Title

THE REAL McCOIL: A method for the concurrent estimation of the complexity of infection and SNP allele frequency for malaria parasites

Journal

PLOS Computational Biology, 13(1)

ISSN

1553-734X

Authors

Chang, Hsiao-Han
Worby, Colin J
Yeka, Adoke
et al.

Publication Date

2017

DOI

10.1371/journal.pcbi.1005348

Peer reviewed

PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

THE REAL McCOIL: A method for the concurrent estimation of the complexity of infection and SNP allele frequency for malaria parasites

Hsiao-Han Chang[1]*, Colin J. Worby[1], Adoke Yeka[2,3], Joaniter Nankabirwa[3,4], Moses R. Kamya[3,4], Sarah G. Staedke[5], Grant Dorsey[6], Maxwell Murphy[6], Daniel E. Neafsey[7], Anna E. Jeffreys[8], Christina Hubbart[8], Kirk A. Rockett[8,9], Roberto Amato[9], Dominic P. Kwiatkowski[8,9], Caroline O. Buckee[1☯], Bryan Greenhouse[6☯]

1 Center for Communicable Disease Dynamics, Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States, **2** Makerere University School of Public Health, College of Health Sciences, Kampala, Uganda, **3** Infectious Disease Research Collaboration, Kampala, Uganda, **4** Department of Medicine, Makerere University College of Health Sciences, Kampala, Uganda, **5** London School of Hygiene and Tropical Medicine, London, United Kingdom, **6** Department of Medicine, University of California, San Francisco, San Francisco, California, United States, **7** Genome Sequencing and Analysis Program, Broad Institute, Cambridge, Massachusetts, United States, **8** Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom, **9** Wellcome Trust Sanger Institute, Cambridge, United Kingdom

☯ These authors contributed equally to this work.
* hhchang@hsph.harvard.edu

OPEN ACCESS

Citation: Chang H-H, Worby CJ, Yeka A, Nankabirwa J, Kamya MR, Staedke SG, et al. (2017) THE REAL McCOIL: A method for the concurrent estimation of the complexity of infection and SNP allele frequency for malaria parasites. PLoS Comput Biol 13(1): e1005348. doi:10.1371/journal.pcbi.1005348

Editor: Mercedes Pascual, University of Chicago, UNITED STATES

Received: September 14, 2016

Accepted: January 5, 2017

Published: January 26, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: Research reported in this article was supported by the National Institutes of Health, National Institute of General Medical Sciences (U54GM088558) https://www.nigms.nih.gov, the National Institute of Allergy and Infectious Diseases as part of the International Centers of Excellence in Malaria Research (ICMER) program

Abstract

As many malaria-endemic countries move towards elimination of *Plasmodium falciparum*, the most virulent human malaria parasite, effective tools for monitoring malaria epidemiology are urgent priorities. *P. falciparum* population genetic approaches offer promising tools for understanding transmission and spread of the disease, but a high prevalence of multiclone or polygenomic infections can render estimation of even the most basic parameters, such as allele frequencies, challenging. A previous method, *COIL*, was developed to estimate complexity of infection (COI) from single nucleotide polymorphism (SNP) data, but relies on monogenomic infections to estimate allele frequencies or requires external allele frequency data which may not available. Estimates limited to monogenomic infections may not be representative, however, and when the average COI is high, they can be difficult or impossible to obtain. Therefore, we developed *THE REAL McCOIL*, Turning HEterozygous SNP data into Robust Estimates of ALelle frequency, via Markov chain Monte Carlo, and Complexity Of Infection using Likelihood, to incorporate polygenomic samples and simultaneously estimate allele frequency and COI. This approach was tested via simulations then applied to SNP data from cross-sectional surveys performed in three Ugandan sites with varying malaria transmission. We show that *THE REAL McCOIL* consistently outperforms *COIL* on simulated data, particularly when most infections are polygenomic. Using field data we show that, unlike with *COIL*, we can distinguish epidemiologically relevant differences in COI between and within these sites. Surprisingly, for example, we estimated high average COI in a peri-urban subregion with lower transmission intensity, suggesting that many of

(U19AI089674) https://www.niaid.nih.gov/, and the Bill and Melinda Gates Foundation (OPP1132226) http://www.gatesfoundation.org/. The sequencing, analysis, informatics and management of the MalariaGEN Community Project and Pf3k Project are supported by the Wellcome Trust through Sanger Institute core funding (098051) and a Strategic Award (090770/Z/09/Z) https://wellcome.ac.uk/. JN is supported by the NURTURE which is funded by the National Institutes of Health (D43TW010132) https://www.nih.gov/. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funders. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

these cases were imported from surrounding regions with higher transmission intensity. *THE REAL McCOIL* therefore provides a robust tool for understanding the molecular epidemiology of malaria across transmission settings.

Author Summary

Monitoring malaria epidemiology is critical for evaluating the impact of interventions and designing strategies for control and elimination. Population genetics has been used to inform malaria epidemiology, but it is limited by the fact that a fundamental metric needed for most analyses—the frequency of alleles in a population—is difficult to estimate from blood samples containing more than one genetically distinct parasite (polygenomic infections). A widely used approach has been to restrict analysis to monogenomic infections, which may represent a biased subset and potentially ignores a large amount of data. Therefore, we developed a new analytical approach that uses data from all infections to simultaneously estimate allele frequency and the number of distinct parasites within each infection. The method, called *THE REAL McCOIL*, was evaluated using simulations and was then applied to data from cross-sectional surveys performed in three regions of Uganda. Simulations demonstrated accurate performance, and analyses of samples from Uganda using *THE REAL McCOIL* revealed epidemiologically relevant differences within and between the three regions that previous methods could not. *THE REAL McCOIL* thus facilitates population genetic analysis when there are polygenomic infections, which are common in many malaria endemic areas.

This is a *PLOS Computational Biology* Methods Paper.

Introduction

Malaria has declined significantly over the past decade, but continues to cause half a million deaths annually [1]. Calls for elimination have shifted research efforts towards developing new approaches for transmission reduction, including the identification of source and sink regions and hotspots that sustain transmission [2–4]. *Plasmodium falciparum* population genetic tools are increasingly being used to inform these efforts [5–12] and have been proposed as a means to establish the direction of parasite flows and to determine elimination status both by identifying the source of imported infections and by establishing that no local transmission is occurring [13–17]. However, in malaria-endemic regions, infections are frequently characterized by multiple different genotypes (polygenomic infections), which makes interpreting genetic data challenging. As a result, population genetic analyses of malaria parasites have often been limited to monogenomic infections, greatly reducing the utility of available data and potentially introducing biases into results.

Rapid technological developments have led to a proliferation of approaches for characterizing malaria parasite genomes, each with different implications for cost, suitability for field samples across a range of transmission settings, and applicability to different research questions [5,18–23]. Many genotyping approaches are based on a small number of single nucleotide

polymorphisms (SNPs). SNP data are cheap and straightforward to obtain from commonly used dried blood spot (DBS) samples, collected in a variety of field settings, and remain the most common approach for genotyping studies. However, a high prevalence of polygenomic infections can render estimation of even the most basic parameters from SNP data, such as population allele frequencies, difficult.

Population allele frequencies are usually estimated from monogenomic infections [6,7,24], because of the challenge of estimating the true proportion of each lineage from heterozygous SNP loci resulting from high complexity of infection (COI, the number of clones in an individual). However, constraining data sets to only monogenomic infections may introduce systematic biases because these infections may not be representative. Such constraint also greatly limits the precision of estimates when the majority of samples are polygenomic. It is common to use the proportion of heterozygous calls in each individual or the fraction of polygenomic infections to compare genetic diversity between populations [6,7,16,25–27]. However, the complexity of infection underlying polygenomic infections can vary dramatically, and the probability of a particular locus being heterozygous will depend on its allele frequency in the population. *COIL* (estimating *COI* using likelihood), was recently developed to provide a more quantitative measure of genetic diversity [28], but unless supplied with external allele frequency data, relies on monogenomic infections to estimate allele frequencies and is therefore problematic when a large fraction of infections are polygenomic. While external allele frequency data can be obtained from parasite population genomic data such as the Pf3K project (http://www.malariagen.net/projects/pf3k), these estimates are only available in specific locations, and may exhibit considerable heterogeneity in space and time.

Here we introduce a new Bayesian approach, Turning HEterozygous SNP data into Robust Estimates of ALelle frequency, via Markov chain Monte Carlo, and Complexity Of Infection using Likelihood (*THE REAL McCOIL)*, to additionally incorporate polygenomic samples, using Markov chain Monte Carlo methods to simultaneously estimate allele frequency and COI. We tested two versions of our method on a series of simulations and then applied it to data on 105 SNP loci in 868 samples from cross-sectional surveys in three regions of varying endemicity in Uganda [29–31]. The allele frequencies estimated by our new approach were used to calculate F_{ST} [32], a measure of genetic differentiation between sites, and F_{WS} [33], a measure of the within-host genetic diversity. These results demonstrate the utility of *THE REAL McCOIL* to obtain accurate estimates of COI and allele frequency from SNP data, which can be used to characterize genetic diversity and perform population genetic analyses of parasite populations even in very high transmission settings.

Materials and methods

Ethics statement

The cross sectional survey was approved by IRBs at the University of California, San Francisco (#11–07138) and SOMREC at Makerere University, Uganda (#2011–203).

Methods to estimate population allele frequency and complexity of infection

We developed a Markov chain Monte Carlo (MCMC) method to simultaneously estimate population allele frequency for each SNP and COI for each individual. Since estimating COI and allele frequencies are highly related to each other, our approach explored the uncertainty of both at the same time, and by doing so, incorporated information from polygenomic infections. Assuming there are n individuals and k loci, the parameters to be estimated include

complexity of infection for each individual ($M = [m_1, m_2, \ldots, m_n]$) and population allele frequency for each locus ($P = [p_1, p_2, \ldots, p_k]$). We used the data in two ways: a categorical method, in which we considered SNP at locus j of individual i, B_{ij}, to be heterozygous or homozygous (0 [homozygous minor allele], 0.5 [heterozygous], 1 [homozygous major allele]), and a proportional method, in which the proportion of major allele at locus j of individual i, S_{ij}, was calculated from the relative signal intensity for each allele ($S_{ij} = \frac{A_{1ij}}{A_{1ij}+A_{2ij}}$, where A_1 and A_2 represent the signal intensity of major and minor allele that are obtained from Sequenom or similar types of SNP assays, respectively [34]). The notations are summarized in Table A in S1 File. Similar to *COIL*, *THE REAL McCOIL* assumed that different loci are independent, that different samples are independent and polygenomic infections are obtained from multiple independent infections, and that the samples were collected from a single homogeneous population.

Categorical method: Modeling heterozygous/homozygous calls. The likelihood of observing heterozygous/homozygous calls depends on COI, population allele frequency, and the probability of erroneously calling homozygous loci heterozygous (e_1) and conversely calling heterozygous loci homozygous (e_2). We have

$$L(M, P|B_O) = P(B_O|M, P) = \prod_{i=1}^{n}\prod_{j=1}^{k}\sum_{B_{T_{ij}}\in\{0,0.5,1\}} P(B_{Oij}|B_{Tij})P(B_{Tij}|m_i, p_j), \quad (1)$$

where B_{Tij} and B_{Oij} represent the true and observed heterozygosity at locus j of individual i (B_{Tij} and $B_{Oij} \in [0, 0.5, 1]$). We specify $P(B_{Oij}|B_{Tij})$ to take the following form (Table 1), depending on the values of B_{Tij} and B_{Oij}:and

$$P(B_{Tij}|m_i, p_j) = \begin{cases} {p_j}^{m_i} & \text{if } B_{Tij} = 1, \\ (1-p_j)^{m_i} & \text{if } B_{Tij} = 0, \\ 1-{p_j}^{m_i}-(1-p_j)^{m_i} & \text{if } B_{Tij} = 0.5. \end{cases} \quad (2)$$

We assumed uniform priors for M and P and updated them sequentially using a Metropolis-Hastings algorithm over $N = 100{,}000$ iterations, excluding the initial burn-in 1000 iterations to obtain the posterior distributions of M and P. If e_1 and e_2 were not pre-specified, *THE REAL McCOIL* estimated their posterior distributions along with M and P. The details of the sampling procedure are described in Text A in S1 File.

Proportional method: Modeling frequency data. The likelihood of obtaining the raw frequency of signals is composed of the observational model (f, the likelihood of observed frequency of signals given true within-host allele frequency) and the likelihood of true within-

Table 1. The observational model for categorical method.

		B_{Tij}		
		0	**0.5**	**1**
B_{Oij}	**0**	$1-e_1$	$e_2/2$	0
	0.5	e_1	$1-e_2$	e_1
	1	0	$e_2/2$	$1-e_1$

doi:10.1371/journal.pcbi.1005348.t001

host allele frequency (g) as follows:

$$L(M,P|S_O) = P(S_O \mid M,P) = \prod_{i=1}^{n}\prod_{j=1}^{k} P(S_{Oij} \mid m_i, p_j)$$
$$= \prod_{i=1}^{n}\prod_{j=1}^{k} \begin{pmatrix} f(S_{Oij} \mid S_{Tij}=0)g(S_{Tij}=0 \mid m_i,p_j) \\ +\int_{0<S_{Tij}<1} f(S_{Oij} \mid S_{Tij})g(S_{Tij} \mid m_i,p_j)d_{S_{Tij}} \\ +f(S_{Oij} \mid S_{Tij}=1)g(S_{Tij}=1 \mid m_i,p_j) \end{pmatrix} \quad (3)$$

where S_{Tij} and S_{Oij} represent the true and observed frequency of major allele at locus j of individual i ($0 \leq S_{Tij}, S_{Oij} \leq 1$). Consistent with other population genetic approaches [35], we assumed that each observation S_{Oij} was drawn from a normal distribution with the mean equal to the true frequency S_{Tij} and variance equal to $\sigma^2 = \frac{\varepsilon_{est}}{\sqrt{{A_{1ij}}^2+{A_{2ij}}^2}}$, where ε_{est} represents the overall level of measurement error. The variance decreased with the intensity of the signal ($I = \sqrt{{A_{1ij}}^2 + {A_{2ij}}^2}$). To exclude the values outside of [0, 1], we assumed point mass at 0 and 1 and their densities were obtained by integrating values from $-\infty$ to 0 and from 1 to ∞, respectively.

That is,

$$f(S_{Oij} \mid S_{Tij}) = \begin{cases} \Phi\left(\frac{-S_{Tij}}{\sigma}\right) & \text{if } S_{Oij}=0 \\ \phi\left(\frac{S_{Oij}-S_{Tij}}{\sigma}\right) & \text{if } 0<S_{Oij}<1 \\ \Phi\left(\frac{S_{Tij}-1}{\sigma}\right) & \text{if } S_{Oij}=1 \end{cases} \quad (4)$$

where Φ and ϕ are the cumulative distribution function and the probability density function of the standard normal distribution.

The density of the true within-host frequency was composed of a continuous distribution and point masses at 0 and 1 as follows:

$$g(S_{Tij} \mid m_i, p_j) = \begin{cases} (1-p_j)^{m_i} & \text{if } S_{Tij}=0 \\ \left({p_j}^{m_i} - (1-p_j)^{m_i}\right)\text{Beta}(S_{T_{ij}}; \alpha_{m_i p_j}, \beta_{m_i p_j}) & \text{if } 0<S_{Tij}<1 \\ {p_j}^{m_i} & \text{if } S_{Tij}=1 \end{cases} \quad (5)$$

where $\text{Beta}(x; \alpha_{m_i,p_j}, \beta_{m_i,p_j})$ denotes the probability density function of the Beta distribution evaluated at x. The shape and scale parameters, $\alpha_{m_i p_j}$ and $\beta_{m_i p_j}$, respectively, depend on the complexity of infection (m_i) and population allele frequency (p_j), and were obtained by fitting the simulated data. We estimated values for $\alpha_{m_i p_j}$ and $\beta_{m_i p_j}$ pre-analysis, using simulated data to fit Beta distributions for a range of values for m_i and p_j. To do this, we simulated the within-host allele frequency distribution for given values of m_i and p_j by sampling a single allele for each infection from a Bernoulli distribution with p_j and mixing these alleles with the relative contributions sampled from a uniform distribution as follows: sampling (m_i −1) numbers from a uniform distribution, ordering these numbers to obtain $u_{(1)}, u_{(2)}, \cdots, u_{(m_i-1)}$, and mixing alleles using the proportions equal to the difference

between them, $u_{(1)} - 0$, $u_{(2)} - u_{(1)}$, $\cdots$, $u_{(m_i-1)} - u_{(m_i-2)}$, $1 - u_{(m_i-1)}$. Biologically, this means the proportion of either lineage can be any value between 0 and 1 with equal probability when $m_i = 2$. We then fit a Beta distribution to the resulting empirical distribution to obtain fitted values $\hat{\alpha}_{m_i p_j}$ and $\hat{\beta}_{m_i p_j}$. We performed this for each combination of m and p, where m ranged from 2 to 25 and p ranged from 0.01 to 0.99. As a continuous variable, we rounded observed values of p to the second decimal point to correspond to our discrete simulation range, selecting the appropriate $\alpha_{m_i p_j}$ and $\beta_{m_i p_j}$ to calculate the likelihood. Fig A in S1 File shows some examples of the distribution of simulated within-host allele frequencies with the fitted Beta distribution given m and p. While the fitted Beta parameters were obtained by simulating the ratio of mixing from a uniform distribution, the method performed well when the ratio of mixing was sampled from an exponential distribution, and *THE REAL McCOIL* can incorporate any fitted Beta distributions the users provide. We assumed uniform priors and updated P, M, S_T sequentially using a Metropolis-Hastings algorithm over $N = 100{,}000$ iterations, excluding the initial burn-in 1000 iterations to obtain posterior distributions of P and M. If ε_{est} was not pre-specified, *THE REAL McCOIL* estimated its posterior distribution along with P and M. The details of sampling procedure are described in Text A in S1 File.

Simulations

We sampled COI of each individual from a zero-truncated Poisson distribution with mean $\bar{m}$, and population allele frequency of each locus from a uniform distribution $U(0, 1)$. For each individual, we independently sampled allele(s) for each locus from Bernoulli (p_j). We determined the relative proportion of different lineages within the host by sampling the proportion of each infection from a uniform distribution $U(0, 1)$. For comparison, we additionally tried sampling from a truncated exponential distribution with the rate $\lambda = 1$. After obtaining within-host allele frequency (S_{Tij}), we drew S_{Oij} from a normal distribution with mean = S_{Tij} and variance $\sigma^2 = \frac{\varepsilon}{I}$, where ε represents the level of measurement error. We sampled the intensity of the signal I for each locus of each individual from the sum of a Poisson distribution with average $\bar{I} = 8$ and a normal distribution with mean = 0 and variance = 0.25. Simulations were designed to represent the type of raw data obtained from Sequenom or similar types of SNP assays, where an intensity value is obtained for each potential allele [34]. If the intensity of signal was smaller than I_{min}, we assumed the data were missing. We obtained the intensities of two alleles, A_1 and A_2, by $A_1 = I\frac{S_O}{\sqrt{{S_O}^2+(1-S_O)^2}}$ and $A_2 = I\frac{1-S_O}{\sqrt{{S_O}^2+(1-S_O)^2}}$, and determined heterozygous calls or homozygous calls by the relative intensity of signals of two alleles, which was characterized by $\arctan\left(\frac{A_1}{A_2}\right)$, the angle in polar coordinate system. The SNP was called as heterozygous if $\arctan\left(\frac{A_1}{A_2}\right)$ was within (d_1, d_2) and homozygous otherwise (Fig B in S1 File). For simulated data with measurement error $\varepsilon > 0$, we used $(d_1, d_2) = (5, 85)$. For real data, (d_1, d_2) was determined by expert review of each locus as described below.

We compared the performance of the categorical and proportional versions of our method to *COIL*, assessing the difference in parameter estimates and variation. We simulated violations of the model assumptions, specifically independence among loci, independence among parasite lineages within the same host, and a single, homogeneous population. Dependence among loci was simulated by different proportions of loci (p) that were linked. We simulated relatedness (r) among lineages within the same host by sampling alleles either from an existing lineage within the same host (with probability r) or from the population (with probability (1-r)). We simulated two equally sized subpopulations with either the same or different

average COI and with various levels of difference in allele frequencies and treated them as one single population to test the robustness of the assumption that the population was well-mixed. We also simulated missing data and populations with COI up to 20.

Genotyping of field samples

Dried blood spot samples were obtained from representative cross-sectional surveys performed in 2012 and 2013 as part of the East African International Centers of Excellence in Malaria Research (ICEMR) program. Surveys were performed in each of three sub-counties in Uganda: Nagongera in Tororo District, Kihihi in Kanungu District, and Walukuba in Jinja District. Details of these surveys, along with entomological and cohort data from the same sites have been published [29,31,36,37]. In brief, 200 households from each sub-county were randomly selected from a census population, and all children and an age-stratified sample of adults were enrolled from each household. All samples taken from individuals with evidence of asexual parasitemia by microscopy were selected for Sequenom SNP genotyping, and an age-stratified subset were also selected for merozoite surface protein 2 (*msp2*) genotyping. The Sequenom assay consisted of 128 SNPs selected to be polymorphic and at intermediate/high frequency in multiple popluations (https://www.malariagen.net/projects/p-falciparum-community-project). After removing variants with elevated missing rate, we retained 105 SNPs (see S1 Table for SNP data) and three of them are in known drug resistance loci. Samples were genotyped according to the relative intensity of the two alleles, as previously described [21]. Genotyping of *msp2* was performed with alleles sized by capillary electrophoresis, as previously described [38]. The number of unique alleles were called by a single, expert reader, with allele counts > 5 grouped into a single category due to difficulties in accurately distinguishing artifacts from true alleles at high complexities of infection.

Data analysis

After excluding samples with more than 25% missing SNP data and loci with more than 20% missing data from the analysis, the numbers of individuals included were 462 (71%) [Nagongera], 48 (51%) [Walukuba], and 74 (59%) [Kihihi], and the numbers of loci were 63 (60%) [Nagongera], 49 (47%) [Walukuba], and 52 (50%) [Kihihi]. After these cutoffs, only the analysis of Nagongera included one drug resistance locus, and others included none. We used a permutation test with $N = 10{,}000$ to compare estimated COI between groups because there were many ties. In the analysis, we assumed that error rates e_1 and e_2 were both 0.05 and $\varepsilon_{est} = 0.02$. F_{WS} was calculated by $(1-H_W/H_S)$, where H_W and H_S are $2p_W(1-p_W)$ and $2p_S(1-p_S)$ respectively and p_w and p_s are within-host allele frequency and population allele frequency respectively [33]. The H_W/H_S ratio was estimated by performing linear regression between H_W and H_S with fixed intercept = 0.

Results

Simultaneously estimating allele frequencies and the complexity of infection

We simulated data of 100 SNPs from populations with an average COI of 3, 5 and 7 and sample size of 100, and compared estimates of COI and allele frequencies using *COIL* and *THE REAL McCOIL*. When average COI was 3, all three methods estimated COI well, although allele frequency estimates from *COIL* were less precise than *THE REAL McCOIL* (mean absolute deviation [MAD] = 0.077 [*COIL*], 0.019 [*THE REAL McCOIL categorical*], 0.019 [*THE REAL McCOIL proportional*], Mann-Whitney test p-value $< 2\times10^{-16}$) (Fig 1). When average

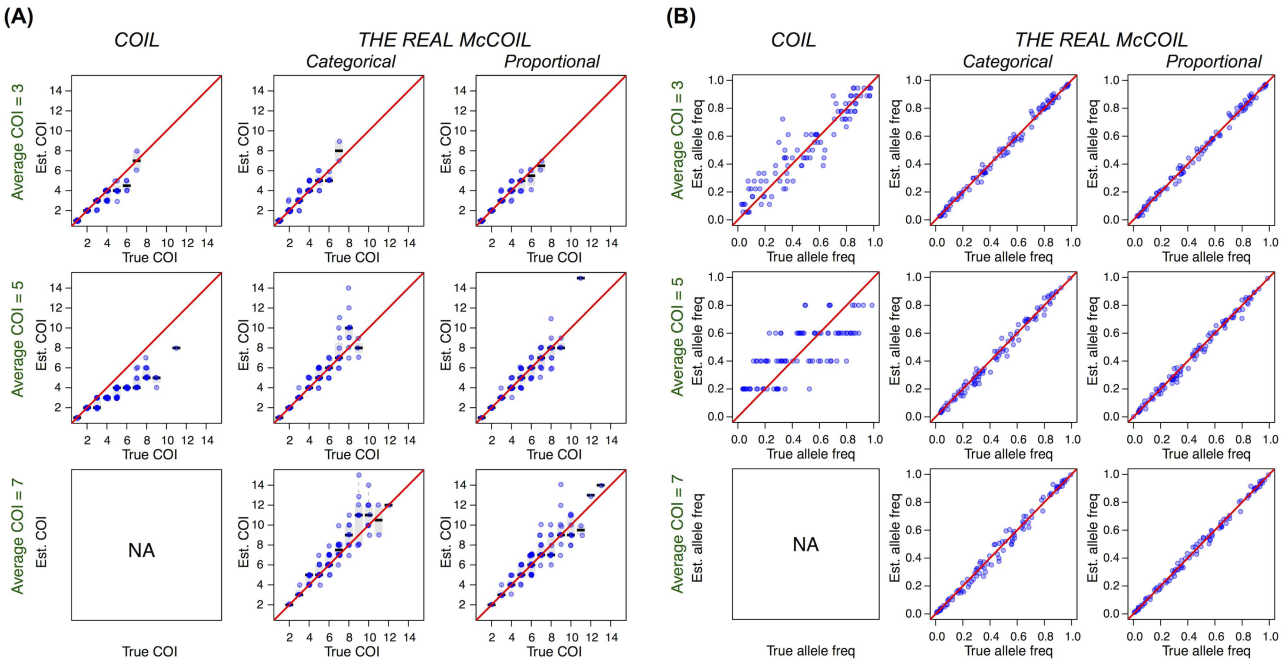

Fig 1. True vs. estimated values of COI (A) and allele frequencies (B) using *COIL* and *THE REAL McCOIL*. Each blue dot represents a sample. The black bar and the grey box show the median and 25% to 75% quantile. *THE REAL McCOIL* estimated allele frequencies and COI better than *COIL*, especially when the average COI was high and the majority of infections were polygenomic.

doi:10.1371/journal.pcbi.1005348.g001

COI was 5, however, *COIL* did not estimate COI or allele frequencies accurately (MAD = 1.45 [COI] and 0.15 [allele frequency]), and when COI was 7, it was unable to estimate allele frequencies due to a lack of monogenomic infections. In contrast to *COIL*, which consistently underestimated or failed to estimate COI in populations with greater numbers of polygenomic infections, *THE REAL McCOIL* estimated both COI and allele frequencies well even when COI was high (for categorical and proportional methods, respectively: COI = 5, MAD = 0.61, 0.45 [COI] and 0.024, 0.019 [allele frequency]; COI = 7, MAD = 0.86, 0.79 [COI] and 0.025, 0.015 [allele frequency]). Thus, the ability of *THE REAL McCOIL* to jointly estimate allele frequencies and COI from all available data resulted in considerably improved performance in estimates of both quantities, especially when the average COI was high.

Furthermore, we compared the performance of the categorical and proportional methods when we included measurement error in simulations of observed within-host allele frequency. The categorical method modeled measurement error by incorporating the probability of calling homozygous loci heterozygous (e_1) and vice versa (e_2) in the likelihood equation, and the proportional method modeled measurement error by assuming that the difference between true and observed within-host allele frequencies decreased with the intensity of the signals, and was proportional to the error parameter (ε_{est}). Fig C (A)(C) in S1 File shows that measurement error resulted in a systematic bias in estimates of COI. However, this bias was relatively minor and fairly robust to misspecification of measurement error, especially when the proportional method was used. In addition, allele frequencies were accurately estimated by both methods (Fig C (B)(E) in S1 File). If parameters for measurement error were not specified, *THE REAL McCOIL* fit them as part of the MCMC. Fig C (D)(F) in S1 File shows that the probability that the 95% credible interval contained the true COI when error parameters were fitted was higher than those when error parameters were greatly mis-specified.

Sensitivity analysis

We next simulated specific violations of the model assumptions to test the robustness of our approach. In particular, we examined the impact of linkage disequilibrium between loci, genetic relatedness of parasites within an individual host, and relatedness between subsets of individuals within the overall population (population substructure). When a proportion of loci (p) were completely linked, COI was slightly overestimated (Fig D in S1 File). When different lineages in the same host were not independent, COI was underestimated and the level of underestimation of COI increased with the level of relatedness (r) (Fig E in S1 File). When we treated two subpopulations as one population, COI was underestimated and the difference between true and estimated COI increased with the difference in the average of COI and the difference in allele frequencies between two subpopulations (Fig F in S1 File). Of these three violations of model assumptions, only a high degree of relatedness between parasites within an individual host resulted in substantial bias in estimates of COI, and none substantially affected estimates of population allele frequencies. Genotyping of real samples often results in missing data; both methods performed well even when 50% of the data were missing (Fig G in S1 File). Furthermore, we tested how the number of loci influences the performance of estimating COI. While the probability that 95% credible interval contained the true COI did not change with the number of loci, the average difference between true and estimated COI decreased (Fig H in S1 File). *THE REAL McCOIL* provided unbiased estimates even when COI was very high (e.g. 15–20), despite the uncertainty of the estimates increasing with true COI (Fig I in S1 File).

Complexity of infection and allele frequencies in three regions of Uganda

We next applied *THE REAL McCOIL* to data on 105 SNPs generated from smear positive individuals identified in cross-sectional surveys in three regions of Uganda [36,37] and compared results obtained from *THE REAL McCOIL* to those using *COIL*. Both categorical and proportional methods were applied and showed consistent results; for simplicity we therefore present only results from the categorical method.

Nagongera, Kihihi, and Walukuba have been shown to have transmission intensities varying by approximately 100 fold, with entomological inoculation rates recently measured at 310, 32, and 2.8 infectious bites per person year, respectively [29]. Using *COIL*, the estimated COI was relatively low, with little difference between the 3 sites (median COI = 2 [Nagongera], 2 [Walukuba], and 1.5 [Kihihi]) (Fig 2A). In contrast, results from *THE REAL McCOIL* show that the COI in Nagongera and Walukuba were similar, and much higher than that in Kihihi (median COI = 5 [Nagongera], 4.5 [Walukuba], and 1 [Kihihi])(Fig 2A, Table B in S1 File and S2 Table). These differences between sites were not captured by *COIL* because of its dependence on monogenomic infections to obtain estimates of allele frequencies, which were rare in these individuals. We also compared our results to COI estimated using another standard method, *msp2* typing, which was performed on a subset of the samples (Fig J in S1 File). Unlike *THE REAL McCOIL*, however, *msp2* typing estimated similar COI in Walukuba and Kihihi (p-value = 0.49) (Fig 2A). *msp2* encodes an antigen that elicits strong antibody responses, and this discrepancy may be due to complex population structure arising from immune selection. The difference may also result from the resolution of *msp2* typing, which is constrained to COI $\leq$ 5 [39], or the fact that it is a single marker, rather than a collection of genome-wide markers.

The high COI observed in the lowest transmission site of Walukuba was unexpected but reflected clear differences in the proportion of heterozygous calls, which was similar between Nagongera and Walukuba and lower in Kihihi (Fig K in S1 File). The distributions of age and parasite density were similar between the sites, and thus unlikely to explain these differences (Fig L and Fig M in S1 File). We calculated F_{WS}, an inverse measure of outcrossing [33,40],

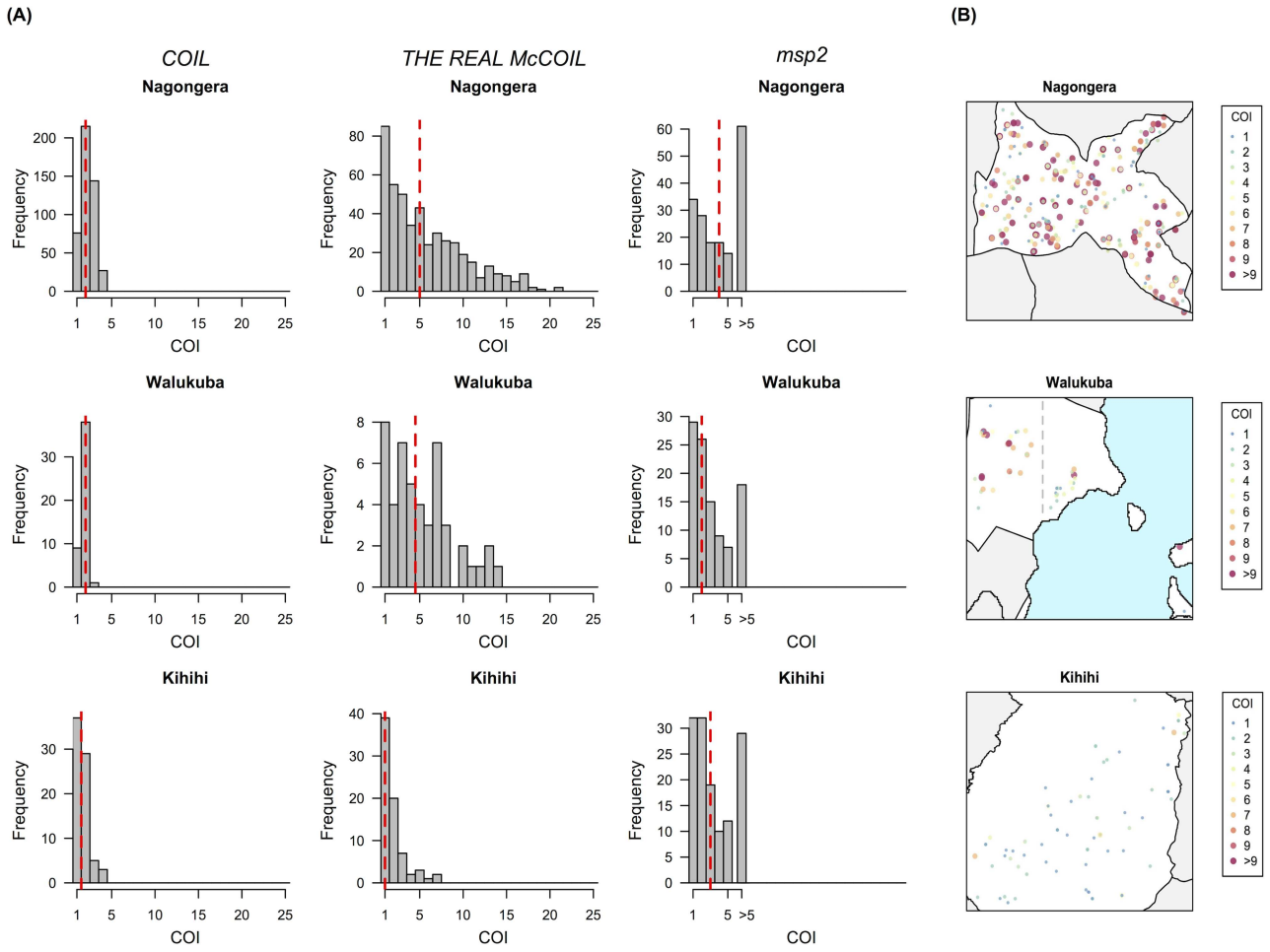

Fig 2. Estimates of COI in Nagongera, Walukuba, and Kihihi. (A) Estimates of COI by *COIL*, *THE REAL McCOIL*, and *msp2*. For *THE REAL McCOIL*, the point estimates of COI shown are medians from the posterior distributions. The COI estimated by *THE REAL McCOIL* in Nagongera and Walukuba were similar, and much higher than that in Kihihi (median COI = 5 [Nagongera], 4.5 [Walukuba], and 1 [Kihihi]; permutation test, *p*-values = 0.158 [Nagongera vs. Walukuba], 0.002 [Nagongera vs. Kihihi], 0.0006 [Walukuba vs. Kihihi]). Allele counts > 5 in *msp2* typing were grouped into a single category due to difficulties in accurately distinguishing artifacts from true alleles at high complexities of infection. The dashed red lines represent the medians of COI in three regions. **(B)** The spatial distribution of estimated COI by *THE REAL McCOIL* in three regions. Small random noise was added to the location of samples in the map. COI of samples collected from the West of Walukuba was higher than those from the East of Walukuba (medians = 5 [West] and 3 [East], *p*-value = 0.027).

doi:10.1371/journal.pcbi.1005348.g002

and found that it was significantly negatively associated with our COI estimates (Fig 3; Pearson's correlation test between log(COI) and F_{WS}, $\rho = -0.93$ [Nagongera], −0.94 [Walukuba], and −0.95 [Kihihi], *p*-values $< 2.2\times10^{-16}$ for all). F_{WS} in Nagongera and Walukuba are similar and lower than that in Kihihi, suggesting that the level of outcrossing is smallest in Kihihi, which is consistent with the pattern of COI.

We also examined the relationship between COI and epidemiological and geographical factors within each site. In Nagongera, COI in young children increased with age until peaking at age 7, and then decreased; sample sizes for the other two sites were too small to estimate trends (Fig N in S1 File). Interestingly, parasite density was negatively correlated with COI after adjusting for age (partial correlation $r = -0.15$ [Nagongera], −0.27 [Walukuba], −0.23 [Kihihi], *p*-values = 0.0011 [Nagongera], 0.058 [Walukuba], 0.043 [Kihihi]). This negative association was most pronounced in those aged 3–10 years in Nagongera (Fig O in S1 File), and may

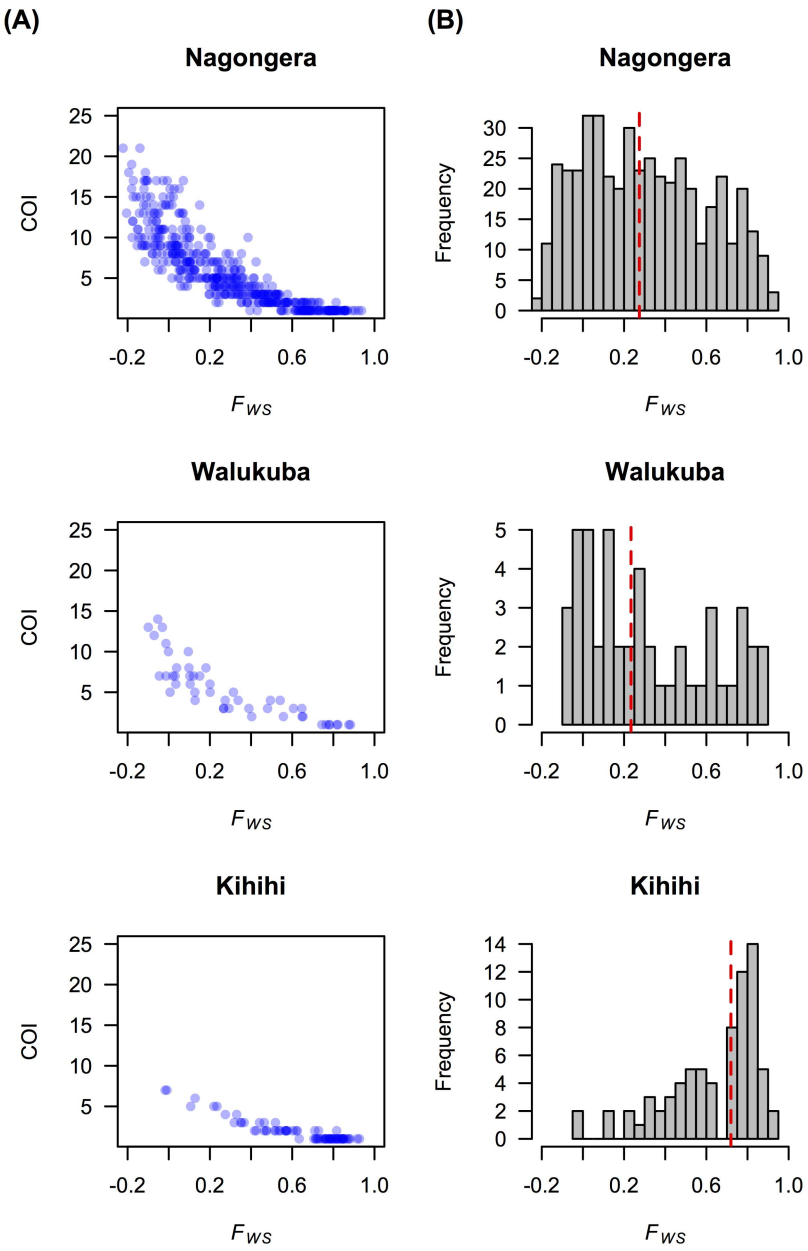

Fig 3. F_{WS}. (A) Estimated COI by *THE REAL McCOIL* was negatively associated with F_{WS}. **(B)** F_{WS} in Kihihi was higher than Nagongera and Walukuba. The F_{WS} values shown were calculated using population allele frequencies estimated from categorical method of *THE REAL McCOIL*.

doi:10.1371/journal.pcbi.1005348.g003

reflect the dominance of particular clones in acute, high-density infections. No differences in COI were observed between households with or without Insecticide Treated Nets (ITNs), or between sampling years.

In Kihihi, elevation and COI were negatively associated ($r = -0.259$, p-value = 0.026), consistent with the previously identified negative associations between elevation and mosquito density, the incidence of malaria, and serological evidence of exposure [41]. Interestingly, the unexpectedly high COI observed in Walukuba was largely driven by samples collected from the West of this sub-county, (Fig 2B; medians = 5 [West] and 3 [East], p-value = 0.027). We

have previously noted that mosquito densities in Walukuba are lower in the West, which is closer to urban centers, as compared to the East, which is a fishing village comprised largely of makeshift wooden housing [42]. One potential explanation for this seemingly paradoxical finding—high COI in the lowest transmission part of the lowest transmission site–is that a substantial proportion of these infections were imported from areas of higher transmission, where parasite populations are more diverse and co-transmission of multiple genetically distinct parasites is more likely.

Finally, we compared allele frequencies from each of the three sites to determine whether there was any evidence of population differentiation. We found little genetic differentiation between sites measured based on our estimated allele frequencies (F_{ST} ranged from 0.004 to 0.04; Table C in S1 File and S3 Table), although Kihihi, which is somewhat geographically isolated, had slightly higher F_{ST} with respect to the other two sites.

Discussion

Despite the availability of increasingly efficient genotyping technologies for molecular epidemiology, the prevalence of polygenomic infections in malaria-endemic regions hinders the estimation of basic population genetic parameters for *Plasmodium falciparum*. While *COIL* can estimate COI using allele frequencies from monogenomic infections or external data, direct estimation of allele frequencies from all samples is a preferable approach, particularly when no relevant frequency data are available and sample size is sufficient to overcome stochastic sampling error. *THE REAL McCOIL* accomplishes this by incorporating information from polygenomic infections to simultaneously estimate COI and population allele frequencies. We show through detailed simulations that our approach is robust to most model assumptions and can readily handle missing data. In addition, *THE REAL McCOIL* can utilize raw SNP genotyping data, allowing the method to be robust to errors in allele calling. Analysis of genotyping data from Uganda show that *THE REAL McCOIL* is able to identify nuances in field data that previous methods could not. In particular, compared with *msp2* genotyping or applying *COIL* to SNP data, we identified much higher average COI overall and epidemiologically relevant variation between and within study sites.

Through a number of simulations, we show that results obtained from *THE REAL McCOIL* are robust to assumptions that loci are independent and that the parasite population is homogeneous. As would be expected, a high degree of relatedness between parasites within an individual host resulted in substantial downward bias in estimates of COI. This is not trivial, as parasites in some epidemiological settings may be closely related within a host, e.g. due to co-transmission [43]. Fortunately, we found that this bias follows a clear linear pattern and can either be corrected if the level of relatedness is known, estimated directly from the data, or can at least be given reasonable bounds (Text B in S1 File). While estimating the level of relatedness may be challenging, enough information may be present in the data to do so in some cases, as demonstrated by a recent paper which estimated this parameter from sequence-read data [44]. *THE REAL McCOIL* can also be applied to read-based SNP data, and in theory can be extended to estimate relatedness. While we note that the most obvious model for measurement error in sequence-read data is a binomial distribution (Text C in S1 File), a normal distribution as applied in our current version offers a reasonable approximation and has computational advantages.

Genotyping of one or a few highly polymorphic antigen markers, such as *msp1* and *msp2*, is currently the most common method for determining COI [45,46]. The use of capillary electrophoresis has improved resolution of alleles, but due to the creation of PCR artifacts it is still difficult to accurately measure COI > 5 [38]. Deep sequencing of antigens such as *csp* is an

alternative approach [47,48]. However, with all of these approaches, immune selection on these genes within individuals and in a population can bias estimates of COI in ways which are difficult to predict [49,50]. Since loci under different types of selection can evolve independently in the presence of recombination, the diversity and geographic distribution of loci under immune selection may not be the same as observed among SNP loci. Both recombination rate and immune selection pressure will vary systematically with transmission intensity, resulting in complex associations between different genetic markers. Therefore, multiple genetic lineages defined by SNP panels may be associated with few *msp2* alleles, or vice versa, depending on the transmission setting and selective environment. In addition, if lineages within the host are related, using multiple markers across the genome is more likely to detect multiple lineages than using one region of the genome. F_{WS}, based on the difference between within-host and population heterozygosity, is a related metric used to quantify within-host diversity [33]. While F_{WS} is correlated with COI, the metric is conceptually different because it is influenced by both the relative proportions of lineages within the host and population allele frequencies [21,33,40]. *estMOI* [51] uses phasing information from sequence reads and the number of unique allelic combinations to estimate COI but requires deep sequencing data and can be biased by sequencing error. Some methods that use SNP data to estimate haplotype frequencies also simultaneously estimate COI [52,53]. However, current haplotype-based methods can only consider a limited number of loci (~7) because the number of possible haplotypes quickly expands with the number of loci. We expect that *THE REAL McCOIL* is better at estimating COI than these methods because it can incorporate a much larger number of SNPs. Moreover, COI estimated from *THE REAL McCOIL* could be used as a prior in tools estimating haplotype frequencies.

Application of *THE REAL McCOIL* to genotyping data from Uganda allowed us to calculate allele frequencies and F_{ST}, which was not possible to do from the raw data or using *COIL* due to the high proportion of heterozygous calls. *THE REAL McCOIL* also provided estimates of COI for all sites, which demonstrated associations with epidemiologic factors not identified using *msp2* genotyping. Interestingly, we identified a high COI in the lowest transmission site, potentially indicating importation of parasites from higher transmission areas. Although the possibility remains that recent transmission reduction left complex, chronic infections in its wake, explaining the high COI observed in Walukuba, the simplest explanation is that these infections were imported from high transmission settings nearby. Additionally, our results demonstrated that COI increased with age until age 7, and subsequently decreased, consistent with studies based on *msp1* and/or *msp2* typing [54–59]. Previous studies reported inconsistent associations between COI and parasite density for children $>$ 2 years old (positive [55,58,60], none [54,61], or negative [62]). We observed a negative association between COI and parasite density in children aged 3–10 in Nagongera. Although higher parasite density may help detect more strains within the host [63–65], the detection of minority strains may be more influenced by relative proportions of the strains [39]. Individuals with high parasite densities may be relatively immunologically naïve and have one or few lineages dominating the infection [66]. Lower parasite densities may be associated with partial immunity and parasite persistence, and consequently the accumulation of parasite lineages [67–71]. Also, parasite lineages are more likely to persist and accumulate in people with low parasite density because they are less likely to have clinical symptoms [70,72] and be treated. The discrepancy between studies can be due to different genetic markers, different transmission setting and immune levels, different contribution of co-transmission vs. superinfections, or some combination of these factors.

In summary, *THE REAL McCOIL* facilitates population genetic analysis of SNP data from polygenomic infections, which are common in many transmission settings and may predominate even in low transmission settings. Population allele frequency, which was previously

difficult to estimate if the majority of samples were polygenomic, can be estimated by *THE REAL McCOIL*, allowing downstream analysis that requires frequencies, such as estimating F_{ST}, $F_{WS,}$ and effective population size (N_e) [32,33,73]. *THE REAL McCOIL* is not only limited to *P. falciparum*, but can also be applied to other parasite species with polygenomic infections [74], including *Plasmodium vivax* [75]. Codes for *THE REAL McCOIL* are available on GitHub (https://github.com/Greenhouse-Lab/THEREALMcCOIL).

Supporting information

S1 File. Supporting information. Supplementary texts, figures and tables.
(PDF)

S1 Table. SNP data.
(TXT)

S2 Table. The 95% credible intervals of COI of samples from Uganda.
(TXT)

S3 Table. The 95% credible intervals of allele frequencies.
(TXT)

Acknowledgments

We thank Aimee Taylor and Rachel Daniels for helpful discussions.

Author Contributions

Conceptualization: HHC COB BG.

Formal analysis: HHC.

Methodology: HHC CJW DEN.

Resources: AY JN MRK SGS GD MM AEJ CH KAR RA DPK BG.

Supervision: COB BG.

Writing – original draft: HHC COB BG.

Writing – review & editing: CJW GD DEN RA.

References

1. World Health Organization (2015) World Malaria Report 2015. Geneva, Switzerland: World Health Organization.
2. Bousema T, Drakeley C, Gesase S, Hashim R, Magesa S, et al. (2010) Identification of hot spots of malaria transmission for targeted malaria control. J Infect Dis 201: 1764–1774. doi: 10.1086/652456 PMID: 20415536
3. Bousema T, Griffin JT, Sauerwein RW, Smith DL, Churcher TS, et al. (2012) Hitting hotspots: spatial targeting of malaria for control and elimination. PLoS Med 9: e1001165. doi: 10.1371/journal.pmed.1001165 PMID: 22303287
4. Moonen B, Cohen JM, Snow RW, Slutsker L, Drakeley C, et al. (2010) Operational strategies to achieve and maintain malaria elimination. Lancet 376: 1592–1603. doi: 10.1016/S0140-6736(10)61269-X PMID: 21035841
5. Carlton JM, Volkman SK, Uplekar S, Hupalo DN, Pereira Alves JM, et al. (2015) Population Genetics, Evolutionary Genomics, and Genome-Wide Studies of Malaria: A View Across the International Centers of Excellence for Malaria Research. Am J Trop Med Hyg 93: 87–98.

6. Daniels R, Chang HH, Sene PD, Park DC, Neafsey DE, et al. (2013) Genetic surveillance detects both clonal and epidemic transmission of malaria following enhanced intervention in Senegal. PLoS One 8: e60780. doi: 10.1371/journal.pone.0060780 PMID: 23593309
7. Daniels RF, Schaffner SF, Wenger EA, Proctor JL, Chang HH, et al. (2015) Modeling malaria genomics reveals transmission decline and rebound in Senegal. Proc Natl Acad Sci U S A 112: 7067–7072. doi: 10.1073/pnas.1505691112 PMID: 25941365
8. Conway DJ (2007) Molecular epidemiology of malaria. Clin Microbiol Rev 20: 188–204. doi: 10.1128/CMR.00021-06 PMID: 17223628
9. Malaria GENPfCP (2016) Genomic epidemiology of artemisinin resistant malaria. Elife 5.
10. Miotto O, Amato R, Ashley EA, MacInnis B, Almagro-Garcia J, et al. (2015) Genetic architecture of artemisinin-resistant Plasmodium falciparum. Nat Genet 47: 226–234. doi: 10.1038/ng.3189 PMID: 25599401
11. Mobegi VA, Loua KM, Ahouidi AD, Satoguina J, Nwakanma DC, et al. (2012) Population genetic structure of Plasmodium falciparum across a region of diverse endemicity in West Africa. Malar J 11: 223. doi: 10.1186/1475-2875-11-223 PMID: 22759447
12. Nkhoma SC, Nair S, Al-Saai S, Ashley E, McGready R, et al. (2013) Population genetic correlates of declining transmission in a human pathogen. Mol Ecol 22: 273–285. doi: 10.1111/mec.12099 PMID: 23121253
13. Obaldia N 3rd, Baro NK, Calzada JE, Santamaria AM, Daniels R, et al. (2015) Clonal outbreak of Plasmodium falciparum infection in eastern Panama. J Infect Dis 211: 1087–1096. doi: 10.1093/infdis/jiu575 PMID: 25336725
14. Patel JC, Taylor SM, Juliao PC, Parobek CM, Janko M, et al. (2014) Genetic Evidence of Importation of Drug-Resistant Plasmodium falciparum to Guatemala from the Democratic Republic of the Congo. Emerg Infect Dis 20: 932–940. doi: 10.3201/eid2006.131204 PMID: 24856348
15. Wei G, Zhang L, Yan H, Zhao Y, Hu J, et al. (2015) Evaluation of the population structure and genetic diversity of Plasmodium falciparum in southern China. Malar J 14: 283. doi: 10.1186/s12936-015-0786-0 PMID: 26194795
16. Escalante AA, Ferreira MU, Vinetz JM, Volkman SK, Cui L, et al. (2015) Malaria Molecular Epidemiology: Lessons from the International Centers of Excellence for Malaria Research Network. Am J Trop Med Hyg 93: 79–86. doi: 10.4269/ajtmh.15-0005 PMID: 26259945
17. Greenhouse B, Smith DL (2015) Malaria genotyping for epidemiologic surveillance. Proc Natl Acad Sci U S A 112: 6782–6783. doi: 10.1073/pnas.1507727112 PMID: 26016526
18. Anderson TJ, Haubold B, Williams JT, Estrada-Franco JG, Richardson L, et al. (2000) Microsatellite markers reveal a spectrum of population structures in the malaria parasite Plasmodium falciparum. Mol Biol Evol 17: 1467–1482. PMID: 11018154
19. Daniels R, Volkman SK, Milner DA, Mahesh N, Neafsey DE, et al. (2008) A general SNP-based molecular barcode for Plasmodium falciparum identification and tracking. Malar J 7: 223. doi: 10.1186/1475-2875-7-223 PMID: 18959790
20. Chang HH, Park DJ, Galinsky KJ, Schaffner SF, Ndiaye D, et al. (2012) Genomic sequencing of Plasmodium falciparum malaria parasites from Senegal reveals the demographic history of the population. Mol Biol Evol 29: 3427–3439. doi: 10.1093/molbev/mss161 PMID: 22734050
21. Manske M, Miotto O, Campino S, Auburn S, Almagro-Garcia J, et al. (2012) Analysis of Plasmodium falciparum diversity in natural infections by deep sequencing. Nature 487: 375–379. doi: 10.1038/nature11174 PMID: 22722859
22. Mobegi VA, Duffy CW, Amambua-Ngwa A, Loua KM, Laman E, et al. (2014) Genome-wide analysis of selection on the malaria parasite Plasmodium falciparum in West African populations of differing infection endemicity. Mol Biol Evol 31: 1490–1499. doi: 10.1093/molbev/msu106 PMID: 24644299
23. Haasl RJ, Payseur BA (2011) Multi-locus inference of population structure: a comparison between single nucleotide polymorphisms and microsatellites. Heredity (Edinb) 106: 158–171.
24. Sisya TJ, Kamn'gona RM, Vareta JA, Fulakeza JM, Mukaka MF, et al. (2015) Subtle changes in Plasmodium falciparum infection complexity following enhanced intervention in Malawi. Acta Trop 142: 108–114. doi: 10.1016/j.actatropica.2014.11.008 PMID: 25460345
25. Chang HH, Meibalan E, Zelin J, Daniels R, Eziefula AC, et al. (2016) Persistence of Plasmodium falciparum parasitemia after artemisinin combination therapy: evidence from a randomized trial in Uganda. Sci Rep 6: 26330. doi: 10.1038/srep26330 PMID: 27197604
26. Echeverry DF, Nair S, Osorio L, Menon S, Murillo C, et al. (2013) Long term persistence of clonal malaria parasite Plasmodium falciparum lineages in the Colombian Pacific region. BMC Genet 14: 2. doi: 10.1186/1471-2156-14-2 PMID: 23294725

27. Johnston WT, Mutalima N, Sun D, Emmanuel B, Bhatia K, et al. (2014) Relationship between Plasmodium falciparum malaria prevalence, genetic diversity and endemic Burkitt lymphoma in Malawi. Sci Rep 4: 3741. doi: 10.1038/srep03741 PMID: 24434689

28. Galinsky K, Valim C, Salmier A, de Thoisy B, Musset L, et al. (2015) COIL: a methodology for evaluating malarial complexity of infection using likelihood from single nucleotide polymorphism data. Malar J 14: 4. doi: 10.1186/1475-2875-14-4 PMID: 25599890

29. Kilama M, Smith DL, Hutchinson R, Kigozi R, Yeka A, et al. (2014) Estimating the annual entomological inoculation rate for Plasmodium falciparum transmitted by Anopheles gambiae s.l. using three sampling methods in three sites in Uganda. Malar J 13: 111. doi: 10.1186/1475-2875-13-111 PMID: 24656206

30. Okello PE, Van Bortel W, Byaruhanga AM, Correwyn A, Roelants P, et al. (2006) Variation in malaria transmission intensity in seven sites throughout Uganda. Am J Trop Med Hyg 75: 219–225. PMID: 16896122

31. Kamya MR, Arinaitwe E, Wanzira H, Katureebe A, Barusya C, et al. (2015) Malaria transmission, infection, and disease at three sites with varied transmission intensity in Uganda: implications for malaria control. Am J Trop Med Hyg 92: 903–912. doi: 10.4269/ajtmh.14-0312 PMID: 25778501

32. Weir BS, Cockerham CC (1984) Estimating F-Statistics for the Analysis of Population Structure. Evolution 38: 1358–1370.

33. Auburn S, Campino S, Miotto O, Djimde AA, Zongo I, et al. (2012) Characterization of within-host Plasmodium falciparum diversity using next-generation sequence data. PLoS One 7: e32891. doi: 10.1371/journal.pone.0032891 PMID: 22393456

34. Ross P, Hall L, Smirnov I, Haff L (1998) High level multiplex genotyping by MALDI-TOF mass spectrometry. Nat Biotechnol 16: 1347–1351. doi: 10.1038/4328 PMID: 9853617

35. Wen X, Stephens M (2010) Using Linear Predictors to Impute Allele Frequencies from Summary or Pooled Genotype Data. Ann Appl Stat 4: 1158–1182. PMID: 21479081

36. Nankabirwa JI, Yeka A, Arinaitwe E, Kigozi R, Drakeley C, et al. (2015) Estimating malaria parasite prevalence from community surveys in Uganda: a comparison of microscopy, rapid diagnostic tests and polymerase chain reaction. Malar J 14: 528. doi: 10.1186/s12936-015-1056-x PMID: 26714465

37. Yeka A, Nankabirwa J, Mpimbaza A, Kigozi R, Arinaitwe E, et al. (2015) Factors associated with malaria parasitemia, anemia and serological responses in a spectrum of epidemiological settings in Uganda. PLoS One 10: e0118901. doi: 10.1371/journal.pone.0118901 PMID: 25768015

38. Gupta V, Dorsey G, Hubbard AE, Rosenthal PJ, Greenhouse B (2010) Gel versus capillary electrophoresis genotyping for categorizing treatment outcomes in two anti-malarial trials in Uganda. Malar J 9: 19. doi: 10.1186/1475-2875-9-19 PMID: 20074380

39. Greenhouse B, Myrick A, Dokomajilar C, Woo JM, Carlson EJ, et al. (2006) Validation of microsatellite markers for use in genotyping polyclonal Plasmodium falciparum infections. Am J Trop Med Hyg 75: 836–842. PMID: 17123974

40. Murray L, Mobegi VA, Duffy CW, Assefa SA, Kwiatkowski DP, et al. (2016) Microsatellite genotyping and genome-wide single nucleotide polymorphism-based indices of Plasmodium falciparum diversity within clinical infections. Malar J 15: 275. doi: 10.1186/s12936-016-1324-4 PMID: 27176827

41. Helb DA, Tetteh KK, Felgner PL, Skinner J, Hubbard A, et al. (2015) Novel serologic biomarkers provide accurate estimates of recent Plasmodium falciparum exposure for individuals and communities. Proc Natl Acad Sci U S A 112: E4438–4447. doi: 10.1073/pnas.1501705112 PMID: 26216993

42. Kigozi SP, Pindolia DK, Smith DL, Arinaitwe E, Katureebe A, et al. (2015) Associations between urbanicity and malaria at local scales in Uganda. Malar J 14: 374. doi: 10.1186/s12936-015-0865-2 PMID: 26415959

43. Nkhoma SC, Nair S, Cheeseman IH, Rohr-Allegrini C, Singlam S, et al. (2012) Close kinship within multiple-genotype malaria parasite infections. Proc Biol Sci 279: 2589–2598. doi: 10.1098/rspb.2012.0113 PMID: 22398165

44. O'Brien JD, Iqbal Z, Wendler J, Amenga-Etego L (2016) Inferring Strain Mixture within Clinical Plasmodium falciparum Isolates from Genomic Sequence Data. PLoS Comput Biol 12: e1004824. doi: 10.1371/journal.pcbi.1004824 PMID: 27362949

45. Snounou G, Beck HP (1998) The use of PCR genotyping in the assessment of recrudescence or reinfection after antimalarial drug treatment. Parasitol Today 14: 462–467. PMID: 17040849

46. Viriyakosol S, Siripoon N, Petcharapirat C, Petcharapirat P, Jarra W, et al. (1995) Genotyping of Plasmodium falciparum isolates by the polymerase chain reaction and potential uses in epidemiological studies. Bull World Health Organ 73: 85–95.

47. Neafsey DE, Juraska M, Bedford T, Benkeser D, Valim C, et al. (2015) Genetic Diversity and Protective Efficacy of the RTS,S/AS01 Malaria Vaccine. N Engl J Med 373: 2025–2037. doi: 10.1056/NEJMoa1505819 PMID: 26488565

48. Bailey JA, Mvalo T, Aragam N, Weiser M, Congdon S, et al. (2012) Use of massively parallel pyrosequencing to evaluate the diversity of and selection on Plasmodium falciparum csp T-cell epitopes in Lilongwe, Malawi. J Infect Dis 206: 580–587. doi: 10.1093/infdis/jis329 PMID: 22551816

49. Ferreira MU, Hartl DL (2007) Plasmodium falciparum: worldwide sequence diversity and evolution of the malaria vaccine candidate merozoite surface protein-2 (MSP-2). Exp Parasitol 115: 32–40. doi: 10.1016/j.exppara.2006.05.003 PMID: 16797008

50. Escalante AA, Lal AA, Ayala FJ (1998) Genetic polymorphism and natural selection in the malaria parasite Plasmodium falciparum. Genetics 149: 189–202. PMID: 9584096

51. Assefa SA, Preston MD, Campino S, Ocholla H, Sutherland CJ, et al. (2014) estMOI: estimating multiplicity of infection using parasite deep sequencing data. Bioinformatics 30: 1292–1294. doi: 10.1093/bioinformatics/btu005 PMID: 24443379

52. Li X, Foulkes AS, Yucel RM, Rich SM (2007) An expectation maximization approach to estimate malaria haplotype frequencies in multiply infected children. Stat Appl Genet Mol Biol 6: Article33. doi: 10.2202/1544-6115.1321 PMID: 18052916

53. Takala SL, Smith DL, Stine OC, Coulibaly D, Thera MA, et al. (2006) A high-throughput method for quantifying alleles and haplotypes of the malaria vaccine candidate Plasmodium falciparum merozoite surface protein-1 19 kDa. Malar J 5: 31. doi: 10.1186/1475-2875-5-31 PMID: 16626494

54. Smith T, Beck HP, Kitua A, Mwankusye S, Felger I, et al. (1999) Age dependence of the multiplicity of Plasmodium falciparum infections and of other malariological indices in an area of high endemicity. Trans R Soc Trop Med Hyg 93 Suppl 1: 15–20.

55. Peyerl-Hoffmann G, Jelinek T, Kilian A, Kabagambe G, Metzger WG, et al. (2001) Genetic diversity of Plasmodium falciparum and its relationship to parasite density in an area with different malaria endemicities in West Uganda. Trop Med Int Health 6: 607–613. PMID: 11555426

56. Ntoumi F, Contamin H, Rogier C, Bonnefoy S, Trape JF, et al. (1995) Age-dependent carriage of multiple Plasmodium falciparum merozoite surface antigen-2 alleles in asymptomatic malaria infections. Am J Trop Med Hyg 52: 81–88. PMID: 7856831

57. Konate L, Zwetyenga J, Rogier C, Bischoff E, Fontenille D, et al. (1999) Variation of Plasmodium falciparum msp1 block 2 and msp2 allele prevalence and of infection complexity in two neighbouring Senegalese villages with different transmission conditions. Trans R Soc Trop Med Hyg 93 Suppl 1: 21–28.

58. Henning L, Schellenberg D, Smith T, Henning D, Alonso P, et al. (2004) A prospective study of Plasmodium falciparum multiplicity of infection and morbidity in Tanzanian children. Trans R Soc Trop Med Hyg 98: 687–694. doi: 10.1016/j.trstmh.2004.03.010 PMID: 15485698

59. Branch OH, Takala S, Kariuki S, Nahlen BL, Kolczak M, et al. (2001) Plasmodium falciparum genotypes, low complexity of infection, and resistance to subsequent malaria in participants in the Asembo Bay Cohort Project. Infect Immun 69: 7783–7792. doi: 10.1128/IAI.69.12.7783-7792.2001 PMID: 11705960

60. Vafa M, Troye-Blomberg M, Anchang J, Garcia A, Migot-Nabias F (2008) Multiplicity of Plasmodium falciparum infection in asymptomatic children in Senegal: relation to transmission, age and erythrocyte variants. Malar J 7: 17. doi: 10.1186/1475-2875-7-17 PMID: 18215251

61. Agyeman-Budu A, Brown C, Adjei G, Adams M, Dosoo D, et al. (2013) Trends in multiplicity of Plasmodium falciparum infections among asymptomatic residents in the middle belt of Ghana. Malar J 12: 22. doi: 10.1186/1475-2875-12-22 PMID: 23327681

62. Kidima W, Nkwengulila G (2015) Plasmodium falciparum msp2 Genotypes and Multiplicity of Infections among Children under Five Years with Uncomplicated Malaria in Kibaha, Tanzania. J Parasitol Res 2015: 721201. doi: 10.1155/2015/721201 PMID: 26770821

63. Contamin H, Fandeur T, Bonnefoy S, Skouri F, Ntoumi F, et al. (1995) PCR typing of field isolates of Plasmodium falciparum. J Clin Microbiol 33: 944–951. PMID: 7790466

64. Mayor A, Saute F, Aponte JJ, Almeda J, Gomez-Olive FX, et al. (2003) Plasmodium falciparum multiple infections in Mozambique, its relation to other malariological indices and to prospective risk of malaria morbidity. Trop Med Int Health 8: 3–11. PMID: 12535242

65. Guerra-Neira A, Rubio JM, Royo JR, Ortega JC, Aunon AS, et al. (2006) Plasmodium diversity in non-malaria individuals from the Bioko Island in Equatorial Guinea (West Central-Africa). Int J Health Geogr 5: 27. doi: 10.1186/1476-072X-5-27 PMID: 16784527

66. Childs LM, Buckee CO (2015) Dissecting the determinants of malaria chronicity: why within-host models struggle to reproduce infection dynamics. J R Soc Interface 12: 20141379. doi: 10.1098/rsif.2014.1379 PMID: 25673299

67. Felger I, Smith T, Edoh D, Kitua A, Alonso P, et al. (1999) Multiple Plasmodium falciparum infections in Tanzanian infants. Trans R Soc Trop Med Hyg 93 Suppl 1: 29–34.

68. al-Yaman F, Genton B, Reeder JC, Anders RF, Smith T, et al. (1997) Reduced risk of clinical malaria in children infected with multiple clones of Plasmodium falciparum in a highly endemic area: a prospective community study. Trans R Soc Trop Med Hyg 91: 602–605. PMID: 9463681

69. Farnert A, Rooth I, Svensson, Snounou G, Bjorkman A (1999) Complexity of Plasmodium falciparum infections is consistent over time and protects against clinical disease in Tanzanian children. J Infect Dis 179: 989–995. doi: 10.1086/314652 PMID: 10068596

70. Doolan DL, Dobano C, Baird JK (2009) Acquired immunity to malaria. Clin Microbiol Rev 22: 13–36, Table of Contents. doi: 10.1128/CMR.00025-08 PMID: 19136431

71. Smith T, Felger I, Tanner M, Beck HP (1999) Premunition in Plasmodium falciparum infection: insights from the epidemiology of multiple infections. Trans R Soc Trop Med Hyg 93 Suppl 1: 59–64.

72. Ali H, Ahsan T, Mahmood T, Bakht SF, Farooq MU, et al. (2008) Parasite density and the spectrum of clinical illness in falciparum malaria. J Coll Physicians Surg Pak 18: 362–368. doi: 06.2008/JCPSP.362368 PMID: 18760048

73. Pamilo P, Varvio-Aho SL (1980) On the estimation of population size from allele frequency changes. Genetics 95: 1055–1057. PMID: 17249052

74. Balmer O, Tanner M (2011) Prevalence and implications of multiple-strain infections. Lancet Infect Dis 11: 868–878. doi: 10.1016/S1473-3099(11)70241-9 PMID: 22035615

75. Friedrich LR, Popovici J, Kim S, Dysoley L, Zimmerman PA, et al. (2016) Complexity of Infection and Genetic Diversity in Cambodian Plasmodium vivax. PLoS Negl Trop Dis 10: e0004526. doi: 10.1371/journal.pntd.0004526 PMID: 27018585