A genomic platform for epidemiological surveillance and vaccine antigen discovery 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 are under strong selective pressure to diversify in terms of antigenicity. We present a sequencing and data analysis platform for epidemiological surveillance and discovery of indel-rich vaccine antigens by long-read circular consensus sequencing (CCS) in multiclonal pathogen 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 of the indel-rich gene merozoite surface protein 2 of *Plasmodium falciparum* 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 from demultiplexed FASTQ files. This platform can be easily adapted to other polymorphic antigens of interest in *Plasmodium* and other human pathogens.

Keywords: Genomic surveillance, long-read sequencing, antigen discovery, epidemiology

48 There were 241 million cases of malaria and 627 thousand deaths from the disease in 2020[1]. An effective
49 vaccine against malaria is a cost-effective and urgently needed strategy to reduce the disease burden and
50 potentially eradicate malaria in the future[2]. RTS,S/AS01 (Mosquirix™) is the first malaria vaccine to be
51 recently endorsed by the World Health Organization for use in children[3]. Nevertheless, the protection
52 conferred by RTS,S/AS01 is strain-specific[4], therefore limiting the effectiveness of this vaccine in regions
53 where *P. falciparum* circulates with a high genetic and antigenic diversity[5]. New approaches to survey the
54 pool of genetic diversity in highly polymorphic antigens from *P. falciparum* and other pathogens will
55 accelerate the development of strain-transcending vaccines.
56
57 Merozoite surface protein 2 (*msp2*) is highly expressed in the blood stage of *P. falciparum*[6]. The protein
58 has two highly conserved domains in the N and the C termini flanking a highly polymorphic central region.
59 The locus is a hotspot for indels of up to 300bp[7], making its sequencing and assembly, especially for
60 multiclonal infections, a challenging task using short-read sequencing[8]. Insertions, deletions or tandem
61 repeats larger than the read lengths in short-read sequencing represent another significant mapping and
62 assembly challenge.
63
64 Here, we present a multiplex long-read amplicon sequencing platform for the accurate genotyping of
65 indel-rich antigens using the malaria vaccine candidate MSP2 as a proof of concept. The use of this circular
66 consensus sequencing (CCS) technology to obtain single-base resolution data on polymorphic antigens
67 will facilitate the implementation of a comprehensive genomic surveillance strategy for infectious
68 diseases caused by highly diverse pathogens and will allow the in-depth structural and antigenic
69 characterization of vaccine candidates capable of eliciting strain-transcending immunity.
70
71

72 **Results**

73 ***CCS on pools of amplicons provides insights into complex infections with high sequence accuracy***

74 Combining nested PCR, asymmetric barcoding of amplicons and CCS, we developed a platform for the
75 genotyping by sequencing of loci rich in structural variants, in up to 384 pooled controls and clinical
76 isolates (Fig. 1). With an average read length of 28300 bp and a mean insert size of 862 bp, every *msp2*
77 amplicon was sequenced 32 times on average, generating circular consensus reads where most
78 sequencing errors (previously estimated to be one error every 13048 bp in CCS reads[9]) are properly
79 amended. We observed no correlation between template concentration in a series of mock infection
80 controls including single *msp2* variants and the number of reads resulting from those samples
81 (Supplementary Fig. 1). Similarly, there was no correlation between sample parasite density (parasites/µl)
82 and the number of CCS reads for microscopy-positive samples in the Tanzania cohort (Supplementary Fig.
83 2).

84

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

86 Size and sequence variant calling for clinical isolates was based on the accuracy of variant detection in
87 synthetic mock infection controls. These controls corresponded to 235 single or mixed serial dilutions
88 including various molar ratios of synthetic reference sequences for *msp2* HB3 (LR131339 REGION:
89 250760..251530), CD01 (PfCD01_020011700), Dd2 (PfDd2_020009600), SN01 (PfSN01_020009800), KE01
90 (PfKE01_020009300), SD01 (PfSD01_020012300), GN01 (PfGN01_020012100), 7G8 (Pf7G8_020011500),
91 GB4 (PfGB4_020009800) and 3D7 (PF3D7_0206800), that were barcoded separately and pooled together
92 before adaptor ligation and sequencing (Supplementary Table 1). Serial dilution of single-variant mock
93 infections showed that the correct size (Fig. 2a) and sequence (Fig. 2b) variant for reference alleles can be
94 detected with read coverages higher than 50% at concentrations of 10 copies/µl for all the variants and
95 down to 1 copy/µl for 6 out of 10. We observed that all samples with barcode bc1034, including the 100

96 copies/µl dilution for variants SD01, GN01, 7G8, GB4, CD01, Dd2, SN01 and KE01, did not yield reads above
97 background levels (Supplementary Fig. 3) and were therefore not included in the sensitivity analysis.
98
99 Quantification of left-over noise after the noise-reduction module (see below) revealed that the mean
100 read coverage for the most common size variant miscall in 235 mock infections was 3.5% of reads (SD 5.4,
101 95% CI 2.8-4.2%) (Fig. 2c). A downstream analysis of sequence variants for the same pool of mock
102 infections showed that the mean read coverage for the most common sequence variant miscall was 4.5%
103 of the reads (SD 5.9, 95% CI 3.8-5.3%) (Fig. 2d). The mean read coverage for size variant miscalls at
104 different molar ratios and dilutions in the synthetic mock infections (Fig. 2c) allowed us to setup a variant
105 calling threshold of 14% (the mean read coverage of the most prevalent variant miscall in all the mock
106 infections plus 2 standard deviations (SD)). Similarly, a coverage threshold was defined in the calling of
107 nucleotide sequence variants (Fig. 2d) equivalent to 16% or the mean coverage of the most common
108 sequence miscall in the synthetic controls plus 2 SD. For the same 235 mock infections, a total of 1140
109 correct size variant calls were expected; nevertheless, 926 (81.2%) of those were detected in the dataset
110 with a total of 360 above the noise threshold (31.2% of the expected, false negative rate (FNR): 0.688).
111 Furthermore, 7407 size variant miscalls were detected, from which only 9 had read coverages above or
112 equal to the established 14% threshold (False positive rate (FPR): 0.001). As for sequence variants, 1312
113 correct calls were expected in the 235 mock infections. From these, 467 (35.6%) were detected with 342
114 above the noise threshold (26.1% of the expected, FNR: 0.739). In addition, 8082 sequence variant miscalls
115 were detected and only 8 of those showed read coverages above the corresponding variant calling cutoff
116 of 16% (FPR: 0.001). Finally, size and sequence variant calling thresholds at FPRs of 0.01 and 0.05 were
117 calculated (Supplementary Table 2). The selection of FPR significantly affects the number of variants that
118 can be detected in individual isolates, with less stringent FPRs (lower variant calling thresholds) resulting
119 in a closer match between the number of variants (clones in a clinical isolate, else referred as multiplicity

120 of infection (MOI)) expected and observed, however compromising the sequence accuracy of the variants
121 called (Supplementary Fig. 4).
122
123 ***Size variant genotyping for clinical isolates can be simulated by in silico PCR on CCS reads***
124 Even though sequencing with single nucleotide resolution is the optimal tool to assess isolate complexity
125 and relatedness in molecular epidemiological surveys[10, 11], most of the available genetic diversity data for
126 *P. falciparum* relies on the genotyping of size variants in *msp2*, *msp1* and *glurp*[12]. *In silico* PCR on CCS reads
127 with the widely used family-specific (IC1 and FC27) primers for *msp2*[13] was applied to call size variants on
128 isolates which could then be compared to data previously generated by nested PCR (using the same
129 primers) and analysis of fragment sizes by capillary electrophoresis (CE)[14]. Fully sequenced and annotated
130 genomes for a number *P. falciparum* strains became available (PlasmoDB[15]) in the years following the
131 original *msp2* genotyping publication[13] and we used these sequences to design forward oligos that
132 incorporate new base substitutions (Supplementary Fig. 5 and Supplementary Table 3).
133
134 *In silico* PCR on the library of synthetic mock infections using the original and the newly designed primers
135 followed by size variant calling showed that the eight size variants expected (three different sequence
136 variants correspond to FC292) were detected. In addition, a clinical isolate corresponding to an FC316
137 variant and spiked into some of the mock infection samples was found in 10 samples at an FPRs of 0.001
138 (Fig. 3a). Only three miscalls were detected in 5 out of the 321 (1.6%) mock infections analyzed using the
139 same FPR. Lastly, *in silico* PCR for size variant genotyping in clinical isolates that were sequenced in a
140 separate library, showed that 52 (54.7%) out of 95 isolates were IC1-positive (Fig. 3b) while 59 (62.1%)
141 were FC27-positive (Fig. 3c). Sixteen samples (16.8%) were found to be double positive for IC1 and FC27.
142

143 Even though amplification conditions vary widely in the PCR programs for size variant genotyping by CE
144 and the barcoding of *msp2* amplicons, we decided to compare the specific size variants being called by CE
145 and CCS on 24 *msp2*-positive samples from Tanzania that were randomly selected. A 7bp plus or minus
146 window in the CE calls was used to account for the size variant binning required when using this method[14].
147 We found that for 15 out of 24 *msp2*-positive samples (62.5%, Supplementary Table 4) there was at least
148 one match in the size variants called by the two methods with high relative fluorescence units (RFU>300)
149 and read coverage values (≥14% of reads in the sample).
150
151 ***Isolate-derived msp2 sequences allow the analysis of antigen structural conservation and the***
152 ***construction of isolate phylogenies***
153 Finally, isolate phylogenies, analysis of B-cell epitope conservation and prediction of T cell epitopes were
154 computed for the same 95 *P. falciparum* isolates. Conservation analysis of unique MSP2 variants of *P.*
155 *falciparum* in the IC1 (n = 37) and FC27 (n = 34) families for isolates from Tanzania shows that there are
156 well defined regions which are highly conserved in the two families (Fig. 4a) and that could be targeted
157 by vaccine-elicited antibodies to confer strain-transcending protection. We used the consensus sequences
158 derived from the multiple alignment Fourier transform (MAFFT) of the MSP2 variants present in the
159 isolates to map linear epitopes bound by well-characterized monoclonal antibodies (mAbs) raised against
160 the 3D7 and FC27 variants[16]. We observed that epitopes present in conserved regions (6D8, 1F7/6C9/9H4
161 and 4D11/9G8) can also be present in the variable regions of some of the clinical isolates sequenced (Fig.
162 4a). Furthermore, sequences of a polymorphic locus like *msp2* are instrumental in the construction of
163 phylogenies describing the degree of relatedness between different *P. falciparum* isolates. We show that
164 *msp2* sequences from clinical isolates can be used to construct phylogenies for the two families of the
165 gene (Fig. 4b and Fig. 4c).
166

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

169 To explore how sequence diversity could be affecting the immunogenicity of different antigen variants,
170 we computed binding predictions for all the MSP2 15-mer peptides from the same Tanzania isolates
171 against the widely representative HLA class II alleles DRB1*03:01, DRB1*07:01, DRB1*15:01, DRB3*01:01,
172 DRB3*02:02, DRB4*01:01 and DRB5*01:01 using IEDB-AR[17]. 126444 T cell epitopes were found in unique
173 variants of MSP2 sequenced for Tanzania isolates (Supplementary Table 5). From these, 2217 peptides
174 (Supplementary Table 6) were non-redundant while 29 epitopes out of the 2217 were predicted as strong
175 binders based on an adjusted rank equal or lower than 1% (99th binding percentile) (Supplementary Table
176 7).

177

178 ***An iterative analysis pipeline streamlines the processing and interpretation of epidemiological and***
179 ***antigenic diversity data from demultiplexed FASTQ files***

180 In order to streamline the analysis of *msp2* genotyping data, a fully-automated pipeline was assembled in
181 Galaxy, a language-agnostic platform that integrates tools written in R, Python and other programming
182 languages (https://github.com/dfplazag/CCS-Pfal-msp2.git). We observed that the size variant miscalls
183 with the highest read counts corresponded to sequences up to 3 bases longer or shorter than the expected
184 correct calls (Supplementary Fig. 6a). To reduce the read coverage for these miscalls, an iterative noise
185 reduction strategy was applied where sequences called on a first iteration of the pipeline were used to
186 remove from the original dataset highly similar sequences (up to 3 mismatches based on the 4-mismatch
187 difference for the *msp2*s in the strains CD01 and HB3, Supplementary Fig. 6b) which were also up to 3
188 nucleotides longer or shorter than the sequence called. This approach resulted in a read count reduction
189 of more than 50% for 13 out of the 22 most prevalent size variant miscalls in the dataset (Supplementary
190 Fig. 6c) and a small increase in the number of correct size variant calls from 348 without the noise

191 reduction module to 360 in the set of 235 mock infections. The pipeline is composed of two workflows.
192 The first workflow (Supplementary Fig. 7) takes a collection of demultiplexed FASTQ files and a table with
193 sample names to deliver a tabular dataset including every read assigned to a sample for all the samples in
194 the library and a table with the number of reads for each sample. Workflow 1 also calculates the baseline
195 read count for a sample to be considered *msp2*-positive based on the mean number of reads miss-
196 assigned to water controls plus 2 SDs. Outputs from workflow 1, in addition to a FASTA file with the *msp2*
197 ortholog in *P. billcollinsi* and a second FASTA with the size variant genotyping oligos for *msp2* designed by
198 us and others are the inputs for a second analysis workflow (Fig. 5a). Workflow 2 (Supplementary Fig. 8)
199 runs in two iterations with the noise reduction module filtering out common size variant miscalls between
200 the first and the second iterations (Fig. 5b). While both iterations call size and sequence variants, only
201 iteration two proceeds to construct nucleotide and protein sequence alignments and phylogenies,
202 compute read coverage matrices for size variants in all the samples genotyped, predict T cell epitopes and
203 calculate MOI values for every sample. Workflow 2 can be easily customized to the study of other
204 polymorphic antigens in *P. falciparum* and other pathogens.
205
206 **Discussion**
207 Here, we introduce a long-read amplicon sequencing platform that we believe can be transformative in
208 unlocking the genetic diversity of antigens rich in structural variants. Our platform is also ideal to target
209 and assess the diversity in polymorphic antigens of relevance for the development of immunity to human
210 pathogens for which vaccines are not currently available.
211
212 We show that sensitivity and specificity for the CCS genotyping platform can be accurately estimated using
213 complex synthetic mock infections. Most errors introduced during sequencing are amended when
214 individual sequences resulting from each polymerase pass on the circularized template are aligned while

215 producing consensus reads. Moreover, the size and sequence variants called for the synthetic sequences
216 allowed us to define coverage thresholds that can be adjusted to eliminate PCR or sequencing artifacts
217 from the dataset at different FPRs. For single-variant mock infections, we called the correct size and
218 sequence at concentrations of 10 copies/µl in all the 10 variants tested and down to 1 copy/µl in 6 out of
219 the 10.
220
221 Multiclonal infections represent a significant challenge in the genotyping of pathogen isolates by short-
222 read sequencing. While a logistically cumbersome approach consists of the sorting and sequencing of
223 individual pathogens present in the isolate[18], long-read sequencing of complete amplicons circumvents
224 the need for sorting and sequencing of individual clones or the use of algorithms that try to differentiate
225 and assemble short reads derived from a mix of variants[19]. The use of CCS in the determination of infection
226 complexity showed that relaxing the stringency in the calling of variants results in a much better match
227 between the expected and the observed number of co-infecting alleles, although compromising the
228 sequence accuracy of the variants called. Variant calling thresholds can therefore be adjusted depending
229 on the intended application, with higher FPRs (e.g. in average, seven clones can be identified for infections
230 with expected MOI: 10 at an FPR: 0.05) used for epidemiological surveillance and lower FPRs (e.g. in
231 average, one clone can be identified for infections with expected MOI: 10 at an FPR: 0.001) applied to
232 antigen discovery and vaccine development.
233
234 Phylogeny reconstruction and the study of multiclonal infections are now made possible by the availability
235 of amplicon sequencing data. Phylogenetic analysis of clones recovered from clinical isolates will be a
236 pivotal tool in the genomic surveillance of infectious disease outbreaks. The ease and low cost of
237 sequencing amplicons for thousands of isolates will also allow the study of population structure and gene
238 flows at a scale that is beyond the current data generation and analysis capacity of whole-genome

239 sequencing approaches[20]. In malaria, the use of highly polymorphic *msp2* sequences as clone-specific
240 "fingerprints" will allow the study of recrudescence and treatment failure with unprecedented resolution
241 joining the battery of markers already in use for sequencing-based follow-up of drug resistance in clinical
242 trials such as *cpp*, *msp7*[12], *ama1-D3*[21], *cpmp*[22] and *csp*[4].
243
244 The analysis of antigenic diversity will support the development of highly polymorphic antigens as vaccine
245 candidates in two different fronts. First, by identifying fully conserved regions that can then be targeted
246 by the host immune response therefore conferring strain-transcending protection. Second, by informing
247 the production of multivalent cocktails including variable domains for dozens or hundreds of antigens.
248
249 The high-throughput analysis of T cell epitopes present in indel-rich antigens can be used to study host-
250 pathogen dynamics with exquisite resolution when coupled to the genotyping of HLA[23, 24] for the individual
251 harboring specific antigen variants.
252
253 A complete survey on the antigenic repertoire of highly polymorphic antigens such as MSP2 could inform
254 the development of a strain-transcending polyvalent antigen cocktails. mRNA as an antigen delivery
255 system has been proven to be effective in eliciting powerful and protective immune responses against
256 SARS-CoV-2[25]. In the future, antigen delivery in the form of codon-optimized mRNAs or alternative
257 delivery platforms for multivalent antigen formulations might be capable of eliciting immunity against
258 dozens or even hundreds of circulating antigen variants in a single vaccine shot imitating, in a way, the
259 protection against clinical disease conferred by repeated exposure to multiple infections.
260
261 The implementation of Galaxy for cheap, high-performance computing, facilitates the understanding and
262 sharing of non-sensitive isolate data amongst epidemiologists, wet lab scientists and computational

263 biologists alike. Analysis pipelines and data histories can be easily shared since all the storage and
264 computation takes place in the cloud. Furthermore, the decentralization of storage and computation
265 guarantees the robustness of the data processing and analysis infrastructure. Last but not least, cloud
266 computing infrastructure provides an unparalleled platform to foster international collaborations where
267 analyses can be ran and monitored in real time by collaborators anywhere in the world[26].
268
269 In conclusion, our long-read platform for amplicon sequencing, provides an instrumental tool for the study
270 of complex infections and can be easily adapted to the genotyping of other antigens rich in structural
271 variants. Furthermore, the detailed structural information gained from the sequencing of polymorphic
272 vaccine candidates can now inform the design of polyvalent vaccines capable of transcending the limited
273 protection conferred by formulations targeting individual antigen variants.
274
275 **References**

276 1. WHO World malaria report 2021. (2021).
277 2. Sauboin, C., Van Bellinghen, L.A., Van De Velde, N. & Van Vlaenderen, I. Economic Impact of
278 Introducing the RTS,S Malaria Vaccine: Cost-Effectiveness and Budget Impact Analysis in 41
279 Countries. *MDM Policy Pract* **4**, 2381468319873324 (2019).
280 3. Sinnis, P. & Fidock, D.A. The RTS,S vaccine-a chance to regain the upper hand against malaria?
281 *Cell* **185**, 750-754 (2022).
282 4. Neafsey, D.E. et al. Genetic Diversity and Protective Efficacy of the RTS,S/AS01 Malaria Vaccine.
283 *N Engl J Med* **373**, 2025-2037 (2015).
284 5. Pringle, J.C. et al. RTS,S/AS01 malaria vaccine mismatch observed among Plasmodium
285 falciparum isolates from southern and central Africa and globally. *Sci Rep* **8**, 6622 (2018).

286 6. Otto, T.D. et al. New insights into the blood-stage transcriptome of Plasmodium falciparum
287 using RNA-Seq. *Mol Microbiol* **76**, 12-24 (2010).
288 7. Mwingira, F. et al. Plasmodium falciparum msp1, msp2 and glurp allele frequency and diversity
289 in sub-Saharan Africa. *Malar J* **10**, 79 (2011).
290 8. Chaisson, M.J.P. et al. Multi-platform discovery of haplotype-resolved structural variation in
291 human genomes. *Nat Commun* **10**, 1784 (2019).
292 9. Wenger, A.M. et al. Accurate circular consensus long-read sequencing improves variant
293 detection and assembly of a human genome. *Nat Biotechnol* **37**, 1155-1162 (2019).
294 10. Black, A., MacCannell, D.R., Sibley, T.R. & Bedford, T. Ten recommendations for supporting open
295 pathogen genomic analysis in public health. *Nat Med* **26**, 832-841 (2020).
296 11. Neafsey, D.E., Taylor, A.R. & MacInnis, B.L. Advances and opportunities in malaria population
297 genomics. *Nat Rev Genet* (2021).
298 12. Gruenberg, M., Lerch, A., Beck, H.P. & Felger, I. Amplicon deep sequencing improves
299 Plasmodium falciparum genotyping in clinical trials of antimalarial drugs. *Sci Rep* **9**, 17790
300 (2019).
301 13. Snounou, G. et al. Biased distribution of msp1 and msp2 allelic variants in Plasmodium
302 falciparum populations in Thailand. *Trans R Soc Trop Med Hyg* **93**, 369-374 (1999).
303 14. Liljander, A. et al. Optimization and validation of multi-coloured capillary electrophoresis for
304 genotyping of Plasmodium falciparum merozoite surface proteins (msp1 and 2). *Malar J* **8**, 78
305 (2009).
306 15. Aurrecoechea, C. et al. PlasmoDB: a functional genomic database for malaria parasites. *Nucleic*
307 *Acids Res* **37**, D539-543 (2009).
308 16. Adda, C.G. et al. Antigenic characterization of an intrinsically unstructured protein, Plasmodium
309 falciparum merozoite surface protein 2. *Infect Immun* **80**, 4177-4185 (2012).

310 17. Dhanda, S.K. et al. IEDB-AR: immune epitope database-analysis resource in 2019. *Nucleic Acids*
311 *Res* **47**, W502-W506 (2019).
312 18. Nair, S. et al. Single-cell genomics for dissection of complex malaria infections. *Genome Res* **24**,
313 1028-1038 (2014).
314 19. Zhong, D., Koepfli, C., Cui, L. & Yan, G. Molecular approaches to determine the multiplicity of
315 Plasmodium infections. *Malar J* **17**, 172 (2018).
316 20. Amambua-Ngwa, A. et al. Major subpopulations of Plasmodium falciparum in sub-Saharan
317 Africa. *Science* **365**, 813-816 (2019).
318 21. Lerch, A. et al. Longitudinal tracking and quantification of individual Plasmodium falciparum
319 clones in complex infections. *Sci Rep* **9**, 3333 (2019).
320 22. Lerch, A. et al. Development of amplicon deep sequencing markers and data analysis pipeline for
321 genotyping multi-clonal malaria infections. *BMC Genomics* **18**, 864 (2017).
322 23. Carlson, J.M. et al. Impact of pre-adapted HIV transmission. *Nat Med* **22**, 606-613 (2016).
323 24. Cotton, L.A. et al. Genotypic and functional impact of HIV-1 adaptation to its host population
324 during the North American epidemic. *PLoS Genet* **10**, e1004295 (2014).
325 25. Abbasi, J. COVID-19 and mRNA Vaccines-First Large Test for a New Approach. *JAMA* **324**, 1125-
326 1127 (2020).
327 26. Langmead, B. & Nellore, A. Cloud computing for genomic data analysis and collaboration. *Nat Rev*
328 *Genet* **19**, 208-219 (2018).
329
330
331
332
333

334 **Methods**

335 ***Clinical P. falciparum isolates***

336 Venous blood samples were collected in Nyamisati, Pwani Region, Tanzania, in 1999 and 2016 where the
337 population has been followed longitudinally since 1985. At the time of the 1999 survey, parasite
338 prevalence was 64.8% by qPCR[27]. By 2016, prevalence had decreased to 10.2%[28]. Ninety six individuals
339 were included from the 1994 survey: 80 previously shown to be qPCR positive for *P. falciparum* malaria
340 and 16 who tested negative[29]. The 25 isolates included from the 2016 survey were positive by microscopy
341 for *P. falciparum*. This study was approved by the Regional Ethical Committee in Stockholm, Sweden, and
342 the Ethical Review Board of the National Institute for Medical Research in Tanzania.

343

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

345 To assess the specificity and sensitivity of sequence and variant calling, genes encoding *msp2* in the *P.*
346 *falciparum* reference strains HB3 (LR131339 REGION: 250760..251530, FC27, 292 bp CE variant), CD01
347 (PfCD01_020011700, FC27, 292 bp CE variant), Dd2 (PfDd2_020009600, FC27, 412 bp CE variant), SN01
348 (PfSN01_020009800, FC27, 292 bp CE variant), KE01 (PfKE01_020009300, FC27, 373 bp CE variant), SD01
349 (PfSD01_020012300, IC, 434 bp CE variant), GN01 (PfGN01_020012100, IC, 539 bp CE variant), 7G8
350 (Pf7G8_020011500, IC, 485 bp CE variant), GB4 (PfGB4_020009800, IC, 533 bp CE variant) and 3D7
351 (PF3D7_0206800, IC, 473 bp CE variant) were chemically synthesized (GeneScript) and inserted into the
352 pUC57 vector by EcoRV (GAT | ATC) blunt ligation. Mixed mock infections were prepared by diluting 1, 10,
353 100, 1000, 10000 or 100000 copies/µl of pUC57 with the different *msp2* cassettes in water to simulate
354 MOIs of 1 to 10 (Supplementary Table 1). Each mock infection was assigned to a unique barcode
355 combination.

356

357

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

359 For all the clinical isolates, DNA was purified from frozen packed cells in EDTA using QIAamp DNA blood
360 mini kit (Qiagen) following the manufacturer instructions. Concentrations were measured by Nanodrop
361 and the DNA was stored at -80C before use. Taking advantage of the high degree of conservation in the
362 5' and 3' ends of the *msp2* gene, one previously described primer for 5' annealing (msp2_fw: 5'-
363 ATGAAGGTAATTAAAACATTGTCTATTATA-3')[13] that encodes *msp2*'s START codon, as well as a newly
364 designed 3' primer (msp2_rv2: 5'-TTATATGAATATGGCAAAAGATAAAACAA-3') that anneals to the end of
365 the *msp2* open reading frame and that encodes *msp2*'s STOP codon were used for a first round of
366 amplification in 15µl reactions. Six units of Phusion Hot Start II DNA Polymerase (Thermo Scientific), 1x
367 Phusion HF Buffer (Thermo Scientific), 0.5µM msp2_fw, 0.5µM msp2_rv2 and 1µl purified DNA were
368 mixed in individual wells of 96-well PCR plates. The first amplification program included 5 steps: 1) 98°C
369 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-
370 4.

371

372 ***Asymmetrical sample barcoding and pooling***

373 Low parasitemias of *P. falciparum* had been previously shown to require nested amplification to
374 specifically detect rare PCR templates[30]. For this, barcoded oligos were designed for a second round of
375 nested amplification. Forty standard PacBio barcodes from the Sequel_RSII_96_barcodes_v1 collection
376 (barcodes bc1001 to bc1020 in the forward primer and barcodes bc1031 to bc1050 in the reverse) were
377 added to the 5' end of high-performance liquid chromatography (HPLC)-purified primers annealing
378 directly downstream or upstream of msp2_F and msp2_rv2, respectively (Supplementary Table 8).
379 Combinations of these primers allowed the barcoding of 384 individual samples in four 96-well PCR plates
380 (Supplementary Fig. 9) for each CCS library that was processed. A reaction volume of 50µl was used for
381 the nested amplification. Twenty units of Phusion Hot Start II DNA Polymerase (Thermo Scientific), 1x

382 Phusion HF Buffer (Thermo Scientific), 0.5 µM forward and reverse primers and 1µl template (product
383 from the first PCR) were used for amplicon barcoding. A PCR program with 7 steps was used: 1) 98°C for
384 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;
385 with steps 2-4 ran in 5 cycles (annealing of the 3' region of the barcoding primer only) followed by 35
386 cycles of steps 5 and 6 (annealing along the entire length of the barcoding primer). After barcoding,
387 selected wells containing synthetic mock infection and negative controls, were ran on an agarose gel to
388 verify the presence or the lack thereof of PCR products. The 12 wells from individual plate rows were
389 pooled and cleaned (QIAquick PCR Purification Kit, QIAGEN), and the 30µl elutions resulting from every
390 prep were combined and vortexed thoroughly. Fragment sizes for clean-pooled samples were verified by
391 agarose gel electrophoresis. Concentration, as well as 260/280 and 260/230 nm ratios, were measured in
392 a Thermo Scientific Multiskan GO.
393
394 ***Single Molecule Real Time (SMRT) library construction for CCS and Lima demultiplexing***
395 PacBio libraries were produced using the SMRTbell™ Template Prep Kit 1.0 according to manufacturer's
396 instructions. In brief, 500 ng of DNA (amplicon) was purified using PB AMPure beads before DNA damage
397 repair and end-repair followed by ligation of hair-pin adaptors to generate a SMRTbell™ library for circular
398 consensus sequencing. The library was then subjected to exo treatment and PB AMPure bead wash for
399 clean-up. Each 384-plex library was sequenced in a single SMRTcell™ 1M for the PacBio Sequel I
400 instrument using a Sequel 3.0 polymerase and 480 minute movie time, allowing for approx. 32 passes of
401 the polymerase on the barcoded *msp2* products (16 passes on the positive strand + 16 passes on the
402 negative strand). CCS reads were assigned to individual samples and primers were computationally
403 trimmed using the Lima package (in asymmetric mode) that is included in the SMRT Link toolbox (version
404 10.0.0.108728). Isolates collected in Nyamisati were sequenced twice to assess the reproducibility in the

405 calling of size and sequence variants. FASTQ files were deposited in the European Nucleotide Archive[31]
406 (Study PRJEB46950, samples ERS12142522 to ERS12143174).
407
408 ***In silico PCR for size variant genotyping***
409 To assess the presence of size variants for the FC27 and IC1 families of *msp2* that could be compared to
410 previous size variant genotyping surveys using nested PCR and CE-based fragment sizing[14], FASTQ files for
411 individual samples were collapsed into an individual tabular file with the unique read identifier in column
412 1 and the read sequence in column 2. Positive or negative strand reads were identified and filtered into
413 two separate tables based on the presence of the tags 5'-AATTTCTTTATTTTTGTTACC-*read*-
414 GTTTTAATTTCAGCAACAC-3' or 5'-GTGTTGCTGAAATTAAAAC-*read*-GGTAACAAAAATAAAGAAATT-3',
415 respectively. These tags correspond to the annealing regions on the *msp2* sequence for the barcoding
416 oligos used during library construction on the positive or the negative strands of the gene. The two tables
417 were converted to FASTA and negative strand reads were reverse-complemented and concatenated tail-
418 to-head to the reads from the positive strand into a single FASTA file. A blastn-short search optimized for
419 sequences shorter than 50 bases was ran on this FASTA file[32, 33], using the sequence from the eight oligos
420 widely used for size variant genotyping of *msp2*[13] and the ten new primers designed for this study as
421 queries. The primers designed for this study incorporate base substitutions observed for the annealing
422 sites in the *msp2* variants of reference strains ML01, GN01, IT, SD01, TG01, GB4, KE01, Dd2, CD01 and HB3
423 (Supplementary Table 3). Expectation value cutoff was set to 0.001. Since *msp2* is a low complexity locus[34],
424 low complexity regions were not filtered out for the analysis. Maximum number of hits to include in the
425 output dataset was set to 1'000'000. 95% was used as identity cutoff and gaps were not allowed. A
426 minimum query coverage per alignment was adjusted to 100%. All the alignments meeting the expected
427 value criteria (-max_hsps) were included in the final output. Output format (-outfmt) was set to Tabular
428 (first 12 columns in Supplementary Table 9). This file was split into 18 separate subfiles, each containing

429 the mapping results for individual oligos based on the Query accession column using the *Filter Tabular*
430 function (version 2.0.0.) from Galaxy, filtering by regex expression matching and including lines matching
431 the name of the 18 genotyping oligos (Supplementary Table 3). The results for forward and reverse oligos
432 were inner joined (Both 1st & 2nd file) with the *Join two files* wrapper using the saccver column (unique
433 CCS read identifiers) as the key. Annealing strand for both forward and reverse oligos was calculated
434 following equation (1):
435

436 $$(1)\ Annealing\ strand\ = \frac{ssend\ (End\ of\ alignment\ in\ subject) - sstart\ (Start\ of\ alignment\ in\ subject)}{|ssend - sstart|}$$

437

438 A stringent CCS read filtering process was applied with the following criteria: 1) Number of alignment
439 mismatches (mismatch) = 0 for both forward and reverse oligos, 2) qend = 30 for both forward and reverse
440 oligos, 3) No alignment gaps allowed (gapopen = 0) and 4) Forward and reverse oligos map to opposite
441 strands of the CCS read reflected as a positive strand value for the forward oligo and a negative strand
442 value for the reverse oligo. Size variant calling was computed for each filtered CCS read using equation
443 (2):

444 $$(2)\ Size\ variant\ (in\ bp)\ = \frac{Rvstrand\ \times\ (sstartFw\ -\ sstartRv)}{|Rvstrand|} + 1$$

445 Where:

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

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

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

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

452 (complete linkage) using Gene Cluster 3.0 with Euclidean as the distance metric. Heat maps were
453 visualized using TreeView 3[35].
454
455 ***Size variant genotyping of msp2 by nested PCR and CE***
456 In order to validate size variant calls from *in silico* PCRs on CCS reads, we performed size variant
457 genotyping of *msp2* by nested PCR and fragment sizing by CE as previously described[14] in a selection of 24
458 *P. falciparum*-positive isolates. In brief, 20µl reactions to amplify the *msp2* locus were prepared by mixing
459 0.02 units/µL of AmpliTaq DNA polymerase (ThermoFisher Scientific), 1x AmpliTaq Buffer II (ThermoFisher
460 Scientific), 250nM msp2_fw, 250nM msp2_rv, 2.0mM MgCl2, 125µM dNTP mix and 3µl purified DNA in
461 individual wells of 96-well PCR plates. This first amplification program included 6 steps: 1) 95°C for 5 min,
462 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
463 cycles of steps 2-4. Next, nested PCR was used to separately amplify VIC- or 6-Carboxyfluorescein-labeled
464 IC1- or FC27-specific products, respectively, in 20µl reactions. These included 0.05 (IC1 reaction) or 0.02
465 (FC27 reaction) units/µL of AmpliTaq DNA polymerase (ThermoFisher Scientific), 1x AmpliTaq® Buffer II
466 (ThermoFisher Scientific), 300 (IC1 reaction) or 125 (FC27 reaction) nM of forward and reverse 5'-labeled
467 primers (Snounou G *et al,* 1999 in Supplementary Table 3), 1mM MgCl2, 125µM dNTP mix and 1 µl PCR
468 product from the first reaction. The nested amplification program included 6 steps: 1) 95°C for 5 min, 2)
469 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
470 steps 2-4. Products lengths from the nested reaction were then analyzed in an ABI 3730 PRISM® DNA
471 analyzer. Size variants were called using the GeneMapper® Software 5 and a threshold of 300 relative
472 fluorescence units (RFU).
473
474
475

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

477 All sample-labelled barcode-free reads from clinical isolates were concatenated to the sequence of
478 msp2_fw and the reverse-complemented sequence of msp2_rv2 in the 5' and 3' ends, respectively. A
479 FASTA file was created for unique combinations of samples and non-redundant (in the same sample)
480 nucleotide sequences passing the coverage cutoffs for size and sequence variant calling. Sequences were
481 aligned with MAFFT[36] using the E-INS-i accuracy-oriented method that is suitable for sequences with large
482 unalignable regions (Gap extend penalty 0.0, gap opening penalty 1.53) and sequences in the resulting
483 alignment were reordered according to pairwise similarity. The reference sequence for the ortholog of
484 *msp2* present in *P. billcollinsi* (PlasmoDB: PBILCG01_0209400)[15, 37] was added to the alignment and used
485 as the tree root in the phylogenetic analysis that followed. The ortholog of *msp2* in *P. billcollinsi* is the only
486 one in the subgenus *Laverania* that is closely related to the *P. falciparum* clade where no subfamily (IC1
487 or FC27) can be identified from the presence of annealing sites for size variant genotyping oligos
488 (Supplementary Fig. 10). Maximum likelihood trees were constructed with IQ-TREE[38] using 1000 bootstrap
489 replicates (-bb). 1000 replicates were used for the SH-like approximate likelihood ratio test (SH-aLRT, -
490 alrt) that was calculated for individual tree branches. Trees were visualized and annotated using iTOL v6[39].
491 Branches supported by more than 80% of the bootstrap replicates were highlighted in the resulting trees.
492 In order to study protein sequence conservation, non-redundant reads in the library passing the variant
493 calling cutoffs were translated using Galaxy *transeq* version 5.0.0 in frame one before computing IC- and
494 FC27-specific alignments with MAFFT (E-INS-I method, gap extend penalty for group-to-group alignment:
495 0.123, gap opening penalty at group-to-group alignment: 1.53) using a BLOSUM62 matrix. Alignments
496 were visualized with Jalview[40]. Valdar[41] conservation scores for every residue in the resulting alignments
497 were calculated with AACon[42]. The linear epitopes of well-characterized mouse monoclonal antibodies
498 against the 3D7 and FC27 variants of MSP2[16] were mapped to the alignment consensus sequences for the
499 IC1 and FC27 families.

500 ***Prediction of T cell epitopes on MSP2 variants in the Immune Epitope Database and Analysis Resource***
501 ***(IEDB-AR)***
502 To test the usefulness of the obtained protein sequences in the study of T cell responses specific to MSP2
503 variants, HLA class II binding predictions were computed with the MHC-II Binding prediction tool from
504 IEDB-AR[17] for unique combinations of samples and protein sequences using a seven HLA allele set
505 described to capture 50% of the immune response (DRB1*03:01, DRB1*07:01, DRB1*15:01, DRB3*01:01,
506 DRB3*02:02, DRB4*01:01, DRB5*01:01)[43]. This set has been validated previously in the identification of
507 dominant epitopes independent of ethnicity or HLA variability[44]. Peptide lengths were set to 15 amino
508 acids and epitopes were identified using the consensus method combining four prediction algorithms (NN-
509 align[45], SMM-align[46], CombLib and Sturniolo[47]) with top performances[48]. Epitopes with an adjusted rank
510 equal or lower to 1% (99th percentile of binders) were defined as high-affinity binders[44].
511
512 **Construction of an end-to-end pipeline for the analysis of *msp2* diversity, epidemiology and antigenicity**
513 To accelerate data processing, variant calling and the generation of downstream epidemiological and
514 antigen biochemistry insights, we designed a fully automated Galaxy-based analysis pipeline made of two
515 separate workflows that run on tandem (Fig. 5). The first workflow takes as inputs a collection of
516 demultiplexed FASTQ files (one file per sample in the sequencing library) and a renaming two-column
517 tabular file with the name of each FASTQ file in the collection in the first column and the names of the
518 matching samples in the second. This first workflow renames the FASTQ files of the sequencing library,
519 counts the number of reads per sample and calculates the minimum number of reads required for a
520 sample to be considered *msp2*-positive based on the average number of reads assigned to the different
521 water controls in the library plus 2 standard deviations. A downstream workflow two takes as inputs a
522 tabular file with the collapsed renamed collection (output from workflow 1), a tabular file with the number
523 of reads per sample (output from workflow 1), a FASTA file with the ortholog gene of *msp2* present in *P.*

524 *billcollinsi*[37] (to be used as tree root in the construction of isolate phylogenies) and a FASTA file with the
525 sequences of the oligos used for size variant genotyping. This second workflow filters out *msp2*-negative
526 samples, classifies reads as belonging to the IC1 or FC27 families of *msp2* and calculates size variants for
527 each individual read. A suit of epidemiologically relevant outputs are provided by the pipeline including
528 sequence- and size variant-based number of co-infecting clones (MOI) per sample and maximum
529 likelihood phylogenetic trees for the *msp2* variants present in the library. Finally, the pipeline completes
530 the *msp2* sequence for each read in the dataset by adding to the 5' and 3' ends the corresponding oligos
531 used in the first PCR of the library construction protocol and translates the reads into protein sequences
532 that are then used to predict T cell epitopes, construct multiple sequence alignments and calculate linear
533 epitope abundance matrices. Nine versions of workflow two are available using all the variant calling
534 cutoff combinations from Supplementary Table 2 in order to match FPRs for size and sequence variant
535 calling of 0.001, 0.01 and 0.05. All pipelines and accessory files can be found in GitHub
536 (https://github.com/dfplazag/CCS-Pfal-msp2.git).
537

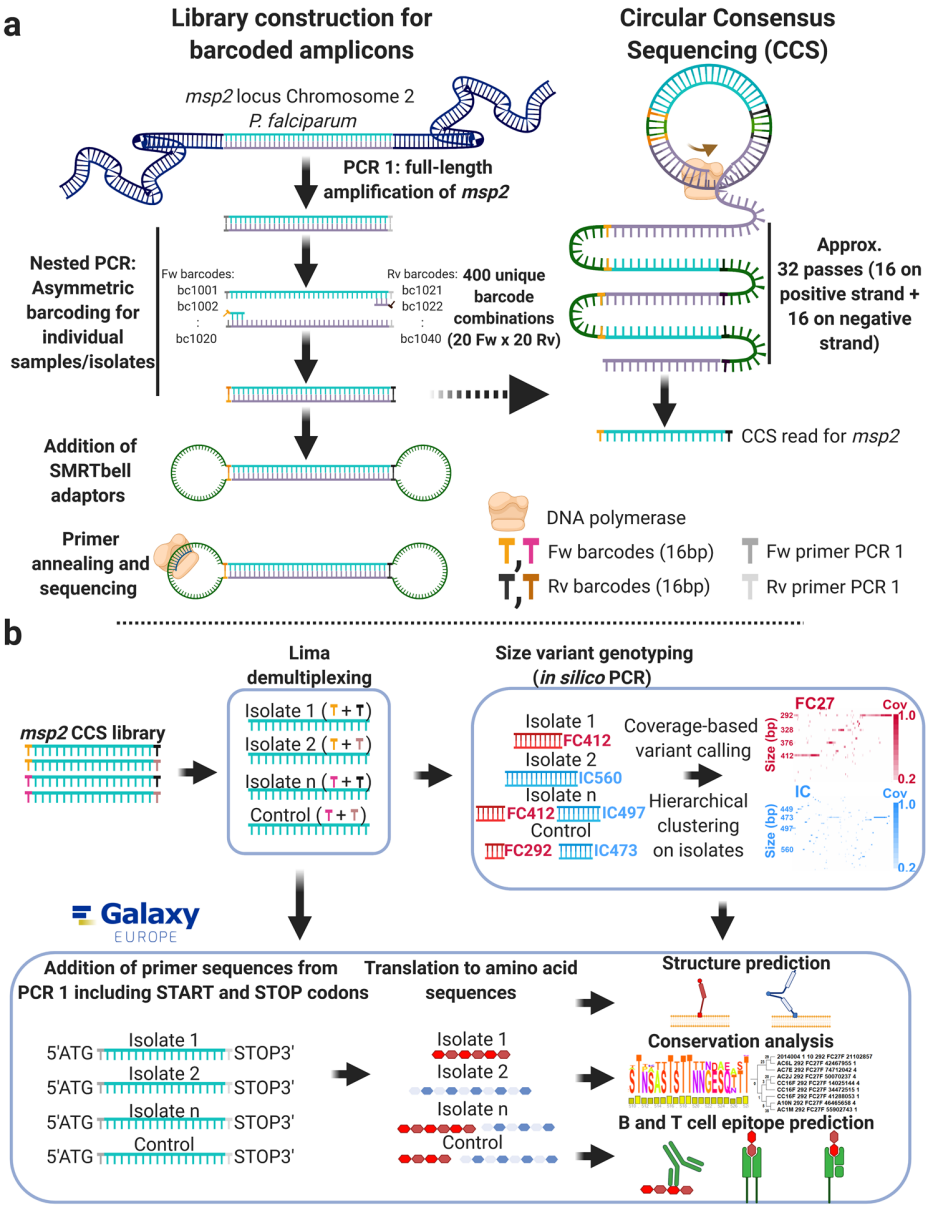
538 **Acknowledgements**

539 The authors would like to thank the study participants and research teams contributing to the clinical
540 malaria isolates collected in Tanzania. The project was funded by the Swedish Research Council (VR MH
541 2018-04468 and 2018-02668) and the Stockholm County Council (Project grant SLL20150135 and FoUI-
542 953118). The authors would also like to acknowledge the National Genomics Infrastructure (NGI) /
543 Uppsala Genome Center and UPPMAX for providing assistance in sequencing and computational
544 infrastructure. Work performed at NGI / Uppsala Genome Center has been funded by RFI / VR and Science
545 for Life Laboratory, Sweden. The data handling and storage were enabled by resources provided by the
546 Swedish National Infrastructure for Computing (SNIC) at the Uppsala Multidisciplinary Center for
547 Advanced Computational Science partially funded by the Swedish Research Council (VR 2018-05973). The

548 Galaxy server that was used for some calculations is in part funded by Collaborative Research Centre 992
549 Medical Epigenetics (DFG grant SFB 992/1 2012) and German Federal Ministry of Education and Research
550 (BMBF grants 031 A538A/A538C RBC, 031L0101B/031L0101C de.NBI-epi, 031L0106 de.STAIR (de.NBI)).
551 Finally, we are grateful to Dr. Carlos Fernando Suárez for the insightful discussions on the population
552 dynamics of circulating T cell epitopes from *P. falciparum*.
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571

572 **Figures and legends**

573 **Fig. 1**



574

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

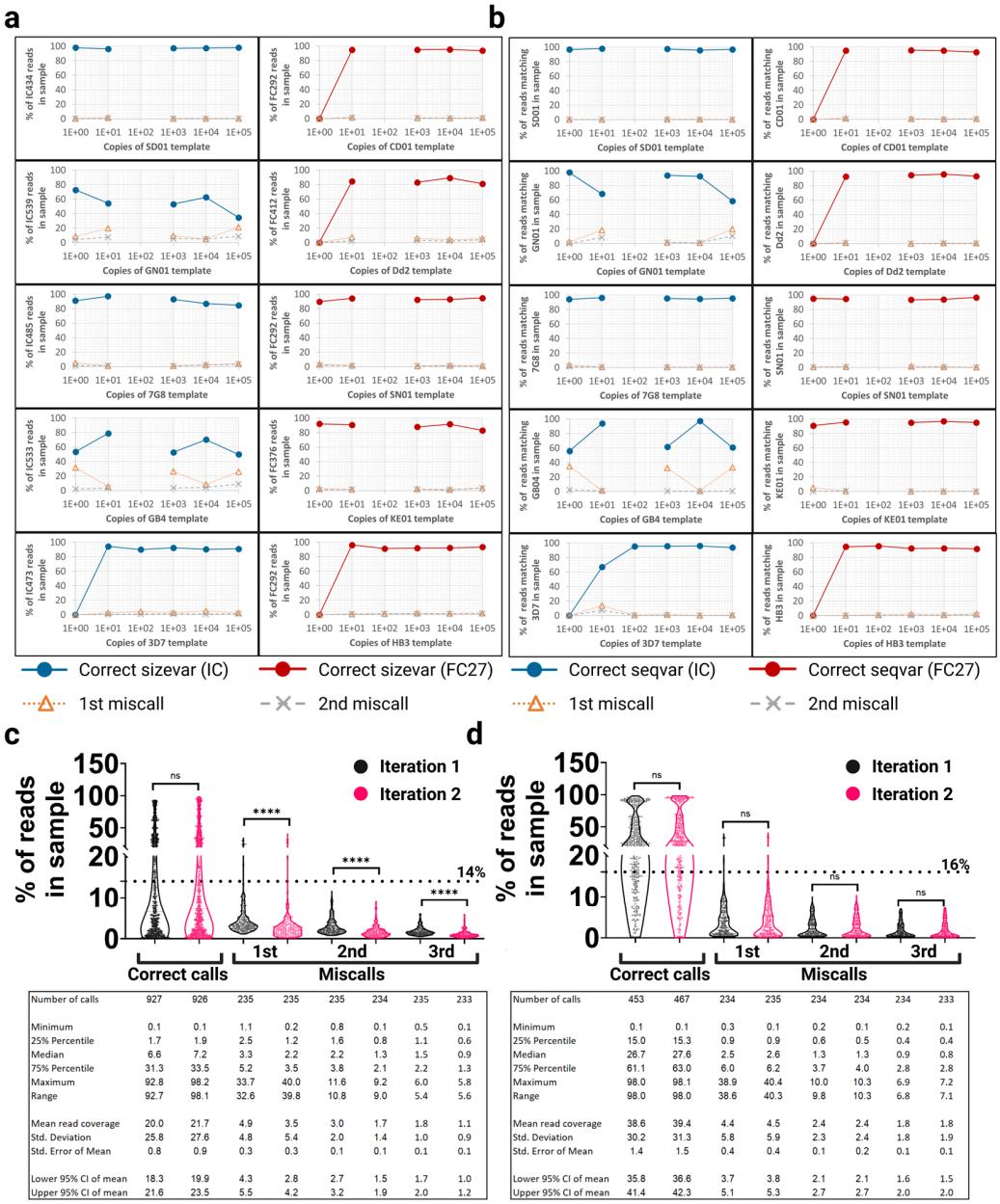
576 **a,** Full-length *msp2* was amplified in a first round of PCR that used primers annealing to the conserved 5'

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

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

579 downstream and upstream from the annealing sites for the forward and reverse oligos, respectively, used
580 in the first PCR reaction. Asymmetric amplicon barcoding allowed for the accurate pooling and
581 computational identification of up to 384 individual samples using a combination of 40 different barcodes.
582 Thereafter, SMRTbell adaptors were coupled to the barcoded amplicons resulting in single-stranded
583 circularized constructs containing both strands of *msp2*. Circularized *msp2* amplicons were sequenced in
584 32 polymerase passes (16 on each strand) and the resulting sequences for each amplicon were aligned to
585 produce circular consensus reads. **b,** Circular consensus sequences (CCSs) were de-multiplexed with Lima
586 and a combined FASTA file was created with all these indexed CCSs reads. Size variant genotyping was
587 done by *in silico* PCR on this FASTA file using the IC1 and FC27 primer sequences designed by Snounou *et*
588 *al*, 1999, in addition to new primers designed for this study as queries in a Blastn search against the
589 dataset. Isolates and size variants were hierarchically clustered and visualized in two sequencing coverage
590 heatmaps for IC1 and FC27. In addition, CCS reads were bioinformatically coupled to the primers from the
591 first PCR to construct isolate phylogenies or be translated into amino acid sequences to allow the
592 prediction of structural elements, the study of sequence conservation, as well as the presence of B and T
593 cell epitopes.
594
595
596
597
598
599
600

601 **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

602

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

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

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

606 3D7 (IC473), CD01 (FC292), Dd2 (FC412), SN01 (FC292), KE01 (FC376) and HB3 (FC292) were prepared and

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

622

623

624

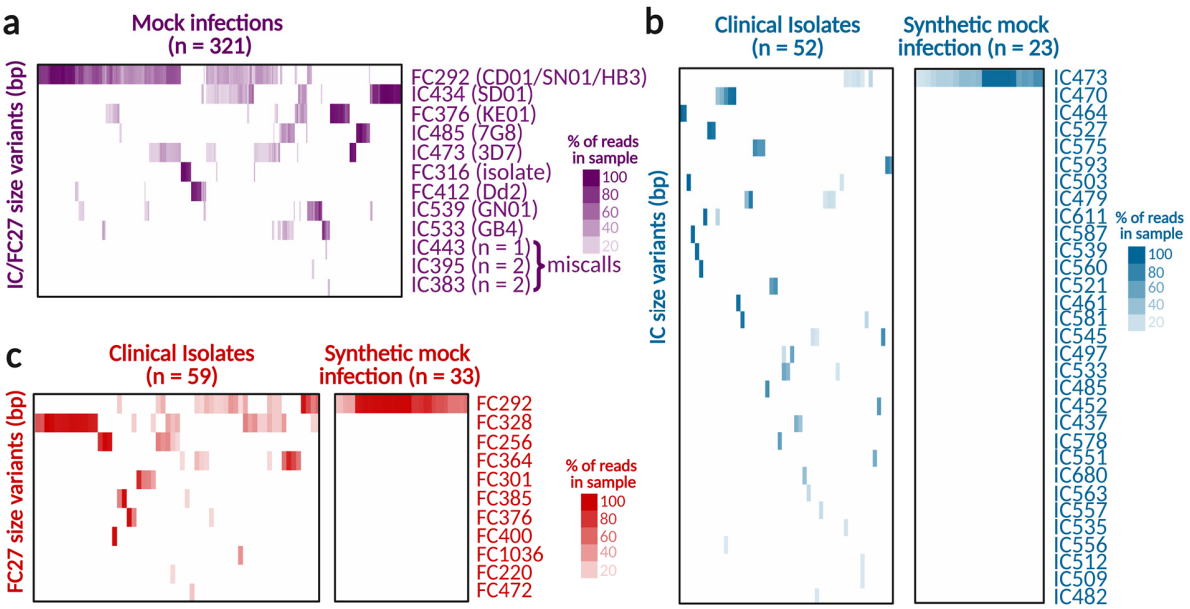
625

626

627

628

629 **Fig. 3**



630

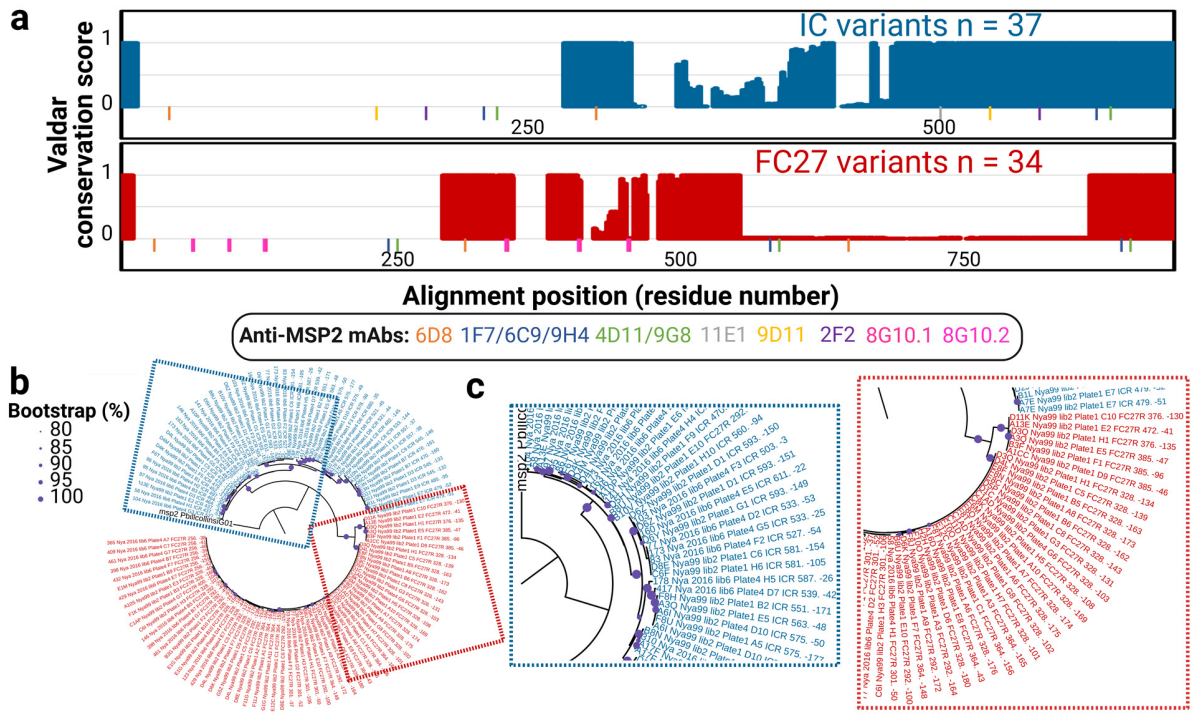
631 **Fig. 3. Size variant genotyping simulated by *in silico* PCR on CCS reads from synthetic mock infections**
632 **and clinical isolates**. **a,** Variant calling for 321 mock infection controls. Mock infections were prepared
633 with 10 synthetic gene variants of *msp2* (in parenthesis) in concentrations of 1 to 100000 copies per
634 reaction. Infections ranged in complexity from 1 to 10 variants per mixture. One clinical isolate (FC316)
635 was spiked-in and detected in 11 samples included in the pool. IC443, IC395 and IC383 are miscalls present
636 in 5 out of 321 mock infections (1.6%). **b,** IC1 and **c,** FC27 size variant genotyping for 95 clinical isolates
637 collected in Nyamisati, Tanzania, in 1999 and 2016. Size variant calling is shown for synthetic mock
638 infections used as positive controls in the CCS libraries where the isolates were sequenced and including
639 a mix of one IC1 (3D7, IC473) and one FC27 (HB3, FC292) variant. Isolates, mock infections and size variants
640 were grouped by complete hierarchical clustering using Euclidian as the distance metric.

641

642

643

644 **Fig. 4**



645

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

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

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

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

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

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

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

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

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

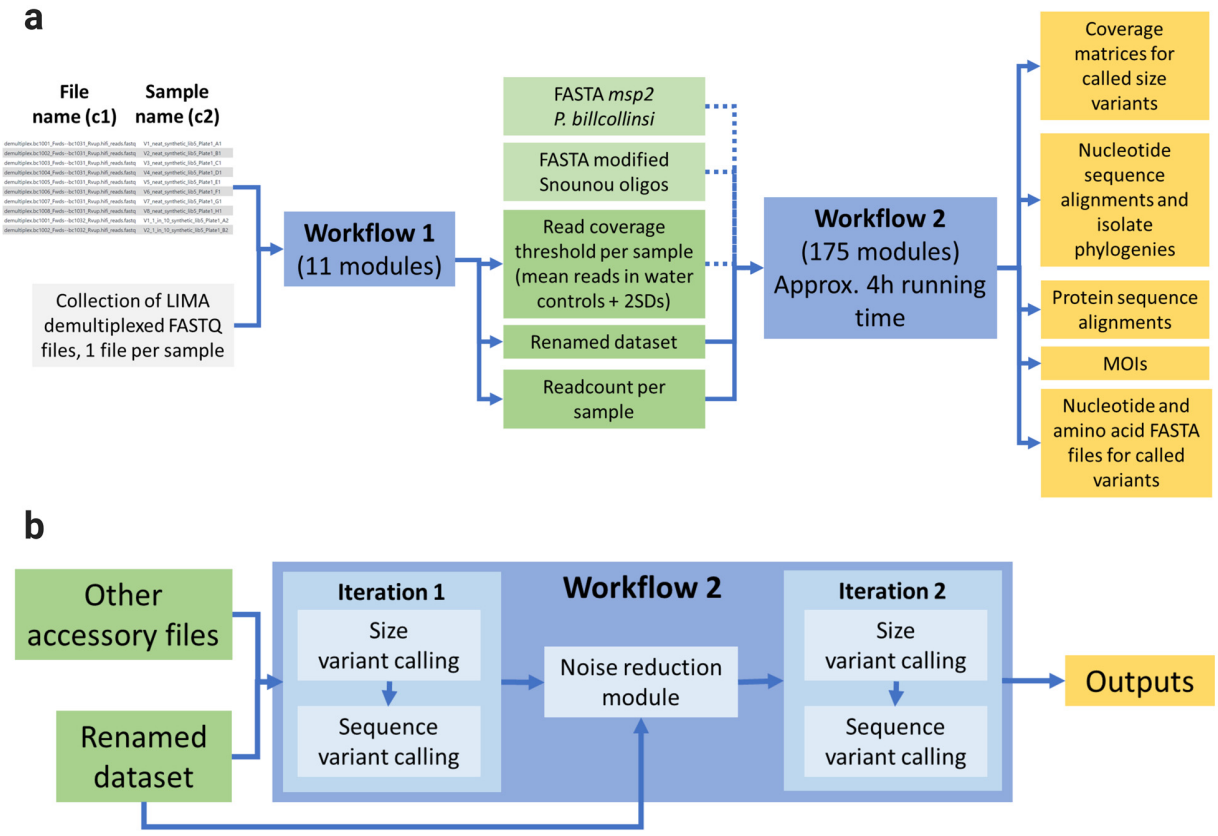
655 (*Laverania* subgenus). **c,** The detailed structure of the trees is shown in the dashed frame panels.

656

657

658 **Fig. 5**

659



660

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

670

671 **Online References**

672 27. Farnert, A. et al. Epidemiology of malaria in a village in the Rufiji River Delta, Tanzania: declining
673 transmission over 25 years revealed by different parasitological metrics. *Malar J* **13**, 459 (2014).
674 28. Yman, V. et al. Persistent transmission of Plasmodium malariae and Plasmodium ovale species in
675 an area of declining Plasmodium falciparum transmission in eastern Tanzania. *PLoS Negl Trop*
676 *Dis* **13**, e0007414 (2019).
677 29. Shokoples, S.E., Ndao, M., Kowalewska-Grochowska, K. & Yanow, S.K. Multiplexed real-time PCR
678 assay for discrimination of Plasmodium species with improved sensitivity for mixed infections. *J*
679 *Clin Microbiol* **47**, 975-980 (2009).
680 30. Snounou, G. et al. High sensitivity of detection of human malaria parasites by the use of nested
681 polymerase chain reaction. *Mol Biochem Parasitol* **61**, 315-320 (1993).
682 31. Toribio, A.L. et al. European Nucleotide Archive in 2016. *Nucleic Acids Res* **45**, D32-D36 (2017).
683 32. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
684 33. Cock, P.J., Chilton, J.M., Gruning, B., Johnson, J.E. & Soranzo, N. NCBI BLAST+ integrated into
685 Galaxy. *Gigascience* **4**, 39 (2015).
686 34. Feng, Z.P. et al. Abundance of intrinsically unstructured proteins in P. falciparum and other
687 apicomplexan parasite proteomes. *Mol Biochem Parasitol* **150**, 256-267 (2006).
688 35. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-
689 wide expression patterns. *Proc Natl Acad Sci U S A* **95**, 14863-14868 (1998).
690 36. Katoh, K. & Standley, D.M. MAFFT multiple sequence alignment software version 7:
691 improvements in performance and usability. *Mol Biol Evol* **30**, 772-780 (2013).
692 37. Otto, T.D. et al. Genomes of all known members of a Plasmodium subgenus reveal paths to
693 virulent human malaria. *Nat Microbiol* **3**, 687-697 (2018).

694 38. Nguyen, L.T., Schmidt, H.A., von Haeseler, A. & Minh, B.Q. IQ-TREE: a fast and effective
695 stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**, 268-274
696 (2015).
697 39. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments.
698 *Nucleic Acids Res* **47**, W256-W259 (2019).
699 40. Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M. & Barton, G.J. Jalview Version 2--a
700 multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189-1191
701 (2009).
702 41. Valdar, W.S. Scoring residue conservation. *Proteins* **48**, 227-241 (2002).
703 42. MacGowan, S.A. et al. The Dundee Resource for Sequence Analysis and Structure Prediction.
704 *Protein Sci* **29**, 277-297 (2020).
705 43. Paul, S. et al. Development and validation of a broad scheme for prediction of HLA class II
706 restricted T cell epitopes. *J Immunol Methods* **422**, 28-34 (2015).
707 44. Grifoni, A. et al. Targets of T Cell Responses to SARS-CoV-2 Coronavirus in Humans with COVID-
708 19 Disease and Unexposed Individuals. *Cell* **181**, 1489-1501 e1415 (2020).
709 45. Jensen, K.K. et al. Improved methods for predicting peptide binding affinity to MHC class II
710 molecules. *Immunology* **154**, 394-406 (2018).
711 46. Nielsen, M., Lundegaard, C. & Lund, O. Prediction of MHC class II binding affinity using SMM-
712 align, a novel stabilization matrix alignment method. *BMC Bioinformatics* **8**, 238 (2007).
713 47. Sturniolo, T. et al. Generation of tissue-specific and promiscuous HLA ligand databases using
714 DNA microarrays and virtual HLA class II matrices. *Nat Biotechnol* **17**, 555-561 (1999).
715 48. Wang, P. et al. A systematic assessment of MHC class II peptide binding predictions and
716 evaluation of a consensus approach. *PLoS Comput Biol* **4**, e1000048 (2008).