frontiers | Frontiers in Epidemiology

TYPE Original Research
PUBLISHED 23 September 2022
DOI 10.3389/fepid.2022.943625

Check for updates

OPEN ACCESS

EDITED BY
Jonathan J. Juliano,
University of North Carolina at Chapel Hill, United States

REVIEWED BY
Robert Verity,
Imperial College London,
United Kingdom
Aimee Taylor,
Institut Pasteur, France

*CORRESPONDENCE
Henri Christian Junior Tsoungui Obama
tsoungui@hs-mittweida.de

SPECIALTY SECTION
This article was submitted to Infectious Disease Epidemiology, a section of the journal Frontiers in Epidemiology

RECEIVED 13 May 2022
ACCEPTED 26 August 2022
PUBLISHED 23 September 2022

CITATION
Tsoungui Obama HCJ and Schneider KA (2022) A maximum-likelihood method to estimate haplotype frequencies and prevalence alongside multiplicity of infection from SNP data.
Front. Epidemiol. 2:943625.
doi: 10.3389/fepid.2022.943625

A maximum-likelihood method to estimate haplotype frequencies and prevalence alongside multiplicity of infection from SNP data

Henri Christian Junior Tsoungui Obama* and Kristan Alexander Schneider

Department of Applied Computer- and Biosciences, University of Applied Sciences Mittweida, Mittweida, Germany

The introduction of genomic methods facilitated standardized molecular disease surveillance. For instance, SNP barcodes in *Plasmodium vivax* and *Plasmodium falciparum* malaria allows the characterization of haplotypes, their frequencies and prevalence to reveal temporal and spatial transmission patterns. A confounding factor is the presence of multiple genetically distinct pathogen variants within the same infection, known as multiplicity of infection (MOI). Disregarding ambiguous information, as usually done in *ad-hoc* approaches, leads to less confident and biased estimates. We introduce a statistical framework to obtain maximum-likelihood estimates (MLE) of haplotype frequencies and prevalence alongside MOI from malaria SNP data, i.e., multiple biallelic marker loci. The number of model parameters increases geometrically with the number of genetic markers considered and no closed-form solution exists for the MLE. Therefore, the MLE needs to be derived numerically. We use the Expectation-Maximization (EM) algorithm to derive the maximum-likelihood estimates, an efficient and easy-to-implement algorithm that yields a numerically stable solution. We also derive expressions for haplotype prevalence based on either all or just the unambiguous genetic information and compare both approaches. The latter corresponds to a biased *ad-hoc* estimate of prevalence. We assess the performance of our estimator by systematic numerical simulations assuming realistic sample sizes and various scenarios of transmission intensity. For reasonable sample sizes, and number of loci, the method has little bias. As an example, we apply the method to a dataset from Cameroon on sulfadoxine-pyrimethamine resistance in *P. falciparum* malaria. The method is not confined to malaria and can be applied to any infectious disease with similar transmission behavior. An easy-to-use implementation of the method as an R-script is provided.

KEYWORDS

complexity of infection (COI), multiplicity of infection (MOI), EM-algorithm, drug resistance, malaria, sulfadoxine-pyrimethamine (SP), resistance markers, haplotype phasing

1. Introduction

With ever-decreasing costs, genomic/molecular technologies are commonly supporting traditional means of disease surveillance. Rather than clinical data, demographic, or behavioral risk factors, molecular methods provide information on a fine-grained scale that allows to reconstruct sources and routes of disease transmission by reverse engineering or to identify and monitor specific pathogen variants, for instance those associated with drug resistance [cf. (1, 2)]. Improvements in molecular technologies and bioinformatics facilitate the collection of genetic/molecular data on temporal and spatial scales [cf. (3, 4)].

On the epidemiological scale, characterizing pathogen variants on a genomic level, allows, e.g., to identify the emergence of variants resistant to vaccines or therapeutics. Further, by monitoring their prevalence in time and space, paths of transmission can be reconstructed. This might point to weak points in disease control and prevention [cf. (5)]. For the two most relevant species of human malaria, *P. falciparum* and *P. vivax*, even SNP barcodes were developed to standardize molecular surveillance [cf. (6)].

On the individual scale, genomic characterization of pathogen variants can be informative on the clinical pathogenesis of the disease [cf. (7, 8)]. Namely, the presence of drug-resistant variants or the interaction of genetically distinct variants within an infection can influence disease outcomes. Different pathogen variants can assemble in an infection due to co-transmission or independent infective events as a consequence of multiple infectious contacts [cf. (9)]. In malaria this is commonly referred to as multiplicity of infection (MOI) or complexity of infection (COI) (9–12). The concept of MOI and COI is particularly well-recognized in malaria as it is informative on transmission intensities and disease exposure [cf. (11, 13)]. Despite their recognition, MOI or COI are not uniformly defined in the literature [cf. (14)]. Here, we will define MOI in terms of a statistical framework. Although, the concept of MOI and the framework presented here are applicable to a variety of infectious diseases, we have applications to malaria in mind.

A common problem—well recognized in malaria—is the characterization of several different pathogen variants within an infection, because molecular methods do not yield phased genetic information [cf. (15, 16)]. In fact, molecular characterization of pathogen variants (typically haplotypes) and MOI are intrinsically coupled. In practice, two main approaches emerged. The first are *ad-hoc* approaches, which avoid the need to phase molecular information, by disregarding infections with multiple pathogen variants. These approaches are simple to apply at the cost of dismissing the full potential of molecular surveillance. The second are based on formal statistical models. The theoretical background of these methods is sophisticated and applications require some expertise in programming or bioinformatics. For malaria, several such methods have been developed. (Importantly, these methods are in general not restricted to malaria.) The most common ones have a similar underlying statistical framework and are based either on maximum-likelihood (ML) estimation [cf. (14, 17–20)], or Bayesian methods, e.g., classical Bayesian estimation [cf. (21, 22)], or Bayesian hidden Markov models [cf. (23, 24)]. Several methods to estimate MOI and haplotype frequencies are available as software tools [cf. the model comparison in (22)]. For instance, (18) provides a ML method to estimate the distribution of MOI and allele frequencies at one or two genetic markers, assuming that MOI follows either a (conditional) Poisson or (conditional) negative binomial distribution. In the case of the conditional Poisson distribution the method was further developed by (20, 25), who also provided efficient implementations. A user-friendly implementation which allows flexible data handling is provided by the R package MLMOI [cf. (26)]. A bias-corrected ML approach was provided by (14). MalHaploFreq [cf. (27)] uses a ML approach to estimate the distribution of haplotypes characterized by up to three biallelic loci (e.g., SNPs) and the distribution of MOI, which is assumed to follow either a Poisson, conditional Poisson, or negative binomial distribution. This approach was generalized to an arbitrary number of SNPs for the cases of the Poisson and conditional Poisson distribution [cf. (28)]. It makes use of the expectation-maximization algorithm (EM algorithm) to derive the ML estimates. However, the algorithm was neither derived in an explicit and efficient way, nor was an implementation made available. A different ML approach to estimate haplotype frequencies based on several approximations that guarantee numerical feasibility was suggested [cf. (29)]. This method also provides estimates for MOI, but based on several simplifications. Bayesian approaches include the method based on Gibbs sampling by (22) to estimate haplotype frequencies from SNP data including an error model, the Metropolis-Hastings algorithm of (30, 31) to estimate frequencies of haplotypes which are not restricted to biallelic loci. However, this approach requires heuristic estimates of MOI. For biallelic loci, the program COIL offers a classical Bayesian approach to estimate MOI (COI) [cf. (32)]. THE REAL McCOIL [cf. (33)] is a generalization based on the Metropolis-Hastings algorithm. It estimates MOI and minor-allele frequencies at uncorrelated SNPs in two different ways. Importantly, ML-based methods and Bayesian methods should yield consistent results, as both involve the likelihood function. In the strict sense, ML approaches provide point estimates for parameters, whereas Bayesian methods provide posterior distributions for parameters of interest. Agreement should be particularly strong if the prior distribution is uninformative, or if the prior distribution gives substantial weight to the true parameters and the data set is representative. Discrepancies between the methods are expected if: (i) the data is an “outlier” and the prior gives substantial weight to the true parameters (in which case, essentially all information is excerpted from the prior, which is more reliable than the posterior); (ii) the data is reliable, but the prior gives too much weight to the wrong parameters.

As genomics data is becoming more common in diseases like malaria, methods capable to handle such data became popular. For example, an approach to estimate MOI from deep-sequencing data is provided in (34). Moreover, methods considering relatedness of pathogen variants within infections become increasingly popular [cf. (35–40)]. Models such as DeploidIBD [cf. (40)] estimate the number and proportions of haplotypes in an infection alongside their identity-by-descent (IBD) profiles.

Here, we use a ML approach to estimate the frequencies of haplotypes, determined by n biallelic loci, and the distribution of MOI, assuming a conditional Poisson distribution. As with related methods, MOI is defined as the number of super-infections (i.e., independent infectious events during the course of the disease under the assumption of no co-transmission of pathogen variants). The proposed method is intended for a moderate number of loci, i.e., n should not be so large that individual haplotypes will be characterized. In particular, we employ the EM algorithm as in (28). While the general form provided by (28) is easily derived, a more explicit form is combinatorically involved and complicated. Here, we provide such an explicit form. This is the foundation of an efficient implementation of the algorithm. Such an implementation is provided as an easy-to-use R script. Importantly, we derive expressions for prevalence of haplotypes, i.e., the probability that a given haplotype occurs in an infection. Prevalence is mediated by MOI, i.e., it is derived from the haplotype frequency distribution and the distribution of MOI. If primarily interested in disease outcomes rather than the population genetics of the pathogen, prevalence is more relevant than the frequency distribution of haplotypes. Prevalence is notoriously difficult to estimate, especially from unphased molecular data. Namely, a statistical model, as the one presented here, is required for its estimation. Without such a model, *ad-hoc* estimates can be made from samples without ambiguity regarding haplotype phasing [e.g., as done in some of the analyses in (41)]. Such estimates are however biased. We assess the performance of the estimator of MOI, frequencies, prevalence, and *ad-hoc* approximations of prevalence in terms of bias and variance by numerical simulations.

As an example, we apply the method to estimate the frequency of malaria haplotypes associated with resistance against sulfadoxine-pyrimethamine (SP). Specifically, we apply the method to molecular data obtained from malaria-positive blood samples collected in Cameroon at two time points [cf. (42)].

We start with a formulation of the underlying statistical framework and a clear definition of MOI. Readers not focused on mathematical rigor shall feel free to move directly to the result section. Formal proofs and derivations are provided in the Mathematical Appendix.

2. Methods

A formal description of the statistical model is presented here. The model extends the method of (18), further developed by (25) to estimate multiplicity of infection (MOI) defined as the number of super-infections (i.e., independent infectious events during the course of the disease under the assumption of no co-transmissions/co-infections) and allele frequencies at a single-marker locus. Here, we extend the method to an arbitrary number of marker loci each with two alleles, e.g., single nucleotide polymorphisms (SNPs) as used in *P. vivax* or *P. falciparum* barcodes [cf. (6)], to estimate the haplotypes frequency distribution and MOI. As pointed out in (43), the assumption of no co-infections is not too strict. More precisely, the model approximately also holds if co- and super-infections occur.

2.1. Statistical model

Consider pathogen haplotypes, denoted $\boldsymbol{h}$, characterized by n biallelic markers. At each locus, the wildtype allele is coded by 0 and the mutant allele by 1. Hence, a haplotype is represented by a 0-1-vector indicating its allelic configuration, i.e., $\boldsymbol{h} = (h_1, \ldots, h_n)$, with $h_k \in \{0, 1\}$. A total of $H = 2^n$ haplotypes are possible. The set of all possible haplotypes is thus given by $\boldsymbol{h} \in \mathscr{H} = \{0, 1\}^n$.

Each haplotype $\boldsymbol{h}$ (0-1-vector) corresponds to a binary representation of the numbers $1, \ldots, 2^n$, namely to $[\boldsymbol{h}]_2 : = 1 + \sum_{l=1}^{n} h_l 2^{l-1}$. (As an example for $n = 4$ the haplotype $(1, 0, 0, 1)$ corresponds to the number $1+1\cdot2^0+0\cdot2^1+0\cdot2^2+1\cdot2^3 = 1+1+8 = 10$). We order haplotypes according to that representation.

The frequency of haplotype $\boldsymbol{h}$, denoted by $p_{\boldsymbol{h}}$, is its relative abundance in the pathogen population (assessed at a particular census point). For example, in the case of malaria, the frequency of a haplotype is its relative abundance in the sporozoite population in the mosquitoes' salivary glands (20). Collectively, the frequencies form the vector $\boldsymbol{p} : = (p_{\boldsymbol{h}})_{\boldsymbol{h}\in\mathscr{H}} = (p_1, \ldots, p_H)$, where $p_k = p_{\boldsymbol{h}}$ if $[\boldsymbol{h}]_2 = k$. In practice, $p_k = 0$ for several haplotypes, since not all 2^n possible haplotypes will be present in the pathogen population.

It is assumed that at each infective event, exactly one pathogen haplotype is transmitted. However, individuals can get (super-) infected several times during one disease episode. The number of (super-) infections during one disease episode is referred here to as multiplicity of infection (MOI). We treat the terms complexity of infection (COI) and MOI synonymously here. The term super-infections refers here to

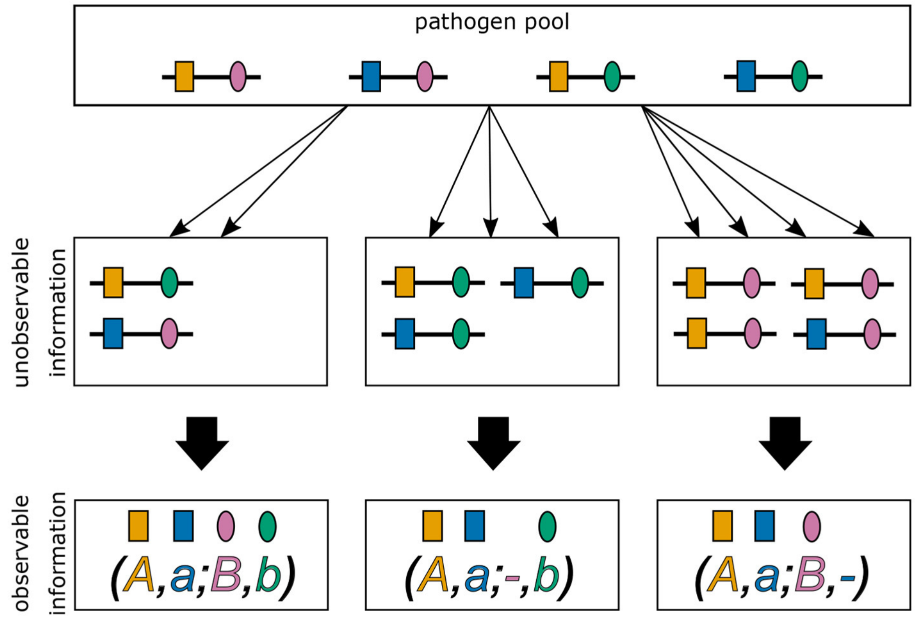

FIGURE 1
Ambiguity in haplotype information for two biallelic loci. Illustrated are three different infections from the same pathogen population. The first infection **(middle left)** describes a super-infection with two haplotypes, i.e., MOI = 2. The corresponding observation **(bottom left)** provides only unphased (i.e., ambiguous) haplotype information. It is impossible to reconstruct with certainty the haplotypes actually present in the infection and the corresponding MOI. The second infection **(middle)**, illustrates a super-infection with two haplotypes transmitted one and two times, respectively, i.e., MOI = 3. From the observed information, the haplotypes present in the infections can be unambiguously phased. However, MOI remains unknown. The last infection **(middle right)** is similar to the second, however with MOI = 4.

independent infective events without co-transmissions, i.e., only one pathogen variant is transmitted. In contrast, a co-infection is one infective event during which several pathogen variants are co-transmitted. However, the model is still approximately applicable if co-transmissions occur [cf. (43)].

Assuming that infections are rare and independent, MOI follows a Poisson distribution. When considering only disease-positive individuals, MOI follows a conditional (or positive) Poisson distribution. Thus, the probability to be (super-) infected exactly m times (MOI $= m$) is given by [see (25)]

$$\kappa_m = \frac{1}{e^{\lambda}-1}\frac{\lambda^m}{m!}, \quad m = 1, 2, 3, \ldots. \tag{1a}$$

Note that, a zero-inflated Poisson distribution [cf. (44)] yields the same conditional Poisson distribution. Therefore, $m \sim \text{CPoiss}(\lambda)$, where λ is the parameter characterizing the conditional Poisson distribution.

The probability generating function (PGF) of the conditional Poisson distribution is given by

$$G(z) := \mathbb{E}[z^m] = \sum_{m=1}^{\infty} \kappa_m z^m = \frac{e^{\lambda z}-1}{e^{\lambda}-1}, \tag{1b}$$

and the mean MOI is given by (see 25)

$$\psi := \mathbb{E}[m] = \frac{\lambda}{1-e^{-\lambda}}. \tag{1c}$$

At each infective event, exactly one haplotype, randomly chosen from the pathogen population, is transmitted to the host. Given, an individual is super-infected m times (MOI $= m$), the process of infection corresponds to multinomially sampling from the pathogen population. If $m_{\boldsymbol{h}}$ is the number of times an individual was infected with haplotype $\boldsymbol{h}$ (necessarily $|\boldsymbol{m}| := \sum_{\boldsymbol{h}\in\mathscr{H}} m_{\boldsymbol{h}} = m_1 + \ldots + m_H = m$), the infection is subsumed by the vector $\boldsymbol{m} := (m_{\boldsymbol{h}})_{\boldsymbol{h}\in\mathscr{H}} = (m_1, \ldots, m_H)$. Therefore, given MOI m, infection $\boldsymbol{m}$ (with $|\boldsymbol{m}| = m$) occurs with probability

$$P(\boldsymbol{m} \mid m) = \binom{m}{\boldsymbol{m}} \boldsymbol{p}^{\boldsymbol{m}}, \tag{2}$$

where $\binom{m}{\boldsymbol{m}} = \frac{m!}{m_1! \ldots m_H!}$, and $\boldsymbol{p}^{\boldsymbol{m}} = p_1^{m_1} \ldots p_H^{m_H}$.

In practice, the vector $\boldsymbol{m}$ and even MOI m are unobservable from a clinical specimen. In addition, assays to determine genetic information usually do not yield full haplotype information, i.e., if multiple different haplotypes are present within an infection, assays yield ambiguous genetic information due to the lack of phasing (cf. Figure 1) (15). It is assumed here, that only the absence/presence of alleles at every locus is assessable. [Notably, in the case of phased data, haplotypes as defined here, would be equivalent to alleles at a single multi-allelic marker locus and can be analyzed with the methods of (18, 20, 25).]

We denote allelic information of an infection by a vector $\boldsymbol{x} = (x_1, \ldots, x_n)$, where x_k is the set of alleles detected at locus k in the sample. It is assumed that all alleles that are present in an infection are actually detected and that alleles are not erroneously detected. Therefore, x_k equals one of the sets $\{0\}$, $\{1\}$, or $\{0, 1\}$, corresponding to, respectively, the presence of the wildtype, the mutant, or both alleles at locus k. The set of all possible observations is $\mathscr{O} : = (\{\{0\}, \{1\}, \{0, 1\}\})^n$. There are a total of 3^n possible observations. Several different infections $\boldsymbol{m}$ can yield the same observation $\boldsymbol{x}$ ($\boldsymbol{m} \rightarrow \boldsymbol{x}$). The set of all infections with MOI equal to m which yield observation $\boldsymbol{x}$ is denoted by

$$M_{\boldsymbol{x}}^{(m)} : = \{\boldsymbol{m} \mid \boldsymbol{m} \rightarrow \boldsymbol{x}, |\boldsymbol{m}| = m\}. \tag{3a}$$

Furthermore, we denote the set of all haplotypes which are compatible with observation $\boldsymbol{x}$ (i.e., the set of all haplotypes that could potentially be present in the underlying infection) by

$$A_{\boldsymbol{x}} : = \{\boldsymbol{h} = (h_1, \ldots, h_n) \mid h_k \in x_k \text{ for all } k\}. \tag{3b}$$

Let us denote the set of sub-observations of observation $\boldsymbol{x}$, i.e., all observations with at most the same alleles detected at each locus as in $\boldsymbol{x}$ (cf. Supplementary Figure 1), by

$$\mathscr{A}_{\boldsymbol{x}} : = \{\boldsymbol{y} = (y_1, \ldots, y_n) \mid y_k \subseteq x_k \text{ for all } k\}. \tag{3c}$$

If $\boldsymbol{y}$ is a sub-observation of $\boldsymbol{x}$, i.e., $\boldsymbol{y} \in \mathscr{A}_{\boldsymbol{x}}$, we write $\boldsymbol{y} \preceq \boldsymbol{x}$. Note that "$\preceq$" defines a partial order on the set of possible observations. We further define the proper sub-observation $\boldsymbol{y}$ of $\boldsymbol{x}$ by

$$\boldsymbol{y} \prec \boldsymbol{x} \ :\Leftrightarrow \ \boldsymbol{y} \preceq \boldsymbol{x} \wedge \boldsymbol{x} \neq \boldsymbol{y}. \tag{4}$$

Using this notation the set $M_{\boldsymbol{x}}^{(m)}$ is rewritten as

$$\begin{aligned} M_{\boldsymbol{x}}^{(m)} =&\{\boldsymbol{m} \mid m_{\boldsymbol{h}} = 0 \text{ if } \boldsymbol{h} \notin A_{\boldsymbol{x}}, \\ &|\boldsymbol{m}| = m\} \setminus \bigcup_{\boldsymbol{y} \prec \boldsymbol{x}} \{\boldsymbol{m} \mid m_{\boldsymbol{h}} = 0 \text{ if } \boldsymbol{h} \notin A_{\boldsymbol{y}}, \ |\boldsymbol{m}| = m\}, \end{aligned} \tag{5}$$

where $A_{\boldsymbol{y}}$ is defined as (3) but for the proper sub-observation $\boldsymbol{y}$ rather than for the original observation $\boldsymbol{x}$. Given an infection with MOI m, the probability of observing $\boldsymbol{x}$ is

$$P(\boldsymbol{x} \mid m) = \frac{P(\boldsymbol{x}, m)}{\kappa_m}. \tag{6a}$$

Therefore,

$$\begin{aligned} P(\boldsymbol{x}, m) =&P\left(M_{\boldsymbol{x}}^{(m)}\right) = \sum_{\boldsymbol{m} \in M_{\boldsymbol{x}}^{(m)}} P(\boldsymbol{m}) = \sum_{\boldsymbol{m} \in M_{\boldsymbol{x}}^{(m)}} P(\boldsymbol{m}|m)\kappa_m \\ =&\kappa_m \sum_{\boldsymbol{m} \in M_{\boldsymbol{x}}^{(m)}} \binom{m}{\boldsymbol{m}} \boldsymbol{p}^{\boldsymbol{m}}. \end{aligned} \tag{6b}$$

The probability of observation $\boldsymbol{x}$ becomes

$$P(\boldsymbol{x}) = \sum_{m=1}^{\infty} P(\boldsymbol{x}, m) = \sum_{m=1}^{\infty} \kappa_m \sum_{\boldsymbol{m} \in M_{\boldsymbol{x}}^{(m)}} \binom{m}{\boldsymbol{m}} \boldsymbol{p}^{\boldsymbol{m}}. \tag{6c}$$

Henceforth, to simplify the notation we will use $P_{\boldsymbol{x}}$ to denote the probability of observing $\boldsymbol{x}$ instead of $P(\boldsymbol{x})$. By using (3a) and the inclusion-exclusion principle the inner sum on the right-hand side of (6c) can be rewritten as

$$\begin{aligned} \sum_{\boldsymbol{m} \in M_{\boldsymbol{x}}^{(m)}} \binom{m}{\boldsymbol{m}} \boldsymbol{p}^{\boldsymbol{m}} =& \sum_{\substack{\boldsymbol{m} : |\boldsymbol{m}| = m \\ m_{\boldsymbol{h}} = 0 \text{ if } \boldsymbol{h} \notin A_{\boldsymbol{x}}}} \binom{m}{\boldsymbol{m}} \boldsymbol{p}^{\boldsymbol{m}} \\ &+ \sum_{\boldsymbol{y} \prec \boldsymbol{x}} (-1)^{\sum\limits_{k=1}^{n} \left(|x_k| - |y_k|\right)} \sum_{\substack{\boldsymbol{m} : |\boldsymbol{m}| = m \\ m_{\boldsymbol{h}} = 0 \text{ if } \boldsymbol{h} \notin A_{\boldsymbol{y}}}} \binom{m}{\boldsymbol{m}} \boldsymbol{p}^{\boldsymbol{m}}, \end{aligned} \tag{7a}$$

where $|x_k|$ and $|y_k|$ are, respectively, the cardinals of x_k, and y_k. Let $N_{\boldsymbol{x}}$ denote the number of loci in observation $\boldsymbol{x}$ at which both alleles were detected, i.e., $N_{\boldsymbol{x}} = |\{k \mid |x_k| = 2\}|$. The number of loci with a single allele detected is then $n - N_{\boldsymbol{x}}$. Hence, the number of alleles detected in observation $\boldsymbol{x}$ is given by $2N_{\boldsymbol{x}} + n - N_{\boldsymbol{x}} = n + N_{\boldsymbol{x}}$. We hence obtain

$$\sum_{k=1}^{n} \left(|x_k| - |y_k|\right) = N_{\boldsymbol{x}} - N_{\boldsymbol{y}}, \tag{7b}$$

and

$$\begin{aligned} \sum_{\boldsymbol{m} \in M_{\boldsymbol{x}}^{(m)}} \binom{m}{\boldsymbol{m}} \boldsymbol{p}^{\boldsymbol{m}} =& \sum_{\substack{\boldsymbol{m} : |\boldsymbol{m}| = m \\ m_{\boldsymbol{h}} = 0 \text{ if } \boldsymbol{h} \notin A_{\boldsymbol{x}}}} \binom{m}{\boldsymbol{m}} \boldsymbol{p}^{\boldsymbol{m}} \\ &+ \sum_{\boldsymbol{y} \prec \boldsymbol{x}} (-1)^{N_{\boldsymbol{x}} - N_{\boldsymbol{y}}} \sum_{\substack{\boldsymbol{m} : |\boldsymbol{m}| = m \\ m_{\boldsymbol{h}} = 0 \text{ if } \boldsymbol{h} \notin A_{\boldsymbol{y}}}} \binom{m}{\boldsymbol{m}} \boldsymbol{p}^{\boldsymbol{m}} \\ =& \sum_{\boldsymbol{y} \preceq \boldsymbol{x}} (-1)^{N_{\boldsymbol{x}} - N_{\boldsymbol{y}}} \sum_{\substack{\boldsymbol{m} : |\boldsymbol{m}| = m \\ m_{\boldsymbol{h}} = 0 \text{ if } \boldsymbol{h} \notin A_{\boldsymbol{y}}}} \binom{m}{\boldsymbol{m}} \boldsymbol{p}^{\boldsymbol{m}} \\ =& \sum_{\boldsymbol{y} \in \mathscr{A}_{\boldsymbol{x}}} (-1)^{N_{\boldsymbol{x}} - N_{\boldsymbol{y}}} \sum_{\substack{\boldsymbol{m} : |\boldsymbol{m}| = m \\ m_{\boldsymbol{h}} = 0 \text{ if } \boldsymbol{h} \notin A_{\boldsymbol{y}}}} \binom{m}{\boldsymbol{m}} \boldsymbol{p}^{\boldsymbol{m}}. \end{aligned} \tag{7c}$$

Therefore, the probability of observing $\boldsymbol{x}$ in (6c) becomes

$$P_{\boldsymbol{x}} = \sum_{m=1}^{\infty} \kappa_m \sum_{\boldsymbol{y} \in \mathscr{A}_{\boldsymbol{x}}} (-1)^{N_{\boldsymbol{x}} - N_{\boldsymbol{y}}} \sum_{\substack{\boldsymbol{m} : |\boldsymbol{m}| = m \\ m_{\boldsymbol{h}} = 0 \text{ if } \boldsymbol{h} \notin A_{\boldsymbol{y}}}} \binom{m}{\boldsymbol{m}} \boldsymbol{p}^{\boldsymbol{m}}. \tag{8a}$$

By the multinomial theorem, we have

$$\sum_{\substack{\boldsymbol{m} : |\boldsymbol{m}| = m \\ m_{\boldsymbol{h}} = 0 \text{ if } \boldsymbol{h} \notin A_{\boldsymbol{y}}}} \binom{m}{\boldsymbol{m}} \boldsymbol{p}^{\boldsymbol{m}} = \left(\sum_{\boldsymbol{h} \in A_{\boldsymbol{y}}} p_{\boldsymbol{h}}\right)^{m}. \tag{8b}$$

Using the PGF (8a) becomes

$$\begin{aligned} P_{\boldsymbol{x}} &= \sum_{m=1}^{\infty} \kappa_m \sum_{\boldsymbol{y}\in\mathscr{A}_{\boldsymbol{x}}} (-1)^{N_{\boldsymbol{x}}-N_{\boldsymbol{y}}} \left(\sum_{\boldsymbol{h}\in A_{\boldsymbol{y}}} p_{\boldsymbol{h}} \right)^m \\ &= \sum_{\boldsymbol{y}\in\mathscr{A}_{\boldsymbol{x}}} (-1)^{N_{\boldsymbol{x}}-N_{\boldsymbol{y}}} \sum_{m=1}^{\infty} \kappa_m \left(\sum_{\boldsymbol{h}\in A_{\boldsymbol{y}}} p_{\boldsymbol{h}} \right)^m \qquad (8c) \\ &= \sum_{\boldsymbol{y}\in\mathscr{A}_{\boldsymbol{x}}} (-1)^{N_{\boldsymbol{x}}-N_{\boldsymbol{y}}} G\left(\sum_{\boldsymbol{h}\in A_{\boldsymbol{y}}} p_{\boldsymbol{h}} \right). \end{aligned}$$

This probability depends on the model parameters λ, which appears in the PGF, and the vector of haplotype frequencies $\boldsymbol{p}$. Hence, the parameter space of the model is

$$\Theta := \mathbb{R}^+ \times \mathcal{S}_H = \left\{ (\lambda, \boldsymbol{p}) \mid \lambda > 0 \text{ and } \boldsymbol{p} \in \mathcal{S}_H \right\}, \qquad (9)$$

where, $\mathcal{S}_H := \left\{ (p_1, \ldots, p_H) \,\middle|\, \sum_{k=1}^{H} p_k = 1 \text{ and } p_k \geq 0, \text{ for all } k \right\}$ is the $H-1$-dimensional simplex.

The true parameter vector $\boldsymbol{\theta} = (\lambda, \boldsymbol{p})$ is unknown and has to be estimated from empirical data. Assume a dataset $\mathscr{X}$ consisting of N observations $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)}$, where the notation $\boldsymbol{x}^{(j)} = \left(x_1^{(j)}, \ldots, x_n^{(j)} \right)$ is used for the jth observation. For the dataset $\mathscr{X}$, let $n_{\boldsymbol{x}}$ be the number of times observation $\boldsymbol{x}$ is made. Naturally,

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

Using (8c), the likelihood function of the parameter $\boldsymbol{\theta} = (\lambda, \boldsymbol{p})$ given the data $\mathscr{X}$ is given by

$$\mathcal{L}_{\mathscr{X}}(\boldsymbol{\theta}) = \prod_{j=1}^{N} P_{\boldsymbol{x}^{(j)}} = \prod_{\boldsymbol{x}\in\mathscr{O}} \left(\sum_{\boldsymbol{y}\in\mathscr{A}_{\boldsymbol{x}}} (-1)^{N_{\boldsymbol{x}}-N_{\boldsymbol{y}}} G\left(\sum_{\boldsymbol{h}\in A_{\boldsymbol{y}}} p_{\boldsymbol{h}} \right) \right)^{n_{\boldsymbol{x}}}. \qquad (10)$$

Hence, the log-likelihood function becomes

$$\begin{aligned} \ell_{\mathscr{X}}(\boldsymbol{\theta}) &= \log\left(\mathcal{L}_{\mathscr{X}}(\boldsymbol{\theta}) \right) \\ &= \sum_{\boldsymbol{x}\in\mathscr{O}} n_{\boldsymbol{x}} \log\left(\sum_{\boldsymbol{y}\in\mathscr{A}_{\boldsymbol{x}}} (-1)^{N_{\boldsymbol{x}}-N_{\boldsymbol{y}}} G\left(\sum_{\boldsymbol{h}\in A_{\boldsymbol{y}}} p_{\boldsymbol{h}} \right) \right). \qquad (11) \end{aligned}$$

To obtain the maximum-likelihood estimate (MLE) $\hat{\boldsymbol{\theta}} = (\hat{\lambda}, \hat{\boldsymbol{p}})$ the log-likelihood function needs to be maximized. The complexity of the log-likelihood function does not permit a closed solution, and must be maximized numerically. For this purpose the expectation-maximization (EM)-algorithm will be used (17). This will be discussed in Section 3.2.

2.1.1. Confidence intervals

To ascertain uncertainty of the estimates, confidence intervals (CIs) can be derived. A straightforward approach is to derive bootstrap CIs. The simplest type of boostrap CIs are the non-parametric percentile CIs [cf. (45), Chapter 13]. To obtain a $(1-\alpha)$% bootstrap CIs from a dataset $\mathscr{X}$ of sample size N, we sample B (e.g., $B = 10{,}000$) datasets $\mathscr{X}_1, \ldots, \mathscr{X}_B$, each of sample size N with replacement from $\mathscr{X}$. For each dataset $\mathscr{X}_b$, we obtain the MLEs $\hat{\boldsymbol{\theta}}^{(b)}$. For the desired parameter θ_k, the $\frac{\alpha}{2}$% and $(1-\frac{\alpha}{2})$% percentiles, $\hat{\theta}^*_{k,\frac{\alpha}{2}}$ and $\hat{\theta}^*_{k,(1-\frac{\alpha}{2})}$, respectively, are determined from the sequence $\hat{\boldsymbol{\theta}}_k^{(1)}, \ldots, \hat{\boldsymbol{\theta}}_k^{(B)}$. The $(1-\alpha)$% CI is then given by

$$\left(\hat{\theta}^*_{k,\frac{\alpha}{2}}, \hat{\theta}^*_{k,1-\frac{\alpha}{2}} \right). \qquad (12)$$

Clearly, more advanced bootstrap CIs, e.g., bias-corrected and accelerated (BCa) boostsrap CIs [cf. (45), Chapter 14] or parametric bootstrap CIs [cf. (45), Chapter 12] can be calculated similarly.

2.1.2. Assessing bias and variance of the estimator

MLEs have desirable asymptotic properties, i.e., for large sample size. In practice, sample size is often limited, and the quality of the estimator needs to be investigated under finite sample sizes. Because no explicit solution exists for the MLE, its performance in terms of bias and variance needs to be investigated by numerical simulations.

Bias and variance of the MLE will be affected by: (i) sample size N; (ii) the number of considered loci n, i.e., the genetic architecture; (iii) the value of the MOI parameter λ; (iv) the frequency distribution of haplotypes $\boldsymbol{p}$.

To investigate the properties of the MLE for a representative range of parameters we proceeded as follows (parameters used in the simulation study are described below and summarized in Table 1). For a set of parameters (N, n, λ, $\boldsymbol{p}$) we generated $K = 100{,}000$ datasets $\mathscr{X}_1, \ldots, \mathscr{X}_K$ of size N according to the model (8c). For each dataset $\mathscr{X}_k$, the MLE $(\hat{\boldsymbol{\theta}}_k) = (\hat{\lambda}_k, \hat{\boldsymbol{p}}^{(k)})$ was calculated. From each $\hat{\lambda}_k$ the mean MOI $\hat{\psi}_k$ was calculated according to (1c). The bias and variance of the mean MOI ψ were estimated as

$$\text{bias}(\hat{\psi}) = \overline{\psi} - \psi, \qquad (13a)$$

and

$$\text{Var}(\hat{\psi}) = \frac{1}{K-1} \sum_{k=1}^{K} (\hat{\psi}_k - \bar{\psi})^2, \qquad (13b)$$

where

$$\overline{\psi} = \frac{1}{K} \sum_{k=1}^{K} \hat{\psi}_k. \qquad (13c)$$

TABLE 1 Summary of model parameters chosen for the simulations to assess the estimator's performance.

Parameter	Description			Value
K	Number of simulated datasets			$100,000$
n	Number of loci markers			$2, 5, 10$
N	Sample size			$50, 100, 150, 200, 500$
λ	MOI parameter			$0.1, 0.25, 0.5, 1, 1.5, 2, 2.5$
			Symmetric	**Skewed**
$\boldsymbol{p}$	Hapl. freq. (simulated data)	$n = 2$:	$p_1 = \ldots = p_4 = \frac{1}{4}$	$p_1 = 0.7,$ $p_2 = \ldots = p_4 = 0.1$
		$n = 5$:	$p_1 = \ldots = p_{32} = \frac{1}{32}$	$p_1 = 0.7,$ $p_2 = \ldots = p_{32} = \frac{0.3}{31}$

Freq., Frequency; Hapl., Haplotype.

To allow comparisons between different parameter ranges it is more appropriate to consider the relative bias and coefficient of variation which are independent of the scale, i.e.,

$$\frac{\text{bias}(\hat{\psi})}{\psi}, \tag{14a}$$

and

$$\frac{\sqrt{\text{Var}(\hat{\psi})}}{\psi}. \tag{14b}$$

For each haplotype frequency $p_{\boldsymbol{h}}$, bias and variance were defined in the same way with obvious modifications.

2.1.2.1. Genetic architecture

Considering the number of biallelic loci, for the simulations we assumed $n = 2$, 5 to perform systematic investigations of the estimator. The number of possible haplotypes was then 4 and 32, respectively, for $n = 2$, 5. As the number of loci increases due to the curse of dimensionality it becomes too exhaustive to perform systematic investigations. Hence, in addition, we chose two specific distributions for $n = 10$, which correspond to distributions of drug-resistant haplotypes which were previously empirically estimated. Importantly, the method is not limited to just 10 loci.

2.1.2.2. MOI parameter

Concerning the MOI parameter we chose $\lambda = 0.1, 0.25, 0.5, 1, 1.5, 2, 2.5$, corresponding to a mean MOI $\psi = 1.05, 1.13, 1.27, 1.58, 1.93, 2.31, 2.72$. In the case of malaria, this corresponds to low transmission $\psi < 1.27$, intermediate transmission $1.27 \leq \psi < 1.93$ and high transmission $\psi \geq 1.93$ (14).

2.1.2.3. Haplotype frequency distribution

The following haplotype frequency distributions $\boldsymbol{p}$ were chosen. First, a completely uniform (balanced) distribution was chosen, i.e., each of the $H = 2^n$ haplotype had the same frequency,

$$p_1 = \ldots = p_H = \frac{1}{H}. \tag{15a}$$

Second, a skewed distribution with one predominant haplotype was chosen. The frequency of the predominant haplotype was chosen to be 70%, while the remaining haplotypes all had the same frequency. In particular, we chose

$$p_1 = 0.7, p_2 = \ldots = p_H = \frac{0.3}{H-1}. \tag{15b}$$

For $n = 2$ this yielded, $p_1 = 0.7, p_2 = p_3 = p_4 = 0.1$ and for $n = 5$ loci $p_1 = 0.7, p_2 = \ldots = p_{32} = 0.0097$.

Third, we chose specific empirical distributions for the case $n = 10$. The reason is that the dimension of the parameter space becomes high ($H = 1,024$), but most haplotypes will not be realized in a population. Specifically, we assumed two haplotype frequency distributions that correspond to empirically estimated distributions of *P. falciparum* malaria haplotypes. These haplotypes were characterized by $n = 10$ SNPs associated with resistance to sulfadoxine-pyrimethamine (SP). The two haplotype frequency distributions were estimated from a population in Siaya County, Kenya, respectively, in 2005 and 2010 (see (46)).

In 2005, the frequencies of detected haplotypes were $p_2 = 0.055$, $p_5 = 0.016$, $p_6 = 0.171$, $p_{11} = 0.015$, $p_{13} = 0.024$, $p_{14} = 0.719$, respectively, while in 2010 they were $p_1 = 0.007$, $p_3 = 0.015$, $p_6 = 0.084$, $p_7 = 0.006$, $p_{14} = 0.791$, $p_{15} = 0.007$, $p_{16} = 0.009$, $p_{19} = 0.081$, respectively (see Table 2).

2.1.2.4. Sample size

Sample size is crucial to the performance of an estimator. To evaluate the effect of sample size, we constructed datasets of size $N = 50, 100, 150, 200, 500$. In malaria $N = 50 - 150$ are typical sample sizes. The large sample size $N = 500$, which is

TABLE 2 Frequencies of SP-resistant haplotypes from the Kenyan data used for simulation study.

Haplotype		Frequency p	
		Years	
dhfr	*dhps*	2005	2010
NCN	SGEA	–	0.007
NRN	SGEA	0.055	–
ICN	SAKA	–	0.015
ICN	SGKA	0.016	–
ICN	SGEA	0.171	0.084
ICN	AAKA	–	0.006
IRN	SAKA	0.015	–
IRN	SGKA	0.024	–
IRN	SGEA	0.719	0.791
IRN	SGEG	–	0.007
IRN	AAKA	–	0.009
IRN	AGEA	–	0.081

The first column shows the amino acid sequence at codons 51, 59, and 108 at the *dhfr* locus and the second column shows the amino acid sequence at codons 436, 437, 540, and 581 at the *dhps* locus.

becoming more common in malaria, but might still be infeasible for low transmission areas. Nevertheless, considering $N = 500$ helps understand the asymptotic behavior of the estimator.

We used R (47) to implement the simulation study and create the graphical outputs. The code is available at: https://github.com/Maths-against-Malaria/MultiLociBiallelicModel.git.

2.2. Data application

As an application, we estimated the frequency of malaria haplotypes associated with resistance against SP. The data was taken from (42, 48, 49) and is described there in detail. In short, it was collected in Yaoundé, Cameroon in 2001/2002 and 2004/2005. Mutations at codons 51, 59, 108, and 164 at the *dhfr* locus on chromosome 4 and 436, 437, 540, 581, and 613 at the *dhps* locus on chromosome 8 were determined either by direct sequencing or pyrosequencing. Due to missing data, we included 165 samples from 2001/2002 and 165 samples from 2004/2005.

3. Results

For molecular surveillance of parasite haplotypes, obtaining adequate estimates of haplotype frequencies is crucial. From a clinical point of view, the occurrence of particular haplotypes in infections, i.e., the prevalence of haplotypes, is more relevant. Importantly, frequency and prevalence are not the same, as the latter is mediated by MOI [cf. (43)]. First, we clarify the relationship between frequency, prevalence, and MOI. Second, an efficient algorithm for estimating MOI, haplotype frequencies, and prevalence is provided. Finally, the properties of the estimator are investigated numerically.

3.1. Prevalence and similar quantities

The frequency of haplotype $\boldsymbol{h}$, i.e., $p_{\boldsymbol{h}}$, defines its relative abundance in the pathogen population. According to the underlying model, $p_{\boldsymbol{h}}$ is the probability that haplotype $\boldsymbol{h}$ is transmitted at a given infective event. However, several infective events (super-infections) can cause an infection, so that the probability that haplotype $\boldsymbol{h}$ is transmitted at any infective event exceeds $p_{\boldsymbol{h}}$. The probability that haplotype $\boldsymbol{h}$ is present in an infection is called its prevalence.

Typically, molecular assays do not provide phased information. Hence, if several haplotypes are present within an infection, it is ambiguous which haplotypes are actually infecting (cf. Figure 1).

Observations that carry only one haplotype are called single infections, i.e., an observation $\boldsymbol{x}$ is a single infection if $|x_k| = 1 \ \forall k$. Molecular information from single infections are unambiguous. However, even for single infections MOI is unobservable (since it is unclear how many times the host was super-infected with the same haplotype).

Infections with two or more distinct haplotypes are called multiple infections. The resulting molecular information is ambiguous, except in the case of exactly two super-infecting haplotypes that differ at only one locus. We call such super-infections unambiguous multiple infections. Namely, $\tilde{\boldsymbol{x}}$ is an unambiguous multiple infection if there exist a unique locus k such that $|\tilde{x}_k| = 2$ and $|\tilde{x}_l| = 1 \ \forall \, l \neq k$. Clearly, MOI—as for every sample—is unobservable.

Since haplotype information is ambiguous in multiple infections, to assess the prevalence of haplotypes sometimes only unambiguous samples are considered in practice. Here, we first define prevalence in general (unobservable prevalence). Second, we derive the probability to observe a given haplotype in unambiguous samples (conditional prevalence).

3.1.1. Prevalence

Since haplotype information is typically unavailable from molecular assays, haplotypes are per se not observable in molecular samples. To emphasize this fact we call the probability that a haplotype occurs in an infection "unobservable prevalence." The unobservable prevalence of $\boldsymbol{h}$ is denoted by $q_{\boldsymbol{h}}$ and the probability that haplotype $\boldsymbol{h}$ does not occur in infection by $q_{-\boldsymbol{h}} = 1 - q_{\boldsymbol{h}}$. The unobservable prevalence is hence

$$q_{\boldsymbol{h}} = 1 - q_{-\boldsymbol{h}} = 1 - \sum_{m=1}^{\infty} \kappa_m \sum_{\substack{\boldsymbol{m}\,:\,|\boldsymbol{m}|=m \\ m_{\boldsymbol{h}}=0}} \binom{m}{\boldsymbol{m}} \boldsymbol{p}^{\boldsymbol{m}}. \tag{16a}$$

By using the multinomial theorem one obtains

$$\sum_{\substack{\boldsymbol{m}\,:\,|\boldsymbol{m}|=m \\ m_{\boldsymbol{h}}=0}} \binom{m}{\boldsymbol{m}} \boldsymbol{p}^{\boldsymbol{m}} = \Big(\sum_{\boldsymbol{i} \in \mathscr{H}} p_{\boldsymbol{i}} - p_{\boldsymbol{h}} \Big)^m = (1 - p_{\boldsymbol{h}})^m. \tag{16b}$$

This yields

$$\begin{aligned} q_{\boldsymbol{h}} &= 1 - \sum_{m=1}^{\infty} \kappa_m (1 - p_{\boldsymbol{h}})^m = 1 - G(1 - p_{\boldsymbol{h}}) \\ &= 1 - \frac{e^{\lambda(1-p_{\boldsymbol{h}})} - 1}{e^{\lambda} - 1}. \end{aligned} \tag{16c}$$

Hence, prevalence is derived readily from the MOI parameter λ and the frequencies. Given the frequency of haplotype $\boldsymbol{h}$ its unobservable prevalence increases with increasing MOI (increasing λ).

The unobservable prevalence $q_{\boldsymbol{h}}$ always exceeds the frequency $p_{\boldsymbol{h}}$ of haplotype $\boldsymbol{h}$. The higher transmission intensities, the more does prevalence exceed frequency. This is illustrated in Figures 10–13. In the limit of $\lambda \to 0$, i.e., every infection is a single infection with MOI $m = 1$, prevalence and frequency coincide. Distinguishing between frequency and prevalence is hence particularly important in areas of seasonal disease transmission, where the parameter λ will fluctuate between seasons.

3.1.2. Conditional prevalence

Because of ambiguity of haplotype information in multiple infections, it is impossible to identify the number of samples containing haplotype $\boldsymbol{h}$ in a dataset. In practice, often only unambiguous samples are considered, to determine prevalence. Here, we derive the corresponding quantity in the underlying framework, i.e., the prevalence of haplotype $\boldsymbol{h}$, conditioned on observing only unambiguous data. The quantity is referred to as "conditional prevalence."

We denote the set of all possible unambiguous observations by $\tilde{\mathscr{O}}$. The conditional prevalence is

$$r_{\boldsymbol{h}|\tilde{\mathscr{O}}} := P(\boldsymbol{h} \mid \tilde{\mathscr{O}}) = \frac{P(\boldsymbol{h}, \tilde{\mathscr{O}})}{P(\tilde{\mathscr{O}})} = \frac{r_{\boldsymbol{h}}}{P(\tilde{\mathscr{O}})}, \tag{17}$$

where $r_{\boldsymbol{h}} := P(\boldsymbol{h}, \tilde{\mathscr{O}})$ is the probability to observe haplotype $\boldsymbol{h}$ in an unambiguous observation, and $P(\tilde{\mathscr{O}})$ is the probability of unambiguous observations. For each haplotype $\boldsymbol{h}$, let $U_{\boldsymbol{h}}$ be the set of all haplotypes $\boldsymbol{i}$, which yield unambiguous observations with $\boldsymbol{h}$ (i.e., if only $\boldsymbol{h}$ and $\boldsymbol{i}$ are present in an infection, it is an unambiguous infection). Note that there are exactly n haplotypes $\boldsymbol{i}$ such that $\boldsymbol{i} \in U_{\boldsymbol{h}}$. Formally, we have

$$U_{\boldsymbol{h}} := \{\boldsymbol{i} \in \mathscr{H} \mid \exists!\, k : \boldsymbol{i}_k \neq \boldsymbol{h}_k\}. \tag{18}$$

The quantity $r_{\boldsymbol{h}}$ is obtained as the sum of the probabilities of multiple infections with only one haplotype $\boldsymbol{i} \in U_{\boldsymbol{h}}$ and $\boldsymbol{h}$, or single infections with haplotype $\boldsymbol{h}$. An unambiguous observation with $\boldsymbol{h}$ and MOI m is obtained by randomly sampling $m_{\boldsymbol{i}}$ times $\boldsymbol{i} \in U_{\boldsymbol{h}}$ and $m_{\boldsymbol{h}} = m - m_{\boldsymbol{i}}$ times $\boldsymbol{h}$. Note, if $m_{\boldsymbol{h}} = m$ or $m_{\boldsymbol{i}} = m$, a single infection with $\boldsymbol{h}$ or $\boldsymbol{i}$ is obtained, respectively. The latter are irrelevant for the prevalence of $\boldsymbol{h}$. One obtains

$$r_{\boldsymbol{h}} := \sum_{m=1}^{\infty} \kappa_m \sum_{\boldsymbol{i} \in U_{\boldsymbol{h}}} \sum_{m_{\boldsymbol{i}}=1}^{m-1} \binom{m}{m_{\boldsymbol{i}}} p_{\boldsymbol{i}}^{m_{\boldsymbol{i}}} p_{\boldsymbol{h}}^{m-m_{\boldsymbol{i}}} + \sum_{m=1}^{\infty} \kappa_m p_{\boldsymbol{h}}^m. \tag{19a}$$

As shown in Section Prevalence estimates of the Mathematical Appendix, this quantity simplifies to

$$r_{\boldsymbol{h}} = \sum_{\boldsymbol{i} \in U_{\boldsymbol{h}}} \big[G(p_{\boldsymbol{h}} + p_{\boldsymbol{i}}) - G(p_{\boldsymbol{i}}) \big] - (n-1) G(p_{\boldsymbol{h}}), \tag{19b}$$

where G is the PGF (1b).

The probability of all unambiguous observations $P(\tilde{\mathscr{O}})$ is derived in Section Prevalence estimates of the Mathematical Appendix and is given by

$$P(\tilde{\mathscr{O}}) = \sum_{\boldsymbol{h} \in \mathscr{H}} \left[\frac{1}{2} \sum_{\boldsymbol{i} \in U_{\boldsymbol{h}}} \big[G(p_{\boldsymbol{h}} + p_{\boldsymbol{i}}) - G(p_{\boldsymbol{i}}) \big] - \Big(\frac{n}{2} - 1 \Big) G(p_{\boldsymbol{h}}) \right]. \tag{20}$$

Hence, the prevalence of $\boldsymbol{h}$ conditioned on ambiguous observations is given by

$$r_{\boldsymbol{h}|\tilde{\mathscr{O}}} = \frac{\sum_{\boldsymbol{i} \in U_{\boldsymbol{h}}} \big[G(p_{\boldsymbol{h}} + p_{\boldsymbol{i}}) - G(p_{\boldsymbol{i}}) \big] - (n-1) G(p_{\boldsymbol{h}})}{\sum_{\boldsymbol{j} \in \mathscr{H}} \left[\frac{1}{2} \sum_{\boldsymbol{i} \in U_{\boldsymbol{j}}} \big[G(p_{\boldsymbol{j}} + p_{\boldsymbol{i}}) - G(p_{\boldsymbol{i}}) \big] - \Big(\frac{n}{2} - 1 \Big) G(p_{\boldsymbol{j}}) \right]}. \tag{21}$$

The conditional prevalence also exceeds the frequencies. Its value, however, seems to be closer to the frequencies $p_{\boldsymbol{h}}$ than that of the unobserved prevalence $q_{\boldsymbol{h}}$. The reason is that a mixed infection with haplotype $\boldsymbol{h}$ is much more likely to be ambiguous than unambiguous. Particularly, the fraction of single infections with haplotype $\boldsymbol{h}$ in unambiguous infections is disproportionately higher than in ambiguous infections. This is more pronounced if the genetic architecture of haplotypes consist of more loci (larger n). While this is true in theory, in real samples, when considering a large number of loci, unambiguous observations are increasingly unlikely (cf. Figures 10–13). The reason is that most haplotypes are not realized in a real population, which can be characterized by the presence of a few haplotypes which differ at multiple loci.

3.1.3. Relative unambiguous prevalence

Due to unobservable information, a statistical model is required to obtain estimates for frequencies. However, in practice, "*ad-hoc*" estimates are popular if statistical methods are not available [cf. (50)]. Frequency estimates can be obtained, by first disregarding all ambiguous observations, calculate the empirically observed unambiguous prevalence of all haplotypes, and finally normalizing them - here we refer to this as the "relative unambiguous prevalence." To assess how accurate such estimates are, this quantity can be expressed in terms of the statistical model introduced here, namely

$$f_{\boldsymbol{h}} := \frac{r_{\boldsymbol{h}|\tilde{\mathcal{O}}}}{\sum\limits_{\boldsymbol{j}\in\mathscr{H}} r_{\boldsymbol{j}|\tilde{\mathcal{O}}}}. \tag{22a}$$

Using (21) this can be rewritten as

$$f_{\boldsymbol{h}} = \frac{\sum\limits_{\boldsymbol{i}\in U_{\boldsymbol{h}}} \big[G(p_{\boldsymbol{h}} + p_{\boldsymbol{i}}) - G(p_{\boldsymbol{i}})\big] - (n-1)G(p_{\boldsymbol{h}})}{\sum\limits_{\boldsymbol{j}\in\mathscr{H}}\left[\sum\limits_{\boldsymbol{i}\in U_{\boldsymbol{j}}} \big(G(p_{\boldsymbol{j}} + p_{\boldsymbol{i}}) - G(p_{\boldsymbol{i}})\big) - (n-1)G(p_{\boldsymbol{j}})\right]}. \tag{22b}$$

Not surprisingly, the relative unambiguous prevalence resembles the frequency $p_{\boldsymbol{h}}$ of haplotpye $\boldsymbol{h}$ better than either the unobserved prevalence $q_{\boldsymbol{h}}$ or the unambiguous prevalence $r_{\boldsymbol{h}|\tilde{\mathcal{O}}}$. However, whether it is larger or smaller than the true frequency, $f_{\boldsymbol{h}}$ depends on the genetic architecture and the MOI distribution. In general there is no clear straightforward pattern, rendering the relative unambiguous prevalence an inadequate proxy for frequencies (cf. Figures 10–13).

3.2. Maximization of the likelihood function with the EM-algorithm

The maximum-likelihood (ML) method is employed here to obtain estimates for haplotype frequencies and the distribution of MOI. The likelihood function (11) derived in Section 2 does not permit a closed solution and has to be maximized numerically. A convenient and efficient method to maximize the likelihood function is the expectation maximization (EM)-algorithm. It is a two-step recursive method to find maximum likelihood estimates (MLEs). The steps of the EM-algorithm are: (i) the expectation (E) step (in which the expectation of the log-likelihood as a function of the unknown parameters conditioned on the parameter choice at the current iteration step is found), and (ii) the maximization (M) step (during which the function obtained at the E step is maximized with respect to the unknown parameters). The parameters obtained during the maximization step are then used in the next expectation step, and the algorithm is repeated until convergence. The algorithm is derived in the Mathematical Appendix in section Deriving the EM-algorithm.

In the present case the EM-algorithm leads to a two-step iterative procedure. The algorithm starts by choosing arbitrary initial values $\lambda^{(0)}$ and $\boldsymbol{p}^{(0)}$ for the Poisson parameter and haplotype frequencies. In step $t+1$ the frequency estimates $\boldsymbol{p}^{(t+1)}$ are derived as

$$p_{\boldsymbol{h}}^{(t+1)} = \frac{C_{\boldsymbol{h}}^{(t)}}{\sum\limits_{\boldsymbol{h}\in\mathscr{H}} C_{\boldsymbol{h}}^{(t)}}, \tag{23a}$$

where

$$C_{\boldsymbol{h}}^{(t)} = p_{\boldsymbol{h}}^{(t)} \sum_{\boldsymbol{x}\in\mathcal{O}} n_{\boldsymbol{x}} \frac{\sum\limits_{\boldsymbol{y}\in\mathscr{A}_{\boldsymbol{x}}} (-1)^{N_{\boldsymbol{x}}-N_{\boldsymbol{y}}} G'_{\lambda_t}\Big(\sum\limits_{\boldsymbol{i}\in A_{\boldsymbol{y}}} p_{\boldsymbol{i}}^{(t)}\Big) I_{A_{\boldsymbol{y}}}(\boldsymbol{h})}{\sum\limits_{\boldsymbol{y}\in\mathscr{A}_{\boldsymbol{x}}} (-1)^{N_{\boldsymbol{x}}-N_{\boldsymbol{y}}} G_{\lambda_t}\Big(\sum\limits_{\boldsymbol{i}\in A_{\boldsymbol{y}}} p_{\boldsymbol{i}}^{(t)}\Big)}, \tag{23b}$$

$$I_{A_{\boldsymbol{y}}}(\boldsymbol{h}) = \begin{cases} 1 & \text{if } \boldsymbol{h} \in Ay, \\ 0 & \text{if } \boldsymbol{h} \notin Ay. \end{cases} \tag{23c}$$

and $G_{\lambda_t}(z)$ is the PGF (1b) with parameter λ_t. The parameter λ_{t+1} is obtained by iterating the equation

$$x_{\tau+1} = x_\tau - \frac{x_\tau - \frac{B_t}{N}(1 - e^{-x_\tau})}{1 + x_\tau - \frac{x_\tau}{1 - e^{-x_\tau}}}, \tag{23d}$$

where

$$B_t = \sum_{\boldsymbol{x}\in\mathcal{O}} n_{\boldsymbol{x}} \frac{\sum\limits_{\boldsymbol{y}\in\mathscr{A}_{\boldsymbol{x}}} (-1)^{N_{\boldsymbol{x}}-N_{\boldsymbol{y}}} \sum\limits_{\boldsymbol{h}\in A_{\boldsymbol{y}}} p_{\boldsymbol{h}}^{(t)} G'_{\lambda_t}\Big(\sum\limits_{\boldsymbol{i}\in A_{\boldsymbol{y}}} p_{\boldsymbol{i}}^{(t)}\Big)}{\sum\limits_{\boldsymbol{y}\in\mathscr{A}_{\boldsymbol{x}}} (-1)^{N_{\boldsymbol{x}}-N_{\boldsymbol{y}}} G_{\lambda_t}\Big(\sum\limits_{\boldsymbol{i}\in A_{\boldsymbol{y}}} p_{\boldsymbol{i}}^{(t)}\Big)}, \tag{23e}$$

starting from $x_0 = \lambda_t$, until numerical convergence is reached. In particular, the iteration stops once $|x_{\tau+1} - x_\tau| < \varepsilon$ holds, by setting $\lambda_{t+1} = x_{\tau+1}$.

The EM-algorithm terminates once numerical convergence is reached. This is defined to be the case if $|\lambda_{t+1} - \lambda_t| + \|p_{\boldsymbol{h}}^{(t+1)} - p_{\boldsymbol{h}}^{(t)}\|_2 < \varepsilon$. The MLE are obtained as

$$\hat{p}_{\boldsymbol{h}} = p_{\boldsymbol{h}}^{(t+1)} \quad \text{and} \quad \hat{\lambda} = \lambda_{t+1}. \tag{24}$$

In practice, the EM-algorithm converges within a few iterations. Notably, it can be implemented efficiently. Because of the efficient implementation, bootstrap confidence intervals (CIs) can be readily obtained, as described in Confidence intervals.

An implementation of the EM-algorithm and the bootstrap CIs in R is available as Supplementary material. The code is also available at: https://github.com/Maths-against-Malaria/MultiLociBiallelicModel.git.

3.3. Using a plug-in estimate for the Poisson parameter

Estimates of MOI might be unreliable in the case of unbalanced haplotype frequency distributions, as they typically occur for drug-resistance associated haplotypes in malaria. To compensate for this, the number of loci considered can be increased. In the case of resistance-associated haplotypes a typical choice would be a set of unlinked neutral marker loci, which are likely to have balanced frequencies.

However, adding additional loci can lead to three problems in practice. First, due to the curse of dimensionality, the number of parameters to be estimated becomes so large that haplotypes are characterized at the individual level. This can be compensated by marginalizing the frequency estimates with regard to the set of loci of interest. Second, poor data quality can lead to missing data entries, so that the number of samples, which have information at all loci (the original and the additional set), is substantially smaller than if the sets of loci would be considered separately. Similarly, one set of loci might just have been process for a sub-sample. Third, molecular information might, for both sets of loci, have been performed for different sets of samples. In any of these cases, one could estimate the Poisson parameter λ based only on the additional set of loci and use this estimate as a plug-in estimate to obtain the haplotype frequency distribution for the original set of loci.

If one prefers to use a plug-in estimate for the Poisson parameter λ, the EM-algorithm can be adapted. This adaptation is derived in The EM-algorithm using a plug-in estimate for the Poisson parameter. This algorithm is also implemented as an R script (see the available User manual in the Supplementary material).

Notably, it is advisable to use the whole available data, rather than plug-in estimates, unless one of the situations outlined above applies. The reason is that as a general guideline, estimates should be based if possible on the full information being available.

3.4. Estimating samplewise MOI

Once the population-level MOI parameter λ and the haplotype frequencies $\boldsymbol{p}$ have been estimated as $\hat{\lambda}$ and $\hat{\boldsymbol{p}}$, these can be used as plug-in estimates to infer the true MOI m for a sample $\boldsymbol{x}$. In line with maximum-likelihood estimation, a natural estimate $\hat{m}$ is the value of m which maximizes that has the highest probability given the observation $\boldsymbol{x}$ and the plug-in estimates $\hat{\lambda}$ and $\hat{\boldsymbol{p}}$. More precisely, the samplewise estimate $\hat{m}$ of MOI is

$$\hat{m} = \arg\max_{m} P(\mathrm{MOI} = m|\boldsymbol{x};\, \hat{\lambda}, \hat{\boldsymbol{p}}), \tag{25}$$

where $P(\mathrm{MOI} = m|\boldsymbol{x}; \hat{\lambda}, \hat{\boldsymbol{p}})$ is defined in Samplewise MOI of the Mathematical Appendix. The samplewise MOI estimates are implemented in the R script available at https://github.com/Maths-against-Malaria/MultiLociBiallelicModel.git. A numerical example is found in the User manual.

3.5. Data application

As an application, we estimated MOI and haplotype frequencies of malaria parasites associated with resistance to sulfadoxine-pyrimethamine (SP) in Cameroon in 2001/2002 and 2004/2005. MOI was estimated to be intermediate at both time points, and slightly decreased in 2004/2005. The estimated MOI parameters were $\hat{\lambda} = 0.9397$ (95% CI: 0.7492, 1.1551) for 2001/2002 and $\hat{\lambda} = 0.8645$ (95% CI: 0.6496, 1.0928) for 2004/2005. This slight decrease is in accordance with the downward trend in the number of reported cases in Cameroon from 1992 to 2005, sustained by programs like the "Roll Back Malaria" program [cf. (51)].

The estimates of haplotype frequencies are presented in Table 3 (confidence intervals for the frequencies estimates are omitted to improve readability). The drug sensitive wildtype and those with single mutations decreased in frequency between the two time points, whereas strongly resistant haplotypes with triple mutations on *dhfr* and double mutations at *dhps* increased in frequency. This is not surprising considering that SP drug pressure was high during that time. Namely, chloroquine was officially removed as first line therapy in Cameroon in 2002, whereas amodiaquine and SP became first- and second-line treatments [cf. (42)]. Although, the combination of artesunate and amodiaquine became the official therapy for uncomplicated *P. falciparum* malaria in 2004, it was not widely used until 2007 [cf. (42)]. In particular, the frequency of the highly resistant haplotypes 51**I**/59**R**/108**N**/I164—S436/437**G**/K540/A581/A613 increased from 38 (95% CI: 31.27, 44.37%) to 46% (95% CI: 39.28, 53.44%) (see Table 3), whereas the less resistant haplotypes characterized by just two mutations in *dhfr* decreased in frequency.

Although the *ad-hoc* frequency estimates given by the relative unambiguous prevalence (see above) are close to the MLE, some frequency estimates differ substantially. For instance, the MLEs for the frequency of the highly resistant haplotype 51**I**/59**R**/108**N**/I164—436**A**/A437/K540/A581/A613 at both time points are, respectively, 24.5% (95% CI: 18.98, 30.41%) and 25.2% (95% CI: 19.53, 30.93%), the corresponding *ad-hoc* estimates are 19.7 and 20.9% (see Table 3). This is not surprising because, this haplotype is likely to occur in mixed infections with the predominant haplotype, which would be disregarded by the *ad-hoc* estimates.

Obvious is the difference between frequency and prevalence estimates. The frequencies of the highly resistant haplotype 51**I**/59**R**/108**N**/I164—S436/437**G**/K540/A581/A613 at both

TABLE 3 Frequencies estimates of SP-resistant haplotypes from the Cameroonian data.

Haplotype		**Frequency** p		**Prevalence** p		**Cond. prevalence** p		***Ad-hoc* freq.** p	
		Years		**Years**		**Years**		**Years**	
dhfr	***dhps***	**01/02**	**04/05**	**01/02**	**04/05**	**01/02**	**04/05**	**01/02**	**04/05**
NCSI	SAKAA	0.0190	0.00468	0.0290	0.0070	0.0168	0.0037	0.0204	0.0063
NCSI	**A**AKAA	0.0604	0.00548	0.0905	0.0082	0.0521	0.0043	0.0476	0.0063
NCSI	S**G**KAA	0.0178	0.00965	0.0272	0.0144	0.0151	0.0075	0.0136	0.0063
IC**N**I	S**G**KAA	0.0164	$< 10^{-12}$	0.0250	$< 10^{-12}$	0.0194	$< 10^{-12}$	0.0272	−−
N**RN**I	SAKAA	0.0095	0.01025	0.0145	0.0152	0.0084	0.0084	0.0136	0.0127
N**RN**I	**A**AKAA	0.0122	$< 10^{-12}$	0.0188	$< 10^{-12}$	0.0130	$< 10^{-12}$	0.0136	−−
N**RN**I	S**G**KAA	0.0475	0.02179	0.0717	0.0322	0.0583	0.0257	0.0612	0.0253
N**RN**I	**AG**KAA	0.0134	0.00935	0.0205	0.0139	0.0124	0.0082	0.0204	0.0127
IRNI	SAKAA	0.0271	0.04085	0.0413	0.0600	0.0387	0.0564	0.0340	0.0506
IRNI	**A**AKAA	0.2447	0.25216	0.3372	0.3384	0.2493	0.2515	0.1973	0.2089
IRNI	S**G**KAA	0.3792	0.46288	0.4921	0.5698	0.4325	0.5202	0.3741	0.4494
IRNI	SAKA**T**	0.0265	0.00776	0.0404	0.0116	0.0230	0.0063	0.0204	0.0063
IRNI	**AG**KAA	0.0606	0.12062	0.0909	0.1711	0.0867	0.1714	0.0884	0.1519
IRNI	S**G**KA**T**	0.0165	$< 10^{-12}$	0.0252	$< 10^{-12}$	0.0198	$< 10^{-12}$	0.0204	−−

The first column shows the amino acid sequence at codons 51, 59, 108, and 164 at the *dhfr* locus and the second column shows the amino acid sequence at codons 436, 437, 540, 581, and 613 at the *dhps* locus. A total of 50 and 36 haplotypes were estimated to have strictly positive frequency in 2001/2002 and 2004/2005, respectively. Only haplotypes with an estimated frequency >0.01 in 2001/2002 or 2004/2005 are reported. The remaining columns show the MLE of the respective frequencies, the estimated prevalence, the estimated conditional prevalence, and the *ad-hoc* estimate (relative unambiguous prevalence) for the haplotype frequencies.
Cond., Conditional; Freq., Frequency.

time points—38 and 46%—are substantially lower than the prevalences, estimated to be 49 and 57%, respectively (see Table 3). However, the estimates for the conditional prevalence (which are not recommendable) are only slightly smaller than the prevalence estimates and amount to 43 and 52%. Since MOI is intermediate, the discrepancy is not expected to be too large.

The same is true for the other highly resistant haplotype 51**I**/59 **R**/108**N**/I164—436**A**/A437/K540/A581/ A613. The frequency estimates are 24.5% (95% CI: 18.59, 30.07%) and 25.2% (95% CI: 18.84, 30.95%), whereas the prevalence estimates are 33.7 and 33.8%. Although MOI is intermediate, the discrepancy between the prevalence and conditional prevalence estimates (24.9 and 25.2% at the two time points) are quite large (cf. Table 3).

3.6. Performance of the estimator

Ideally an estimator is (i) unbiased, i.e., it is accurate and (ii) precise, i.e., it has low variance. The minimal variance of an unbiased estimator is given by the Cramér-Rao lower bound. Typically, MLEs have good asymptotic properties. They are asymptotically unbiased and efficient (they asymptotically attain the minimal possible variance). Despite these desirable asymptotic properties, the quality of MLEs has to be investigated under finite sample sizes. If bias is small and the variance of the estimator is close to the Cramér-Rao lower bound, one has confidence that the estimator is "optimal." The quality of an estimator is measured by the mean squared error (MSE)

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + (\mathrm{Bias}(\hat{\theta}))^2, \tag{26}$$

which is the sum of the estimators variance and squared bias.

Since, the MLE here has no closed solution, bias and variance cannot be studied analytically. We therefore perform a numerical simulation to study the estimator's bias and variance.

3.6.1. Bias of the estimator

To compare bias across a range of different parameter values, we consider a 'dimensionless' quantity, namely the relative bias (14a) in percent, i.e., the bias of the estimator in percent of the true value of the parameter.

3.6.1.1. Relative bias for haplotypes frequencies estimates

The estimator is typically unbiased, with no noticeable effect of sample size if the true haplotype distribution is symmetric independently of the number of considered loci n, i.e., all haplotypes are equally abundant in the pathogen population (see Figure 2). This is intuitively expected, because the haplotypes are interchangeable in this case. Deviations from the true frequency distribution occur only due to random sampling. Although

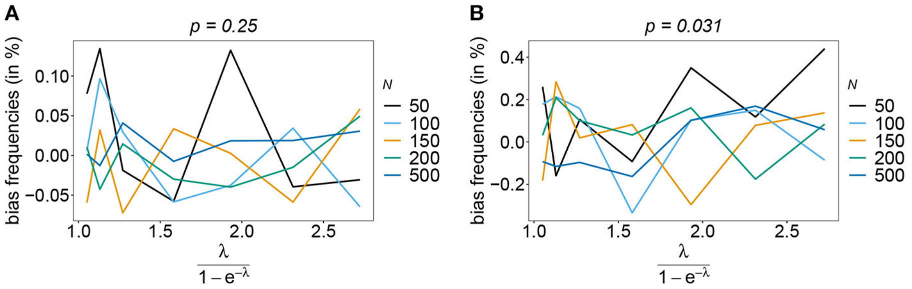

FIGURE 2
Bias of frequencies estimates—the symmetric case. Shown is the bias of the frequency estimates in % as a function of the mean MOI (i.e., for a range of Poisson parameters). The symmetric haplotype frequency distributions (cf. Table 1) for $n = 2$ **(A)** and $n = 5$ **(B)** are assumed. In both panels, only the bias for the first haplotype is shown (in the symmetric case, all haplotypes are equivalent and bias looks similarly). Colors correspond to different sample sizes.

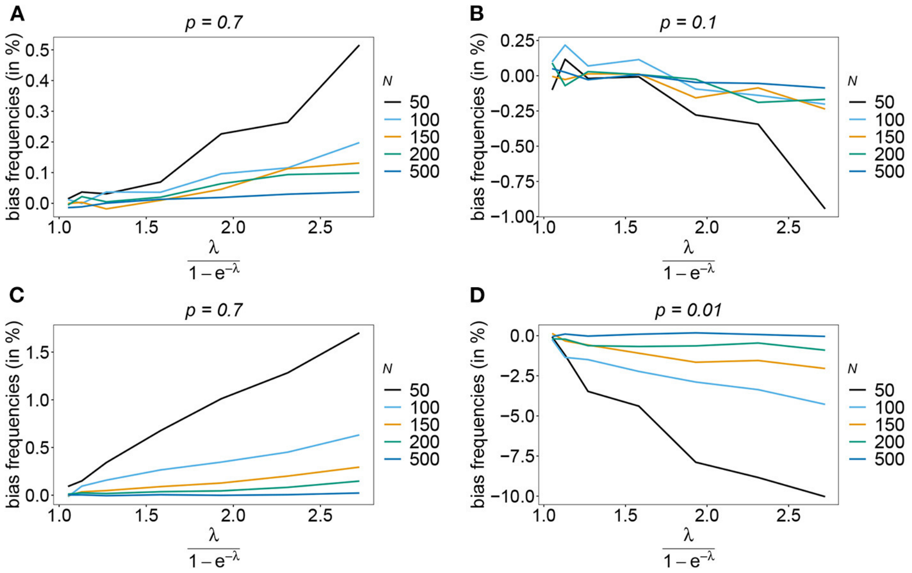

FIGURE 3
Bias of frequencies estimates—the unbalanced case. Shown is the bias of the frequency estimates in % as a function of the mean MOI (i.e., for a range of Poisson parameters). The skewed haplotype frequency distributions (cf. Table 1) for $n = 2$ **(A,B)** and $n = 5$ **(C,D)** are assumed. In both cases only the bias for the predominant haplotype and one underrepresented haplotype are shown (all underrepresented haplotypes are equivalent and bias looks similarly). Colors correspond to different sample sizes.

random effects are more pronounced for small sample size and larger n (also seen from the variation in Figure 2), in terms of bias this effect averages out. However, it will affect the MLE's variance (see below).

Also, if the underlying haplotype frequency distribution is skewed, the estimator has low bias (see Figure 3). Bias (in relative terms) is highest for haplotypes with low frequencies. These tend to be underrepresented in datasets. On the contrary, the frequencies of predominant haplotypes will be overestimated, as these tend to be over-represented in datasets. This is particularly true for high MOI ($\psi > 1.8$) and small sample size. Bias vanishes with increasing sample size. The estimates can be considered almost unbiased for $N \geq 150$. Note that the bias of rare haplotypes will only be large in relative terms, not in absolute terms. In practice, large sample size is required to detect rare haplotypes. Bias tends to be larger, if a larger number of loci is considered (compare Figures 3A,B with Figures 3C,D). The reason is that the number of possible haplotypes is increasing geometrically with larger n and in the numerical examples a large number of haplotypes with low frequencies are assumed.

The number of possible haplotypes increases geometrically with the number of considered loci. For $n = 10$, 1,024 possible haplotypes exist. In practice only a fraction of the possible haplotypes circulate in the population. Figures 4, 5

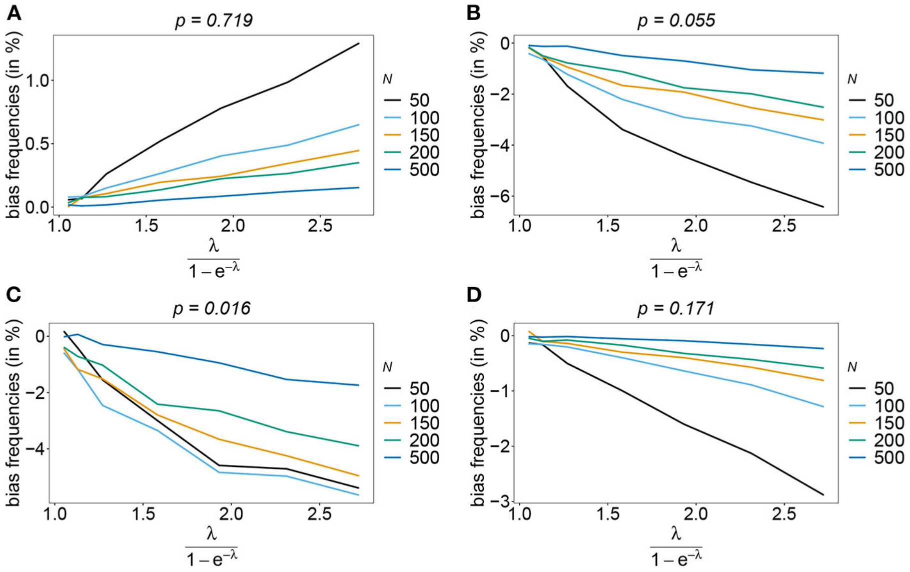

FIGURE 4
Bias of frequencies estimates—Kenya data 2005. Shown is the bias of the frequency estimates in % as a function of the mean MOI (i.e., for a range of Poisson parameters). The haplotype frequency distributions (cf. Table 2) for $n = 10$ are assumed. The bias for the predominant haplotype and few underrepresented haplotype are shown **(A–D)**, the corresponding haplotype frequencies are shown at the top of the panels. Colors correspond to different sample sizes.

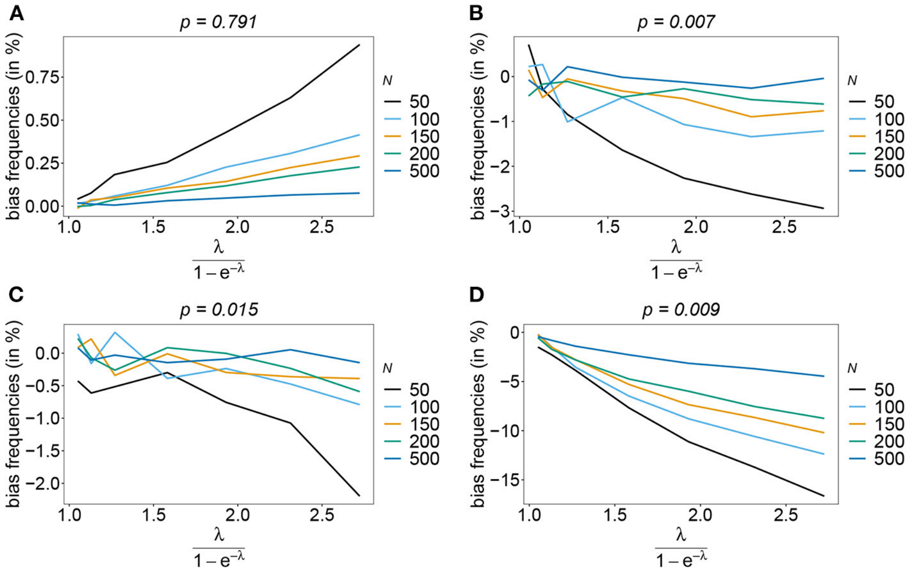

FIGURE 5
Bias of frequencies estimates—Kenya data 2010. Shown is the bias of the frequency estimates in % as a function of the mean MOI (i.e., for a range of Poisson parameters). The haplotype frequency distributions (cf. Table 2) for $n = 10$ are assumed. The bias for the predominant haplotype and few underrepresented haplotype are shown **(A–D)**, the corresponding haplotype frequencies are shown at the top of the panels. Colors correspond to different sample sizes.

show the bias of haplotype frequencies assuming the frequency distributions of malaria haplotypes estimated in Kenya. These are two rather unbalanced frequency distributions. Also in these cases, the frequency estimates have little bias that vanishes with increasing sample size. Again the frequencies of predominant haplotypes tend to be overestimated, while those of rare haplotypes tend to be underestimated. Bias increases with increasing MOI. The reason is that super-infections are common, and rare haplotypes will be unlikely to occur in single infections. However, they will likely occur together in mixed

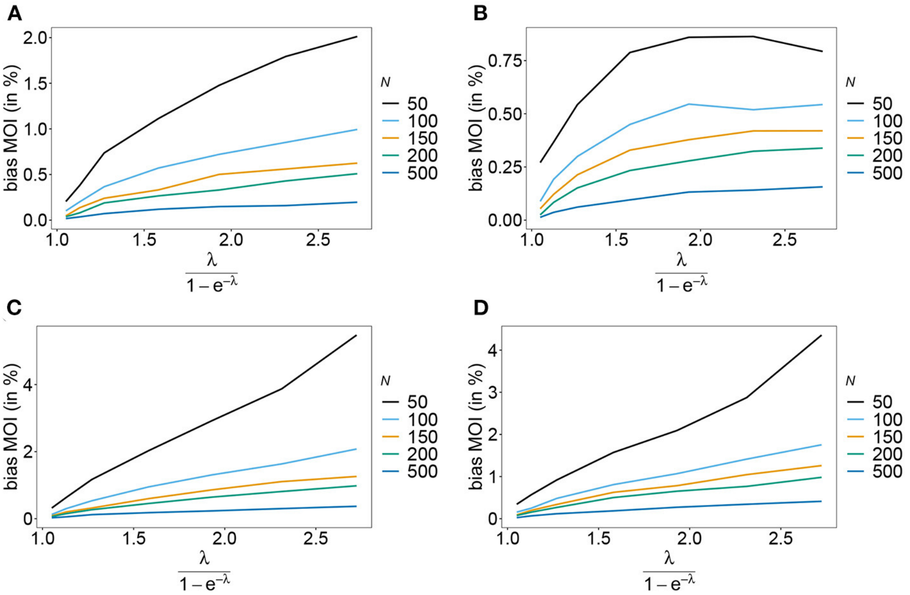

FIGURE 6
Bias of MOI estimates. Shown is the bias of the mean MOI estimates ψ in % as a function of the true mean MOI (i.e., for a range of Poisson parameters). Symmetric haplotype frequency distributions (Table 1) are assumed for $n = 2$ **(A)** and $n = 5$ **(B)**, whereas skewed distributions are assumed in **(C,D)**, for $n = 2$ and $n = 5$, respectively. Colors correspond to different sample sizes.

infections with predominant types, resulting in ambiguous information. As a consequence, the estimator yields positive frequency estimates for haplotypes that are not circulating in the population, and thereby understimates those rare haplotype that are actually present. (Note that in practice, due to ambiguity, it is typically impossible to determine which haplotypes are actually circulating in the population.)

3.6.1.2. Relative bias for MOI parameter estimates

Rather than evaluating the bias of the MOI parameter λ, bias is evaluated in term of the empirically more relevant mean MOI ψ. For a given n, the estimator has relatively little bias irrespective of the frequency distribution and true value of λ. In general, the estimator overestimates the true parameter. The reason is that λ is positive, and can be overestimated but not underestimated by arbitrary amounts. In general MLEs are sensitive to outliers. Here, particularly for large λ, rare over-representations of multiple infections in the data, lead to substantial overestimates. This is more likely to occur for small sample sizes and large n. Consequently, bias is increasing as a function of λ and decreasing as a function of sample size N. Typically, bias is decreasing for larger n, because the amount of information contained in a dataset increases. However, also sample size has to be adequate for larger n to appropriately represent the haplotype distribution. Not surprisingly, bias is higher for more skewed frequency distributions. This is because, rare haplotypes will be underrepresented and single infections with rare haplotypes are unlikely to be observed in a dataset, particularly for large ψ (see Figures 6, 7). In general, bias is small for samples of size $N \geq 150$.

3.6.2. Variance of the estimator

The estimator's variance was assessed in terms of the coefficient of variation (14b). This is a dimensionless quantity, which allows comparisons across a range of parameter values.

As expected the estimator's variance decreases with increasing sample size N, because datasets reflect the underlying population more accurately, leading to less fluctuations between different realizations of datasets. The variance of the MOI estimator is relatively small (see Figures 8, 9). Not surprisingly, it increases with n (compare Figures 8A,C with Figures 8B,D and see Figure 9). The number of haplotypes increases with n, and hence, less information is available for each single haplotype. Thus, for larger n, due to the curse of dimensionality, the underlying population is less adequately represented in a dataset of given sample size N.

The variance of the frequency estimates has similar properties (not shown).

3.6.3. Prevalence and relative prevalence

Sometimes the prevalence of haplotypes is empirically more important than their frequencies. For example, in the case of drug resistance, clinically it is more relevant to assess the probability that a patient is infected with a resistant haplotype,

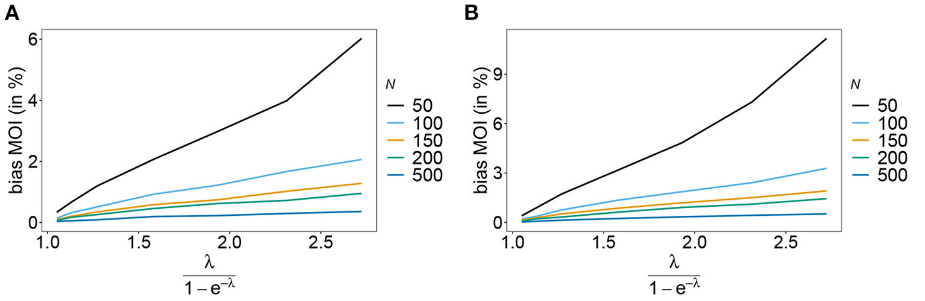

FIGURE 7
Bias of MOI estimates. As in Figure 6 but for $n = 10$ for the haplotype estimated for antimalarial drug resistance in Kenya (see Table 2) in 2005 **(A)** and 2010 **(B)**.

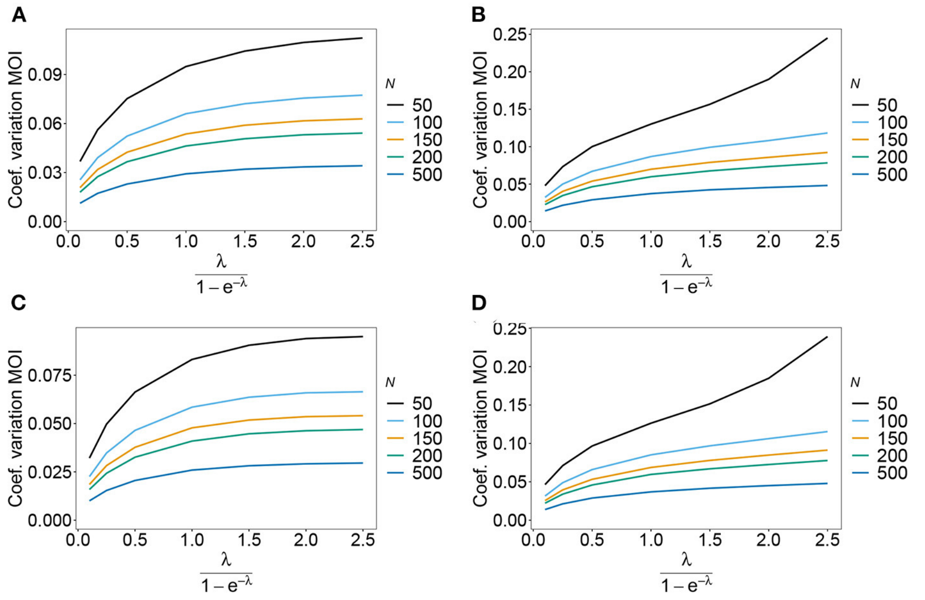

FIGURE 8
Variance of MOI estimates. Shown is the variance of the mean MOI estimates ψ in % as a function of the true mean MOI (i.e., for a range of Poisson parameters). The symmetric haplotype frequency distribution is assumed respectively for $n = 2$ and $n = 5$ **(A,B)** as well as the skewed haplotype frequency distributions **(C,D)** (cf. Table 1). Colors correspond to different sample sizes.

rather than its frequency in the parasite population. Due to multiple infections, the absence and presence of haplotypes in infections is in general ambiguous. Hence, prevalence is an unobservable quantity (unobservable prevalence). However, estimates for (unobservable) prevalence are readily derived as plug-in estimates from the MLE and (16c). In fact, these estimates are very accurate (Figures 10–13), except if haplotypes are rare and sample size is small, in which case the prevalence is underestimated (cf. Figures 12B, 13B).

In practice, prevalence is often estimated as the conditional prevalence (21). This quantity substantially underestimates the (unobservable) prevalence (Figures 10–13). Hence, it should not be used as a proxy for prevalence. However, although the conditional prevalence is not recommendable, it can be accurately recaptured from the MLE and (21).

In the absence of a statistical model, haplotype frequencies can be estimated by normalizing the conditional prevalences of the haplotypes (relative unambiguous prevalence) (22b). These are undesirable *ad-hoc* estimates, because they are biased (Figures 10–13). Particularly unfortunate is that bias depends on MOI. The larger the mean MOI, the larger the bias. However, it is rather unpredictable whether a particular haplotype's frequency is over- or underestimated (Figures 10–13). This depends on the true haplotype frequencies and MOI. The relative bias can be accurately obtained as a plug-in estimate from the MLE and (22b).

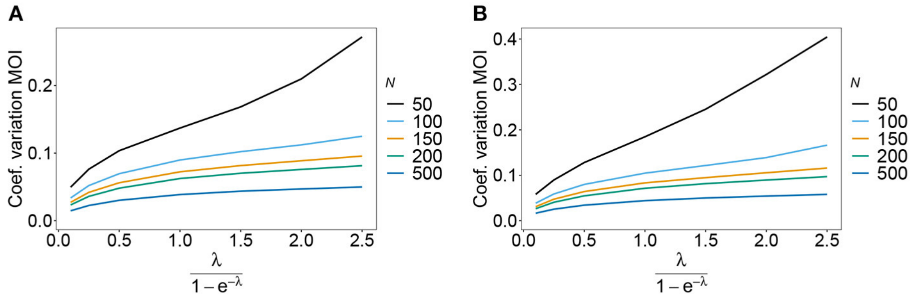

FIGURE 9
Variance of MOI estimates. Shown is the variance of the mean MOI estimates ψ in % as a function of the true mean MOI (i.e., for a range of Poisson parameters). The haplotype frequency distributions for $n = 10$ are assumed, respectively, for the year 2005 **(A)** and 2010 **(B)** (cf. Table 2). Colors correspond to different sample sizes.

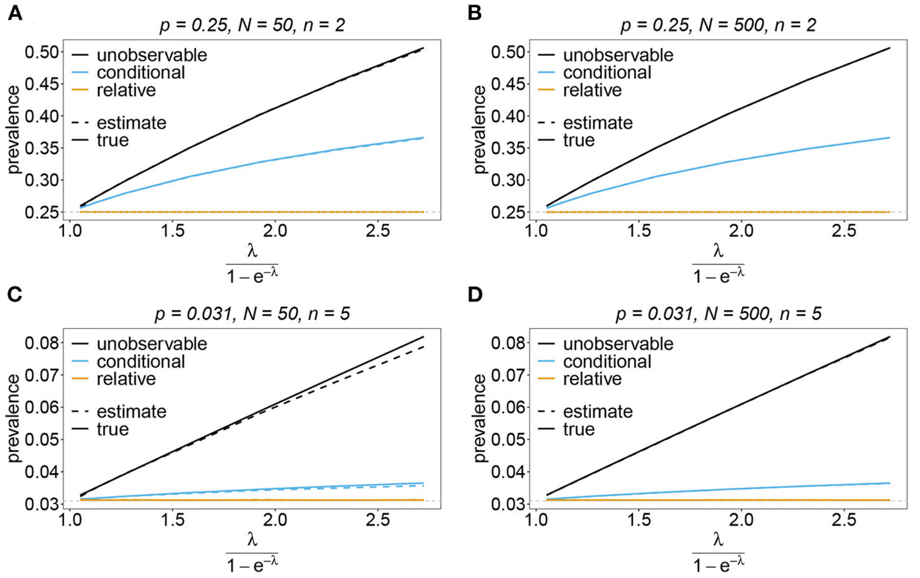

FIGURE 10
Prevalence estimates—the symmetric case. Shown is the prevalence of the haplotypes as a function of the mean MOI (i.e., for a range of Poisson parameters). The symmetric haplotype frequency distributions (cf. Table 1) for $n = 2$ **(A,B)** and $n = 5$ **(C,D)** are assumed. In both cases of n only the prevalence estimates for the first haplotype are shown for a small ($N = 50$) and big ($N = 500$) sample size (in the symmetric case all haplotypes are equivalent and prevalence looks similarly). Colors correspond to different prevalence models. The solid line show the true prevalence and the dashed line the estimates.

4. Discussion

Public health strategies are increasingly relying on genomic/molecular surveillance to monitor infectious diseases [cf. (2, 3)]. This is particularly true for malaria, where molecular surveillance is a standard approach to monitor pathogen variants, which are associated with drug resistance, or jeopardize reliable diagnostics (e.g., *P. falciparum* variants with deletions in the HRP2/3 genes, which can lead to false-negative rapid diagnostic test results [cf. (52)]. Moreover, patterns of transmission, disease exposure, or the evolutionary genetics can be ascertained by studying genomic/molecular data [cf. (53–55)].

A usual problem in molecular surveillance in the context of malaria is the presence of several genetically distinct parasite haplotypes within an infection—this is particularly common in areas of high transmission [cf. (41, 54)]. Unfortunately, in such cases, usual molecular methods provide only ambiguous information concerning the haplotypes present in infections [cf. (15)]. Namely, molecular information is typically unphased. In the case of, e.g., antimalarial drug resistance, precise estimates of the frequency of particular haplotypes and the likelihood that they are observed in an infection (prevalence) are required. This requires sophisticated statistical models that resolve the underlying ambiguity in the observations.

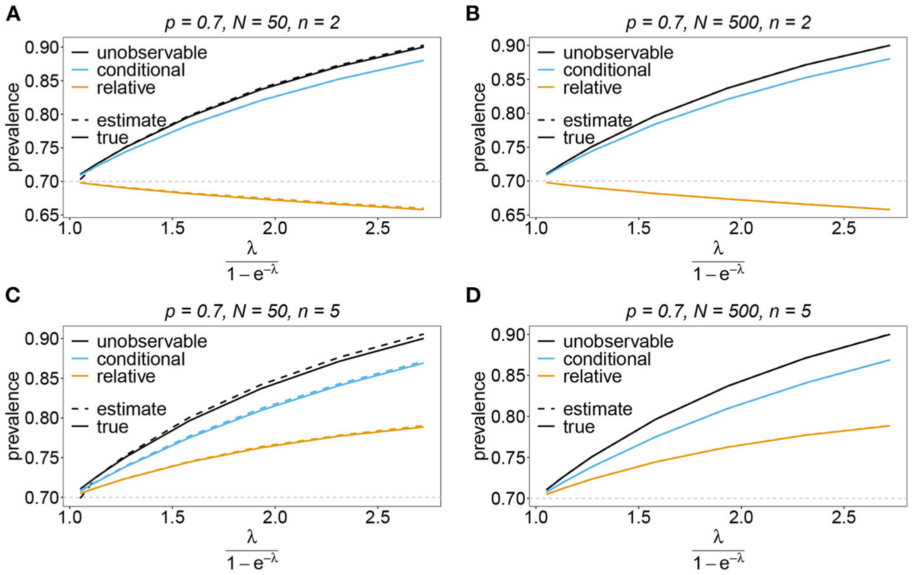

FIGURE 11
Prevalence estimates—the unbalanced case. Shown is the prevalence of the predominant haplotypes as a function of the mean MOI (i.e., for a range of Poisson parameters). The skewed haplotype frequency distributions (cf. Table 1) for $n = 2$ **(A,B)** and $n = 5$ **(C,D)** are assumed. In both cases of n only the prevalence estimates are shown for a small ($N = 50$) and big ($N = 500$) sample size. Colors correspond to different prevalence models. The solid lines show the true prevalence and the dashed lines the estimates.

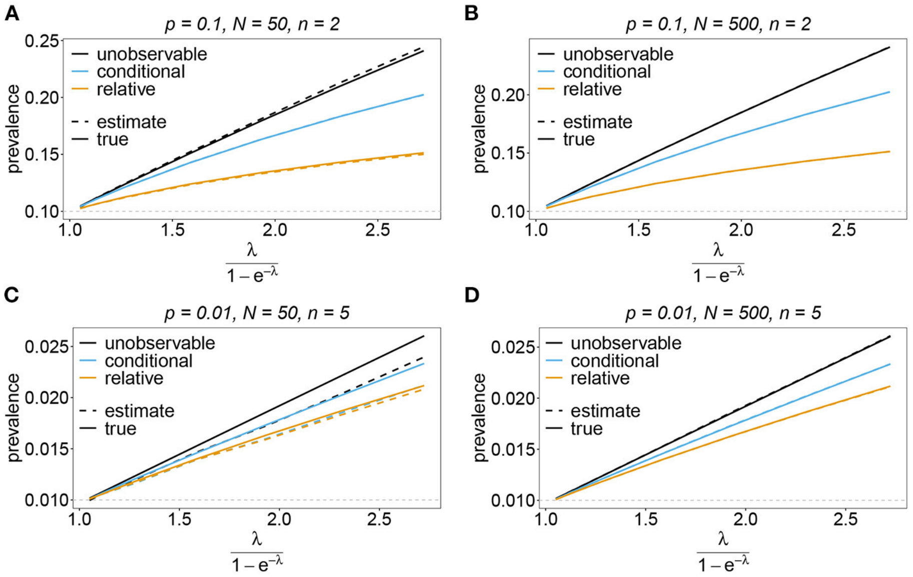

FIGURE 12
Prevalence estimates—the unbalanced case: Shown is the prevalence of the underrepresented haplotypes as a function of the mean MOI (i.e., for a range of Poisson parameters). The skewed haplotype frequency distributions (cf. Table 1) for $n = 2$ **(A,B)** and $n = 5$ **(C,D)** are assumed. In both cases of n only the prevalence of one of the underrepresented haplotypes estimates are shown for a small ($N = 50$) and big ($N = 500$) sample size (all underrepresented haplotypes are equivalent and prevalence looks similarly). Colors correspond to different prevalence models. The solid lines show the true prevalence and the dashed lines the estimates.

Here, we introduced a statistical model to estimate the frequencies and prevalence of pathogen haplotypes from molecular data. More precisely, hosts can be super-infected several times with the same or different pathogenic variants. The number of super-infections is referred to as multiplicity of infection (MOI). Concerning the genetic architecture

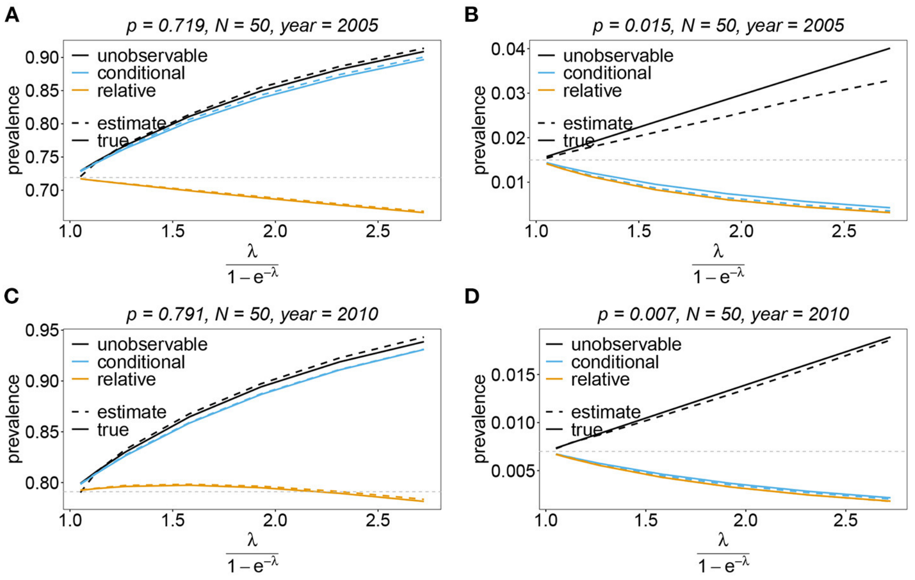

FIGURE 13
Prevalence estimates—the unbalanced case: Shown is the prevalence of the dominant and one underrepresented haplotype for the years 2005 **(A,B)** and 2010 **(C,D)** as a function of the mean MOI (i.e., for a range of Poisson parameters). The haplotype frequency distributions (cf. Table 2) for $n = 10$ are assumed. The prevalence estimates are shown for $N = 50$. Colors correspond to different prevalence models. The solid lines show the true prevalence and the dashed lines the estimates.

of pathogen variants, we assumed the pathogen to be haploid, and haplotypes to be determined by *n* biallelic loci. Typical applications are malaria parasites associated with drug resistance, where loci correspond to specific codons in one or more genes, e.g., codons 51, 59, 108, 164 in the *dhfr* locus and codons 436, 437, 540, 581, 613 in the *dhps* locus of *P. falciparum* associated with resistance to sulfadoxine-pyrimethamine (SP). The method is intended to derive frequencies of haplotypes which are "aggregates," e.g., certain drug resistant haplotypes. It is not intended to characterize parasites at the individual level. Hence, the number of loci (n) should not be so large, that each haplotype occurs only once in the overall sample. In our simulations, we used up to $n = 10$. However, the method is not limited to 10 loci. Importantly, n is the number of loci, which are found polymporphic in the data, as monomorphic loci can be dropped.

We suggested a maximum-likelihood estimation of haplotype frequencies and the distribution of MOI, assuming an underlying conditional Poisson distribution. As in (28), we employ the expectation-maximization (EM) algorithm to derive the maximum-likelihood estimate (MLE). However, Li et al. (28) provided only a general form, which can be derived numerically by brute force. We derived a more explicit version that allows an efficient implementation, which is provided as an easy-to-use R script. Importantly, based on the statistical framework, we provided explicit expressions for prevalence.

Based on this, the MLE can be used as a plug-in statistic, to provide estimates for prevalence. For instance, in the context of antimalarial drug resistance, the prevalence of drug-resistant haplotypes is a more relevant quantity regarding disease outcomes. The frequency of a haplotype is its relative abundance in the pathogen population, whereas prevalence is the probability that the haplotype occurs in an infection. Estimating prevalence is notoriously difficult. If transmission is low, i.e., the average MOI is small, prevalence and frequency almost coincide. However, if transmission is high (large average MOI), prevalence can be substantially higher than frequency. If MOI is high, several pathogen haplotypes commonly occur in infections. However, if molecular methods provide only unphased information, they do not allow to directly observe haplotypes in such infections [cf. (16)]. Therefore, a statistical model is required to resolve this ambiguity. *Ad-hoc* methods to estimate prevalence do not require an explicit statistical model. However, they must necessarily be based only on unambiguous information. We investigated the deviations of conditional prevalence (conditioned on unambiguous information) and prevalence and concluded that *ad-hoc* approximations to prevalence might substantially underestimate the true prevalence. Indeed it might be highly problematic for treatment policies if prevalence of drug-resistant pathogen variants are underestimated.

Here, we applied the method to an empirical dataset of mutations associated with SP resistance from Cameroon in the years 2001/2002 and 2004/2005. Furthermore, we investigated the performance of the MLE in terms of bias and variance. In general, the estimates of the haplotype frequencies have little bias. Bias tends to be higher for higher average MOI. Moreover, in relative terms, bias is higher for rare haplotypes. The MOI parameter has a higher bias than the frequency estimates, particularly if average MOI is high. However, bias decreases quickly with sample size. Also, the variance of the estimator, in terms of the coefficient of variation tends to be small. Due to the good performance of the estimator, also the estimates of prevalence are reliable.

Although the method performs well, it has certain limitations. So far it (i) is restricted to SNP data (or biallelic data); (ii) does not account for missing information; (iii) does not incorporate errors in the data (due to the molecular methods used to generate the data); (iv) does not take relatedness between pathogen variants and co-transmissions into account. The first three limitations are however justified by the curse of dimensionality. Assuming n loci with three instead of two alleles would result in 3^n rather than 2^n possible haplotypes. For $n = 5$ loci this amounts to 243 rather than 32 haplotypes. Importantly, the biallelic genetic architecture is justified by the popularity of SNP data. In the case of malaria, sample sizes of $N = 50$ to $N = 500$ are realistic. With $n = 10$ loci, 1,024 haplotypes are possible. Hence, the number of parameters would by far exceed the sample size. Importantly, in practice, not all of the 1,024 possible haplotypes are compatible with the data. Hence, only a subset of haplotypes is relevant, rendering a realistic sample size to be adequate. However, when aiming to incorporate missing information and errors, all possible haplotypes have to be considered in a statistical model, with the majority of them being irrelevant. An exact statistical model will be hopelessly over-parameterized if the number of considered loci is large. Therefore, approximate models would need to be considered, which disregard infrequent haplotypes. The fourth limitation deserves particular attention. With genomic data becoming more available in molecular surveillance, methods have been developed to account for relatedness of pathogen variants within infections [cf. (35–40)]. Genetic relatedness is informative on transmission dynamics, i.e., knowledge of whether pathogen variants co-occurring in a single infection are identical by state or identical by descent, or whether they were co-transmitted together rather than sequentially is informative on possible routes of transmission (35, 40). Such methods however require genomic information or at least larger SNP barcodes. Although genomic sequencing is becoming more affordable, obtaining such data requires efforts and resources, which are still not feasible in many settings. Note that in vector-borne diseases like malaria, genetic relatedness is only partially informative on transmission dynamics. Namely, it is unclear whether relatedness is caused by some mosquitoes infecting many hosts, or whether mosquitoes get infected by many hosts. Particularly, the admixture of the vector and host populations influence relatedness of the pathogen within the hosts and vectors. To an extreme, if the disease is transmitted mainly within households, i.e., the vectors are not well mixed with the host population, high relatedness of pathogen variants within infections is expected, independently of the overall transmission intensity. On the other hand, if infections within households are uncommon, less relatedness is expected. Hence, relatedness might be more informative on the routes of transmission than on transmission intensities themselves. In the latter case, the proposed approach to estimate MOI seems preferable. Some methods also estimate the relative abundance of haplotypes in an infection (40). Such information is important if there is evidence that the pathogenesis of the disease depends on the interactions of pathogen variants within the infection—especially, if the emphasis is on the clinical manifestation of the disease rather than on the pathogen population level. Notably, all methods that consider relatedness have their limitations. Since (due to the curse of dimensionality) they are not haplotype based, typically independence of genetic markers is assumed. For applications such as drug resistance in malaria, this assumption is not justified, such that a haplotype based approach as proposed here seems more appropriate. In vector-borne diseases, one of the main advantage of the proposed method, is that it does not require an explicit model of vector-host dynamics. Incorporating relatedness, tailored to the characteristics of the disease, would require imposing such model to be accurate. This, however, would require several assumptions, which might yet be poorly justified by empirical evidence. In case co-transmission of pathogen variants are important, it would be interesting to ascertain how well the proposed method performs. However, such assessment is notoriously difficult, because a true model for transmission needs to be specified based on empirical evidence.

In conclusion, we provided a method to estimate haplotype frequencies and prevalences alongside the distribution of MOI from malaria genetic data. The estimator shows convenient statistical properties and can be efficiently implemented. The estimator is implemented in an easy-to-use R script available on Github at https://github.com/Maths-against-Malaria/MultiLociBiallelicModel.git.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author/s.

Author contributions

HT contributed in the study design, carried out the mathematical analysis, numerical implementation, data analysis, performed numerical simulations, created the graphics, and wrote the manuscript. KS contributed in designing the study, the mathematical analysis, supervised the numerical implementation, helped to design the numerical simulations, and participated in writing and correcting the manuscript.

Both authors contributed to the article and approved the submitted version.

Funding

This work was supported by grants of the German Academic Exchange (DAAD; https://www.daad.de/de/; Project-ID 57417782, Project-ID: 57599539), the Sächsisches Staatsministerium für Wissenschaft, Kultur und Tourismus and Sächsische Aufbaubank – Förderbank (SMWK-SAB; https://www.smwk.sachsen.de/; https://www.sab.sachsen.de/; project Innovationsvorhaben zur Profilschärfung an Hochschulen für angewandte Wissenschaften, Project-ID 100257255; project Innovationsvorhaben zur Profilschärfung 2022, Project-ID: 100613388), the Federal Ministry of Education and Research (BMBF) and the DLR (Project-ID 01DQ20002; https://www.bmbf.de/; https://www.dlr.de/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments

The authors are grateful to Dr. Andrea M. McCollum and Dr. Venkatachalam Udhayakumar for sharing the molecular dataset on SP resistance from Cameroon. The helpful comments and corrections of Meraj Hashemi, Pierre Marie Ngougoue Ngougoue, Nessma Adil Mahmoud Yousif, and Arlinda Sadiku are gratefully acknowledged. The manuscript was improved with the help of constructive comments from three reviewers and the editors, to whom the authors are grateful.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fepid.2022.943625/full#supplementary-material

References

1. Horstmann DM. Importance of disease surveillance. *Prevent Med.* (1974) 3:436–42. doi: 10.1016/0091-7435(74)90003-6

2. Krishna B. Disease surveillance: the bedrock for control and prevention. *Indian J Crit Care Med.* (2021) 25:745–6. doi: 10.5005/jp-journals-10071-23908

3. Richards CL, Iademarco MF, Atkinson D, Pinner RW, Yoon P, MacKenzie WR, et al. Advances in public health surveillance and information dissemination at the centers for disease control and prevention. *Publ Health Rep.* (2017) 132:403–10. doi: 10.1177/0033354917709542

4. Gwinn M, MacCannell DR, Khabbaz RF. Integrating advanced molecular technologies into public health. *J Clin Microbiol.* (2017) 55:703–14. doi: 10.1128/JCM.01967-16

5. Lo SW, Jamrozy D. Genomics and epidemiological surveillance. *Nat Rev Microbiol.* (2020) 18:478. doi: 10.1038/s41579-020-0421-0

6. Fola AA, Kattenberg E, Razook Z, Lautu-Gumal D, Lee S, Mehra S, et al. SNP barcodes provide higher resolution than microsatellite markers to measure plasmodium vivax population genetics. *Malar J.* (2020) 19:375. doi: 10.1186/s12936-020-03440-0

7. Bah SY, Morang'a CM, Kengne-Ouafo JA, Amenga-Etego L, Awandare GA. Highlights on the application of genomics and bioinformatics in the fight against infectious diseases: challenges and opportunities in Africa. *Front Genet.* (2018) 9:575. doi: 10.3389/fgene.2018.00575

8. Sun BB, Kurki MI, Foley CN, Mechakra A, Chen CY, Marshall E, et al. Genetic associations of protein-coding variants in human disease. *Nature.* (2022) 603:95–102. doi: 10.1038/s41586-022-04394-w

9. Zhong D, Koepfli C, Cui L, Yan G. Molecular approaches to determine the multiplicity of plasmodium infections. *Malar J.* (2018) 17:172. doi: 10.1186/s12936-018-2322-5

10. Pacheco MA, Lopez-Perez M, Vallejo AF, Herrera S, Arévalo-Herrera M, Escalante AA. Multiplicity of infection and disease severity in plasmodium Vivax. *PLoS Neglect Trop Dis.* (2016) 10:e0004355. doi: 10.1371/journal.pntd.0004355

11. Earland D, Buchwald AG, Sixpence A, Chimenya M, Damson M, Seydel KB, et al. Impact of multiplicity of *Plasmodium falciparum* infection on clinical disease in Malawi. *Am J Trop Med Hyg.* (2019) 101:412–5. doi: 10.4269/ajtmh.19-0093

12. Friedrich LR, Popovici J, Kim S, Dysoley L, Zimmerman PA, Menard D, et al. Complexity of infection and genetic diversity in Cambodian plasmodium Vivax. *PLoS Neglect Trop Dis.* (2016) 10:e0004526. doi: 10.1371/journal.pntd.0004526

13. 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. *Paras Vect.* (2020) 13:427. doi: 10.1186/s13071-020-04302-z

14. Hashemi M, Schneider KA. Bias-corrected maximum-likelihood estimation of multiplicity of infection and lineage frequencies. *PLoS ONE.* (2021) 16:e0261889. doi: 10.1371/journal.pone.0261889

15. Miar Y, Sargolzaei M, Schenkel FS. A comparison of different algorithms for phasing haplotypes using Holstein cattle genotypes and pedigree data. *J Dairy Sci.* (2017) 100:2837–49. doi: 10.3168/jds.2016-11590

16. Xu Z, Dixon JR. Genome reconstruction and haplotype phasing using chromosome conformation capture methodologies. *Brief Funct Genomics.* (2019) 19:139–50. doi: 10.1093/bfgp/elz026

17. Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol.* (1995) 12:921–7.

18. Hill WG, Babiker HA. Estimation of numbers of malaria clones in blood samples. *Proc R Soc B Biol Sci.* (1995) 262:249–57. doi: 10.1098/rspb.1995.0203

19. Hawley ME, Kidd KK. HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. *J Hered.* (1995) 86:409–11. doi: 10.1093/oxfordjournals.jhered.a111613

20. Schneider KA. Large and finite sample properties of a maximum-likelihood estimator for multiplicity of infection. *PLoS ONE*. (2018) 13:e0194148. doi: 10.1371/journal.pone.0194148

21. Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet*. (2001) 68:978–89. doi: 10.1086/319501

22. Wigger L, Vogt JE, Roth V. Malaria haplotype frequency estimation. *Stat Med*. (2013) 32:3737–51. doi: 10.1002/sim.5792

23. Rastas P, Koivisto M, Mannila H, Ukkonen E. A hidden Markov technique for haplotype reconstruction. In: Casadio R, Myers G, editors. *Algorithms in Bioinformatics. Lecture Notes in Computer Science*. Berlin; Heidelberg: Springer (2005). p. 140–51. doi: 10.1007/11557067_12

24. Druet T, Georges M. A Hidden Markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. *Genetics*. (2010) 184:789–98. doi: 10.1534/genetics.109.108431

25. Schneider KA, Escalante AA. A likelihood approach to estimate the number of co-infections. *PLoS ONE*. (2014) 9:e97899. doi: 10.1371/journal.pone.0097899

26. Hashemi M, Schneider K. *MLMOI: Estimating Frequencies, Prevalence and Multiplicity of Infection*. CRAN (2020). Available online at: https://CRAN.R-project.org/package=MLMOI

27. Hastings IM, Smith TA. MalHaploFreq: a computer programme for estimating malaria haplotype frequencies from blood samples. *Malar J*. (2008) 7:130. doi: 10.1186/1475-2875-7-130

28. 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) 6:33. doi: 10.2202/1544-6115.1321

29. 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) 15:430. doi: 10.1186/s12936-016-1473-5

30. 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:e42496. doi: 10.1371/journal.pone.0042496

31. 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. doi: 10.1186/1475-2875-13-102

32. 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. doi: 10.1186/1475-2875-14-4

33. Chang HH, 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:e1005348. doi: 10.1371/journal.pcbi.1005348

34. Assefa SA, Preston MD, Campino S, Ocholla H, Sutherland CJ, Clark TG. estMOI: estimating multiplicity of infection using parasite deep sequencing data. *Bioinformatics*. (2014) 30:1292–4. doi: 10.1093/bioinformatics/btu005

35. 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) 279:2589–98. doi: 10.1098/rspb.2012.0113

36. 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:e1005923. doi: 10.1371/journal.pcbi.1005923

37. 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) 27:93–103.e4. doi: 10.1016/j.chom.2019.12.001

38. Neafsey DE, Taylor AR, MacInnis BL. Advances and opportunities in malaria population genomics. *Nat Rev Genet*. (2021) 22:502–17. doi: 10.1038/s41576-021-00349-5

39. Dia A, Cheeseman IH. Single-cell genome sequencing of protozoan parasites. *Trends Parasitol*. (2021) 37:803–14. doi: 10.1016/j.pt.2021.05.013

40. 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. doi: 10.7554/eLife.40845

41. 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. doi: 10.1186/1475-2875-11-77

42. 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) 52:4089–97. doi: 10.1128/AAC.00623-08

43. Schneider KA. Charles Darwin meets ronald ross: a population-genetic framework for the evolutionary dynamics of malaria. In: Teboh-Ewungkem MI, Ngwa GA, editors. *Infectious Diseases and Our Planet*. Cham: Springer International Publishing (2021). p. 149–91. doi: 10.1007/978-3-030-50826-5_6

44. Neal AT. Distribution of clones among hosts for the lizard malaria parasite plasmodium mexicanum. *PeerJ*. (2021) 9:e12448. doi: 10.7717/peerj.12448

45. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. New York, NY: Chapman and Hall; CRC Press (1994). doi: 10.1201/9780429246593

46. Pacheco MA, Schneider KA, Cheng Q, Munde EO, Ndege C, Onyango C, et al. Changes in the frequencies of *Plasmodium falciparum dhps* and *dhfr* drug-resistant mutations in children from Western Kenya from 2005 to 2018: the rise of Pfdhps S436H. *Malar J*. (2020) 19:378. doi: 10.1186/s12936-020-03454-8

47. Ihaka R, Gentleman R. R: a language for data analysis and graphics. *J Comput Graph Stat*. (1996) 5:299–314. doi: 10.1080/10618600.1996.10474713

48. Basco LK, Ringwald P. Molecular epidemiology of malaria in Cameroon. XXIV. Trends of *in vitro* antimalarial drug responses in Yaounde, Cameroon. *Am J Trop Med Hyg*. (2007) 76:20–6. doi: 10.4269/ajtmh.2007.76.20

49. Tahar R, Basco LK. Molecular epidemiology of malaria in cameroon. XXVI. Twelve-year *in vitro* and molecular surveillance of pyrimethamine resistance and experimental studies to modulate pyrimethamine resistance. *Am J Trop Med Hyg*. (2007) 77:221–7. doi: 10.1016/j.actatropica.2007.04.008

50. Browning SR, Browning BL. Haplotype phasing: existing methods and new developments. *Nat Rev Genet*. (2011) 12:703–14. doi: 10.1038/nrg3054

51. Nabarro DN, Tayler EM. The "roll back malaria" campaign. *Science*. (1998) 280:2067–8. doi: 10.1126/science.280.5372.2067

52. Gamboa D, Ho MF, Bendezu J, Torres K, Chiodini PL, Barnwell JW, et al. A large proportion of *P. falciparum* isolates in the Amazon region of Peru lack pfhrp2 and pfhrp3: implications for malaria rapid diagnostic tests. *PLoS ONE*. (2010) 5:e8091. doi: 10.1371/journal.pone.0008091

53. Schneider KA, Kim Y. An analytical model for genetic hitchhiking in the evolution of antimalarial drug resistance. *Theor Popul Biol*. (2010) 78:93–108. doi: 10.1016/j.tpb.2010.06.005

54. Pacheco MA, Forero-Pe na 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) 19:176. doi: 10.1186/s12936-020-03247-z

55. Pava Z, Puspitasari AM, Rumaseb A, Handayuni I, Trianty L, Utami RAS, et al. Molecular surveillance over 14 years confirms reduction of plasmodium vivax and falciparum transmission after implementation of artemisinin-based combination therapy in Papua, Indonesia. *PLoS Neglect Trop Dis*. (2020) 14:e0008295. doi: 10.1371/journal.pntd.0008295