A genomic platform for surveillance and antigen discovery in *Plasmodium spp.* using long-read amplicon sequencing

David Fernando Plaza[1,2,ŧ], Julia Zerebinski[1,2], Ioanna Broumou[1,2], Maximilian Julius Lautenbach[1,2], Billy Ngasala[3], Christopher Sundling[1,2], Anna Färnert[1,2]

[1] Division of Infectious Diseases, Department of Medicine Solna and Center for Molecular Medicine, Karolinska Institutet, Stockholm, Sweden

[2] Department of Infectious Diseases, Karolinska University Hospital, Stockholm, Sweden

[3] Muhimbili University of Health and Allied Sciences, Dar es Salaam, Tanzania

[ŧ] e-mail: david.plaza@ki.se

Abstract

Many vaccine candidate proteins in the malaria parasite *Plasmodium falciparum* are under strong selective pressure to diversify in terms of antigenicity. We present a sequencing and data analysis platform for the genomic and epidemiological surveillance of the indel-rich antigen MSP2 from *P. falciparum* using long-read circular consensus sequencing (CCS) in multiclonal malaria isolates. Our platform uses 40 PCR primers to asymmetrically barcode and identify multiclonal infections in pools of up to 384 samples. We validated the method using 235 mock infections combining 10 synthetic variants at different concentrations and infection complexities, as well as 95 isolates from *P. falciparum*-infected residents of Nyamisati, Tanzania. We also constructed a fully automated analysis pipeline that streamlines the processing and interpretation of epidemiological and antigenic diversity data for *msp2* from demultiplexed FASTQ files. Finally, this platform can be easily adapted to other polymorphic antigens of interest in *Plasmodium* or any other human pathogen.

Keywords: Genomic surveillance, long-read sequencing, antigen discovery, malaria epidemiology, merozoite surface protein 2

49 **Main**

50 There were 241 million cases of malaria and 627 thousand deaths from the disease reported in 2020[1]. An
51 effective vaccine against malaria is a cost-effective and urgently needed strategy to reduce the disease
52 burden and potentially eradicate malaria in the future[2]. RTS,S/AS01 (Mosquirix™) is the first malaria
53 vaccine to be recently endorsed by the World Health Organization for use in children[3]. Nevertheless, the
54 protection conferred by RTS,S/AS01 is strain-specific[4], therefore limiting the effectiveness of this vaccine
55 in regions where *P. falciparum* circulates with a high genetic and antigenic diversity[5]. R21/Matrix-M,
56 another vaccine candidate based on the circumsporozoite protein of the pre-erythrocytic stage, has more
57 recently shown promising higher-level efficacy, though still partial and the longevity of protection is
58 uncertain[6]. Polymorphic antigens expressed in the blood stage of the parasite, such as the merozoite
59 surface protein 1 (MSP1), MSP2 MSP3, apical membrane antigen 1 (AMA1) and erythrocyte binding
60 antigen 175 (EBA175), amongst others[7], have been studied in clinical trials for their potential as malaria
61 vaccines, and several of them are widely used as molecular markers in the characterization of parasite
62 populations and the definition of treatment failures in drug trials[8]. New approaches to survey the pool of
63 genetic diversity in these and other antigens will accelerate the development of strain-transcending
64 vaccines against *P. falciparum* malaria.

65

66 MSP2 is a 19 to 34 kDa protein (depending on the allele)[9] that is highly expressed in the blood stage of *P.*
67 *falciparum*[10]. The protein has two highly conserved domains in the N and the C termini comprising nearly
68 half of the sequence and flanking a highly polymorphic central region. The locus coding for this central
69 region is a hotspot for indels of up to 300bp[11], making its sequencing and assembly, especially for
70 multiclonal infections, a challenging task using traditional short-read sequencing[12]. Insertions, deletions
71 or tandem repeats with lengths larger than the read lengths in short-read sequencing libraries represent
72 another significant mapping and assembly challenge.

73

74 Here, we present a multiplex long-read amplicon sequencing-based platform for the accurate genotyping
75 of the malaria vaccine candidate MSP2. The use of this circular consensus sequencing (CCS) technology to
76 obtain single-base resolution data on polymorphic *P. falciparum* antigens will facilitate the
77 implementation of a comprehensive genomic surveillance strategy for malaria and will allow the in-depth
78 structural and antigenic characterization of MSP2. In addition, a complete evaluation of the genetic
79 diversity of candidate antigens associated to naturally acquired immunity against malaria is the first step
80 to further develop these parasite proteins as vaccine candidates capable of eliciting strain-transcending
81 immunity. Finally, our genomic surveillance platform can easily be adapted to facilitate the genotyping of
82 any structurally variable antigen in other human pathogens.

83

84 **Results**

85 ***CCS on pools of msp2 amplicons provides insights into complex malaria infections with high sequence***
86 ***accuracy***

87 Combining nested PCR, asymmetric barcoding of amplicons and CCS, we developed a platform for the
88 genotyping by sequencing of *msp2*, a locus rich in structural variants, in up to 384 pooled controls and
89 malaria clinical isolates (Fig. 1). With an average read length of 28300 bp and a mean insert size of 862 bp,
90 every *msp2* amplicon was sequenced 32 times on average, generating circular consensus reads where
91 most sequencing errors (previously estimated to be one error every 13048 bp in CCS reads[13]) are properly
92 amended. We observed no correlation between template concentration in a series of mock infection
93 controls including single *msp2* variants and the number of reads resulting from those samples
94 (Supplementary Fig. 1). Similarly, there was no correlation between sample parasite density (parasites/µl)
95 and the number of CCS reads for microscopy-positive samples in the Tanzania cohort (Supplementary Fig.
96 2).

97

98 ***Variant calling specificity and sensitivity can be estimated from synthetic mock infection controls***

99 Size and sequence variant calling for clinical isolates was based on the accuracy of variant detection in
100 synthetic mock infection controls. These controls corresponded to 235 single or mixed serial dilutions
101 including various molar ratios of synthetic reference sequences for *msp2* HB3 (LR131339 REGION:
102 250760..251530), CD01 (PfCD01_020011700), Dd2 (PfDd2_020009600), SN01 (PfSN01_020009800), KE01
103 (PfKE01_020009300), SD01 (PfSD01_020012300), GN01 (PfGN01_020012100), 7G8 (Pf7G8_020011500),
104 GB4 (PfGB4_020009800) and 3D7 (PF3D7_0206800), that were barcoded separately and pooled together
105 before adaptor ligation and sequencing (Supplementary Table 1). Serial dilution of single-variant mock
106 infections showed that the correct size (Fig. 2a) and sequence (Fig. 2b) variant for reference *msp2* alleles
107 can be detected with read coverages higher than 50% at concentrations of 10 copies/µl for all the variants
108 and down to 1 copy/µl for 6 out of 10 (SD01, GN01, 7G8, GB4, SN01 and KE01). We observed that all
109 samples with barcode bc1034, including the 100 copies/µl dilution for variants SD01, GN01, 7G8, GB4,
110 CD01, Dd2, SN01 and KE01, did not yield reads above background levels (Supplementary Fig. 3) and were
111 therefore not included in the sensitivity analysis.

112

113 Quantification of left-over noise after the noise-reduction module (see below) revealed that the mean
114 read coverage for the most common size variant miscall in 235 mock infections was 3.5% of reads (SD 5.4,
115 95% CI 2.8-4.2%) (Fig. 2c). A downstream analysis of sequence variants for the same pool of mock
116 infections showed that the mean read coverage for the most common sequence variant miscall was 4.5%
117 of the reads (SD 5.9, 95% CI 3.8-5.3%) (Fig. 2d). The mean read coverage for size variant miscalls at
118 different molar ratios and dilutions in the synthetic mock infections (Fig. 2c) allowed us to setup a variant
119 calling threshold of 14% (the mean read coverage of the most prevalent variant miscall in all the mock
120 infections plus 2 standard deviations (SD)). Similarly, a coverage threshold was defined in the calling of

121 nucleotide sequence variants (Fig. 2d) equivalent to 16% or the mean coverage of the most common
122 sequence miscall in the synthetic controls plus 2 SD. For the same 235 mock infections, a total of 1140
123 correct size variant calls were expected; nevertheless, 926 (81.2%) of those were detected in the dataset
124 with a total of 360 above the noise threshold (31.2% of the expected, false negative rate (FNR): 0.688).
125 Furthermore, 7407 size variant miscalls were detected, from which only 9 had read coverages above or
126 equal to the established 14% threshold (False positive rate (FPR): 0.001). As for sequence variants, 1312
127 correct calls were expected in the 235 mock infections. From these, 467 (35.6%) were detected with 342
128 above the noise threshold (26.1% of the expected, FNR: 0.739). In addition, 8082 sequence variant miscalls
129 were detected and only 8 of those showed read coverages above the corresponding variant calling cutoff
130 of 16% (FPR: 0.001). Finally, size and sequence variant calling thresholds at FPRs of 0.01 and 0.05 were
131 calculated (Supplementary Table 2). The selection of FPR significantly affects the number of variants that
132 can be detected in individual isolates, with less stringent FPRs (lower variant calling thresholds) resulting
133 in a closer match between the number of variants (clones in a clinical isolate, else referred as multiplicity
134 of infection (MOI)) expected and observed, however compromising the sequence accuracy of the variants
135 called (Supplementary Fig. 4).
136
137 ***Size variant genotyping for clinical isolates can be simulated by in silico PCR on CCS reads***
138 Even though sequencing with single nucleotide resolution is the optimal tool to assess isolate complexity
139 and relatedness in molecular epidemiological surveys[14, 15], most of the available genetic diversity data for
140 *P. falciparum* relies on the genotyping of size variants in *msp2*, *msp1* and *glurp*[16] due to its low cost and
141 ease of use. *In silico* PCR on CCS reads with the widely used family-specific (IC1 and FC27) primers for
142 *msp2*[17] was applied to call size variants on *P. falciparum* isolates which could then be compared to data
143 previously generated by nested PCR (using the same primers) and analysis of fragment sizes by capillary
144 electrophoresis (CE)[18]. Fully sequenced and annotated genomes for a number *P. falciparum* strains

145 became available (PlasmoDB[19]) in the years following the original *msp2* genotyping publication[17] and we
146 used these sequences to design forward oligos that incorporate new base substitutions (Supplementary
147 Fig. 5 and Supplementary Table 3).
148
149 *In silico* PCR on the library of synthetic mock infections using the original and the newly designed primers
150 followed by size variant calling with the thresholds previously defined showed that the eight size variants
151 expected (three different sequence variants correspond to FC292) can be detected amongst the collection
152 of mock infections. In addition, a clinical isolate corresponding to an FC316 variant and spiked into some
153 of the mock infection samples could be detected in 10 samples at an FPRs of 0.001 (Fig. 3a). Only three
154 miscalls were found in 5 out of the 321 (1.6%) mock infections analyzed using the same FPR. Lastly, *in*
155 *silico* PCR for size variant genotyping in clinical isolates that were sequenced in a separate library, showed
156 that 52 (54.7%) out of 95 isolates were IC1-positive (Fig. 3b) while 59 (62.1%) were FC27-positive (Fig. 3c);
157 with some variants such as FC292 and FC328 being present in 21 (35%) and 29 (49%) of the FC27-positive
158 isolates sequenced, respectively. Sixteen samples (16.8%) were found to be double positive for IC1 and
159 FC27.
160
161 Even though amplification conditions vary widely in the PCR programs for size variant genotyping by CE
162 and the barcoding of *msp2*-derived amplicons, we decided to compare the specific size variants being
163 called by CE and CCS on 24 *msp2*-positive samples from Tanzania that were randomly selected. A 7bp plus
164 or minus window in the CE calls was used to account for the size variant binning required when using this
165 method[18]. We found that for 15 out of 24 *msp2*-positive (62.5%, Supplementary Table 4) samples there
166 was at least one match in the size variants called by the two methods with high relative fluorescence units
167 (RFU>300) and read coverage values (≥14% of reads in the sample).
168

169 ***Isolate-derived msp2 sequences allow the analysis of antigen structural conservation and the***
170 ***construction of isolate phylogenies***

171 Finally, isolate phylogenies, analysis of B-cell epitope conservation and prediction of isolate-specific T cell
172 epitopes were computed for the same 95 *P. falciparum* isolates. One of the strengths of having sequences
173 for *msp2* compared to the size variant data provided by the CE genotyping method is the possibility of
174 running conservation analysis of protein sequences that can, on the one hand, be used in the development
175 of MSP2 as a vaccine candidate, and on the other hand, allow the description of phylogenetic relatedness
176 between isolates. Conservation analysis of unique MSP2 variants of *P. falciparum* in the IC1 (n = 37) and
177 FC27 (n = 34) families for isolates from Tanzania shows that there are well defined regions which are highly
178 conserved in the two families (Fig. 4a) and that could be targeted by vaccine-elicited antibodies to confer
179 strain-transcending protection. We used the consensus sequences derived from the multiple alignment
180 Fourier transform (MAFFT) of the MSP2 variants present in the isolates to map linear epitopes bound by
181 well-characterized anti-MSP2 monoclonal antibodies (mAbs) raised against the 3D7 and FC27 variants of
182 the antigen[20]. We observed that epitopes present in conserved regions of these two specific variants (6D8,
183 1F7/6C9/9H4 and 4D11/9G8) can also be present in the variable regions of some of the clinical isolates
184 sequenced (Fig. 4a). Furthermore, sequences of a polymorphic locus like *msp2* are instrumental in the
185 construction of phylogenies describing the degree of relatedness between different *P. falciparum* isolates.
186 We show that *msp2* sequences from clinical isolates can be used to construct phylogenies for the two
187 families of the gene (Fig. 4b and Fig. 4c).

188

189 ***Non-redundant T cell epitopes with wide human population coverage were detected in the MSP2***
190 ***variants from a set of clinical isolates***

191 To explore how sequence diversity in MSP2 could be affecting the immunogenicity of different antigen
192 variants, we computed binding predictions for all the 15-mer peptides derived from the same Tanzania

193 isolates against the widely representative HLA class II alleles DRB1*03:01, DRB1*07:01, DRB1*15:01,
194 DRB3*01:01, DRB3*02:02, DRB4*01:01 and DRB5*01:01 using the MHC-II Binding prediction tool from
195 IEDB-AR[21]. 126444 T cell epitopes predicted to bind to these HLA class II alleles were found in unique
196 variants of MSP2 sequenced for Tanzania isolates (Supplementary Table 5). From these, 2217 peptides
197 (Supplementary Table 6) were found to be non-redundant while 29 epitopes out of these 2217 were
198 predicted as strong HLA class II binders based on an adjusted rank equal or lower than 1% (99th binding
199 percentile) (Supplementary Table 7).
200
201 ***An iterative analysis pipeline streamlines the processing and interpretation of epidemiological and***
202 ***antigenic diversity data from demultiplexed FASTQ files***
203 In order to streamline the analysis of *msp2* genotyping data, a fully-automated pipeline was assembled in
204 the bioinformatics analysis interface Galaxy, a language-agnostic platform that integrates tools written in
205 R, Python and other programming languages (https://github.com/dfplazag/CCS-Pfal-msp2.git). We
206 observed that the size variant miscalls with the highest read counts corresponded to sequences up to 3
207 bases longer or shorter than the expected correct calls (Supplementary Fig. 6a). In order to reduce the
208 read coverage for these miscalls, an iterative noise reduction strategy was applied where sequences called
209 on a first iteration of the pipeline were used to remove from the original dataset highly similar sequences
210 (up to 3 mismatches based on the 4-mismatch difference for the *msp2*s in the strains CD01 and HB3,
211 Supplementary Fig. 6b) which were also up to 3 nucleotides longer or shorter than the sequence called.
212 This approach resulted in a read count reduction of more than 50% for 13 out of the 22 most prevalent
213 size variant miscalls in the dataset (Supplementary Fig. 6c) and a small increase in the number of correct
214 size variant calls from 348 without the noise reduction module to 360 in the set of 235 mock infections.
215 The resulting pipeline is composed of two workflows. The first workflow (Supplementary Fig. 7) takes as
216 inputs a Galaxy collection of demultiplexed FASTQ files and a tabular file with the sample name pre-

217 assigned to every FASTQ file; delivering in return, a tabular dataset collapsed to include every read
218 assigned to a sample for all the samples in the library and a table with the number of reads for each
219 sample. Workflow 1 also calculates the baseline read count for a sample to be considered *msp2*-positive
220 based on the mean number of reads miss-assigned to water controls plus 2 SDs. Outputs from workflow
221 1, in addition to a FASTA file with the sequence of the *msp2* ortholog in *P. billcollinsi* and a second FASTA
222 file with all the size variant genotyping oligos for *msp2* designed by us and others are the inputs for a
223 second more comprehensive analysis workflow composed of 175 processing modules (Fig. 5a). Workflow
224 2 (Supplementary Fig. 8) runs in two iterations with the noise reduction module previously described
225 filtering out common size variant miscalls between the first and the second iterations (Fig. 5b). While both
226 iterations call size and sequence variants, only iteration two proceeds to construct nucleotide and protein
227 sequence alignments and phylogenies, compute read coverage matrices for size variants in all the samples
228 genotyped, predict T cell epitopes and calculate MOI values for every sample. Finally, workflow two can
229 be easily customized to the study of other polymorphic antigens in *P. falciparum* or other pathogens.
230
231 **Discussion**
232 Here, we introduce a long-read amplicon sequencing platform that we believe can be transformative in
233 unlocking the genetic diversity of antigens rich in structural variants for the malaria parasite *P. falciparum*.
234 Our platform is also ideal to target and assess the diversity in polymorphic antigens of relevance for the
235 development of immunity to other human pathogens for which vaccines are not currently available.
236
237 We show that sensitivity and specificity for the CCS genotyping platform can be accurately estimated using
238 synthetic mock infections that include different molar ratios of up to 10 *msp2* variants of the IC1 and FC27
239 families. Most errors introduced during sequencing are amended when individual sequences resulting
240 from each polymerase pass on the circularized template are aligned while producing consensus reads.

241 Moreover, the size and sequence variants called for the synthetic sequences allowed us to define coverage
242 thresholds that can be adjusted to eliminate PCR or sequencing artifacts from the dataset at different
243 FPRs. For single-variant mock infections, we called the correct size and sequence at concentrations of 10
244 copies/µl in all the 10 variants tested and down to 1 copy/µl in 6 out of the 10 (4 IC1 and 2 FC27). This
245 difference in sensitivity for the detection of variants of the IC1 and FC27 families should be considered
246 when genotyping isolates of low parasitemia.
247
248 We applied Primer-BLAST[22] to predict the annealing coordinates for CE genotyping oligos. These
249 coordinates were then used to calculate the theoretical PCR product lengths that would result from using
250 the oligos on the CCS reads as templates. Despite the different amplification strategies and variant calling
251 approaches used by the two platforms, we showed that 62.5% of the *msp2*-positive isolates genotyped
252 by CCS and CE, have at least one matching size variant between the two methods. The most appealing
253 possibility derived from the sequencing of *msp2* lies in the construction of isolate relatedness models for
254 which CE data is inadequate; nevertheless, CE validation of the CCS-based size variant calling indicates
255 that CCS data could be at least partially compared to previous *msp2* diversity surveys carried out by CE.
256
257 Multiclonal infections represent a significant challenge in the genotyping of malaria isolates by short-read
258 sequencing. While a logistically cumbersome approach consists of the sorting and sequencing of individual
259 parasites present in the isolate[23], long-read sequencing of complete amplicons circumvents the need for
260 sorting and sequencing of individual clones or the use of algorithms that try to differentiate and assemble
261 short reads derived from a mix of variants[24]. The use of CCS in the determination of infection complexity
262 showed that relaxing the stringency in the calling of variants results in a much better match between the
263 expected and the observed number of co-infecting alleles, although compromising the sequence accuracy
264 of the variants called. Variant calling thresholds can therefore be adjusted depending on the intended

265 application, with higher FPRs (e.g. in average, seven clones can be identified for infections with expected
266 MOI: 10 at an FPR: 0.05) used for epidemiological surveillance and lower FPRs (e.g. in average, one clone
267 can be identified for infections with expected MOI: 10 at an FPR: 0.001) applied to antigen discovery and
268 vaccine development.
269
270 Phylogeny reconstruction and the study of complex multiclonal infections are now made possible by the
271 availability of amplicon sequencing data. Phylogenetic analysis of *P. falciparum* clones recovered from
272 clinical isolates will be a pivotal tool in the genomic surveillance of malaria outbreaks for exposed
273 communities. The ease and low cost of sequencing *msp2* amplicons for thousands of isolates will also
274 allow the study of population structure and gene flows for *P. falciparum* at a scale that is beyond the
275 current data generation and analysis capacity of whole-genome sequencing approaches[25]. In addition, the
276 use of highly polymorphic *msp2* sequences as clone-specific "fingerprints" will allow the study of
277 recrudescence and treatment failure with unprecedented resolution joining the battery of markers
278 already in use for sequencing-based follow-up of drug resistance in clinical trials such as *cpp*, *msp7*[16],
279 *ama1-D3*[26], *cpmp*[27] and *csp*[4].
280
281 Sequence conservation analysis of MSP2 showed that well characterized epitopes in the conserved
282 domains of the protein can also be present in the variable domains of some of the isolates sequenced,
283 possibly by sequence duplication. The analysis of antigenic diversity will support the development of MSP2
284 as a vaccine candidate against malaria in two different fronts. On the one hand, by identifying fully
285 conserved regions of MSP2 that can then be targeted by the host immune response therefore conferring
286 strain-transcending protection. On the other hand, by informing the production of a multivalent antigen
287 cocktail including variable domains of dozens or hundreds of clones.
288

289 The high-throughput analysis of T cell epitopes present in MSP2 can be used to study host-pathogen
290 dynamics with exquisite resolution when coupled to the genotyping of HLA class II[28, 29] in the individual
291 harboring specific MSP2 variants. The use of Galaxy for the analysis of human genomic data in studies
292 where host and parasite are sequenced will require the local installation of Galaxy in a firewall-protected
293 server without internet access, allowing the safe handling of sensitive patient genomic data.
294
295 Early efforts at developing MSP2 as a vaccine candidate were built upon observations showing that
296 antibodies against this protein are highly correlated to the development of naturally acquired immunity
297 against clinical malaria[30]. Nonetheless, these vaccine prototypes failed to capture the vast genetic
298 diversity of MSP2 and likely conferred limited protection against heterologous infections[31, 32]. These
299 results suggest that a complete survey on the antigenic repertoire of MSP2 could inform the development
300 of a strain-transcending polyvalent antigen cocktail. mRNA as an antigen delivery system has been proven
301 to be effective in eliciting powerful and protective immune responses against SARS-CoV-2[33]. In the future,
302 antigen delivery in the form of codon-optimized mRNAs or alternative delivery platforms for multivalent
303 MSP2 formulations might be capable of eliciting immunity against dozens or even hundreds of circulating
304 antigen variants of *P. falciparum* in a single vaccine shot imitating, in a way, the protection against clinical
305 malaria conferred by repeated exposure to multiple infections.
306
307 The implementation of Galaxy for cheap, high-performance computing, facilitates the understanding and
308 sharing of non-sensitive isolate data amongst epidemiologists, wet lab scientists and computational
309 biologists alike in both non-endemic regions and regions of high malaria transmission. Analysis pipelines
310 and data histories can be easily shared since all the storage and computation takes place in the cloud.
311 Furthermore, the decentralization of storage and computation guarantees the robustness of the data
312 processing and analysis infrastructure. Last but not least, cloud computing infrastructure provides an

313 unparalleled platform to foster international collaborations where analyses can be ran and monitored in
314 real time by collaborators anywhere in the world[34].
315
316 In conclusion, our long-read platform for amplicon sequencing, provides an instrumental tool for the study
317 of complex infections and can be easily adapted to the genotyping of other *P. falciparum* antigens rich in
318 structural variants, such as MSP1, MSP3, GLURP, CSP and AMA1. Furthermore, the detailed structural
319 information gained from the sequencing of polymorphic vaccine candidates can now inform the design of
320 polyvalent vaccines capable of transcending the limited protection conferred by formulations targeting
321 individual *P. falciparum* antigen variants.
322
323 **References**

324 1. WHO World malaria report 2021. (2021).
325 2. Sauboin, C., Van Bellinghen, L.A., Van De Velde, N. & Van Vlaenderen, I. Economic Impact of
326 Introducing the RTS,S Malaria Vaccine: Cost-Effectiveness and Budget Impact Analysis in 41
327 Countries. *MDM Policy Pract* **4**, 2381468319873324 (2019).
328 3. Sinnis, P. & Fidock, D.A. The RTS,S vaccine-a chance to regain the upper hand against malaria?
329 *Cell* **185**, 750-754 (2022).
330 4. Neafsey, D.E. et al. Genetic Diversity and Protective Efficacy of the RTS,S/AS01 Malaria Vaccine.
331 *N Engl J Med* **373**, 2025-2037 (2015).
332 5. Pringle, J.C. et al. RTS,S/AS01 malaria vaccine mismatch observed among Plasmodium
333 falciparum isolates from southern and central Africa and globally. *Sci Rep* **8**, 6622 (2018).
334 6. Datoo, M.S. et al. Efficacy of a low-dose candidate malaria vaccine, R21 in adjuvant Matrix-M,
335 with seasonal administration to children in Burkina Faso: a randomised controlled trial. *Lancet*
336 **397**, 1809-1818 (2021).

337 7. Schwartz, L., Brown, G.V., Genton, B. & Moorthy, V.S. A review of malaria vaccine clinical
338 projects based on the WHO rainbow table. *Malar J* **11**, 11 (2012).
339 8. WHO Methods for surveillance of antimalarial drug efficacy. (WHO Press, 2009).
340 9. Smythe, J.A. et al. Structural diversity in the Plasmodium falciparum merozoite surface antigen
341 2. *Proc Natl Acad Sci U S A* **88**, 1751-1755 (1991).
342 10. Otto, T.D. et al. New insights into the blood-stage transcriptome of Plasmodium falciparum
343 using RNA-Seq. *Mol Microbiol* **76**, 12-24 (2010).
344 11. Mwingira, F. et al. Plasmodium falciparum msp1, msp2 and glurp allele frequency and diversity
345 in sub-Saharan Africa. *Malar J* **10**, 79 (2011).
346 12. Chaisson, M.J.P. et al. Multi-platform discovery of haplotype-resolved structural variation in
347 human genomes. *Nat Commun* **10**, 1784 (2019).
348 13. Wenger, A.M. et al. Accurate circular consensus long-read sequencing improves variant
349 detection and assembly of a human genome. *Nat Biotechnol* **37**, 1155-1162 (2019).
350 14. Black, A., MacCannell, D.R., Sibley, T.R. & Bedford, T. Ten recommendations for supporting open
351 pathogen genomic analysis in public health. *Nat Med* **26**, 832-841 (2020).
352 15. Neafsey, D.E., Taylor, A.R. & MacInnis, B.L. Advances and opportunities in malaria population
353 genomics. *Nat Rev Genet* (2021).
354 16. Gruenberg, M., Lerch, A., Beck, H.P. & Felger, I. Amplicon deep sequencing improves
355 Plasmodium falciparum genotyping in clinical trials of antimalarial drugs. *Sci Rep* **9**, 17790
356 (2019).
357 17. Snounou, G. et al. Biased distribution of msp1 and msp2 allelic variants in Plasmodium
358 falciparum populations in Thailand. *Trans R Soc Trop Med Hyg* **93**, 369-374 (1999).

359 18. Liljander, A. et al. Optimization and validation of multi-coloured capillary electrophoresis for
360 genotyping of Plasmodium falciparum merozoite surface proteins (msp1 and 2). *Malar J* **8**, 78
361 (2009).
362 19. Aurrecoechea, C. et al. PlasmoDB: a functional genomic database for malaria parasites. *Nucleic*
363 *Acids Res* **37**, D539-543 (2009).
364 20. Adda, C.G. et al. Antigenic characterization of an intrinsically unstructured protein, Plasmodium
365 falciparum merozoite surface protein 2. *Infect Immun* **80**, 4177-4185 (2012).
366 21. Dhanda, S.K. et al. IEDB-AR: immune epitope database-analysis resource in 2019. *Nucleic Acids*
367 *Res* **47**, W502-W506 (2019).
368 22. Ye, J. et al. Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction.
369 *BMC Bioinformatics* **13**, 134 (2012).
370 23. Nair, S. et al. Single-cell genomics for dissection of complex malaria infections. *Genome Res* **24**,
371 1028-1038 (2014).
372 24. Zhong, D., Koepfli, C., Cui, L. & Yan, G. Molecular approaches to determine the multiplicity of
373 Plasmodium infections. *Malar J* **17**, 172 (2018).
374 25. Amambua-Ngwa, A. et al. Major subpopulations of Plasmodium falciparum in sub-Saharan
375 Africa. *Science* **365**, 813-816 (2019).
376 26. Lerch, A. et al. Longitudinal tracking and quantification of individual Plasmodium falciparum
377 clones in complex infections. *Sci Rep* **9**, 3333 (2019).
378 27. Lerch, A. et al. Development of amplicon deep sequencing markers and data analysis pipeline for
379 genotyping multi-clonal malaria infections. *BMC Genomics* **18**, 864 (2017).
380 28. Carlson, J.M. et al. Impact of pre-adapted HIV transmission. *Nat Med* **22**, 606-613 (2016).
381 29. Cotton, L.A. et al. Genotypic and functional impact of HIV-1 adaptation to its host population
382 during the North American epidemic. *PLoS Genet* **10**, e1004295 (2014).

383 30. Osier, F.H. et al. New antigens for a multicomponent blood-stage malaria vaccine. *Sci Transl Med*
384 **6**, 247ra102 (2014).
385 31. Feng, G. et al. Human Immunization With a Polymorphic Malaria Vaccine Candidate Induced
386 Antibodies to Conserved Epitopes That Promote Functional Antibodies to Multiple Parasite
387 Strains. *J Infect Dis* **218**, 35-43 (2018).
388 32. Genton, B. et al. A recombinant blood-stage malaria vaccine reduces Plasmodium falciparum
389 density and exerts selective pressure on parasite populations in a phase 1-2b trial in Papua New
390 Guinea. *J Infect Dis* **185**, 820-827 (2002).
391 33. Abbasi, J. COVID-19 and mRNA Vaccines-First Large Test for a New Approach. *JAMA* **324**, 1125-
392 1127 (2020).
393 34. Langmead, B. & Nellore, A. Cloud computing for genomic data analysis and collaboration. *Nat Rev*
394 *Genet* **19**, 208-219 (2018).
395
396
397
398
399
400
401
402
403
404
405
406

407 **Methods**

408 ***Clinical P. falciparum isolates***

409 Venous blood samples were collected in Nyamisati, Pwani Region, Tanzania, in 1999 and 2016 where the
410 population has been followed longitudinally since 1985. At the time of the 1999 survey, parasite
411 prevalence was 64.8% by qPCR[35]. By 2016, prevalence had decreased to 10.2%[36]. Ninety six individuals
412 were included from the 1994 survey: 80 previously shown to be qPCR positive for *P. falciparum* malaria
413 and 16 who tested negative by qPCR using the multiplexed TaqMan method for *Plasmodium* species
414 identification previously described[37]. The 25 isolates included from the 2016 survey were positive by
415 microscopy for *P. falciparum*. This study was approved by the Regional Ethical Committee in Stockholm,
416 Sweden, and the Ethical Review Board of the National Institute for Medical Research in Tanzania.

417

418 ***Construction of mock infections including synthetic sequences from 10 reference msp2 variants***

419 To assess the specificity and sensitivity of sequence and variant calling, genes encoding *msp2* in the *P.*
420 *falciparum* reference strains HB3 (LR131339 REGION: 250760..251530, FC27, 292 bp CE variant), CD01
421 (PfCD01_020011700, FC27, 292 bp CE variant), Dd2 (PfDd2_020009600, FC27, 412 bp CE variant), SN01
422 (PfSN01_020009800, FC27, 292 bp CE variant), KE01 (PfKE01_020009300, FC27, 373 bp CE variant), SD01
423 (PfSD01_020012300, IC, 434 bp CE variant), GN01 (PfGN01_020012100, IC, 539 bp CE variant), 7G8
424 (Pf7G8_020011500, IC, 485 bp CE variant), GB4 (PfGB4_020009800, IC, 533 bp CE variant) and 3D7
425 (PF3D7_0206800, IC, 473 bp CE variant) were chemically synthesized (GeneScript) and inserted into the
426 pUC57 vector by EcoRV (GAT | ATC) blunt ligation. Mixed mock infections were prepared by diluting 1, 10,
427 100, 1000, 10000 or 100000 copies/µl of pUC57 with the different *msp2* cassettes in water to simulate
428 MOIs of 1 to 10 (Supplementary Table 1). Each mock infection was assigned to a unique barcode
429 combination.

430

431

432 ***DNA extraction and complete amplification of the msp2 open reading frame***

433 For all the clinical isolates, DNA was purified from frozen packed cells in EDTA using QIAamp DNA blood
434 mini kit (Qiagen) following the manufacturer instructions. Concentrations were measured by Nanodrop
435 and the DNA was stored at -80C before use. Taking advantage of the high degree of conservation in the
436 5' and 3' ends of the *msp2* gene, one previously described primer for 5' annealing (msp2_fw: 5'-
437 ATGAAGGTAATTAAAACATTGTCTATTATA-3')[17] that encodes *msp2*'s START codon, as well as a newly
438 designed 3' primer (msp2_rv2: 5'-TTATATGAATATGGCAAAAGATAAAACAA-3') that anneals to the end of
439 the *msp2* open reading frame and that encodes *msp2*'s STOP codon were used for a first round of
440 amplification in 15µl reactions. Six units of Phusion Hot Start II DNA Polymerase (Thermo Scientific), 1x
441 Phusion HF Buffer (Thermo Scientific), 0.5µM msp2_fw, 0.5µM msp2_rv2 and 1µl purified DNA were
442 mixed in individual wells of 96-well PCR plates. The first amplification program included 5 steps: 1) 98°C
443 for 30s, 2) 98°C for 10s, 3) 58.1°C for 30s, 4) 72°C for 40s and 5) 72°C for 5min, with 40 cycles of steps 2-
444 4.

445

446 ***Asymmetrical sample barcoding and pooling***

447 Low parasitemias of *P. falciparum* had been previously shown to require nested amplification to
448 specifically detect rare PCR templates[38]. For this, barcoded oligos were designed for a second round of
449 nested amplification. Forty standard PacBio barcodes from the Sequel_RSII_96_barcodes_v1 collection
450 (barcodes bc1001 to bc1020 in the forward primer and barcodes bc1031 to bc1050 in the reverse) were
451 added to the 5' end of high-performance liquid chromatography (HPLC)-purified primers annealing
452 directly downstream or upstream of msp2_F and msp2_rv2, respectively (Supplementary Table 8).
453 Combinations of these primers allowed the barcoding of 384 individual samples in four 96-well PCR plates
454 (Supplementary Fig. 9) for each CCS library that was processed. A reaction volume of 50µl was used for

455 the nested amplification. Twenty units of Phusion Hot Start II DNA Polymerase (Thermo Scientific), 1x
456 Phusion HF Buffer (Thermo Scientific), 0.5 µM forward and reverse primers and 1µl template (product
457 from the first PCR) were used for amplicon barcoding. A PCR program with 7 steps was used: 1) 98°C for
458 30s, 2) 98°C for 10s, 3) 50°C for 30s, 4) 72°C for 40s, 5) 98°C for 10s, 6) 72°C for 40s and 7) 72°C for 5 min;
459 with steps 2-4 ran in 5 cycles (annealing of the 3' region of the barcoding primer only) followed by 35
460 cycles of steps 5 and 6 (annealing along the entire length of the barcoding primer). After barcoding,
461 selected wells containing synthetic mock infection and negative controls, were ran on an agarose gel to
462 verify the presence or the lack thereof of PCR products. The 12 wells from individual plate rows were
463 pooled and cleaned (QIAquick PCR Purification Kit, QIAGEN), and the 30µl elutions resulting from every
464 prep were combined and vortexed thoroughly. Fragment sizes for clean-pooled samples were verified by
465 agarose gel electrophoresis. Concentration, as well as 260/280 and 260/230 nm ratios, were measured in
466 a Thermo Scientific Multiskan GO.
467
468 ***Single Molecule Real Time (SMRT) library construction for CCS and Lima demultiplexing***
469 PacBio libraries were produced using the SMRTbell™ Template Prep Kit 1.0 according to manufacturer's
470 instructions. In brief, 500 ng of DNA (amplicon) was purified using PB AMPure beads before DNA damage
471 repair and end-repair followed by ligation of hair-pin adaptors to generate a SMRTbell™ library for circular
472 consensus sequencing. The library was then subjected to exo treatment and PB AMPure bead wash for
473 clean-up. Each 384-plex library was sequenced in a single SMRTcell™ 1M for the PacBio Sequel I
474 instrument using a Sequel 3.0 polymerase and 480 minute movie time, allowing for approx. 32 passes of
475 the polymerase on the barcoded *msp2* products (16 passes on the positive strand + 16 passes on the
476 negative strand). CCS reads were assigned to individual samples and primers were computationally
477 trimmed using the Lima package (in asymmetric mode) that is included in the SMRT Link toolbox (version
478 10.0.0.108728). Isolates collected in Nyamisati were sequenced twice to assess the reproducibility in the

479 calling of size and sequence variants. FASTQ files were deposited in the European Nucleotide Archive[39]
480 (Study PRJEB46950, samples ERS12142522 to ERS12143174).
481
482 ***In silico PCR for size variant genotyping***
483 To assess the presence of size variants for the FC27 and IC1 families of *msp2* that could be compared to
484 previous size variant genotyping surveys using nested PCR and CE-based fragment sizing[18], FASTQ files for
485 individual samples were collapsed into an individual tabular file with the unique read identifier in column
486 1 and the read sequence in column 2. Positive or negative strand reads were identified and filtered into
487 two separate tables based on the presence of the tags 5'-AATTTCTTTATTTTTGTTACC-*read*-
488 GTTTTAATTTCAGCAACAC-3' or 5'-GTGTTGCTGAAATTAAAAC-*read*-GGTAACAAAAATAAAGAAATT-3',
489 respectively. These tags correspond to the annealing regions on the *msp2* sequence for the barcoding
490 oligos used during library construction on the positive or the negative strands of the gene. The two tables
491 were converted to FASTA and negative strand reads were reverse-complemented and concatenated tail-
492 to-head to the reads from the positive strand into a single FASTA file. A blastn-short search optimized for
493 sequences shorter than 50 bases was ran on this FASTA file[40, 41], using the sequence from the eight oligos
494 widely used for size variant genotyping of *msp2*[17] and the ten new primers designed for this study as
495 queries. The primers designed for this study incorporate base substitutions observed for the annealing
496 sites in the *msp2* variants of reference strains ML01, GN01, IT, SD01, TG01, GB4, KE01, Dd2, CD01 and HB3
497 (Supplementary Table 3). Expectation value cutoff was set to 0.001. Since *msp2* is a low complexity locus[42],
498 low complexity regions were not filtered out for the analysis. Maximum number of hits to include in the
499 output dataset was set to 1'000'000. 95% was used as identity cutoff and gaps were not allowed. A
500 minimum query coverage per alignment was adjusted to 100%. All the alignments meeting the expected
501 value criteria (-max_hsps) were included in the final output. Output format (-outfmt) was set to Tabular
502 (first 12 columns in Supplementary Table 9). This file was split into 18 separate subfiles, each containing

503 the mapping results for individual oligos based on the Query accession column using the *Filter Tabular*
504 function (version 2.0.0.) from Galaxy, filtering by regex expression matching and including lines matching
505 the name of the 18 genotyping oligos (Supplementary Table 3). The results for forward and reverse oligos
506 were inner joined (Both 1st & 2nd file) with the *Join two files* wrapper using the saccver column (unique
507 CCS read identifiers) as the key. Annealing strand for both forward and reverse oligos was calculated
508 following equation (1):
509

510 $$(1)\ Annealing\ strand\ = \frac{ssend\ (End\ of\ alignment\ in\ subject) - sstart\ (Start\ of\ alignment\ in\ subject)}{|ssend - sstart|}$$

511

512 A stringent CCS read filtering process was applied with the following criteria: 1) Number of alignment
513 mismatches (mismatch) = 0 for both forward and reverse oligos, 2) qend = 30 for both forward and reverse
514 oligos, 3) No alignment gaps allowed (gapopen = 0) and 4) Forward and reverse oligos map to opposite
515 strands of the CCS read reflected as a positive strand value for the forward oligo and a negative strand
516 value for the reverse oligo. Size variant calling was computed for each filtered CCS read using equation
517 (2):

518 $$(2)\ Size\ variant\ (in\ bp)\ = \frac{Rvstrand\ \times\ (sstartFw\ -\ sstartRv)}{|Rvstrand|} + 1$$

519 Where:

520 sstartFw = Start of alignment in subject (CCS read) for the forward primer (ICF or FC27F)

521 sstartRv = Start of alignment in subject (CCS read) for the reverse primer (ICR or FC27R)

522 Rvstrand = Annealing strand of the Rv oligo (ICR or FC27R)

523 Collections of called size variants and their supporting read coverages (proportion of reads in the sample)
524 for all the samples genotyped were collapsed into read coverage matrices using the *Column Join On*
525 *Multiple Datasets* function on Galaxy. Samples and their called size variants were hierarchically clustered

526 (complete linkage) using Gene Cluster 3.0 with Euclidean as the distance metric. Heat maps were
527 visualized using TreeView 3[43].
528
529 ***Size variant genotyping of msp2 by nested PCR and CE***
530 In order to validate size variant calls from *in silico* PCRs on CCS reads, we performed size variant
531 genotyping of *msp2* by nested PCR and fragment sizing by CE as previously described[18] in a selection of 24
532 *P. falciparum*-positive isolates. In brief, 20µl reactions to amplify the *msp2* locus were prepared by mixing
533 0.02 units/µL of AmpliTaq DNA polymerase (ThermoFisher Scientific), 1x AmpliTaq Buffer II (ThermoFisher
534 Scientific), 250nM msp2_fw, 250nM msp2_rv, 2.0mM MgCl2, 125µM dNTP mix and 3µl purified DNA in
535 individual wells of 96-well PCR plates. This first amplification program included 6 steps: 1) 95°C for 5 min,
536 2) 58°C for 2 min, 3) 72°C for 2 min, 4) 94°C for 1 min, 5) 58°C for 2 min and 6) 72°C for 5 min with 25
537 cycles of steps 2-4. Next, nested PCR was used to separately amplify VIC- or 6-Carboxyfluorescein-labeled
538 IC1- or FC27-specific products, respectively, in 20µl reactions. These included 0.05 (IC1 reaction) or 0.02
539 (FC27 reaction) units/µL of AmpliTaq DNA polymerase (ThermoFisher Scientific), 1x AmpliTaq® Buffer II
540 (ThermoFisher Scientific), 300 (IC1 reaction) or 125 (FC27 reaction) nM of forward and reverse 5'-labeled
541 primers (Snounou G *et al,* 1999 in Supplementary Table 3), 1mM MgCl2, 125µM dNTP mix and 1 µl PCR
542 product from the first reaction. The nested amplification program included 6 steps: 1) 95°C for 5 min, 2)
543 58°C for 1 min, 3) 72°C for 1 min, 4) 94°C for 30s, 5) 58°C for 1 min and 6) 72°C for 5 min with 23 cycles of
544 steps 2-4. Products lengths from the nested reaction were then analyzed in an ABI 3730 PRISM® DNA
545 analyzer. Size variants were called using the GeneMapper® Software 5 and a threshold of 300 relative
546 fluorescence units (RFU).
547
548
549

550 ***Phylogenetic and sequence conservation analysis of isolates in the library pool***

551 All sample-labelled barcode-free reads from clinical isolates were concatenated to the sequence of
552 msp2_fw and the reverse-complemented sequence of msp2_rv2 in the 5' and 3' ends, respectively. A
553 FASTA file was created for unique combinations of samples and non-redundant (in the same sample)
554 nucleotide sequences passing the coverage cutoffs for size and sequence variant calling. Sequences were
555 aligned with MAFFT[44] using the E-INS-i accuracy-oriented method that is suitable for sequences with large
556 unalignable regions (Gap extend penalty 0.0, gap opening penalty 1.53) and sequences in the resulting
557 alignment were reordered according to pairwise similarity. The reference sequence for the ortholog of
558 *msp2* present in *P. billcollinsi* (PlasmoDB: PBILCG01_0209400)[19, 45] was added to the alignment and used
559 as the tree root in the phylogenetic analysis that followed. The ortholog of *msp2* in *P. billcollinsi* is the only
560 one in the subgenus *Laverania* that is closely related to the *P. falciparum* clade where no subfamily (IC1
561 or FC27) can be identified from the presence of annealing sites for size variant genotyping oligos
562 (Supplementary Fig. 10). Maximum likelihood trees were constructed with IQ-TREE[46] using 1000 bootstrap
563 replicates (-bb). 1000 replicates were used for the SH-like approximate likelihood ratio test (SH-aLRT, -
564 alrt) that was calculated for individual tree branches. Trees were visualized and annotated using iTOL v6[47].
565 Branches supported by more than 80% of the bootstrap replicates were highlighted in the resulting trees.
566 In order to study protein sequence conservation, non-redundant reads in the library passing the variant
567 calling cutoffs were translated using Galaxy *transeq* version 5.0.0 in frame one before computing IC- and
568 FC27-specific alignments with MAFFT (E-INS-I method, gap extend penalty for group-to-group alignment:
569 0.123, gap opening penalty at group-to-group alignment: 1.53) using a BLOSUM62 matrix. Alignments
570 were visualized with Jalview[48]. Valdar[49] conservation scores for every residue in the resulting alignments
571 were calculated with AACon[50]. The linear epitopes of well-characterized mouse monoclonal antibodies
572 against the 3D7 and FC27 variants of MSP2[20] were mapped to the alignment consensus sequences for the
573 IC1 and FC27 families.

574 ***Prediction of T cell epitopes on MSP2 variants in the Immune Epitope Database and Analysis Resource***
575 ***(IEDB-AR)***
576 To test the usefulness of the obtained protein sequences in the study of T cell responses specific to MSP2
577 variants, HLA class II binding predictions were computed with the MHC-II Binding prediction tool from
578 IEDB-AR[21] for unique combinations of samples and protein sequences using a seven HLA allele set
579 described to capture 50% of the immune response (DRB1*03:01, DRB1*07:01, DRB1*15:01, DRB3*01:01,
580 DRB3*02:02, DRB4*01:01, DRB5*01:01)[51]. This set has been validated previously in the identification of
581 dominant epitopes independent of ethnicity or HLA variability[52]. Peptide lengths were set to 15 amino
582 acids and epitopes were identified using the consensus method combining four prediction algorithms (NN-
583 align[53], SMM-align[54], CombLib and Sturniolo[55]) with top performances[56]. Epitopes with an adjusted rank
584 equal or lower to 1% (99th percentile of binders) were defined as high-affinity binders[52].
585
586 **Construction of an end-to-end pipeline for the analysis of *msp2* diversity, epidemiology and antigenicity**
587 To accelerate data processing, variant calling and the generation of downstream epidemiological and
588 antigen biochemistry insights, we designed a fully automated Galaxy-based analysis pipeline made of two
589 separate workflows that run on tandem (Fig. 5). The first workflow takes as inputs a collection of
590 demultiplexed FASTQ files (one file per sample in the sequencing library) and a renaming two-column
591 tabular file with the name of each FASTQ file in the collection in the first column and the names of the
592 matching samples in the second. This first workflow renames the FASTQ files of the sequencing library,
593 counts the number of reads per sample and calculates the minimum number of reads required for a
594 sample to be considered *msp2*-positive based on the average number of reads assigned to the different
595 water controls in the library plus 2 standard deviations. A downstream workflow two takes as inputs a
596 tabular file with the collapsed renamed collection (output from workflow 1), a tabular file with the number
597 of reads per sample (output from workflow 1), a FASTA file with the ortholog gene of *msp2* present in *P.*

598 *billcollinsi*[45] (to be used as tree root in the construction of isolate phylogenies) and a FASTA file with the
599 sequences of the oligos used for size variant genotyping. This second workflow filters out *msp2*-negative
600 samples, classifies reads as belonging to the IC1 or FC27 families of *msp2* and calculates size variants for
601 each individual read. A suit of epidemiologically relevant outputs are provided by the pipeline including
602 sequence- and size variant-based number of co-infecting clones (MOI) per sample and maximum
603 likelihood phylogenetic trees for the *msp2* variants present in the library. Finally, the pipeline completes
604 the *msp2* sequence for each read in the dataset by adding to the 5' and 3' ends the corresponding oligos
605 used in the first PCR of the library construction protocol and translates the reads into protein sequences
606 that are then used to predict T cell epitopes, construct multiple sequence alignments and calculate linear
607 epitope abundance matrices. Nine versions of workflow two are available using all the variant calling
608 cutoff combinations from Supplementary Table 2 in order to match FPRs for size and sequence variant
609 calling of 0.001, 0.01 and 0.05. All pipelines and accessory files can be found in GitHub
610 (https://github.com/dfplazag/CCS-Pfal-msp2.git).
611

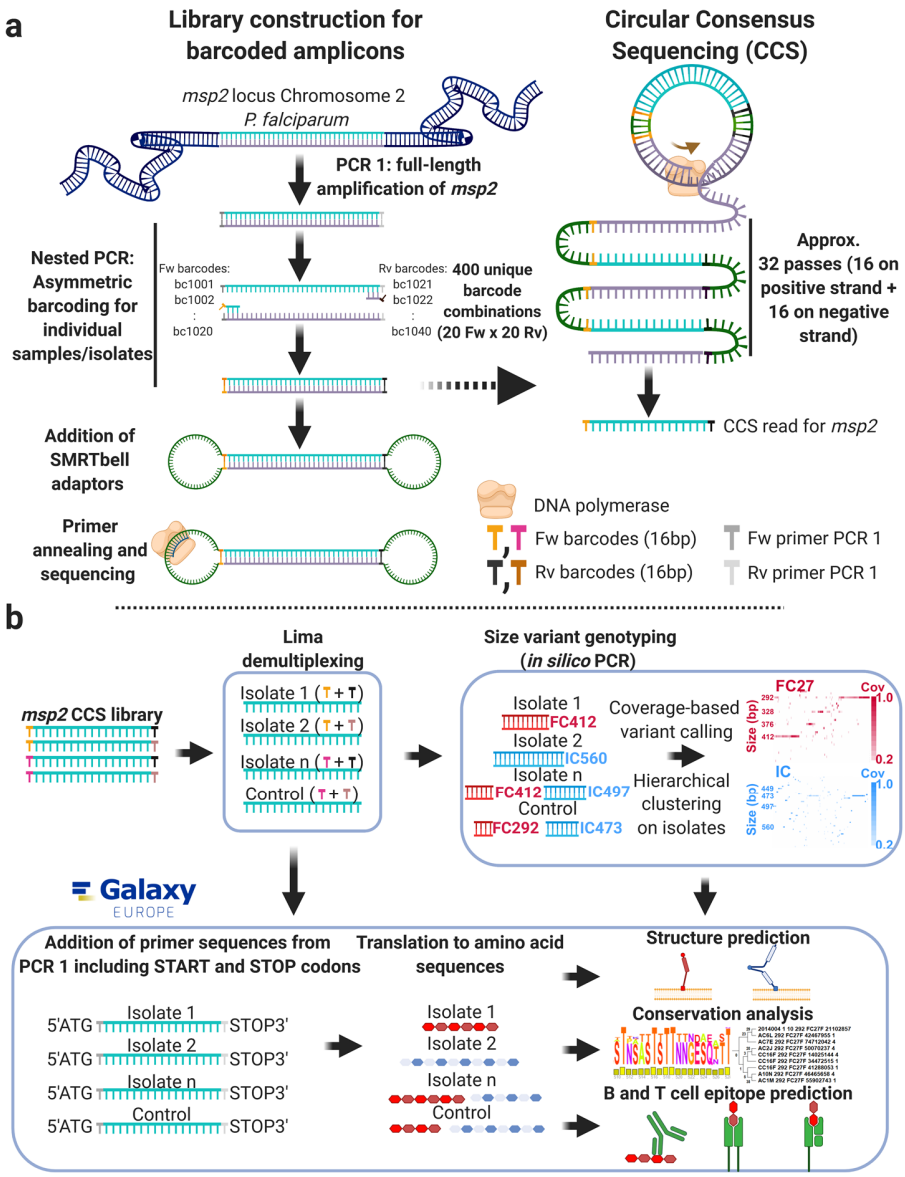
612 **Acknowledgements**

613 The authors would like to thank the study participants and research teams contributing to the clinical
614 malaria isolates collected in Tanzania. The project was funded by the Swedish Research Council (VR MH
615 2018-04468 and 2018-02668) and the Stockholm County Council (Project grant SLL20150135 and FoUI-
616 953118). The authors would also like to acknowledge the National Genomics Infrastructure (NGI) /
617 Uppsala Genome Center and UPPMAX for providing assistance in sequencing and computational
618 infrastructure. Work performed at NGI / Uppsala Genome Center has been funded by RFI / VR and Science
619 for Life Laboratory, Sweden. The data handling and storage were enabled by resources provided by the
620 Swedish National Infrastructure for Computing (SNIC) at the Uppsala Multidisciplinary Center for
621 Advanced Computational Science partially funded by the Swedish Research Council (VR 2018-05973). The

Galaxy server that was used for some calculations is in part funded by Collaborative Research Centre 992 Medical Epigenetics (DFG grant SFB 992/1 2012) and German Federal Ministry of Education and Research (BMBF grants 031 A538A/A538C RBC, 031L0101B/031L0101C de.NBI-epi, 031L0106 de.STAIR (de.NBI)). Finally, we are grateful to Dr. Carlos Fernando Suárez for the insightful discussions on the population dynamics of circulating T cell epitopes from *P. falciparum*.

646 **Figures and legends**

647 **Fig. 1**



648

649 **Fig. 1. Library construction and analysis pipeline for *msp2* genotyping by circular consensus sequencing.**

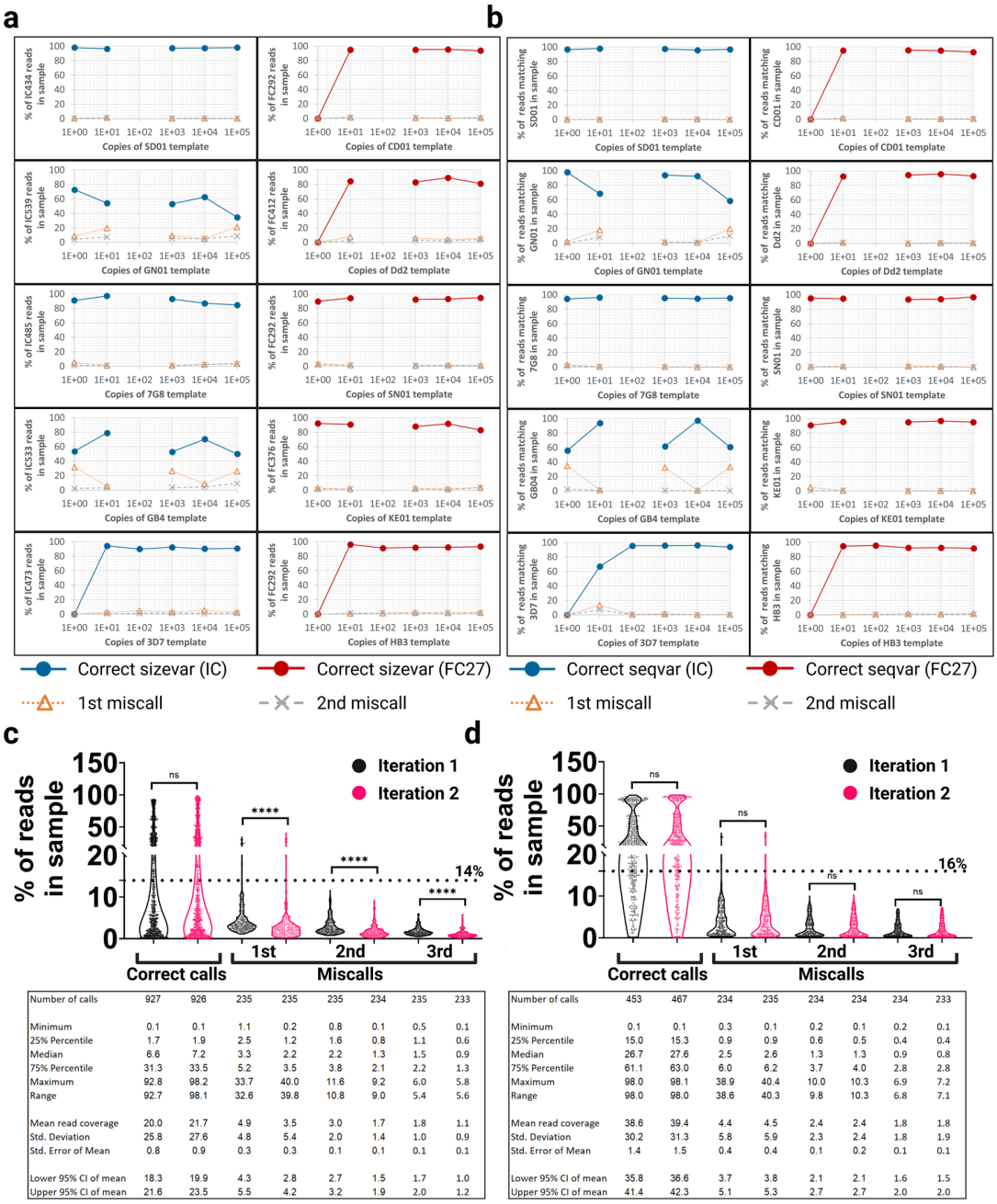
650 **a,** Full-length *msp2* was amplified in a first round of PCR that used primers annealing to the conserved 5'

651 and 3' ends of the *msp2* gene. In the nested PCR that followed, *msp2* amplicons were asymmetrically

652 barcoded using primers from the conserved 5' and 3' regions of the gene but annealing directly

653 downstream and upstream from the annealing sites for the forward and reverse oligos, respectively, used
654 in the first PCR reaction. Asymmetric amplicon barcoding allowed for the accurate pooling and
655 computational identification of up to 384 individual samples using a combination of 40 different barcodes.
656 Thereafter, SMRTbell adaptors were coupled to the barcoded amplicons resulting in single-stranded
657 circularized constructs containing both strands of *msp2*. Circularized *msp2* amplicons were sequenced in
658 32 polymerase passes (16 on each strand) and the resulting sequences for each amplicon were aligned to
659 produce circular consensus reads. **b,** Circular consensus sequences (CCSs) were de-multiplexed with Lima
660 and a combined FASTA file was created with all these indexed CCSs reads. Size variant genotyping was
661 done by *in silico* PCR on this FASTA file using the IC1 and FC27 primer sequences designed by Snounou *et*
662 *al*, 1999, in addition to new primers designed for this study as queries in a Blastn search against the
663 dataset. Isolates and size variants were hierarchically clustered and visualized in two sequencing coverage
664 heatmaps for IC1 and FC27. In addition, CCS reads were bioinformatically coupled to the primers from the
665 first PCR to construct isolate phylogenies or be translated into amino acid sequences to allow the
666 prediction of structural elements, the study of sequence conservation, as well as the presence of B and T
667 cell epitopes.
668
669
670
671
672
673
674

675 **Fig. 2**



Number of calls	927	926	235	235	235	234	235	233
Minimum	0.1	0.1	1.1	0.2	0.8	0.1	0.5	0.1
25% Percentile	1.7	1.9	2.5	1.2	1.6	0.8	1.1	0.6
Median	6.6	7.2	3.3	2.2	2.2	1.3	1.5	0.9
75% Percentile	31.3	33.5	5.2	3.5	3.8	2.1	2.2	1.3
Maximum	92.8	98.2	33.7	40.0	11.6	9.2	6.0	5.8
Range	92.7	98.1	32.6	39.8	10.8	9.0	5.4	5.6
Mean read coverage	20.0	21.7	4.9	3.5	3.0	1.7	1.8	1.1
Std. Deviation	25.8	27.6	4.8	5.4	2.0	1.4	1.0	0.9
Std. Error of Mean	0.8	0.9	0.3	0.3	0.1	0.1	0.1	0.1
Lower 95% CI of mean	18.3	19.9	4.3	2.8	2.7	1.5	1.7	1.0
Upper 95% CI of mean	21.6	23.5	5.5	4.2	3.2	1.9	2.0	1.2

Number of calls	453	467	234	235	234	234	234	233
Minimum	0.1	0.1	0.3	0.1	0.2	0.1	0.2	0.1
25% Percentile	15.0	15.3	0.9	0.9	0.6	0.5	0.4	0.4
Median	26.7	27.6	2.5	2.6	1.3	1.3	0.9	0.8
75% Percentile	61.1	63.0	6.0	6.2	3.7	4.0	2.8	2.8
Maximum	98.0	98.1	38.9	40.4	10.0	10.3	6.9	7.2
Range	98.0	98.0	38.6	40.3	9.8	10.3	6.8	7.1
Mean read coverage	38.6	39.4	4.4	4.5	2.4	2.4	1.8	1.8
Std. Deviation	30.2	31.3	5.8	5.9	2.3	2.4	1.8	1.9
Std. Error of Mean	1.4	1.5	0.4	0.4	0.1	0.2	0.1	0.1
Lower 95% CI of mean	35.8	36.6	3.7	3.8	2.1	2.1	1.6	1.5
Upper 95% CI of mean	41.4	42.3	5.1	5.3	2.7	2.7	2.0	2.0

676

677 **Fig. 2. Sensitivity and specificity in the calling of size and sequence variants, as well as variant calling**

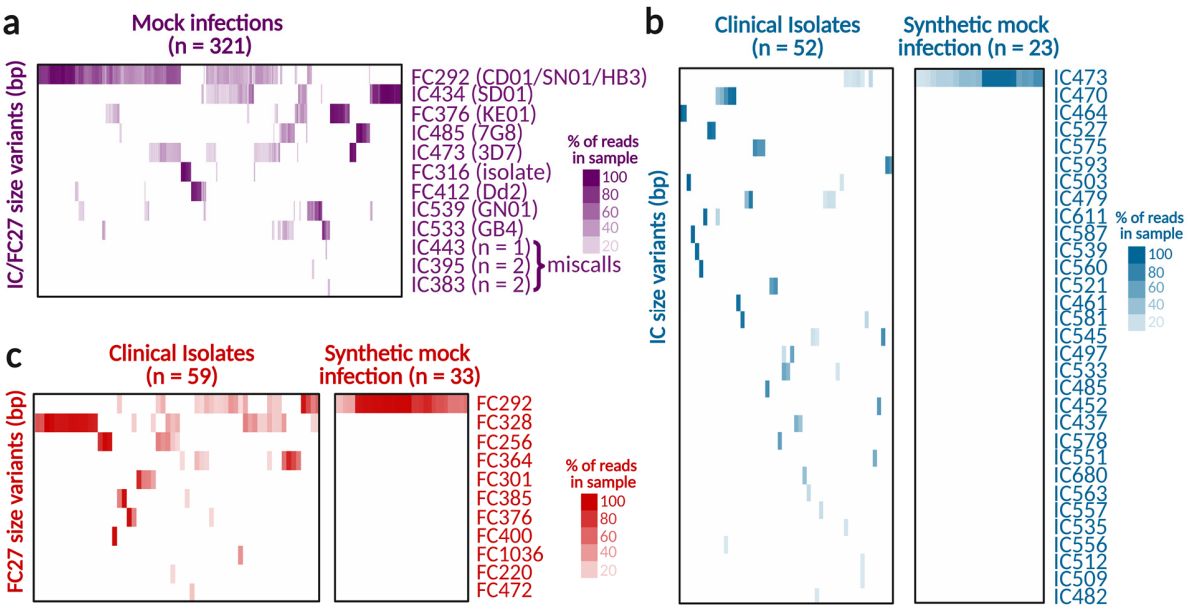
678 **thresholds can be established using mock infections with synthetic constructs of *msp2*.** Serial dilutions

679 of pure synthetic *msp2* from the reference lines SD01 (IC434), GN01 (IC539), 7G8 (IC485), GB4 (IC533),

680 3D7 (IC473), CD01 (FC292), Dd2 (FC412), SN01 (FC292), KE01 (FC376) and HB3 (FC292) were prepared and

individually barcoded (each step in the dilution series with a unique barcoding combination), pooled together and sequenced. Dataset noise was measured as the read coverage for the most frequent miscalls in each sample. The correct **a,** size and **b,** sequence variants were called for a range of template concentrations (number of template copies) in single-clone mock infections. Coverage for the correct variant (Correct sizevar or Correct seqvar), the most common variant miscall (1st miscall) and the second most common variant miscall (2nd miscall) in the IC1 (in blue) and FC27 (in red) families of *msp2* are shown. Percentage of reads per sample for **c,** size and **d,** sequence variant correct calls and miscalls present in a set of 235 synthetic mock infections of different complexity (from 1 to 10 clones per infection mixed in different molar ratios to emulate possible infection scenarios found in clinical isolates). Read coverage was measured in the first and second iterations of the analysis workflow, showing that the noise reduction module introduced between the two iterations significantly reduces the percentage of reads in the sample assigned to the three most frequent size variant miscalls (Mann-Whitney pairwise comparison, **** P value < 0.0001). Mean and SD are shown. Based on these, thresholds for variant calling were set to the mean coverage of the most common miscall plus two standard deviations, being these 14% and 16% for size and sequence variants (shown as dotted lines in **c,** and **d,**), respectively.

703 **Fig. 3**



704

705 **Fig. 3. Size variant genotyping simulated by *in silico* PCR on CCS reads from synthetic mock infections**

706 **and clinical isolates**. **a,** Variant calling for 321 mock infection controls. Mock infections were prepared

707 with 10 synthetic gene variants of *msp2* (in parenthesis) in concentrations of 1 to 100000 copies per

708 reaction. Infections ranged in complexity from 1 to 10 variants per mixture. One clinical isolate (FC316)

709 was spiked-in and detected in 11 samples included in the pool. IC443, IC395 and IC383 are miscalls present

710 in 5 out of 321 mock infections (1.6%). **b,** IC1 and **c,** FC27 size variant genotyping for 95 clinical isolates

711 collected in Nyamisati, Tanzania, in 1999 and 2016. Size variant calling is shown for synthetic mock

712 infections used as positive controls in the CCS libraries where the isolates were sequenced and including

713 a mix of one IC1 (3D7, IC473) and one FC27 (HB3, FC292) variant. Isolates, mock infections and size variants

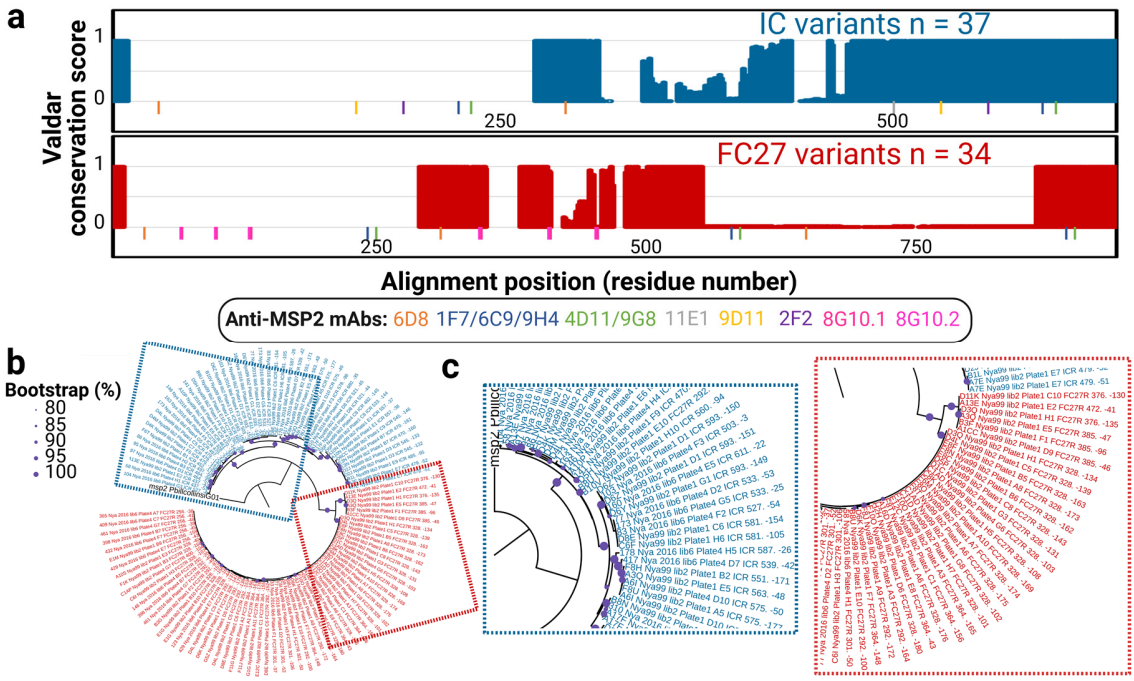
714 were grouped by complete hierarchical clustering using Euclidian as the distance metric.

715

716

717

718 **Fig. 4**



719

720 **Fig. 4. Conservation of MSP2 protein sequences and phylogenetic analysis of malaria isolates based on**

721 ***msp2***. Amino acid and nucleotide sequences were aligned with MAFFT. **a,** Valdar conservation scores were

722 plotted for every residue position in the alignment for the IC1 and FC27 families. MSP2 regions potentially

723 targeted by broadly-neutralizing (strain-transcending) antibodies show high Valdar conservation scores.

724 The location of epitopes recognized by previously characterized mouse monoclonal antibodies is shown

725 under each conservation plot. **b,** Maximum likelihood phylogenetic tree for 51 and 62 unique clones of

726 the IC1 (in blue) and FC27 (in red) families, respectively, found in isolates collected in Tanzania in 1999

727 and 2016. Tree branches supported by 80% or more of the bootstrapping replicates are marked with

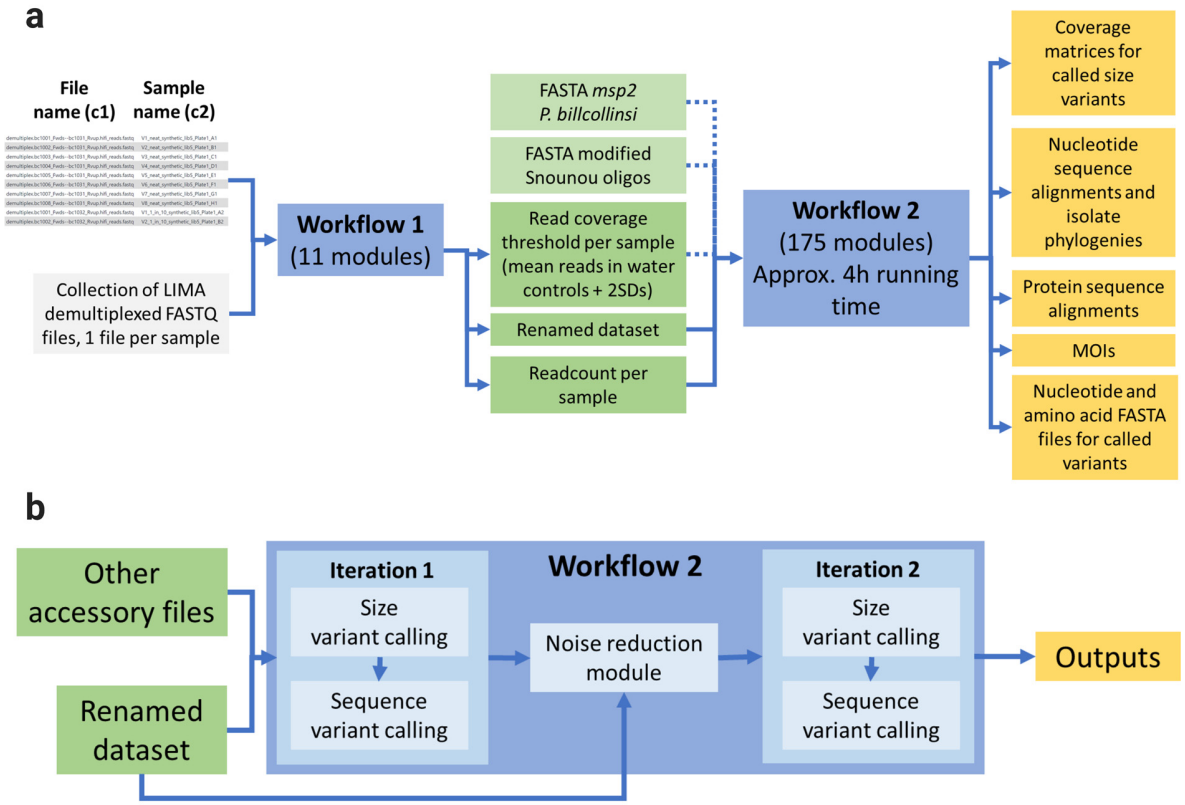
728 purple circles. The tree was rooted using the ortholog of *msp2* present in the species *P. billcollinsi*

729 (*Laverania* subgenus). **c,** The detailed structure of the trees is shown in the dashed frame panels.

730

731

732 **Fig. 5**



733

734 **Fig. 5. An iterative analysis pipeline streamlines the processing and interpretation of epidemiological**
735 **and antigenic diversity data from demultiplexed FASTQ files. a,** General structure of the end-to-end
736 analysis pipeline constructed for *msp2* sequencing data in the bioinformatics analysis interface Galaxy.
737 The pipeline uses as sole inputs a collection of demultiplexed FASTQ files, and three accessory files used
738 for size variant genotyping, file renaming and the construction of isolate phylogenies. Different pipeline
739 outputs are also shown (see also Supplementary Fig. 7 and Supplementary Fig. 8) **b,** A more detailed
740 representation of workflow 2 shows the use of iterative noise reduction for the removal from the dataset
741 of common miscalls resulting from the addition or deletion of up to three bases during PCR amplification
742 or sequencing.

743

744 **Online References**
745

746 35. Farnert, A. et al. Epidemiology of malaria in a village in the Rufiji River Delta, Tanzania: declining
747 transmission over 25 years revealed by different parasitological metrics. *Malar J* **13**, 459 (2014).
748 36. Yman, V. et al. Persistent transmission of Plasmodium malariae and Plasmodium ovale species in
749 an area of declining Plasmodium falciparum transmission in eastern Tanzania. *PLoS Negl Trop*
750 *Dis* **13**, e0007414 (2019).
751 37. Shokoples, S.E., Ndao, M., Kowalewska-Grochowska, K. & Yanow, S.K. Multiplexed real-time PCR
752 assay for discrimination of Plasmodium species with improved sensitivity for mixed infections. *J*
753 *Clin Microbiol* **47**, 975-980 (2009).
754 38. Snounou, G. et al. High sensitivity of detection of human malaria parasites by the use of nested
755 polymerase chain reaction. *Mol Biochem Parasitol* **61**, 315-320 (1993).
756 39. Toribio, A.L. et al. European Nucleotide Archive in 2016. *Nucleic Acids Res* **45**, D32-D36 (2017).
757 40. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
758 41. Cock, P.J., Chilton, J.M., Gruning, B., Johnson, J.E. & Soranzo, N. NCBI BLAST+ integrated into
759 Galaxy. *Gigascience* **4**, 39 (2015).
760 42. Feng, Z.P. et al. Abundance of intrinsically unstructured proteins in P. falciparum and other
761 apicomplexan parasite proteomes. *Mol Biochem Parasitol* **150**, 256-267 (2006).
762 43. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-
763 wide expression patterns. *Proc Natl Acad Sci U S A* **95**, 14863-14868 (1998).
764 44. Katoh, K. & Standley, D.M. MAFFT multiple sequence alignment software version 7:
765 improvements in performance and usability. *Mol Biol Evol* **30**, 772-780 (2013).
766 45. Otto, T.D. et al. Genomes of all known members of a Plasmodium subgenus reveal paths to
767 virulent human malaria. *Nat Microbiol* **3**, 687-697 (2018).

768 46. Nguyen, L.T., Schmidt, H.A., von Haeseler, A. & Minh, B.Q. IQ-TREE: a fast and effective
769 stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**, 268-274
770 (2015).
771 47. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments.
772 *Nucleic Acids Res* **47**, W256-W259 (2019).
773 48. Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M. & Barton, G.J. Jalview Version 2--a
774 multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189-1191
775 (2009).
776 49. Valdar, W.S. Scoring residue conservation. *Proteins* **48**, 227-241 (2002).
777 50. MacGowan, S.A. et al. The Dundee Resource for Sequence Analysis and Structure Prediction.
778 *Protein Sci* **29**, 277-297 (2020).
779 51. Paul, S. et al. Development and validation of a broad scheme for prediction of HLA class II
780 restricted T cell epitopes. *J Immunol Methods* **422**, 28-34 (2015).
781 52. Grifoni, A. et al. Targets of T Cell Responses to SARS-CoV-2 Coronavirus in Humans with COVID-
782 19 Disease and Unexposed Individuals. *Cell* **181**, 1489-1501 e1415 (2020).
783 53. Jensen, K.K. et al. Improved methods for predicting peptide binding affinity to MHC class II
784 molecules. *Immunology* **154**, 394-406 (2018).
785 54. Nielsen, M., Lundegaard, C. & Lund, O. Prediction of MHC class II binding affinity using SMM-
786 align, a novel stabilization matrix alignment method. *BMC Bioinformatics* **8**, 238 (2007).
787 55. Sturniolo, T. et al. Generation of tissue-specific and promiscuous HLA ligand databases using
788 DNA microarrays and virtual HLA class II matrices. *Nat Biotechnol* **17**, 555-561 (1999).
789 56. Wang, P. et al. A systematic assessment of MHC class II peptide binding predictions and
790 evaluation of a consensus approach. *PLoS Comput Biol* **4**, e1000048 (2008).

791