Neutral vs. non-neutral genetic footprints of *Plasmodium falciparum* multiclonal infections

Frédéric Labbé[1], Qixin He[2], Qi Zhan[1], Kathryn E. Tiedje[3,4], Dionne C. Argyropoulos[3,4], Mun Hua Tan[3,4], Anita Ghansah[5], Karen P. Day[3,4], Mercedes Pascual[1,6*]

1. Department of Ecology and Evolution, The University of Chicago, 1101 E 57th Street, Chicago, IL 60637, United States
2. Department of Biological Sciences, Purdue University, West Lafayette, IN, United States
3. School of BioSciences, Bio21 Institute, The University of Melbourne, Melbourne, Australia
4. Department of Microbiology and Immunology, Bio21 Institute and Peter Doherty Institute, The University of Melbourne, Melbourne, Australia
5. Department of Parasitology, Noguchi Memorial Institute for Medical Research, College of Health Science, University of Ghana, Legon, Ghana
6. Santa Fe Institute, Santa Fe, NM, United States

* Corresponding author

E-mail: pascualmm@uchicago.edu (MP)

Abstract

At a time when effective tools for monitoring malaria control and eradication efforts are crucial, the increasing availability of molecular data motivates their application to epidemiology. The multiplicity of infection (MOI), defined as the number of genetically distinct parasite strains co-infecting a host, is one key epidemiological parameter for evaluating malaria interventions. Estimating MOI remains a challenge for high-transmission settings where individuals typically carry multiple co-occurring infections. Several quantitative approaches have been developed to estimate MOI, including two cost-effective ones relying on molecular data: i) THE REAL McCOIL method is based on putatively neutral single nucleotide polymorphism loci, and ii) the *var*coding method is a fingerprinting approach that relies on the diversity and limited repertoire overlap of the *var* multigene family encoding the major *Plasmodium falciparum* blood-stage antigen PfEMP1 and is therefore under selection. In this study, we assess the robustness of the MOI estimates generated with these two approaches by simulating *P. falciparum* malaria dynamics under three transmission conditions using an extension of a previously developed stochastic agent-based model. We demonstrate that these approaches are complementary and best considered across distinct transmission intensities. While *var*coding can underestimate MOI, it allows robust estimation, especially under high-transmission where repertoire overlap is extremely limited from frequency-dependent selection. In contrast, THE REAL McCOIL often considerably overestimates MOI, but still provides reasonable estimates for low- and moderate-transmission. As many countries pursue malaria elimination targets, defining the most suitable approach to estimate MOI based on sample size and local transmission intensity is highly recommended for monitoring the impact of intervention programs.

Author Summary

Despite control and elimination efforts, malaria continues to be a serious public health threat especially in high-transmission regions. Molecular tools for evaluating these efforts include those seeking to estimate multiplicity (or complexity) of infection (MOI), the number of genetically distinct parasite strains co-infecting a host, a key epidemiological parameter. MOI estimation remains challenging in high-transmission regions where hosts typically carry multiple co-infections by *Plasmodium falciparum*. THE REAL McCOIL and the *var*coding are two cost-effective methods relying on distinct parts of the parasite genome, those respectively under neutrality and selection. The more recent *var*coding approach relies on the *var* multigene family encoding for the major blood-stage antigen and contributing to a complex immune evasion strategy of the parasite. We compare the performance of the two methods by simulating disease dynamics under different transmission intensities with a stochastic agent-based model tracking infection by different parasite genomes and immune memory in individual hosts, then sampling resulting infections to estimate MOI. Although THE REAL McCOIL provides reasonable estimates for low- and moderate-transmission, *var*coding allows more robust estimates especially under high-transmission. Defining the most suitable approach to estimate MOI based on local transmission intensity is highly recommended for hyper-diverse pathogens such as malaria.

Introduction

Malaria deaths have steadily and significantly declined over the period 2000–2019 in response to control and elimination efforts [1]. However, malaria continues to be a serious threat causing approximately half a million deaths in 2019, especially among young children in high-transmission endemic regions in Africa. In these regions, infections are characterized by multiple genetically distinct *Plasmodium* parasite genotypes simultaneously infecting a host. Multiplicity of infection (MOI), also known as complexity of infection (COI), is defined as the number of genetically distinct parasite strains co-infecting a single host [2]. Multiclonal infections (i.e., MOI > 1) can be the result of a single bite by a mosquito transmitting more than one genetic parasite strain or independent bites by infected mosquitoes (also termed superinfection). The number of co-infections is associated with transmission intensity, clinical risk, age and immunity [3–6].

Given the potential relevance of MOI to malaria surveillance, various approaches have been developed to estimate MOI from clinical samples. As *Plasmodium* parasites reproduce asexually as haploid stages when they infect humans, signatures of polymorphic genotypes are evidence of multiclonal infections. While any highly polymorphic marker is thus suitable for estimating MOI, it remains a challenge to accurately measure MOI in malaria-endemic areas where multiclonal infections are common. The most common approach for determining MOI involves size-polymorphic antigenic markers, such as *msp1*, *msp2*, *msp3*, *glurp*, *ama1*, and *csp*, that can be amplified by PCR and determined by capillary electrophoresis or agarose gel [7]. Similarly, microsatellites, also termed simple sequence repeat (SSR), are another type of size-polymorphic marker that can be amplified by PCR to estimate MOI by determining the number of alleles

detected [5,8–12]. However, despite improved resolution of allele detection by capillary electrophoresis, these approaches based on size-polymorphisms usually involve a certain degree of subjective interpretation, are unable to discriminate alleles of similar sizes [13], and create PCR artifacts resulting in inconsistent results, especially for MOI > 5 [14,15]. As an alternative to size-polymorphic markers, other methods of determining MOI have focused on reconstructing haplotypes from genotyping or sequencing data (e.g., estMOI, FWS, and DEploid) [16–19]. Whole-genome sequencing is currently not a cost-effective approach when MOI is the main interest of a study, and these haplotype-reconstruction approaches are computationally intensive, resulting in limited MOI reliability for highly complex infections [20]. Finally, two cost-effective molecular approaches, known as THE REAL McCOIL [21] and *var*coding [22,23], have been more recently developed to identify and track MOI with standard laboratory equipment. They differ in important ways as they rely on contrasting parts of the genome, respectively under neutrality and immune selection.

As many possible genotypes exist among a combination of several genome-wide single nucleotide polymorphisms (SNPs), methods of determining MOI have focused on neutral SNP data to discriminate among strains [24]. Galinsky *et al.* (2015) developed the COIL approach to estimate MOI from a panel of bi-allelic SNP data, but this method relies on monoclonal infections (MOI = 1) for estimation of allele frequencies or requires external allele frequency data. As external allele frequency data may only be available for specific locations and be heterogeneous in space and time, an analytical approach, called THE REAL McCOIL (Turning HEterozygous SNP data into Robust Estimates of ALelle frequency, via Markov chain Monte Carlo, and Complexity Of Infection using Likelihood), was developed to simultaneously estimate the MOI

within a human host and the allele frequencies in the population based on a panel of SNPs [21]. Using simulations and 105 SNP data from cross-sectional surveys in Uganda, Chang *et al.* (2017) showed that THE REAL McCOIL approach improved performance in estimates of both quantities, despite the uncertainty of these estimates increasing with true MOI.

The more recent *var*coding approach (also termed *var* genotyping or *var* fingerprinting) [22,23], employs the highly polymorphic sequences encoding the immunogenic DBLα domain of PfEMP1 (*Plasmodium falciparum* erythrocyte membrane protein 1), the major surface antigen of the blood stage of infection [26]. During an infection, PfEMP1 molecules are exported by the parasite to the surface of the infected erythrocytes, where they influence virulence of the disease and become a target of the adaptive immune system [27]. The multigene family known as *var* encodes variants of this surface antigen which can reach tens of thousands of variants in endemic populations [22,28–33]. The sequential expression of a set of up to 60 *var* genes per parasite (hereafter, a repertoire) leads to immune evasion, prolongs infection duration, and establishes chronic infections enabling onward transmission [34,35]. Immune evasion is particularly important in high transmission regions where *var* repertoires are composed of largely distinct sets of *var* genes [22,28–31,36]. This non-random composition of *var* repertoires has been shown to result from negative frequency-dependent immune selection. Largely non-overlapping *var* repertoires enhance survival in semi-immune hosts, in accordance with earlier models of parasite competition for hosts via specific immunity [37] and more recent deep molecular sampling of local populations and computational theory [22,30,36]. This feature of *var* population structure allows distinct strains to accumulate in the blood of human hosts. The extensive diversity of the *var* gene family together with the very low percentage of *var* genes shared between parasites facilitate

measuring MOI by amplifying, pooling, sequencing, and counting the number of DBLα types in a host. This feature of *var* population genetics is the basis of the fingerprinting concept of *var*coding. From a single PCR with degenerate primers and amplicon sequencing, the method counts unique DBLα types per infection. It is not based on assigning haplotypes but instead, it assumes a set number of types per genome based on control data accounting for PCR sampling errors to calculate MOI [22,23].

As *var*coding does not require haplotype construction, we propose that this method is particularly well suited for high transmission where REAL McCOIL has shown some limitations due to the bi-allelic nature of the SNP calling [21]. To evaluate the relative performance of these two contrasting approaches to estimate MOI across different transmission settings, this study simulates malaria transmission under low, moderate, and high-transmission using an extended agent-based model (ABM). We specifically extend a previously developed stochastic computational model to incorporate neutral bi-allelic SNPs which, together with the *var* genes, can be used for MOI estimation. Depending on transmission intensity, we ask whether one of these approaches is more accurate than the other. When most infections are multiclonal (i.e., MOI > 1), we demonstrate that THE REAL McCOIL and the *var*coding approaches tend to overestimate and underestimate the MOI, respectively. Moreover, while the high diversity of the *var* gene family allows robust MOI estimation with the *var*coding approach, especially across high-transmission settings, THE REAL McCOIL provides reasonable estimates across low- and moderate-transmission settings where the *var*coding can be limited by partially overlapping *var* repertoires. We discuss the limitations and advantages of these two approaches to determine the multiplicity of malaria parasite infection as well as their implications for malaria surveillance.

Results

Accurate estimation of multiplicity of infection is important for evaluating current intervention strategies against malaria and thus defining or adapting future ones. We evaluated two recently developed MOI estimation approaches by simulating malaria transmission using an extended ABM, and sampling, estimating, and comparing MOI under three different transmission settings (i.e., “low”, “moderate”, or “high”).

For each simulation under low-, moderate-, or high-transmission intensity, 2000 individuals were sampled at the end of the wet season (S1 Fig). On the one hand, the number of sampled individuals per age class is consistent with the age distribution and the size of each age class (S1A and S1B Figs). On the other hand, the number of infected sampled individuals was significantly and negatively correlated with the age of the hosts, as expected (Pearson correlation test; high-transmission: r = -0.81, *P*-value < 2.2e-16; moderate-transmission: r = -0.63, *P*-value = 3.0e-10; low-transmission: r = -0.65, *P*-value = 9.5e-11). The hosts between 0 and 5 years old thus exhibit the highest number of infections with an average of 64 ± 11 (mean ± SD), 267 ± 17, and 377 ± 18 sampled hosts for the low-, moderate-, and high-transmission simulations, respectively (S1C Fig). As expected, the number of sampled hosts is higher for simulations with high-transmission settings (939 ± 344) than for those with moderate (314 ± 151) and low-transmission settings (88 ± 36) (S1C Fig). Consistently, the average EIR and prevalence were also higher for simulations with high-transmission settings than for those with low- or moderate-transmission settings (S2 Fig). For instance, low-transmission setting simulations have an average of 1.1 ± 0.2 infectious bites per host per year and 0.04 ± 0.02 infected cases, whereas moderate-transmission simulations have an

175 average of 6.8 ± 0.2 infectious bites per host per year and 0.16 ± 0.08 infected cases, and high-
176 transmission simulations have an average of 21.6 ± 0.2 infectious bites per host per year and 0.47
177 ± 0.17 infected cases, reflecting highly contrasting malaria transmission intensities. We note that
178 in our simulations "infectious bites" were computed as infectious contact events experienced by a
179 host, and that for generality purposes, we set the transmissibility probability specifying whether
180 such contact results in infection to 0.5 (S1 Table). Thus, for comparison purposes with empirical
181 values, the EIR values obtained in the simulations should be divided by such probability.
182 Consistently with these epidemiological and genetic diversity statistics, the true MOI distribution
183 generated by the simulations is also significantly different among the three levels of transmission
184 (t-test: *P*-value < 2.2e-16; average true MOI of 1.08 ± 0.30, 1.61 ± 0.95, and 3.84 ± 3.31 for the
185 low-, moderate-, and high-transmission simulations, respectively) (Fig 1B).
186

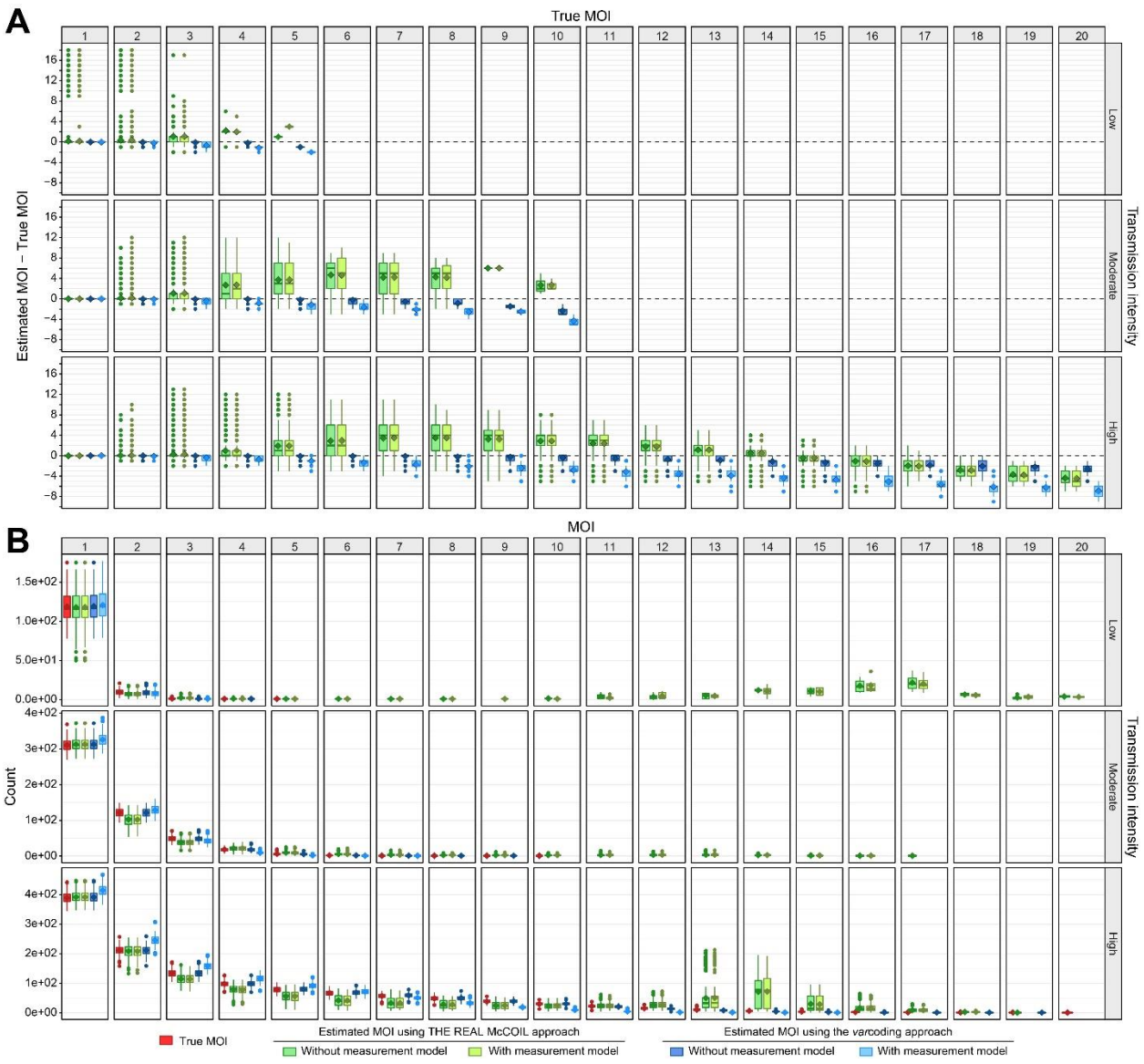

Fig 1: Multiplicity of infection (MOI). For each category, the horizontal central solid line represents the median, the diamond represents the mean, the box represents the interquartile range (IQR) from the 25th to 75th centiles, the whiskers indicate the most extreme data point, which is no more than 1.5 times the interquartile range from the box, and the dots show the outliers, i.e., the points beyond the whiskers. The upper, middle, and lower row panels show correspond to simulations under low-, moderate-, and high-transmission settings, respectively (S1 and S2 Tables). **A)** Accuracy of MOI estimates, defined as the difference between estimated and true MOI

per host. While null values highlight accurate MOI estimates (indicated by a dashed black horizontal line), the positive and negative values highlight over- and under-estimation, respectively. Estimates with the neutral SNP-based approach (THE REAL McCOIL) are indicated in green, and those with the *var* gene-based approach (*var*coding) are indicated in blue. The dark and light green or blue colors indicate respectively MOI estimations made without and with a measurement model (Fig 2). The column panels show differences for specific true MOI values. **B)** Population distribution of the estimated and true MOI per host from the simulated "true" values and those estimated with the methods indicated by the colors similar to panel A. For high transmission, the distribution obtained with THE REAL McCOIL shows a more pronounced tail than that from the simulated infections, with a secondary peak around MOI = 14. Note that the method considerably over-estimates individual MOI below that value but then under-estimates above it (panel A). Thus, these opposite trends compensate each other to some extent in the population distribution, producing nevertheless a deviation at high values. The *var*coding method provides a good representation of the "true" distribution from the simulations, and of the individual values in general, with a consistent tendency to underestimate when sampling error is taken into account.

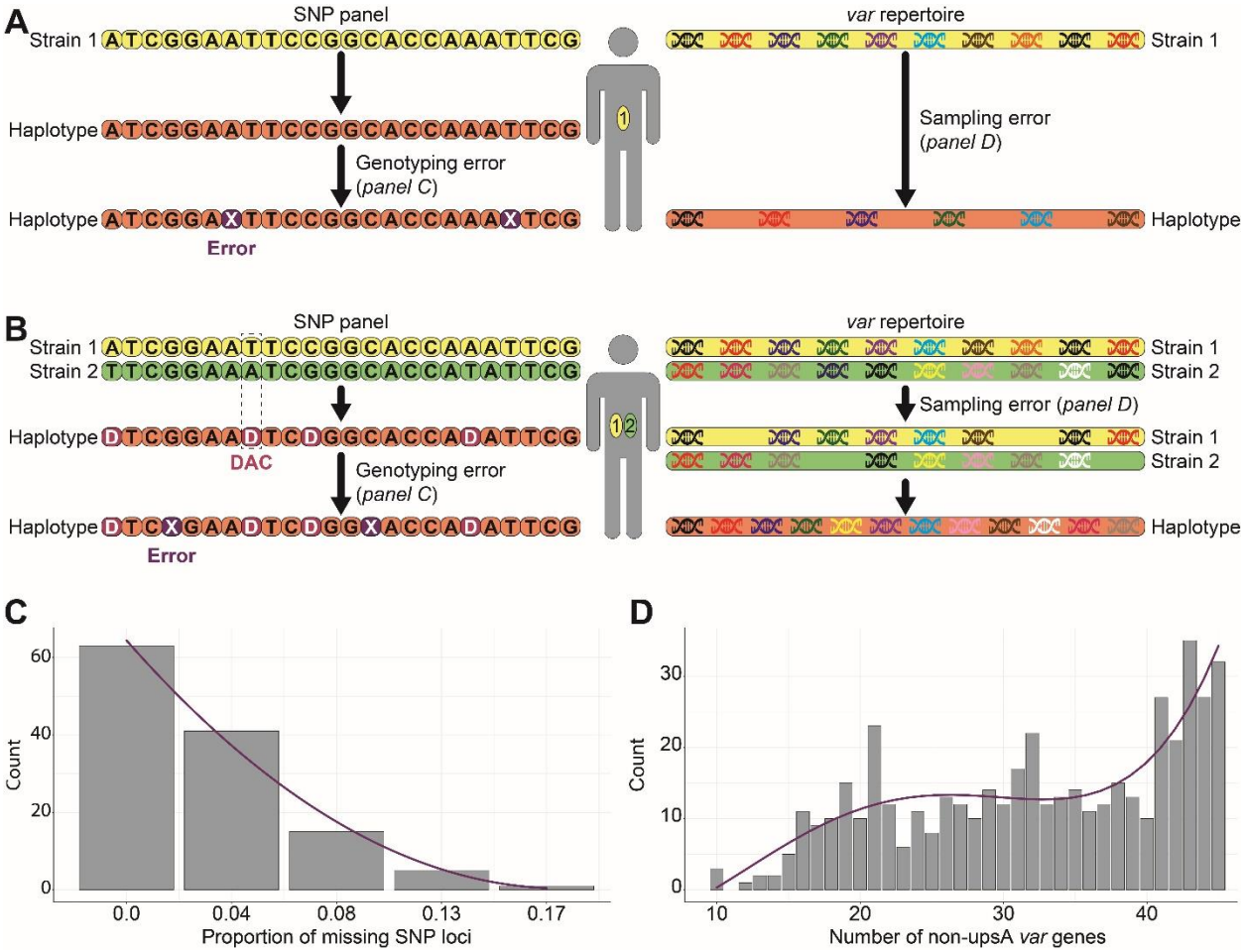

Fig 2: Measurement models. A) and **B)** Schematic diagrams of the SNP and *var* measurement models for a host infected by one (MOI = 1) or two (MOI = 2) genetically distinct *P. falciparum* strains, respectively. To account for potential SNP genotyping failures, we randomly replaced the host genotypes with missing data (X). This replacement was implemented by using the distribution illustrated in panel C. When MOI is high, the frequency of double allele calls (DACs) is also high (Fig 3). To account for *var* gene potential sequencing errors, we sub-sampled the number of *var* genes per repertoire. This sub-sampling was implemented by using the distribution illustrated in panel D. For simplicity, the *var* repertoire in these two examples only consists of 10 *var* genes despite each migrant parasite genome consists of a repertoire of 45 *var* genes in the simulations (S1 Table). **C)** Histogram of the proportion of missing SNP loci per host haplotype from a panel

of 24 bi-allelic SNP loci. The genotypes were previously obtained from monoclonal infections sampled during one cross-sectional survey made in 2015 in the Bongo District, in northern Ghana. The purple curves show the best curves that fit the data using the adjusted R-squared. **D)** Histogram of the number of non-upsA (i.e., upsB and upsC) DBLα *var* gene types per repertoire. The molecular sequences were previously sequenced from monoclonal infections, i.e., hosts infected by a single *P. falciparum* strain (MOI = 1), sampled during six cross-sectional surveys made from 2012 to 2016 in the Bongo District, in northern Ghana.

Estimates with THE REAL McCOIL approach

THE REAL McCOIL approach based on neutral SNP data tends to overestimate MOI when true values range from 3 to 14 (Fig 1A). In contrast, this approach tends to underestimate MOI for true MOI above ~14, values only found in hosts without any single minor allele calls (Figs 2A and 3). Underestimated MOI can differ from the true MOI by up to 2, 3, and 7 co-infections for the low-, moderate-, and high-transmission simulations, respectively. Interestingly, while most hosts with MOI < 3 show accurate MOI estimates, some can differ from the true MOI by up to 18, 12, and 13 co-infections for the low-, moderate-, and high-transmission simulations, respectively. The inaccuracy of the MOI estimates based on THE REAL McCOIL approach, defined as the absolute differences between estimated and true MOI per host, is significantly and positively correlated with the true MOI (*P*-values < 2.2e-16; r = 0.07, r = 0.59, and r = 0.55 for the low-, moderate-, and high-transmission simulations, respectively).

As the proportion of double allele calls (DACs) per host is significantly and positively correlated with the true MOI (*P*-value < 2.2e-16, r = 0.82) (Fig 3), inaccuracy is also significantly and positively correlated with the proportion of DACs per host (*P*-value < 2.2e-16, r = 0.61). Consistently, inaccuracy is significantly and negatively correlated with the proportion of single allele calls (minor and major alleles) per host SNP haplotype (*P*-value < 2.2e-16, r = -0.61) (Fig 3). Inaccuracy is significantly and negatively correlated with the number of SNPs (*P*-value < 2.2e-16, r = -0.15), but estimated MOI with the highest number of SNPs (105 SNPs) can still differ from the true MOI by up to 18 (S3A Fig). Surprisingly, the MOI estimates show higher accuracy for simulations generated with distinct initial SNP allele frequencies but did not seem influenced by the presence of linked SNP loci (S4 Fig). Overall, despite including slightly fewer hosts with low MOI and significantly more hosts with high MOI (Fig 1B), the average estimated MOI was quite similar to that of the true MOI (*P*-values < 2.2e-16; average estimated MOI of 1.27 ± 1.60, 1.93 ± 1.93, and 4.80 ± 4.60 for the low-, moderate-, and high-transmission simulations, respectively). However, due to the combination of overestimated and underestimated MOI values, the distribution showed a second peak of high density around a MOI of 15, which is absent from the true MOI distribution. This peak can correspond to a maximum of ~200 hosts in some simulations.

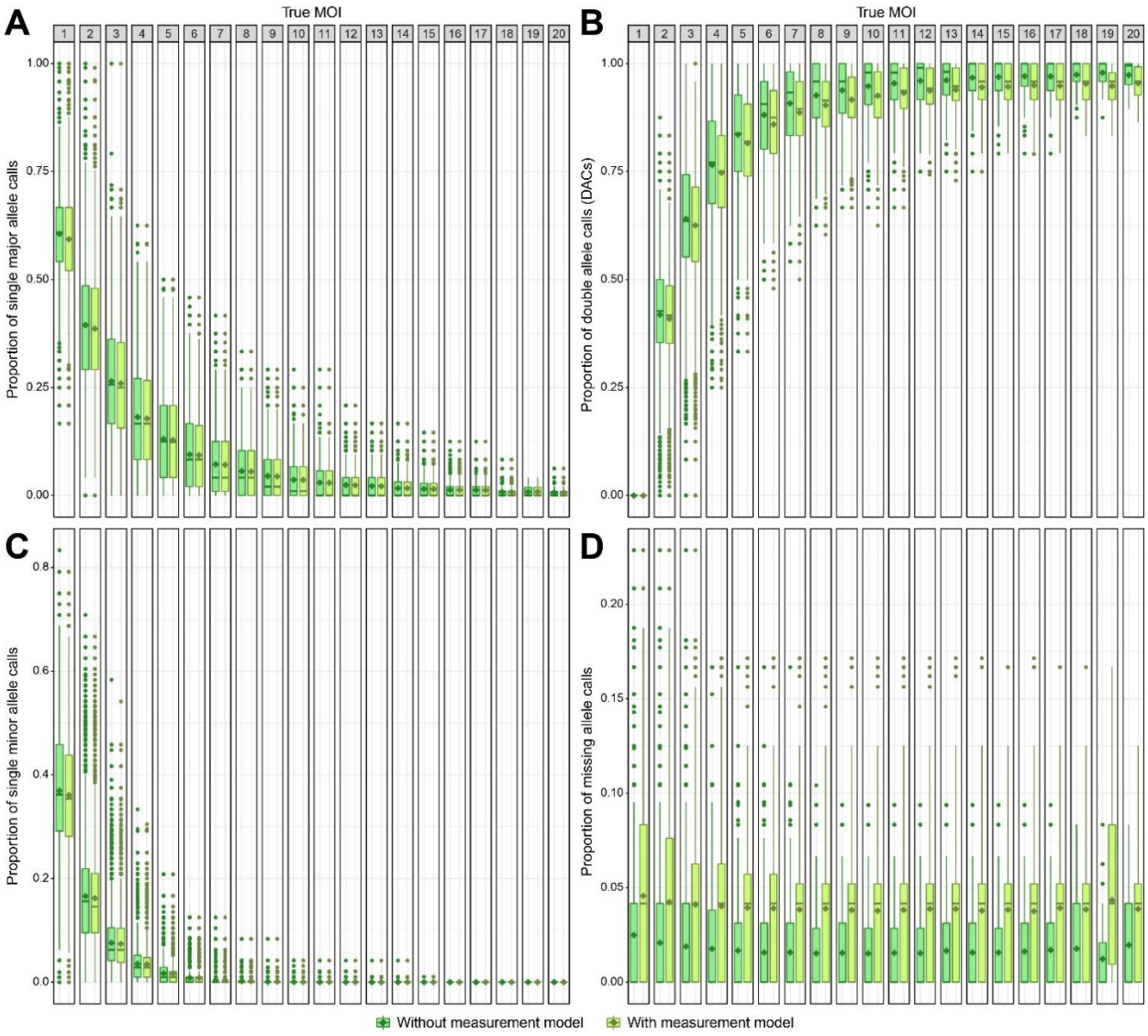

Fig 3: Proportion of SNP calls (genotypes) per host SNP haplotype. A) Single major allele calls; **B)** Double allele calls (DACs); **C)** Single minor allele calls; **D)** Missing allele calls. The column panels show the proportions for specific true MOI values. The dark and light green colors indicate the proportion of calls made without and with a measurement model, respectively (Fig 2). For each category, the horizontal central solid line represents the median, the diamond represents the mean, the box represents the interquartile range (IQR) from the 25th to 75th centiles, the whiskers indicate the most extreme data point, which is no more than 1.5 times the interquartile range from the box, and the dots show the outliers, i.e., the points beyond the whiskers.

Consideration of a SNP measurement model, which accounts for potential genotyping failures by randomly replacing some SNP genotypes with missing values, only slightly decreases the accuracy of the MOI estimates based on THE REAL McCOIL approach, relative to MOI estimates made without the measurement model (Figs 1A and 2).

Subsampling the individuals from 2000 to 500 did not reduce the accuracy of the estimated MOI with or without the measurement model (S5 Fig). However, subsampling to 200 individuals significantly increased the number of SNP loci with a MAF < 10%, especially for the low- (7 ± 8 loci) and moderate-transmission setting simulations (3 ± 3 loci). Consequently, due to the high number of SNP loci, and thus individuals, that could not be considered in THE REAL McCOIL analysis, MOI could not have been estimated for any of the low-transmission setting simulations when a subsampling of 200 individuals was applied (S6 Fig). Moreover, while THE REAL McCOIL approach could provide MOI estimates for subsamples of the moderate- and high-transmission setting simulations, the subsampling of 200 individuals significantly reduced the accuracy of these MOI estimates made with or without the measurement model.

Simulations under low- and moderate-transmission settings show more accurate MOI estimates than the simulations under high-transmission settings (Fig 1A). For the low-transmission setting, the less accurate MOI estimates (> 13) were generated in only four simulations (i.e., one replicate from runs 10, 11, 22, and 23) (Fig 1B and S2 Table). These results were explained by highly inaccurate estimated MAF. Indeed, the inaccuracy of the estimated MOI was significantly and positively correlated with the inaccuracy of the estimated MAF per simulation, defined as the

sum of the absolute differences between estimated and true MAF per SNP locus (*P*-values < 2.2e-16; r = 0.12 and r = 0.13 for estimates made with or without measurement model, respectively). Interestingly, although THE REAL McCOIL approach generated very inaccurate MAF estimates for a few low-transmission simulations, it typically produced very accurate estimates of the MAF regardless of transmission settings (S7 Fig). Consistent with the MOI results, the accuracy of the MAF estimates per simulation also increases with the number of SNP loci. The inaccuracy of THE REAL McCOIL MAF estimates per locus, defined as the absolute differences between estimated and true MAF per locus, is significantly and negatively correlated with the true MAF, the proportion of DACs, and the proportion of single minor allele calls per locus (S3 Table). Moreover, this inaccuracy of the MAF estimations is also significantly but positively correlated with the proportion of missing allele calls and single major allele calls per locus (S3 Table).

Estimates with the *var*coding approach

Consistently with an increasing probability of overlapping *var* repertoires for hosts with MOI above 1, the *var*coding approach, which uses the number of *var* genes to estimate MOI, tends to slightly underestimate the MOI (Fig 1A). Its inaccuracy, defined as the absolute difference between estimated and true MOI per host, is significantly and positively correlated with the true MOI (*P*-values < 2.2e-16; r = 0.23, r = 0.26, and r = 0.54 for the low-, moderate-, and high-transmission simulations, respectively). Therefore, MOI values estimated for the hosts with the highest true MOI (i.e., 5, 10, and 20 co-infections for the low-, moderate-, and high-transmission simulations, respectively), differ from the true MOI by up to 1, 3, and 5 co-infections for the low-, moderate-, and high-transmission simulations, respectively. Overall, despite exhibiting slight

deviations, with more hosts at low MOI and fewer hosts at high MOI, the distribution of the estimated MOI based on the *var*coding approach is quite similar to that of the true MOI (*P*-value < 2.2e-16; average estimated MOI of 1.08 ± 0.29; *P*-value = 0.10, 1.60 ± 0.93; *P*-value = 4.9e-05, and 3.74 ± 3.14 for the low-, moderate-, and high-transmission simulations, respectively) (Fig 1B).

As expected, the *var* genes measurement model, which accounts for potential sampling errors by sub-sampling the number of *var* genes per strain, reduced the number of available distinct *var* genes per host and increased inaccuracy (Figs 1A and 2). MOI estimates from simulated data can now differ from the true MOI by up to 2, 5, and 9 co-infections for the low-, moderate-, and high-transmission simulations, respectively. Overall, the distribution of the estimated MOI remained quite similar to that of the true MOI (*P*-values < 2.2e-16; average estimated MOI of 1.07 ± 0.26, 1.49 ± 0.75, and 3.05 ± 2.30 for the low-, moderate-, and high-transmission simulations, respectively), even though it accentuated some of the small deviations we described in the absence of measurement error, namely more hosts with low MOI and fewer hosts with high MOI (Fig 1B).

As the *var*coding approach estimates MOI using individual host level information, subsampling the dataset, from 2000 to 500 or 200 individuals, did not reduce accuracy regardless of consideration of measurement error (S5 and S6 Figs).

The high-transmission setting simulations resulted in more accurate MOI estimates than the low- and moderate-transmission simulations (Fig 1A). This is consistent with the *var* repertoires for simulations under high-transmission settings exhibiting a lower average PTS (0.049 ± 0.001), i.e., less overlap, than the *var* repertoires for simulations under low- and moderate-transmission

settings (0.138 ± 0.027 and 0.063 ± 0.004, respectively) (Figs 4C, S8, and S9). The genetic structure of the parasite population can also be analyzed using networks whose nodes are *var* repertoires, and the weighted edges correspond to the degree of overlap between these repertoires (Fig 4A) [36]. Consistent with the average PTS, the similarity networks for simulations under high-transmission settings did not group the *var* repertoires into well-defined modules while the similarity networks for simulations under high-transmission settings did. This finding was also captured using three-node motifs across the *var* repertoire similarity networks, which showed a lower proportion of reciprocal motifs (i.e., A ↔ B ↔ C ↔ A) for simulations under high-transmission settings (91.4%) than for simulations under low- and moderate-transmission settings (99.8% and 100.0%, respectively) (Figs 4A, 4B, and S8). Altogether, as simulated *P. falciparum* strains under high-transmission settings shared less similar *var* repertoires than those under low- or moderate-transmission settings, the *var*coding approach results in more accurate MOI estimates under high-transmission settings than under the lower transmission ones (Figs 2A, 4, and S8).

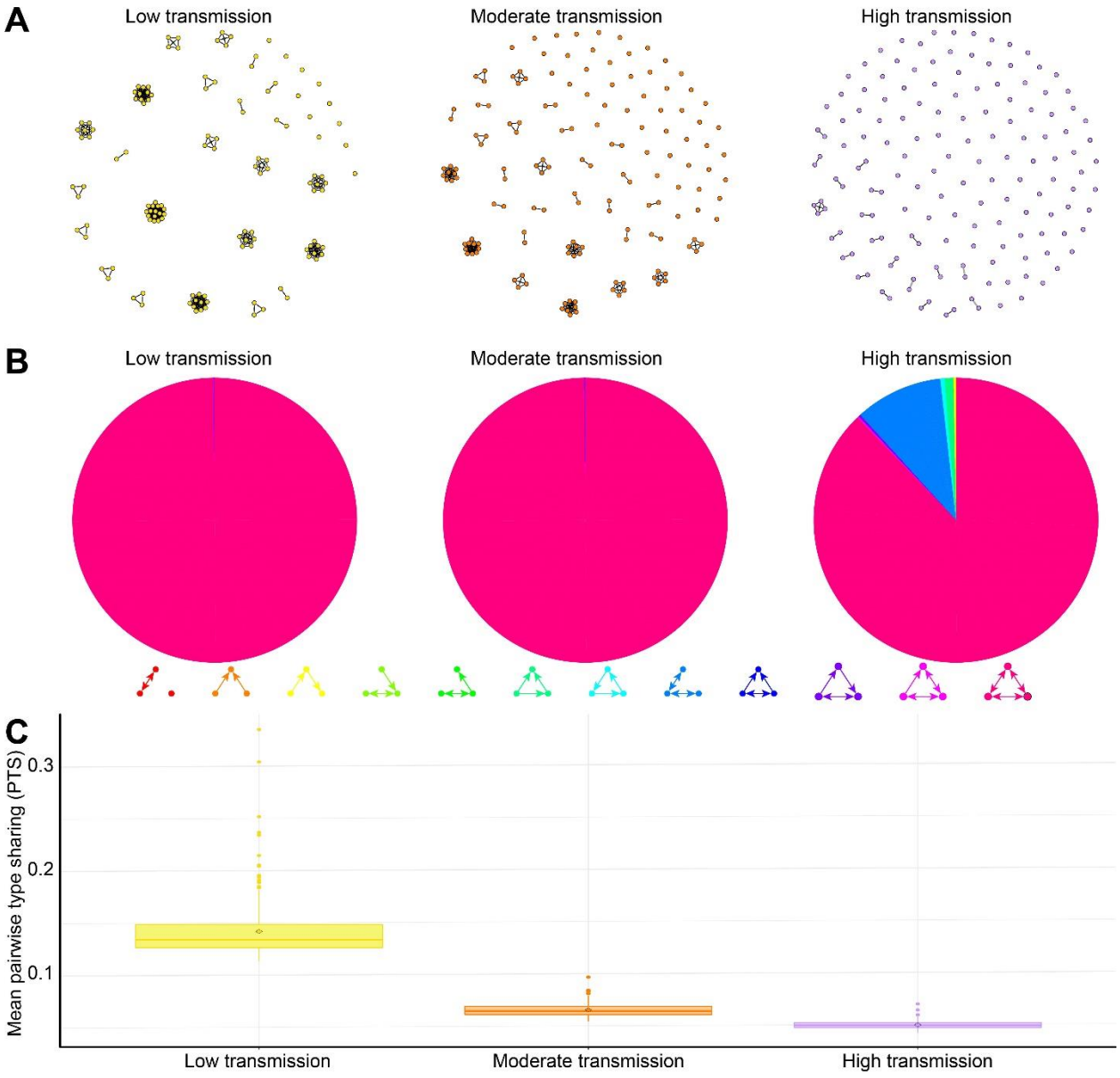

Fig 4: Population structure using network properties. Comparisons of repertoire similarity networks of 150 randomly sampled parasite *var* repertoires generated from a one-time point under low, moderate, and high-transmission settings (i.e., one replicate of runs 1, 25, and 49, respectively; S1 and S2 Tables). Only the top 1% of edges are drawn and used in the analysis. **A)** Similarity networks where nodes are *var* repertoires, weighted edges encode the degree of overlap between the *var* genes contained in these repertoires, and the direction of an edge indicates the asymmetric competition between repertoires. **B)** Distributions of the average proportion of

occurrences of three-node graph motifs across the repertoire similarity networks. **C)** Distribution of the mean pairwise type sharing (PTS) between *var* repertoires. For each category, the horizontal central solid line represents the median, the diamond represents the mean, the box represents the interquartile range (IQR) from the 25th to 75th centiles, the whiskers indicate the most extreme data point, which is no more than 1.5 times the interquartile range from the box, and the dots show the outliers, i.e., the points beyond the whiskers.

Comparison of THE REAL McCOIL and the *var*coding approaches

As described above, each method performs better under specific conditions, such as transmission setting and sampling size. However, for any given simulation, THE REAL McCOIL approach never reached the level of accuracy of *var*coding (Fig 1). First, for simulations under low- and moderate-transmission settings, THE REAL McCOIL approach could generate highly inaccurate MAF, due to the small proportion of infected hosts sampled from the participants, which can result in more inaccurate MOI estimates than those generated with *var*coding. Second, under high-transmission settings, THE REAL McCOIL approach showed a combination of MOI overestimates and underestimates for true MOI under or above ~14 co-infections, respectively (Fig 1). This introduces biases in opposite directions, which can compensate to some extent and artificially provide a reasonable overall population distribution. In contrast, the *var*coding approach showed consistent increasing underestimation of the MOI for increasing true MOI as expected from an increasing *var* repertoire overlap between strains, Third, for a similar sample size (i.e., 2000, 500, or 200 individuals), the *var*coding approach always provided more accurate MOI estimates than THE REAL McCOIL approach (S5 and S6 Figs). This was especially

observed for the smaller sample sizes (i.e., 200 samples), which are quite commonly used for malaria surveillance. On the one hand, THE REAL McCOIL, which relies on information at the population level, could not provide MOI estimates due to the limited number of SNP loci that could be considered in the analysis or could provide less accurate MAF and MOI estimates. On the other hand, the *var*coding, which only relies on the information at the individual level to estimate MOI, consistently provided comparable MOI estimates independently of the sample size. In summary, the accuracy of the estimated MOI was dependent on the transmission setting, the approach used to characterize the multiplicity of malaria parasite infection, and the sample size.

Discussion

When *P. falciparum* transmission is high, which is common within malaria-endemic regions of the world, simulations showed that THE REAL McCOIL approach provided less robust MOI estimation than the *var*coding approach. The former approach tends to overestimate the MOI for hosts with low and moderate true MOI (under ~14 co-infections) and to underestimate the MOI for hosts with high true MOI (above ~14 co-infections). The high proportion of DACs and low proportion of single major and minor allele calls in a host SNP haplotype (barcode) seemed to be the origin of this inaccuracy in the high-transmission simulations. It is interesting to note that the combination of underestimated and overestimated MOI values allowed THE REAL McCOIL approach to generate a fairly accurate average estimated MOI but for the wrong reasons. Therefore, caution should be taken with using this approach when malaria transmission is moderate to high. In particular, considerable overestimates in the population result in a secondary peak in the distribution. In contrast, the low PTS values at high transmission due to the high diversity of *var* genes and selection for reduced repertoire overlap enabled accurate MOI estimates via *var*coding by amplifying, pooling, sequencing, and counting the number of DBLα types in a host. Despite harboring a high proportion of distinct *var* genes, repertoires can still be partially overlapping, sharing similar *var* genes. This limited overlap can result in a reduced number of *var* genes being identified on the basis of their DBLα types, which leads to the potential underestimation of MOI. Because these regions of the parasite genome can be sometimes challenging to access, the resulting sampling errors can reduce the reliability of the methodology leading to consistent underestimation of MOI, as shown with the simulations that included a realistic measurement error based on empirical data. Interestingly, simulations including a measurement error based on the distribution

of the number of non-upsA DBLα *var* gene types per 3D7 laboratory isolate significantly improved the accuracy of the *var* coding MOI estimates, highlighting the importance of high-density isolates when estimating MOI (S10 Fig). Thus, the *var*coding approach provides a cost-effective approach for evaluating MOI under high-transmission conditions, with an underestimation bias introduced however by measurement error from the sampling of the *var* genes.

In low or moderate malaria transmission regions, both THE REAL McCOIL and *var*coding approaches can provide reasonable MOI estimates. As bi-allelic SNP data can be relatively cheap and straightforward to obtain, THE REAL McCOIL method, which can be applied to any parasite isolates with multiclonal infections [21], appears cost-effective for evaluating MOI but could nevertheless introduce biases in determining malaria elimination status by underestimating the effectiveness of the interventions. However, the method required a minimum number of sampled hosts to reasonably estimate MOI, which is not the case for *var*coding. Caution should thus be taken when defining the most suitable sample size while keeping this method cost-effective. Given the high accuracy of the *var*coding approach when measurement error was not incorporated, future work will address correcting for *var* repertoire overlap within a single host to improve MOI estimation.

Surprisingly, although THE REAL McCOIL approach assumes that genotyped SNP loci do not exhibit significant LD, the simulations performed with linked SNP loci did not show less reliable MOI estimation. Linked SNP loci may require longer simulations than the ones considered here (with thousands of generations) to show substantial bias in estimates of MOI. The categorical method of THE REAL McCOIL approach was very sensitive to the parameter controlling the

upper bound for MOI (maxCOI). While a maximum MOI of 20 was applied in the simulations, reducing this upper bound for the low- and moderate-transmission simulations to the highest true MOI values observed for these settings (to 5 and 10 for the low and moderate-transmission simulations, respectively) significantly improved the accuracy of the estimated MOI (S11 Fig). Defining the most suitable MOI upper bound from previous reports could thus be useful to provide more accurate future estimates when using this approach. However, this solution can be circular and therefore impractical.

THE REAL McCOIL approach provided highly accurate MAF estimates for low- transmission intensities and reasonably accurate ones in moderate- and high-transmission intensities. While MAF estimates were robust with as few as 24 SNPs, their accuracy was improved by increasing the number of SNPs genotyped. Most population genetic analyses of malaria parasites rely on monoclonal infections, which reduces the amount of data and produces MAF estimates that may not be representative and thus introduce potential biases. Therefore, despite often significantly overestimating the MOI, THE REAL McCOIL approach could also facilitate population genetic analyses of the malaria parasite by properly estimating the MAF and other related statistics, including the effective population size (N_e), the F_{ST}, and the F_{WS} [38–40].

An additional source of measurement error, not considered by our simulations, concerns the difficulty of properly sampling all the strains simultaneously infecting a particular host due to low parasitemia, the parasite load in the host blood. Indeed, multiclonal infections can potentially result in a reduction of the parasitemia of particular strains [41]. Consequently, strains with low parasitemia could be highly diluted in clinical samples and thus have a lower probability of being

properly sequenced and/or genotyped. Moreover, infections (monoclonal or multiclonal) with low total parasitemia could be missed by the PCR detection approach, which can identify infections with as low as one parasite per μL, and by the commonly used microscopy detection approach, which cannot detect infections with lower than 4-10 parasites per μL [42,43]. Synchronicity of clones in the 48 hour life cycle on alternate days also leads to underestimation of MOI unless repeat daily sampling [44] or every three days is done [45]. These issues could therefore contribute to a more substantial underestimation of the MOI than the ones highlighted in this study.

Another simplification in our ABM was the lack of SNP mutations. Evolution of the neutral part of the parasite genome may influence MOI estimation over long time scales, but should play a minor role for the shorter time periods relevant to epidemiology unless associated with selective sweeps. Finally, our model did not incorporate the different sequence groupings of *var* genes, which can be classified based on their chromosomal position and semi-conserved upstream promoter sequences (ups) into different groups, upsA and non-upsA [46–48]. Case-control studies have reported that while upsA *var* genes are preferentially expressed in children with cerebral and/or severe malaria, non-upsA *var* genes have been associated with asymptomatic infections and clinical cases of malaria [49–57]. This absence of *var* types in our ABM may explain the higher prevalence within the younger age class in the simulations (i.e., 0-5 years old), compared to an observed higher prevalence typically within the 6-10 and 11-20 years old age classes in the empirical data [58].

Although MOI is a useful epidemiological marker to evaluate the efficacy of malaria intervention efforts, properly characterizing multiclonal infections remains a challenge, especially

for high transmission. This work demonstrates that THE REAL McCOIL and the *var*coding approaches provide complementary methodologies to determine MOI across distinct transmission settings. In particular, the high diversity of the *var* gene family and low overlap of *var* gene repertoires between parasites, especially under high-transmission intensity, allows robust MOI estimation with the *var*coding method, despite a tendency for underestimation originating mainly from sampling error. Reliance of THE REAL McCOIL on bi-allelic neutral SNPs limits application at high transmission, with the introduction of a secondary peak in the tail of the population distribution, considerable over-estimates of individual MOI, and opposite signs in the deviations, for both under- and over-estimates in different ranges of true values. The method provides reasonable estimates across low- and moderate-transmission settings where the *var*coding approach could be limited by partially overlapping *var* repertoires. Considering local transmission intensity is thus highly recommended when defining the most suitable marker and/or MOI estimation approach to evaluate the impact of malaria control and elimination campaigns. The highly diverse multigene *var* family under immune selection provides a handle to complexity of infection at high transmission.

Materials and Methods

Agent-Based Model (ABM)

Malaria transmission was modeled with an extended implementation of an agent-based, discrete-event, stochastic model in continuous time [36,59]. Here, we briefly describe the agent-based model (ABM) implemented in Julia (varmodel3) which is based on previous C++ implementations (i.e. varmodel and varmodel2) [36,59]. While the previous implementation of the stochastic ABM was adapted from the next-reaction method which optimizes the Gillespie first-reaction method, this implementation uses a simpler Gillespie algorithm [60,61]. The ABM tracks the infection history and immune memory of each host and its parameters and symbols are summarized in S1 Table. We modeled a local population of 10000 individuals, and a global *var* gene pool whose size acts as a proxy for regional parasite diversity. The simulations are initialized with 20 migrant infections from this regional pool to seed local population transmission and grow local gene diversity to a stationary equilibrium. Each migrant parasite genome consists of a specific combination (i.e. repertoire) of 45 *var* genes. The size of the repertoire was based on the median number of non-upsA DBLα sequences identified in our 3D7 laboratory isolate [22,23]. This grouping of *var* genes is defined based on their semi-conserved upstream promoter sequences (ups) (i.e. upsA and non-upsA (upsB and upsC)) [46–48]. Although each parasite carries both types of *var* genes in a fairly constant proportion [26,62], the MOI estimation method we consider here focuses on the non-upsA DBLα sequences as they were ~20X more diverse and less conserved among repertoires than the upsA DBLα sequences. Therefore, for simplicity purposes, our model considered only those types. Each *var* gene itself is represented as a linear combination of two

524 epitopes, i.e. parts of the molecule that act as antigens and are targeted by the immune system
525 [26,36,63]. The *var* genes in a repertoire are expressed sequentially and the infection ends when
526 the whole repertoire is depleted. The duration of the active period of a *var* gene, and thus of the
527 infection, is determined by the number of unseen epitopes. When a *var* gene is deactivated, the
528 host adds the deactivated *var* gene epitopes to its immunity memory. Specific immunity toward a
529 given epitope experiences a loss rate from host immunity memory, and re-exposure is therefore
530 required to maintain it. The local population is open to immigration from the regional pool.
531
532 Our model extension allows us to keep track of the neutral part of each migrant parasite genome
533 assembled by sampling one of the two possible alleles (labeled as 0 or 1) at each of a defined
534 number of neutral bi-allelic SNPs (S1 Table). While the extended model can generate
535 homogeneous initial SNP allele frequencies by sampling the migrant alleles with an identical
536 probability from the regional pool (i.e., 0.5), it can also generate distinct initial SNP allele
537 frequencies by sampling the migrant alleles from the regional pool with distinct probabilities that
538 sum up to one (e.g. 0.2 and 0.8) and are randomly picked from a defined range (e.g., [0.1-0.9]).
539
540 Seasonality was implemented in the transmission rate parameter to represent monthly
541 variability in mosquito bites [59,64]. The model does not explicitly incorporate mosquito vectors
542 but considers instead an effective contact rate (hereafter, the transmission rate) which determines
543 the times of local transmission events (exponentially distributed). At these times, a donor and a
544 recipient host are selected randomly. To mimic meiotic recombination which happens within the
545 mosquito during the sexual reproduction stage of the parasite, strains that are selected for a
546 transmission event have a probability $P_r = 1 - 1 / N_s$ (where N_s is the number of strains transmitted

to the donor) to become a recombinant strain [36]. To generate a recombinant *var* repertoire, a random set of *var* genes is sampled from a pool containing the two sets of *var* genes from the original genomes. Similarly, to generate the neutral part of a recombinant parasite, a random allele is sampled for each bi-allelic SNP. Moreover, to allow for linkage disequilibrium (LD) across the neutral part of the genome, neutral bi-allelic SNPs can be non-randomly associated and co-segregate as defined in a matrix of LD coefficients indicating the probability that pairs of linked SNPs will co-segregate during the meiotic recombination (S2 Table).

Experimental design

We explored how distinct transmission settings influence MOI estimation with the two different approaches. Specifically, we compared three transmission intensities corresponding to "low" (prevalence of 1-10%), "moderate" (prevalence of 10-35%), and "high" (prevalence ≥35%), implemented with different transmission rates (5.0e-05, 7.5e-05, and 1.0e-04, respectively), and initial gene pool sizes (500, 2000, and 10000, respectively) (S1 and S2 Tables) [65]. As the sensitivity of the SNP-based methods increases with the number of SNP loci, and as Chang *et al.* (2017) retained 105 SNP loci to test THE REAL McCOIL approach, we performed these simulations using 24, 48, 96, and 105 SNP loci for the three transmission settings (S1 and S2 Tables) [21,24,25]. As THE REAL McCOIL approach assumes that distinct parasite lineages in multiclonal infections are unrelated and that genotyped SNP loci do not exhibit significant LD, we performed the simulations with homogenous initial SNP allelic frequencies and with unlinked bi-allelic SNP loci (S1 and S2 Tables). However, as allelic frequencies can be heterogeneous in space and time, we also performed simulations with distinct initial allelic frequencies, and with 8% and

16% of linked SNP loci clustered into one or two groups, respectively. This design results in 72 distinct combinations of parameters (i.e., runs) and we ran 10 replicates per combination with a maximum MOI of 20 (S1 and S2 Tables). Simulations were run for 85 years to get beyond the initial transient dynamics in which *var* gene diversity and parasite population structure are established. For each simulation, we calculated the epidemiological summary statistics, including the number of hosts, the prevalence, and the entomological inoculation rate (EIR). In addition, 2000 individuals were randomly sampled to analyze the true MOI and the parasite genetic and allelic diversity patterns. The simulated data were collected during the last year at 300 days (i.e., November), corresponding to the end of the wet season (high-transmission season) in the Bongo District, a malaria-endemic area of Northern Ghana. Details on the area and population have been previously described [23,58].

MOI estimation

While the "true" MOI per host was directly extracted from the simulations, the estimated MOI was obtained for each host using the two distinct approaches. First, the MOI per host was estimated from the simulated neutral SNP data using THE REAL McCOIL approach v.2 [21]. We performed the categorical method of THE REAL McCOIL with a minor allele frequency (MAF) of 10% for a SNP to be considered and an upper bound of 20 for MOI, keeping all other parameters to their default values (a burn-in period of 10^3 iterations, a total of 10^4 Markov chain Monte Carlo (MCMC) iterations, a minimum number of 20 genotypes for an individual to be considered, a minimum number of 20 samples for a SNP to be considered, an initial MOI of 15, and a probability of calling single allele loci double allele loci and of calling double allele loci single allele loci of

593 0.05 which were estimated with MOI and the allele frequencies) [21,66]. Second, the MOI per
594 host was also estimated from the simulated *var* genes data by counting the total number of distinct
595 *var* genes within each host and by dividing it by the size of one repertoire, here 45 as estimated
596 from control data using repeat samplings of the *var* genes of 3D7 with the *var*coding protocol [23].
597
598 To account for measurement error in both approaches, a measurement model was
599 implemented. First, to account for potential SNP genotyping failures, we applied a measurement
600 model that randomly replaces the host genotypes with missing data, reducing the number of
601 available data for THE REAL McCOIL approach (Fig 2). This replacement was implemented by
602 using the distribution of the proportion of missing genotypes per monoclonal infections from a
603 panel of 24 bi-allelic SNP loci which was previously obtained during one cross-sectional survey
604 in 2015 in the Bongo District in Ghana at the end of the wet season [24] (Fig 1C). For each host,
605 some SNP loci were thus replaced with missing genotypes according to a weight reflecting the
606 proportion of missing genotype counts density function. Second, to account for *var* gene potential
607 sampling errors, we applied a measurement model that sub-samples the number of *var* genes per
608 strain, resulting in a reduction of the total number of *var* genes per host (Fig 2). This sub-sampling
609 was implemented by exploring the distribution of the number of non-upsA DBLα *var* gene types
610 per monoclonal infection for which molecular sequences were previously obtained during six
611 cross-sectional surveys between 2012 and 2016 in the Bongo District in Ghana at the end of the
612 wet season [22,23,36,59] (Fig 1D). For each strain, the number of *var* genes was sub-sampled
613 according to a weight reflecting the *var* gene counts density function.
614

MOI estimations were carried out with and without measurement error. To better reflect what is typically done for malaria surveillance, we also estimated the MOI after subsampling the simulated dataset, from 2000 to 500 or 200 individuals. T-tests were used to compare true and estimated MOI distributions. All t-test comparisons were considered statistically significant when P-value ≤ 0.05.

Repertoire similarity networks

To evaluate the similarity of parasites in the population, pairwise type sharing (PTS) was calculated between all repertoire pairs (regardless of the host in which they are encountered) as $PTS_{ij} = 2n_{ij} / (n_i + n_j)$, where n_i and n_j are the number of unique *var* genes within each repertoire i and j and n_{ij} is the total number of *var* genes shared between repertoires i and j [28]. In addition, the genetic structure of the *P. falciparum* population was also analyzed using similarity networks based on *var* composition. Similarity networks were built in which nodes are *var* repertoires, weighted edges encode the degree of overlap between the *var* genes contained in these repertoires, and the direction of an edge indicates the asymmetric competition between repertoires, i.e., whether one repertoire can outcompete the other [36,67]. To introduce directional edges, we calculated the genetic similarity of repertoire i to repertoire j as $S_{ij} = (N_i \cap N_j) / Ni$, where N_i and N_j are the number of unique *var* genes in repertoires i and j, respectively. To focus on the *var* repertoires with the strongest overlap, only the top 1% of edges are drawn and used in network analysis.

Data availability

The agent-based stochastic simulator of malaria dynamics and the processing scripts to reproduce all the figures are stored and annotated on GitHub: https://github.com/pascualgroup/varmodel3. The SNP data used for this analysis are available in Dryad at https://doi.org/10.5061/dryad.jsxksn0bp. The DBLα sequences used for this analysis are available in GenBank under BioProject Number: PRJNA 396962.

Acknowledgments

This research was initially supported by the Fogarty International Center at the National Institutes of Health (Program on the Ecology and Evolution of Infectious Diseases), grant R01-TW009670. Funding was provided by the joint NIH-NSF-NIFA Ecology and Evolution of Infectious Disease award R01-AI149779. We thank the participants, communities, and the Ghana Health Service in BD, Ghana, for their willingness to participate in this study. We would like to acknowledge the programming assistance of Edward B. Baskerville. We appreciate the support of the University of Chicago through the computational resources of the Midway cluster.

References

1. World Health Organization. World malaria report 2021. World Health Organization; 2021.
2. Kolakovich KA, Ssengoba A, Wojcik K, Tsuboi T, Al-Yaman F, Alpers M, et al. *Plasmodium vivax*:favored gene frequencies of the merozoite surface protein-1 and the multiplicity of infection in a malaria endemic region. Experimental Parasitology. 1996;83: 11–18. doi:10.1006/expr.1996.0044
3. Conway DJ, Roper C, Oduola AMJ, Arnot DE, Kremsner PG, Grobusch MP, et al. High recombination rate in natural populations of *Plasmodium falciparum*. Proceedings of the National Academy of Sciences. 1999;96: 4506–4511. doi:10.1073/pnas.96.8.4506
4. Konaté L, Zwetyenga J, Rogier C, Bischoff E, Fontenille D, Tall A, et al. 5. Variation of *Plasmodium falciparum* msp1 block 2 and msp2 allele prevalence and of infection complexity in two neighbouring Senegalese villages with different transmission conditions. Transactions of The Royal Society of Tropical Medicine and Hygiene. 1999;93: 21–28. doi:10.1016/S0035-9203(99)90323-1
5. Anderson TJC, Haubold B, Williams JT, Estrada-Franco§ JG, Richardson L, Mollinedo R, et al. Microsatellite markers reveal a spectrum of population structures in the malaria parasite *Plasmodium falciparum*. Molecular Biology and Evolution. 2000;17: 1467–1482. doi:10.1093/oxfordjournals.molbev.a026247
6. Conway DJ. Molecular epidemiology of malaria. Clinical Microbiology Reviews. 2007;20: 188–204. doi:10.1128/CMR.00021-06
7. Felger I, Tavul L, Kabintik S, Marshall V, Genton B, Alpers M, et al. *Plasmodium falciparum*: extensive polymorphism in merozoite surface antigen 2 alleles in an area with

endemic malaria in Papua New Guinea. Experimental Parasitology. 1994;79: 106–116. doi:10.1006/expr.1994.1070

8. Morlais I, Nsango SE, Toussile W, Abate L, Annan Z, Tchioffo MT, et al. *Plasmodium falciparum* mating patterns and mosquito infectivity of natural isolates of gametocytes. PLOS ONE. 2015;10: e0123777. doi:10.1371/journal.pone.0123777

9. Mwangi JM, Omar SA, Ranford-Cartwright LC. Comparison of microsatellite and antigen-coding loci for differentiating recrudescing *Plasmodium falciparum* infections from reinfections in Kenya. International Journal for Parasitology. 2006;36: 329–336. doi:10.1016/j.ijpara.2005.10.013

10. Su X, Wellems TE. Toward a high-resolution *Plasmodium falciparum* linkage map: polymorphic markers from hundreds of simple sequence repeats. Genomics. 1996;33: 430–444. doi:10.1006/geno.1996.0218

11. Sutton PL, Torres LP, Branch OH. Sexual recombination is a signature of a persisting malaria epidemic in Peru. Malaria Journal. 2011;10: 329. doi:10.1186/1475-2875-10-329

12. Zhong D, Afrane Y, Githeko A, Yang Z, Cui L, Menge DM, et al. *Plasmodium falciparum* genetic diversity in western Kenya highlands. Am J Trop Med Hyg. 2007;77: 1043–1050.

13. Brody JR, Calhoun ES, Gallmeier E, Creavalle TD, Kern SE. Ultra-fast high-resolution agarose electrophoresis of DNA and RNA using low-molarity conductive media. BioTechniques. 2004;37: 598–602. doi:10.2144/04374ST04

14. Contamin H, Fandeur T, Bonnefoy S, Skouri F, Ntoumi F, Mercereau-Puijalon O. PCR typing of field isolates of *Plasmodium falciparum*. J Clin Microbiol. 1995;33: 944–951. doi:10.1128/jcm.33.4.944-951.1995

15. Gupta V, Dorsey G, Hubbard AE, Rosenthal PJ, Greenhouse B. Gel versus capillary electrophoresis genotyping for categorizing treatment outcomes in two anti-malarial trials in Uganda. Malar J. 2010;9: 19. doi:10.1186/1475-2875-9-19

16. Assefa SA, Preston MD, Campino S, Ocholla H, Sutherland CJ, Clark TG. estMOI: estimating multiplicity of infection using parasite deep sequencing data. Bioinformatics. 2014;30: 1292–1294. doi:10.1093/bioinformatics/btu005

17. Hill WG, Babiker HA. Estimation of numbers of malaria clones in blood samples. Proceedings of the Royal Society of London Series B: Biological Sciences. 1995;262: 249–257. doi:10.1098/rspb.1995.0203

18. O'Brien JD, Amenga-Etego L, Li R. Approaches to estimating inbreeding coefficients in clinical isolates of *Plasmodium falciparum* from genomic sequence data. Malaria Journal. 2016;15: 473. doi:10.1186/s12936-016-1531-z

19. Zhu SJ, Almagro-Garcia J, McVean G. Deconvolution of multiple infections in *Plasmodium falciparum* from high throughput sequencing data. Bioinformatics. 2018;34: 9–15. doi:10.1093/bioinformatics/btx530

20. Lerch A, Koepfli C, Hofmann NE, Messerli C, Wilcox S, Kattenberg JH, et al. Development of amplicon deep sequencing markers and data analysis pipeline for genotyping multi-clonal malaria infections. BMC Genomics. 2017;18: 864. doi:10.1186/s12864-017-4260-y

21. Chang H-H, Worby CJ, Yeka A, Nankabirwa J, Kamya MR, Staedke SG, et al. THE REAL McCOIL: A method for the concurrent estimation of the complexity of infection and SNP allele frequency for malaria parasites. PLOS Computational Biology. 2017;13: e1005348. doi:10.1371/journal.pcbi.1005348

22. Ruybal-Pesántez S, Tiedje KE, Pilosof S, Tonkin-Hill G, He Q, Rask TS, et al. Age-specific patterns of DBLα var diversity can explain why residents of high malaria transmission areas remain susceptible to *Plasmodium falciparum* blood stage infection throughout life. International Journal for Parasitology. 2022 [cited 11 May 2022]. doi:10.1016/j.ijpara.2021.12.001

23. Tiedje KE, Oduro AR, Bangre O, Amenga-Etego L, Dadzie SK, Appawu MA, et al. Indoor residual spraying with a non-pyrethroid insecticide reduces the reservoir of *Plasmodium falciparum* in a high-transmission area in northern Ghana. PLOS Global Public Health. 2022. doi:10.1371/journal.pgph.0000285

24. Daniels R, Volkman SK, Milner DA, Mahesh N, Neafsey DE, Park DJ, et al. A general SNP-based molecular barcode for *Plasmodium falciparum* identification and tracking. Malar J. 2008;7: 223. doi:10.1186/1475-2875-7-223

25. Galinsky K, Valim C, Salmier A, de Thoisy B, Musset L, Legrand E, et al. COIL: a methodology for evaluating malarial complexity of infection using likelihood from single nucleotide polymorphism data. Malaria Journal. 2015;14: 4. doi:10.1186/1475-2875-14-4

26. Rask TS, Hansen DA, Theander TG, Pedersen AG, Lavstsen T. *Plasmodium falciparum* erythrocyte membrane protein 1 diversity in seven genomes – divide and conquer. PLOS Computational Biology. 2010;6: e1000933. doi:10.1371/journal.pcbi.1000933

27. Bull PC, Lowe BS, Kortok M, Molyneux CS, Newbold CI, Marsh K. Parasite antigens on the infected red cell surface are targets for naturally acquired immunity to malaria. Nat Med. 1998;4: 358–360. doi:10.1038/nm0398-358

28. Barry AE, Leliwa-Sytek A, Tavul L, Imrie H, Migot-Nabias F, Brown SM, et al. Population genomics of the immune evasion (*var*) genes of *Plasmodium falciparum*. PLOS Pathogens. 2007;3: e34. doi:10.1371/journal.ppat.0030034

29. Chen DS, Barry AE, Leliwa-Sytek A, Smith T-A, Peterson I, Brown SM, et al. A molecular epidemiological study of *var* gene diversity to characterize the reservoir of *Plasmodium falciparum* in Humans in Africa. PLOS ONE. 2011;6: e16629. doi:10.1371/journal.pone.0016629

30. Day KP, Artzy-Randrup Y, Tiedje KE, Rougeron V, Chen DS, Rask TS, et al. Evidence of strain structure in *Plasmodium falciparum var* gene repertoires in children from Gabon, West Africa. PNAS. 2017;114: E4103–E4111. doi:10.1073/pnas.1613018114

31. Ruybal-Pesántez S, Tiedje KE, Tonkin-Hill G, Rask TS, Kamya MR, Greenhouse B, et al. Population genomics of virulence genes of *Plasmodium falciparum* in clinical isolates from Uganda. Sci Rep. 2017;7: 11810. doi:10.1038/s41598-017-11814-9

32. Tessema SK, Nakajima R, Jasinskas A, Monk SL, Lekieffre L, Lin E, et al. Protective immunity against severe malaria in children is associated with a limited repertoire of antibodies to conserved PfEMP1 variants. Cell Host & Microbe. 2019;26: 579-590.e5. doi:10.1016/j.chom.2019.10.012

33. Tonkin-Hill G, Ruybal-Pesántez S, Tiedje KE, Rougeron V, Duffy MF, Zakeri S, et al. Evolutionary analyses of the major variant surface antigen-encoding genes reveal population structure of *Plasmodium falciparum* within and between continents. PLOS Genetics. 2021;17: e1009269. doi:10.1371/journal.pgen.1009269

34. Scherf A, Hernandez-Rivas R, Buffet P, Bottius E, Benatar C, Pouvelle B, et al. Antigenic variation in malaria: in situ switching, relaxed and mutually exclusive transcription of var

genes during intra-erythrocytic development in *Plasmodium falciparum*. EMBO J. 1998;17: 5418–5426. doi:10.1093/emboj/17.18.5418

35. Su X, Heatwole VM, Wertheimer SP, Guinet F, Herrfeldt JA, Peterson DS, et al. The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of *Plasmodium falciparum*-infected erythrocytes. Cell. 1995;82: 89–100. doi:10.1016/0092-8674(95)90055-1

36. He Q, Pilosof S, Tiedje KE, Ruybal-Pesántez S, Artzy-Randrup Y, Baskerville EB, et al. Networks of genetic similarity reveal non-neutral processes shape strain structure in *Plasmodium falciparum*. Nat Commun. 2018;9: 1817. doi:10.1038/s41467-018-04219-3

37. Gupta S, Ferguson N, Anderson R. Chaos, persistence, and evolution of strain structure in antigenically diverse infectious agents. Science. 1998;280: 912–915. doi:10.1126/science.280.5365.912

38. Auburn S, Campino S, Miotto O, Djimde AA, Zongo I, Manske M, et al. Characterization of within-host *Plasmodium falciparum* diversity using next-generation sequence data. PLOS ONE. 2012;7: e32891. doi:10.1371/journal.pone.0032891

39. Pamilo P, Varvio-Aho S-L. On the estimation of population size from allele frequency changes. Genetics. 1980;95: 1055–1057.

40. Weir BS, Cockerham CC. Estimating F-statistics for the analysis of population structure. Evolution. 1984;38: 1358–1370. doi:10.2307/2408641

41. Peyerl-Hoffmann G, Jelinek T, Kilian A, Kabagambe G, Metzger WG, Von Sonnenburg F. Genetic diversity of *Plasmodium falciparum* and its relationship to parasite density in an area with different malaria endemicities in West Uganda. Tropical Medicine & International Health. 2001;6: 607–613. doi:10.1046/j.1365-3156.2001.00761.x

786 42. Chen I, Clarke SE, Gosling R, Hamainza B, Killeen G, Magill A, et al. "Asymptomatic"
787 malaria: a chronic and debilitating infection that should be treated. PLOS Medicine. 2016;13:
788 e1001942. doi:10.1371/journal.pmed.1001942
789 43. Okell LC, Bousema T, Griffin JT, Ouédraogo AL, Ghani AC, Drakeley CJ. Factors
790 determining the occurrence of submicroscopic malaria infections and their relevance for
791 control. Nat Commun. 2012;3: 1237. doi:10.1038/ncomms2241
792 44. Farnert A, Snounou G, Rooth I, Bjorkman A. Daily dynamics of*Plasmodium falciparum*
793 subpopulations in asymptomatic children in a holoendemic area. Am J Trop Med Hyg.
794 1997;56: 538–547. doi:10.4269/ajtmh.1997.56.538
795 45. Bruce MC, Donnelly CA, Packer M, Lagog M, Gibson N, Narara A, et al. Age- and species-
796 specific duration of infection in asymptomatic malaria infections in Papua New Guinea.
797 Parasitology. 2000;121 (Pt 3): 247–256. doi:10.1017/s0031182099006344
798 46. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, et al. Genome sequence of
799 the human malaria parasite *Plasmodium falciparum*. Nature. 2002;419: 498–511.
800 doi:10.1038/nature01097
801 47. Lavstsen T, Salanti A, Jensen AT, Arnot DE, Theander TG. Sub-grouping of *Plasmodium*
802 *falciparum* 3D7 var genes based on sequence analysis of coding and non-coding regions.
803 Malar J. 2003;2: 27. doi:10.1186/1475-2875-2-27
804 48. Kraemer SM, Kyes SA, Aggarwal G, Springer AL, Nelson SO, Christodoulou Z, et al.
805 Patterns of gene recombination shape *var* gene repertoires in *Plasmodium falciparum*:
806 comparisons of geographically diverse isolates. BMC Genomics. 2007;8: 45.
807 doi:10.1186/1471-2164-8-45

808 49. Falk N, Kaestli M, Qi W, Ott M, Baea K, Corteés A, et al. Analysis of *Plasmodium*
809 *falciparum var* genes expressed in children from Papua New Guinea. The Journal of
810 Infectious Diseases. 2009;200: 347–356. doi:10.1086/600071
811 50. Jensen ATR, Magistrado P, Sharp S, Joergensen L, Lavstsen T, Chiucchiuini A, et al.
812 *Plasmodium falciparum* associated with severe childhood malaria preferentially expresses
813 PfEMP1 encoded by group A *var* genes. Journal of Experimental Medicine. 2004;199: 1179–
814 1190. doi:10.1084/jem.20040274
815 51. Kaestli M, Cockburn IA, Cortés A, Baea K, Rowe JA, Beck H-P. Virulence of malaria is
816 associated with differential expression of *Plasmodium falciparum var* gene subgroups in a
817 case-control study. The Journal of Infectious Diseases. 2006;193: 1567–1574.
818 doi:10.1086/503776
819 52. Kalmbach Y, Rottmann M, Kombila M, Kremsner PG, Beck H-P, Kun JFJ. Differential *var*
820 gene expression in children with malaria and antidromic effects on host gene expression. The
821 Journal of Infectious Diseases. 2010;202: 313–317. doi:10.1086/653586
822 53. Kyriacou HM, Stone GN, Challis RJ, Raza A, Lyke KE, Thera MA, et al. Differential *var*
823 gene transcription in *Plasmodium falciparum* isolates from patients with cerebral malaria
824 compared to hyperparasitaemia. Molecular and Biochemical Parasitology. 2006;150: 211–
825 218. doi:10.1016/j.molbiopara.2006.08.005
826 54. Normark J, Nilsson D, Ribacke U, Winter G, Moll K, Wheelock CE, et al. PfEMP1-DBL1α
827 amino acid motifs in severe disease states of *Plasmodium falciparum* malaria. PNAS.
828 2007;104: 15835–15840. doi:10.1073/pnas.0610485104
829 55. Rottmann M, Lavstsen T, Mugasa JP, Kaestli M, Jensen ATR, Müller D, et al. Differential
830 expression of *var* gene groups is associated with morbidity caused by *Plasmodium*

falciparum infection in Tanzanian children. Infection and Immunity. 2006 [cited 20 Jan 2022]. doi:10.1128/IAI.02073-05

56. Warimwe GM, Fegan G, Musyoki JN, Newton CRJC, Opiyo M, Githinji G, et al. Prognostic indicators of life-threatening malaria are associated with distinct parasite variant antigen profiles. Science Translational Medicine. 2012 [cited 20 Jan 2022]. doi:10.1126/scitranslmed.3003247

57. Warimwe GM, Keane TM, Fegan G, Musyoki JN, Newton CRJC, Pain A, et al. *Plasmodium falciparum var* gene expression is modified by host immunity. PNAS. 2009;106: 21801–21806. doi:10.1073/pnas.0907590106

58. Tiedje KE, Oduro AR, Agongo G, Anyorigiya T, Azongo D, Awine T, et al. Seasonal variation in the epidemiology of asymptomatic *Plasmodium falciparum* infections across two catchment areas in Bongo District, Ghana. The American Journal of Tropical Medicine and Hygiene. 2017;97: 199–212. doi:10.4269/ajtmh.16-0959

59. Pilosof S, He Q, Tiedje KE, Ruybal-Pesántez S, Day KP, Pascual M. Competition for hosts modulates vast antigenic diversity to generate persistent strain structure in *Plasmodium falciparum*. PLOS Biology. 2019;17: e3000336. doi:10.1371/journal.pbio.3000336

60. Gibson MA, Bruck J. Efficient Exact Stochastic Simulation of Chemical Systems with Many Species and Many Channels. J Phys Chem A. 2000;104: 1876–1889. doi:10.1021/jp993732q

61. Gillespie DT. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. Journal of Computational Physics. 1976;22: 403–434. doi:10.1016/0021-9991(76)90041-3

852 62. Buckee CO, Recker M. Evolution of the multi-domain structures of virulence genes in the
853 Human malaria parasite, *Plasmodium falciparum*. PLOS Computational Biology. 2012;8:
854 e1002451. doi:10.1371/journal.pcbi.1002451
855 63. Bull PC, Buckee CO, Kyes S, Kortok MM, Thathy V, Guyah B, et al. *Plasmodium falciparum*
856 antigenic variation. Mapping mosaic *var* gene sequences onto a network of shared, highly
857 polymorphic sequence blocks. Molecular Microbiology. 2008;68: 1519–1534.
858 doi:10.1111/j.1365-2958.2008.06248.x
859 64. White MT, Griffin JT, Churcher TS, Ferguson NM, Basáñez M-G, Ghani AC. Modelling the
860 impact of vector control interventions on *Anopheles gambiae* population dynamics. Parasites
861 Vectors. 2011;4: 153. doi:10.1186/1756-3305-4-153
862 65. Organization WH. A framework for malaria elimination. World Health Organization; 2017.
863 66. R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R
864 Foundation for Statistical Computing; 2019. Available: https://www.R-project.org
865 67. He Q, Pilosof S, Tiedje KE, Day KP, Pascual M. Frequency-dependent competition between
866 strains imparts persistence to perturbations in a model of Plasmodium falciparum malaria
867 transmission. Front Ecol Evol. 2021;9. doi:10.3389/fevo.2021.633263
868

Supporting information captions

S1 Fig. Host age distribution, and number of sampled individuals and hosts per age class. For each category, the horizontal central solid line represents the median, the diamond represents the mean, the box represents the interquartile range (IQR) from the 25th to 75th centiles, the whiskers indicate the most extreme data point which is no more than 1.5 times the interquartile range from the box, and the dots show the outliers, i.e. the points beyond the whiskers. **A)** Age distribution of the sampled individuals. **B)** Number of sampled individuals per age class. **C)** Number of sampled hosts per age class. Upper (yellow), middle (orange), and lower (purple) panels correspond to simulations under low-, moderate-, and high-transmission settings, respectively (S1 and S2 Tables). Values were split into five age classes, i.e. 0-5, 6-10, 11-20, 21-39, and $\geq$ 40 years.

S2 Fig. Prevalence and entomological inoculation rate (EIR) per transmission intensity. For each category, the horizontal central solid line represents the median, the diamond represents the mean, the box represents the interquartile range (IQR) from the 25th to 75th centiles, the whiskers indicate the most extreme data point which is no more than 1.5 times the interquartile range from the box, and the dots show the outliers, i.e. the points beyond the whiskers. **A)** Prevalence; **B)** EIR. Statistics calculated for simulations under low-, moderate-, and high-transmission settings are indicated in yellow, orange, and purple, respectively (S1 and S2 Tables).

S3 Fig. Initial number of SNPs and accuracy of the multiplicity of infection (MOI) estimates determined with THE REAL McCOIL approach. The accuracy of MOI estimates is defined as

the differences between estimated and true MOI per host. While null values highlight accurate MOI estimates (indicated by a dashed black horizontal line), the positive and negative values highlight over- and under-estimation, respectively. The dark and light green colors indicate respectively MOI estimations made without and with a measurement model (Fig 2). For each category, the horizontal central solid line represents the median, the diamond represents the mean, the box represents the interquartile range (IQR) from the 25th to 75th centiles, the whiskers indicate the most extreme data point which is no more than 1.5 times the interquartile range from the box, and the dots show the outliers, i.e. the points beyond the whiskers. **A)** Accuracy of MOI estimates per true MOI. **B)** Accuracy of MOI estimates per transmission intensity (S1 and S2 Tables).

S4 Fig. SNP properties and accuracy of the multiplicity of infection (MOI) estimates determined with THE REAL McCOIL approach. The accuracy of MOI estimates is defined as the differences between estimated and true MOI per host. While null values highlight accurate MOI estimates (indicated by a dashed black horizontal line), the positive and negative values highlight over- and under-estimation, respectively. For each category, the horizontal central solid line represents the median, the diamond represents the mean, the box represents the interquartile range (IQR) from the 25th to 75th centiles, the whiskers indicate the most extreme data point which is no more than 1.5 times the interquartile range from the box, and the dots show the outliers, i.e. the points beyond the whiskers. The dark and light green colors indicate respectively MOI estimations made without and with a measurement model (Fig 2).

S5 Fig: Reliability of the multiplicity of infection (MOI) estimations when subsampling 25% of the sampled individuals (i.e. 500 individuals). For each category, the horizontal central solid line represents the median, the diamond represents the mean, the box represents the interquartile range (IQR) from the 25th to 75th centiles, the whiskers indicate the most extreme data point which is no more than 1.5 times the interquartile range from the box, and the dots show the outliers, i.e. the points beyond the whiskers. The upper, middle, and lower row panels correspond to simulations under low-, moderate-, and high-transmission settings, respectively (S1 and S2 Tables). **A)** Accuracy of MOI estimates, defined as the difference between estimated and true MOI per host. While null values highlight accurate MOI estimates (indicated by a dashed black horizontal line), the positive and negative values highlight over- and under-estimation, respectively. Estimates with the neutral SNP-based approach (THE REAL McCOIL) are indicated in green, and those with the *var* gene-based approach (*var*coding) are indicated in blue. The dark and light green or blue colors indicate respectively MOI estimations made without and with a measurement model (Fig 2). The column panels show differences for specific true MOI values. **B)** Population distribution of the estimated and true MOI per host from the simulated "true" values and those estimated with the methods indicated by the colors similar to panel A.

S6 Fig: Reliability of the multiplicity of infection (MOI) estimations when subsampling 10% of the sampled individuals (i.e. 200 individuals). For each category, the horizontal central solid line represents the median, the diamond represents the mean, the box represents the interquartile range (IQR) from the 25th to 75th centiles, the whiskers indicate the most extreme data point which is no more than 1.5 times the interquartile range from the box, and the dots show the outliers, i.e. the points beyond the whiskers. The upper, middle, and lower row panels correspond to

simulations under low-, moderate-, and high-transmission settings, respectively (S1 and S2 Tables). **A)** Accuracy of MOI estimates, defined as the difference between estimated and true MOI per host. While null values highlight accurate MOI estimates (indicated by a dashed black horizontal line), the positive and negative values highlight over- and under-estimation, respectively. Estimates with the neutral SNP-based approach (THE REAL McCOIL) are indicated in green, and those with the *var* gene-based approach (*var*coding) are indicated in blue. The dark and light green or blue colors indicate respectively MOI estimations made without and with a measurement model (Fig 2). The column panels show differences for specific true MOI values. **B)** Population distribution of the estimated and true MOI per host from the simulated "true" values and those estimated with the methods indicated by the colors similar to panel A.

S7 Fig. Accuracy of the minor allele frequency (MAF) estimates per locus determined with THE REAL McCOIL approach. The accuracy of MAF estimates per locus is defined as the differences between estimated and true MAF per locus. While null values highlight accurate MAF estimates per locus (indicated by a dashed black horizontal line), the positive and negative values highlight over- and under-estimation, respectively. For each category, the horizontal central solid line represents the median, the diamond represents the mean, the box represents the interquartile range (IQR) from the 25th to 75th centiles, the whiskers indicate the most extreme data point which is no more than 1.5 times the interquartile range from the box, and the dots show the outliers, i.e. the points beyond the whiskers. The dark and light green colors indicate respectively MAF estimations made without and with a measurement model (Fig 2). Upper, middle, and lower panels correspond to simulations under low-, moderate-, and high-transmission settings, respectively (S1 and S2 Tables).

960

961 **S8 Fig. Population structure using repertoire similarity network properties.** Comparisons of
962 repertoire similarity networks of 150 randomly sampled parasite *var* repertoires generated from a
963 one-time point under low, moderate, and high-transmission settings (S1 and S2 Tables). Only the
964 top 1% of edges are drawn and used in the analysis. The upper panel shows the distribution of the
965 mean pairwise type sharing (PTS) per run. For each category, the horizontal central solid line
966 represents the median, the diamond represents the mean, the box represents the interquartile range
967 (IQR) from the 25th to 75th centiles, the whiskers indicate the most extreme data point which is
968 no more than 1.5 times the interquartile range from the box, and the dots show the outliers, i.e. the
969 points beyond the whiskers. The lower panel shows the distributions of the proportion of
970 occurrences of three-node graph motifs across the repertoire similarity networks.

971

972 **S9 Fig. Pairwise type sharing (PTS).** For each category, the horizontal central solid line
973 represents the median, the diamond represents the mean, the box represents the interquartile range
974 (IQR) from the 25th to 75th centiles, and the whiskers indicate the most extreme data point. **A)**
975 Distribution of the PTS per transmission intensity. **B)** Distribution of the PTS per run.

976

977 **S10 Fig. Reliability of the multiplicity of infection (MOI) estimations when simulations**
978 **include a measurement error based on the distribution of the number of non-upsA DBL*α***
979 ***var* gene types per 3D7 laboratory isolates for the *var* coding approach.** For each category, the
980 horizontal central solid line represents the median, the diamond represents the mean, the box
981 represents the interquartile range (IQR) from the 25th to 75th centiles, the whiskers indicate the
982 most extreme data point which is no more than 1.5 times the interquartile range from the box, and

the dots show the outliers, i.e. the points beyond the whiskers. The upper, middle, and lower row panels correspond to simulations under low-, moderate-, and high-transmission settings, respectively (S1 and S2 Tables). **A)** Accuracy of MOI estimates, defined as the differences between estimated and true MOI per host. While null values highlight accurate MOI estimates (indicated by a dashed black horizontal line), the positive and negative values highlight over- and under-estimation, respectively. Estimates with the neutral SNP-based approach (THE REAL McCOIL) are indicated in green, and those with the *var* gene-based approach (*var*coding) are indicated in blue. The dark and light blue or green colors indicate respectively MOI estimates made without and with a measurement model (Fig 2). The column panels show differences for specific true MOI values. **B)** Population distribution of the estimated and true MOI per host from the simulated "true" values and those estimated with the methods indicated by the colors similar to panel A.

S11 Fig. Reliability of the multiplicity of infection (MOI) estimations when THE REAL McCOIL approach using an upper bound for MOI of 5, 10, and 20 for the low-, moderate-, and high-transmission simulations, respectively. For each category, the horizontal central solid line represents the median, the diamond represents the mean, the box represents the interquartile range (IQR) from the 25th to 75th centiles, the whiskers indicate the most extreme data point which is no more than 1.5 times the interquartile range from the box, and the dots show the outliers, i.e. the points beyond the whiskers. The upper, middle, and lower row panels correspond to simulations under low-, moderate-, and high-transmission settings, respectively (S1 and S2 Tables). **A)** Accuracy of MOI estimates, defined as the differences between estimated and true MOI per host. While null values highlight accurate MOI estimates (indicated by a dashed black

horizontal line), the positive and negative values highlight over- and under-estimation, respectively. The estimated MOI using the *var* genes based approach (i.e. *var* coding) are indicated in blue, and the estimated MOI using the neutral SNPs based approach (i.e. THE REAL McCOIL) are indicated in green. The dark and light blue or green colors indicate respectively MOI estimates made without and with a measurement model (Fig 2). The column panels show differences for specific true MOI values. **B)** Population distribution of the estimated and true MOI per host from the simulated "true" values and those estimated with the methods indicated by the colors similar to panel A.

S1 Table. Epidemiological and genetic parameters used in the stochastic simulations.

S2 Table. Epidemiological and genetic distinct parameters per run.

S3 Table. Pearson correlation coefficients between the inaccuracy of the minor allele frequency (MAF) per locus estimated with THE REAL McCOIL approach (defined as the absolute differences between estimated and true MAF per locus), and the locus properties.