Res* 2017, **45**(D1):D190-D199.
- 662 52. Eilbeck K, Moore B, Holt C, Yandell M: **Quantitative measures for the management**
663 **and comparison of annotated genomes**. *BMC Bioinformatics* 2009, **10**:67.
- 664 53. Waterhouse RM, Seppey M, Simao FA, Manni M, Ioannidis P, Klioutchnikov G,
665 Kriventseva EV, Zdobnov EM: **BUSCO Applications from Quality Assessments to**
666 **Gene Prediction and Phylogenomics**. *Mol Biol Evol* 2018, **35**(3):543-548.
- 667 54. Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simao FA, Zdobnov
668 EM: **OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial**
669 **and viral genomes for evolutionary and functional annotations of orthologs**. *Nucleic*
670 *Acids Res* 2019, **47**(D1):D807-D811.
- 671 55. Kelly S, Maini PK: **DendroBLAST: approximate phylogenetic trees in the absence of**
672 **multiple sequence alignments**. *PLoS One* 2013, **8**(3):e58537.
- 673 56. Kumar S, Stecher G, Li M, Knyaz C, Tamura K: **MEGA X: Molecular Evolutionary**
674 **Genetics Analysis across Computing Platforms**. *Mol Biol Evol* 2018, **35**(6):1547-1549.
- 675 57. Jones DT, Taylor WR, Thornton JM: **The rapid generation of mutation data matrices**
676 **from protein sequences**. *Comput Appl Biosci* 1992, **8**(3):275-282.

677 **Acknowledgements**

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

681

682 **Funding**

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

685

686 **Author information**

687 Affiliations

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

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

692

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

695 Cihan Oguz, Sue Huse & Justin Lack

696

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

699 Lu Xia

700

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

702 Jack Chen

703

704 **Contributions**

705 CZ, LX, Y-CP, JW, MS, parasite infection of mice, DNA preparation, data analysis; JC, PacBio

706 sequencing, CO, SH, JL, data analysis and writing; CAL and X-zS supervision and writing; Xz-S

707 project conception.

708

709 **Corresponding author**

710 Correspondence to Xinzhuan Su.

711

712 **Ethics declarations**

713 **Ethics approval and consent to participate**

714 *Plasmodium yoelii nigeriensis* N67 is a malaria parasite that infects red blood cells of rodents,

715 including mice. This study involves infection of C57BL/6 mice with the parasite to obtain

716 parasite genomic DNA for sequencing. The experiments were performed in accordance with the

717 protocol approved (approval #LMVR11E) by the Institutional Animal Care and Use Committee

718 (IACUC) at the National Institute of Allergy and Infectious Diseases (NIAID) following the

719 guidelines of the Public Health Service Policy on Humane Care and Use of Laboratory Animals

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

722

723 Consent for publication

724 Not applicable.

725

726 Conflict of interests

727 The authors declare that they have no competing interests.

728

729 **Supplementary information**

730 **Additional file 1. Fig. S1-2.**

731 **Additional file 2. Table S1.**

732 **Additional file 3. Table S2.**

733 **Additional file 4. Table S3.**

734 **Additional file 5. Table S4.**

735 **Additional file 6. Table S5.**

736 **Additional file 7. Table S6.**

737 **Additional file 8. Table S7.**

738 **Additional file 9. Table S8.**

739 **Additional file 10. Table S9.**

740 **Additional file 11. Table S10.**

741 **Additional file 12. Table S11.**

742