

1 **Analytic optimization of *Plasmodium falciparum* marker gene haplotype recovery**
2 **from amplicon deep sequencing of complex mixtures**

3

4 **Authors**

5 Zena Lapp, PhD¹

6 Elizabeth Freedman²

7 Kathie Huang²

8 Christine F Markwalter, PhD¹

9 Andrew A Obala, PhD³

10 Wendy Prudhomme-O'Meara, PhD^{1,2}

11 Steve M Taylor, MD MPH^{1,2}

12

13 ¹Duke Global Health Institute, Duke University, Durham, NC, USA

14 ²Division of Infectious Diseases, School of Medicine, Duke University, Durham, NC, USA

15 ³School of Medicine, College of Health Sciences, Moi University, Eldoret, Kenya

16

17

18 **Short title:** *P. falciparum* marker gene haplotype recovery from AmpSeq of complex mixtures

19 **Abstract**

20 Molecular epidemiologic studies of malaria parasites commonly employ amplicon deep
21 sequencing (AmpSeq) of marker genes derived from dried blood spots (DBS) to answer public
22 health questions related to topics such as transmission and drug resistance. As these methods
23 are increasingly employed to inform direct public health action, it is important to rigorously
24 evaluate the risk of false positive and false negative haplotypes derived from clinically-relevant
25 sample types. We performed a control experiment evaluating haplotype recovery from AmpSeq
26 of 5 marker genes (*ama1*, *csp*, *msp7*, *sera2*, and *trap*) from DBS containing mixtures of DNA
27 from 1 to 10 known *P. falciparum* reference strains across 3 parasite densities in triplicate
28 (n=270 samples). While false positive haplotypes were present across all parasite densities and
29 mixtures, we optimized censoring criteria to remove 83% (148/179) of false positives while
30 removing only 8% (67/859) of true positives. Post-censoring, the median pairwise Jaccard
31 distance between replicates was 0.83. We failed to recover 35% (477/1365) of haplotypes
32 expected to be present in the sample. Haplotypes were more likely to be missed in low-density
33 samples with <1.5 genomes/ μ L (OR: 3.88, CI: 1.82-8.27, vs. high-density samples with \geq 75
34 genomes/ μ L) and in samples with lower read depth (OR per 10,000 reads: 0.61, CI: 0.54-0.69).
35 Furthermore, minority haplotypes within a sample were more likely to be missed than dominant
36 haplotypes (OR per 0.01 increase in proportion: 0.96, CI: 0.96-0.97). Finally, in clinical samples
37 the percent concordance across markers for multiplicity of infection ranged from 40%-80%.
38 Taken together, our observations indicate that, with sufficient read depth, haplotypes can be
39 successfully recovered from DBS while limiting the false positive rate.

40 Introduction

41 Malaria parasite surveillance and molecular epidemiologic studies increasingly employ as a
42 genotyping approach amplicon deep sequencing (AmpSeq) of short polymorphic fragments of
43 parasite DNA. Depending on which segments of the genome are sequenced, this approach
44 returns haplotypes that can be used to estimate complexity of infection (1), investigate
45 transmission between hosts (2,3), evaluate the prevalence and incidence of markers of drug
46 resistance (4–6), and classify recurrent infections following drug treatment as reinfections or
47 recrudescences (7,8). As a result of these diverse use cases of AmpSeq, there is a broad need
48 for practical and empirically-derived approaches to maximize haplotype recovery and mitigate
49 the risks of false genotypes.

50
51 Prior groups have evaluated the accuracy of haplotype recovery from mixtures containing DNA
52 from known *P. falciparum* strains across a range of available tools and parameters, and
53 reported that strains present in low proportions are likely to be missed (2,9) and that false
54 positive haplotypes often have lower read depth (10,11). In a large analysis of complex mixtures
55 of up to five reference strains, recovery of two markers was compared using four haplotype
56 calling tools (11). They found that fewer haplotypes are recovered from samples with less *P.*
57 *falciparum* DNA, that haplotypes with a lower read count were more frequently false positives,
58 and that the four different haplotype calling tools performed similarly. What remains unexplored
59 by these prior reports are investigations of haplotype recovery from samples with three key
60 features of field studies: i) prepared and processed as dried blood spots (DBS), ii) present
61 across a range of densities reflective of infections that are typically observed in field studies,
62 and iii) harboring genomes from a large range of *P. falciparum* strains, which in natural
63 infections can exceed 15 (2).

64

65 We evaluated the accuracy of the recovery of diverse *P. falciparum* haplotypes from DBS
66 harboring simple and complex mixtures of parasite genomes. To do so, we prepared mixtures of
67 up to 10 parasite strains at known proportions and across three parasite density categories, and
68 amplified and sequenced each in triplicate with MiSeq across polymorphic segments of five
69 distinct markers (*ama1*, *csp*, *msp7*, *sera2*, and *trap*). With these reads, we investigated the
70 influence of parasite density, genomic complexity, and haplotype censoring criteria on the
71 removal of false positive haplotypes, the sensitivity and precision of haplotype discovery, inter-
72 replicate variability, and the ability to recover expected haplotypes at each locus.

73 **Materials and Methods**

74 **Mock Infection design**

75 We selected five targets of interest in the *P. falciparum* genome that have been used in prior
76 AmpSeq studies (2,7,12): *ama1*, *csp*, *msp7*, *sera2*, and *trap*. We amplified by PCR using the
77 reference primers for each (**Table 1**) from each of ten reference *P. falciparum* strain genomic
78 DNAs (gDNAs): MRA-102G (3D7), MRA-150G (Dd2), MRA-152G (7g8), MRA-155G (HB3),
79 MRA-159G (K1), MRA-176G (V1/S), MRA-1169G (Tanzania), MRA-915G (FUP UGANDA-Palo
80 Alto), MRA-309G (FCB), and MRA-731G (FCR3/Gambia). The products of each were Sanger
81 sequenced to determine the reference sequence for each strain.

82

83 **Table 1: Marker-specific reference primers**

Marker	Forward primer	Reverse primer
<i>ama1</i>	TCAGGGAAATGTCCAGTATTTG	GGACCATTATTTTCTTGAGCTG
<i>csp</i>	TTAAGGAACAAGAAGGATAATACCA	AAATGACCCAAACCGAAATG
<i>msp7</i>	ATGAACAAGAGATATCAACACA	TTAAATTGTTTCATGGTATTCCTTA
<i>sera2</i>	TACTTTCCCTTGCCCTTGTG	CACTACAGATGAATCTGCTACAGGA
<i>trap</i>	TCCAGCACATGCGAGTAAAG	AAACCCGAAAATAAGCACGA

84

85 For each reference strain and marker, Unipro UGENE v42 (13) was used to map forward and
86 reverse reads from Sanger sequencing to the respective marker gene. The trimming quality
87 threshold and mapping minimum similarity were set to zero. The sequences were manually
88 trimmed and, where discrepancies in base calls were observed between forward and reverse
89 reads, bases were called manually.

90

91 Five mock polygenomic infections and a 3D7-only mock infection were created by making
92 control mixtures that combined 1 ng/μl gDNA stocks of the distinct parasite reference strains in
93 known percentages ranging from 1% to 100% (**Figure 1A**). Each control mixture was serially
94 diluted in uninfected whole blood, and dried blood spots (DBSs) were made for each of the 11
95 dilutions per mixture. DBS were singly punched into individual wells of a deep 96-well plate, and
96 a modified Chelex-100 protocol (3) was used to make gDNA extracts. These were then tested in
97 duplicate with a duplex pfr364/human β-tubulin quantitative PCR (qPCR) assay that estimated
98 parasite densities using a standard curve generated with extracts from control DBS at dilutions
99 of *P. falciparum* 3D7 ranging from 0.1 to 2000 parasites/μL of whole blood (14). Control mixture
100 extracts were assigned to one of three parasite density ranges (low, <1.5 genomes/μl; medium,
101 1.5-75 genomes/μl; and high, ≥75 genomes/μl) and pooled by mixture at each density range for
102 a total 18 pools (6 mixtures x 3 densities) to be used as templates for subsequent PCR
103 amplification.

104

105 **Library preparation and sequencing**

106 Each mixture template was prepared for sequencing according to qPCR Ct-value as described
107 in (3). Then, from each mixture template, we amplified at the target segments of *ama1*, *csp*,
108 *msp7*, *sera2*, and *trap* in individual reactions in triplicate using a nested PCR strategy. Library
109 preparation for sequencing followed described methods (15) with the following exceptions:
110 PCR1 reactions included 300nM of each primer and 7 μl of template gDNA when extract Ct was
111 < 28 (high density), 18 μl when $28 \leq Ct < 34$ (medium density), and 15 μl concentrated extract
112 when $Ct \geq 34$ (low density). PCR 1 cycling conditions were $95C \times 3' \rightarrow (98C \times 20s \rightarrow 62C \times$
113 $15s \rightarrow 72C \times 20s) \times 8 \rightarrow (98C \times 20s \rightarrow 70C \times 15s \rightarrow 72C \times 20s) \times 27 \rightarrow 72C \times 1'$. PCR 2
114 reactions included 2 μl of template when gDNA pool extract Ct was < 28, and 8 μl of template
115 when Ct was ≥ 28. The resulting dual-indexed libraries were then pooled and purified as

116 previously described (15) before sequencing on an Illumina MiSeq (v3 300PE) platform. Raw
117 sequences have been deposited under BioProject PRJNAXXX.

118

119 **Haplotype recovery**

120 We used Snakemake v 7.20.0 (16) to build an integrated pipeline for haplotype recovery,
121 BRAVA (Basic and Rigorous Amplicon Variant Analyzer; [https://github.com/duke-malaria-](https://github.com/duke-malaria-collaboratory/BRAVA)
122 [collaboratory/BRAVA](https://github.com/duke-malaria-collaboratory/BRAVA)) in order to trim, filter, and map reads, and thence call haplotypes. Primers
123 and adapters of amplicon deep sequencing reads for each marker were removed using
124 Cutadapt v4.1 (17). These reads were trimmed using Trimmomatic v0.38 (18); this removed the
125 leading and trailing bases below a Phred quality score of 10, removed all nucleotides from the 3'
126 end after the quality of the read falls below an average Phred quality score of 15 over a sliding
127 window of 4 nucleotides, and dropped reads with fewer than 80 nucleotides. Remaining reads
128 were mapped to the 3D7 reference genome using BBmap v39.01 (19). Reads were then further
129 filtered and trimmed using the R package DADA2 v1.20.0 (20) function filterAndTrim with a
130 maximum number of expected errors (maxEE) equal to 1. Values ranging from 2 to 10 were
131 tested for the truncQ parameter in filterAndTrim, which truncates reads at the first instance of a
132 quality score \leq truncQ. The optimal value was determined to be the value that maximized the
133 number of reads used for haplotype calling (10); the haplotypes that were output when using
134 this value of truncQ were used for all subsequent analyses. Next, the learnErrors function was
135 used to learn error rates, the dada function was used to remove sequencing errors and identify
136 haplotypes, and the removeBimeraDenovo function was used to remove chimeras. All
137 haplotypes returned by DADA2 were included for analysis.

138

139 **Categorization of haplotypes**

140 For each locus in each sample, we categorized each haplotype returned by DADA2 into one of
141 three groups:

- 142 1. Expected haplotype: A haplotype with an identical sequence to that of a template sequence
143 expected to be observed in the sequenced library. These were considered *true positive*
144 *haplotypes*.
- 145 2. Haplotype arising from systematic error: A haplotype with a sequence or read depth that we
146 did not expect to observe in the sequenced library, but which was observed across all three
147 replicates for at least one density bin. These were suspected to be truly present owing to
148 either inadvertent introduction to mixtures during gDNA preparation or the presence of
149 multiple haplotypes in the original source gDNA. Haplotypes arising from systematic error
150 were removed from the analysis prior to screening for optimal thresholds for haplotype
151 censoring, as we suspected that these template strains were truly present in the library that
152 was sequenced and therefore shouldn't be expected to be corrected by applying filtering
153 criteria.
- 154 3. Haplotype arising from random error: A haplotype that we did not expect to observe in the
155 sequenced library, and that was not consistently present across replicates for any mixture-
156 density combination. These were considered *false positive haplotypes*.

157

158 **Identification of optimal thresholds for haplotype censoring**

159 We evaluated the efficacy of four common metrics used to censor haplotypes: i) the depth of
160 reads within a sample supporting a haplotype (read depth), ii) the proportion of reads within a
161 sample supporting a haplotype (read proportion), iii) the ratio of abundances of pairs of
162 haplotypes within a sample with a Hamming distance of one (read ratio), and iv) the length
163 difference of the returned haplotype relative to that of the expected reference strain (length
164 difference). As mentioned above, haplotypes arising from systematic error were removed prior
165 to evaluating these criteria. All reference strain haplotypes for all loci were identical in length to
166 the 3D7 haplotype, except one *msp7* haplotype that was 3 base pairs shorter. Thus, we defined
167 this censoring criterion as follows: the difference in length between the observed haplotype and

168 the 3D7 reference haplotype must be equal to 0, -3, or 3 (i.e. one codon may be inserted or
169 deleted). For the other 3 censoring criteria, we used Youden's J statistic to identify optimal
170 thresholds across all possible thresholds of the criterion and corresponding confidence intervals
171 with the `coords` and `ci.coords` functions from the R package `pROC` v1.18.0 (21). Because the
172 importance of retaining true positive haplotypes vs. removing false positive haplotypes varies
173 depending on the use case, this statistic was computed using three different ways to weight
174 false negative vs. false positive classifications: equal weight to false negatives and false
175 positives, 2x the weight to false negatives, and 2x the weight to false positives. To evaluate
176 censoring criteria, we used the optimal criteria based on false negatives having 2x the cost of
177 false positives.

178

179 **Risk factor analysis for missing haplotypes**

180 We performed a bivariate and multivariate logistic regression to investigate risk factors for
181 haplotype missingness in R using the `glmer` function in `lme4` v1.1.32 (22). Missing haplotypes
182 were defined as those that were not observed in the sample prior to the application of any
183 haplotype censoring criteria. The outcome was the presence or absence of the haplotype in the
184 un-censored haplotypes, and covariates were target, starting proportion of the reference
185 template strain, read depth (per 10,000 reads), parasite density, and expected number of
186 distinct haplotypes present in the sample. A random intercept was included for each mixture-
187 density combination. Low-density mix C samples were excluded from this analysis as they
188 exhibited signatures of contamination from a high-density sample.

189

190 **Clinical sample analysis**

191 Ten *P. falciparum*-positive DBS collected in a field study in Webuye, Kenya that were previously
192 sequenced at the *ama1* and *csp* loci (2) were sequenced at the *msp7*, *sera2*, and *trap* loci.
193 These samples were selected from those that were high-density and had MOIs >1 at both *ama1*

194 and *csp* loci (using previously defined haplotype calls and censoring criteria (2)). Haplotypes for
195 newly sequenced loci were called with the pipeline described above, using the same method as
196 for *ama1* and *csp*. All haplotypes were censored using the identified optimal censoring criteria.

197

198 **Ethical statement**

199 The field study in which the clinical samples were collected was approved by institutional review
200 boards of Moi University (2017/36) and Duke University (Pro00082000). All participants or
201 guardians provided written informed consent, and those over age 8 years provided additional
202 assent.

203

204 **Data analysis and visualization**

205 All data were analyzed and visualized using R v4.2.1 (23) in RStudio v2022.12.0+353 (24) with
206 the following packages: *msa* v1.28.0 (25), *tidyverse* v2.0.0 (26), *readxl* v1.4.2 (27), *ape* v5.7.1
207 (28), *regentrans* v1.0.0 (29), *reshape2* v1.4.4 (30), *scales* v1.2.1 (31), *cowplot* v1.1.1 (32),
208 *ggupset* v0.3.0 (33), *broom.mixed* v0.2.9.4 (34), *ggpmisc* v0.5.2 (35), *ggpubr* v0.6.0 (36), and
209 *ggtext* v0.1.2 (37). Variant positions in the amplified portion of each marker gene were extracted
210 from Pf6k (38) VCF files and tallied. We compared read depths of true and false positive
211 haplotypes, and median multiplicities of infection, using a Wilcoxon test, and number of
212 haplotypes censored by density using a Fisher's exact test. Code and data for this manuscript
213 can be found at https://github.com/duke-malaria-collaboratory/brava_validation.

214 Results

215 Mixtures, reference strains, deep sequencing, and haplotype calling

216 We sequenced five previously-developed AmpSeq marker genes: *ama1*, *csp*, *msp7*, *sera2*, and
217 *trap* (**Table 2**), and generated for sequencing 6 mock infections harboring mixtures of gDNA
218 from between 1 and 10 distinct parasite reference strains (**Figure 1A**) to approximate the
219 polygenomic nature of many infections in high-transmission areas. Not all marker genes were
220 unique to a strain; a total of 37 distinct haplotypes were present across the 10 strains and 5
221 markers. Pairwise single nucleotide variant (SNV) distance varied between strains and markers
222 (median: 4, range: 0-15; **Figure 1B**).

223
224 For each marker gene, each of the 6 mixtures was sequenced from dilution pools corresponding
225 to low (<1.5 genomes/ μ L), medium (1.5-75 genomes/ μ L) and high (\geq 75 genomes/ μ L) parasite
226 density bins in triplicate, tallying to 1365 expected haplotype occurrences across 270
227 sequenced samples. We obtained analyzable reads for 257/270 samples, with differences in the
228 absolute yield of read counts between low (4.3 million), medium (8.0 million), and high (10.2
229 million) density samples. This general observation held for each individual marker, save for *trap*
230 and *msp7* which returned moderate read amounts irrespective of parasite density bin (**Figure**
231 **1C**). Overall, we observed across the five loci and 257 samples 1292 haplotype occurrences
232 (**Figure 1D**), for which the median read depth was 1542.

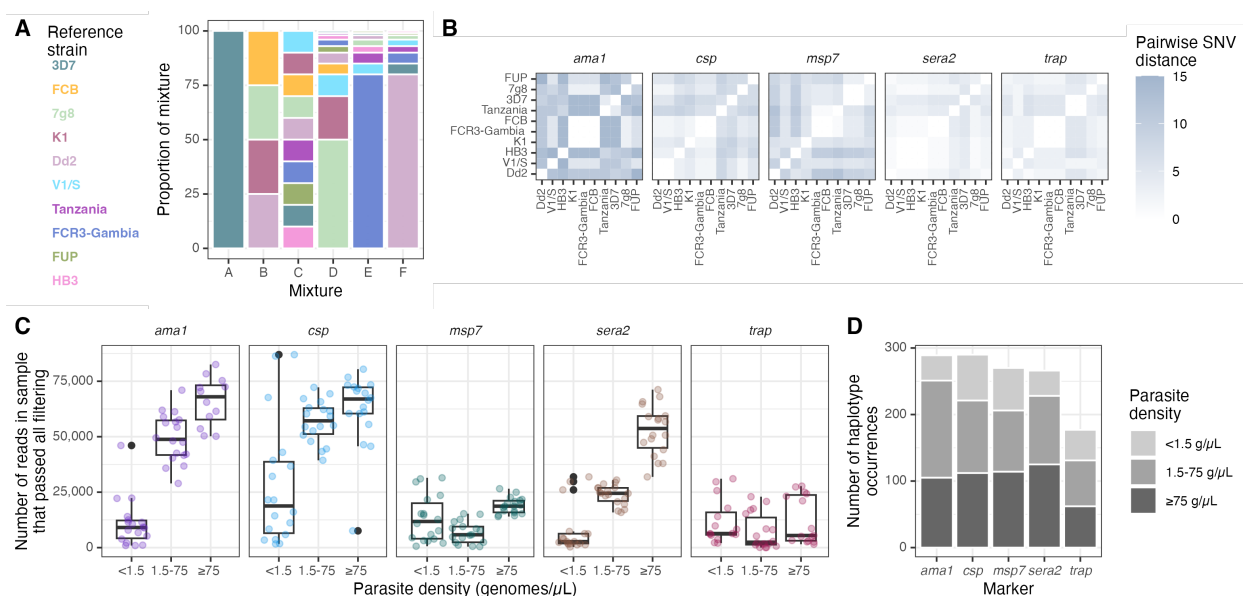
233

234 Table 2: Marker gene characteristics

Target	Stage expressed	3D7 gene ID	Chromosome	3D7 coordinates amplified	Sequence length	3D7 GC content	Number of Pf6k variant positions
<i>ama1</i>	Blood	PF3D7_1133400	11	1294312- 1294613	300	27%	49 (16%)

<i>csp</i>	Liver	PF3D7_0304600	03	221351- 221640	288	29%	53 (18%)
<i>msp7</i>	Blood	PF3D7_1335100	13	1419236- 1419567	330	25%	53 (16%)
<i>sera2</i>	Blood	PF3D7_0207900	02	320762- 321022	259	41%	62 (24%)
<i>trap</i>	Liver	PF3D7_1335900	13	1465058- 1465379	320	31%	46 (14%)

235



236

237 **Figure 1: Overview of mixtures, reference strains, and sequence yield. (A)** Overview of
 238 mixtures A through F, each composed of various proportions of gDNA from the listed *P.*
 239 *falciparum* reference strains (colors). **(B)** Pairwise single nucleotide variant (SNV) distances
 240 between reference haplotypes of each of the marker genes obtained by Sanger sequencing. **(C)**
 241 Number of reads in each sample by parasite density bin, faceted by marker gene. **(D)** Total
 242 number of pre-censored haplotype occurrences for each marker across all mixtures and
 243 replicates, colored by parasite density bin. Note that *ama1* was sequenced separately from the

244 other markers so read depth cannot be directly compared between *ama1* and other markers.

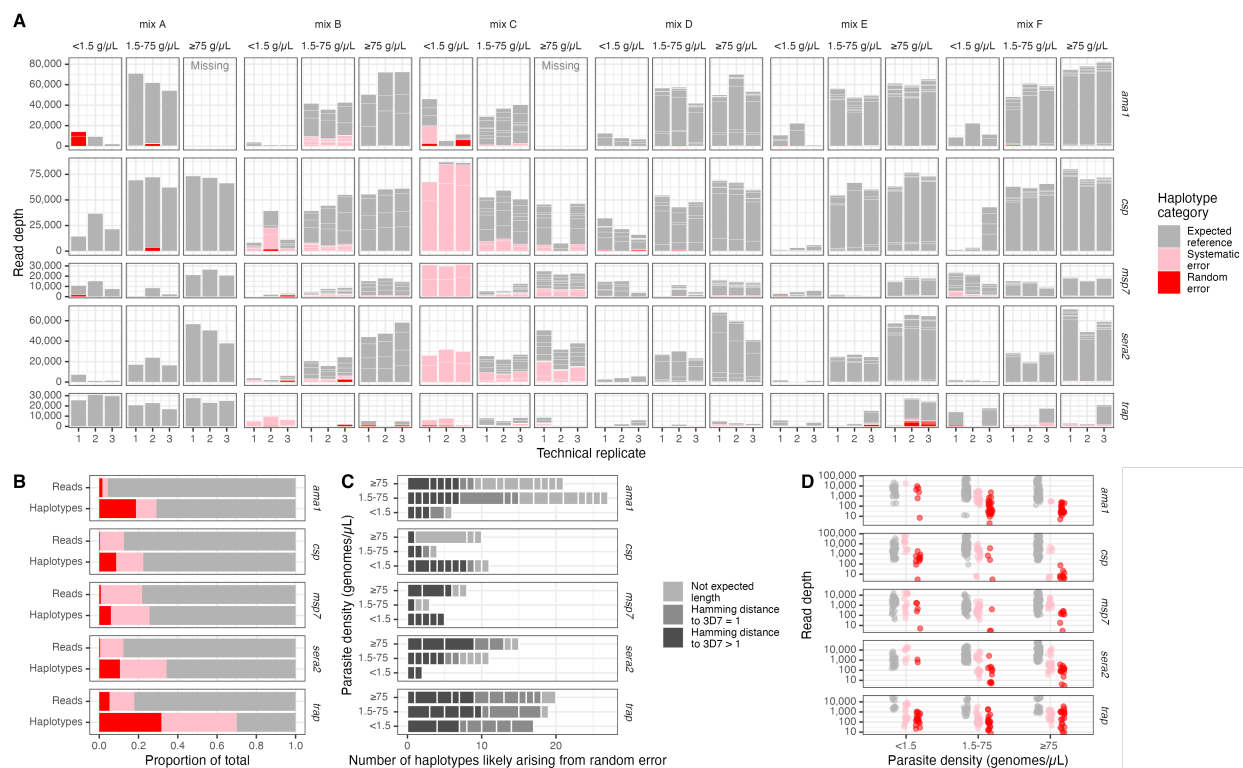
245 g/ μ L = genomes/ μ L.

246

247 **False positive haplotypes**

248 We first investigated false positive haplotype occurrences across samples. Within each sample,
249 we categorized each observed haplotype as expected to be present in the sample (true positive,
250 n=859/1292, 66%), likely cryptically present in the original mixture (systematic error;
251 n=254/1292, 20%), or likely arising from random error (false positive, n=179/1292, 14%) (**Figure**
252 **2A**). Only 1% of reads that passed filtering supported haplotypes that were categorized as false
253 positives. We observed this trend of proportionately few reads supporting proportionately more
254 false positive haplotypes across both markers and parasite density bins (**Figure 2B**).

255 Furthermore, the percentage of false positive haplotypes was relatively similar across parasite
256 density bin (12-16%), although for *ama1* and *sera2*, there were fewer false positive haplotypes
257 for low-density templates (**Figure 2C**). False positive haplotypes were often not the correct
258 sequence length, were often only one nucleotide different from a reference sequence in the
259 sample (**Figure 2C**), and had lower read depths than haplotypes we expected to observe
260 (median=104 vs. 2393, Wilcox p < 0.001; **Figure 2D**).

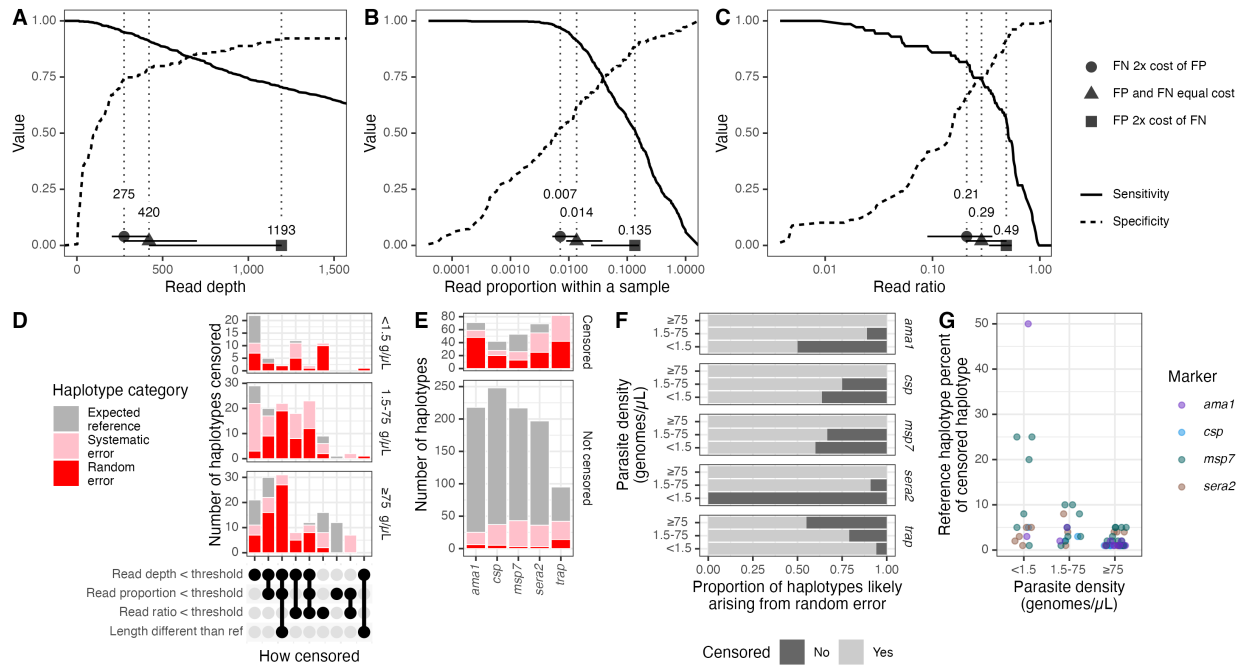


261
 262 **Figure 2: Overview of false positive haplotypes. (A)** Sample-level haplotype overview.
 263 Stacked boxes in each column represent observed haplotypes from reads that passed filtering,
 264 categorized as those expected in the reference (gray), arising from systematic error (pink), or
 265 from random error (red). Box heights indicate the number of reads supporting the haplotype. **(B)**
 266 Proportion of reads and of haplotypes by marker categorized as expected reference, systematic
 267 error, and random error. **(C)** False positive haplotypes by marker categorized by unexpected
 268 length and by SNV distance to the 3D7 reference sequence. Hamming distances were only
 269 computed for haplotypes identical in length to the 3D7 reference sequence. **(D)** Read depth for
 270 expected (gray), systematic error (pink) and random error (red) haplotypes by parasite density
 271 bin and by marker. g/ μ L = genomes/ μ L.

272
 273 **Evaluating haplotype censoring criteria**

274 We next evaluated, in our dataset, the effectiveness of four important threshold criteria typically
 275 applied to remove false positives from AmpSeq data: read depth, read proportion, read ratio of

276 similar haplotypes, and haplotype length. The optimal thresholds had large confidence intervals
277 and varied depending on how much weight was given to false positive vs. false negative
278 haplotypes (**Figure 3A-C**). Prioritizing the inclusion of true positive haplotypes over the removal
279 of false positive haplotypes, optimal thresholds were 275 for read depth (95% CI: [204-420];
280 sensitivity = 0.95 [0.90-0.99]; specificity = 0.52 [0.46-0.68]), 0.007 for read proportion (95% CI:
281 [0.005-0.014]; sensitivity = 0.97 [0.91-0.99]; specificity = 0.54 [0.47-0.69]), and 0.21 for read
282 ratio (95% CI: [0.09-0.36]; sensitivity = 0.82 [0.72-0.93]; specificity = 0.67 [0.44-0.67]). Using
283 these criteria, across all targets 975/1292 (75%) haplotype occurrences remained
284 corresponding to 59/124 (48%) distinct haplotypes, yielding at least one uncensored haplotype
285 in 254/257 (99%) samples. Specifically, these thresholds censored 148/179 (83%) random error
286 haplotypes, 102/254 (40%) systematic error haplotypes, and 67/859 (8%) expected reference
287 haplotypes (**Figure 3D-E**). Of the 179 random error haplotypes, 75% fell under the read
288 threshold, 54% fell under the proportion threshold, 30% fell under the within-sample ratio
289 threshold, and 28% had a length different than the reference strains. Furthermore, for all
290 markers but *trap*, fewer false positive haplotypes were successfully censored in lower parasite
291 density bins (Fisher's exact $p < 0.01$, **Figure 3F**), yielding more false positives post-censoring in
292 low- (11) compared to medium- (6) and high-density (0) parasite bins. Of the censored true
293 positive haplotypes, over half (39/67; 58%) were from high-density templates, and only 5/67
294 (7%) made up $\geq 10\%$ of the original mixture (**Figure 3G**).



295

296 **Figure 3: Optimization and application of censoring criteria. (A-C)** Sensitivity and specificity

297 across ranges of tested thresholds for haplotype **(A)** read depth, **(B)** read proportion, and **(C)**

298 ratio within a sample between haplotypes with a Hamming distance of 1. **(D)** Count of censored

299 haplotypes by the criterion by which they were censored and by density of parasites in DBS

300 sample. The majority of censored haplotypes were non-reference haplotypes and fell under the

301 identified read depth threshold. **(E)** Numbers of censored and uncensored haplotypes by

302 haplotype category and by marker. **(F)** Proportion of uncensored (light grey) and censored (dark

303 grey) haplotypes likely arising from random error, by parasite density bin and marker. **(G)**

304 Reference haplotype percent of censored haplotypes. No reference haplotypes were censored

305 out for *trap*. g/μL = genomes/μL. FN = false negative; FP = false positive. g/μL = genomes/μL.

306

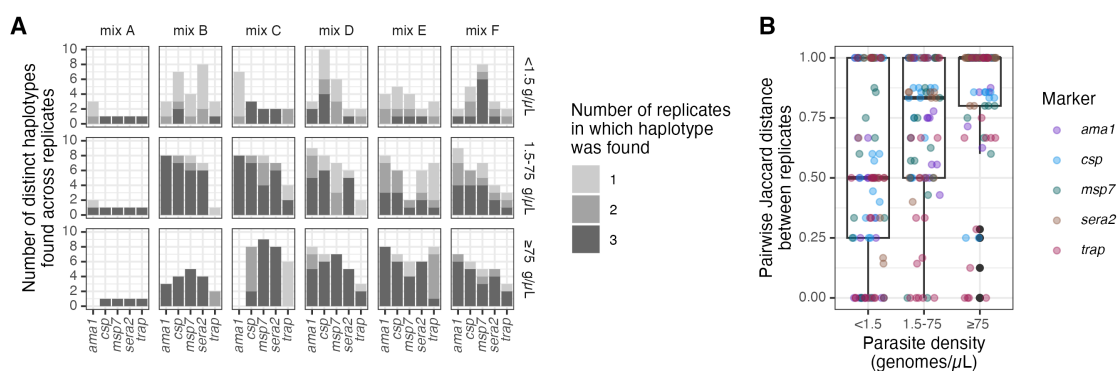
307 Inter-replicate variability

308 To evaluate the consistency with which haplotypes were returned, we measured inter-replicate

309 variability post-censoring. Overall, 58% of haplotypes were observed in all 3 replicates, 18% in 2

310 replicates, and 24% in 1 replicate. Haplotypes were more consistently returned in all three

311 replicates for high-density samples (76% of the time) compared to medium- (61% of the time)
 312 and low-density samples (30% of the time) (**Figure 4A**). Consistent with this, in high-density
 313 samples Jaccard distances between replicates were higher (median = 1, IQR = 0.2) compared
 314 to medium- (median = 0.83, IQR = 0.5) and low-density samples (median = 0.5, IQR = 0.75)
 315 (**Figure 4B**).
 316

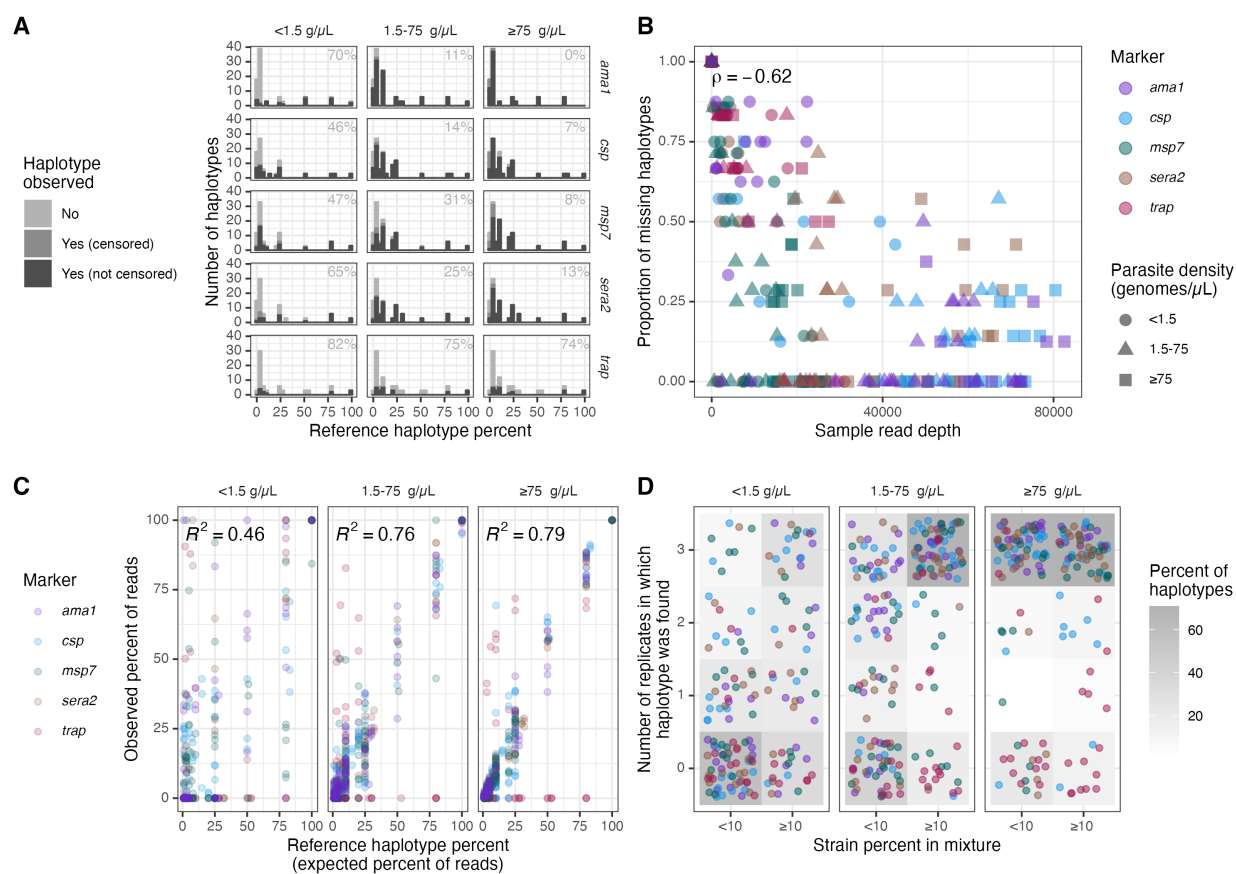


317
 318 **Figure 4: Inter-replicate variability. (A)** Number of replicates in which each haplotype was
 319 found (color) by mix, parasite density bin, and target. **(B)** Pairwise Jaccard distance between
 320 replicates by parasite density bin, colored by marker. g/μL = genomes/μL.

321
 322 **Missing haplotypes**

323 Of the 1365 haplotype occurrences expected to be present across all samples, we did not
 324 recover 477 (35%). Thus, we next investigated factors associated with missing haplotypes. As
 325 expected, haplotype proportion within a sample was inversely associated with missingness, with
 326 each increase of 0.01 in proportion associated with a 4% reduction in the likelihood of being
 327 missed (OR: 0.96, 95% CI: 0.96-0.97), even when controlling for marker, density bin, number of
 328 reads in the sample, and expected number of haplotypes (**Figure 5A; Table 3**). Additionally, for
 329 all markers except *trap*, <15% of haplotypes were missed from high-density samples, while
 330 >45% were missed from low-density samples (**Fig 5A**). Overall read depth for a sample was
 331 negatively correlated with the proportion of haplotypes that were missing from the sample

332 (Spearman's rho = -0.62; **Figure 5B**). Furthermore, within a sample, observed and expected
 333 read proportions were correlated, although there was high stochasticity, particularly for the low-
 334 density samples (**Figure 5C**). Finally, in high-density samples only 30/166 (18%) haplotypes
 335 were not recovered in any replicates, while in low-density samples 78/158 (49%) were not
 336 recovered in any replicates (**Figure 5D**).
 337



338
 339 **Figure 5: Summary of missing haplotypes. (A)** Numbers of missing haplotypes (light grey),
 340 observed but censored haplotypes (medium grey), and observed haplotypes (dark grey) in
 341 individual samples by marker and parasite density bin. The number in each facet indicates the
 342 percentage of missing haplotypes. All subsequent panels in this figure consider observed but
 343 censored haplotypes as missing. **(B)** Correlation between the overall read depth of a sample
 344 and proportion of all expected haplotypes within a mixture that were not successfully recovered.

345 Color indicates marker, and shape indicates parasite density. Spearman's rho = -0.62. **(C)**
 346 Correlation between proportions of expected and observed haplotypes within individual samples
 347 by parasite density bin, colored by marker. **(D)** Number of replicates in which the haplotype was
 348 found by binned strain percent in the original mixture (present at <10% or ≥10%). Each point is
 349 a haplotype colored by marker. The grey color beneath the points indicates the percent of
 350 haplotypes across all targets and mixtures in a given strain percent bin that were observed in
 351 the corresponding number of replicates. Low-density mix C samples were excluded from this
 352 figure as they exhibited signatures of contamination from a high-density sample. g/μL =
 353 genomes/μL.

354

355 **Table 3: Risk factors for haplotype missingness**

Feature	Term	Bivariate Odds Ratio (95% CI)	Multivariate Odds Ratio (95% CI)*
Haplotype proportion (per 0.01 increase)		0.98 (0.97-0.98); p=4.3e-11	0.96 (0.96-0.97); p=6e-17
Target	<i>ama1</i>	REF	REF
	<i>csp</i>	0.76 (0.49-1.18); p=0.23	0.90 (0.54-1.51); p=0.7
	<i>msh7</i>	1.24 (0.81-1.89); p=0.32	0.40 (0.23-0.68); p=8e-04
	<i>sera2</i>	1.68 (1.1-2.57); p=0.016	1.05 (0.63-1.77); p=0.8
	<i>trap</i>	21.37 (13.02-35.08); p=1e-33	6.13 (3.13-12.03); p=1e-07
Density, genomes/μL	≥75	REF	REF
	1.5-75	1.62 (0.76-3.45); p=0.21	1.47 (0.75-2.88); p=0.3
	<1.5	6.27 (2.87-13.69); p=3.9e-06	3.88 (1.82-8.27); p=5e-04
Read depth (per 10,000 reads)		0.57 (0.53-0.62); p=5.7e-40	0.61 (0.54-0.69); p=3e-15

Feature	Term	Bivariate	Multivariate
		Odds Ratio (95% CI)	Odds Ratio (95% CI)*
Expected number of haplotypes		0.32 (0.24-0.44); p=1.2e-12	1.08 (0.91-1.27); p=0.4

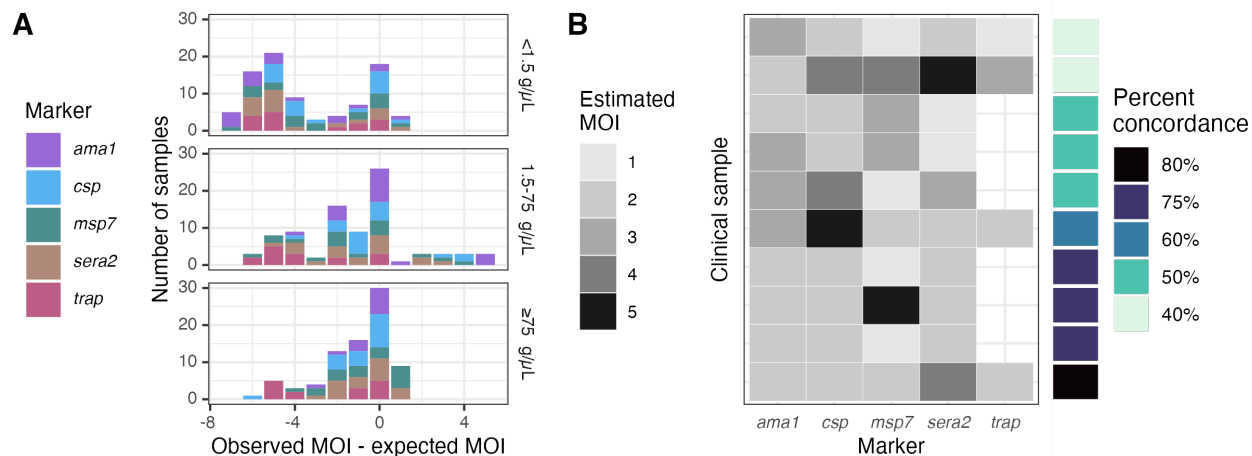
356 * Covariates included were haplotype proportion, target, parasite density, read depth, and
357 expected number of haplotypes

358

359 **Estimating multiplicity of infection based on marker haplotype diversity**

360 We next compared the expected multiplicity of infection (MOI) to the observed MOI after
361 censoring, with MOI expressed as the number of haplotypes observed at each individual
362 marker. Relative to the expected MOIs, the observed MOIs were equal 29% (74/254) of the
363 time, lower 61% (154/254) of the time, and higher only 10% (26/254) of the time. MOIs were
364 more likely to be underestimated in low-density samples (median observed-expected MOI = -4
365 for low-density samples vs. -1 for medium- and high- samples, Wilcox $p < 0.001$; **Figure 6A**).
366 We performed a similar comparison using 10 high-density *P. falciparum* infections collected as
367 DBS through a recent field study in Western Kenya, in order to capture a broader naturally-
368 occurring diversity of marker haplotypes (2). Using the optimal censoring criteria defined above,
369 we observed 142 haplotypes across all samples and markers, of which 36 (25%) were
370 censored. The range of MOIs was 1-5 for each marker. No marker consistently estimated the
371 highest or lowest MOI, and the percent concordance ranged from 40% to 80% (**Figure 6B**).

372



373

374 **Figure 6: Estimated multiplicities of infection (MOIs) based on each marker haplotype. (A)**

375 Observed minus expected MOIs for mixtures post-censoring. **(B)** Estimated MOIs in clinical

376 samples. Haplotypes were censored according to the optimal criteria identified above, giving

377 false negatives 2x the cost of false positives. Not all *trap* clinical samples returned sequences.

378 Percent concordance was computed for each sample as the percentage of markers for which

379 the estimated MOI was equal to the mode MOI.

380 Discussion

381 AmpSeq is an increasingly popular tool for molecular epidemiologic studies of *P. falciparum*
382 collected on DBS, which necessitates rigorous haplotype recovery from field samples. We
383 prepared DBS containing mixtures of gDNA from reference *P. falciparum* strains, amplified and
384 sequenced polymorphic segments of 5 common marker genes in triplicate, and quantified the
385 performance of haplotype recovery using a range of metrics. We observed that high sample
386 read depth was associated with enhanced recovery of most haplotypes present in the original
387 sample, and that censoring criteria based on read depth, read proportion, read ratio, and
388 haplotype length can effectively remove most false positive haplotypes while retaining most true
389 positive haplotypes. Thus, for use-cases which involve high-density samples or samples
390 sequenced at high read depth, rigorous recovery can be achieved for multiple markers.

391
392 Consistent with prior studies (2,9,11), we observed that the likelihood of haplotype recovery is
393 enhanced by higher parasite density and by a larger proportion of an individual haplotype within
394 a mixture. In particular, the consistency with which we observed haplotypes across replicates
395 was higher in high-density samples compared to low-density samples. However, we further
396 observed that, independent of parasite density and reference haplotype proportion, successful
397 haplotype recovery was further associated with a higher overall sample read depth. The ability
398 to recover haplotypes constituting a minority population within a parasitemia with an overall low
399 density is an important goal for many use cases of AmpSeq. Namely, therapeutic efficacy
400 studies of antimalarials use active case detection to screen for recurrence of parasites, and
401 frequently capture low-density infections with multiple strains which must then be compared to
402 those in the initial infection in order to distinguish reinfection from recrudescence infection (7).
403 Additionally, studies of transmission networks in highly endemic settings in which low-density,
404 asymptomatic infections predominate also benefit from comprehensive profiling of strains within
405 mixtures in order to ascertain parasite relatedness between hosts (2). In these and similar use

406 cases, the likelihood of detecting minority haplotypes can be improved by maximizing per-
407 sample read depth, such as by limiting multiplexing and selecting maximal sequencing platform
408 output.

409
410 We observed very different optimal censoring thresholds depending on how we weighted the
411 relative importance of false positive and false negative haplotypes, which highlights the need to
412 select censoring criteria suitable for the primary study objective. Penalizing false negative
413 haplotypes more than false positive haplotypes yielded haplotype censoring criteria that still
414 managed to remove most false positive haplotypes while retaining high sensitivity. Furthermore,
415 these criteria were consistent with thresholds that others have used and reported in the
416 literature (read depth: 204-420, read proportion: 0.005-0.014, read ratio: 0.09-0.36) (2,11).

417
418 We observed inconsistency in performance between markers with respect to false positives,
419 censoring, missingness, and MOI. Pre-censoring, false-positive haplotypes were rarely
420 recovered for *msh7* but common for *ama1* and *trap*. However, post-censoring the number of
421 false positives was relatively low for all markers but *trap*. Fewer haplotypes were recovered for
422 *sera2* and *trap* overall. Furthermore, there was no consistent trend across a limited set of
423 clinical samples of marker-specific MOI, suggesting that MOI estimates based upon a single
424 marker may frequently underestimate the true MOI of a sample, as previously described (11).
425 Since most markers returned largely correct haplotype calls across a range of mixtures and
426 parasite density bins, choice of marker may depend not only on marker performance but also
427 other factors such as the biological question of interest (e.g. transmission, vaccine development,
428 etc.).

429
430 Despite controlled laboratory conditions, we observed signatures of both systematic and
431 random error. Systematic error may have resulted from two different sources. First, it is possible

432 that multiple haplotypes were present in the original template strains and were missed during
433 Sanger sequencing. Second, systematic error could arise from contamination during gDNA
434 extraction. Owing to the high-throughput manner of DBS processing, using 96-well plates, it is
435 unfortunate but expected that we observe contamination in a small minority of samples included
436 in a sequencing run. This highlights the importance of meticulous laboratory work and thoughtful
437 controls, particularly because these haplotypes are less likely to be removed by censoring
438 criteria owing to their presence in the original template. In contrast, random error may arise due
439 to PCR stochasticity and polymerase error in low-input next-generation sequencing libraries
440 (39). This is also inevitable, and the censoring criteria described here successfully removed
441 many haplotypes arising from these technical errors.

442
443 Our study had several limitations. First, we created the mixtures from gDNA rather than from
444 intracellular DNA; therefore, the composition of the solution from which DNA was amplified was
445 slightly less complex than that from clinical samples. However, as we extracted DNA from DBS,
446 our results provide a closer approximation to clinical samples than previous studies. Second, we
447 did not attempt to censor haplotypes arising from systematic error because the commonly used
448 censoring criteria assessed here assume that false positive haplotypes arise from random
449 rather than systematic error. Third, this study focused on in silico recovery of haplotypes, and
450 replicates were drawn from the same gDNA extract pools. Thus, variability occurring due to
451 extraction is not accounted for in these data. However, our results provide useful insight into
452 variation and random errors occurring at the amplification and sequencing steps.

453

454

455 **Conclusions**

456 We observed that *P. falciparum* haplotypes from multiple different targets can be successfully
457 recovered from DBS, that in the majority of cases these haplotypes are recovered across

458 replicates, and that censoring criteria already used by the community remove most false

459 positive haplotypes while retaining high sensitivity.

460 **Funding**

461 This work was supported by the National Institute of Allergy and Infectious Diseases
462 (R01AI146849 to W.P.-O. and S.M.T. and K01AI175527 to C.F.M.)

463

464 **Acknowledgements**

465 We thank Jenna DeCurzio for helping with the preparation of samples for sequencing, as well
466 as Laura-Leigh Rowlette and Fangfei Ye at the Duke University Sequencing & Genomic
467 Technologies Shared Resource for performing sequencing and preliminary processing of
468 sequenced reads. *P. falciparum* strains 3D7 (MRA-102, contributed by Daniel J. Carucci), FUP
469 UGANDA-PALO ALTO (MRA-915, contributed by T. Sam-Yellowe), Dd2 (MRA-150G,
470 contributed by David Walliker), 7G8 (MRA-152G, contributed by David Walliker), HB3 (MRA-
471 155G, contributed by Tom Wellems), K1 (MRA-159G, contributed by Dennis Kyle), V1/S (MRA-
472 176G, contributed by Dennis Kyle), Tanzania (MRA-1169G, contributed by Michal Fried), FCB
473 (MRA-309G, contributed by Tom Wellems), and FCR3/Gambia (MRA-731G, contributed by
474 William Trager) were obtained from BEI Resources, NIAID, NIH.

475

476 **Author contributions**

477 ZL, EF, CFM, WPO, and SMT conceptualized the project. ZL, EF, and CFM developed
478 methodology. EF performed the investigation. ZL and KH developed software. ZL performed
479 data curation, formal analysis, and visualization. AAO, WPO, and SMT provided resources,
480 supervision, project administration, and funding acquisition. ZL, EF, and SMT wrote the original
481 draft. All authors reviewed and edited the manuscript.

482 **References**

- 483 1. Miller RH, Hathaway NJ, Kharabora O, Mwandagaliwa K, Tshetu A, Meshnick SR, et al. A
484 deep sequencing approach to estimate Plasmodium falciparum complexity of infection (COI)
485 and explore apical membrane antigen 1 diversity. *Malaria Journal*. 2017 Dec 16;16(1):490.
- 486 2. Sumner KM, Freedman E, Abel L, Obala A, Pence BW, Wesolowski A, et al. Genotyping
487 cognate Plasmodium falciparum in humans and mosquitoes to estimate onward transmission
488 of asymptomatic infections. *Nat Commun*. 2021 Feb 10;12(1):909.
- 489 3. Markwalter CF, Menya D, Wesolowski A, Esimit D, Lokoel G, Kipkoech J, et al. Plasmodium
490 falciparum importation does not sustain malaria transmission in a semi-arid region of Kenya.
491 *PLOS Global Public Health*. 2022 Aug 10;2(8):e0000807.
- 492 4. Niba PTN, Nji AM, Chedjou JPK, Hansson H, Hocke EF, Ali IM, et al. Evolution of
493 Plasmodium falciparum antimalarial drug resistance markers post-adoption of artemisinin-
494 based combination therapies in Yaounde, Cameroon. *International Journal of Infectious*
495 *Diseases*. 2023 Jul 1;132:108–17.
- 496 5. Osoi V, Akinyi M, Wamae K, Kimenyi KM, de Laurent Z, Ndwiga L, et al. Targeted Amplicon
497 Deep Sequencing for Monitoring Antimalarial Resistance Markers in Western Kenya.
498 *Antimicrobial Agents and Chemotherapy*. 2022 Mar 10;66(4):e01945-21.
- 499 6. Olukosi AY, Ajibaye O, Omoniwa O, Oresanya O, Oluwagbemiga AO, Ujuju C, et al. Baseline
500 prevalence of molecular marker of sulfadoxine/pyrimethamine resistance in Ebonyi and Osun
501 states, Nigeria: amplicon deep sequencing of dhps-540. *Journal of Antimicrobial*
502 *Chemotherapy*. 2023 Mar 2;78(3):788–91.

- 503 7. Gruenberg M, Lerch A, Beck HP, Felger I. Amplicon deep sequencing improves Plasmodium
504 falciparum genotyping in clinical trials of antimalarial drugs. *Sci Rep*. 2019 Nov
505 28;9(1):17790.
- 506 8. Castañeda-Mogollón D, Toppings NB, Kamaliddin C, Lang R, Kuhn S, Pillai DR. Amplicon
507 Deep Sequencing Reveals Multiple Genetic Events Lead to Treatment Failure with
508 Atovaquone-Proguanil in Plasmodium falciparum. *Antimicrobial Agents and Chemotherapy*.
509 2023 May 8;67(6):e01709-22.
- 510 9. Lerch A, Koepfli C, Hofmann NE, Messerli C, Wilcox S, Kattenberg JH, et al. Development of
511 amplicon deep sequencing markers and data analysis pipeline for genotyping multi-clonal
512 malaria infections. *BMC Genomics*. 2017 Nov 13;18(1):864.
- 513 10. Hathaway NJ, Parobek CM, Juliano JJ, Bailey JA. SeekDeep: single-base resolution de
514 novo clustering for amplicon deep sequencing. *Nucleic Acids Research*. 2018 Feb
515 28;46(4):e21.
- 516 11. Early AM, Daniels RF, Farrell TM, Grimsby J, Volkman SK, Wirth DF, et al. Detection of
517 low-density Plasmodium falciparum infections using amplicon deep sequencing. *Malaria*
518 *Journal*. 2019 Jul 1;18(1):219.
- 519 12. LaVerriere E, Schwabl P, Carrasquilla M, Taylor AR, Johnson ZM, Shieh M, et al.
520 Design and implementation of multiplexed amplicon sequencing panels to serve genomic
521 epidemiology of infectious disease: a malaria case study [Internet]. *Infectious Diseases*
522 (except HIV/AIDS); 2021 Sep [cited 2021 Sep 27]. Available from:
523 <http://medrxiv.org/lookup/doi/10.1101/2021.09.15.21263521>
- 524 13. Okonechnikov K, Golosova O, Fursov M, the UGENE team. Unipro UGENE: a unified
525 bioinformatics toolkit. *Bioinformatics*. 2012 Apr 15;28(8):1166–7.

- 526 14. Taylor SM, Sumner KM, Freedman B, Mangeni JN, Obala AA, Prudhomme O'Meara W.
527 Direct Estimation of Sensitivity of Plasmodium falciparum Rapid Diagnostic Test for Active
528 Case Detection in a High-Transmission Community Setting. *Am J Trop Med Hyg.* 2019
529 Dec;101(6):1416–23.
- 530 15. Nelson CS, Sumner KM, Freedman E, Saelens JW, Obala AA, Mangeni JN, et al. High-
531 resolution micro-epidemiology of parasite spatial and temporal dynamics in a high malaria
532 transmission setting in Kenya. *Nat Commun.* 2019 Dec 9;10(1):5615.
- 533 16. Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine.
534 *Bioinformatics.* 2012 Oct 1;28(19):2520–2.
- 535 17. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing
536 reads. *EMBnet.journal.* 2011 May 2;17(1):10–2.
- 537 18. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence
538 data. *Bioinformatics.* 2014 Aug 1;30(15):2114–20.
- 539 19. Bushnell B. BMap: A Fast, Accurate, Splice-Aware Aligner [Internet]. Lawrence
540 Berkeley National Lab. (LBNL), Berkeley, CA (United States); 2014 Mar [cited 2023 Jul 14].
541 Report No.: LBNL-7065E. Available from: <https://www.osti.gov/biblio/1241166>
- 542 20. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2:
543 High-resolution sample inference from Illumina amplicon data. *Nat Methods.* 2016
544 Jul;13(7):581–3.
- 545 21. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: an open-
546 source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics.*
547 2011 Mar 17;12(1):77.

- 548 22. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using
549 lme4. *Journal of Statistical Software*. 2015 Oct 7;67:1–48.
- 550 23. R Core Team. R: A Language and Environment for Statistical Computing [Internet].
551 Vienna, Austria: R Foundation for Statistical Computing; 2021. Available from: [https://www.R-](https://www.R-project.org/)
552 [project.org/](https://www.R-project.org/)
- 553 24. RStudio | Open source & professional software for data science teams [Internet]. [cited
554 2022 Apr 1]. Available from: <https://www.rstudio.com/>
- 555 25. Bodenhofer U, Bonatesta E, Horejš-Kainrath C, Hochreiter S. msa: an R package for
556 multiple sequence alignment. *Bioinformatics*. 2015 Dec 15;31(24):3997–9.
- 557 26. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to
558 the Tidyverse. *Journal of Open Source Software*. 2019 Nov 21;4(43):1686.
- 559 27. Wickham H, Bryan J, attribution) Rs (Copyright holder of all R code and all C code
560 without explicit copyright, code) MK (Author of included R, code) KV (Author of included
561 libxls, code) CL (Author of included libxls, et al. readxl: Read Excel Files [Internet]. 2019
562 [cited 2022 Feb 16]. Available from: <https://CRAN.R-project.org/package=readxl>
- 563 28. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and
564 evolutionary analyses in R. *Bioinformatics*. 2019 Feb 1;35(3):526–8.
- 565 29. Hoffman S, Lapp Z, Wang J, Snitkin ESY 2022. regentrans: a framework and R package
566 for using genomics to study regional pathogen transmission. *Microbial Genomics*.
567 8(1):000747.
- 568 30. Wickham H. Reshaping Data with the reshape Package. *Journal of Statistical Software*.
569 2007 Nov 13;21(1):1–20.

- 570 31. Wickham H, Seidel D, RStudio. scales: Scale Functions for Visualization [Internet]. 2020
571 [cited 2022 Apr 1]. Available from: <https://CRAN.R-project.org/package=scales>
- 572 32. Wilke CO. cowplot: Streamlined Plot Theme and Plot Annotations for “ggplot2” [Internet].
573 2019 [cited 2020 Apr 15]. Available from: <https://CRAN.R-project.org/package=cowplot>
- 574 33. Ahlmann-Eltze C. ggupset: Combination Matrix Axis for “ggplot2” to Create “UpSet” Plots
575 [Internet]. 2020 [cited 2023 Jul 14]. Available from:
576 <https://cran.rstudio.com/web/packages/ggupset/index.html>
- 577 34. Bolker B, Robinson D, Menne D, Gabry J, Buerkner P, Hua C, et al. broom.mixed:
578 Tidying Methods for Mixed Models [Internet]. 2019 [cited 2023 Jul 14]. Available from:
579 <https://CRAN.R-project.org/package=broom.mixed>
- 580 35. Aphalo PJ, Slowikowski K. ggpmisc: Miscellaneous Extensions to “ggplot2” [Internet].
581 2018 [cited 2023 Jul 14]. Available from: <https://CRAN.R-project.org/package=ggpmisc>
- 582 36. Kassambara A. ggpubr: “ggplot2” Based Publication Ready Plots [Internet]. 2018 [cited
583 2023 Jul 14]. Available from: <https://CRAN.R-project.org/package=ggpubr>
- 584 37. Wilke CO. ggtext: Improved Text Rendering Support for “ggplot2” [Internet]. 2020 [cited
585 2022 Apr 1]. Available from: <https://CRAN.R-project.org/package=ggtext>
- 586 38. An open dataset of Plasmodium falciparum ... | Wellcome Open Research [Internet].
587 [cited 2023 Jul 14]. Available from: <https://wellcomeopenresearch.org/articles/6-42>
- 588 39. Keschull JM, Zador AM. Sources of PCR-induced distortions in high-throughput
589 sequencing data sets. *Nucleic Acids Res.* 2015 Dec 2;43(21):e143.

590