RESEARCH ARTICLE

Large and finite sample properties of a maximum-likelihood estimator for multiplicity of infection

Kristan Alexander Schneider*

Department CB, University of Applied Sciences Mittweida, Mittweida, Germany

* kristan.schneider@hs-mittweida.de

Abstract

Reliable measures of transmission intensities can be incorporated into metrics for monitoring disease-control interventions. Genetic (molecular) measures like multiplicity of infection (MOI) have several advantages compared with traditional measures, e.g., R_0. Here, we investigate the properties of a maximum-likelihood approach to estimate MOI and pathogen-lineage frequencies. By verifying regulatory conditions, we prove asymptotical unbiasedness, consistency and efficiency of the estimator. Finite sample properties concerning bias and variance are evaluated over a comprehensive parameter range by a systematic simulation study. Moreover, the estimator's sensitivity to model violations is studied. The estimator performs well for realistic sample sizes and parameter ranges. In particular, the lineage-frequency estimates are almost unbiased independently of sample size. The MOI estimate's bias vanishes with increasing sample size, but might be substantial if sample size is too small. The estimator's variance matrix agrees well with the Cramér-Rao lower bound, even for small sample size. The numerical and analytical results of this study can be used for study design. This is exemplified by a malaria data set from Venezuela. It is shown how the results can be used to determine the necessary sample size to achieve certain performance goals. An implementation of the likelihood method and a simulation algorithm for study design, implemented as an R script, is available as S1 File alongside a documentation (S2 File) and example data (S3 File).

OPEN ACCESS

Citation: Schneider KA (2018) Large and finite sample properties of a maximum-likelihood estimator for multiplicity of infection. PLoS ONE 13(4): e0194148. https://doi.org/10.1371/journal.pone.0194148

Editor: Ryan E. Wiegand, Centers for Disease Control and Prevention, UNITED STATES

Received: September 16, 2017

Accepted: February 26, 2018

Published: April 9, 2018

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

Funding: This work was funded by the SMWK-SAB project "Innovationsvorhaben zur Profilschärfung an Hochschulen für angewandte Wissenschaften" (Project number 100257255).

Competing interests: The author has declared that no competing interests exist.

Introduction

The decline of malaria incidence in sub-Saharan Africa and elsewhere shifted the focus of health authorities in many countries towards elimination. This renders the need to evaluate the effectiveness of control programs to reduce transmission more urgent. Indeed, codifying a set of metrics, suitable to easily and reliably measure the impact of new and existing control interventions on malaria transmission, is highly desirable. Of particular interest are metrics, capable to monitor changes in exposure and transmission intensity. A recent book chapter [1] extensively reviewed 11 metrics of malaria transmission with regard to precision, accuracy, methods of collection and cost efficiency. While the entomological inoculation rate (EIR) and the basic reproduction number R_0 are still the gold standards to measure transmission in malaria, molecular metrics such as multiplicity of infection (MOI) and molecular force of infection (mFOI) emerged as most appropriate. The relevance of identifying suitable metrics to quantify transmission is not restricted to malaria, but applies equally to other infectious diseases. Notably, incidence of MOI or superparasitism per se is epidemiologically an important metric of exposure in infectious diseases [2–11].

MOI refers to the number of super-infections of a disease, typically visible by the occurrence of multiple genetic variants ('lineages') within an infection. This is indicative of transmission dynamics as it reflects the overlap of several genetic variants due to multiple infectious contacts. Hence, MOI relates to intra-host dynamics [12], i.e., the dynamics of interactions among different 'lineages' within infections, and its derived pathogenic and epidemiological consequences. The concept of MOI is closely related to that of complexity of infection [13, 14].

Intra-host dynamics have been the subject of several theoretical and experimental investigations exploring a broad spectrum of scenarios over the last decades [12, 15–18]. Importantly, intra-host dynamics affect the spread of parasite lineages with adaptive mutations conferring resistance to antimicrobial agents or that allow the evasion of immune and/or vaccine-mediated protection [19, 20]. Currently, this is of particular importance in malaria, as the spread of artemisinin tolerance/resistance is threatening to challenge control efforts [21–23]. In summary, following or measuring MOI is essential whenever epidemiological inferences are influenced by or theoretical models depend on intra-host dynamics.

Although for a given pathogen it is relatively easy to measure the number of distinctive pathogenic lineages in models and experimental settings (e.g., [24]), it is not possible to define a universal framework of MOI that is appropriate for the vast spectrum of genetic architectures observed in pathogen organisms. For instance, viruses like HIV accumulate mutations at a rate that allows for the use of phylogenetic base methods [25]. On the contrary, eukaryotic parasites such as *Plasmodium*, *Trypanosoma*, *Toxoplasma*, and *Schistosoma* [26, 27] and bacteria such as *Mycobacterium* [28] evolve at a rate at which it is possible to determine a stable number of genetically distinct lineages during the course of an infection given a set of genetic markers. Even when restricted to such pathogens, due to difficulties naturally arising from confounding factors in ecological and epidemiological investigations [2, 3, 5, 28, 29], the concept of MOI is enervated. Estimating MOI and frequency spectra was dominated by ad-hoc methods, which are intuitive but typically introduce a bias, which cannot be quantified due to the lack of a statistical framework. E.g., in the context of malaria a patient's MOI is estimated as the maximum number of distinct alleles detected in a blood sample from one (e.g. [30, 31]), two (e.g. [32]) or several marker loci (e.g. [33]). Being a lower bound for the number of parasite haplotypes in a sample, this underestimates MOI (particularly if only a few markers are considered), with bias depending on the haplotype-frequency spectrum. On the other hand, MOI might by overestimated as a result of data artifacts (sequencing errors, wrong STR calls)—particularly with a large number of genetic markers considered. MOI was also estimated as the cumulative number of lineages (alleles) identified across samples divided by the number of disease-positive samples. This corresponds to the sum of the empirical prevalences of the respective lineages (e.g. [34]). Some authors additionally reported MOI as the average number of observed alleles separately for different marker loci (e.g. [35]). Alternatively, MOI was estimated as the average number of alleles at several marker loci (e.g. [36]). Also the number of polymorphic SNPs was used as an indirect measure for MOI by [37].

Other methods, try to estimate MOI using a statistical framework that does not simultaneously provide frequency estimates but rather suggest the use of ad-hoc frequency estimates (e.g. [13, 38]). Often haplotype (lineage) frequencies are estimated from single-infections (those in which only one haplotype is found; e.g. [37]). This method is justified if MOI is low with most infections being single infections. Increasing MOI raises uncertainty in each sample because super-infections with the same haplotype cannot be distinguished from single infections. Hence, sample size is decreased as multiple infections become more likely, which are excluded from the estimate. This approach is particularly unfortunate because samples containing most information about MOI (multiple infections) are excluded. Similar problems arise if single infections are employed for frequency estimation along those, in which only two distinct haplotypes occur (e.g. [39]). Another ad-hoc approach to estimate allele frequencies, which accounts for multiple infections by giving all alleles found at a marker in a multiple infection the same weight, was employed by [40]. Although sample size is not artificially reduced, this method does not properly consider the interaction of frequency spectra and MOI. A further alternative is to employ only the predominant lineage (haplotype) in infections for frequency estimation (cf. [34]).

Hence, a formal statistical framework that allows the estimation of the actual number of lineages and other approximations to MOI that facilitates and/or considers confounding factors is indispensable to avoid ad-hoc methods.

In the context of malaria (and related diseases) such a framework was introduced by [41] and further developed by [42] and [43]. More precisely, a maximum-likelihood framework was developed to estimate MOI and the frequency of pathogen ‘lineages’ from molecular data obtained from a collection of blood samples of disease-positive patients.

Notably, several alternative methods, based on essentially the same statistical framework, have been proposed. However, all have some limitations or use heuristic approximations. Common to all is that they focus on specific applications and the mathematical properties have not been studied in detail. The computer program MalHaploFreq by [44] uses a maximum-likelihood approach to estimate frequencies of haplotypes consisting of up to 3 SNPs, without providing an estimate of MOI. The latter is also not provided by the Gibbs-sampling method used in [45] to estimate haplotype frequencies, again designed for SNP data, but allowing for more than 3 SNPs. The same is true for the Metropolis-Hastings algorithm used in [46], which is not restricted to SNP data. The EM-algorithm adapted in [47] allows for many SNPs but frequency estimates are based on an approximation considering only the most frequent haplotypes. Heuristic estimates of MOI are required in the Metropolis Hastings algorithm of [48] for frequency estimates not restricted to SNPs. The EM and MCMC algorithms in [49], which are based on several approximations to make them numerically tractable, again focus on SNP data but provide MOI estimates. The program COIL [13] uses a likelihood approach for MOI estimation from SNP data. The algorithm however requires ad-hoc estimates of marginal frequencies, SNPs to be uncorrelated, and assumes a maximum possible MOI of 5. An Improvement, THE REAL McCOIL [14] adapts a Metropolis-Hastings algorithm to estimate MOI and minor-allele frequencies at uncorrelated SNPs in two different ways. The program estMOI [50] requires deep sequencing data. The formal statistical framework in [42] uses an EM algorithm for SNP data.

Here, we will further investigate the framework of [41, 43] and the concept of MOI with regard to the criteria pointed out by [1]. First, we will prove a number of regulatory conditions, which imply asymptotic unbiasedness, strong consistency and efficiency of the maximum-likelihood estimate. In addition to these asymptotic properties we numerically investigate the estimator’s finite sample properties by conducting a systematic simulation study to quantify the estimator’s bias (accuracy) and variance (precision). Importantly, we also investigate the estimator’s robustness to model violations.

First, we summarize the methods and derive analytic results. Then, we will describe the simulation study and summarize its outcome. As an illustration we will take a closer look at a malaria data set from Venezuela, previously published in [51].

An implementation of the likelihood method in R, which can be readily applied to molecular data, is available as supporting information, alongside a simulation algorithm for study design to determine the necessary sample size to achieve given accuracy and precision goals of the estimation (S1, S2 and S3 Files).

Materials and methods

Here, we briefly summarize the maximum-likelihood method to estimate the average MOI, proposed by [41] and [43].

Model background

Assume n different 'lineages' $A_1, \ldots, A_n$ of a pathogen, e.g., n alleles at a marker locus (or haplotypes in a non-recombining region), which circulate in a given population and are found in N blood samples of infected individuals—or more generally N clinical specimens. A blood sample can contain multiple lineages reflecting super-infections. Regarding the lineages, it is assumed that they are characterized by markers (SNPs or STRs), whose frequencies do not change too rapidly in the population. Because the frequency spectrum of markers linked to genes under selection might change rapidly or might be very skewed, we have neutral markers in mind. The n lineages considered are those that contribute to infection, not new variants that are generated by mutation inside hosts (and 'fail' to participate in transmission). The frequencies of the n lineages are denoted by $\boldsymbol{p} = (p_1, \ldots, p_n)$. The frequency vector $\boldsymbol{p}$ is an element of the $(n - 1)$-dimensional simplex $\mathcal{S}_n : \{(x_1, \ldots, x_n) | \sum_{i=1}^{n} x_1 = 1 \text{ and } x_i > 0 \text{ for all } i\}$.

Infective events are assumed to be rare and independent, i.e., already infected persons are not more or less likely to get infected (super-infected) than uninfected persons. With these assumptions the number of infections per individual is Poisson distributed, or more precisely conditionally Poisson (or positively Poisson) distributed since only infected individuals are considered (Eq S30). The conditional Poisson distribution is characterized by a single parameter λ. Jointly we denote the parameters by $\boldsymbol{\theta} = (\lambda, \boldsymbol{p}) \in \Theta := \mathbb{R}^+ \times \text{ int}\,\mathcal{S}_n$, where int $\mathcal{S}_n$ denotes the interior of the $(n - 1)$-dimensional simplex. In other words the parameters satisfy $\lambda > 0, \sum_{i=1}^{n} p_i = 1$ and $0 < p_k < 1$ for $k = 1, \ldots, n$.

It is further assumed that at each infective event one lineage is drawn randomly from the pathogen population (according to the lineages' frequencies) to infect the individual. Hence, if an individual is infected exactly m times (which is a conditional Poisson random number), m lineages are drawn according to a multinomial distribution with parameters m and $p_1, \ldots, p_n$. This yields a vector $(m_1, \ldots, m_n)$, where m_k is the number of times the individual was infected with lineage A_k. Clearly, $m_1 + \ldots + m_n = m$, because the individual is infected exactly m times. Looking at a blood sample, only the absence and presence of lineages is observed, but it is impossible to reconstruct the values m_k or even m. Even if only one lineage is found, it is unclear how many times the individual was infected with this lineage. Hence, looking at a blood sample only a 0–1 vector $\boldsymbol{i} = (i_1, \ldots, i_n)$ is observed, indicative of the present lineages, i.e., $i_k = 1$ if $m_k > 0$ and $i_k = 0$ if $m_k = 0$. The probability of observing an infection with configuration $\boldsymbol{i}$ is

$$Q_{\boldsymbol{i}} = \frac{1}{e^{\lambda} - 1} \prod_{j=1}^{n} (e^{\lambda p_j} - 1)^{i_j} \tag{1}$$

according to [43]. Notably, the probabilistic model defined by (1) is identifiable.

Remark 1 *The probability functions* (1) *defined by any two values of* $\boldsymbol{\theta} \in \Theta$ *are distinct.*

The proof is presented in Appendix A in S4 File. It is important to point out that p_k is the relative frequency of lineage A_k in not its prevalence. The former refers to the relative abundance of the lineage in the parasite population, the latter to the probability that the lineage is present in an infection (within the population of disease-positive individuals).

Remark 2 *Lineage* A_k*'s prevalence, i.e., the probability of observing* A_k *in a disease-positive blood sample, is*

$$q_{\{k\}} := q_k = \sum_{\substack{\boldsymbol{i} \in \{0,1\}^n \setminus \{\boldsymbol{0}\}: \\ i_k = 1}} Q_{\boldsymbol{i}} = \frac{e^{\lambda}(e^{\lambda p_k} - 1)}{e^{\lambda p_k}(e^{\lambda} - 1)}. \tag{2}$$

A proof is provided in Appendix B in S4 File.

A data set obtained from N blood samples consists of N 0-1-vectors, indicating which lineages are detected. We denote the jth blood sample by $\boldsymbol{x}_j = (x_{1j}, \ldots, x_{nj})$. Collectively, the data is denoted by $\boldsymbol{X}$. Further, N_k is the number of samples in which lineage A_k is detected, i.e., $N_k = \sum_{j=1}^{N} x_{kj}$. Under the outlined model, the log-likelihood of observing data $\boldsymbol{X}$ is given by

$$L = L(\lambda, \boldsymbol{p}) = L(\lambda, \boldsymbol{p} | \boldsymbol{X}) = -N \log (e^{\lambda} - 1) + \sum_{k=1}^{n} N_k \log (e^{\lambda p_k} - 1) \tag{3}$$

(cf. [43]). Obviously, $(N, N_1, \ldots, N_k)$ form a sufficient statistic for the data $\boldsymbol{X}$. The value N_k/N is the observed prevalence of lineage k.

Maximum-likelihood estimate

The maximum-likelihood estimate (MLE) for $(\lambda, \boldsymbol{p})$ exists and is uniquely defined except in two irregular situations. In the first, only one lineage is found in each blood sample, i.e., $\sum_{k=1}^{n} N_k = N$, i.e., there is no indication of super-infections. In the second, at least one lineage is found in every blood sample, i.e., $N_k = N$ for at least one k. We can state (cf. [43]):

Remark 3 *Except in irregular situations, the MLE* $\hat{\boldsymbol{\theta}} = (\hat{\lambda}, \hat{\boldsymbol{p}})$ *is given by*

$$\hat{p}_k = -\frac{1}{\hat{\lambda}} \log \left(1 - \frac{N_k}{N}(1 - e^{-\hat{\lambda}})\right), \tag{4a}$$

where $\hat{\lambda}$ *is found by iterating*

$$\lambda_{t+1} = \lambda_t - \frac{\lambda_t + \sum_{k=1}^{n} \log \left(1 - \frac{N_k}{N}(1 - e^{-\lambda_t})\right)}{1 - \sum_{k=1}^{n} \frac{N_k}{N e^{\lambda_t} - N_k(e^{\lambda_t} - 1)}}, \tag{4b}$$

which converges monotonically at quadratic rate from any initial value $\lambda_1 \geq \hat{\lambda}$. *Hence, it is guaranteed to find the MLE as long as the initial value* λ_1 *is chosen to be sufficiently large.*

If $N = \sum_{k=1}^{n} N_k$ *and* $N_k \neq N$ *for all* k, $\hat{\lambda} = 0$ *and* $\hat{p}_k = \frac{N_k}{N}$. *If* $N_k = N$ *for at least one* k, *the MLE does not exist* ("$\hat{\lambda} = \infty$").

An implementation of the algorithm above is provided as an R script (S1 File) alongside a documentation (S2 File).

Because the MLE does not exist if $N_k = N$ or $\sum_{k=1}^{n} N_k = N$, for study design it is important to minimize the probability of obtaining irregular data. Moreover, one might prefer conditioning the likelihood on a regular data set for analytical investigations. In Appendix B in S4 File the probability of observing irregular data is calculated to be

$$q := \frac{1}{(1-e^{-\lambda})^N}\left(1 - \prod_{j=1}^{n}\left(1-(1-e^{-\lambda p_j})^N\right)\right) + \left(\sum_{j=1}^{n} Q_{\boldsymbol{e}_j}\right)^N - \sum_{j=1}^{n} Q_{\boldsymbol{e}_j}^N \tag{5}$$

($\boldsymbol{e}_j$ denote the standard base vectors). Clearly, this probability vanishes as $N \to \infty$. However, if N and λ are small and the lineage frequencies are very skewed, observing irregular data is likely.

The problem of irregular data can be avoided by imposing restrictions on the parameter space, except if only one lineage is observed in the data, i.e., $\sum_{k=1}^{n} N_k = N$ and $N_j = N$ for some j. The MLE can be adapted as follows.

Result 1 *Assume the true parameter* $\boldsymbol{\theta}_0$ *lies within the interior of the compact set* $\hat{\Theta} = [\lambda_{\min}, \lambda_{\max}] \times \mathcal{S}_n$. *The maximum-likelihood estimate is given by Remark 3 if* $N_k \neq N$ *for all* k, $\sum_{k=1}^{n} N_k > N$ *and* $\hat{\lambda} \in [\lambda_{\min}, \lambda_{\max}]$. *If* $\hat{\lambda} < \lambda_{\min}$ *or* $\sum_{k=1}^{n} N_k = N$ *(but* $N_j \neq N$ *for all* j*), the MLE is given by* $\hat{\boldsymbol{\theta}} = (\lambda_{\min}, \hat{\boldsymbol{p}})$, *where*

$$\hat{p}_k = -\frac{1}{\lambda_{\min}} \log\left(1 - \frac{N_k \lambda_{\min}}{\hat{\beta}}\right), \tag{6a}$$

where $\hat{\beta}$ *is found by iterating*

$$\beta_{t+1} = \beta_t - \beta_t \frac{\lambda_{\min} + \sum_{k=1}^{n} \log\left(1 - \frac{N_k \lambda_{\min}}{\beta_t}\right)}{\sum_{k=1}^{n} \frac{N_k \lambda_{\min}}{\beta_t - N_k \lambda_{\min}}}, \tag{6b}$$

which is guaranteed to converge from any initial value β_1 *satisfying* $\max_{k=1,\ldots,n} N_k \lambda_{\min} < \beta_1 < \hat{\beta}$. *If* $\hat{\lambda} > \lambda_{\max}$ *or* $N_k = N$ *for any* k *(but* $\sum_{k=1}^{n} N_k > N$*), the MLE is given by* $\hat{\boldsymbol{\theta}} = (\lambda_{\max}, \hat{\boldsymbol{p}})$, *where* $\hat{\beta}$ *is given by* (6) *with* $\lambda_{\max}$ *replaced by* $\lambda_{\min}$.

If only one lineage is present in the data, i.e., $\sum_{k=1}^{n} N_k = N$ *and* $N_j = N$ *for some* j, *the MLE is not unique, more precisely any estimate* $(\lambda, \boldsymbol{e}_j)$ *is equally likely.*

A proof is found in Appendix B in S4 File.

Results

Large-sample properties

Usually MLEs have attractive limiting properties under relatively weak conditions. To prove these here it is more convenient to regard the admissible parameter space as a subset of $\mathbb{R}^n$.

This is achieved by eliminating one of the redundant frequencies. We set $p_n = 1 - \sum_{k=1}^{n-1} p_k$, $\boldsymbol{\vartheta} = (\lambda, p_1, \ldots, p_{n-1})$ and $\tilde{\Theta} = \{(\lambda, p_1, \ldots, p_{n-1}) \mid \lambda_{\min} \leq \lambda \leq \lambda_{\max},\ 0 < p_k \,\forall\, k \text{ and } \sum_{k=1}^{n-1} p_k < 1\}$.

Let $\hat{\boldsymbol{\vartheta}}$ denote the corresponding MLE, which, of course, equals $\hat{\boldsymbol{\theta}}$ with the last component dropped. In the new parameter space the Fisher information matrix is derived as follows.

Result 2 *The Fisher information matrix,* $I_N(\boldsymbol{\vartheta}) := -\mathbb{E}\frac{\partial^2 L}{\partial \boldsymbol{\vartheta}^2}$, *is given by*

$$I_{1,1} = \frac{-Ne^{\lambda}}{(e^{\lambda}-1)^2} + \frac{Ne^{\lambda}}{e^{\lambda}-1} \sum_{k=1}^{n} \frac{p_k^2}{e^{\lambda p_k}-1}\,, \tag{7a}$$

$$I_{1,k+1} = I_{k+1,1} = \frac{N\lambda}{1-e^{-\lambda}} \left(\frac{p_k}{e^{\lambda p_k}-1} - \frac{p_n}{e^{\lambda p_n}-1} \right) \quad \textit{for } k = 1, \ldots, n-1\,, \tag{7b}$$

$$I_{k+1,k+1} = \frac{N\lambda^2}{1-e^{-\lambda}} \left(\frac{1}{e^{\lambda p_k}-1} + \frac{1}{e^{\lambda p_n}-1} \right) \quad \textit{for } k = 1, \ldots, n-1\,, \tag{7c}$$

$$I_{k+1,j+1} = \frac{N\lambda^2}{1-e^{-\lambda}} \frac{1}{e^{\lambda p_n}-1} \quad \textit{for } k, j = 1, \ldots, n-1,\ k \neq j\,. \tag{7d}$$

The information matrix is derived in Appendix C in S4 File and two of its important properties are proved (cf. Theorem 1 and Theorem 2 in S4 File), namely:

Remark 4 *The Fisher information matrix is positive definite and satisfies*

$$\boldsymbol{I}_N = -\mathbb{E}\frac{\partial^2 L}{\partial \boldsymbol{\vartheta}^2} = \mathbb{E}\left(\left(\frac{\partial L}{\partial \boldsymbol{\vartheta}} \right)^T \cdot \left(\frac{\partial L}{\partial \boldsymbol{\vartheta}} \right) \right). \tag{8}$$

Indeed for the MLE we can state the following result.

Result 3 *The MLE specified in Result 1 is*

(i) strongly consistent, i.e. $\hat{\boldsymbol{\theta}} \overset{a.s.}{\to} \boldsymbol{\theta}_0$,

(ii) asymptotically unbiased, i.e., $\mathbb{E}\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}_0$,

(iii) efficient, i.e., $(\mathrm{Var}\ \hat{\boldsymbol{\theta}})V \to I_{n\times n}$,

(iv) asymptotically normally distributed, i.e., $\hat{\boldsymbol{\theta}} \sim \mathcal{N}(\boldsymbol{\theta}_0, V)$,

where the asymptotic covariance matrix $V = (v_{ij})$ *is the Cramér-Rao lower bound given by*

$$v_{11} = \frac{(e^{\lambda}-1)^2}{Ne^{\lambda}}\,\frac{C}{e^{\lambda}-1-C}\,, \tag{9a}$$

$$v_{1j} = v_{j1} = \frac{(e^{\lambda}-1)^2}{\lambda Ne^{\lambda}}\,\frac{e^{\lambda p_j}-1-p_jC}{e^{\lambda}-1-C}\,, \tag{9b}$$

$$v_{ii} = \frac{(e^{\lambda}-1)^2}{\lambda^2 Ne^{\lambda}}\left(\frac{e^{\lambda p_i}-1}{e^{\lambda}-1} + \frac{p_i^2C - 2p_i(e^{\lambda p_i}-1) + \frac{(e^{\lambda p_i}-1)^2}{e^{\lambda}-1}}{e^{\lambda}-1-C}\right), \tag{9c}$$

$$v_{ij} = \frac{(e^{\lambda}-1)^2}{\lambda^2 Ne^{\lambda}}\,\frac{p_ip_jC - p_i(e^{\lambda p_j}-1) - p_j(e^{\lambda p_i}-1) + \frac{(e^{\lambda p_i}-1)(e^{\lambda p_j}-1)}{e^{\lambda}-1}}{e^{\lambda}-1-C}\,, \tag{9d}$$

for $i, j = 2, \ldots n + 1$, $i \neq j$ *and*

$$C = \sum_{k=1}^{n}(e^{\lambda p_k}-1) \tag{9e}$$

Proof. First, the true parameter $\boldsymbol{\theta}_0$ lies in the interior of Θ. Hence, a compact subset $\hat{\Theta} \subsetneq \Theta$ exists, such that $\boldsymbol{\theta}_0 \in \text{ int}\,\hat{\Theta}$. By eliminating the redundant variable p_n this is equivalent to $\boldsymbol{\vartheta}_0 \in \text{ int }\ \Theta^1 \subseteq \mathbb{R}^n$, where $\Theta^1 \subseteq \tilde{\Theta}$ is compact. Second, the model is identifiable according to Remark 1. Third, the first three derivatives of the log-likelihood function with respect to the parameters exist and $\frac{1}{N}\frac{\partial^3 L}{\partial \boldsymbol{\vartheta}^3}$ is uniformly bounded on Θ^1 according to Remark 1. Fourth, the Fisher information satisfies (8) and is positive definite. Hence, the regulatory conditions given in [52] (Chapter 4, p.118) are satisfied. These imply strong consistency, asymptotical unbiasedness and efficiency of $\hat{\boldsymbol{\vartheta}}$ and hence $\hat{\boldsymbol{\theta}}$. The Cramér-Rao lower bound is derived in Appendix D in S4 File.

The mean MOI is given by $\psi = \frac{\lambda}{1-e^{-\lambda}}$ rather than by the Poisson parameterλ and might be preferable. Since MLEs are transformation respecting, $\hat{\psi} = \frac{\hat{\lambda}}{1-e^{-\hat{\lambda}}}$ holds. Also the Cramér-Rao bound needs some adjustment (see Appendix E in S4 File).

Remark 5 *The Cramér-Rao bound* $\tilde{V}$ *of the MLE* $(\hat{\psi}, \hat{p}_1, \ldots, \hat{p}_n)$ *is given by*

$$\tilde{v}_{1,1} = \frac{e^{\lambda}(e^{\lambda}-\lambda-1)^2}{N(e^{\lambda}-1)^2}\,\frac{C}{e^{\lambda}-1-C}\,, \tag{10a}$$

$$\tilde{v}_{1,j} = \frac{e^{\lambda}-\lambda-1}{\lambda N}\,\frac{e^{\lambda p_j}-1-p_jC}{e^{\lambda}-1-C}\,, \tag{10b}$$

$$\tilde{v}_{ij} = v_{ij}\,, \tag{10c}$$

where $i, j = 2, \ldots n + 1$ *and* v_{ij} *an* C *are given by* (9c), (9d) *and* (9e).

Finite sample properties of the MOI estimate

The desirable properties of the MLE hold only in the large-sample limit given that the parametric model (1) is correct. In practice, the MLE's quality depends on (i) the model's fit,

(ii) the true parameters, and (iii) sample size. To investigate these dependencies, we conduct a systematic numerical study. All numerical investigations were performed in R version 3.1.0 [53]. A detailed description is found in Appendix F in S4 File. The main R code for the simulations is provided as supporting information (S1 File), adapted for users to run their own simulations. All detailed results are provided as S5 File.

Mean and median bias. The MLE for MOI, $\hat{\psi} = \frac{\hat{\lambda}}{1-e^{-\hat{\lambda}}}$ is—as typically for MLEs—biased, however as shown analytically bias vanishes as sample size N increases (Fig 1A–1D and S5 File). The maximum bias is about 4%. There is a tendency of overestimating the true parameter in a non-linear fashion. As ψ increases, bias first decreases until $\psi \approx 1.2$, and then starts to increase almost linearly (Fig 1 and S5 File). Overestimation occurs on average because ψ is bounded from below by 1, whereas it has no upper bound. Estimates for ψ will be occasionally much too large while they cannot be much too small. This is particularly likely for very small and large ψ. For these reasons it seems better to use the median bias as a proxy for the estimate's accuracy (see below).

Bias vanishes quickly as N increases. A considerable reduction occurs when sample size is increased from 40 to 50. For $N = 100$ bias is already low, but there is still a remarkable gain by increasing sample size to 150–still a realistic sample size. Bias typically stays below $\approx 0.5\%$ then and almost vanishes for small values of ψ. Increasing sample size to 200, yields an improvement mainly for values beyond $\psi \approx 1.7$, which however is a rather extreme parameter range in practice. Bias almost vanishes for $N \geq 300$, at least for moderate values of ψ (Fig 1 and S5 File).

While the overall pattern described above is valid regardless of the lineage-frequency distribution, higher skewness leads to increased bias (compare Fig 1A with 1B)–particularly, if one lineage is dominating whereas others have frequency of $\approx 5\%$. The effect is not too strong if $N \geq 100$, in which case bias stays below 5%. This is not surprising, because a highly frequent lineage super-infects with high probability. In a sample this cannot be detected as a super-infection, and thus leads to a small underestimation of ψ. However, sometimes low-frequency lineages are over-represented (they will occur together with high-frequency lineages), leading occasional to huge over-estimates. Consequently, rare outliers create bias. Since the median is robust against outliers, skewed frequency distributions do not affect the median bias.

Bias decreases with an increasing number of lineages (Fig 1 and S5 File) while the qualitative pattern described above remains unchanged. The reason is that for larger n, super-infections are more accurately represented. Namely, for given n super-infections with $n + 1$ or more lineages cannot be identified. Although this leads more likely to underestimates, occasionally huge over-estimates occur. Over-represented super-infections are interpreted as large MOI due to the underlying Poisson model. For these reasons and those mentioned above, for very skewed frequency spectra (Fig 1 and S5 File), bias increases drastically, especially if sample size is small. For small ψ bias increases up to 15%. For extremely large ψ bias becomes negative, because samples indicative of super-infections with several lineages become rare, as the same lineages are infecting multiple times.

Slight model violations do not change the overall pattern of bias much (cf. Fig 2A–2C). Bias tends to increase faster with ψ but typically stays less than 2% for $N \geq 150$ and realistic values of ψ. For the most extreme parameters bias stays below 15%. It tends to be a little higher for the shifted binomial model (cf. Appendix F in S4 File) than for the other two alternatives, which is not surprising since this model constitutes the largest model violation among them. Assuming the uniform model—a radical model violation—changes the overall pattern (cf. Fig 2 and S5 File). As ψ increases, bias first increases (from typically 5%–10%) slightly and then

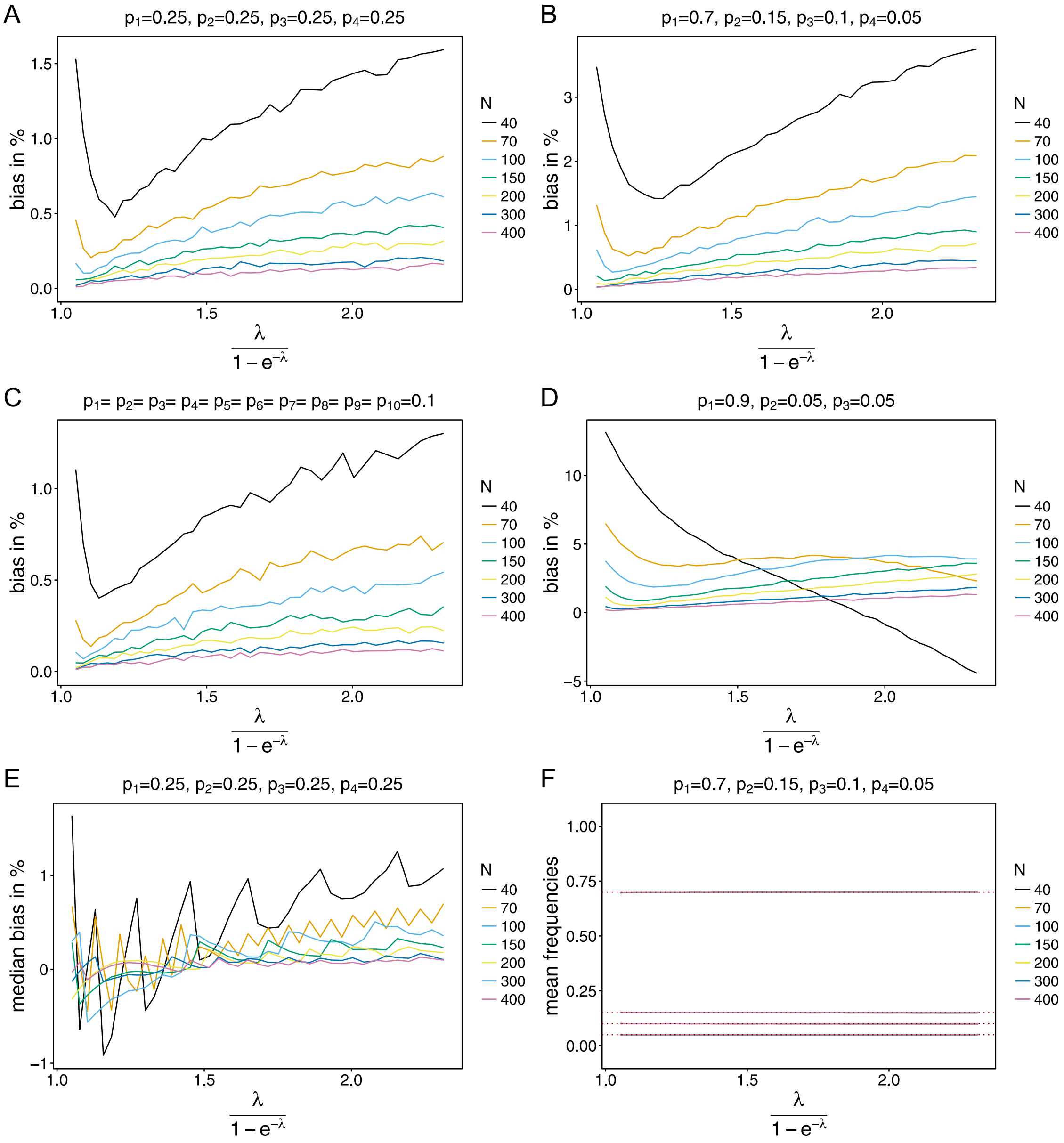

Fig 1. Bias for the conditional Poisson model. (A)-(D) Shown is the bias of the MLE $\hat{\psi}$ in percent as a function of the true parameter ψ based on simulated data created by the conditional Poisson model. Each panel assumes different n and lineage frequency distributions $\boldsymbol{p}$ shown at the top of each panel. Each line is for a different sample size N. (E) Shown is the median bias rather than the mean bias. (F) Average estimates for the lineage frequencies $p_1 = 0.7$, $p_2 = 0.15$, $p_3 = 0.1$ and $p_4 = 0.05$ (marked by the red-dotted horizonal lines) for different sample sizes. Since the lineage-frequency estimates are almost unbiased—independently of the sample size—the lines corresponding to different N almost coincide. Hence, only the top lines (purple lines for $N = 200$) at $p_1 = 0.7$, $p_2 = 0.15$, $p_3 = 0.1$ and $p_4 = 0.05$ and the dotted red lines on top are visible.

https://doi.org/10.1371/journal.pone.0194148.g001

becomes strongly negative—however for an extreme parameter range. Bias improves if sample size increases, for more lineages (larger n), and more balanced lineage-frequency distributions.

Concluding, bias is moderate if sample size is sufficiently large $N \geq 150$, which seems to be a reasonable compromise between feasibility and accuracy.

Bias, measured by the median, is usually by an order of magnitude smaller (compare Fig 1A with 1E), for the reasons mentioned above. In fact it is almost absent, particularly if sample

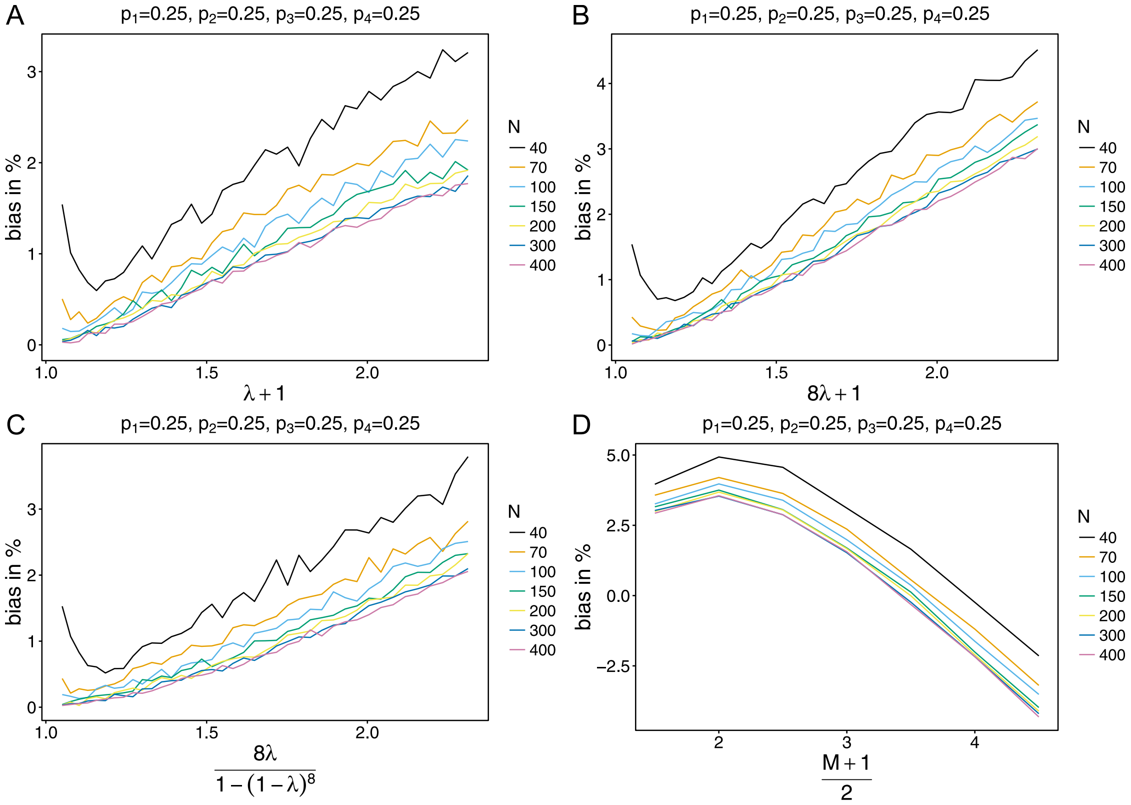

Fig 2. Dependence on true model. Shown is the bias as a function of the true parameter ψ, for $n = 4$ and a balanced lineage-frequency distribution. The underlying true models are the shifted Poisson, shifted binomial, conditional binomial and uniform models (cf. Appendix F in S4 File) in panels (A), (B), (C), (D), respectively.

https://doi.org/10.1371/journal.pone.0194148.g002

size is sufficiently large ($N \geq 100$) and $\psi \leq 1.7$. Although the maximum median bias (for small sample size, extremely skewed lineage frequencies, and extremely high MOI) is up to 15%, it is typically much lower. Hence, median bias can be regarded as being absent if sample size is ≥ 100.

It should be pointed out that bias is reported conditional on regular data here. The probability of irregular data satisfying $\sum_{k=1}^{n} N_k = N$ increases as λ and N decrease. Such data result in estimates of $\hat{\lambda} = 0$ or $\hat{\lambda} = \lambda_{\text{min}}$, which are close to the true value. Therefore, if bias is not conditioned on $\sum_{k=1}^{n} N_k > N$, it almost vanishes for small λ, but depends on the choice of λ_{min}. As the asymptotic results on the estimator assume regular data, we decided to present bias only conditional on regular data.

Variance. Variation of the estimator of ψ is best measured relative to the mean, i.e., by the coefficient of variation (CV). For most parameter combinations, the CV increases with ψ in a Holing-Type-II fashion (see Fig 3 in S5 File). This is particularly true for balanced lineage-frequency distributions. The CV decreases with increasing sample size and typically stays below 20% for $N = 40$ and on the order of 5% for $N = 400$. The CV is smaller for a larger number of lineages n. For skewed lineage-frequency distributions, the CV might be largest for intermediate ψ and small sample size $N \leq 100$ (cf. Fig 3D). Not surprisingly, these results are robust against model violations; in fact the CVs cannot be distinguished visually. The observed

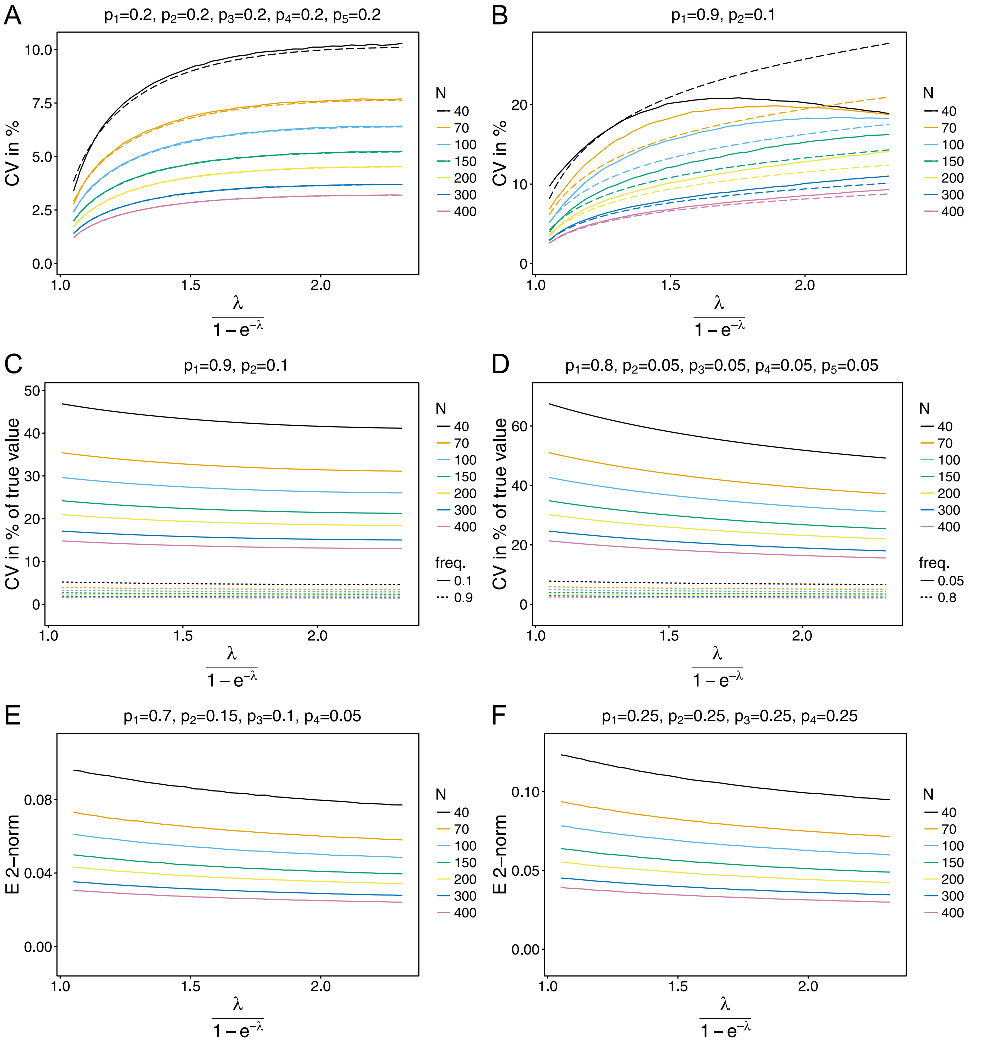

Fig 3. Measures of variation. (A)-(B) CV of $\hat{\psi}$ in % (i.e. ×100) as a function of ψ for the conditional Poisson model. The dashed line is the respective prediction based on the Cramér-Rao lower bound. Almost identical pictures are obtained for the other models (conditional Poisson, shifted Poisson, conditional binomial and shifted binomial). Panels are for different $\boldsymbol{p}$ (with different lineage numbers n). (C)-(D) CV for lineage frequencies. Shown is the theoretical prediction which is almost indistinguishable from the curves obtained by simulation, for all models. Panels are for different $\boldsymbol{p}$ (with different lineage numbers n). (E)-(F) Average Euclidian distance of the MLE $\hat{\boldsymbol{p}}$ and the true parameter $\boldsymbol{p}$. Shown are the curves obtained from the conditional Poisson model. Panels are for different $\boldsymbol{p}$ (with different lineage numbers n).

https://doi.org/10.1371/journal.pone.0194148.g003

pattern can be explained from the results concerning bias. If the true ψ is small, bias is created mainly by rare large overestimates, while any underestimate will be close to the true value. This hardly affects the (empirical) variance of $\hat{\psi}$, which is small by nature if the true value of ψ is small. Thus, relative to the average estimate, which is an overestimate, variance is small. For larger true values of ψ, variance increases naturally. Moreover, an underestimate $\hat{\psi}$ will no longer be very close to the true value, which further contributes to an increased variance. Finally, relative to the average $\hat{\psi}$, which now properly reflects the true value, the CV is large than for small ψ. This effect does not result in a linear increase of the CV, because the average $\hat{\psi}$ increases disproportionately compared with its variance.

To evaluate the estimator's efficiency the CV can be compared to its theoretical prediction, i.e., to the square root of the Cramèr-Rao lower bound (given by Eq 10a) divided by the true value of ψ. For at least intermediate sample size $N \geq 150$, the CV is very close to its theoretical prediction unless the lineage-frequencies are extremely skewed. In the latter case the CV is typically much smaller than its prediction for large ψ (see Fig 3D).

Notably, variance is rather large compared with bias. An approximate 95% confidence interval for the estimator relative to its true value is obtained as the relative bias ±2×CV.

Concluding, the variation is typically almost identical to its theoretical minimum for $n \geq 3$, $N \geq 100$ and balanced lineage-frequencies. For $N \geq 150$, realistic ψ and not too skewed $\boldsymbol{p}$, the CV is on the order of 5%.

Finite sample properties of the frequency estimates

Bias. The picture for the frequency estimates is much simpler than for the MOI parameter (Fig 1F). They are (almost) unbiased, independently of the sample size (N), the number of lineages (n), skewness of the lineage-frequency distribution, and true underlying model (S5 File). Whereas the MLE's performance of the MOI parameter depends on all these (at least to some extend), the frequency estimates are robust against these factors (S5 File).

Variance. There are several variance measures for frequency estimates $\hat{\boldsymbol{p}}$. The simplest approach is to consider the variances of each lineage frequency separately. The advantage is that the variances are comparable with the Cramér-Rao bounds, which perfectly matches the empirical variance. However, as the number of lineages n increases, these measures become tedious, especially since one is typically interested in functions or summaries of the distribution, rather than in the frequencies separately.

Like the bias, the variances of the estimates do not depend on the true underlying model. Not surprisingly, the variance is typically small, particularly if the true frequencies are small. Hence, it is more appropriate to consider the coefficient of variation, which implies only quantitative but no qualitative changes.

Since the variance is robust to model violations, so is the CV. Fig 3 shows the theoretical prediction of the CV, i.e., $\frac{\sqrt{\tilde{I}^{-1}_{k,k}}}{p_k}$. The observed CV, i.e., square root of sample variance divided by sample mean of $\hat{p}_k$, almost perfectly matches the theoretical prediction and it cannot be distinguished between the different underlying models. However, the CV depends on the number of lineages (n), sample size (N), MOI (ψ) and skewness of the allele frequency distribution. The CV decreases with increasing sample size and ψ. This is rather intuitive, since the data contains increasingly more information. For large ψ, more super-infections occur, which contain more information about the lineage frequencies, hence not only large N but also $\sum_{k=1}^{n} N_k$ imply more information per lineage. The CV tends to increase with an increasing number of lineage (n) and decreasing lineage-frequency. Again this is not surprising, because rare lineages hardly occur in a data set and this lack of information results in an increased variance. The CV is around 15% for balanced lineage frequencies. However, for $n = 10$ and lineage frequencies of 0.01, the CV increase to a substantial amount of 150% (S5 File). Nevertheless, for the major lineage frequencies, the CV typically stays under 10% if $N \geq 150$.

Considering the CV is conservative, because one is typically interested in summary statistics of the lineage frequencies such as heterozygosity (cf. [54]). Such statistics will be insensitive against various factors, since small lineage frequencies—which have a large relative, but a very small absolute error—hardly affect them. Variance measured by the average Euclidian distance to the true parameter gives similar results (Fig 3E and 3F). The average distance is robust against model violations and decreases with increasing ψ and sample size N for the reasons mentioned above. The average Euclidian distance also increases with an increasing number of lineages. However, unlike the CV, it decreases for more skewed lineage-frequency distributions (Fig 3E and 3F), because the absolute error of minor lineages hardly affects the distance measures.

Data application

The malaria dataset from [51] is examined more closely. From 97 malaria positive blood samples, 56 microsatellite loci were assayed, where 12 of the 56 marker loci can be considered selectively neutral. Hence, these are appropriate for the ML methods (cf. [43]).

Four of the 12 selected marker loci violated the assumption $\sum_{k=1}^{n} N_k > N$ (irregular data). In other words a third of the neutral markers did not provide sufficient information, a fact which can be avoided in the future by using the results provided here. Namely, it is possible to calculate a proxy of the necessary sample size to avoid irregular data sets.

First, note that typically some STR markers fail to amplify in a number of blood samples. It is not obvious how to proceed with the resulting missing data. Namely, a sample with a missing value cannot be considered disease free, since only disease-positive samples are analyzed. Moreover, it is not clear whether missing values are completely random or if there is a non-ignorable dependence between missing values, markers and repeat length, e.g., certain markers will amplify better than others or the probability for at least one lineage to amplify is higher if several variants super-infect. Hence, missing data depends on many confounding factors, which are difficult or impossible to determine. A pragmatic approach to handle missing data is to treat each marker as a data set and ignore, for each marker, those sample with a missing value. Consequently, this results in a different sample size for each marker.

Having decided on how to proceed with missing values, the MLEs for the lineage frequencies can be used as an estimate of the true frequencies. This can also be done in case that $\sum_{k=1}^{n} N_k = N$ (cf. Result 3). The median of the MLEs, $\lambda_{me} = 0.7213$, (derived from those 8 marker loci to which the method is applicable) can be used as an estimate proxy for the true parameter. This is justified because the median should be almost unbiased. Based on $\hat{\boldsymbol{p}}$ and λ_{me} the probabilities of obtaining an irregular data set (Eq 5) is listed in Table 1. Clearly, these probabilities are substantially large. Based on these estimates, the probability that at least four of the 12 markers yield irregular data is ≈ 28% (if all 12 markers are independent, which is a valid assumption for neutral markers). To reduce the probability of obtaining irregular data in every marker (independently) below 5%, sample size needs to be increased to $N = 300$. The probability that four or more markers yield irregular data is then $\approx 3.4 \times 10^{-6}$. However, the probability to obtain at least one irregular data set is still ≈ 13%. Increasing sample size to $N = 400$ reduced this probability to less than 5%.

Table 1 also shows the square root of the Cramér-Rao lower bound. Notice, that this is the lower bound for ψ. Hence, it should be compared with the respective median $\psi_{me} = 1.037$. The corresponding coefficient of variation is about 5%. Finally, the bias of the MLEs of ψ can be approximately obtained by looking up the bias for similar parameters from the simulation study. The maximum estimate of the bias (1.4%) is obtained for marker *J6*. Mostly, bias is as low as 0.5%.

Summarizing, a sample size of $N = 97$ is sufficient to obtain accurate and precise estimates, as justified by the proxies of bias obtained from the simulation study and the Cramér-Rao lower bounds based on Table 1. The quality of the estimates is also reflected by the fact that all

Table 1. Results for Venezuela data. Shown are sample size N, MLEs for $\hat{\boldsymbol{p}}$ and $\hat{\lambda}$, an estimate for the square root of the Cramér-Rao lower bound for ψ, the probabilities of obtaining an irregular data set derived from (5) using the estimates $\hat{\boldsymbol{p}}$ and λ_{me} for the actual sample size N (q), sample size $N = 300$ (q_{300}) and $N = 400$ (q_{400}), respectively. The estimate λ_{me} is the median of the 8 estimates for which $\hat{\lambda} > 0$ (regular data). In pairwise comparisons these eight estimates were not found to be significantly different at a 5% level based on pairwise likelihood-ratio tests provided in [43].

locus	N	$\hat{\boldsymbol{p}}$	$\hat{\lambda}$	$\sqrt{CR}$	q	q_{300}	q_{400}	bias
U5	7	(0.78, 0.09, 0.08, 0.01, 0.01, 0.01, 0.01, 0.01)	0.055	0.065	0.265	0.015	0.005	≈ 0.5%
K6	95	(0.785, 0.205, 0.01)	0.06	0.065	0.305	0.025	0.005	≈ 0.5%
L1	96	(0.53, 0.47)	0.085	0.055	0.175	0.005	0.	≈ 0.8%
c4	74	(0.63, 0.36, 0.015)	0.055	0.065	0.275	0.005	0.	≈ 0.5%
b3	88	(0.6, 0.26, 0.13, 0.01)	0.12	0.055	0.165	0	0.	≈ 1%
fr13	97	(0.72, 0.27, 0.01)	0.1	0.06	0.235	0.01	0.005	≈ 0.5%
ps6	97	(0.64, 0.36)	0.09	0.055	0.195	0.005	0.	≈ 0.8%
ps7	81	(0.6, 0.28, 0.11, 0.01)	0.045	0.055	0.195	0	0.	≈ 0.5%
J3	96	(0.625, 0.28, 0.085, 0.01)	0	0.055	0.16	0.005	0.	≈ 0.5%
J6	96	(0.845, 0.085, 0.075)	0	0.075	0.38	0.05	0.02	≈ 1.4%
U6	97	(0.65, 0.28, 0.07)	0	0.055	0.175	0.005	0.	≈ 0.5%
L4	91	(0.69, 0.12, 0.09, 0.075, 0.01, 0.01)	0	0.055	0.195	0.005	0.	≈ 0.8%

https://doi.org/10.1371/journal.pone.0194148.t001

8 'regular markers' yield estimates which are not statistically significant in pairwise comparisons (cf. [43]). However, this sample size is insufficient to guarantee with high probability that the data for each marker is regular. To guarantee regular data for 12 markers with at least 95% probability a sample size of 400 is necessary. Such calculations can already be performed during study design, if there are some vague ideas about the true parameters. Of note, it might be difficult to collect $N = 400$ samples in a low transmission area like Venezuela. In this case, a sample size of $N = 100$ should be sufficient, but many molecular markers should be assayed.

Discussion

A central goal of infectious-disease control programs is the reduction of the circulating pathogen's population size. Understanding the genetic changes associated with diminishing population size may provide valuable metrics to monitor success of control interventions. The reason is that population-genetic parameters reflect transmission intensities more accurately than incidence data—at least they will complement incidence data. Two quantities are starting to be more recognized in this context in epidemiology [1, 55], molecular force of infection and multiplicity of infection (MOI). The potential gain of incorporating such genetic/molecular information to infer transmission compared to traditional measures, e.g., entomological inoculation rate (EIR) or basic reproduction numbers (R_0), which are notoriously difficult to estimate, is starting to be realized [1, 55].

The aim of this article was to obtain a better understanding for the approach of [41] and [43] to estimate MOI and "lineage" frequency spectra in infections. A detailed description on how MOI relates to quantities such as molecular force of infection, EIR or R_0 can be found in [1]. While MOI might be built into a metric for monitoring transmission, accurate estimates of lineage-frequency spectra are desirable for monitoring the evolutionary dynamics of an endemic disease and for calculating frequency-based statistics. The method explored here is applies to diseases, for which infections are rare and independent events, and the course of the disease is relatively short. More precisely, de novo mutations should accumulate rarely within an infection. Hence, the method will not be applicable to pathogens like HIV, but to diseases like malaria.

To further investigate the properties of the maximum-likelihood approach of [43], we conducted a comprehensive numerical robustness study, which was complemented by additional analytical findings. (An implementation of the method and the simulation algorithm that can readily visualize the outcomes is available as S1 File). The study was designed under the criteria outlined in [1]. Particularly, we wanted to quantify the quality of the MOI estimate and the estimates for lineage frequencies in terms of bias and variance—with regard to different parameters and model violations. The method's advantage is its simplicity. Namely, to calculate the estimates from N blood samples or clinical specimens, in which n lineages are detected, it suffices to determine the numbers N_k of blood samples in which lineage k is found. The MLE can be calculated from the numbers $N_1, \ldots, N_n$ and N.

We proved usual attractive properties (asymptotic unbiasedness, strong consistency, efficiency) for the MLE, which also has good finite sample properties. The MLE yields reliable results if sample size is at least moderately large ($N \geq 150$). The method performs better if the lineage-frequency spectrum is not too skewed, MOI (ψ) is small, and more lineages are circulating (larger n) for the following reasons. If the same lineage is super-infecting multiple times, information is lost, as it is indistinguishable whether this lineage was infecting just once or more than once. This loss of information occurs with higher probability if the lineage-frequency spectrum is skewed, MOI is high, or the number of different lineages is small. In other words, not just sample size N, but also $\sum_{k=1}^{n} N_k$ is a measure of how much information is available. With balanced frequency spectra $\sum_{k=1}^{n} N_k$ will tend to be larger than for unbalanced ones, resulting in more reliable estimates.

The MLE is only asymptotically unbiased. This is true particularly for the MLE for MOI (as measured by the average bias), whereas the estimates for the lineage frequencies are (almost) unbiased—even under model violations. The MOI parameter is biased, because it has a lower but no upper bound and unlikely data lead to disproportionately large estimates. However, the MOI estimate is almost unbiased, if bias is measured by the median. This is not true with model violations, although the median bias is still smaller than the mean bias. Therefore, the method appears nevertheless applicable if sample size is sufficiently large ($N \leq 150$).

In general, the variance of the estimates is small—however it is large compared with bias. Especially, the Cramér-Rao bounds are good predictors for the estimator's variance, again regardless of model violations. Particularly, if there is some prior idea of the true parameters' ranges, these bounds are extremely helpful for effective study design (with respect to sample size and properties of the genetic/molecular markers) to achieve a certain precision goal. Since the estimator's variance is close to the Cramér-Rao bounds, there is not much room for improvement for the estimator. However, the estimator might be generalized to include information from several genetic markers at the same time. A simple ad-hoc approach to reduce variance would be to average several MOI estimates from different (uncorrelated) markers. This and more sophisticated methods are subject to future work.

The fact that the lineage frequencies' estimates are unbiased and the small variances seem promising to use these estimates for genetic statistics. Note, however, that the coefficient of variation might be rather large for minor lineages, which is not surprising since they are likely either under- or over-represented. Hence, it might be problematic to use such estimates in genetic statistics, which rely on minor allele frequencies. Statistics such as heterozygosity or statistics that use ranked frequency spectra should nevertheless be rather robust.

Unfortunately, the method will not always be applicable. It is required that the data contains at least one detectable super-infection $\sum_{k=1}^{n} N_k > N$ and that no lineage is found in all samples $0 < N_k < N$ for all k. Violations of these requirements cannot be easily resolved by just adding pseudo-counts, because estimates would be very sensitive to the exact details of the adjustment. Nevertheless, we calculated the probability to obtain a sample to which the method is inapplicable. These results can again be used for study design to guarantee with high probability that the method will be applicable.

This issue was demonstrated with a malaria data set from Venezuela [51]. Particularly, the MLE for MOI was obtained from several neutral microsatellite markers as well as their respective allele frequency spectra. The method was not applicable to some markers since super-infections were not observed (irregular data sets). This is not surprising although sample size is moderate ($N = 97$) considering that the samples were taken in an area of low transmission. A sample size of at least $N = 300$ would have been necessary to guarantee regular data with 95% probability for each marker separately. To guarantee with 95% probability that no marker yields irregular data, sample size needs to be at least $N = 400$. Such considerations should be taken into account during study design.

A drawback of the maximum-likelihood-approach studied here is its dependence on the Poisson assumption. This assumption can in principal be relaxed by assuming e.g. a negative binomial distribution for the number of infectious events (infective contacts that cause an infection), which was in fact done by [41]. A negative binomial distribution arises if the number of infections is distributed heterogeneously across the population. More precisely, if the population consists of infinitely many patches, within which the number of infectious events are Poisson distributed and the Poisson parameters across patches follow a Γ-distribution. This is justified for malaria by the queuing model of [56]. However, even if the number of infectious events is negatively-binomially and not Poisson distributed, as suggested by empirical evidence (e.g. [57]), taking into account the fact that not every infectious event is infective, the number of infective events might be accurately approximated by a Poisson distribution. So far, an analytical treatment as presented here and in [43] has not be established under the assumption of a negative binomial distribution and is subject to future research.

Finally, it should be mentioned that an alternative to the frequentist method further investigated here is provided by the Bayesian framework of [13, 14]. Agreement of both approaches would underline the quality of both approaches. However, the comparison lies beyond the scope of the present study.

Supporting information

S1 File. R script. Programming code and implementation of likelihood method as R script.
(R)

S2 File. Documentation. Documentation of the R script (S1 File).
(PDF)

S3 File. Example data. ZIP archive containing example data sets in various formats.
(ZIP)

S4 File. Appendix. Appendix A-F.
(PDF)

S5 File. Additional figures. Additional figures showing detailed results.
(PDF)

Acknowledgments

The author thanks Andrea M. McCollum for sharing the Venezuela data set. The valuable comments of editor Ryan E. Wiegand and the reviewers Don Klinkenberg and Mark D. Preston that helped improving this work are gratefully acknowledged.

Author Contributions

Conceptualization: Kristan Alexander Schneider.

Formal analysis: Kristan Alexander Schneider.

Investigation: Kristan Alexander Schneider.

Methodology: Kristan Alexander Schneider.

Project administration: Kristan Alexander Schneider.

Writing – original draft: Kristan Alexander Schneider.

References

1. Tusting LS, Bousema T, Smith DL, Drakeley C. Chapter Three—Measuring Changes in Plasmodium falciparum Transmission: Precision, Accuracy and Costs of Metrics. Advances in Parasitology. 2014; 84:151–208. http://dx.doi.org/10.1016/B978-0-12-800099-1.00003-X.
2. Wacker M, Turnbull L, Walker L, Mount M, Ferdig M. Quantification of multiple infections of Plasmodium falciparum in vitro. Malaria Journal. 2012; 11(1):180. https://doi.org/10.1186/1475-2875-11-180 PMID: 22646748
3. Matussek A, Stark L, Dienus O, Aronsson J, Mernelius S, Löfgren S, et al. Analyzing Multiclonality of Staphylococcus aureus in Clinical Diagnostics Using spa-Based Denaturing Gradient Gel Electrophoresis. Journal of Clinical Microbiology. 2011; 49(10):3647–3648. https://doi.org/10.1128/JCM.00389-11 PMID: 21832023
4. Balmer O, Tanner M. Prevalence and implications of multiple-strain infections. The Lancet Infectious Diseases. 2011; 11(11):868–878. http://dx.doi.org/10.1016/S1473-3099(11)70241-9. PMID: 22035615
5. Vu-Thien H, Hormigos K, Corbineau G, Fauroux B, Corvol H, Moissenet D, et al. Longitudinal survey of Staphylococcus aureus in cystic fibrosis patients using a multiple-locus variable-number of tandem-repeats analysis method. BMC Microbiology. 2010; 10(1):24. https://doi.org/10.1186/1471-2180-10-24 PMID: 20105324
6. Tognazzo M, Schmid-Hempel R, Schmid-Hempel P. Probing Mixed-Genotype Infections II: High Multiplicity in Natural Infections of the Trypanosomatid, Crithidia bombi, in Its Host, Bombus spp. PLoS ONE. 2012; 7(11):e49137. https://doi.org/10.1371/journal.pone.0049137 PMID: 23145099
7. Pinkevych M, Petravic J, Bereczky S, Rooth I, Färnert A, Davenport MP. Understanding the Relationship Between Plasmodium falciparum Growth Rate and Multiplicity of Infection. The Journal of Infectious Diseases. 2015; 211(7):1121–1127. https://doi.org/10.1093/infdis/jiu561 PMID: 25301957
8. Amoah LE, Nuvor SV, Obboh EK, Acquah FK, Asare K, Singh SK, et al. Natural antibody responses to Plasmodium falciparum MSP3 and GLURP(R0) antigens are associated with low parasite densities in malaria patients living in the Central Region of Ghana. Parasites & Vectors. 2017; 10:395–. https://doi.org/10.1186/s13071-017-2338-7
9. Brown CM, Bidle KD. Attenuation of virus production at high multiplicities of infection in Aureococcus anophagefferens. Virology. 2014; 466(Supplement C):71–81. https://doi.org/10.1016/j.virol.2014.07.023. PMID: 25104555
10. Marasini S, Chang DY, Jung JH, Lee SJ, Cha HL, Suh-Kim H, et al. Effects of Adenoviral Gene Transduction on the Stemness of Human Bone Marrow Mesenchymal Stem Cells. Molecules and Cells. 2017; 40(8):598–605. https://doi.org/10.14348/molcells.2017.0095 PMID: 28835020
11. Nishimura T, Tamizu E, Uno S, Uwamino Y, Fujiwara H, Nishio K, et al. hsa-miR-346 is a potential serum biomarker of Mycobacterium avium complex pulmonary disease activity. Journal of Infection and Chemotherapy. 2017; 23(10):703–708. https://doi.org/10.1016/j.jiac.2017.07.015 PMID: 28827075
12. Alizon S, de Roode JC, Michalakis Y. Multiple infections and the evolution of virulence. Ecology Letters. 2013; 16(4):556–567. https://doi.org/10.1111/ele.12076 PMID: 23347009

13. Galinsky K, Valim C, Salmier A, de Thoisy B, Musset L, Legrand E, et al. COIL: a methodology for evaluating malarial complexity of infection using likelihood from single nucleotide polymorphism data. Malaria Journal. 2015; 14(1):1–9. https://doi.org/10.1186/1475-2875-14-4

14. 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 Computational Biology. 2017; 13(1):e1005348–. https://doi.org/10.1371/journal.pcbi.1005348 PMID: 28125584

15. Frank SA. A Kin Selection Model for the Evolution of Virulence. Proceedings of the Royal Society of London Series B: Biological Sciences. 1992; 250(1329):195–197. https://doi.org/10.1098/rspb.1992.0149 PMID: 1362989

16. Lively C. Evolution of virulence: coinfection and propagule production in spore-producing parasites. BMC Evolutionary Biology. 2005; 5(1):64. https://doi.org/10.1186/1471-2148-5-64 PMID: 16281984

17. Schjørring S, Koella JC. Sub-lethal effects of pathogens can lead to the evolution of lower virulence in multiple infections. Proceedings of the Royal Society of London Series B: Biological Sciences. 2003; 270(1511):189–193. https://doi.org/10.1098/rspb.2002.2233 PMID: 12590759

18. Ben-Ami F, Mouton L, Ebert D. The effects of multiple infections on the expression and evolution of virulence in a Daphnia-endoparasite system. Evolution. 2008; 62(7):1700–1711. https://doi.org/10.1111/j.1558-5646.2008.00391.x PMID: 18384658

19. Schneider KA, Kim Y. An analytical model for genetic hitchhiking in the evolution of antimalarial drug resistance. Theor Popul Biol. 2010; 78(2):93–108. https://doi.org/10.1016/j.tpb.2010.06.005 PMID: 20600206

20. Klein EY, Smith DL, Laxminarayan R, Levin S. Superinfection and the evolution of resistance to antimalarial drugs. Proc Biol Sci. 2012; 279(1743):3834–3842. https://doi.org/10.1098/rspb.2012.1064 PMID: 22787024

21. Straimer J, Gnädig NF, Witkowski B, Amaratunga C, Duru V, Ramadani AP, et al. K13-propeller mutations confer artemisinin resistance in Plasmodium falciparum clinical isolates. Science. 2015; 347 (6220):428–431. https://doi.org/10.1126/science.1260867 PMID: 25502314

22. Winzeler E, Manary M. Drug resistance genomics of the antimalarial drug artemisinin. Genome Biology. 2014; 15(11):544. https://doi.org/10.1186/s13059-014-0544-6 PMID: 25470531

23. Hawkes M, Conroy AL, Kain KC. Spread of Artemisinin Resistance in Malaria. New England Journal of Medicine. 2014; 371(20):1944–1945. https://doi.org/10.1056/NEJMc1410735 PMID: 25390755

24. Ben-Ami F, Routtu J. The expression and evolution of virulence in multiple infections: the role of specificity, relative virulence and relative dose. BMC Evolutionary Biology. 2013; 13(1):97. https://doi.org/10.1186/1471-2148-13-97 PMID: 23641899

25. Poon AFY, Swenson LC, Bunnik EM, Edo-Matas D, Schuitemaker H, van't Wout AB, et al. Reconstructing the dynamics of HIV evolution within hosts from serial deep sequence data. PLoS Comput Biol. 2012; 8(11):e1002753. https://doi.org/10.1371/journal.pcbi.1002753 PMID: 23133358

26. Theron A, Sire C, Rognon A, Prugnolle F, Durand P. Molecular ecology of Schistosoma mansoni transmission inferred from the genetic composition of larval and adult infrapopulations within intermediate and definitive hosts. Parasitology. 2004; 129(Pt 5):571–585. https://doi.org/10.1017/S0031182004005943 PMID: 15552402

27. Lindström I, Sundar N, Lindh J, Kironde F, Kabasa JD, Kwok OCH, et al. Isolation and genotyping of Toxoplasma gondii from Ugandan chickens reveals frequent multiple infections. Parasitology. 2008; 135:39–45. PMID: 17892617

28. Cohen T, van Helden PD, Wilson D, Colijn C, McLaughlin MM, Abubakar I, et al. Mixed-Strain Mycobacterium tuberculosis Infections and the Implications for Tuberculosis Treatment and Control. Clinical Microbiology Reviews. 2012; 25(4):708–719. https://doi.org/10.1128/CMR.00021-12 PMID: 23034327

29. Thanapongpichat S, McGready R, Luxemburger C, Day N, White N, Nosten F, et al. Microsatellite genotyping of Plasmodium vivax infections and their relapses in pregnant and non-pregnant patients on the Thai-Myanmar border. Malaria Journal. 2013; 12(1):275. https://doi.org/10.1186/1475-2875-12-275 PMID: 23915022

30. Awaga KL, Missihoun TD, Karou SD, Djadou KE, Chabi NW, Akati A, et al. Genetic diversity and genotype multiplicity of Plasmodium falciparum infections in symptomatic individuals in the maritime region of Togo. Trop Med Int Health. 2012; 17(2):153–160. https://doi.org/10.1111/j.1365-3156.2011.02913.x PMID: 22074288

31. Kateera F, Nsobya SL, Tukwasibwe S, Mens PF, Hakizimana E, Grobusch MP, et al. Malaria case clinical profiles and Plasmodiumfalciparum parasite genetic diversity: a cross sectional survey at two sites of different malaria transmission intensities in Rwanda. Malaria Journal. 2016; 15:237. https://doi.org/10.1186/s12936-016-1287-5 PMID: 27113354

32. Kobbe R, Neuhoff R, Marks F, Adjei S, Langefeld I, von Reden C, et al. Seasonal variation and high multiplicity of first Plasmodium falciparum infections in children from a holoendemic area in Ghana, West Africa. Trop Med Int Health. 2006; 11(5):613–619. https://doi.org/10.1111/j.1365-3156.2006.01618.x PMID: 16640613

33. Nabet C, Doumbo S, Jeddi F, Konaté S, Manciulli T, Fofana B, et al. Genetic diversity of Plasmodium falciparum in human malaria cases in Mali. Malaria Journal. 2016; 15(1):353. https://doi.org/10.1186/s12936-016-1397-0 PMID: 27401016

34. Mohd Abd Razak MR, Sastu UR, Norahmad NA, Abdul-Karim A, Muhammad A, Muniandy PK, et al. Genetic Diversity of Plasmodium falciparum Populations in Malaria Declining Areas of Sabah, East Malaysia. PLoS ONE. 2016; 11(3):e0152415. https://doi.org/10.1371/journal.pone.0152415 PMID: 27023787

35. Mahdi Abdel Hamid M, Elamin AF, Albsheer MMA, Abdalla AAA, Mahgoub NS, Mustafa SO, et al. Multiplicity of infection and genetic diversity of Plasmodium falciparum isolates from patients with uncomplicated and severe malaria in Gezira State, Sudan. Parasites & Vectors. 2016; 9:362. https://doi.org/10.1186/s13071-016-1641-z

36. Weir W, Karagenç T, Gharbi M, Simuunza M, Aypak S, Aysul N, et al. Population diversity and multiplicity of infection in Theileria annulata. International Journal for Parasitology. 2011; 41(2):193–203. https://doi.org/10.1016/j.ijpara.2010.08.004 PMID: 20833170

37. Sisya TJ, Kamn'gona RM, Vareta JA, Fulakeza JM, Mukaka MFJ, Seydel KB, et al. Subtle changes in Plasmodium falciparum infection complexity following enhanced intervention in Malawi. Acta Trop. 2015; 142:108–114. https://doi.org/10.1016/j.actatropica.2014.11.008 PMID: 25460345

38. Takala SL, Smith DL, Stine OC, Coulibaly D, Thera MA, Doumbo OK, et al. A high-throughput method for quantifying alleles and haplotypes of the malaria vaccine candidate Plasmodium falciparum merozoite surface protein-1 19 kDa. Malaria Journal. 2006; 5:31–31. https://doi.org/10.1186/1475-2875-5-31 PMID: 16626494

39. 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 Neglected Tropical Diseases. 2016; 10(3):e0004526–. https://doi.org/10.1371/journal.pntd.0004526 PMID: 27018585

40. McCollum AM, Schneider KA, Griffing SM, Zhou Z, Kariuki S, Ter-Kuile F, et al. Differences in selective pressure on dhps and dhfr drug resistant mutations in western Kenya. Malar J. 2012; 11:77. https://doi.org/10.1186/1475-2875-11-77 PMID: 22439637

41. Hill WG, Babiker HA. Estimation of Numbers of Malaria Clones in Blood Samples. Proceedings of the Royal Society of London Series B: Biological Sciences. 1995; 262(1365):249–257. https://doi.org/10.1098/rspb.1995.0203 PMID: 8587883

42. Li X, Foulkes AS, Yucel RM, Rich SM. An expectation maximization approach to estimate malaria haplotype frequencies in multiply infected children. Statistical applications in genetics and molecular biology. 2007; 6(1). https://doi.org/10.2202/1544-6115.1321 PMID: 18052916

43. Schneider KA, Escalante AA. A Likelihood Approach to Estimate the Number of Co-Infections. PLoS ONE. 2014; 9(7):e97899. https://doi.org/10.1371/journal.pone.0097899 PMID: 24988302

44. Hastings IM, Smith TA. MalHaploFreq: A computer programme for estimating malaria haplotype frequencies from blood samples. Malaria Journal. 2008; 7(1):130. https://doi.org/10.1186/1475-2875-7-130 PMID: 18627599

45. Wigger L, Vogt JE, Roth V. Malaria haplotype frequency estimation. Stat Med. 2013; 32(21):3737–3751. https://doi.org/10.1002/sim.5792 PMID: 23609602

46. Taylor A, Flegg J, Nsobya S, Yeka A, Kamya M, Rosenthal P, et al. Estimation of malaria haplotype and genotype frequencies: a statistical approach to overcome the challenge associated with multiclonal infections. Malaria Journal. 2014; 13(1):102. https://doi.org/10.1186/1475-2875-13-102 PMID: 24636676

47. Kuk AYC, Li X, Xu J. An EM algorithm based on an internal list for estimating haplotype distributions of rare variants from pooled genotype data. BMC Genet. 2013; 14:82. https://doi.org/10.1186/1471-2156-14-82 PMID: 24034507

48. Ross A, Koepfli C, Li X, Schoepflin S, Siba P, Mueller I, et al. Estimating the numbers of malaria infections in blood samples using high-resolution genotyping data. PLoS One. 2012; 7(8):e42496. https://doi.org/10.1371/journal.pone.0042496 PMID: 22952595

49. Ken-Dror G, Hastings IM. Markov chain Monte Carlo and expectation maximization approaches for estimation of haplotype frequencies for multiply infected human blood samples. Malar J. 2016; 15(1):430. https://doi.org/10.1186/s12936-016-1473-5 PMID: 27557806

50. Assefa SA, Preston MD, Campino S, Ocholla H, Sutherland CJ, Clark TG. estMOI: estimating multiplicity of infection using parasite deep sequencing data. Bioinformatics. 2014; 30(9):1292–1294. https://doi.org/10.1093/bioinformatics/btu005 PMID: 24443379

51. McCollum AM, Mueller K, Villegas L, Udhayakumar V, Escalante AA. Common origin and fixation of Plasmodium falciparum dhfr and dhps mutations associated with sulfadoxine-pyrimethamine resistance in a low-transmission area in South America. Antimicrob Agents Chemother. 2007; 51(6):2085–2091. https://doi.org/10.1128/AAC.01228-06 PMID: 17283199

52. Davison AC. Statistical Models (Cambridge Series in Statistical and Probabilistic Mathematics). Cambridge University Press; 2003.

53. R Core Team. R: A Language and Environment for Statistical Computing; 2014. Available from: http://www.R-project.org/.

54. Nei M. Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics. 1978; 89(3):583–590. PMID: 17248844

55. Mueller I, Schoepflin S, Smith TA, Benton KL, Bretscher MT, Lin E, et al. Force of infection is key to understanding the epidemiology of Plasmodium falciparum malaria in Papua New Guinean children. Proceedings of the National Academy of Sciences of the United States of America. 2012; 109(25):10030–10035. https://doi.org/10.1073/pnas.1200841109 PMID: 22665809

56. Smith D, Hay S. Endemicity response timelines for Plasmodium falciparum elimination. Malaria Journal. 2009; 8(1):87. https://doi.org/10.1186/1475-2875-8-87 PMID: 19405974

57. Kilama M, Smith DL, Hutchinson R, Kigozi R, Yeka A, Lavoy G, et al. Estimating the annual entomological inoculation rate for Plasmodium falciparum transmitted by Anopheles gambiae s.l. using three sampling methods in three sites in Uganda. Malaria Journal. 2014; 13(1):111–. https://doi.org/10.1186/1475-2875-13-111 PMID: 24656206