PLOS Computational Biology

RESEARCH ARTICLE

Molecular surveillance of multiplicity of infection, haplotype frequencies, and prevalence in infectious diseases

Henri Christian Junior Tsoungui Obama[1,2*], **Kristan Alexander Schneider**[1,3,4]

1 Department of Applied Computer- and Biosciences, University of Applied Sciences Mittweida, Mittweida, Germany, **2** Department of Mathematics, Chemnitz University of Technology, Chemnitz, Germany, **3** Center for Global Health, Department of Internal Medicine, University of New Mexico, Albuquerque, New Mexico, United States of America, **4** Translational Informatics Division, Department of Internal Medicine, University of New Mexico, Albuquerque, New Mexico, United States of America

* christian.tsoungui@aims-cameroon.org

OPEN ACCESS

Citation: Tsoungui Obama HCJ, Schneider KA (2026) Molecular surveillance of multiplicity of infection, haplotype frequencies, and prevalence in infectious diseases. PLoS Comput Biol 22(7): e1012955. https://doi.org/10.1371/journal.pcbi.1012955

Editor: James M McCaw, The University of Melbourne, AUSTRALIA

Received: March 10, 2025

Accepted: March 11, 2026

Published: July 17, 2026

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1012955

Abstract

Background

The presence of multiple different pathogen variants within the same infection, referred to as multiplicity of infection (MOI), confounds molecular disease surveillance in diseases such as malaria. Specifically, if molecular/genetic assays yield unphased data, MOI causes ambiguity concerning pathogen haplotypes. Hence, statistical models are required to infer haplotype frequencies and MOI from ambiguous data. Such methods must apply to a general genetic architecture (i.e., multiple, multiallelic markers), when aiming to condition secondary analyses, e.g., population genetic measures such as heterozygosity or linkage disequilibrium, on the background of variants of interest, e.g., drug-resistance associated haplotypes.

Methods and findings

A statistical method to estimate MOI and pathogen haplotype frequencies, assuming a general genetic architecture, is introduced. The statistical model is formulated and the relation between haplotype frequency, prevalence and MOI is explained. Because no closed solution exists for the maximum-likelihood estimate, the expectation-maximization (EM) algorithm is used to derive the maximum-likelihood estimate. The asymptotic variance of the estimator (inverse Fisher information) is derived. This yields a lower bound for the variance of the estimated model parameters (Cramér-Rao lower bound; CRLB). By numerical simulations, it is shown that the bias of the estimator decreases with sample size, and that its covariance is well approximated by the inverse Fisher information, suggesting that the estimator is asymptotically unbiased and efficient. Computational performance is evaluated using empirical datasets, suggesting that the method is appropriate for up to thirteen polymorphic

Data availability statement: All data are in the manuscript and/or supporting information files, as well as online repositories. Namely, GitHub at https://github.com/Maths-against-Malaria/generalModel.git, where the material will be maintained, and Zenodo at https://doi.org/10.5281/zenodo.14553429.

Funding: o This study was supported by Sächsisches Staatsministerium für Wissenschaft und Kunst (SMWK)/Sächsische Aufbaubank (SAB), Project-ID 100257255 to KAS, by Deutscher Akademischer Austauschdienst (DAAD), Project-ID 57417782 to KAS, by Bundesministerium für Bildung und Forschung (BMBF)/ Deutsches Zentrum für Luft- und Raumfahrt (DLR), 01DQ20002 to KAS, by Europäischer Sozialfond (ESF)/ Sächsisches Staatsministerium für Wissenschaft und Kunst (SMWK)/Sächsische Aufbaubank (SAB), Project-ID 100257255 to KAS, Project-ID 100310497 to KAS and by Deutsche Forschungsgemeinschaft (DFG), Project-ID 656983 to KAS. 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.

markers. As an application, a dataset from Cameroon concerning anti-malarial drug resistance is analyzed, showing how the method can be utilized to derive population genetic measures associated with haplotypes of interest.

Conclusion

The proposed method has desirable statistical properties and is adequate for handling molecular data consisting of moderate number of multiallelic molecular markers. The EM-algorithm provides a stable iteration to numerically calculate the maximum-likelihood estimates. An implementation of the algorithm alongside a detailed documentation is provided in Supporting information S1 Data.

Author summary

Malaria annually causes 263 million infections and 596,000 deaths. Control efforts are challenged by factors like spreading drug resistance. Monitoring pathogen variants at the genetic level (molecular surveillance), especially those linked to drug resistance, is a public health priority. A major challenge is the presence of multiple, genetically distinct pathogen variants (characterized by several genetic markers) within infections (multiplicity of infection). Because genetic assays do not provide phased information in this context, ambiguity in reconstructing the actual variants present in an infection arises. This challenge is not limited to malaria. Probabilistic methods are required to phase genetic data, i.e., to reconstruct the pathogen variants present in infections. As such, we introduce a statistical method to estimate the distribution of pathogen variants at the population level from unphased molecular data obtained from disease-positive specimens. This is a combinatorially difficult task, as the number of possible genetic variants grows exponentially with the amount of genetic information included. Although the method applies to data with an arbitrary genetic architecture (i.e., multiple multiallelic markers), its application is constrained by computational limitations. The method's adequacy is explored and used to analyze a malaria dataset from Cameroon to guide applications. A stable numerical implementation is provided.

Introduction

Epidemiological surveillance of infectious diseases is increasingly augmented by molecular/genetic methods. This facilitates a shift from symptom-based to pathogen-specific diagnostics and allows monitoring of specific pathogen variants of interest, such as variants associated with antimicrobial resistance, and their routes of transmission. Molecular surveillance has become popular for a variety of pathogens of interest [1], due to advances in molecular/genetic methods.

Pathogen variants are typically characterized by their allelic configuration at certain loci/markers, e.g., by SNP barcodes, a microsatellite (STR) profile, or

microhaplotypes. The presence of several pathogen variants within an infection is common in many diseases. For instance, in malaria infections, it is well-recognized that distinct pathogen variants can be transmitted (by the same or different mosquitoes), a phenomenon commonly referred to as complexity of infection (COI) or multiplicity of infection (MOI) [2]. Especially, in malaria, MOI is recognized, because it scales with transmission intensities, however not necessarily in a linear way [3,4]. Additionally, MOI in malaria has far-reaching implications, as it mediates the amount of recombination between parasite variants and thereby changes the population genetics underlying malaria evolution, e.g., with regard to drug resistance [5]. Due to MOI in malaria, it is also important to distinguish between a variant's frequency (its relative abundance) and its prevalence (the probability of observing it in an infection) [2]. The latter is of clinical and epidemiological relevance, the former is of relevance for molecular surveillance, e.g., when studying the spread of drug resistance.

MOI is unfortunately ambiguously defined in the literature, with discrepancies between verbal and formal definitions, and the actual quantity referred to as MOI. These differences are discussed in detail in [2]. Namely, MOI is formally defined in most statistical models as the number of super-infections with the same or different pathogen variants, assuming that exactly one variant is transmitted at each infective event (cf. [2]). In the following, this definition of MOI is used. Verbally, MOI is often referred to as the number of different pathogen variants within an infection, which is a derived quantity from the formal definition of MOI. Note that co-transmission of different pathogen variants (co-infections) is ignored in the formal definition of MOI. Estimating MOI is typically coupled with estimating the frequency spectra of pathogen variants. Several methods have been proposed, often in the context of malaria, although these methods are not specific to this disease. Ad-hoc methods, e.g., [6,7], are straightforward to apply but typically biased (cf. [2] for a detailed discussion). Specifically, in [7], all samples with evidence of multiple infections are disregarded, effectively leading to overestimation of predominant and underestimation of rare pathogen variants. In [6] samples with unambiguous information are retained, and all pathogen variants in an infection are given the same weight. Although less biased than the former method, the latter leads to overestimation of rare variants. The bias of these methods increases with MOI. Alternative methods relying on formal statistical frameworks are more or less based on the same probabilistic model (cf. [2]). Differences occur in: (i) the quantities to be estimated (whether MOI or a derived parameter is estimated; cf. [2]); (ii) the genetic architecture allowed by these approaches, (e.g., single marker [8,9], two multiallelic markers [8,10], a set of biallelic markers [11,12], or multiple multiallelic markers [13,14]); (iii) whether MOI is estimated explicitly [15,16]; (iv) whether the probabilistic model uses approximations [11,17–20]; (v) the assumptions regarding the underlying distribution of MOI [2,21]; (vi) whether heuristic plug-in estimates are required for some parameters [22]; (vii) whether bias-corrections are applied [23]; (viii) the way missing data is handled [24]; and (ix) whether a Bayesian (e.g., [11,14,16]) or a frequentist approach (e.g., [8,9,12]) is being used.

Estimators for MOI and variant frequencies typically do not have an explicit form and have to be calculated numerically. In the frequentist case, this is often achieved by using the EM-algorithm [10,12,17,18,21,24–26]. Importantly, the asymptotic properties of the estimators were studied in detail for some methods – mainly in the simplest cases assuming a genetic architecture consisting of a single marker [23,24,27]. In these cases it can be shown that the probabilistic models fall into the class of exponential families, proving the existence, uniqueness, asymptotic unbiasedness, efficiency, and consistency of the estimators. Such investigations were complemented by numerical simulations to study the finite-sample properties. The same approach was used for more complicated underlying genetic architectures [10,12]. For the simple genetic architectures, it was also shown that the maximum likelihood estimates of MOI and variant frequencies coincide with the moment estimator for MOI and variants prevalence [27]. Notably, Bayesian and maximum-likelihood estimators should be in agreement if an uninformative prior distribution of model parameters is assumed. In the strict sense, prior distributions need to be inferred from an independent data source in a Bayesian framework, which is not always possible. A strong discrepancy between Bayesian and maximum-likelihood-based methods indicates that the underlying dataset does not reflect the information from the prior distributions well, thereby identifying a potential problem.

Estimates of haplotype frequencies and MOI can be further utilized as plug-in estimates for population genetic inferences, such as detecting selection patterns or population structure. E.g., quantities such as heterozygosity or pattern of linkage-disequilibrium can be derived, which are informative on selection processes, such as in the case of drug resistance.

Standard population genetic analyses to mine patterns of selection are concerned with completed selective sweeps, i.e., the mutations of interest replaced the wildtypes. In the case of antimicrobial resistance, one is concerned with ongoing selective processes, i.e., not all variants of interest reached fixation in the population. Hence, signatures of selection are distorted. It is therefore necessary to condition population genetic analyses on the subpopulation which is characterized by variants of interest at certain markers (loci). Many methods to estimate MOI and variant frequencies are insufficient in this context, as they are applicable only in the case of too restrictive genetic architectures. This is true for methods that apply only to single markers or biallelic markers.

Here, the maximum likelihood methods from [10,12] to estimate MOI and haplotype frequencies are extended to be applicable to a general genetic architecture. (In this context, “general” refers to the number of genetic markers and the number of alleles per markers. However, all markers have to be present in the specimens, which excludes deletions, certain inclusions, inversions or chromosomal rearrangements.) Because the method is capable of estimating frequencies of haplotypes characterized by multiple multiallelic loci, such estimates can be utilized for further population genetic analyses.

Readers are advised to start with section Methods and then move towards the results section. The underlying statistical model is derived in Methods under the assumption that MOI follows a conditional (positive) Poisson distribution (i.e., only disease-positive samples are considered). A number of illustrations help to facilitate the understanding of the mathematical notation (which is summarized for convenience in Table 1). A notation is used which unifies the preceding work of [9,10,12] which assumes a single multiallelic molecular marker, multiple biallelic markers, and two multiallelic markers, respectively. As for the simple genetic architectures the current model is too complex to allow for a closed-form solution for the maximum likelihood estimate (MLE). Hence, the expectation-maximization (EM) algorithm is employed. The derivation of the algorithm has the same structure as in [10,12] with some modifications and is hence presented for the sake of completeness in the Supporting information S1 Appendix in section "Deriving the EM-algorithm". Only the final iterative algorithm is presented in Results. There, Table 1 provides an overview of the notation for those not interested in the technical details but only the results. Table 1 and the figures in Methods should be consulted to facilitate readability. Furthermore, in Methods expressions for haplotype prevalence are derived, which are epidemiologically more relevant than frequencies. Particularly, the interplay between the MOI distribution, frequencies, and prevalence becomes clear from the formulae. Importantly, prevalence cannot be directly observed from molecular data, since typically only unphased haplotype information is available. Consequently, it is in general ambiguous which haplotypes (variants) are present in an infection. In ad hoc approaches (see above), sometimes only unambiguous information is considered (e.g., [6,7]) to derive estimates for haplotype prevalence/frequencies. To facilitate the relations with such methods, also formulae for the prevalence of haplotypes in unambiguous observations are derived.

Furthermore, the asymptotic covariances between model parameters and their transform (mean MOI; haplotype prevalence) are derived. Two alternative approaches are discussed in S1 Appendix.

The finite sample properties of the estimator are explored by numerical simulations. These complement the results of [10,12,27], which in special cases indicate that the proposed method has little bias and that the estimator’s variance is close to the Cramér-Rao lower bound. Moreover, the computational efficiency of the model is estimated by analyzing the Pf8 dataset from the MalariaGen project [28]. This suggests that computational time increases exponentially with the maximum number of polymorphic markers in samples.

Special emphasis is given to data application. Specifically, a *P. falciparum* dataset concerned with markers associated with sulfadoxine-pyrimethamine (SP) resistance collected in Yaoundé, Cameroon is analyzed. First, the frequencies

Table 1. Summary of key notation: Listed are the most important notations used throughout the manuscript. $\mathscr{P}(\boldsymbol{X})$ denotes the powerset of set $\boldsymbol{X}$. Abbreviations: PGF... probability generating function.

Notation	Descriptions	Structure	See also
L	Number of loci (markers)	$L \in \{2, 3, 4, \ldots\}$	–
n_k	Number of alleles at locus (marker) k	$n_k \in \{1, 2, 3, \ldots\}$	–
H	Number of possible haplotypes	$H = n_1 \cdot \ldots \cdot n_L$	–
$\boldsymbol{h}$	Vector representing haplotype by its allelic configuration	$\boldsymbol{h} \in \prod_{k=1}^{L} \{1, \ldots, n_k\}$	Fig 11B
$r(\boldsymbol{h})$	Rank (label) of haplotype $\boldsymbol{h}$	$r(\boldsymbol{h}) \in \{1, 2, \ldots, H\}$	(30a), Fig 11B
$\mathscr{H}$	Set of all possible haplotypes	$\mathscr{H} = \prod_{k=1}^{L} \{1, \ldots, n_k\}$	Fig 11A
$\mathcal{S}_H$	$(H-1)$-dimensional simplex	$\mathcal{S}_H := \mathbb{R}^+ \times (0, 1)^H$	–
$p_{\boldsymbol{h}}$	Frequency of haplotype $\boldsymbol{h}$	$p_{\boldsymbol{h}} \in [0, 1]$	–
$\boldsymbol{p}$	Vector of haplotype frequencies	$(p_{\boldsymbol{h}})_{\boldsymbol{h} \in \mathscr{H}} \in \mathcal{S}_H$	–
λ	MOI parameter (conditional Poisson distribution)	$\lambda > 0$	–
β	Lagrange multiplier (nuisance parameter)	$\beta \in \mathbb{R}$	–
$\boldsymbol{\theta}$	Vector of model parameters	$\boldsymbol{\theta} = (\lambda, \boldsymbol{p}) \in \mathbb{R}^+ \times \mathcal{S}_H,$	–
$\tilde{\boldsymbol{\theta}}$	Vector of model parameters (higher-dimensional space)	$\tilde{\boldsymbol{\theta}} = (\lambda, \boldsymbol{p}, \beta) \in \mathbb{R}^+ \times \mathcal{S}_H \times \mathbb{R}$	–
Θ	Model parameter space	$\Theta = \mathbb{R}^+ \times \mathcal{S}_H$	–
ψ	Mean MOI (mean of positive Poisson distribution)	$\psi \geq 1$	
$\boldsymbol{\vartheta}$	Transformed parameter vector	$(\psi, \boldsymbol{p}) \in [1, \infty) \times \mathcal{S}_H$	–
$\tilde{\boldsymbol{\vartheta}}$	Transformed parameter vector (higher-dimensional space)	$(\psi, \boldsymbol{p}, \beta) \in [1, \infty) \times \mathcal{S}_H \times \mathbb{R}$	–
$q_{\boldsymbol{h}}$	Prevalence of haplotype $\boldsymbol{h}$	$q_{\boldsymbol{h}} \in [0, 1]$	(4c)
$\boldsymbol{q}$	Vector of prevalences	$(q_{\boldsymbol{h}})_{\boldsymbol{h} \in \mathscr{H}} \in [0, 1]^H$	
m	MOI, number of independent infections	$m \in \{1, 2, 3, \ldots\}$	Fig 12
κ_m	Probability of MOI m (positive Poisson distribution)	$\kappa_m \in [0, 1]$	(32a)
$G(z)$ or $G_\lambda(z)$	PGF of MOI distribution (with parameter λ)	$G,\ G_\lambda : [0, 1] \to [0, 1]$	(48b)
$m_{\boldsymbol{h}}$	Number of times haplotype $\boldsymbol{h}$ is transmitted	$m_{\boldsymbol{h}} \in \{0, \ldots, m\}$	–
$\boldsymbol{m}$	Vector subsuming an infection	$\boldsymbol{m} = (m_{\boldsymbol{h}})_{\boldsymbol{h} \in \mathscr{H}}$	Fig 1
$\lvert\boldsymbol{m}\rvert$	Norm of vector $\boldsymbol{m}$	$\sum_{\boldsymbol{h} \in \mathscr{H}} m_{\boldsymbol{h}}$	–
$\boldsymbol{x}$ (or $\boldsymbol{y}$)	Observation, set-valued vector, with the k-th element,	$\boldsymbol{x},\ \boldsymbol{y} \in \prod_{k=1}^{L} \left(\mathscr{P}(\{1, \ldots, n_k\}) \setminus \{\emptyset\}\right)$	Figs 1 and 13
x_k (or y_k)	x_k (or y_k) being the set of alleles observed at locus k k-th element of an observation $\boldsymbol{x}$ (or $\boldsymbol{y}$)	$x_k,\ y_k \in \mathscr{P}(\{1, \ldots, n_k\}) \setminus \{\emptyset\}$	–
$\mathscr{O}$	Set of all possible observations	$\prod_{k=1}^{L} \left(\mathscr{P}(\{1, \ldots, n_k\}) \setminus \{\emptyset\}\right)$	–
$\boldsymbol{m} \to \boldsymbol{x}$	Infection $\boldsymbol{m}$ leads to observation $\boldsymbol{x}$	–	Fig 1
$A_{\boldsymbol{x}}$ (or $A_{\boldsymbol{y}}$)	Set of all haplotypes compatible with observation $\boldsymbol{x}$ (or $\boldsymbol{y}$)	$A_{\boldsymbol{x}},\ A_{\boldsymbol{y}} \in \mathscr{P}(\mathscr{H}) \setminus \{\emptyset\}$	Fig 1

(Continued)

Table 1. (Continued)

Notation	Descriptions	Structure	See also
$\mathscr{A}_{\boldsymbol{x}}$ (or $\mathscr{A}_{\boldsymbol{y}}$)	Set of all sub-observations of observation $\boldsymbol{x}$ (or $\boldsymbol{y}$)	$\mathscr{A}_{\boldsymbol{x}} \in \prod_{k=1}^{L} \left(\mathscr{P}(x_k) \setminus \{\emptyset\}\right)$ (or $\mathscr{A}_{\boldsymbol{y}} \in \prod_{k=1}^{L} \left(\mathscr{P}(y_k) \setminus \{\emptyset\}\right)$)	–
$\boldsymbol{y} \preceq \boldsymbol{x}$	$\boldsymbol{y}$ is a sub-observation of $\boldsymbol{x}$	$\boldsymbol{y} \in \mathscr{A}_{\boldsymbol{x}}$	–
$\boldsymbol{y} \prec \boldsymbol{x}$	$\boldsymbol{y}$ is a proper sub-observation of $\boldsymbol{x}$	$\boldsymbol{y} \in \mathscr{A}_{\boldsymbol{x}} \setminus \{\boldsymbol{x}\}$	–
$M_{\boldsymbol{x}}^{(m)}$	Set of all infections with MOI m that yield observation $\boldsymbol{x}$	$\{\boldsymbol{m} \mid \boldsymbol{m} \to \boldsymbol{x}, \lvert\boldsymbol{m}\rvert = m\}$	–
N	Sample size	Positive integer	–
$\mathscr{X}$	Empirical dataset	–	–
$\mathscr{M}$	Unobserved dataset of infections	–	–
$U_{\boldsymbol{h}}$	Set of haplotypes yielding unambiguous infections with $\boldsymbol{h}$	–	Fig 2
$\boldsymbol{x}^{(k)}$	k-th observation in dataset	–	–
$\mathcal{I}, \mathcal{I}'$	Fisher information and higher-dimensional embedding	–	–

https://doi.org/10.1371/journal.pcbi.1012955.t001

and prevalence of resistance-associated haplotypes are calculated. In a second step, heterozygosity and pairwise LD are determined across a range of microsatellite markers flanking the mutations of interest, conditioned on resistance-associated haplotypes that surpass a minimum frequency threshold of 10%. The data application is intended to showcase possible applications of the proposed method and to discuss its limitations. The necessity of the general genetic architecture becomes clear from the data applications. They are designed to showcase the limitations of the previous work of [10,12,27] (cf. Discussion).

An implementation of the method as R script alongside detailed documentation is provided as supplements and is available via GitHub at https://github.com/Maths-against-Malaria/generalModel.git, where the material will be maintained, and Zenodo at https://doi.org/10.5281/zenodo.14553429 (cf. [29]).

The method proposed here is haplotype-based, without simplifying assumptions such as independence between markers (linkage equilibrium). This applies to multiallelic molecular markers. This includes SNP barcodes, microhaplotypes, or panels of microsatellite markers. As a consequence, the number of involved computations grows exponentially with the number of markers in the dataset. Although in theory the number of molecular markers and the number of circulating variants (alleles) per marker are not restricted, the implementation reasches numerical limits if the number of markers is too large. The method is applicable to up to 13 polymorphic SNP markers (note that in standard SNP barcodes not all SNPs are polymorphic in a dataset - hence it will be applicable to, e.g., the standard 24-SNP barcode of [30]) or up to 10 mutiallelic microsatellite markers on a standard desktop computer. The applicability of the method is discussed in detail.

Results

The results described in this section use a specific notation aiming at facilitating the understanding of the results as well as the underlying derivations presented in the Methods section. Some key notations are summarized in Table 1.

The statistical model is derived in Methods. Here, only the final results are presented. First, the algorithm to derive the maximum likelihood estimates (MLE) of haplotype frequencies and the MOI parameter for the proposed statistical model is presented. Second, estimates of haplotype prevalences, which are epidemiologically and clinically more relevant quantities than frequencies, using the MLEs as plugin estimates are presented. Third, the asymptotic variance of the estimator is derived, i.e., the Cramér-Rao lower bound is derived as the inverse Fisher information. This is achieved by two alternative approaches, shown to be equivalent: (i) by embedding the log-likelihood function in a

higher dimensional space; (ii) by eliminating a redundant parameter. Next, the finite sample properties of the estimator are explored via numerical simulations. Finally, the computational efficiency of the proposed method is ascertained using data from the MalariaGen project [28], and its use is demonstrated on a *P. falciparum* dataset from Cameroon, to estimate heterozygosity and LD across a range of microsatellite markers on the background of haplotypes associated with SP-resistance.

Maximum likelihood estimate

The estimates $\hat{p}_{\boldsymbol{h}}$ and $\hat{\lambda}$ of the model parameters $p_{\boldsymbol{h}}$ and λ, representing the frequency of haplotype $\boldsymbol{h}$ and MOI parameter, respectively, are obtained by maximizing the log-likelihood function (53). Note that even in the simple case of a single locus (for which the model can be written as an exponential family), no closed-form solution exists (cf. [9]), and one has to rely on numerical methods. Here, for this purpose, the expectation-maximization (EM)-algorithm is employed.

The EM-algorithm is an iterative procedure that alternates two steps: (i) the expectation (E) step during which the expectation, with regard to the parameter choice in the current step, of the log-likelihood function is calculated over an unobserved variable and the unknown model parameters; (ii) the maximization (M) step during which the function from the E step is maximized over the unknown parameters. This yields the parameter choice for the next E-step. The whole iteration is repeated until convergence, yielding the MLE. A detailed derivation of the algorithm is described in Supporting information S1 Appendix in section Deriving the EM-algorithm.

Let $\mathscr{H}$ denote the set of all possible haplotypes, $\mathscr{O}$ the set of all possible observations, $\mathscr{A}_{\boldsymbol{x}}$ the set of all sub-observations of observation $\boldsymbol{x}$, $n_{\boldsymbol{x}}$ the number of times observation $\boldsymbol{x}$ occurs in the dataset of sample size N, $A_{\boldsymbol{y}}$ the set of all haplotypes compatible with observation $\boldsymbol{y}$, and $I_{A_{\boldsymbol{y}}}$ the indicator function of set $A_{\boldsymbol{y}}$ (see Table 1). The algorithm begins with an initial choice for the MOI parameter λ_0 and haplotype frequencies $p_{\boldsymbol{h}}^{(0)}$. Given the parameter choices λ_t and $p_{\boldsymbol{h}}^{(t)}$ in step t, the estimates are obtained in step $t+1$ by first updating the haplotype frequencies $p_{\boldsymbol{h}}^{(t+1)}$ as

$$p_{\boldsymbol{h}}^{(t+1)} = \frac{C_{\boldsymbol{h}}^{(t)}}{\sum_{\boldsymbol{h}\in\mathscr{H}} C_{\boldsymbol{h}}^{(t)}}, \tag{1a}$$

where

$$C_{\boldsymbol{h}}^{(t)} = p_{\boldsymbol{h}}^{(t)} \sum_{\boldsymbol{x}\in\mathscr{O}} n_{\boldsymbol{x}} \frac{\sum_{\boldsymbol{y}\in\mathscr{A}_{\boldsymbol{x}}} (-1)^{|\boldsymbol{x}|-|\boldsymbol{y}|} G'\left(\sum_{\boldsymbol{i}\in A_{\boldsymbol{y}}} p_{\boldsymbol{i}}^{(t)}\right) I_{A_{\boldsymbol{y}}}(\boldsymbol{h})}{\sum_{\boldsymbol{y}\in\mathscr{A}_{\boldsymbol{x}}} (-1)^{|\boldsymbol{x}|-|\boldsymbol{y}|} G\left(\sum_{\boldsymbol{i}\in A_{\boldsymbol{y}}} p_{\boldsymbol{i}}^{(t)}\right)}, \tag{1b}$$

and

$$I_{A_{\boldsymbol{y}}}(\boldsymbol{h}) = \begin{cases} 1 & \text{if } \boldsymbol{h} \in A_{\boldsymbol{y}}, \\ 0 & \text{otherwise.} \end{cases} \tag{1c}$$

Then, the MOI parameter λ_{t+1} is obtained by iterating the following equation until convergence:

$$x_{\tau+1} = x_\tau - \frac{x_\tau - \frac{B_t}{N}(1-e^{-x_\tau})}{1 + x_\tau - \frac{x_\tau}{1-e^{-x_\tau}}}, \tag{1d}$$

where

$$B_t = \sum_{\boldsymbol{x}\in\mathscr{O}} n_{\boldsymbol{x}} \frac{\sum_{\boldsymbol{y}\in\mathscr{A}_{\boldsymbol{x}}} (-1)^{|\boldsymbol{x}|-|\boldsymbol{y}|} \sum_{\boldsymbol{h}\in A_{\boldsymbol{y}}} p_{\boldsymbol{h}}^{(t)} G'\left(\sum_{i\in A_{\boldsymbol{y}}} p_i^{(t)}\right)}{\sum_{\boldsymbol{y}\in\mathscr{A}_{\boldsymbol{x}}} (-1)^{|\boldsymbol{x}|-|\boldsymbol{y}|} G\left(\sum_{i\in A_{\boldsymbol{y}}} p_i^{(t)}\right)}, \tag{1e}$$

where $G(z)$ and $G'(z)$ are the probability generation function (PGF; (48b)) of the MOI distribution (conditional Poisson distribution) and its derivative, respectively.

The iteration in (1d) starts with the initial value $x_0 = \lambda_t$ and is repeated until convergence, i.e., until $|x_{\tau+1} - x_\tau| < \varepsilon$ is satisfied. The update of the Poisson parameter in step $t+1$ is then $\lambda_{t+1} = x_{\tau+1}$.

The two iterative steps follow (1) until numerical convergence, i.e., until $|\lambda_{t+1} - \lambda_t| + \|p_{\boldsymbol{h}}^{(t+1)} - p_{\boldsymbol{h}}^{(t)}\|_2 < \varepsilon$. Once this holds, the MLE is obtained as

$$\hat{p}_{\boldsymbol{h}} = p_{\boldsymbol{h}}^{(t+1)} \quad \text{and} \quad \hat{\lambda} = \lambda_{t+1}. \tag{2}$$

Note, the algorithm is guaranteed to converge to a point, at which a maximum of the original likelihood function is attained (not necessarily a global maximum), typically within a few iterations.

The mean MOI ψ is then estimated by evaluating (32b) at the MLE, i.e., as

$$\hat{\psi} = \frac{\hat{\lambda}}{1 - e^{-\hat{\lambda}}}. \tag{3}$$

Considering the practical applicability, the difficulty here is to efficiently implement the rather complicated sums in (1). An implementation is available as an R script and comprehensive manual, provided in Supporting information S1 Data, and a version subject to future updates can be accessed via GitHub at https://github.com/Maths-against-Malaria/generalModel.git, and Zenodo at [29].

Key assumptions and applicability. The statistical model is based on a number of assumptions, which determines its applicability to malaria or other diseases.

First, the distribution of MOI is assumed to follow a conditional Poisson distribution, and is hence fully characterized by a single parameter λ. This is not restrictive, as it has been shown in similar situations that more flexible distirbutions hardly yield improvements [21].

Second, the model incorporates superinfections, while ignoring co-infections (i.e., the co-transmission of multiple pathogen variants at the same infective event). This implies that given an MOI value m, the number of infecting haplotypes is drawn from a multinomial distribution with parameters m (i.e., the realization of a random variable characterized by the Poisson parameter λ) and $\boldsymbol{p}$ (haplotype distribution). Hence, pathogen variants within each infection are independent. Under co-infections co-occurring haplotypes are not independent. As long as the dependency is weak, the super-infection assumption is a valid approximation. However, if – on a population level – the independence assumption of pathogen variants within infections is strongly violated, the proposed model is not applicable. For co-infections, one would need to model the distribution of sets of haplotypes being co-transmitted. In a disease like malaria, this would require additional assumptions regarding mosquito transmission dynamics, largely affected by factors such as vector competence and parasite density within hosts. In this case, both host-vector and vector-host transmission has to be modeled ideally. The former necessitates incorporating mutational models within hosts, as this mediates host-vector transmission.

Third, the method is haplotype-based without making simplifying assumptions, such as independence of markers (linkage equilibrium), which is often assumed in alternative methods, e.g., [14]. The advantage of assuming linkage equilibrium is that the number of model parameters grows linearly with the number of molecular markers. However, this assumption is a simplification which is not justified in all situations. In Particular, when, e.g., considering dense SNP panels, as this can bias results. Without assuming linkage equilibrium, the number of model parameters increases geometrically with the number of molecular markers. Although the method here is in theory applicable to an arbitrary number of multiallelic markers, the increase in model parameters imposes computational limitations on the number of molecular markers. As shown below, the model is applicable within a reasonable running time for up to approximately 10 polymorphic SNPs observed in samples, or about 8 multiallelic markers (such as microsatellites or microhaplotypes, cf. [31–33]). In diseases such as malaria, this is sufficient in many situations for standard SNP panels (e.g., 24 SNPs [30]). Namely, in practice, laboratory assays fail for some markers, not all markers are polymorphic in a given study populations, and the number of polymorphic SNPs in samples is typically much smaller than the number of SNPs considered.

Fourth, the model implicitly assumes missingness at random. Only samples for which molecular information is available at all markers can be included, while samples with missing information are excluded. If there is a justified reason to assume a systematic bias due to the type of molecular data or assays being performed, the missing at random assumption can be an oversimplification.

Fifth, an error model is not included, i.e., the allelic information in samples is assumed to be perfect. This is a pragmatic assumption due to the haplotype-based nature of the method. Incorporating an error model would be computationally intractable, as any observation could results form errors from any other observation. Note, error models are possible under the simplifying assumption of linkage equilibrium, as the probabilistic expressions for observations simplify substantially. However, assuming independent errors, e.g., [14], might not be justified in all situations, e.g., if minor variants fail to be detected. In general, the necessity and type of an error model is assay specific. When considering, e.g., SNP panels, the expected errors vary substantially for SNPs inferred from sequencing data, melting assays, or TaqMan assays. Whereas errors in microhaplotypes obtained from nanopore sequencing are negligible, microsatellites are prone to errors.

Estimation of prevalence

The prevalence of a haplotype is the probability that it occurs in an infection. Importantly, pathogen haplotypes are not directly observed in infections. Rather, the genetic information obtained from molecular assays is typically unphased, which leads to ambiguity if multiple haplotypes are present in an infection (cf. Fig 1). Consequently, the absence and presence of certain haplotypes in an infection is uncertain.

Therefore, two concepts of prevalence are distinguished here: (i) “true prevalence”, which is the probability that a haplotype is present in an infection, and (ii) “conditional prevalence”, which is the probability that a haplotype occurs in an infection, in which it can be unambiguously observed. The distinction is important for practical purposes. In the first case, a statistical model, as the one proposed here, is required to resolve ambiguity in the observable information, while in the latter case, all samples with ambiguous information would be disregarded, and prevalence simply calculated as the relative frequency of the remaining samples, in which a particular haplotype occurs.

To distinguish between unobservable and conditional prevalence, it is necessary to emphasize that for an infection with MOI = m, two outcomes are possible: (i) a “single infection” involving the same pathogen haplotype m times, and (ii) a super-infection involving several pathogen haplotypes, i.e., “multiple infections”. Depending on the infecting pathogen haplotypes, multiple infections yield either ambiguous observations when the alleles of the infecting pathogen haplotypes are different at more than one locus (as the infections illustrated in Fig 1) or unambiguous observations if only the pathogen haplotype are infecting such that their alleles differ at exactly one locus (cf. Fig 2). Note that single infections yield unambiguous observations.

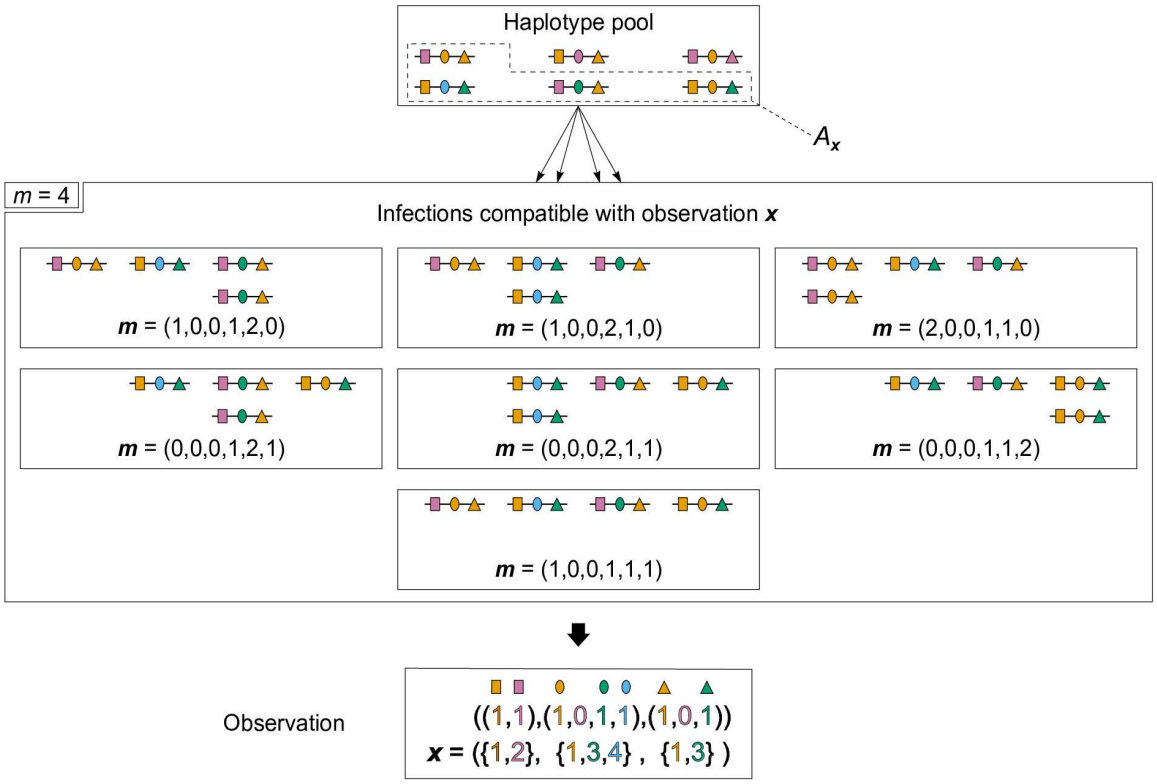

Fig 1. Compatibility between infections and observation: For MOI = 4, the middle panel illustrates all possible infections $\boldsymbol{m} = (m_1, m_2, m_3, m_4, m_5, m_6)$ with the 6 haplotypes circulating in the population (top panel), i.e., $|\boldsymbol{m}| = m = 4$ compatible with observation $\boldsymbol{x}$ (bottom panel). Haplotypes are sampled from a pool (top panel), in which they are ordered from one to six (so that the first and sixth haplotypes are at the top left and bottom right corner, respectively). In the top panel the set $\boldsymbol{A_x}$ of all haplotypes compatible with observation $\boldsymbol{x}$, is indicated.

https://doi.org/10.1371/journal.pcbi.1012955.g001

In the following, first the true prevalence is derived, followed by the derivation of the conditional prevalence, which is conditioned on only observing unambiguous observations, i.e., single infections and unambiguous multiple infections.

True prevalence. For a given haplotype $\boldsymbol{h}$, let $q_{\boldsymbol{h}}$ denote its true prevalence. The derivations follow the steps in [12]. Moreover, assume an infection with observation $\boldsymbol{x}$, and the probability $q_{-\boldsymbol{h}}$ that haplotype $\boldsymbol{h}$ is not in the infection underlying $\boldsymbol{x}$. If κ_m denotes the probability of MOI = m, the true prevalence $q_{\boldsymbol{h}}$ is obtained as

$$q_{\boldsymbol{h}} = 1 - q_{-\boldsymbol{h}} = 1 - \sum_{m=1}^{\infty} \kappa_m \sum_{\substack{\boldsymbol{m}:|\boldsymbol{m}|=m \\ m_{\boldsymbol{h}}=0}} \binom{m}{\boldsymbol{m}} \boldsymbol{p}^{\boldsymbol{m}}, \tag{4a}$$

where $\binom{m}{\boldsymbol{m}}$ and $\boldsymbol{p}^{\boldsymbol{m}}$ or short notations for the multinomial coefficients and power products occurring in a multinomial distribution (cf. Eq. 33). Using the multinomial theorem, the innermost sum yields

$$\sum_{\substack{\boldsymbol{m}:|\boldsymbol{m}|=m \\ m_{\boldsymbol{h}}=0}} \binom{m}{\boldsymbol{m}} \boldsymbol{p}^{\boldsymbol{m}} = \Big(\sum_{\boldsymbol{j} \in \mathscr{H}} p_{\boldsymbol{j}} - p_{\boldsymbol{h}} \Big)^m = (1 - p_{\boldsymbol{h}})^m. \tag{4b}$$

With G being again the PGF (cf. Eq. 48b) one obtains

$$q_{\boldsymbol{h}} = 1 - \sum_{m=1}^{\infty} \kappa_m (1-p_{\boldsymbol{h}})^m = 1 - G(1-p_{\boldsymbol{h}}) = 1 - \frac{e^{\lambda(1-p_{\boldsymbol{h}})} - 1}{e^{\lambda} - 1} = \frac{1 - e^{-\lambda p_{\boldsymbol{h}}}}{1 - e^{-\lambda}}. \tag{4c}$$

Therefore, given the MLEs $\hat{\lambda}$ and $\hat{p}_{\boldsymbol{h}}$ of the MOI parameter and haplotype frequency, respectively, the true prevalence is estimated from (4c) by

$$\hat{q}_{\boldsymbol{h}} = \frac{1 - e^{-\hat{\lambda}\hat{p}_{\boldsymbol{h}}}}{1 - e^{-\hat{\lambda}}}. \tag{5}$$

Conditional prevalence. In practice, it might be desirable to estimate haplotype prevalence based on observations in which haplotype information is unambiguous. In such a case, all ambiguous observations in the dataset are discarded, and the prevalence of a given haplotype $\boldsymbol{h}$ is estimated as the probability of observing $\boldsymbol{h}$, conditioned on observing unambiguous observations.

The set of possible unambiguous observations is a subset of $\mathcal{O}$ and is denoted by $\tilde{\mathcal{O}}$, i.e., $\tilde{\mathcal{O}} \subseteq \mathcal{O}$ (the subset is proper if more than one locus with at least two alleles are considered). The conditional prevalence $q_{\boldsymbol{h}|\tilde{\mathcal{O}}}$ of haplotype $\boldsymbol{h}$ is defined as

$$q_{\boldsymbol{h}|\tilde{\mathcal{O}}} := P(\boldsymbol{h} \mid \tilde{\mathcal{O}}) = \frac{P(\boldsymbol{h}, \tilde{\mathcal{O}})}{P(\tilde{\mathcal{O}})} = \frac{\tilde{q}_{\boldsymbol{h}}}{P(\tilde{\mathcal{O}})}, \tag{6}$$

where, $\tilde{q}_{\boldsymbol{h}}$ defines the probability that $\boldsymbol{h}$ is observed in an unambiguous observation, and $P(\tilde{\mathcal{O}})$ is the probability of all unambiguous observations.

For a given haplotype $\boldsymbol{h}$, to obtain the quantities $\tilde{q}_{\boldsymbol{h}}$ and $P(\tilde{\mathcal{O}})$, one needs to construct the set $U_{\boldsymbol{h}}$ of all haplotypes $\boldsymbol{i}$ that yield unambiguous multiple infections with $\boldsymbol{h}$ (see Fig 2). Recall that unambiguous multiple infections are obtained only for

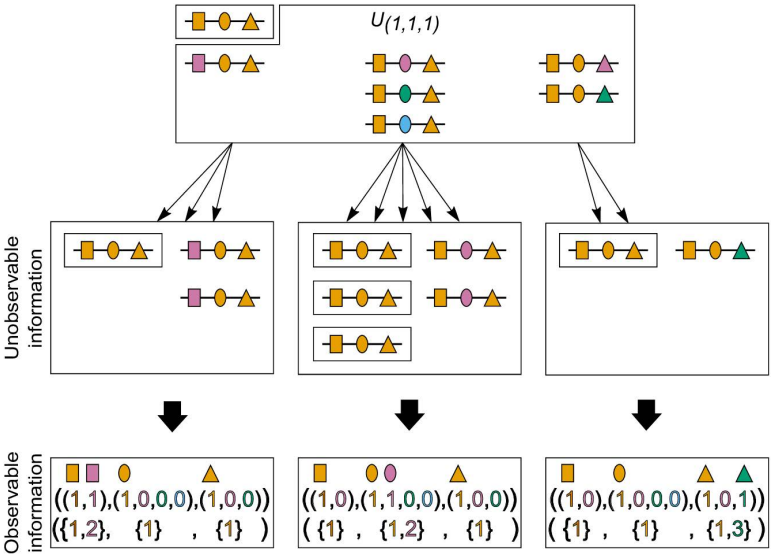

Fig 2. Illustration of unambiguity of infections between $\boldsymbol{h}$ and haplotypes in set $U_{\boldsymbol{h}}$: Shown are multiple infections involving (i) haplotype $\boldsymbol{h}$ = (1, 1, 1), and (ii) a haplotype from the set $U_{(1,1,1)}$ (upper block). In all resulting observations, haplotype information is unambiguous. However, MOI remains unknown.

https://doi.org/10.1371/journal.pcbi.1012955.g002

infections with exactly two pathogen haplotypes, such that their alleles are distinct at exactly one locus. Note that $\boldsymbol{h}$ is not in the set $U_{\boldsymbol{h}}$, i.e., $\boldsymbol{h} \notin U_{\boldsymbol{h}}$. Therefore, the set $U_{\boldsymbol{h}}$ is defined by

$$U_{\boldsymbol{h}} := \left\{ \boldsymbol{i} \in \mathscr{H} \mid \exists ! \, k : \boldsymbol{i}_k \neq \boldsymbol{h}_k \right\} \setminus \{\boldsymbol{h}\}. \tag{7}$$

Considering a genetic architecture with L loci and n_k alleles at locus k, the cardinality of $U_{\boldsymbol{h}}$ is $|U_{\boldsymbol{h}}| = \sum_{k=1}^{L}(n_k - 1) = \sum_{k=1}^{L} n_k - L$, and for any haplotype $\boldsymbol{i}$ such that $\boldsymbol{i} \in U_{\boldsymbol{h}}$, it follows that $\boldsymbol{h} \in U_{\boldsymbol{i}}$.

The probability $\tilde{q}_{\boldsymbol{h}}$ that haplotype $\boldsymbol{h}$ is observed in an unambiguous infection is the sum of the probabilities of all the multiple infections involving haplotype $\boldsymbol{h}$ and exactly one haplotype $\boldsymbol{i}$ in $U_{\boldsymbol{h}}$, and the single infection with $\boldsymbol{h}$. Assume a fixed haplotype $\boldsymbol{h}$, a haplotype $\boldsymbol{i} \in U_{\boldsymbol{h}}$, and an unambiguous observation $\tilde{\boldsymbol{x}} \in \tilde{\mathscr{O}}$ involving $\boldsymbol{h}$. For MOI = m, because $\boldsymbol{i}$ and $\boldsymbol{h}$ are sampled with replacement each of the m times, the following outcomes are possible for $\tilde{\boldsymbol{x}}$: (i) $\boldsymbol{h}$ is sampled all m times, i.e., $m_{\boldsymbol{h}} = m$, yielding a single infection with $\boldsymbol{h}$, (ii) $\boldsymbol{i}$ is sampled all m times, i.e., $m_{\boldsymbol{i}} = m$, yielding a single infection with $\boldsymbol{i}$ and (iii) $\boldsymbol{i}$ and $\boldsymbol{h}$ are sampled $m_{\boldsymbol{i}}$ and $m_{\boldsymbol{h}}$ times, respectively, such that $m = m_{\boldsymbol{h}} + m_{\boldsymbol{i}}$, yielding an unambiguous multiple infection (cf. [12]). Since unambiguous infections with $\boldsymbol{h}$ are of interest, only cases (i) and (iii) are relevant in the derivation of $\tilde{q}_{\boldsymbol{h}}$. Therefore, one has

$$\tilde{q}_{\boldsymbol{h}} := \sum_{m=1}^{\infty} \kappa_m \sum_{\boldsymbol{i} \in U_{\boldsymbol{h}}} \sum_{m_{\boldsymbol{i}}=1}^{m-1} \binom{m}{m_{\boldsymbol{i}}} p_{\boldsymbol{i}}^{m_{\boldsymbol{i}}} p_{\boldsymbol{h}}^{m-m_{\boldsymbol{i}}} + \sum_{m=1}^{\infty} \kappa_m p_{\boldsymbol{h}}^{m}. \tag{8a}$$

In section Prevalence estimates of Supporting information S1 Appendix this quantity is shown to simplify to

$$\tilde{q}_{\boldsymbol{h}} = \sum_{\boldsymbol{i} \in U_{\boldsymbol{h}}} \Big(G(p_{\boldsymbol{h}} + p_{\boldsymbol{i}}) - G(p_{\boldsymbol{i}}) \Big) - \Big(\sum_{k=1}^{L} n_k - L - 1 \Big) G(p_{\boldsymbol{h}}),$$

where G is again the PGF (48b), and L is the number of loci considered.

The probability of all unambiguous observations $P(\tilde{\mathscr{O}})$ is derived in section Prevalence estimates of Supporting information S1 Appendix and is given by

$$P(\tilde{\mathscr{O}}) = \sum_{\boldsymbol{h} \in \mathscr{H}} \left(\frac{1}{2} \sum_{\boldsymbol{i} \in U_{\boldsymbol{h}}} \Big(G(p_{\boldsymbol{h}} + p_{\boldsymbol{i}}) - G(p_{\boldsymbol{i}}) \Big) - \left(\frac{1}{2} \Big(\sum_{k=1}^{L} n_k - L \Big) - 1 \right) G(p_{\boldsymbol{h}}) \right).$$

Hence, the conditional prevalence of $\boldsymbol{h}$ is given by

$$q_{\boldsymbol{h}|\tilde{\mathscr{O}}} = \frac{\sum_{\boldsymbol{i} \in U_{\boldsymbol{h}}} \Big(G(p_{\boldsymbol{h}} + p_{\boldsymbol{i}}) - G(p_{\boldsymbol{i}}) \Big) - \Big(\sum_{k=1}^{L} n_k - L - 1 \Big) G(p_{\boldsymbol{h}})}{\sum_{\boldsymbol{j} \in \mathscr{H}} \left(\frac{1}{2} \sum_{\boldsymbol{i} \in U_{\boldsymbol{j}}} \Big(G(p_{\boldsymbol{j}} + p_{\boldsymbol{i}}) - G(p_{\boldsymbol{i}}) \Big) - \left(\frac{1}{2} \Big(\sum_{k=1}^{L} n_k - L \Big) - 1 \right) G(p_{\boldsymbol{j}}) \right)}. \tag{9}$$

The estimate of conditional prevalence is obtained by plugging the MLEs in (9), i.e.,

$$\hat{q}_{\boldsymbol{h}|\tilde{\mathscr{O}}} = \frac{\sum_{\boldsymbol{i} \in U_{\boldsymbol{h}}} \Big(G(\hat{p}_{\boldsymbol{h}} + \hat{p}_{\boldsymbol{i}}) - G(\hat{p}_{\boldsymbol{i}}) \Big) - \Big(\sum_{k=1}^{L} n_k - L - 1 \Big) G(\hat{p}_{\boldsymbol{h}})}{\sum_{\boldsymbol{j} \in \mathscr{H}} \left(\frac{1}{2} \sum_{\boldsymbol{i} \in U_{\boldsymbol{j}}} \Big(G(\hat{p}_{\boldsymbol{j}} + \hat{p}_{\boldsymbol{i}}) - G(\hat{p}_{\boldsymbol{i}}) \Big) - \left(\frac{1}{2} \Big(\sum_{k=1}^{L} n_k - L \Big) - 1 \right) G(\hat{p}_{\boldsymbol{j}}) \right)}. \tag{10}$$

Asymptotic variance and covariance

It is desirable to find unbiased estimators. In practice, this is hardly possible for complex statistical models. Specifically, maximum-likelihood estimators are typically only asymptotically unbiased, i.e., bias converges to zero as sample size increases to infinity. This also holds true for the present model, which was shown to be biased [10,12,24,27]. For the special case of a single molecular marker, the method was shown to be asymptotically unbiased, because the model falls into the class of exponential families [24]. For the cases of two multiallelic markers and arbitrary many biallelic markers, the model no longer falls into the class of exponential families and the estimator was shown numerically to be asymptotically unbiased [10,12]. In any case, it is also desirable to derive estimators with small variance. For unbiased estimators, the minimum possible variance is given by the Cramér-Rao lower bound (CRLB) [34,35]. Although biased estimators can have a lower variance, in practice, if an asymptotically unbiased estimator has variance close to the CRLB, there is not much hope to improve the estimator (cf. [36]). Hence, the CRLB is defined first, and it is shown numerically that it agrees well with the variance of the estimator. Together with the results of [10,12,24,27] this provides a numerical proof that the MLE of the present model is efficient, i.e., its asymptotic variance reaches the CRLB.

Original model parameters. The CRLB is the inverse expected Fisher information [34,35]. Given the model parameters $\boldsymbol{\theta}$ and the log-likelihood function $L_{\mathscr{X}}(\boldsymbol{\theta})$ in (53), the Fisher information, denoted $\mathcal{I}(\boldsymbol{\theta})$, is the square matrix with entries

$$\mathcal{I}_{i,j} = -\mathbb{E}\left[\frac{\partial^2 L_{\mathscr{X}}}{\partial\theta_i \partial\theta_j}(\boldsymbol{\theta})\right],$$

where θ_i and θ_j are components of the vector $\boldsymbol{\theta} \in \Theta = \mathbb{R}^+ \times \mathcal{S}_H$ (H being the total number of possible haplotypes, cf. Table 1). The expectation is taken with regard to the distribution of the observations $P_{\boldsymbol{x}}$. For the Fisher information it is more convenient to label haplotypes by integers. Each haplotype $\boldsymbol{h}$ is identified by its rank $h = r(\boldsymbol{h})$ (see Table 1). In the context of this model, the parameter space is not an open set of $\mathbb{R}^{H+1}$, since the $(H-1)$-dimensional simplex is not an open set in $\mathbb{R}^H$, as one of the haplotype frequency, e.g., p_H can be written as a function of the others, i.e., $p_H = 1 - \sum_{k=1}^{H-1} p_k$ (p_k denoting the frequency of the k-th haplotype). Results on the Fisher information require the parameter space to be an open set. Here, this can be resolved in two ways. First, one of the redundant parameters, e.g., p_H, can be eliminated by substituting $p_H = 1 - \sum_{k=1}^{H-1} p_k$. Second, the model can be embedded into a higher dimensional space by introducing a nuisance parameter.

Although the first approach seems simple at first sight, typically, calculations are simpler in the second approach. For the present model, this became evident in the special case of a single marker (cf. [9]). Hence, this approach is also pursued for the present general case, and it is formally proved that both approaches are equivalent (see section Fisher information matrix in Supporting information S1 Appendix).

Following [9], we use a Lagrange multiplier as a nuisance parameter, by defining

$$\Lambda_{\mathscr{X}}(\boldsymbol{\theta}, \beta) := \sum_{\boldsymbol{x}\in\mathscr{O}} n_{\boldsymbol{x}} \log\left(\sum_{\boldsymbol{y}\in\mathscr{A}_{\boldsymbol{x}}} (-1)^{|\boldsymbol{x}|-|\boldsymbol{y}|} G\left(\sum_{\boldsymbol{h}\in A_{\boldsymbol{y}}} p_{\boldsymbol{h}}\right)\right) + \beta\left(1 - \sum_{\boldsymbol{h}\in\mathscr{H}} p_{\boldsymbol{h}}\right). \tag{11}$$

The Lagrange multiplier β guarantees that $\Lambda_{\mathscr{X}}$ and the original likelihood function attain their maxima at the same point. Note, $\Lambda_{\mathscr{X}}$ is not a log-likelihood function. It is defined formally for all points $\tilde{\boldsymbol{\theta}} = (\boldsymbol{\theta}, \beta)$ in an open real space, i.e.,

$$\tilde{\boldsymbol{\theta}} = (\boldsymbol{\theta}, \beta) \in \tilde{\Theta} := \mathbb{R}^+ \times (0,1)^H \times \mathbb{R} \subseteq \mathbb{R}^{H+2}. \tag{12}$$

The equivalent of the Fisher information is defined as

$$\tilde{\mathcal{I}}_{i,j} = -\mathbb{E}\left[\frac{\partial^2 \Lambda_{\mathscr{X}}}{\partial\tilde{\theta}_i \partial\tilde{\theta}_j}(\tilde{\boldsymbol{\theta}})\right], \tag{13}$$

where i and j are the i-th and j-th element of the vector $\tilde{\boldsymbol{\theta}}$. The expectation is formally taken with regard to $P_{\boldsymbol{x}}$, which in the higher-dimensional embedding might not be a probability distribution. Importantly, for inverting the above matrix, the rows and columns corresponding to the nuisance parameter β need to be considered.

A detailed description of the derivation of the entries of the Fisher information is presented in section Entries of information matrix in higher dimensional space in Supporting information S1 Appendix. The entries are as follows

$$\begin{aligned}\tilde{\mathcal{I}}_{\lambda,\lambda} = N \sum_{\boldsymbol{x}\in\mathscr{O}} \Bigg(\Bigg(\sum_{\boldsymbol{y}\in\mathscr{A}_{\boldsymbol{x}}} (-1)^{|\boldsymbol{x}|-|\boldsymbol{y}|} G\Big(\sum_{\boldsymbol{h}\in A_{\boldsymbol{y}}} p_{\boldsymbol{h}} \Big) \Bigg)^{-1} \\ \times \Bigg(\sum_{\boldsymbol{y}\in\mathscr{A}_{\boldsymbol{x}}} (-1)^{|\boldsymbol{x}|-|\boldsymbol{y}|} \frac{\partial}{\partial\lambda} G\Big(\sum_{\boldsymbol{h}\in A_{\boldsymbol{y}}} p_{\boldsymbol{h}} \Big) \Bigg)^2 \Bigg),\end{aligned} \tag{14a}$$

where

$$\frac{\partial}{\partial\lambda} G(z) = \frac{z e^{\lambda z}}{e^{\lambda}-1} - e^{\lambda}\frac{e^{\lambda z}-1}{(e^{\lambda}-1)^2}.$$

Additionally,

$$\begin{aligned}\tilde{\mathcal{I}}_{\lambda,p_i} = N \sum_{\boldsymbol{x}\in\mathscr{O}} \Bigg(\Bigg(\sum_{\boldsymbol{y}\in\mathscr{A}_{\boldsymbol{x}}} (-1)^{|\boldsymbol{x}|-|\boldsymbol{y}|} G\Big(\sum_{\boldsymbol{h}\in A_{\boldsymbol{y}}} p_{\boldsymbol{h}} \Big) \Bigg)^{-1} \\ \times \Bigg(\sum_{\boldsymbol{y}\in\mathscr{A}_{\boldsymbol{x}}} (-1)^{|\boldsymbol{x}|-|\boldsymbol{y}|} \frac{\partial}{\partial\lambda} G\Big(\sum_{\boldsymbol{h}\in A_{\boldsymbol{y}}} p_{\boldsymbol{h}} \Big) \Bigg) \\ \times \Bigg(\sum_{\boldsymbol{y}\in\mathscr{A}_{\boldsymbol{x}}} (-1)^{|\boldsymbol{x}|-|\boldsymbol{y}|} \frac{\partial}{\partial p_{\boldsymbol{i}}} G\Big(\sum_{\boldsymbol{h}\in A_{\boldsymbol{y}}} p_{\boldsymbol{h}} \Big) \Bigg) \Bigg),\end{aligned} \tag{14b}$$

with

$$\frac{\partial}{\partial p_{\boldsymbol{i}}} G\Big(\sum_{\boldsymbol{h}\in A_{\boldsymbol{y}}} p_{\boldsymbol{h}} \Big) = \frac{\lambda}{e^{\lambda}-1} \exp\Big(\lambda \sum_{\boldsymbol{h}\in A_{\boldsymbol{y}}} p_{\boldsymbol{h}} \Big) I_{A_{\boldsymbol{y}}}(\boldsymbol{i}),$$

$$I_{A_{\boldsymbol{y}}}(\boldsymbol{i}) = \begin{cases} 1 & \text{if } \boldsymbol{i} \in A_{\boldsymbol{y}}, \\ 0 & \text{otherwise.} \end{cases}$$

Moreover,

$$\tilde{\mathcal{I}}_{\lambda,\beta} = 0, \tag{14c}$$

and

$$\tilde{\mathcal{I}}_{p_i,p_i} = N \sum_{\boldsymbol{x}\in\mathscr{O}} \left(\left(\sum_{\boldsymbol{y}\in\mathscr{A}_{\boldsymbol{x}}} (-1)^{|\boldsymbol{x}|-|\boldsymbol{y}|} G\left(\sum_{\boldsymbol{h}\in A_{\boldsymbol{y}}} p_{\boldsymbol{h}} \right) \right)^{-1} \times \left(\sum_{\boldsymbol{y}\in\mathscr{A}_{\boldsymbol{x}}} (-1)^{|\boldsymbol{x}|-|\boldsymbol{y}|} \frac{\partial}{\partial p_{\boldsymbol{i}}} G\left(\sum_{\boldsymbol{h}\in A_{\boldsymbol{y}}} p_{\boldsymbol{h}} \right) \right)^2 \right). \tag{14d}$$

One also has

$$\tilde{\mathcal{I}}_{p_i,p_j} = N \sum_{\boldsymbol{x}\in\mathscr{O}} \left(\left(\sum_{\boldsymbol{y}\in\mathscr{A}_{\boldsymbol{x}}} (-1)^{|\boldsymbol{x}|-|\boldsymbol{y}|} G\left(\sum_{\boldsymbol{h}\in A_{\boldsymbol{y}}} p_{\boldsymbol{h}} \right) \right)^{-1} \times \left(\sum_{\boldsymbol{y}\in\mathscr{A}_{\boldsymbol{x}}} (-1)^{|\boldsymbol{x}|-|\boldsymbol{y}|} \frac{\partial}{\partial p_{\boldsymbol{i}}} G\left(\sum_{\boldsymbol{h}\in A_{\boldsymbol{y}}} p_{\boldsymbol{h}} \right) \right) \times \left(\sum_{\boldsymbol{y}\in\mathscr{A}_{\boldsymbol{x}}} (-1)^{|\boldsymbol{x}|-|\boldsymbol{y}|} \frac{\partial}{\partial p_{\boldsymbol{j}}} G\left(\sum_{\boldsymbol{h}\in A_{\boldsymbol{y}}} p_{\boldsymbol{h}} \right) \right) \right). \tag{14e}$$

Finally,

$$\tilde{\mathcal{I}}_{p_i,\beta} = 1, \tag{14f}$$

and

$$\tilde{\mathcal{I}}_{\beta,\beta} = 0.$$

Hence, the "Fisher information matrix" has the form

$$\tilde{\mathcal{I}}(\tilde{\boldsymbol{\theta}}) = \begin{pmatrix} \tilde{\mathcal{I}}_{\lambda,\lambda} & \tilde{\mathcal{I}}_{\lambda,p_1} & \cdots & \tilde{\mathcal{I}}_{\lambda,p_H} & 0 \\ \tilde{\mathcal{I}}_{p_1,\lambda} & \tilde{\mathcal{I}}_{p_1,p_1} & \cdots & \tilde{\mathcal{I}}_{p_1,p_H} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \tilde{\mathcal{I}}_{p_H,\lambda} & \tilde{\mathcal{I}}_{p_H,p_1} & \cdots & \tilde{\mathcal{I}}_{p_H,p_H} & 1 \\ 0 & 1 & \cdots & 1 & 0 \end{pmatrix}. \tag{15}$$

The results in Fisher information matrix in Supporting information S1 Appendix show that for any likelihood function, whose parameter space includes a simplex, the CRLB is given by the inverted high-dimensional embedded "Fisher information matrix" with the rows and columns corresponding to the nuisance parameters being disregarded after inversion. Here, this corresponds to disregarding the row and column corresponding to β. This is denoted by

$$\mathrm{CRLB}(\boldsymbol{\theta}) = \tilde{\mathcal{I}}^{-1}(\boldsymbol{\theta}), \tag{16}$$

where the latter indicates that the row and column corresponding to β are disregarded, and the resulting matrix is evaluated at an admissible parameter $\boldsymbol{\theta} \in \Theta$. This gives the asymptotic covariance matrix of the original model parameters λ and $\boldsymbol{p}$.

Mean MOI and haplotype frequencies. In practice, the mean MOI ψ is of more interest than the MOI parameter λ. The inverse Fisher information can be readily used to calculate the asymptotic covariance matrix of the mean MOI and haplotype frequencies. Namely, recall that $\psi = f(\lambda) := \frac{\lambda}{1-e^{-\lambda}}$. Let

$$\tilde{\boldsymbol{\vartheta}} := (\psi, \boldsymbol{p}, \beta) = \big(f(\lambda), \boldsymbol{p}, \beta\big) \tag{17}$$

denote the parameters of interest. Thus, the log-likelihood function in terms of the new parameters becomes

$$\Lambda^*_{\mathscr{X}}(\tilde{\boldsymbol{\vartheta}}) = \Lambda^*_{\mathscr{X}}(\psi, \boldsymbol{p}, \beta) := \Lambda_{\mathscr{X}}\big(f^{-1}(\psi), \boldsymbol{p}, \beta\big). \tag{18}$$

The "Fisher information" of the transformed parameter $\tilde{\boldsymbol{\vartheta}}$ is defined as

$$\mathcal{I}^*_{ij} = -\mathbb{E}\left[\frac{\partial^2 \Lambda^*_{\mathscr{X}}}{\partial\tilde{\vartheta}_i \partial\tilde{\vartheta}_j}(\tilde{\boldsymbol{\vartheta}})\right], \tag{19}$$

where i and j are the i-th and j-th element of $\tilde{\boldsymbol{\vartheta}}$. Straightforward application of the chain rule gives

$$\mathcal{I}^*_{\psi,\psi} = \left(\frac{\partial f(\lambda)}{\partial\lambda}\right)^{-2} \tilde{\mathcal{I}}_{\lambda,\lambda}, \tag{20a}$$

$$\mathcal{I}^*_{\psi,p_i} = \mathcal{I}^*_{p_i,\psi} = \left(\frac{\partial f(\lambda)}{\partial\lambda}\right)^{-1} \tilde{\mathcal{I}}_{\lambda,p_i}, \tag{20b}$$

$$\mathcal{I}^*_{p_i,p_j} = \tilde{\mathcal{I}}_{p_i,p_j}, \tag{20c}$$

$$\mathcal{I}^*_{\psi,\beta} = \mathcal{I}^*_{\beta,\psi} = \left(\frac{\partial f(\lambda)}{\partial\lambda}\right)^{-1} \tilde{\mathcal{I}}_{\beta,\psi}, \tag{20d}$$

$$\mathcal{I}^*_{p_i,\beta} = \mathcal{I}^*_{\beta,p_i} = \tilde{\mathcal{I}}_{\beta,p_i}, \tag{20e}$$

and

$$\mathcal{I}^*_{\beta,\beta} = \tilde{\mathcal{I}}_{\beta,\beta}, \tag{20f}$$

where

$$\frac{\partial f(\lambda)}{\partial\lambda} = \frac{1}{1-e^{-\lambda}} - \frac{\lambda e^{-\lambda}}{(1-e^{-\lambda})^2}.$$

In compact form, this is written as

$$\mathcal{I}^*(\tilde{\boldsymbol{\vartheta}}) = diag\left(\left(\frac{\partial f(\lambda)}{\partial\lambda}\right)^{-1}, 1\ldots, 1\right)\tilde{\mathcal{I}}(\tilde{\boldsymbol{\theta}})diag\left(\left(\frac{\partial f(\lambda)}{\partial\lambda}\right)^{-1}, 1\ldots, 1\right), \tag{21}$$

from which it is evident that the inverse matrix becomes

$$\mathcal{I}^{*-1}(\tilde{\vartheta}) = diag\left(\frac{\partial f(\lambda)}{\partial \lambda}, 1 \ldots, 1\right) \tilde{\mathcal{I}}^{-1}(\tilde{\boldsymbol{\theta}}) diag\left(\frac{\partial f(\lambda)}{\partial \lambda}, 1 \ldots, 1\right). \tag{22}$$

The asymptotic covariance matrix is again given by dropping the row and column corresponding to β in (22).

Mean MOI and prevalences. In practice, haplotype prevalences are sometimes of more interest than frequencies. The asymptotic variance of prevalences is derived from that of the original parameters by similar transformations as above. Namely, recall that the (true) prevalence of haplotype $\boldsymbol{h}$ is given by $q_{\boldsymbol{h}} = \frac{1 - e^{-\lambda p_{\boldsymbol{h}}}}{1 - e^{-\lambda}}$. The transformed parameters

$$\vartheta' := (\psi, q_1, \ldots, q_H, \beta) \tag{23}$$

can be expressed in terms of the original parameters as

$$\vartheta' = \boldsymbol{g}(\lambda, \boldsymbol{p}, \beta), \tag{24}$$

where the parameter transformation can be written as

$$\boldsymbol{g}(\lambda, p_1, \ldots, p_H, \beta) := (f(\lambda), g(\lambda, p_1), \ldots, g(\lambda, p_H), \beta), \tag{25}$$

with f defined as above (17) and

$$g(\lambda, p_h) := q_h = \frac{1 - e^{-\lambda p_h}}{1 - e^{-\lambda}}, \tag{26}$$

(here the equivalence between haplotypes $\boldsymbol{h}$ and their rank h *is used*.)
The log-likelihood function expressed in terms of the transformed parameters becomes

$$\Lambda'_{\mathscr{X}}(\vartheta') = \Lambda'_{\mathscr{X}}(\psi, \boldsymbol{q}, \beta) = \Lambda_{\mathscr{X}}\left(\boldsymbol{g}^{-1}(\psi, \boldsymbol{q}, \beta)\right),$$

where $\boldsymbol{q} := (q_1, \ldots, q_H)$.

As above, application of the chain rule yields the entries of the "Fisher information" of the transformed parameters. Particularly,

$$\mathcal{I}'_{i,j} = -\mathbb{E}\left[\frac{\partial^2 \Lambda'_{\mathscr{X}}}{\partial \vartheta'_i \partial \vartheta'_j}(\vartheta')\right], \tag{27}$$

where i and j are the i-th and j-th element of the transformed parameter vector ϑ'. By using the chain rule, one obtains

$$\begin{aligned}\mathcal{I}'_{\psi,\psi} = &\left(\frac{\partial f(\lambda)}{\partial \lambda}\right)^{-2} \tilde{\mathcal{I}}_{\lambda,\lambda} + 2 \sum_{\boldsymbol{i} \in \mathscr{H}} \left(\frac{\partial g(\lambda, p_{\boldsymbol{i}})}{\partial \lambda} \frac{\partial f(\lambda)}{\partial \lambda}\right)^{-1} \tilde{\mathcal{I}}_{p_{\boldsymbol{i}}, p_{\boldsymbol{i}}} \\ &+ \sum_{\boldsymbol{i} \in \mathscr{H}} \sum_{\boldsymbol{j} \in \mathscr{H}} \left(\frac{\partial g(\lambda, p_{\boldsymbol{i}})}{\partial \lambda} \frac{\partial g(\lambda, p_{\boldsymbol{j}})}{\partial \lambda}\right)^{-1} \tilde{\mathcal{I}}_{p_{\boldsymbol{i}}, p_{\boldsymbol{j}}},\end{aligned} \tag{28a}$$

$$\mathcal{I}'_{\psi,q_h} = \mathcal{I}'_{q_h,\psi} = \left(\frac{\partial g(\lambda,p_h)}{\partial p_h}\frac{\partial f(\lambda)}{\partial \lambda}\right)^{-1}\tilde{\mathcal{I}}_{\lambda,p_h} + \sum_{j\in\mathcal{H}}\left(\frac{\partial g(\lambda,p_h)}{\partial p_h}\frac{\partial g(\lambda,p_j)}{\partial \lambda}\right)^{-1}\tilde{\mathcal{I}}_{p_h,p_h}, \tag{28b}$$

$$\mathcal{I}'_{q_h,q_j} = \mathcal{I}'_{q_j,q_h} = \left(\frac{\partial g(\lambda,p_h)}{\partial p_h}\frac{\partial g(\lambda,p_j)}{\partial p_j}\right)^{-1}\tilde{\mathcal{I}}_{p_h,p_j}, \tag{28c}$$

$$\mathcal{I}'_{q_h,\beta} = \mathcal{I}'_{\beta,q_h} = \left(\frac{\partial g(\lambda,p_h)}{\partial p_h}\right)^{-1}\tilde{\mathcal{I}}_{p_h,\beta}, \tag{28d}$$

$$\mathcal{I}'_{\psi,\beta} = \mathcal{I}'_{\beta,\psi} = 0, \tag{29e}$$

and

$$\mathcal{I}'_{\beta,\beta} = \tilde{\mathcal{I}}_{\beta,\beta}, \tag{28f}$$

where

$$\frac{\partial g(\lambda,p_h)}{\partial \lambda} = \frac{p_h e^{-\lambda p_h}}{1-e^{-\lambda}} - \frac{(1-e^{-\lambda p_h})e^{-\lambda}}{(1-e^{-\lambda})^2},$$

and

$$\frac{\partial g(\lambda,p_h)}{\partial p_h} = \frac{\lambda e^{-\lambda p_h}}{1-e^{-\lambda}}.$$

Note that the Jacobian matrix of the transformation $\boldsymbol{g}:\mathbb{R}^{H+2}\to\mathbb{R}^{H+2}$, with $\tilde{\boldsymbol{\theta}} = (\lambda, \boldsymbol{p}, \beta) \mapsto \boldsymbol{\vartheta}' = \boldsymbol{g}(\lambda, \boldsymbol{p}, \beta)$, denoted by J, allows to rewrite $\mathcal{I}'$ as

$$\mathcal{I}'(\boldsymbol{\vartheta}') = J^T\tilde{\mathcal{I}}(\tilde{\boldsymbol{\theta}})J, \tag{29a}$$

where

$$J := \frac{\partial \boldsymbol{g}(\lambda,p_1,\ldots,p_H,\beta)}{\partial(\psi,q_1,\ldots,q_H,\beta)}.$$

More explicitely,

$$J = \begin{pmatrix} \frac{\partial f(\lambda)}{\partial \lambda} & 0 & 0 & \cdots & 0 & 0 \\ \frac{\partial g(\lambda,p_1)}{\partial \lambda} & \frac{\partial g(\lambda,p_1)}{\partial p_1} & 0 & \cdots & 0 & 0 \\ \frac{\partial g(\lambda,p_2)}{\partial \lambda} & 0 & \frac{\partial g(\lambda,p_2)}{\partial p_2} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{\partial g(\lambda,p_H)}{\partial \lambda} & 0 & 0 & \cdots & \frac{\partial g(\lambda,p_H)}{\partial p_H} & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}^{-1}.$$

Therefore, the inverse of the matrix $\mathcal{I}'$ in (29a) is obtained as

$$\mathcal{I}'^{-1}(\vartheta') = J^{-1}\tilde{\mathcal{I}}^{-1}(\tilde{\theta})(J^{-1})^T. \tag{29b}$$

The inverse J^{-1} of the Jacobian matrix J is just

$$J^{-1} = \begin{pmatrix} \frac{\partial f(\lambda)}{\partial \lambda} & 0 & 0 & \cdots & 0 & 0 \\ \frac{\partial g(\lambda, p_1)}{\partial \lambda} & \frac{\partial g(\lambda, p_1)}{\partial p_1} & 0 & \cdots & 0 & 0 \\ \frac{\partial g(\lambda, p_2)}{\partial \lambda} & 0 & \frac{\partial g(\lambda, p_2)}{\partial p_2} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{\partial g(\lambda, p_H)}{\partial \lambda} & 0 & 0 & \cdots & \frac{\partial g(\lambda, p_H)}{\partial p_H} & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}.$$

The inverse $\mathcal{I}'^{-1}$ is the covariance matrix of the estimator $\vartheta' = \boldsymbol{g}(\lambda, \boldsymbol{p}, \beta)$ and is obtained from (29b).

The observed information is derived in a similar way. It is defined by dropping the expectation in (13), and evaluating at the MLE. Although the derivations are not presented explicitly in Supporting information, the observed information is included in the implementation of the method (see Supporting information S2 Appendix).

Finite sample properties

Typically, maximum-likelihood estimators have convenient statistical properties. The proposed method was shown to have little bias under finite sample assuming the following genetic architectures: multiple biallelic loci [12], a single multiallelic locus [27], and two multiallelic loci [10]. Furthermore, for a single multiallelic locus the method was proven to be asymptotically unbiased as it falls within the class of exponential families [24]. In the general case of multiple multiallelic loci, due to the curse of dimensionality, it is impractical to systematically investigate bias, besides that it would add little to what is already known. Hence, here the finite sample properties of the variance of the estimator are investigated and compared with the asymptotic results of the last section (Asymptotic variance and covariance). Thus, the focus here is on the efficiency of the estimator. To make the discrepancy between the actual variance and the asymptotic predictions of the mean MOI comparable across the parameter range, variation is studied in terms of the coefficient of variation (CV) for mean MOI.

Numerical simulations yield a close agreement between estimated coefficients of variation of both the mean MOI and haplotype frequencies and their theoretical prediction (based on the CRLB) for sample size $N > 50$ (Figs 3 and 4). For a small sample size of $N = 50$ a higher discrepancy is observed, particularly for more complex genetic architectures (compare Fig 3A, 3C, and Fig 3B, 3D). Namely, a larger number of alleles per locus yields an increased number of haplotypes, such that these are likely to have an inaccurate representation of the true haplotype distribution in a dataset with a small sample size ($N = 50$). Consequently, the estimate is sensitive to outliers, resulting in high variance for the MOI estimate. Additionally, the variance of the haplotype frequency estimates increases (compared with the asymptotic approximation; see Fig 4).

In summary, the numerical investigations suggest that the estimators of haplotype frequencies and MOI are efficient, and the asymptotic variance is properly predicted by the CRLB if sample size is sufficiently large ($N \geq 100$). Note that in the context of malaria, sample sizes of $N = 100 - 500$ are realistic.

Computational efficiency

Theoretically the method here is general in the sense that there are no restrictions on the number of markers or alleles per marker. However, in practice the method is restricted by computational limitations. Namely, the number of model

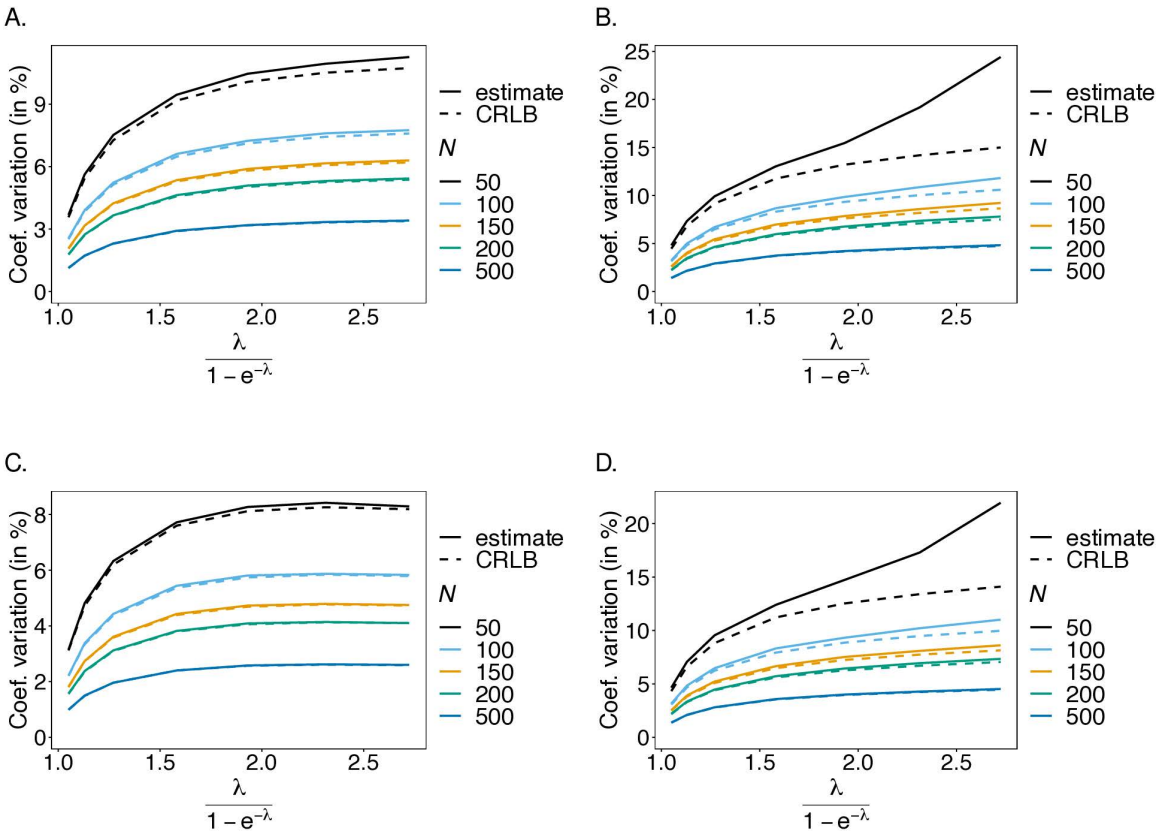

Fig 3. Cramér-Rao lower bound for MOI estimates: Shown is the coefficient of variation (CV) of the mean MOI estimates ψ in % as a function of the true mean MOI (i.e., for a range of MOI parameters). The genetic architecture is $n_1 = n_2 = 2$ in (A, C) and $n_1 = 4$, $n_2 = 7$ in (B, D). The frequency distribution in (A, B) is balanced while that in (C, D) is unbalanced (see Methods). The different sample sizes are represented by the colors, the continuous lines are the CV from the MOI estimates and the dashed lines are the CV from the Cramér-Rao lower bound.

https://doi.org/10.1371/journal.pcbi.1012955.g003

parameters $|\Theta|$ increases "exponentially" with the number of genetic markers. E.g., for n biallelic markers 2^n haplotypes have to be considered. Thus, for $n = 2$, 10, and 100 a total of 4, 1 024, and $1.267651 \cdot 10^{30}$ haplotypes have to be considered, respectively, with 9, 59 049, and $5.153775 \cdot 10^{47}$ possible observations. The total number of observations has to be considered only for the Fisher information (but not necessarily for the observed information), which becomes computationally infeasible. For the computation of the maximum-likelihood estimates (MLEs) all distinct observations in a given dataset and all their sub-observations have to be considered. In the worst case these are all possible observations. In the current implementation, a large number of markers will hence lead to memory overflow. Note, memory overflow, will manifest differently across operating systems. On Windows a memory limit is assigned directly in R (and can be manually increased), such that memory overflow will be indicated as an error. On Linux or macOS R has no internal memory limit, instead it is managed by the operating system itself. Memory overflow manifests in a program crash.

Here, data from the MalariaGen project [28] is used to explore the computational limitations of the method. First, the dataset on antimalarial drug resistance was utilized. The dataset consists of allelic information at $L = 36$ markers, i.e., codons at the *Pfcrt*, *Pfdhfr*, *Pfdhps*, *Pfkelch13* and *Pfmdr1* loci (see Methods), of 33 325 worldwide samples from several studies. This data was chosen because the genetic markers are not restricted to be biallelic. The dataset was divided into separate datasets by country, year, and study. Datasets with more than 20 samples and at least two polymorphic makers were retained, yielding a total of 110 unique datasets (see Methods). Importantly, the number of polymorphic markers per

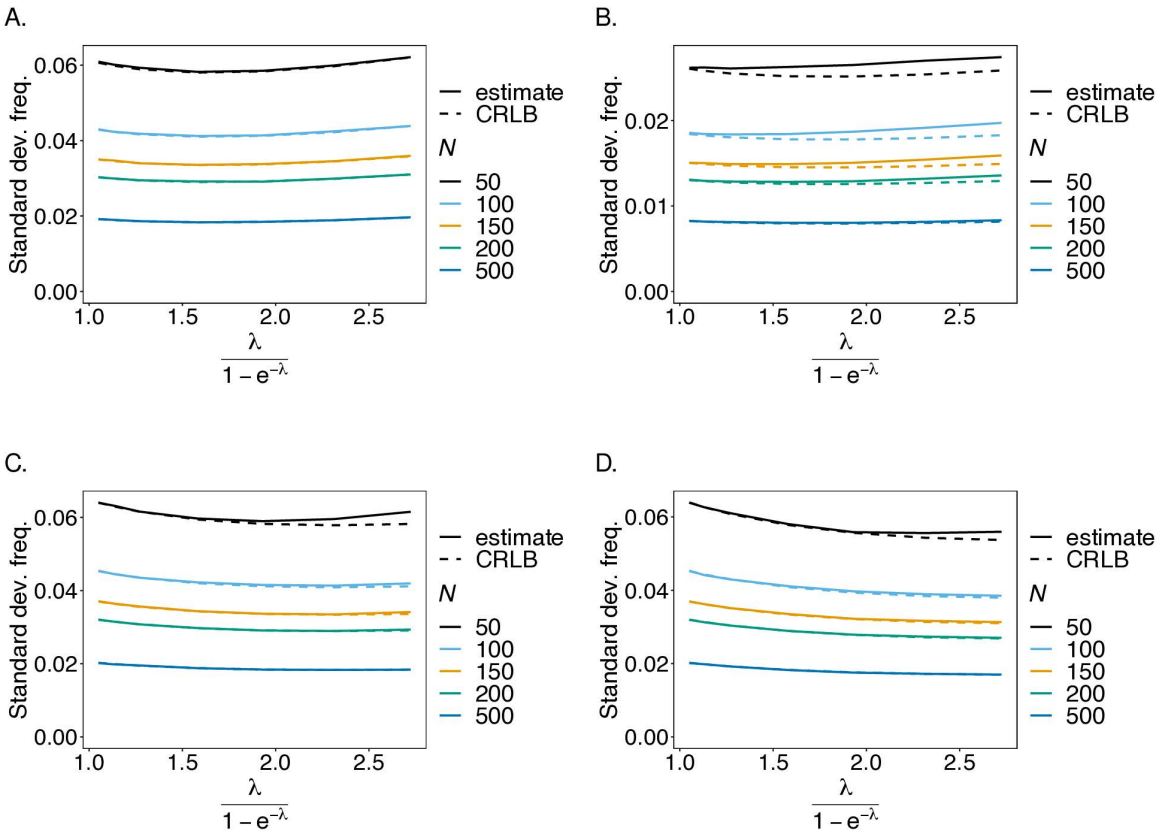

Fig 4. Cramér-Rao lower bound (CRLB) for haplotype frequencies estimates: Shown is the standard deviation (continuous line) alongside the CRLB (dashed line) for the estimates of haplotype frequencies as a function of the true mean MOI (i.e., for a range of MOI parameters). Panels (A, C) correspond to the estimation of a haplotype frequency whose true frequency is $p = 0.25$ assuming the genetic architecture $n_1 = n_2 = 2$. The case of the genetic architecture $n_1 = 4$, $n_2 = 7$ for a dominant haplotype with true frequency $p = 0.7$ is shown in panels (B, D). Notably, the true haplotype frequency distribution in (A, B) is balanced while that in (C, D) is unbalanced (see Methods). The colors indicate the different sample sizes.

https://doi.org/10.1371/journal.pcbi.1012955.g004

dataset ranged from $L = 2$ to $L = 16$. The MLE of MOI and haplotype frequencies was calculated for each dataset and the computational time was recorded.

The method converged for 108 datasets without the occurrence of memory overflow or numerical over- or underflow, specifically for those datasets with up to 13 polymorphic markers. Memory overflow occurred with the remaining 2 datasets with 15 and 16 polymorphic markers, respectively (in our cases R crashed when macOS with 8 GB RAM processor allocated around 30 GB RAM, i.e., 7 GB physical RAM and 23 GB SWAP memory – borrowed from the hard disk, leaving less than 1 GB physical RAM for the operating system to run). For the majority of datasets the method converged within a few seconds to minutes (Fig 5). However, for complex datasets, computational time reached several hours. Computational time increases almost exponentially with the maximum number of polymorphic markers per sample. This is expected since the algebraic operations in the algorithm (1a) and (1d) increases exponentially with the number of polymorphic markers in the dataset. The effect of sample size is not straightforward. With a large sample size, the occurrence of several samples with many polymorphic markers is more likely. Hence, there is a tendency that datasets with a larger sample size require more computational time. However, the exact running time depends on the specific structure of a dataset.

As a second attempt to benchmark computational time in practice, the $N = 228$ samples from a study conducted in Cameroon in 2013 available in the Pf8 dataset of the MalariaGen project were used. The study was picked due to its

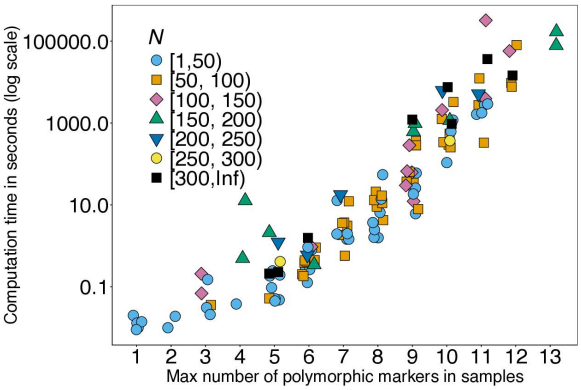

Fig 5. Bench-marking computational time on drug resistance data: Shown is the computational time in seconds (y-axis; log scale) as a function of the maximum number of polymorphic markers per sample (x-axis). Each point shows the time for one of the 108 datasets, for which the method converged. Ranges of sample size are shown in different colors and shapes.

https://doi.org/10.1371/journal.pcbi.1012955.g005

realistic sample size and the fact that transmission was high in the study area, so that sufficient genetic variation is circulating. For these samples, all SNPs within a window ranging from -4.3kb to 4.3kb around *Pfdhfr* on chromosome 4 were extracted. The majority of these SNPs were monomorphic in this dataset. Starting from a window ranging from -0.1kb to 0.1kb around *Pfdhfr*, window size was increased in steps of 0.1kb upstream and downstream (cf. Fig 6). Samples without missing data in the respective window were retained. Hence, sample size decreased with increasing window size (the maximum sample size was $N = 228$ and the minimum $N = 33$). For larger window sizes, the number of polymorphic markers leads to memory overflow (R crashed on macOS as the system tried to allocate too much working memory to R). Notably, in practice sequencing errors need to be eliminated first in such data, which likely decreases the number of polymorphic markers, suggesting that SNPs within larger region can be included. The decay in sample size (cf. Fig 6) can be counteracted by imputing missing values. However, the experiment shows the computational feasibility of the method for relevant applications.

Data application

Estimates of haplotype frequencies and MOI can be utilized for further population genetic analyses. This is illustrated using a *P. falciparum* dataset from Yaoundé, Cameroon [37]. The data was collected in 2001, 2002, 2004, and 2005. In the following analysis, the data is stratified into two groups: years 2001–2002 ($N = 166$) and 2004–2005 ($N = 165$). The dataset consists of markers associated with resistance against sulfadoxine-pyrimethamine (SP) in *P. falciparum*. Resistance to the fast-acting component pyrimethamine is determined by mutations at codons 51, 59, 108, and 164 in the *Pfdhfr* gene, while resistance to the slow-acting partner drug sulfadoxine is determined by mutations at codons 436, 437, 540, 581, and 613 in the *Pfdhps* gene [37].

The following analysis aims to estimate heterozygosity and pairwise-LD across 18 microsatellite markers on chromosomes 4 around *Pfdhfr*, 15 microsatellite markers around *Pfdhps*, as well as 8 neutral markers on chromosomes 2 and 3 [37]. Such analyses are useful to understand the spread of antimalarial drug resistance.

In standard population-genetic analyses, one considers past evolutionary processes, i.e., the variants under selection already reached fixation. This is different in the case of antimalarial drug resistance. Namely, one is interested in ongoing evolutionary processes, in which the mutations of interest have not yet reached fixation. Hence, all analyses need to be conditioned on the variants of interest. Otherwise, patterns of genetic hitchhiking and LD would be obstructed. In the

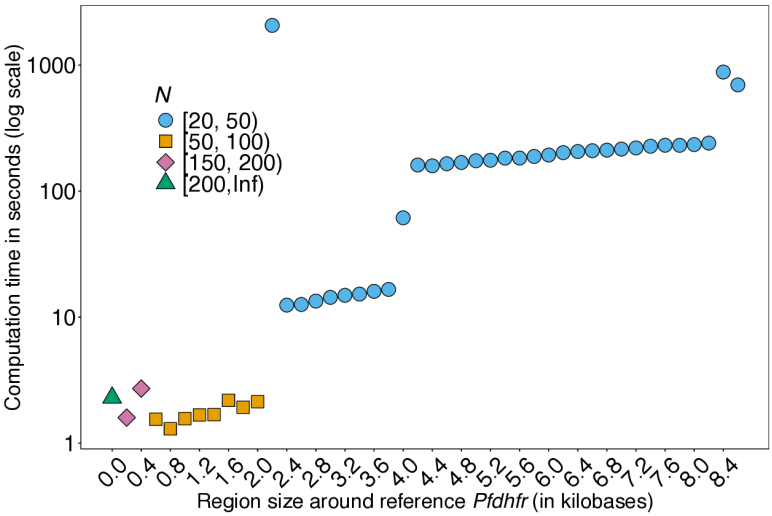

Fig 6. Computational time benchmark on whole genome sequencing data: Shown is the computational time in seconds (y-axis; log scale) as a function of the size of the window around *Pfdhfr* (up- and downstream; x-axis). Each point shows the time for one of the corresponding datasets in a window from 0 to 8.6kb. Ranges of sample size are shown in different colors and shapes.

https://doi.org/10.1371/journal.pcbi.1012955.g006

present case, the variants of interest are haplotypes determined by *Pfdhfr* and *Pfdhps* mutations. To obtain a reasonable sample size, the following analyses are conditioned on haplotypes surpassing a threshold frequency of 10%.

To determine the frequencies of *Pfdhfr* and *Pfdhps* haplotypes, their frequencies were first estimated with the proposed method.

Frequency of resistance-associated haplotypes and MOI. The frequencies of *Pfdhfr* and *Pfdhps* haplotypes at the two-time points are given in Table 2. At *Pfdhfr*, only the triple-mutant 51**I**/59**R**/108**N**/I164, conferring strong resistance, was detected with a frequency higher than 10% at both time points. This suggests widespread resistance to SP in Yaoundé from 2001 to 2005. At *Pfdhps*, the haplotypes 436**A**/A437/K540/A581/A613, S436/437**G**/K540/A581/A613 were detected at a frequency higher than 10% in the years 2001–2002, while the double-mutant haplotype 436**A**/437**G**/K540/A581/A613 also surpassed the 10% threshold in 2004–2005 suggesting ongoing selection for sulfadoxine resistance.

The estimates of the MOI parameter and 95% bootstrap confidence intervals (CIs) are presented in Table 3. Estimates were obtained at both time points, based on *Pfdhfr* markers, and *Pfdhps* markers, as well as the combined *Pfdhfr* and *Pfdhps* markers. The merit of the latter is the inclusion of more information, at the cost of a deflated sample size due to missing data. Because the narrowest CIs are obtained in this case, the gain of information outweighs the reduction in sample size. (This is not a general principle but will depend on the particular dataset.) The widest CIs are obtained for the estimates based only on *Pfdhfr* markers. The reason is the unbalanced haplotype frequency distribution (cf. Table 2).

Between 2001–2002 and 2004–2005, a slight reduction in MOI seems to be observed, although the CIs overlap by a great extent. Only for the estimates based on *Pfdhps* markers, there seems to be a very slight increase in MOI. This is presumably because the haplotype frequency distribution was less balanced in 2004–2005 (which also resulted in slightly wider CIs).

Note that it is preferable to estimate MOI from neutral markers rather than those under selection (cf. [27]). Allele frequency distribution at markers under selection will tend to be more unbalanced, with elevated major allele frequencies. (The selected alleles might even be fixed, or – as in the case of *Pfdhfr* codon 164 or *Pfdhps* codon 540 – the potentially selected mutant has not yet emerged.) Hence, the true underlying MOI will be poorly reflected in observations, resulting in frequent underestimations of MOI – and occasional outliers will lead to substantial overestimates (cf. [10,12]).

Table 2. Frequencies estimates of *P. falciparum* haplotypes associated with SP-resistance from data collected in Cameroon in years 2001-2002, and 2004-2005. Haplotype frequencies above the 10% threshold are marked with an asterisk (*) in both year groups.

	Pfdhfr						*Pfdhps*						
Years	Haplotype (codons)				Freq. (%)	95% CI	Haplotype (codons)					Freq. in %	95% CI
	51	59	108	164			436	437	540	581	613		
2001-2002	N	C	S	I	9.7	5.9–13.7	S	A	K	A	A	6.2	3.0-9-.9
	N	C	**N**	I	2.0	0.4–4.0	**A**	A	K	A	A	33.6*	27.4–40.3
	N	**R**	S	I	0.4	0.0–1.3	S	**G**	K	A	A	45.9*	38.9–52.9
	I	C	**N**	I	2.1	0.4–4.1	S	A	K	A	**S**	2.0	0.3–5.0
	N	**R**	**N**	I	9.8	6.2–13.7	**A**	**G**	K	A	A	7.5	4.0–11.4
	I	**R**	**N**	I	76.0*	69.9–81.9	**A**	A	K	A	**S**	1.3	0.0–2.9
							S	**G**	K	A	**S**	2.7	0.4–5.1
							S	**G**	K	**G**	**S**	0.4	0.0–1.3
2004-2005	N	C	S	I	2.6	0.7–4.9	S	A	K	A	A	6.3	3.1–10.0
	N	C	**N**	I	0.9	0.0–2.4	**A**	A	K	A	A	27.6*	21.6–33.5
	I	C	S	I	0.9	0.0–2.3	S	**G**	K	A	A	50.4*	43.2–57.6
	I	C	**N**	I	0.4	0.0–1.4	S	A	K	A	**S**	0.8	0.0–2.4
	I	**R**	S	I	2.6	0.9–4.5	**A**	**G**	K	A	A	14.5*	9.4–19.9
	N	**R**	**N**	I	3.9	1.6–6.9	**A**	A	K	A	**S**	0.4	0.0–1.9
	I	**R**	**N**	I	88.7*	84.0–93.0							

https://doi.org/10.1371/journal.pcbi.1012955.t002

Table 3. Mean MOI estimates ($\hat{\psi}$) from data collected in Cameroon in years 2001-2002, and 2004-2005.

	Pfdhfr		*Pfdhps*		*Pfdhfr* & *Pfdhps*	
Years	MOI	95% CI	MOI	95% CI	MOI	95% CI
2001-2002	1.58	1.39–1.83	1.44	1.31–1.58	1.54	1.42–1.68
2004-2005	1.42	1.22–1.76	1.45	1.32–1.62	1.49	1.36–1.65

https://doi.org/10.1371/journal.pcbi.1012955.t003

Prevalence of resistance-associated haplotypes. For completeness, also estimates of haplotype prevalence are presented (Table 4). Due to the moderate value of the MOI parameters, frequency and prevalence are similar. Most discrepancies occur for the *Pfdhps* haplotypes in 2001–2002. This is clear from (4c) because the *Pfdhps* have the most balanced frequency distribution.

Although the frequency of the single *Pfdhps* mutant haplotype S436/437**G**/K540/A581/A613 is below 50% in 2001–2002, its prevalence is estimated to be larger than 50%, implying that this variant was present in more than half of the infections.

Genetic hitchhiking. Conditioned on the SP-resistant *Pfdhfr* triple-mutant haplotype 51**I**/59**R**/108**N**/I164, heterozygosity estimates are shown for the years 2001–2002 and 2004–2005 in Fig 7A and 7B, respectively. Heterozygosity was calculated for each microsatellite marker separately to retain the maximum possible sample size. (This is preferable because considering all microsatellite markers at the same time to calculate haplotype frequencies and then marginalizing to obtain allele frequency spectra for each marker would substantially deplete sample size. Namely, only samples without missing data at all loci could be retained. Nevertheless, analyzing markers separately still requires a method that allows for a general genetic architecture, since it is performed conditioned on *Pfdhfr* and *Pfdhps* haplotypes.)

Heterozygosity at markers surrounding *Pfdhfr* is low, compared with that at more distant markers, exposing a clear selective sweep, i.e., traces of selection acting on *Pfdhfr*. Heterozygosity appears slightly higher in 2004–2005 (compare Fig 7A and 7B). Although the differences are within the margin of error, heterozygosity is expected to increase over time.

Table 4. Prevalence estimates of P. falciparum haplotypes associated with SP-resistance from data collected in Cameroon in years 2001-2002, and 2004-2005.

	Pfdhfr						*Pfdhps*						
Years	Haplotype (codons)				Prev. (%)	95% CI	Haplotype (codons)					Prev. in %	95% CI
	51	59	108	164			436	437	540	581	613		
2001-2002	N	C	S	I	14.5	9.1–20.0	S	A	K	A	A	8.7	4.3 – 13.8
	N	C	**N**	I	3.1	0.6–6.1	**A**	A	K	A	A	42.5	34.9–50.1
	N	**R**	S	I	0.6	0.0–2.0	S	**G**	K	A	A	55.5	47.7–63.0
	I	C	**N**	I	3.3	0.7–6.4	S	A	K	A	**S**	2.8	0.5–7.1
	N	**R**	**N**	I	14.7	9.5–20.5	**A**	**G**	K	A	A	10.4	5.6–15.7
	I	**R**	**N**	I	84.2	78.5–89.6	**A**	A	K	A	**S**	1.8	0.0–4.2
							S	**G**	K	A	**S**	3.8	0.6–7.2
							S	**G**	K	**G**	**S**	0.6	0.0–1.8
2004-2005	N	C	S	I	3.6	1.1–6.6	S	A	K	A	A	9.0	4.5–14.3
	N	C	**N**	I	1.2	0.0–3.1	**A**	A	K	A	A	35.9	28.6–43.6
	I	C	S	I	1.2	0.0–3.1	S	**G**	K	A	A	60.2	52.5–68.1
	I	C	**N**	I	0.6	0.0–1.9	S	A	K	A	**S**	1.2	0.0–3.5
	I	**R**	S	I	3.7	1.2–6.8	**A**	**G**	K	A	A	19.8	13.5–26.7
	N	**R**	**N**	I	5.5	2.4–9.2	**A**	A	K	A	**S**	0.6	0.0–2.6
	I	**R**	**N**	I	92.1	87.8–95.8							

https://doi.org/10.1371/journal.pcbi.1012955.t004

A drop in selection pressure could amplify this since treatment policy abandoned SP as first-line treatment in Cameroon in 2004 [37].

Such patterns are not as pronounced in *Pfdhps* but are still visible. The depletion in heterozygosity surrounding *Pfdhps* is slightly more pronounced around the S436/437**G**/K540/A581/A613 compared with the 436**A**/A437/K540/A581/A613 mutation at both time points, which corresponds to stronger selection for the S436/437**G**/K540/A581/A613 haplotype. The hitchhiking pattern flanking *Pfdhps* haplotypes corresponds to "soft selective sweeps" from recurrent mutations [38], i.e., the resistance-associated mutations emerged *de novo* or were imported several times at the background of different haplotypes. Such a soft-sweep pattern (with even higher heterozygosity) is also observed on the background of the 436**A**/437**G**/K540/A581/A613 double mutant (see Fig 7G). It is plausible that the A→**G** mutation at codon 437 occurred on the background of the single-mutant 436**A**/A437/K540/A581/A613, while at the same time, the S→**A** mutation at codon 436 occurred on the background of the S436/437**G**/K540/A581/A613 haplotype, leading to soft-sweeps from independent origins, which therefore retains the levels of genetic variation.

The patterns of genetic hitchhiking in combination with the frequency distribution of *Pfdhfr* and *Pfdhps* haplotypes suggest stronger selection for pyrimethamine resistance. This is intuitive since pyrimethamine is considered the fast-acting component of SP, and targeting *Pfdhps* with sulfonamides alone would not be sufficient for malaria chemotherapy. Pyrimethamine was used as a monotherapy in the 1950s in some countries [39]. This suggests that resistance to pyrimethamine started to evolve earlier than resistance against sulfadoxine.

Pairwise linkage-disequilibrium. Conditional pairwise LD-values are obtained using r^2 (and D' as a comparison). LD is calculated at both time points for each pair of markers conditioned on the *Pfdhfr* triple mutant 51**I**/59**R**/108**N**/I164 at *Pfdhfr* (Fig 8A, 8B), and the *Pfdhps* single mutants 436**A**/A437/K540/A581/A613 (Fig 8C, 8D) and S436/437**G**/K540/A581/A613 (Fig 8E, 8F), as well as on the *Pfdhps* double mutant 436**A**/437**G**/K540/A581/A613 at the second time point (Fig 8G).

Multiallelic LD measures are notoriously difficult to interpret and often uninformative. Conditioned on the *Pfdhfr* triple mutant, only slight LD is observed flanking *Pfdhfr* and *Pfdhps* (Fig 8A, 8B). This suggests that the triple mutant was

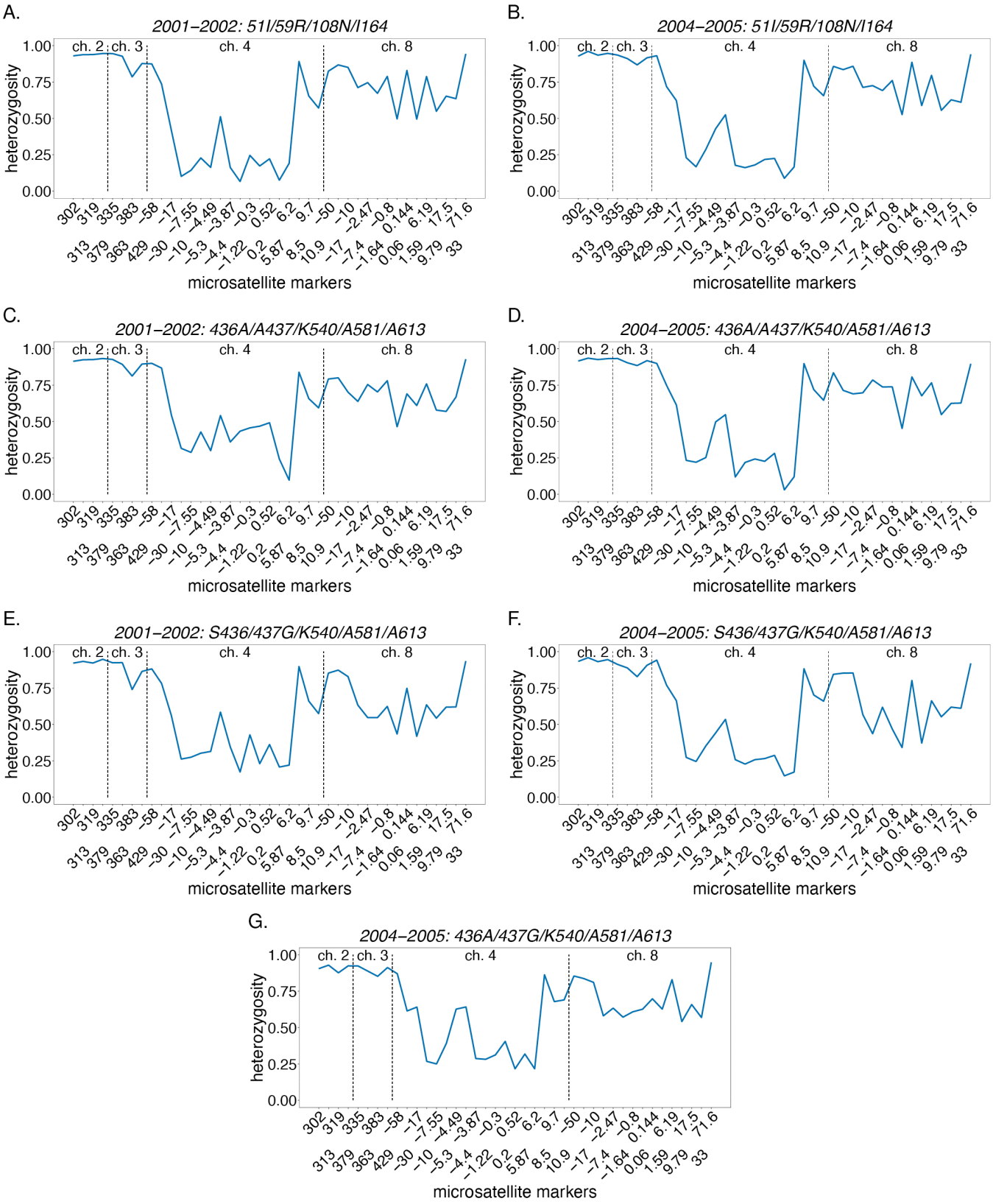

Fig 7. Conditional heterozygosity at microsatellite markers: Shown are estimates of heterozygosity across several microsatellite markers located on chromosomes 2 and 3, as well as on chromosome 4 surrounding *Pfdhfr* and chromosome 8 surrounding *Pfdhps* (chromosomes are separated by the

vertical dashed lines) conditioned on various *Pfdhfr* and *Pfdhps* haplotypes either in the years 2001-2002 or 2004-2005 (shown at the top of each panel). Markers are ordered by their position in kilobases on the chromosomes. On chromosomes 4 and 8, the positions are relative to *Pfdhfr* and *Pfdhps*.

https://doi.org/10.1371/journal.pcbi.1012955.g007

predominant for long enough to allow recombination to restore linkage equilibrium (LE). Note that microsatellite markers are highly variable and fast-evolving, hence recombination and mutation will tend to restore genetic variation. This is particularly true given the results on heterozygosity, which suggest soft selective sweeps around *Pfdhps* variants. This is confirmed by looking at LD conditioned at the *Pfdhps* mutations. Conditioned on these, pairwise LD is more pronounced around *Pfdhfr* than around *Pfdhps*, on the one hand reflecting the soft sweep nature of *Pfdhps* mutations, on the other hand exposing signatures of selection around *Pfdhfr* (most haplotypes are linked to the triple mutant, which is predominant). This suggests that different variants flanking *Pfdhfr* were affected by the soft sweeps at *Pfdhps*. These patterns will overlap and camouflage the traces of selection when conditioning only on the *Pfdhfr* triple mutant. (Note that conditioning on joint *Pfdhfr* and *Pfdhps* haplotypes is impractical here due to missing data, which would substantially reduce sample size.)

The patterns of LD flanking *Pfdhfr* and *Pfdhps* are most pronounced conditioned on the *Pfdhps* double mutant, which (as it is mainly linked to the predominant *Pfdhfr* triple mutant) experiences the highest selective pressure.

Although multiallelic LD in terms of r^2 is hardly informative on the underlying evolutionary process in this case, it is at least intuitive. This is different for the D' measure (Fig 9). In fact, D' leads to counterintuitive results, with the highest levels of LD between neutral markers at chromosomes 2 and 3, and lower LD flanking the loci under selection. These observations are purely an artifact of a small sample size. Namely, a large number of alleles are segregating at these markers. For example, two markers with, e.g., $n_1 = 20$ and $n_2 = 25$ alleles, would lead to 500 possible haplotypes. If all haplotypes had a frequency of 0.002, the markers would be in perfect LE. However, when taking a sample of size $N = 165$, the vast majority of haplotypes will not be sampled, while it is highly likely to sample all alleles at both markers. This leads to high LD values (i.e., most of the terms $\frac{|D_{ij}^{(kl)}|}{D_{\max}^{(kl)}}$ in (64a) equal to 1). Thus, D' will tend to overestimate LD for highly polymorphic markers unless sample size is unrealistically large. This problem does not occur around markers flanking targets of selection, because polymorphism is reduced, i.e., the number of alleles segregating at these markers is substantially lower. This artifact is also seen when comparing D' conditioned on the *Pfdhfr* triple mutant *vs.* other mutants which are subject to weaker selection (Fig 8A, 8B *vs.* 8C-8G). Namely, genetic hitchhiking, which in the case of malaria can be genome-wide [40], reduces polymorphism at neutral markers. This effect is more pronounced, conditioned on variants under stronger selection, thereby mitigating the undesired behavior of D' for highly polymorphic markers in small samples.

Discussion

Due to advancements in molecular/genetic assays over the past decades, molecular disease surveillance has become increasingly feasible and popular. Here, a method to estimate pathogen variant (haplotype) frequencies and multiplicity of infection (MOI) was introduced.

The method extends those of [8,10,12] to accommodate a general genetic architecture, i.e., pathogen variants are characterized by an arbitrary number of multiallelic markers (loci). Following [8–10,12,31], MOI is defined as the number of super-infections with the same or different pathogen variants (cf. [2] for a detailed discussion). This only approximates co-infections, i.e., the co-transmission of different pathogen variants during the same infective event (cf. Fig 10). From a mathematical point of view, the statistical model estimates the distribution of independent pathogen variants within an infection (which is assumed to follow a conditional Poisson distribution). The assumption of independence reflects the concept of super-infections, which are independent, whereas the pathogen variants being co-transmitted are not independent. If co-infections are sufficiently rare, the assumption of independent variants is justified.

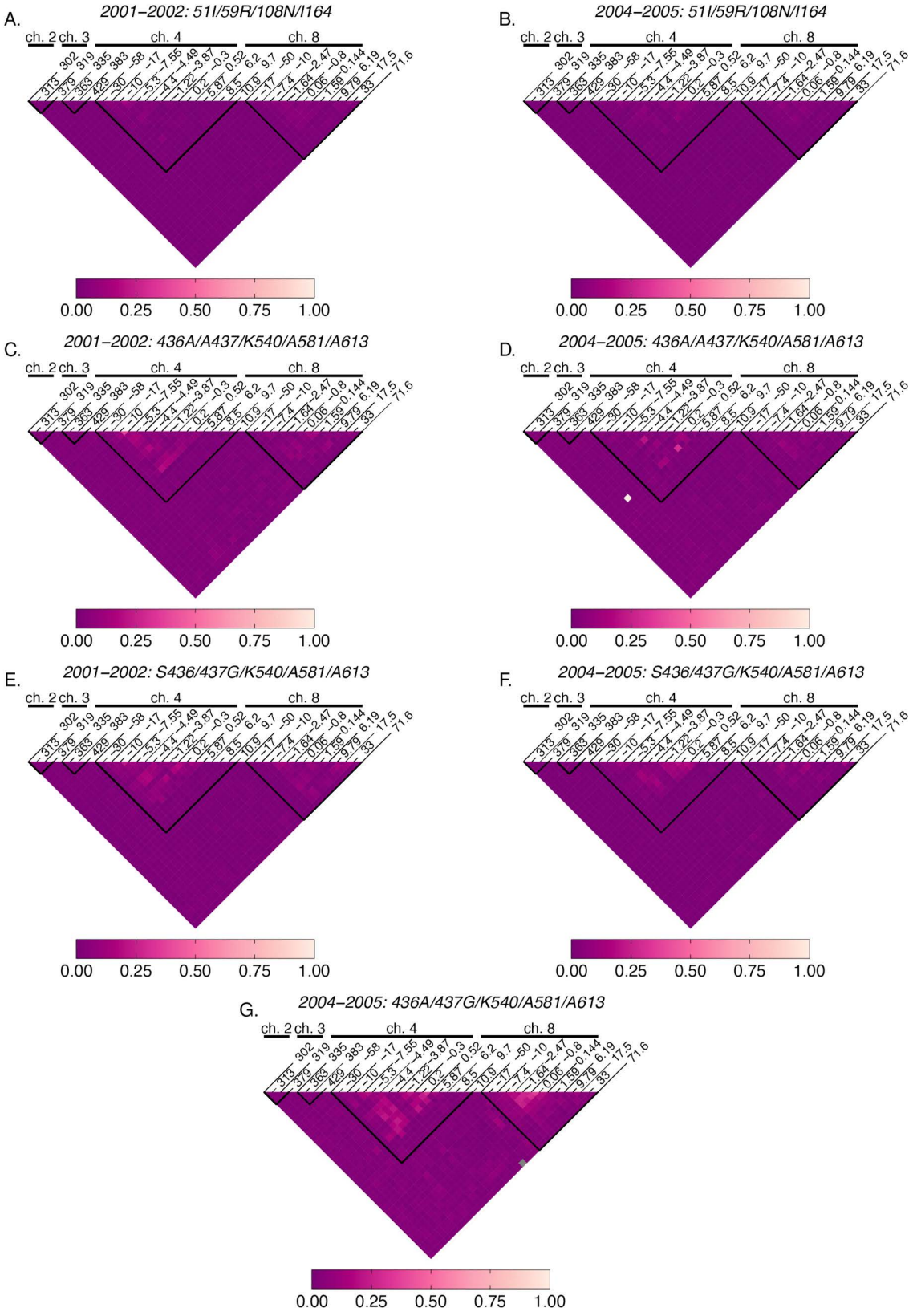

Fig 8. Conditional LD at microsatellite markers: As in Fig 7 but for pairwise-LD values calculated with r^2.

https://doi.org/10.1371/journal.pcbi.1012955.g008

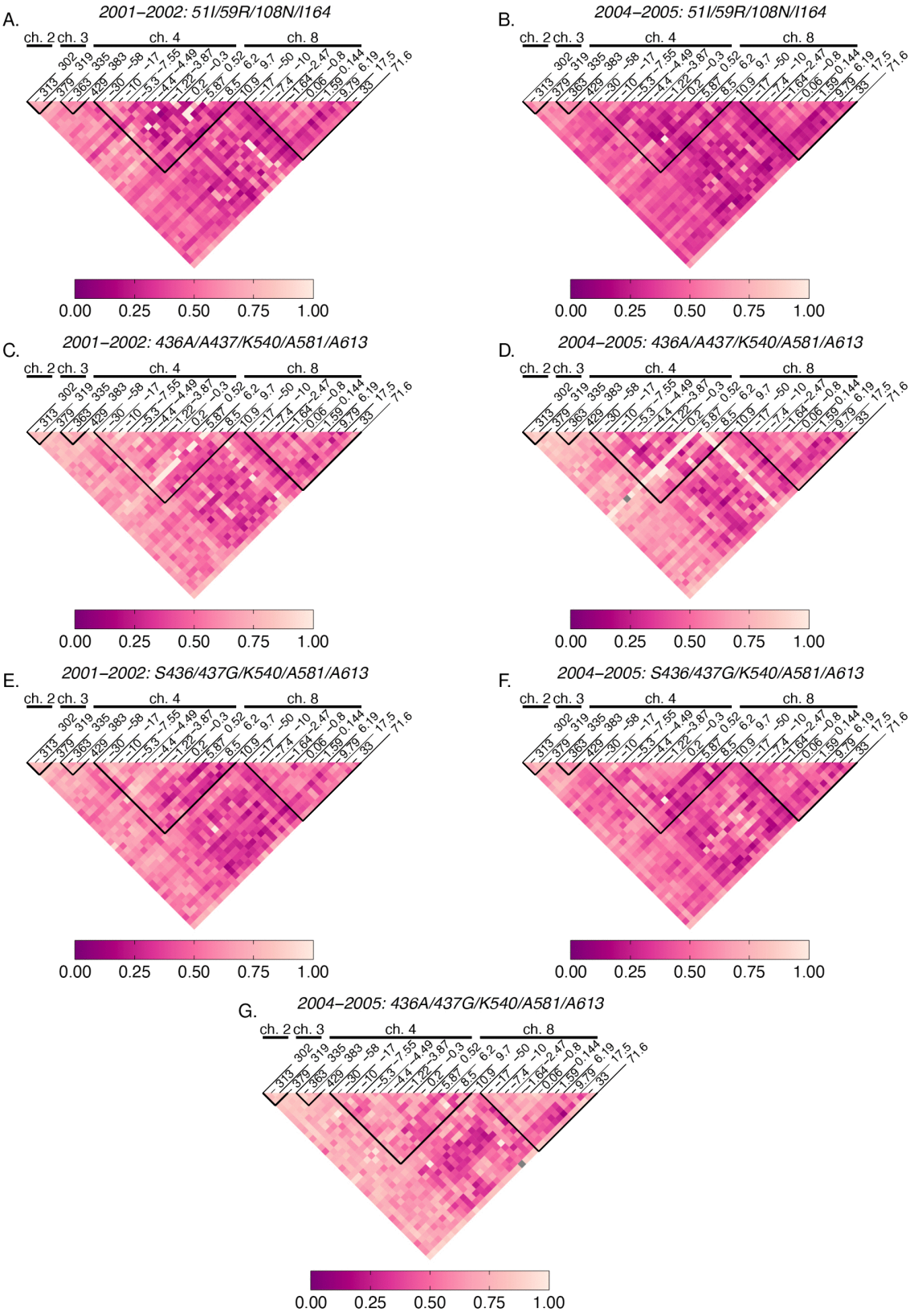

Fig 9. Conditional LD at microsatellite markers: As in Fig 7 but for pairwise-LD values calculated with D'.

https://doi.org/10.1371/journal.pcbi.1012955.g009

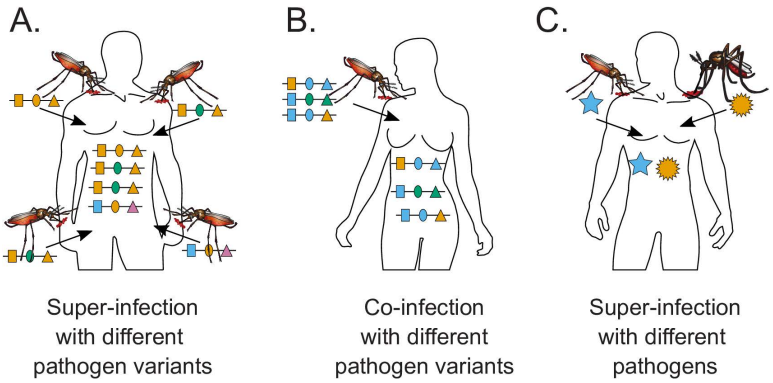

Fig 10. Super- and co-infections: Illustrated is the difference between super- and co-infections in the case of vector-borne diseases. (A) shows 4 super-infections (MOI = 4) with pathogenic variants, i.e., four independent infective events. At each infective event, one pathogenic variant is transmitted. Pathogenic variants are characterized genetically by their allelic expressions (colors) at three positions (shapes) in the genome, which is illustrated by the horizontal lines. Note that MOI = 4, although only three distinct haplotypes are transmitted because two vectors transmit the same pathogenic variant. (B) illustrates a co-infection with three pathogenic variants, i.e., a single infective event at which three pathogenic variants are transmitted. (C) illustrates a super-infection with two different pathogens, illustrated by different shapes, transmitted by different vector species.

https://doi.org/10.1371/journal.pcbi.1012955.g010

The applications provided here are concerned with *P. falciparum*. However, the method is applicable to all human-pathogenic malaria species (cf. [41] for a more detailed discussion). Furthermore, the method is also applicable to infections caused by other eukaryotic parasites such as *Trypanosoma*, *Toxoplasma*, and *Schistosoma*, and bacteria such as *Mycobacterium*, which evolve at a rate at which it is possible to determine a stable number of genetically distinct variants during the course of an infection.

First, the underlying probabilistic model was derived. Because it does not allow a closed-form solution, the EM algorithm was employed to derive an efficient numerical iteration to obtain the maximum likelihood estimates. Furthermore, formulae for prevalence were derived.

Whereas frequency refers to the relative abundance of a pathogen variant in the pathogen population, prevalence refers to the probability that the variant is present in an infection. Prevalence is more relevant for clinical and epidemiological considerations. Although the terms prevalence and frequency are sometimes used synonymously in the literature, they are different. In the absence of multiple infections, frequency and prevalence coincide. However, due to MOI prevalence always exceeds frequency (cf. [2]).

Commonly, molecular/genetic data is not phased, i.e., haplotype information is missing. As a consequence, molecular data is ambiguous regarding the haplotypes actually present in an infection. In the case of malaria, often heuristic ad-hoc approaches are used to derive haplotype prevalence based only on unambiguous information. To relate the formal statistical approach to these heuristic ones, formulae for prevalence from unambiguous observations were also derived here.

The asymptotic covariances, i.e., the inverse Fisher information, were derived for the model parameters (haplotype frequency and MOI parameter) and their transforms in terms of prevalence and mean MOI. Although the Fisher information was derived explicitly, it was not possible to provide analytic expressions for its inverse. Importantly, two alternative approaches to derive the Fisher information were employed. First, a nuisance parameter was introduced to embed the parameter space into a higher-dimensional open set in a real space. Second, a redundant parameter was eliminated (the haplotype frequencies are elements of the simplex and hence dependent) to reduce the parameter space to an open set in a lower-dimensional real space. This leads to different information matrices. It was shown that both approaches are equivalent and that corresponding entries of the inverse information matrices coincide. The same arguments apply to the

observed information. Although it is not explicitly derived here, the steps are explained, and it is incorporated in the implementation in R.

By numerical simulations, [12] and [10] already showed that the method has little bias in special cases, which vanishes with increasing sample size, suggesting that the method is asymptotically unbiased. Here, these results were complemented, focusing on the bias of haplotype frequencies. Furthermore, it was numerically verified that the covariances of the estimator are well approximated by the inverse Fisher information (Cramér-Rao lower bound; CRLB). Simulations showed that the variance of the estimator (for haplotype frequencies and MOI) agrees well with the CRLB even for moderate sample sizes ($N \geq 100$). The largest discrepancies occur if sample size is small ($N < 100$) and MOI is high. This is less problematic in practice since larger sample sizes should be feasible in high transmission areas. The CRLB defines the minimum variance of an unbiased estimator. The fact that the covariance matrix of the estimator is well approximated by the inverse Fisher information for finite sample size suggests that the proposed method has desirable asymptotic properties.

Application of the method was illustrated by analyzing a *Plasmodium falciparum* dataset consisting of genetic markers associated with resistance to sulfadoxine-pyrimethamine (SP), caused by point mutations in the *Pfdhfr* and *Pfdhps* genes. First, the frequencies and prevalence of resistance-associated haplotypes were derived. In a second step, heterozygosity and pairwise linkage disequilibrium (LD) across microsatellite markers flanking the loci associated with resistance were calculated. These calculations were performed conditioned on specific *Pfdhfr* and *Pfdhps* haplotypes. These analyses were performed to showcase that the MOI and haplotype frequency estimates can be utilized for further population genetic analysis. The example also highlighted the pathological behavior of one multiallelic measure of LD (*D*'). Importantly, the analysis required estimates of haplotypes, determined by a general genetic architecture (multiple multiallelic markers). The special cases of the proposed method published earlier [9,10,12] are insufficient for such analyses. The method in [9] allows only for a single multiallelic marker and is only appropriate to estimate allele frequencies and overall heterozygosity (i.e., not stratified on specific mutational backgrounds). In [12], multiple biallelic markers are assumed, which is sufficient to derive frequencies of resistance-associated haplotypes (as long as they are characterized by biallelic markers), but insufficient to derive heterozygosity of multiallelic markers. Pairwise linkage-disequilibria (LD) can be derived using the special case in [10], however since only two multiallelic markers are allowed, pairwise LD can only be calculated for the whole population, but not stratified on a specific mutational background. The generalization here allows for such analysis.

The method is designed for haploid pathogens. This is hardly a restriction given that many pathogens, including viruses, bacteria, protozoa, etc., are haploid. Nevertheless, the method can be extended to be applicable to diploid or polyploid pathogens, although the relevance of such generalizations is limited.

Besides the promising properties of the proposed method, there are several shortcomings. Unlike in the case of a single molecular marker, there is no formal proof of existence, uniqueness, asymptotic unbiasedness, consistency, and efficiency. For a genetic architecture of a single molecular marker, the model can be rewritten as a natural steep exponential family, which proves the desired properties (importantly, existence and uniqueness hold except in degenerate cases, e.g., if one variant is found in all samples). For two or more markers, the sum in (48a) does not factorize accordingly. However, the agreement between the CRLB and the variance of the estimator determined by simulations suggests that the estimator is efficient. Moreover, the fact that bias decreases with sample size suggests asymptotic unbiasedness (which also suggests uniqueness of the estimator).

MOI is assumed to follow a conditional Poisson distribution, which might be inappropriate if MOI is over-dispersed. Intuitively, a negative binomial distribution seems more appropriate. However, in [21] it was pointed out that the likelihood estimation of negative binomial data is problematic, as the model seems to have degenerate behavior. (This readily occurs for maximum likelihood estimation of negative binomially distributed data. Specifically, the MLE will yield the limit of a Poisson distribution, if the data is not sufficiently over-dispersed.) As an alternative, [21] assumed a non-parametric distribution of MOI, which is the most flexible model. Except for extreme cases, this model did not outperform the Poisson model, suggesting that the assumption of Poisson-distributed MOI is justified.

The estimators can be improved by applying a bias correction as in [23] in the case of a single marker. This led to an improvement in the bias of the MOI parameter (the gain for marker frequencies is marginal, as these are hardly biased). Such bias corrections might be tedious in the present model. As an alternative, parametric or non-parametric bootstrap bias corrections can be applied (cf. [42]).

Another shortcoming of the model is its inability to handle missing data. Namely, observations that lack information at one or more markers need to be disregarded. This yields a tradeoff between the confidence in the estimates and the number of molecular markers included. Namely, while the addition of more markers increases information, it also deflates sample size due to missing data and increases the number of model parameters geometrically. For a single marker, [24] proposed a model that incorporates missing information. This method was only superior to ignoring missing information if these were common. Incorporating missing information for a general genetic architecture leads to combinatorial difficulties in practice. In particular, all possible observations would need to be considered. Assuming 10 markers with 10 alleles per marker yields $1.25 \cdot 10^{30}$ possible observations. For the same reason, a mutational model is not included in the current method. For computational feasibility, restrictions and approximations would need to be applied to the probabilistic model to render it numerically feasible. An alternative to incorporate missing data would be to impute missing information. This can be readily achieved by estimating the model parameters first by disregarding missing information, and then using these estimates to impute missing information for each sample separately based on the marginalization of the distribution over the markers with missing information. In the final step, the estimates would be updated by those of the imputed dataset. A similar approach can be conducted to include assay errors. After obtaining estimates from the original data, a parametric bootstrap approach can be used to generate simulated datasets subject to mutational models.

The method presented here is "exact" in the sense that (assuming no errors in the data) it includes all haplotypes which can theoretically exist. Alternative models, e.g., DEploid, assume a reference haplotype panel [19]. For a smaller genetic architecture, it is desirable to allow all haplotypes to avoid biasing the estimates towards higher MOI. For example, consider the case of 3 biallelic markers. If in a reference panel only the haplotypes (1,1,1), (1,2,1), and (2,1,2) are present, the observation ({1,2},{1,2},{1,2}) must have MOI $m \geq 3$. On the contrary, if all possible haplotypes are included, the minimum MOI is $m = 2$ (e.g., haplotypes (1,1,1) and (2,2,2) are infecting). However, the inclusion of all haplotypes becomes unrealistic for a complex genetic architecture. E.g., with 30 STR markers with 10 alleles each, 10^{30} haplotypes are possible. Assuming a maximum number of 10^{12} parasites within an infection and 500 million infections per year, the number of possible haplotypes exceeds the number of haplotypes that have ever existed by orders of magnitude. Hence, there is a tradeoff between when to use an exact or approximate method. Moreover, the exact method will yield some frequency estimates that are well below the numerical precision. These can be safely rounded to zero without jeopardizing the accuracy of the remaining frequencies. As a rule of thumb, whenever the data is numerically feasible for the current model implementation, the use of an "exact" is appropriate.

Concerning numerical feasibility, the method is appropriate for a moderate number of molecular markers, with a moderate number of alleles per marker. Whether the method is feasible for a particular genetic architecture depends crucially on the underlying dataset. As an example, assume 20 biallelic SNPs. Further assume that only two haplotypes are present, the wildtype and mutant (these differ at all 20 SNPs). If both haplotypes occur in an infection, the formula for the resulting observation considers all 1048576 possible haplotypes, which give rise to more than 3.5 billion summands in equation (48a). In practice, such observations do not occur, and a genetic architecture of more than 20 SNPs is feasible.

In any case, the supported genetic architectures are insufficient to study relatedness of haplotypes within an infection (cf. [43–48]). Relatedness is important to distinguish super- from co-infections. Here, co-infections are ignored and only approximated by super-infections (cf. [27] for a more detailed discussion). If co-infections dominate transmission (cf. [44]), the method here is no longer valid. In such a case, the assumption of independent haplotypes within infections is violated. To resolve this problem, knowledge of the distribution of co-infecting haplotypes would be required. Such information could be generated from additionally collecting samples from mosquitoes or from modelling host-vector and vector-host

transmission explicitly. In any case, even if the assumptions of the method are violated, it will mainly affect the estimates of MOI. The reason is that MOI is defined as the number of independent transmission events, and this will be detached from the number of circulating haplotypes in infections if they are frequently co-transmitted. However, phasing of the haplotypes will be less affected. Hence, even if relatedness within infections is important, the method can be used to phase haplotypes and obtain frequency estimates at the population level, but the MOI estimates should be ignored then.

Note that the model could be approximated by considering prior knowledge on haplotypes likely to be detected in the area of the study and disregarding those that likely do not exist. Making such assumptions can lead to substantially more efficient numerical implementations in terms of computational time and the complexity of feasible genetic architectures. Models such as DEploid [19] and coiaf [20] use a similar approach and rely on some prior knowledge of reference panels, and minor allele frequencies, respectively, to estimate haplotype frequencies and complexity of infection (COI) from a large number of biallelic markers (SNPs) in a reasonable time.

The proposed method does not correct for errors in the data, e.g., sequencing errors. While this is a limitation, it also avoids introducing bias. Namely, the method is not restricted to a particular type of molecular/genetic data. Errors occurring in STR and SNP data are different and have to be handled accordingly. Users who have reasons to question the quality of their data are advised to modify their data according to an appropriate mutational model several times and repeat the analysis for each modified dataset. This approach allows to ascertain how sensitive the model estimates are to errors in the data. This is of particular importance when aiming to analyze whole genome sequencing datasets, e.g., the open access dataset of *Plasmodium falciparum* genome variation from the MalariaGen project, consisting of 33 325 worldwide samples (cf. [28]).

Here, a maximum-likelihood approach was pursued. Bayesian alternatives can be readily implemented and should be in good agreement as long as an uninformative prior distribution is used.

Besides the limitations, the proposed method is valuable for population genetic analysis of infectious diseases, similar to malaria, for a limited amount of molecular data, which is frequently collected in practice. An efficient implementation of the model is provided as an R script in the supplement, on GitHub https://github.com/Maths-against-Malaria/generalModel.git and Zenodo at https://doi.org/10.5281/zenodo.14452432 (cf. [29]). Moreover, detailed documentation of the R script with examples is provided.

Methods

Model background

The objective is to estimate the distributions of multiplicity of infection (MOI) and pathogen lineages for infectious diseases similar to malaria. Here, maximum-likelihood estimation is used. Although the following is not confined to malaria, this example is used to motivate the statistical model.

Here, the term multiplicity of infection refers to the number of super-infections during one disease episode, i.e., to the number of independent infectious events with potentially different variants of the same pathogen (see Fig 10A). Importantly, it is assumed that during an infectious event, only one pathogen variant is transmitted. This is different from co-infections, which refer to the co-transmission of several pathogen variants during one infectious event (see Fig 10B). Importantly, super-infections with different pathogen variants rather than with different pathogens are considered (cf. Fig 10A and 10C).

In what follows, the terms “pathogens variants”, “lineages”, and “haplotypes” are used synonymously.

Statistical model

Genetic architecture. Pathogen haplotypes are characterized by their allelic configuration at L marker loci. Haplotypes are denoted by vectors $\boldsymbol{h} = (h_1, \ldots, h_L)$, where h_k represents one of the n_k possible alleles at locus k. This yields a total

of $H = \prod_{k=1}^{L} n_k = n_1 \cdot \ldots \cdot n_L$ possible haplotypes. The set of all possible haplotypes is given by $\mathscr{H} = \prod_{k=1}^{L} \{1, \ldots, n_k\}$ (see illustration Fig 11A).

In the following, it will be convenient to label (rank) haplotypes by the numbers $1, \ldots, H$. The rank of hapotype $\boldsymbol{h}$ is denoted by $r(\boldsymbol{h})$. A natural way of ranking is to assume that $\boldsymbol{h} = (h_1, \ldots, h_L)$ is the mixed-radix representation [49] of the rank $r(\boldsymbol{h})$ with base $(n_1, \ldots, n_L)$ (see Fig 11B). More precisely,

$$r(\boldsymbol{h}) = \sum_{k=1}^{L} (h_k - 1) g_k + 1, \tag{30a}$$

with

$$g_k := \begin{cases} \prod_{j=1}^{k-1} n_j & \text{if } k > 1, \\ 1 & \text{if } k = 1. \end{cases} \tag{30b}$$

Ranking haplotypes has the advantage that each haplotype $\boldsymbol{h}$ can be uniquely identified with its rank $r(\boldsymbol{h})$. While this is not necessary for the model derivation, it is the basis of an efficient model implementation.

The relative abundance of haplotype $\boldsymbol{h}$ in the pathogen's population is denoted by $p_{\boldsymbol{h}}$, or $p_k = p_{\boldsymbol{h}}$ if $\boldsymbol{h}$ is the k-th haplotype, i.e., if $k = r(\boldsymbol{h})$. Collectively, the vector of haplotype frequencies is denoted by

$$\boldsymbol{p} := (p_{\boldsymbol{h}})_{\boldsymbol{h} \in \mathscr{H}} = (p_1, \ldots, p_H).$$

This notation implies that the set of haplotypes is regarded as ordered by the haplotype ranks.

Importantly, the number H of possible haplotypes increases "geometrically" with the number of loci L and alleles n_k, $k \in 1, \ldots, L$. However, only a subset of possible haplotypes will be realized in the pathogen population, i.e., $p_{\boldsymbol{h}} = 0$ for some (or if L large, most) haplotypes. It is assumed that the genetic architecture already subsumes all possible haplotypes so that mutational events do not need to be considered. In malaria, the pathogen population (or rather its reference point) is the population of sporozoites within the mosquitoes' salivary glands [9,10].

Distribution of MOI. MOI, which is sometimes also referred to as complexity of infection (COI) [4,50,51], is defined here as the number of independent infectious events during one disease episode, assuming that exactly one pathogen variant (haplotype) is transmitted at each event (see Fig 12). Notably, MOI does not refer to the number of distinct pathogen variants within an infection but to the number of infectious events, i.e., a host can be infected with the same pathogen variant multiple times (see Fig 12). Because it cannot be decided in practice how often a given pathogen variant infected a host, MOI is an unobservable quantity. Note that this definition of MOI coincides with the one from [8,9] (cf. [2,23] for more discussion).

Let the distribution of MOI on the population level be denoted by

$$\kappa_m := P[\text{MOI} = m]. \tag{31}$$

In the following, only disease-positive individuals are considered, hence $\text{MOI} \geq 1$.

The intuitive assumption is that infective events are rare and independent. This implies that MOI follows a conditional Poisson distribution, i.e.,

$$\kappa_m = \frac{1}{e^{\lambda} - 1} \frac{\lambda^m}{m!}, \quad m \geq 1, \tag{32a}$$

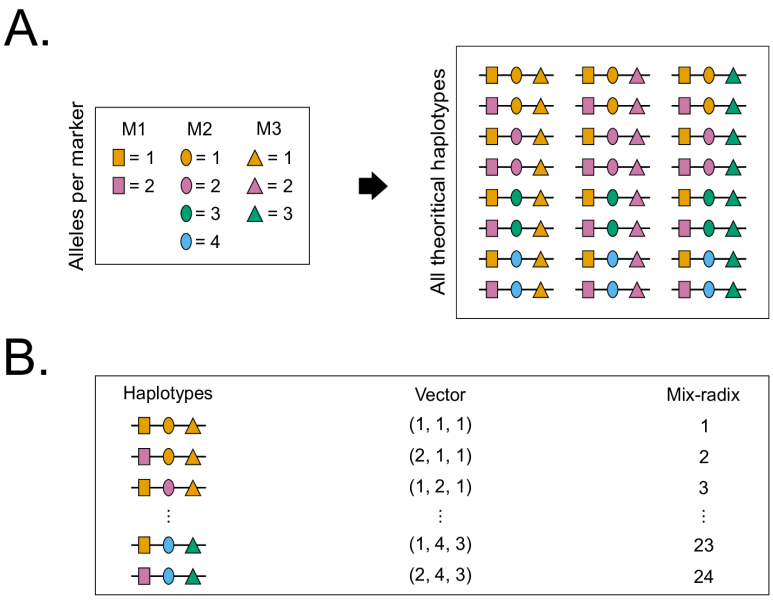

Fig 11. Genetic architecture and haplotype ranking: Panel (A) shows for a given genetic architecture (i.e., three loci with two, four, and three alleles at the first, second, and third locus, respectively), all the haplotypes that could be present in the pathogen population. Each shape represents a locus, and each color is an allele at a given locus. All the possible haplotypes alongside their vector and mix-radix representations, i.e., ranking, are shown in panel (B).

https://doi.org/10.1371/journal.pcbi.1012955.g011

where $\lambda > 0$ is the parameter characterizing the distribution. Clearly, other distributions can be assumed. In malaria, for instance, exposure to mosquitoes might be heterogeneous across different groups in the population. Even if MOI is Poisson distributed in each group, in the total population MOI would then correspond to a mixture of Poisson distributions. In the limit case of infinitely many population strata, such that the Poisson parameter follows a gamma distribution across the strata, MOI would follow a conditional negative binomial distribution. The negative binomial distribution at first might seem more appropriate than the conditional Poisson distribution, as it fits empirical observations of mosquito biting behavior better (cf. [52]). However, as noted in [21], maximum likelihood estimation of the negative binomial distribution is problematic (cf. also [53]) as the maximum-likelihood estimate does not exist if the empirical observations are not overdispersed. (Note also that a negative binomially distributed biting rate does not necessarily imply that the number of infective events is negatively binomially distributed – in fact, these might be well approximated by the Poisson distribution). As an alternative, [21] explored a non-parametric MOI distribution (in the simple case of one marker locus, $L = 1$). This is the most flexible assumption. Their results suggest that even for rather overdispersed true MOI distributions, a statistical model based on the Poisson distribution performs similarly to their non-parametric alternative. As a consequence, in what follows, it is assumed that MOI follows a conditional Poisson distribution.

The mean of the conditional Poisson distribution is given by

$$\psi := \mathbb{E}[\mathrm{MOI}] = \frac{\lambda}{1 - e^{-\lambda}}. \tag{32b}$$

This parameter also characterizes the conditional Poisson distribution and is more intuitive than λ, as it is the average number of times infected individuals are super-infected.

Infections. Since a host can be super-infected multiple times with the same haplotype, an infection corresponds to the MOI vector $\boldsymbol{m} := (m_{\boldsymbol{h}})_{\boldsymbol{h}\in\mathscr{H}}$, where $m_{\boldsymbol{h}}$ is the number of times the host was infected with haplotype $\boldsymbol{h}$. Because only one haplotype is transmitted per infectious event, the MOI value of infection $\boldsymbol{m}$ is $|\boldsymbol{m}| := \sum_{\boldsymbol{h}\in\mathscr{H}} m_{\boldsymbol{h}} = m_1 + \ldots + m_H$.

Super-infections are assumed to be independent, i.e., given a host is infected with the pathogen, the host is not more or less likely to be infected again during the same disease episode. As a consequence, the probability of observing infection $\boldsymbol{m}$, given it has MOI $|\boldsymbol{m}| = m$, follows a multinomial distribution with parameters m and $\boldsymbol{p}$, i.e., $\boldsymbol{m}|m \sim \text{Multi}(m, \boldsymbol{p})$. Hence,

$$P(\boldsymbol{m}\,|\,m) = \binom{m}{\boldsymbol{m}}\boldsymbol{p}^{\boldsymbol{m}} = \frac{m!}{m_1!\ldots m_H!}p_1^{m_1}\ldots p_H^{m_H}, \tag{33}$$

m haplotypes are drawn with replacement from the pathogen population according to their relative abundances $\boldsymbol{p}$.

Observations. As mentioned above, MOI is an unobservable quantity, because an infection $\boldsymbol{m}$ is itself unobservable, as it cannot be distinguished how often each haplotype was infecting. At most, the absence or presence of haplotypes can be observed, i.e., at most sign $m_{\boldsymbol{h}}$ is observable. However, in practice, pathogenic variants within an infection are determined by molecular assays. These typically generate unphased haplotype information, i.e., they generate consensus sequences that only distinguish the allelic variants at each marker locus, rather than determining the actual haplotypes present (see Fig 13 for an illustration). In other words, at each marker locus, only the absence and presence of the alleles are observable.

The observable allelic information in a disease-positive sample is denoted by the vector $\boldsymbol{x} = (x_1, \ldots, x_L)$, where x_k is the set of alleles detected at locus k, i.e.,

$$x_k \in \mathscr{P}(\{1, \ldots, n_k\}) \setminus \{\emptyset\}. \tag{34}$$

Here, $\mathscr{P}$ denotes the power set. Hence, the components of $\boldsymbol{x}$ are sets (corresponding to 0–1 vectors of length n_k; this correspondence is also illustrated in Fig 2). Note that x_k cannot be the empty set, because we only consider disease-positive samples, and we assume the molecular assays to be free of errors (no missing or wrong information, cf. [24]). Therefore, a proper sample is one for which at least one allele is detected at each of the L loci. Consequently, the observational space is given by the set of all possible observations, i.e., as

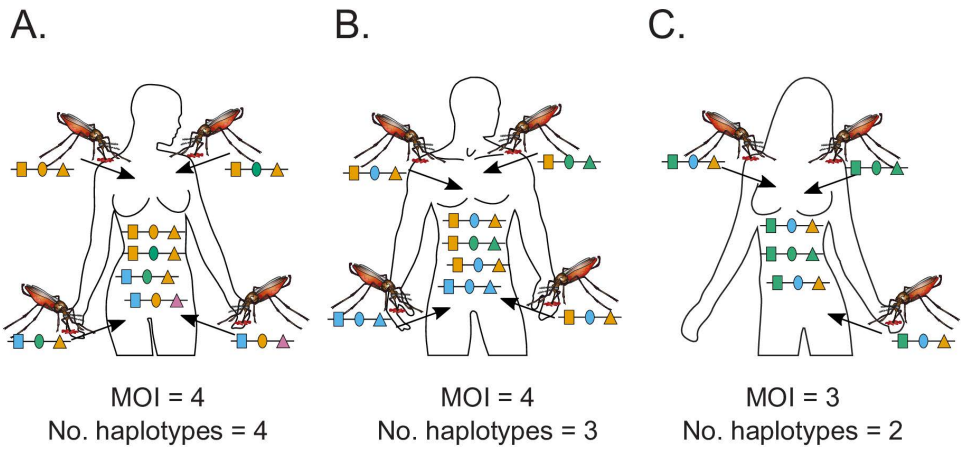

Fig 12. MOI and number of haplotypes: Illustrated are three hypothetical infections with pathogenic variants. Panel (A) shows four super-infections (cf. Fig 10A), i.e., MOI = 4, with four different haplotypes. Panel (B) shows four super-infections (MOI = 4) with three different haplotypes. Panel (C) illustrates three super-infections (MOI = 3) with two different pathogenic variants.

https://doi.org/10.1371/journal.pcbi.1012955.g012

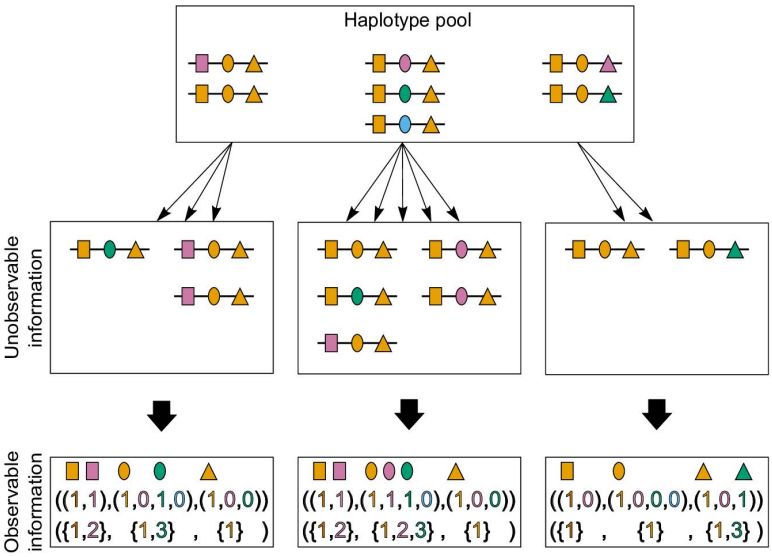

Fig 13. Observable and unobservable information: Illustration shows three different infections from the same pathogen population. The first infection (middle left) describes a super-infection with two haplotypes, i.e., MOI = 2. The haplotype combination generates unphased haplotype information, i.e., ambiguous observation (bottom left). In this case, haplotype present in the observation cannot be reconstructed, as well as the corresponding MOI. The second infection (middle), illustrates a super-infection with four haplotypes, three of which are transmitted once, while the fourth one is transmitted twice, i.e., MOI = 5. The resulting observation and MOI are ambiguous. The last infection (middle right) involves two haplotypes, i.e., MOI = 2, with resulting unphased haplotype information (which is in thic case unambiguous).

https://doi.org/10.1371/journal.pcbi.1012955.g013

$$\mathscr{O} := \prod_{k=1}^{L} \left(\mathscr{P}(\{1, \ldots, n_k\}) \setminus \{\emptyset\} \right). \tag{35}$$

Note that different infections $\boldsymbol{m}$ can yield the same observation $\boldsymbol{x}$ (see Fig 1). Given MOI m, the set of all possible MOI vectors $\boldsymbol{m}$ vectors that lead to $\boldsymbol{x}$ is denoted by

$$M_{\boldsymbol{x}}^{(m)} := \{\boldsymbol{m} \,|\, \boldsymbol{m} \to \boldsymbol{x}, |\boldsymbol{m}| = m\}. \tag{36}$$

Distribution of observations. Two definitions are helpful to derive the probability distribution of the observations $\boldsymbol{x} \in \mathscr{O}$. First, the set of all haplotypes which are compatible with observation $\boldsymbol{x}$, i.e.,

$$A_{\boldsymbol{x}} := \{\boldsymbol{h} \in \mathscr{H} \,|\, h_k \in x_k \text{ for all } k\}. \tag{37a}$$

(When using the correspondence of x_k to a 0–1 vector, $h_k \in x_k$ corresponds to $x_{k,h_k} = 1$, where x_{k,h_k} is the h_k-th component of the 0–1 vector.) Hence, if $\boldsymbol{h} \in A_{\boldsymbol{x}}$, haplotype $\boldsymbol{h}$ is potentially present in the infection underlying observation $\boldsymbol{x}$, while $\boldsymbol{h} \notin A_{\boldsymbol{x}}$ cannot be present. Second, the set of all sub-observations of $\boldsymbol{x}$, i.e.,

$$\mathscr{A}_{\boldsymbol{x}} := \{\boldsymbol{y} \in \mathscr{O} \,|\, y_k \subseteq x_k \text{ for all } k\}. \tag{37b}$$

(When using the correspondence of x_k and y_k to a 0–1 vector, $y_k \subseteq x_k$ corresponds to $x_k \leq y_k$, where the inequality is understood componentwise.) Hence, for a sub-observation $\boldsymbol{y}$ in $\mathscr{A}_{\boldsymbol{x}}$, any haplotype $\boldsymbol{h}$ that is compatible with observation $\boldsymbol{y}$ is also compatible with observation $\boldsymbol{x}$, i.e., if $\boldsymbol{y} \in \mathscr{A}_{\boldsymbol{x}}$, then $\boldsymbol{h} \in A_{\boldsymbol{y}} \Rightarrow \boldsymbol{h} \in A_{\boldsymbol{x}}$.

The probability of observation $\boldsymbol{x}$ given MOI = $m > 0$ is

$$P(\boldsymbol{x} \mid m) = \frac{P(\boldsymbol{x}, m)}{P(m)},\ m \geq 1, \tag{38a}$$

where $P(\boldsymbol{x}, m)$ is the probability of $\boldsymbol{x}$ descending from an infection with MOI = $m > 0$ and $P(m) = \kappa_m$ is the probability of MOI = m. Recall that for $\boldsymbol{m}$ in $M_{\boldsymbol{x}}^{(m)}$, $\boldsymbol{m}$ yields $\boldsymbol{x}$. Hence, the law of total probability gives

$$P(\boldsymbol{x}, m) = P(\boldsymbol{x} \mid m)P(m) = \sum_{\boldsymbol{m} \in M_{\boldsymbol{x}}^{(m)}} P(\boldsymbol{m} \mid m)P(m),\ m \geq 1, \tag{38b}$$

and

$$P_{\boldsymbol{x}} := P(\boldsymbol{x}) = \sum_{m=1}^{\infty} P(\boldsymbol{x}, m) = \sum_{m=1}^{\infty} \kappa_m \sum_{\boldsymbol{m} \in M_{\boldsymbol{x}}^{(m)}} \binom{m}{\boldsymbol{m}} \boldsymbol{p}^{\boldsymbol{m}}. \tag{38c}$$

The probability mass function in (38c) can be further manipulated using the approach described in [12]. Namely, the partial order $\preceq$

$$\boldsymbol{y} \preceq \boldsymbol{x} \ :\Leftrightarrow\ y_k \subseteq x_k \text{ for all } k, \text{ and } \boldsymbol{y} \prec \boldsymbol{x} \ :\Leftrightarrow\ \boldsymbol{y} \preceq \boldsymbol{x} \ \wedge\ \boldsymbol{y} \neq \boldsymbol{x} \tag{39}$$

is defined on the set $\mathscr{O}$. (Since the components of an observation $\boldsymbol{x}$ correspond to 0–1 vectors $y_k \subseteq x_k$ corresponds to $y_k \leq x_k$, where the inequality is understood componentwise.) Note, $\boldsymbol{y} \preceq \boldsymbol{x}$ is equivalent to $\boldsymbol{y} \in \mathscr{A}_{\boldsymbol{x}}$. The set $M_{\boldsymbol{x}}^{(m)}$ can be expressed as

$$M_{\boldsymbol{x}}^{(m)} = \{\boldsymbol{m} \mid m_{\boldsymbol{h}} = 0 \text{ if } \boldsymbol{h} \notin A_{\boldsymbol{x}},\ |\boldsymbol{m}| = m\} \setminus \bigcup_{\boldsymbol{y} \prec \boldsymbol{x}} \{\boldsymbol{m} \mid m_{\boldsymbol{h}} = 0 \text{ if } \boldsymbol{h} \notin A_{\boldsymbol{y}},\ |\boldsymbol{m}| = m\}. \tag{40}$$

Here, infections $\boldsymbol{m}$ in the first set $\{\boldsymbol{m} \mid m_{\boldsymbol{h}} = 0 \text{ if } \boldsymbol{h} \notin A_{\boldsymbol{x}},\ |\boldsymbol{m}| = m\}$ yield either observation $\boldsymbol{x}$ or a proper sub-observation $\boldsymbol{y} \prec \boldsymbol{x}$. The MOI vectors that correspond to such proper sub-observations $\boldsymbol{y}$ are removed by the second term.

Using the above, the inner-sum in (38c) can be rewritten using the inclusion-exclusion principle as

$$\sum_{\boldsymbol{m} \in M_{\boldsymbol{x}}^{(m)}} \binom{m}{\boldsymbol{m}} \boldsymbol{p}^{\boldsymbol{m}} = \sum_{\substack{\boldsymbol{m}:|\boldsymbol{m}|=m \\ m_{\boldsymbol{h}}=0 \text{ if } \boldsymbol{h} \notin A_{\boldsymbol{x}}}} \binom{m}{\boldsymbol{m}} \boldsymbol{p}^{\boldsymbol{m}} + \sum_{\boldsymbol{y} \prec \boldsymbol{x}} (-1)^{|\boldsymbol{x}|-|\boldsymbol{y}|} \sum_{\substack{\boldsymbol{m}:|\boldsymbol{m}|=m \\ m_{\boldsymbol{h}}=0 \text{ if } \boldsymbol{h} \notin A_{\boldsymbol{y}}}} \binom{m}{\boldsymbol{m}} \boldsymbol{p}^{\boldsymbol{m}}, \tag{41}$$

where $|\boldsymbol{x}| = \sum_{k=1}^{L} |x_k|$ and $|\boldsymbol{y}| = \sum_{k=1}^{L} |y_k|$ are respectively, the cardinals of $\boldsymbol{x}$, and $\boldsymbol{y}$. Note that $\boldsymbol{y} \prec \boldsymbol{x} \Rightarrow A_{\boldsymbol{y}} \subsetneq A_{\boldsymbol{x}}$ and $|\boldsymbol{x}| - |\boldsymbol{y}| > 0$. Hence,

$$\begin{aligned} \sum_{\boldsymbol{m} \in M_{\boldsymbol{x}}^{(m)}} \binom{m}{\boldsymbol{m}} \boldsymbol{p}^{\boldsymbol{m}} &= \sum_{\substack{\boldsymbol{m}:|\boldsymbol{m}|=m \\ m_{\boldsymbol{h}}=0 \text{ if } \boldsymbol{h} \notin A_{\boldsymbol{x}}}} \binom{m}{\boldsymbol{m}} \boldsymbol{p}^{\boldsymbol{m}} + \sum_{\boldsymbol{y} \prec \boldsymbol{x}} (-1)^{|\boldsymbol{x}|-|\boldsymbol{y}|} \sum_{\substack{\boldsymbol{m}:|\boldsymbol{m}|=m \\ m_{\boldsymbol{h}}=0 \text{ if } \boldsymbol{h} \notin A_{\boldsymbol{y}}}} \binom{m}{\boldsymbol{m}} \boldsymbol{p}^{\boldsymbol{m}} \\ &= \sum_{\boldsymbol{y} \preceq \boldsymbol{x}} (-1)^{|\boldsymbol{x}|-|\boldsymbol{y}|} \sum_{\substack{\boldsymbol{m}:|\boldsymbol{m}|=m \\ m_{\boldsymbol{h}}=0 \text{ if } \boldsymbol{h} \notin A_{\boldsymbol{y}}}} \binom{m}{\boldsymbol{m}} \boldsymbol{p}^{\boldsymbol{m}} \\ &= \sum_{\boldsymbol{y} \in \mathscr{A}_{\boldsymbol{x}}} (-1)^{|\boldsymbol{x}|-|\boldsymbol{y}|} \sum_{\substack{\boldsymbol{m}:|\boldsymbol{m}|=m \\ m_{\boldsymbol{h}}=0 \text{ if } \boldsymbol{h} \notin A_{\boldsymbol{y}}}} \binom{m}{\boldsymbol{m}} \boldsymbol{p}^{\boldsymbol{m}}. \end{aligned} \tag{42}$$

Therefore, the probability mass function in (38c) becomes

$$P_{\boldsymbol{x}} = \sum_{m=1}^{\infty} \kappa_m \sum_{\boldsymbol{y} \in \mathscr{A}_{\boldsymbol{x}}} (-1)^{|\boldsymbol{x}|-|\boldsymbol{y}|} \sum_{\substack{\boldsymbol{m}:|\boldsymbol{m}|=m \\ m_{\boldsymbol{h}}=0 \text{ if } \boldsymbol{h} \notin A_{\boldsymbol{y}}}} \binom{m}{\boldsymbol{m}} \boldsymbol{p}^{\boldsymbol{m}}. \tag{43}$$

By the multinomial theorem, we have

$$\sum_{\substack{\boldsymbol{m}:|\boldsymbol{m}|=m \\ m_{\boldsymbol{h}}=0 \text{ if } \boldsymbol{h} \notin A_{\boldsymbol{y}}}} \binom{m}{\boldsymbol{m}} \boldsymbol{p}^{\boldsymbol{m}} = \Big(\sum_{\boldsymbol{h} \in A_{\boldsymbol{y}}} p_{\boldsymbol{h}} \Big)^m, \tag{44}$$

which can be replaced in (43) to give

$$P_{\boldsymbol{x}} = \sum_{m=1}^{\infty} \kappa_m \sum_{\boldsymbol{y} \in \mathscr{A}_{\boldsymbol{x}}} (-1)^{|\boldsymbol{x}|-|\boldsymbol{y}|} \Big(\sum_{\boldsymbol{h} \in A_{\boldsymbol{y}}} p_{\boldsymbol{h}} \Big)^m \tag{45}$$

$$= \sum_{\boldsymbol{y} \in \mathscr{A}_{\boldsymbol{x}}} (-1)^{|\boldsymbol{x}|-|\boldsymbol{y}|} \sum_{m=1}^{\infty} \kappa_m \Big(\sum_{\boldsymbol{h} \in A_{\boldsymbol{y}}} p_{\boldsymbol{h}} \Big)^m. \tag{46}$$

The inner-sum in (46) is the probability-generating function (PGF) of the MOI distribution, evaluated at $\sum_{\boldsymbol{h} \in A_{\boldsymbol{y}}} p_{\boldsymbol{h}}$. We denote the probability-generating function by G such that

$$G(z) := \mathbb{E}[z^m] = \sum_{m=1}^{\infty} \kappa_m z^m. \tag{47}$$

Therefore, the probability of observing $\boldsymbol{x}$ is given by

$$P_{\boldsymbol{x}} = \sum_{\boldsymbol{y} \in \mathscr{A}_{\boldsymbol{x}}} (-1)^{|\boldsymbol{x}|-|\boldsymbol{y}|} G\Big(\sum_{\boldsymbol{h} \in A_{\boldsymbol{y}}} p_{\boldsymbol{h}} \Big). \tag{48a}$$

In the case that MOI follows a conditional Poisson distribution (as assumed here),

$$G(z) = \frac{e^{\lambda z} - 1}{e^{\lambda} - 1}. \tag{48b}$$

Under this assumption, the parameter space of the model is

$$\Theta := \mathbb{R}^+ \times \mathcal{S}_H = \left\{ (\lambda, \boldsymbol{p}) \,|\, \lambda > 0 \text{ and } \sum_{\boldsymbol{h}=1}^{H} p_{\boldsymbol{h}} = 1,\ 0 \leq p_{\boldsymbol{h}} \leq 1 \text{ for all } \boldsymbol{h} \right\}, \tag{49}$$

where $\mathcal{S}_H := \left\{ \boldsymbol{p} = (p_1, \ldots, p_H) \,\middle|\, \sum_{k=1}^{H} p_k = 1 \text{ and } 0 \leq p_k \leq 1, \text{ for all } k \right\}$ is an $(H-1)$-dimensional simplex, and λ is a positive real number.

Datasets and likelihood function. A typical molecular dataset $\mathscr{X}$ consists of N samples in the observational space $\mathscr{O}$ (35). I.e., N set-valued vectors with L components corresponding to alleles detected at each of the L markers. The j-th sample is denoted by using a superscript, i.e., $\boldsymbol{x}^{(j)} = \left(x_1^{(j)}, \ldots, x_L^{(j)}\right)$, where $x_k^{(j)}$ is the allelic information at locus k.

Let $n_{\boldsymbol{x}}$ be the number of times each observation $\boldsymbol{x}$ is made in the dataset. Hence,

$$\sum_{\boldsymbol{x}\in\mathscr{O}} n_{\boldsymbol{x}} = N. \tag{50}$$

Consequently, the dataset $\mathscr{X}$ can be represented as the integer-valued vector

$$(n_{\boldsymbol{x}})_{\boldsymbol{x}\in\mathscr{O}}. \tag{51}$$

This representation will be repeatedly used in the following derivations. Note that for increased number of loci L and number of alleles n_k, not all possible observations are detected, i.e., $|\mathscr{O}| = \prod_{k=1}^{L}(2^{n_k}-1)$ exceeds the sample size N. Therefore, $n_{\boldsymbol{x}} = 0$ for most observations $\boldsymbol{x} \in \mathscr{O}$.

Given $\mathscr{X}$, the model parameters $\theta := (\lambda, \boldsymbol{p}) \in \Theta$ can be collectively estimated by maximum likelihood. Because the N samples are assumed to be independent and identically distributed, the likelihood function $\mathcal{L}_{\mathscr{X}}(\theta)$ is

$$\mathcal{L}_{\mathscr{X}}(\theta) = \prod_{\boldsymbol{x}\in\mathscr{X}} P_{\boldsymbol{x}}^{n_{\boldsymbol{x}}} = \prod_{\boldsymbol{x}\in\mathscr{X}} \left(\sum_{\boldsymbol{y}\in\mathscr{A}_{\boldsymbol{x}}} (-1)^{|\boldsymbol{x}|-|\boldsymbol{y}|} G\Big(\sum_{\boldsymbol{h}\in A_{\boldsymbol{y}}} p_{\boldsymbol{h}} \Big) \right)^{n_{\boldsymbol{x}}}, \tag{52}$$

and the log-likelihood function becomes

$$L_{\mathscr{X}}(\theta) = \log\left(\mathcal{L}_{\mathscr{X}}(\theta)\right) = \sum_{\boldsymbol{x}\in\mathscr{O}} n_{\boldsymbol{x}} \log\left(\sum_{\boldsymbol{y}\in\mathscr{A}_{\boldsymbol{x}}} (-1)^{|\boldsymbol{x}|-|\boldsymbol{y}|} G\Big(\sum_{\boldsymbol{h}\in A_{\boldsymbol{y}}} p_{\boldsymbol{h}} \Big) \right). \tag{53}$$

The maximum likelihood estimate (MLE) $\hat{\theta} = (\hat{\lambda}, \hat{\boldsymbol{p}})$ of the model parameters is obtained by maximizing the log-likelihood function. A closed solution for the MLE is typically not possible, particularly for this complex likelihood function. In fact, a closed solution is not even possible in the simplest case of a single molecular marker (cf. [9,24]).

Here, the expectation-maximization (EM)-algorithm is used to derive the MLE. It is an efficient, stable, and fast-converging algorithm to maximize the likelihood function [26]. The algorithm is derived in section Maximum likelihood estimate.

Because the MLE needs to be derived numerically, it is impossible to obtain analytic results about the quality of the estimator. Particularly, bias and variance cannot be determined explicitly. However, these can be assessed numerically as it was done for the special cases of (i) a single molecular marker ($L = 1$) [27], (ii) L biallelic markers [12], and (iii) two multiallelic markers ($L = 2$) [10]. The results suggested that the estimator is asymptotically unbiased, and its asymptotic variance is given by the Cramér-Rao lower bound (CRLB). This is confirmed by asymptotic results in the special case of a single marker ($L = 1$), in which case, the model falls into the exponential-family class. Given these facts, the present estimator can be considered asymptotically unbiased and efficient. However, for the asymptotic variance, the CRLB has to be derived for the present general case.

Numerical investigation of finite sample properties

Finite sample properties of the estimator are investigated by numerical simulations.

Dataset generation. datasets are created following Fig 13. Specifically, for a choice of the parameters λ and $\boldsymbol{p}$, a dataset $\mathscr{X}$ of size N is created by first sampling N MOI values ($m^{(i)}$, $i = 1, \ldots, N$) according to a conditional Poisson distribution. Next, N MOI vectors $\boldsymbol{m}^{(i)}$ ($i = 1, \ldots, N$) are drawn from multinomial distributions with parameters $m^{(i)}$ and $\boldsymbol{p}$.

From each MOI vector $\boldsymbol{m}^{(i)}$, the observation $\boldsymbol{x}^{(i)}$ is created by retaining only the information concerning the absence and presence of alleles at each marker.

For each set of parameters $(\lambda, \boldsymbol{p}, N)$, this procedure is repeated $K = 100\,000$ times to generate K datasets $\mathscr{X}_1, \ldots, \mathscr{X}_K$ of sample size N.

Simulated variance. Given a set of parameters λ, $\boldsymbol{p}$, and N, for each of the K datasets $\mathscr{X}_k$, the MLEs $\hat{\boldsymbol{\theta}}_k = (\hat{\lambda}_k, \hat{\boldsymbol{p}}_k)$ are calculated and the estimate of the mean MOI $\hat{\psi}_k$ is derived. Next, the variance of the estimator for a parameter of interest, $\theta = \psi, p_1, \ldots, p_H$, is calculated as the empirical variance, i.e.,

$$S^2_{\hat{\theta}} = \frac{1}{K-1} \sum_{k=1}^{K} \left(\hat{\theta}^{(k)} - \bar{\theta}\right)^2, \tag{54a}$$

where

$$\bar{\theta} = \frac{1}{K} \sum_{k=1}^{K} \hat{\theta}^{(k)}. \tag{54b}$$

To allow comparison between different true values of the MOI parameter λ, the variation of the mean MOI parameter is reported as the coefficient of variation (CV), i.e.,

$$\frac{\sqrt{S^2_{\hat{\psi}}}}{\psi}, \tag{55a}$$

whereas the variation of the haplotype frequencies is reported in terms of the empirical standard deviation $\sqrt{S^2_{\hat{p}_i}}$. These quantities are compared to the appropriate transforms of the CRLB, i.e., to $\frac{\sqrt{\mathcal{I}^{-1}_{\psi\psi}}}{\psi}$ and $\sqrt{\mathcal{I}^{-1}_{p_i,p_i}}$.

Parameter choices for numerical investigations. The following parameters are chosen for numerical investigations.

Genetic architecture. Two markers ($n = 2$) are assumed. Moreover, for the simulations, $n_1 = n_2 = 2$ and $n_1 = 4, n_2 = 7$ loci are chosen, leading respectively to $H = 4$ and $H = 28$ possible haplotypes.

MOI parameter

The choice of MOI parameters is made to represent different levels of transmission intensities. With malaria in mind, MOI parameters are chosen as $\lambda = 0.1, 0.25, 0.5, 1, 1.5, 2, 2.5$, corresponding to mean MOI $\psi = 1.05, 1.13, 1.27, 1.58, 1.93, 2.31, 2.72$. Note that $\psi < 1.27$ corresponds to low transmission, $1.27 \leq \psi < 1.93$ to intermediate transmission and $\psi \geq 1.93$ to high transmission [23].

Haplotype frequency distribution

The distribution of haplotype frequency $\boldsymbol{p}$ is chosen to mimic two extreme cases:

(i) a uniform (balanced) distribution such that

$$p_1 = \ldots = p_H = \frac{1}{H}, \tag{56a}$$

and (ii) an unbalanced distribution for which one haplotype, e.g., p_1 is predominant with a frequency of 70%, while the remaining haplotypes have the same frequency, i.e.,

$$p_1 = 0.7, p_2 = \ldots = p_H = \frac{0.3}{H-1}. \tag{56b}$$

In particular, for the chosen genetic architectures, this gives $p_1 = 0.7, p_2 = p_3 = p_4 = 0.1$ in the case $n_1 = n_2 = 2$ and $p_1 = 0.7, p_2 = \ldots = p_{28} = \frac{0.3}{27}$ in the case $n_1 = 4$ and $n_2 = 7$.

Sample size

Sample sizes $N = 50, 100, 150, 200, 500$ are chosen for the simulations. Note, in practice, sample size correlates with transmission intensity. In areas of low disease transmission, it is more challenging to achieve a large sample size, whereas this is much easier in areas of high transmission.

Population genetic measures

In a population of size N the expected heterozygosity $H_e^{(l)}$ at locus l with n_l segregating *all*eles $A_i^{(l)}$ $(i = 1, \ldots, n_l)$, each with frequency $f_i^{(l)}$, is given by

$$H_e^{(l)} := \frac{N}{N-1}\left(1 - \sum_{i=1}^{n_l} {f_i^{(l)}}^2\right). \tag{57}$$

Note that $f_i^{(l)}$ is obtained by marginalization of the frequencies of haplotypes containing allele $A_i^{(l)}$ at locus l, i.e.,

$$f_i^{(l)} := \sum_{\substack{\boldsymbol{h} \in \mathscr{H}: \\ h_l = A_i^{(l)}}} p_{\boldsymbol{h}}, \tag{58}$$

the sum of the frequencies of all haplotypes, which carry allele $A_i^{(l)}$ at locus l.

Consider a subset of loci $S \subseteq \{1, \ldots, L\}$ and a specific allelic configuration $\tilde{\boldsymbol{h}} = \left(A_{i_u}^{(u)}\right)_{u \in S}$ $(i_u = 1, \ldots, n_u)$, i.e., a "sub-haplotype" determined by its allelic configuration at the markers in set S. One can calculate the heterozygosity at locus l, conditioned on the background of the sub-haplotype $\tilde{\boldsymbol{h}}$. The marginal frequency of allele $A_i^{(l)}$ $(i = 1, \ldots, n_l)$ at *l*ocus l is given by

$$f_{i|\tilde{\boldsymbol{h}}}^{(l)} := \sum_{\substack{\boldsymbol{h} \in \mathscr{H}: \\ h_l = A_i^{(l)} \\ h_u = \tilde{h}_u \text{ for } u \in S}} p_{\boldsymbol{h}}. \tag{59}$$

The heterozygosity conditioned on $\tilde{\boldsymbol{h}}$ is then given by

$$H_{e|\tilde{\boldsymbol{h}}}^{(l)} := \frac{N}{N-1}\left(1 - \sum_{i=1}^{n_l} {f_{i|\tilde{\boldsymbol{h}}}^{(l)}}^2\right). \tag{60}$$

Pairwise LD can be obtained from several metrics (cf. [54,55]). Here, with no particular preference, two of these commonly used metrics, i.e., r^2 (also known as D^*) and D' are reported. The LD metric $r^{2(kl)}$ between two loci k and l is obtained based on the Hardy-Weinberg heterozygosities $H^{(k)} = 1 - \sum_{i=1}^{n_k} {f_i^{(k)}}^2$ and $H^{(l)} = 1 - \sum_{j=1}^{n_l} {f_j^{(l)}}^2$ at loci k and l, respectively, as

$$r^{2(kl)} := \frac{1}{H^{(k)} H^{(l)}} \sum_{i=1}^{n_k} \sum_{j=1}^{n_l} {D_{ij}^{(kl)}}^2. \tag{61}$$

where $D_{ij}^{(kl)} := f_{ij}^{(kl)} - f_i^{(k)} f_j^{(l)}$ measures the discrepancy between the observed frequency $f_{ij}^{(kl)}$ of sub-haplotype $A_i^{(k)} A_j^{(l)}$ and its expected frequency $f_i^{(k)} f_j^{(l)}$ under random association.

The measure $D'^{(kl)}$ is defined in [54,55] as

$$D'^{(kl)} := \sum_{i=1}^{n_k} \sum_{j=1}^{n_l} f_i^{(k)} f_j^{(l)} \frac{|D_{ij}^{(kl)}|}{D_{\max}^{(kl)}}, \tag{62a}$$

where

$$D_{\max}^{(kl)} := \begin{cases} \min\left(f_i^{(k)} f_j^{(l)}, \left(1 - f_i^{(k)}\right)\left(1 - f_j^{(l)}\right) \right) & \text{if } D_{ij}^{(kl)} < 0, \\ \min\left(f_i^{(k)}\left(1 - f_j^{(l)}\right), f_j^{(l)}\left(1 - f_i^{(k)}\right) \right) & \text{if } D_{ij}^{(kl)} \geq 0. \end{cases} \tag{62b}$$

Restricted to the background of a given sub-haplotype $\tilde{\boldsymbol{h}} = \left(A_{i_u}^{(u)}\right)_{u \in S}$, the LD measures are based on conditional allele frequencies defined as in (59). The conditional LD measures $r^2_{|\tilde{\boldsymbol{h}}}{}^{(kl)}$ and $D'_{|\tilde{\boldsymbol{h}}}{}^{(kl)}$ are defined by

$$r^2_{|\tilde{\boldsymbol{h}}}{}^{(kl)} := \frac{1}{H_{|\tilde{\boldsymbol{h}}}^{(k)} H_{|\tilde{\boldsymbol{h}}}^{(l)}} \sum_{i=1}^{n_k} \sum_{j=1}^{n_l} D^2_{ij|\tilde{\boldsymbol{h}}}, \tag{63a}$$

with

$$H_{|\tilde{\boldsymbol{h}}}^{(k)} = 1 - \sum_{i=1}^{n_k} f_{i|\tilde{\boldsymbol{h}}}^{(k)\,2} \qquad \text{and} \qquad H_{|\tilde{\boldsymbol{h}}}^{(l)} = 1 - \sum_{j=1}^{n_l} f_{j|\tilde{\boldsymbol{h}}}^{(l)\,2}, \tag{63b}$$

and

$$D'_{|\tilde{\boldsymbol{h}}}{}^{(kl)} := \sum_{i=1}^{n_k} \sum_{j=1}^{n_l} f_{i|\tilde{\boldsymbol{h}}}^{(k)} f_{j|\tilde{\boldsymbol{h}}}^{(l)} \frac{|D_{ij|\tilde{\boldsymbol{h}}}^{(kl)}|}{D_{\max|\tilde{\boldsymbol{h}}}^{(kl)}}, \tag{64a}$$

with

$$D_{ij|\tilde{\boldsymbol{h}}}^{(kl)} := f_{ij|\tilde{\boldsymbol{h}}}^{(kl)} - f_{i|\tilde{\boldsymbol{h}}}^{(k)} f_{j|\tilde{\boldsymbol{h}}}^{(l)}, \tag{64b}$$

$$f_{ij|\tilde{\boldsymbol{h}}}^{(kl)} := \sum_{\substack{h \in \mathscr{H}: \\ h_k = A_i^{(k)}, h_l = A_j^{(l)}, \\ h_u = \tilde{h}_u \text{ for } u \in S}} p_{\boldsymbol{h}}, \tag{64c}$$

and

$$D_{\max|\tilde{\boldsymbol{h}}}^{(kl)} := \begin{cases} \min\left(f_{i|\tilde{\boldsymbol{h}}}^{(k)} f_{j|\tilde{\boldsymbol{h}}}^{(l)}, \left(1 - f_{i|\tilde{\boldsymbol{h}}}^{(k)}\right)\left(1 - f_{j|\tilde{\boldsymbol{h}}}^{(l)}\right) \right) & \text{if } D_{ij|\tilde{\boldsymbol{h}}} < 0, \\ \min\left(f_{i|\tilde{\boldsymbol{h}}}^{(k)}\left(1 - f_{j|\tilde{\boldsymbol{h}}}^{(l)}\right), f_{j|\tilde{\boldsymbol{h}}}^{(l)}\left(1 - f_{i|\tilde{\boldsymbol{h}}}^{(k)}\right) \right) & \text{if } D_{ij|\tilde{\boldsymbol{h}}} \geq 0. \end{cases} \tag{64d}$$

Benchmarking computational time

Data from the MalariaGen project [28] was utilized to investigate the performance of the proposed method. In particular, data from the current release of the *P. falciparum* data (Pf8) was used. This dataset contains whole genome information on 33 325 *P. falciparum* specimens and is substantially more comprehensive than the prior version Pf7 [56]. Two numerical experiments were performed.

The first one used the prepared data on drug resistance. We split that data by country, year, and study and retained all resulting datasets with more than 20 samples. We included genetic information from *Pfcrt* codons 72–76 (used as microhaplotype), 93, 97, 218, 220, 271, 326, 333, 353, 356, and 371, *Pfdhfr* codons 16, 51, 59, 108, 164, and 306, *Pfdhps* codons 436, 437, 540, 581, and 613, *Pfmdr1* codons 81, 184, 1034, 1042, 1126, 1246, *Pfexo* codon 415, *Pfarps10* codons 127 and 128, *Pffd* codon 193, *Pfmdr2* codon 484, non-synonymous changes in *Pfkelch13* codons 349–726 (used as microhaplotype), as well as duplication status in *Pfmdr1* and *Pfpm2*. For the purpose here, duplications were treated as alleles, i.e., absence/presence of duplicates. This resulted in a total of $L = 36$ molecular markers. Several of these markers are multiallelic.

The data needed some preliminary steps of data cleaning. Several data entries were marked as low quality. These were included here to increase sample size and the number of polymorphic markers, since the purpose of the numerical experiment was to test the proposed method.

For the second experiment, genetic information for the samples collected in Cameroon in 2013 was used. All SNPs within 4.3kb up- and downstream from *Pfdhfr* on chromosome 4 were extracted. Only those SNPs that were polymorphic in the data were retained. All resulting SNPs were biallelic. The data was then used to create several derived datasets. The first derived dataset included all SNPs within -0.1kb to 0.1kb *Pfdhfr*. This window was then increased in steps of 0.1kb up- and downstream until the final window size of -4.3kb to 4.3kb. In each dataset only samples with complete information were retained. Hence, sample size of the datasets decreased with increasing window size.

In both experiments, the MLE was calculated for each dataset, and the computational time was recorded. All computations were performed using R.4.4.1 on a MacBook Pro (OS: Sequoia 15.5, CPU: Apple M1 8-core CPU with 4 3.2 GHz performance cores and 4 2.1 GHz efficiency cores, RAM: 8 GB, storage: 500 GB).

Supporting information

S1 Appendix. Mathematical Appendix.
(PDF)

S2 Appendix. R-script manual.
(PDF)

S1 Data. Zip file containing an implementation of the model as an R script, a template R script with commands to use for data analysis using the method, and an example molecular dataset.
(ZIP)

Acknowledgments

The authors gratefully acknowledge the African Institute for Mathematical Sciences (AIMS) Cameroon for supporting the meeting between the authors and this research. The many fruitful discussions with friends and colleagues that helped to improve the manuscript were highly appreciated. The constructive comments of two anonymous reviewers are gratefully acknowledged. The content is solely the responsibility of the authors and does not represent the official views of anybody.

Author contributions

Conceptualization: Henri Christian Junior Tsoungui Obama, Kristan Alexander Schneider.

Data curation: Henri Christian Junior Tsoungui Obama.

Formal analysis: Henri Christian Junior Tsoungui Obama, Kristan Alexander Schneider.

Funding acquisition: Kristan Alexander Schneider.

Investigation: Henri Christian Junior Tsoungui Obama, Kristan Alexander Schneider.

Methodology: Henri Christian Junior Tsoungui Obama, Kristan Alexander Schneider.

Project administration: Kristan Alexander Schneider.

Resources: Kristan Alexander Schneider.

Software: Henri Christian Junior Tsoungui Obama, Kristan Alexander Schneider.

Supervision: Kristan Alexander Schneider.

Validation: Henri Christian Junior Tsoungui Obama, Kristan Alexander Schneider.

Visualization: Henri Christian Junior Tsoungui Obama, Kristan Alexander Schneider.

Writing – original draft: Henri Christian Junior Tsoungui Obama, Kristan Alexander Schneider.

Writing – review & editing: Henri Christian Junior Tsoungui Obama, Kristan Alexander Schneider.

References

1. Koudokpon H, Lègba B, Sintondji K, Kissira I, Kounou A, Guindo I, et al. Empowering public health: building advanced molecular surveillance in resource-limited settings through collaboration and capacity-building. Front Health Serv. 2024;4:1289394. https://doi.org/10.3389/frhs.2024.1289394 PMID: 38957804
2. Schneider KA, Tsoungui Obama HCJ, Kamanga G, Kayanula L, Adil Mahmoud Yousif N. The many definitions of multiplicity of infection. Front Epidemiol. 2022;2:961593. https://doi.org/10.3389/fepid.2022.961593 PMID: 38455332
3. Tusting LS, Bousema T, Smith DL, Drakeley C. Measuring changes in Plasmodium falciparum transmission: precision, accuracy and costs of metrics. Adv Parasitol. 2014;84:151–208. https://doi.org/10.1016/B978-0-12-800099-1.00003-X PMID: 24480314
4. Sondo P, Derra K, Rouamba T, Nakanabo Diallo S, Taconet P, Kazienga A, et al. Determinants of Plasmodium falciparum multiplicity of infection and genetic diversity in Burkina Faso. Parasit Vectors. 2020;13(1):427. https://doi.org/10.1186/s13071-020-04302-z PMID: 32819420
5. Sinha A, Kar S, Deora N, Dash M, Tiwari A, Kori L, et al. India-EMBO Lecture Course: Understanding Malaria from Molecular Epidemiology, Population Genetics, and Evolutionary Perspectives. Trends Parasitol. 2023;39(5):307–13. https://doi.org/10.1016/j.pt.2023.02.010
6. McCollum AM, Schneider KA, Griffing SM, Zhou Z, Kariuki S, Ter-Kuile F, et al. Differences in selective pressure on dhps and dhfr drug resistant mutations in western Kenya. Malar J. 2012;11:77. https://doi.org/10.1186/1475-2875-11-77 PMID: 22439637
7. Nash D, Nair S, Mayxay M, Newton PN, Guthmann J-P, Nosten F, et al. Selection strength and hitchhiking around two anti-malarial resistance genes. Proc Biol Sci. 2005;272(1568):1153–61. https://doi.org/10.1098/rspb.2004.3026 PMID: 16024377
8. Hill WG, Babiker HA. Estimation of numbers of malaria clones in blood samples. Proc Biol Sci. 1995;262(1365):249–57. https://doi.org/10.1098/rspb.1995.0203 PMID: 8587883
9. Schneider KA, Escalante AA. A likelihood approach to estimate the number of co-infections. PLoS One. 2014;9(7):e97899. https://doi.org/10.1371/journal.pone.0097899 PMID: 24988302
10. Tsoungui Obama HCJ, Schneider KA. Estimating multiplicity of infection, haplotype frequencies, and linkage disequilibria from multi-allelic markers for molecular disease surveillance. PLoS One. 2025;20(5):e0321723. https://doi.org/10.1371/journal.pone.0321723 PMID: 40424286
11. 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. Malar J. 2015;14:4. https://doi.org/10.1186/1475-2875-14-4 PMID: 25599890
12. Tsoungui Obama HCJ, Schneider KA. A maximum-likelihood method to estimate haplotype frequencies and prevalence alongside multiplicity of infection from SNP data. Front Epidemiol. 2022;2:943625. https://doi.org/10.3389/fepid.2022.943625 PMID: 38455338
13. Li X, Foulkes AS, Yucel RM, Rich SM. An expectation maximization approach to estimate malaria haplotype frequencies in multiply infected children. Stat Appl Genet Mol Biol. 2007;6:Article33. https://doi.org/10.2202/1544-6115.1321 PMID: 18052916
14. 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 Comput Biol. 2017;13(1):e1005348. https://doi.org/10.1371/journal.pcbi.1005348 PMID: 28125584

15. Hastings IM, Smith TA. MalHaploFreq: a computer programme for estimating malaria haplotype frequencies from blood samples. Malar J. 2008;7:130. https://doi.org/10.1186/1475-2875-7-130 PMID: 18627599
16. Wigger L, Vogt JE, Roth V. Malaria haplotype frequency estimation. Stat Med. 2013;32(21):3737–51. https://doi.org/10.1002/sim.5792 PMID: 23609602
17. Kuk AYC, Li X, Xu J. An EM algorithm based on an internal list for estimating haplotype distributions of rare variants from pooled genotype data. BMC Genet. 2013;14:82. https://doi.org/10.1186/1471-2156-14-82 PMID: 24034507
18. Ken-Dror G, Hastings IM. Markov chain Monte Carlo and expectation maximization approaches for estimation of haplotype frequencies for multiply infected human blood samples. Malar J. 2016;15(1):430. https://doi.org/10.1186/s12936-016-1473-5 PMID: 27557806
19. Zhu SJ, Almagro-Garcia J, McVean G. Deconvolution of multiple infections in Plasmodium falciparum from high throughput sequencing data. Bioinformatics. 2018;34(1):9–15. https://doi.org/10.1093/bioinformatics/btx530 PMID: 28961721
20. Paschalidis A, Watson OJ, Aydemir O, Verity R, Bailey JA. coiaf: Directly estimating complexity of infection with allele frequencies. PLoS Comput Biol. 2023;19(6):e1010247. https://doi.org/10.1371/journal.pcbi.1010247 PMID: 37294835
21. Kayanula L, Schneider KA. A non-parametric approach to estimate multiplicity of infection (MOI) and pathogen haplotype frequencies. Front Malar. 2024;2.
22. Ross A, Koepfli C, Li X, Schoepflin S, Siba P, Mueller I, et al. Estimating the numbers of malaria infections in blood samples using high-resolution genotyping data. PLoS One. 2012;7(8):e42496. https://doi.org/10.1371/journal.pone.0042496 PMID: 22952595
23. Hashemi M, Schneider KA. Bias-corrected maximum-likelihood estimation of multiplicity of infection and lineage frequencies. PLoS One. 2021;16(12):e0261889. https://doi.org/10.1371/journal.pone.0261889 PMID: 34965279
24. Hashemi M, Schneider KA. Estimating multiplicity of infection, allele frequencies, and prevalences accounting for incomplete data. PLoS One. 2024;19(3):e0287161. https://doi.org/10.1371/journal.pone.0287161 PMID: 38512826
25. Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data Via the EM Algorithm. J R Stat Soc Series B Stat Methodol. 1977;39(1):1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x
26. Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol. 1995;12(5):921–7. https://doi.org/10.1093/oxfordjournals.molbev.a040269 PMID: 7476138
27. Schneider KA. Large and finite sample properties of a maximum-likelihood estimator for multiplicity of infection. PLoS One. 2018;13(4):e0194148. https://doi.org/10.1371/journal.pone.0194148 PMID: 29630605
28. MalariaGen MM, Abdel Hamid M, Abdelraheem M, Acheampong D, Adam I, Aide P, et al. Pf8: an open dataset of Plasmodium falciparum genome variation in 33,325 worldwide samples [version 1; peer review: 1 approved]. Wellcome Open Res. 2025;10(325).
29. Tsoungui Obama HCJ, Schneider KA. Maths-against-Malaria/generalModel: v1.0.0. Zenodo; 2024.
30. 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. https://doi.org/10.1186/1475-2875-7-223 PMID: 18959790
31. Pacheco MA, Forero-Peña DA, Schneider KA, Chavero M, Gamardo A, Figuera L, et al. Malaria in Venezuela: changes in the complexity of infection reflects the increment in transmission intensity. Malar J. 2020;19(1):176. https://doi.org/10.1186/s12936-020-03247-z PMID: 32380999
32. de Cesare M, Mwenda M, Jeffreys AE, Chirwa J, Drakeley C, Schneider K, et al. Flexible and cost-effective genomic surveillance of P. falciparum malaria with targeted nanopore sequencing. Nat Commun. 2024;15(1):1413. https://doi.org/10.1038/s41467-024-45688-z PMID: 38360754
33. Holzschuh A, Lerch A, Nsanzabana C. Rapid multiplexed nanopore amplicon sequencing to distinguish Plasmodium falciparum recrudescence from new infection in antimalarial drug trials. Sci Rep. 2025;15(1):36941. https://doi.org/10.1038/s41598-025-20925-7 PMID: 41125674
34. Pathak PK. Introduction to Rao (1945) information and the accuracy attainable in the estimation of statistical parameters. In: Kotz S, Johnson NL, editors. Breakthroughs in Statistics: Foundations and Basic Theory. New York (NY): Springer; 1992. p. 227–34.
35. Nielsen F. Cramér-Rao lower bound and information geometry. In: Bhatia R, Rajan CS, Singh AI, editors. Connected at infinity II: A selection of mathematics by Indians. Gurgaon: Hindustan Book Agency; 2013. p. 18–37.
36. Davison AC. Statistical Models. Cambridge: Cambridge University Press; 2003.
37. McCollum AM, Basco LK, Tahar R, Udhayakumar V, Escalante AA. Hitchhiking and selective sweeps of Plasmodium falciparum sulfadoxine and pyrimethamine resistance alleles in a population from central Africa. Antimicrob Agents Chemother. 2008;52(11):4089–97. https://doi.org/10.1128/AAC.00623-08 PMID: 18765692
38. Nair S, Nash D, Sudimack D, Jaidee A, Barends M, Uhlemann A-C, et al. Recurrent gene amplification and soft selective sweeps during evolution of multidrug resistance in malaria parasites. Mol Biol Evol. 2007;24(2):562–73. https://doi.org/10.1093/molbev/msl185 PMID: 17124182
39. Contreras CE, Cortese JF, Caraballo A, Plowe CV. Genetics of drug-resistant Plasmodium falciparum malaria in the Venezuelan state of Bolivar. Am J Trop Med Hygiene. 2002;67(4):400–5.
40. Schneider KA, Kim Y. An analytical model for genetic hitchhiking in the evolution of antimalarial drug resistance. Theor Popul Biol. 2010;78(2):93–108. https://doi.org/10.1016/j.tpb.2010.06.005 PMID: 20600206
41. Schneider KA, Salas CJ. Evolutionary genetics of malaria. Front Genet. 2022;13.
42. Efron B, Tibshirani RJ. An introduction to the bootstrap. New York: Chapman and Hall/CRC; 1994.

43. Nkhoma SC, Nair S, Cheeseman IH, Rohr-Allegrini C, Singlam S, Nosten F, et al. Close kinship within multiple-genotype malaria parasite infections. Proc Biol Sci. 2012;279(1738):2589–98. https://doi.org/10.1098/rspb.2012.0113 PMID: 22398165

44. Wong W, Wenger EA, Hartl DL, Wirth DF. Modeling the genetic relatedness of Plasmodium falciparum parasites following meiotic recombination and cotransmission. PLoS Comput Biol. 2018;14(1):e1005923. https://doi.org/10.1371/journal.pcbi.1005923 PMID: 29315306

45. Nkhoma SC, Trevino SG, Gorena KM, Nair S, Khoswe S, Jett C, et al. Co-transmission of Related Malaria Parasite Lineages Shapes Within-Host Parasite Diversity. Cell Host Microbe. 2020;27(1):93–103.e4. https://doi.org/10.1016/j.chom.2019.12.001 PMID: 31901523

46. Neafsey DE, Taylor AR, MacInnis BL. Advances and opportunities in malaria population genomics. Nat Rev Genet. 2021;22(8):502–17. https://doi.org/10.1038/s41576-021-00349-5 PMID: 33833443

47. Dia A, Cheeseman IH. Single-Cell Genome Sequencing of Protozoan Parasites. Trends in Parasitology. 2021;37(9):803–14.

48. Zhu SJ, Hendry JA, Almagro-Garcia J, Pearson RD, Amato R, Miles A, et al. The Origins and Relatedness Structure of Mixed Infections Vary with Local Prevalence of P. Falciparum Malaria. eLife. 2019;8:e40845.

49. Komerup P. Digit-set conversions: generalizations and applications. IEEE Trans Comput. 1994;43(5).

50. Taylor AR, Flegg JA, Nsobya SL, Yeka A, Kamya MR, Rosenthal PJ, et al. Estimation of malaria haplotype and genotype frequencies: a statistical approach to overcome the challenge associated with multiclonal infections. Malar J. 2014;13:102. https://doi.org/10.1186/1475-2875-13-102 PMID: 24636676

51. 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(9):1292–4. https://doi.org/10.1093/bioinformatics/btu005 PMID: 24443379

52. Irvine MA, Kazura JW, Hollingsworth TD, Reimer LJ. Understanding heterogeneities in mosquito-bite exposure and infection distributions for the elimination of lymphatic filariasis. Proc Biol Sci. 2018;285(1871):20172253. https://doi.org/10.1098/rspb.2017.2253 PMID: 29386362

53. Adamidis K. Theory & Methods: An EM algorithm for estimating negative binomial parameters. Aus NZ J Stat. 1999;41(2):213–21. https://doi.org/10.1111/1467-842x.00075

54. Hedrick PW. Gametic disequilibrium measures: proceed with caution. Genetics. 1987;117(2):331–41. https://doi.org/10.1093/genetics/117.2.331 PMID: 3666445

55. Zhao H, Nettleton D, Dekkers JCM. Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between single nucleotide polymorphisms. Genet Res. 2007;89(1):1–6. https://doi.org/10.1017/S0016672307008634 PMID: 17517154

56. MalariaGEN MM, Abdel Hamid MM, Abdelraheem MH, Acheampong DO, Ahouidi A, Ali M, et al. Pf7: an open dataset of Plasmodium falciparum genome variation in 20,000 worldwide samples. Wellcome Open Res. 2023;8:22. https://doi.org/10.12688/wellcomeopenres.18681.1 PMID: 36864926