OXFORD GENETICS

Dcifer: an IBD-based method to calculate genetic distance between polyclonal infections

Inna Gerlovina (ID) ,* Boris Gerlovin (ID) , Isabel Rodríguez-Barraquer (ID) , Bryan Greenhouse (ID)

EPPIcenter Research Program, Division of HIV, ID and Global Medicine, Department of Medicine, University of California, San Francisco, San Francisco, CA 94143, USA

*Corresponding author: EPPIcenter Research Program, Division of HIV, ID and Global Medicine, Department of Medicine, University of California, San Francisco, 1001 Potrero Avenue, San Francisco, CA 94110, USA. Email: inna.gerlovina@ucsf.edu

Abstract

An essential step toward reconstructing pathogen transmission and answering epidemiologically relevant questions from genomic data is obtaining pairwise genetic distance between infections. For recombining organisms such as malaria parasites, relatedness measures quantifying recent shared ancestry would provide a meaningful distance, suggesting methods based on identity by descent (IBD). While the concept of relatedness and consequently an IBD approach is fairly straightforward for individual parasites, the distance between polyclonal infections, which are prevalent in malaria, presents specific challenges, and awaits a general solution that could be applied to infections of any clonality and accommodate multiallelic (e.g. microsatellite or microhaplotype) and biallelic [single nucleotide polymorphism (SNP)] data. Filling this methodological gap, we present Dcifer (Distance for complex infections: fast estimation of relatedness), a method for calculating genetic distance between polyclonal infections, which is designed for unphased data, explicitly accounts for population allele frequencies and complexity of infection, and provides reliable inference. Dcifer's IBD-based framework allows us to define model parameters that represent interhost relatedness and to propose corresponding estimators with attractive statistical properties. By using combinatorics to account for unobserved phased haplotypes, Dcifer is able to quickly process large datasets and estimate pairwise relatedness along with measures of uncertainty. We show that Dcifer delivers accurate and interpretable results and detects related infections with statistical power that is 2–4 times greater than that of approaches based on identity by state. Applications to real data indicate that relatedness structure aligns with geographic locations. Dcifer is implemented in a comprehensive publicly available software package.

Keywords: genetic relatedness; identity by descent; genetic distance; ancestry; *Plasmodium*; polyclonal infection; microhaplotype

Introduction

Monitoring, effective control, and ultimately elimination of malaria can be accelerated by understanding the dynamics of malaria transmission including evaluation of interventions, identification of sources and sinks, and determining the drivers of sustained transmission. Given the substantial genetic diversity of malaria parasites, genomic data have the potential to illuminate important aspects of epidemiology (World Health Organization 2019). Compared to viruses where mutations are the main source of variation and can be used directly to make temporal inferences, reconstructing transmission for recombining organisms with lower mutation rates requires a different approach. Since genetic recombination between malaria parasites occurs in the mosquito during person-to-person transmission, genetic relatedness can provide information on their shared ancestry and therefore transmission epidemiology at relevant timescales. Consequently, pairwise genetic distance as a measure of relatedness between infections may be more useful and detailed for answering epidemiologic questions than metrics based on comparison between populations (Taylor *et al.* 2017; Wesolowski *et al.* 2018; Chang *et al.* 2019; Tessema *et al.* 2019). By assessing how closely related individual infections are, pairwise distance

can also provide answers to questions such as whether particular infections were more likely to have been acquired locally or imported.

Due to coinfection and super-infection, individuals in endemic areas are often infected with multiple genetically distinct clones simultaneously. These polyclonal infections are the rule rather than the exception for *Plasmodium falciparum* in many endemic areas, even in relatively low transmission settings of sub-Saharan Africa (Roh *et al.* 2019); polyclonality may be even more common for *Plasmodium vivax* (Koepfli *et al.* 2011; White *et al.* 2018). Assessing genetic relatedness between polyclonal infections is more complicated both conceptually and methodologically than doing so for individual parasites. Obtaining phased genotypes of individual parasites from polyclonal infections would present a potential solution, but outside of single-cell sequencing this currently requires the use of statistical methods which are computationally intensive and may have limited accuracy in the absence of informative reference genomes, particularly when more than 2 clones are present (Zhu *et al.* 2019). Even with phased genotypes, a unified summary of relatedness might be useful as a distance measure, so that it could be compared across pairs of infections that may be either monoclonal or have

a higher complexity of infection (COI). Incorporating multiallelic genetic data, i.e. diverse loci with more than 2 variants, can improve the estimation of relatedness between monoclonal infections (Taylor *et al.* 2019) and may offer an even greater improvement over biallelic loci for polyclonal infections (Tessema *et al.* 2022). Fortunately, current technologies make it feasible to efficiently amplify and sequence multiple diverse regions of the *Plasmodium* genome, generating multiallelic data for this purpose (Lerch *et al.* 2017; Aydemir *et al.* 2018; Tessema *et al.* 2022; LaVerriere *et al.* 2022).

Much of the epidemiologically useful information contained in relatedness measures lies in detecting shared ancestry; there is therefore interest in estimating the proportion of genomes that are identical due to descent. Currently available methods based on identity by descent (IBD) for *Plasmodia* are developed for monoclonal infections or are adapted from human genetics: hmmIBD is designed for monoclonal infections and can incorporate multiallelic as well as biallelic loci (Schaffner *et al.* 2018); isoRelate is able to accommodate polyclonal infections (Henden *et al.* 2018), but is limited to biallelic loci and has unclear applicability to infections with COI >2 since it is based on the diploid model. With no existing IBD-based methods to infer a degree of shared ancestry from polyclonal infections using multiallelic data, various suboptimal workarounds are generally employed. For example, some studies have attempted to infer a "dominant strain" from polyclonal infections using within host allele frequencies, while others have excluded polyclonal infections from the analysis altogether. Depending on the proportion of infections that are polyclonal, such procedures may grossly underutilize data or introduce bias to the analysis due to informative missingness. Alternatively, a simple identity by state (IBS) approach has been used (Pringle *et al.* 2019; Tessema *et al.* 2019; Atuh *et al.* 2021); it is convenient and fast but has extensive drawbacks as it produces similarity measures that are not easy to interpret and address relatedness only indirectly (Taylor *et al.* 2019).

To fill the methodological gap, we introduce Dcifer (distance for complex infections: fast estimation of relatedness), a method employing IBD to estimate the level of common ancestry between polyclonal samples. It allows for unphased multiallelic data such as microsatellites or microhaplotypes as well as SNPs, explicitly takes into account COI and population allele frequencies, and does not require densely spaced or linked markers. Focusing on interhost relatedness, we developed a working model that allowed us to define an estimator with desirable statistical properties and formal inference. As the method provides a probabilistic solution to the multitude of possible underlying phased genomes, we used a unified mixed radix incrementing combinatorial algorithm for its implementation as a comprehensive R software package (Gerlovina 2022): github.com/EPPIcenter/dcifer (accessed 2022 Aug 25). Finally, we assessed the performance of Dcifer for estimating relatedness between *P. falciparum* infections using simulations and empirical data.

Methods

The working model we developed is designed to address interhost relatedness and includes assumptions that reduce a complex realistic dependence structure to a simpler model that still allows us to formally define and estimate the quantity of interest without introducing significant bias. The main assumptions can be summarized as the absence of linkage disequilibrium and intrahost relatedness. As the observed data are several levels removed from the random variables we are interested in, the likelihood for

the model accounts for various possible unobserved combinations (phased haplotypes) with multinomial-based probabilities and the use of combinatorics; consequently, numerical methods are used to find a maximum likelihood estimate. The likelihood incorporates population allele frequencies and COI of both samples, and the estimation process explicitly accounts for the fact that alleles present in both infections may match by chance (i.e. be identical by state but not by descent).

Consider 2 infections with COI of n_x, n_y, and a panel of T multiallelic markers. At each locus t, $t = 1, \ldots, T$, there is a set $A_t = \{a_{t,1}, \ldots, a_{t,K_t}\}$ of possible alleles. For convenience, we can arbitrarily order the alleles and map them to the corresponding population allele frequencies $\boldsymbol{\pi} = ((\pi_1), \ldots, (\pi_T)) = ((\pi_{t,k})_{k=1}^{K_t})_{t=1}^{T}$. We assume that the underlying population allele frequencies are the same for both infections.

IBD model for 2 haplotypes

To build up a model for relatedness between polyclonal infections, we first consider 2 haplotypes. Let sequences of random variables $X = (X_1, \ldots, X_T)$, $Y = (Y_1, \ldots, Y_T)$ represent these haplotypes, and let $(IBD_1, \ldots, IBD_T)$ be a sequence of independent identically distributed random variables, where $IBD_t \sim Bernoulli(r)$ and parameter r describes the level of relatedness of the 2 haplotypes (Taylor *et al.* 2019). Let $X_t \sim P_t$, where P_t is a categorical distribution with values in A_t and corresponding probabilities $\pi_{t,1}, \ldots, \pi_{t,K_t}$, $X_t \perp IBD_t$; let Y_t be a random variable such that

$$\begin{cases} Y_t = X_t & \text{if } IBD_t = 1 \\ Y_t \sim P_t & \text{if } IBD_t = 0. \end{cases}$$

Note that X_t and Y_t are interchangeable in this setup, and the joint distribution of X_t and Y_t (marginal and conditional on IBD_t) would not change if they were switched. While IBD_t are i.i.d., (X_t, Y_t) are marginally independent but not identically distributed since P_t is different for each t.

We can also define a random variable $IBS_t \equiv \mathbb{1}(X_t = Y_t)$, with $P(IBS_t = 1) = P(IBD_t = 1) + P(X_t = Y_t, IBD_t = 0)$. In this model, realizations of X and Y could be observed (e.g. if they represent monoclonal infections and there is no genotyping error) but IBD's are unobservable. In contrast, IBS_t are directly observed if X and Y are observed.

Working model for polyclonal infections

Let an $n_x \times T$ matrix $\mathbf{X}$ and $n_y \times T$ matrix $\mathbf{Y}$ represent 2 polyclonal infections with COI of n_x and n_y, with rows of the matrices referring to haplotypes and columns to loci. Thus, $X_i = (X_{i,1}, \ldots, X_{i,T})$, is an i'th haplotype of the first infection, and a column $X_{1,t}, \ldots, X_{n_x,t}$ is a sequence of random variables with values in A_t representing alleles for all the haplotypes at a locus t. Let $S_{x,t} = \{X_{1,t}, \ldots, X_{n_x,t}\}$ denote a multiset (a collection of elements that are not necessarily distinct) of unordered elements of a t'th column of $\mathbf{X}$ and let $U_{x,t} = Supp(S_{x,t}) = \{a_k : a_k \in S_{x,t}\}$ be a set of unique elements in that column; $S_{y,t}$, $U_{y,t}$ are defined similarly. For realizations of $S_{x,t}$ and $U_{x,t}$ we will use notation $s_{x,t}$ and $u_{x,t}$ ($s_{x,t}$ and $s_{y,t}$ are not observed, but $u_{x,t}$ and $u_{y,t}$ are). The model assumes no genotyping error, and the sequences of sets $\boldsymbol{u}_x = (u_{x,1}, \ldots, u_{x,T})$ and $\boldsymbol{u}_y = (u_{y,1}, \ldots, u_{y,T})$ are observed data, for which Dcifer is designed.

There are $\binom{n_x + n_y}{2}$ pairs of malaria strains that can be related. To differentiate IBD_t and parameters of their distributions for different pairs, let $IBD_{x_i,x_j,t}$, $IBD_{y_i,y_j,t}$, and $IBD_{x_i,y_j,t}$ refer to a pair within first infection, a pair within second infection, and a

between-host pair, respectively, and, similarly, let r_{x_i,x_j}, r_{y_i,y_j}, and r_{x_i,y_j} denote corresponding relatedness parameters. If we are only interested in between-host relatedness (which may be the case for many practical applications), we might formulate the goal as "estimating interhost relatedness adjusted for intrahost relatedness," which would condense $\binom{n_x + n_y}{2}$ parameters into some lower-dimensional summary. Usefulness of adjusting for intrahost relatedness can be illustrated by considering a case where an extra haplotype X_{n_x+1}, very closely related to an existing one (say, X_1, with $r_{x_1,x_{n_x+1}} = 0.99$), is added to one of the infections. That would result in essentially doubling X_1's contribution $\sum_{j=1}^{n_y} r_{x_1,y_j}$ to the sum $\sum_{i=1}^{n_x} \sum_{j=1}^{n_y} r_{x_i,y_j}$ of all interhost relatedness parameters, as well as increasing COI (recall that n_x is defined as a number of distinct haplotypes). With such goal as our scientific question, we introduce a simplifying assumption of no intrahost relatedness, which projects a realistic model of unconstrained intrahost and interhost relatedness parameters onto a much smaller model space and allows us to make the problem tractable while aiming to arrive at the same summary estimate as we would if we were able to estimate all the parameters in a bigger model.

For each pair of strains in 2 infections, e.g. i'th strain in the first sample and j'th in the second, let $X_{i,t}$, $Y_{j,t}$, and $IBD_{x_i,y_j,t}$ be the random variables as defined in *IBD Model for 2 Haplotypes*. Then, for the working model for polyclonal infections, we introduce the following assumptions:

1) $r_{x_i,x_j} = 0$ for all $i,j = 1,\ldots,n_x$, $i \neq j$, $r_{y_i,y_j} = 0$ for all $i,j = 1,\ldots,n_y$, $i \neq j$ (no intrahost relatedness) and
2) $IBD_{x_i,y_j,t} \perp IBD_{x_k,y_l,t}$ if $i \neq k$ or $j \neq l$ for all $t = 1,\ldots,T$ (all interhost IBD variables are independent at a given locus).

An important implication of these two assumptions is that any strain in one sample can be related to at most one strain in another: $\sum_{j=1}^{n_y} \mathbb{1}(r_{x_i,y_j} > 0) \leq 1 \forall i = 1,\ldots,n_x$ and $\sum_{i=1}^{n_x} \mathbb{1}(r_{x_i,y_j} > 0) \leq 1 \forall j = 1,\ldots,n_y$. This can be proven by contradiction: since $IBD_{x_1,y_1,t} \perp IBD_{x_2,y_1,t}$, $P(IBD_{x_1,x_2,t} = 1) \geq P(IBD_{x_1,y_1,t} = 1, IBD_{x_2,y_1,t} = 1)$ $= r_{x_1,y_1} r_{x_2,y_1}$. If $r_{x_1,y_1} > 0$ and $r_{x_2,y_1} > 0$, then $r_{x_1,x_2} > 0$, which contradicts assumption 1. For further discussion on the model assumptions, see Supplementary Section 1 in File_1.

Since we can order strains within an infection arbitrarily, and in light of the constraints of the model, we order the haplotypes in 2 infections in such a way that X_1 can only be related to Y_1, X_2 to Y_2, and so on (Fig. 1). In addition, we introduce M—the number of strain pairs that can be related, $M = 1,\ldots,\min(n_x,n_y)$. Then, for brevity, we suppress some of the subscripts and use $r_1,\ldots,r_M$ for $r_{x_1,y_1},\ldots,r_{x_M,y_M}$ and $IBD_{1,t},\ldots,IBD_{M,t}$ for $IBD_{x_1,y_1,t},\ldots IBD_{x_M,y_M,t}$ (note that parameters for all the other IBD variables are zero). The goal of Dcifer is to estimate parameters of the joint distribution of $IBD_{1,t},\ldots,IBD_{M,t}$. Let $\boldsymbol{r} = (r_1,\ldots,r_M)$ denote an estimand, and $\hat{\boldsymbol{r}} = (\hat{r}_1,\ldots,\hat{r}_M)$—its maximum likelihood estimator (MLE):

$$\hat{\boldsymbol{r}} = \underset{\boldsymbol{r} \in [0,1]^M}{\arg\max}\ L(\boldsymbol{r}; \boldsymbol{u}_x, \boldsymbol{u}_y, n_x, n_y, \boldsymbol{\pi}). \tag{1}$$

At each locus t, the likelihood $L(\boldsymbol{r}; \boldsymbol{u}_x, \boldsymbol{u}_y, n_x, n_y, \boldsymbol{\pi})$ needs to account for all the possible combinations of nonunique alleles in both samples (multiple haplotypes will have the same allele if COI is greater than the number of unique alleles). For one sample, this is done by considering a set of all multisets with given support and cardinality (all the $S_{x,t}$ that could have produced $U_{x,t}$, see Fig. 1); we denote a set of all multisets $s_{x,t}$ such that $Supp(s_{x,t}) = u_{x,t}$ and $|s_{x,t}| = n_x$ by $Q_{x,t}$. $P(S_{x,t} = s_{x,t})$ can be calculated using a probability mass function of a multinomial distribution: the number of permutations of $s_{x,t}$ is equal to a multinomial coefficient ("assigning" alleles in s_x to strains, or going from $S_{x,t}$ to $(X_{1,t},\ldots,X_{n_x,t})$), and allele frequencies correspond to event probabilities. Multiplicities of the multiset's elements $a_{t,k} \in A_t$, or the numbers of strains having the same allele, are multinomial random variables. Adopting a short notation for this key component
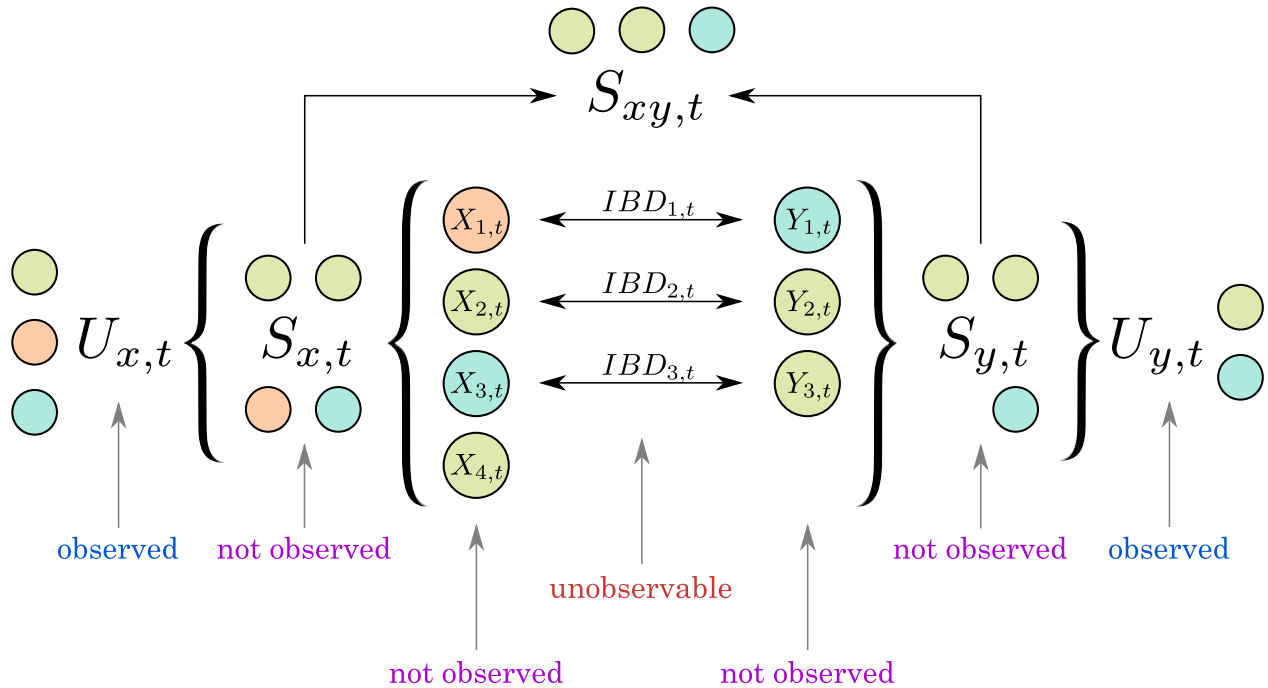

Fig. 1. Working model presented at a single locus t: an example featuring $n_x = 4$, $n_y = 3$, and $M = 3$. Colors of the circles represent alleles; 2 clones in each infection have the same allele. $S_{xy,t} = S_{x,t} \cap S_{y,t}$ is a multiset of shared (nonunique) alleles at a locus t.

of the likelihood, let $g(s^{(n)}; n, (\pi))$ denote a probability mass function for a multinomial distribution $Multinom(n, \pi_1, \ldots, \pi_K)$, where $s^{(n)}$ is a multiset of cardinality n ($|s^{(n)}| = n$) with elements from K categories, and $(\pi) = (\pi_1, \ldots, \pi_K)$ are probabilities for these categories; set $g(\emptyset; 0, (\pi)) = 1$. Next, for given $s_{x,t}$ and $s_{y,t}$, we divide their elements into 3 groups: shared alleles that are identical by descent (say $s^{(m)}$), remaining alleles in $s_{x,t}$ ($s_{x,t} \backslash s^{(m)}$), and remaining alleles in $s_{y,t}$ ($s_{y,t} \backslash s^{(m)}$). The probability of each of these multisets is similarly calculated using multinomial distributions. Supplementary Section 2 in File_1 provides more details and builds up the likelihood from $M = 1$ and $M = 2$. For a general case,

$$L(\boldsymbol{r}; \boldsymbol{u_x}, \boldsymbol{u_y}, n_x, n_y, \boldsymbol{\pi}) = \prod_{t=1}^{T} \sum_{s_x \in Q_{x,t}} \sum_{s_y \in Q_{y,t}} \sum_{IBD_{1,t}=0}^{1} \cdots \sum_{IBD_{M,t}=0}^{1} \left[\prod_{\{i:IBD_{i,t}=1\}} r_i \right]$$
$$\left[\prod_{\{j:IBD_{j,t}=0\}} (1 - r_j) \right] P\left(S_{x,t} = s_x, S_{y,t} = s_y \mid \sum_{l=1}^{M} IBD_{l,t} \right),$$

$$(2)$$

where

$$P\left(S_{x,t} = s_x, S_{y,t} = s_y \mid \sum_{l=1}^{M} IBD_{l,t} = m \right)$$
$$= \begin{cases} 0 & \text{if } m > |s_{xy}| \\ \sum_{s^{(m)} \subseteq s_{xy}} g(s^{(m)}; m, (\pi_t)) \\ \quad \times g(s_x \backslash s^{(m)}; n_x - m, (\pi_t)) \\ \quad \times g(s_y \backslash s^{(m)}; n_y - m, (\pi_t)) & \text{otherwise} \end{cases}$$

and $s_{xy} = s_x \cap s_y$.

When $r_1 = r_2 = \cdots = r_M = r$, the likelihood reduces to

$$L(\boldsymbol{r}; \boldsymbol{u_x}, \boldsymbol{u_y}, n_x, n_y, \boldsymbol{\pi}) = \prod_{t=1}^{T} \sum_{s_x \in Q_{x,t}} \sum_{s_y \in Q_{y,t}} \sum_{m=0}^{M} P_{Binom}(m; M, r)$$
$$P\left(S_{x,t} = s_x, S_{y,t} = s_y \mid \sum_{l=1}^{M} IBD_{l,t} = m \right),$$

$$(3)$$

where $P_{Binom}(m; M, r) = \binom{M}{m} r^m (1 - r)^{(M-m)}$.

While $IBD_{1,t}, \ldots, IBD_{M,t}$ are independent, $\hat{r}_1, \ldots, \hat{r}_M$ are not. This dependence stems from the fact that we do not observe ordered alleles at each locus (or, in other words, phased haplotypes). That also provides intuition for why $r_{total} = \sum_{i=1}^{M} r_i$ is estimated more accurately than individual r_i's : estimating $\sum_{i=1}^{M} IBD_{i,t}$ at a locus t is easier than estimating an actual binary sequence $(IBD_{1,t}, \ldots, IBD_{M,t})$. Another useful observation is that the order of parameter values in $\boldsymbol{r}$ does not affect the value of $L(\boldsymbol{r}; \boldsymbol{u_x}, \boldsymbol{u_y}, n_x, n_y, \boldsymbol{\pi})$, which can be taken into account when the likelihood is evaluated over a grid of $\boldsymbol{r} \in [0, 1]^M$.

Implementation

Calculating the likelihood in (2) requires solving a number of combinatorial problems: finding all the collections of nonunique alleles at a locus that are concordant with observed alleles and COI, finding all the multisets included in a given multiset of shared nonunique alleles, and finding all the possible binary sequences with given constraints for IBD variables. These problems are solved with a unified mixed radix incrementing algorithm (https://github.com/innager/mirsa/tree/v1.0.0; accessed 2022 Aug 25) (Supplementary File_2), which is an extension of an algorithm to generate all n-tuples in (Knuth 2011). As the calculation traverses the combinations described above, multiple

$\boldsymbol{r} = (r_1, \ldots, r_M)$ sequences can be processed at each step, and thus the likelihood for a range of parameter values can be calculated in a single pass. With bounded parameter space, this allows for an efficient way to find MLE by simply calculating the likelihood for an M-dimensional grid of a desired coarseness. The resulting log-likelihood curve or surface can also be useful for inference, especially for procedures based on a likelihood ratio approach, such as testing various hypotheses or determining confidence regions. For a special case of $M = 1$, the log-likelihood can be calculated using Supplementary Equation (2) in File_1, which also admits fast calculation of the score and consequently numerical methods of solving the likelihood equation (Newton's method adapted for bounded parameter space is used in the package).

Inference

Along with an estimate of $\boldsymbol{r}$, Dcifer provides a log-likelihood function, which can serve as a basis for various inferential procedures (for some intuition on the shape of that log-likelihood function, the effect of COI and population allele frequencies on it, and implications for the inference, see Supplementary Section 3 in File_1). In our model, sample size is T, but different loci do not provide the same amount of information (recall that $(X_{i,t}, Y_{j,t})$, $t = 1, \ldots, T$ are independent but not identically distributed); their contribution can be associated with different measures, e.g. heterozygosity. Given these measures and the complexity of the estimator, methods relying on asymptotic approximations should be approached cautiously; still, as the sample size increases, precision of estimation increases as well.

For hypothesis testing and confidence intervals (CI)/regions, we consider common inferential approaches as applied to Dcifer: asymptotic normality, likelihood-ratio statistics, and resampling methods. There are common challenges that affect all 3 approaches: bounded parameter space $[0, 1]^M$ with edge cases not only included but conceptually important, such as a null hypothesis $H_0 : r_1 = \cdots = r_M = 0$ of infections being unrelated; for other cases, sampling distributions for different (even neighboring) parameter values on the interior of the support could be quite different for panels with even fairly large T. Still, some approaches might be better suited for Dcifer, and some may be chosen on the basis of convenience and computational efficiency. For Wald-type CI, observed Fisher information can be easily calculated numerically (and arguably preferred to expected Fisher information; Efron and Hinkley 1978); likelihood-ratio-based CI, while asymptotically equivalent to Wald's (1943), are more robust as they are invariant to parameter transformation that could be used to make the log-likelihood function approximately quadratic at MLE (Beale 1960; Cox and Hinkley 1979, pp. 342–343; Cook and Weisberg 1990; Vander Wiel and Meeker 1990; Meeker and Escobar 1995). Resampling methods include bootstrap and generating a null distribution for hypothesis testing. While there are many advantages this approach provides for finite sampling distributions not yet approaching normality, there is a caveat: if a centered sampling distribution at MLE is not close enough to that at the true value, the inference will be problematic. In addition, inverting quantiles of a bootstrap distribution for CI endpoints can lead to violating the bounds of the parameter space (see Supplementary Section 4 in File_1), as can Wald CI. Simulated null distributions do not suffer from this problem but still rely on various assumptions and might be sensitive to misspecifications as demonstrated in the Results. In contrast, likelihood-ratio confidence regions respect parameter bounds and do not require any additional model assumptions for hypothesis testing.

Likelihood-ratio-based inference is based on Wilks' theorem ([Wilks 1938](#)) and uses the likelihood ratio test statistic, which in the context of Dcifer hypothesis testing with $H_0 : \boldsymbol{r} = \boldsymbol{r}_0$ can be written as $2\,log(L(\hat{\boldsymbol{r}}; \boldsymbol{u_x}, \boldsymbol{u_y}, n_x, n_y, \boldsymbol{\pi})/L(\boldsymbol{r}_0; \boldsymbol{u_x}, \boldsymbol{u_y}, n_x, n_y, \boldsymbol{\pi})) = 2(\ell(\hat{\boldsymbol{r}}) - \ell(\boldsymbol{r}_0))$, where $\ell(\boldsymbol{r}) = log(L(\boldsymbol{r}; \cdot))$, and is approximated by chi-squared distribution $\chi^2(M)$ with M degrees of freedom. The approximate $1 - \alpha$ confidence region consists of the values

$$\{\tilde{\boldsymbol{r}} : \ell(\tilde{\boldsymbol{r}}) \geq \ell(\hat{\boldsymbol{r}}) - 1/2\, q_{M,1-\alpha}\},$$

where $q_{M,1-\alpha}$ is a $(1 - \alpha)$'th quantile of $\chi^2(M)$. As the Wilk's theorem does not apply to border cases, we specifically address these important cases (for which the likelihood-ratio test is still the most powerful; [Neyman and Pearson 1933](#)) and compare the corresponding distribution of the likelihood ratio statistic with $\chi^2(M)$ that no longer approximates it. First, the test at the boundaries is 1-sided while the chi-squared distribution implies 2-sided tests. Accounting for that would mean dividing a *P*-value obtained from the chi-squared distribution by 2 or finding a corresponding critical value for the significance level α. Second, it turns out that even with this adjustment, the resulting *P*-value is still somewhat conservative, and, as shown in the *Results*, the method has excellent error rate control.

Estimating the number of related strain pairs and $\boldsymbol{r}_{\text{total}}$

Parameters of the working model include n_x, n_y, and $\boldsymbol{\pi}$. M, the length of $\boldsymbol{r}$, can be considered a nuisance parameter. In addition, let $M' = \sum_{i=1}^{M} \mathbb{1}(r_i > 0)$ be a number of positively related strain pairs; unlike M, M' can be a quantity of interest. The estimator $\hat{\boldsymbol{r}}$ in [Equation (1)](#) assumes that the model is constrained by given values of n_x, n_y, $\boldsymbol{\pi}$, and M; the likelihood is calculated using these values. However, while n_x, n_y, and $\boldsymbol{\pi}$ are "external" to $\boldsymbol{r}$ and are provided or obtained through other processes, M is inherent to relatedness between 2 infections. Thus, here, we consider a less constrained model where M is not given. In this case, a trivial solution to estimating $\boldsymbol{r}$ would be to set $M = \min(n_x, n_y)$ since $\boldsymbol{r}$'s associated with different $M' \leq M \leq \min(n_x, n_y)$ will only differ in the number of zeros ($r_i = 0$). If we want to estimate M' or $r_{\text{total}} = \sum_{i=1}^{M} r_i$, which are functions of $\boldsymbol{r}$, they can be similarly obtained from $\hat{\boldsymbol{r}}$ as $\hat{M}' = \sum_{i=1}^{M} \mathbb{1}(\hat{r}_i > 0)$ and $\hat{r}_{\text{total}} = \sum_{i=1}^{M} \hat{r}_i$.

In practical applications, this trivial solution can incur high computational cost for higher $\min(n_x, n_y)$, and therefore we propose alternative estimators $\tilde{M}'$, $\tilde{\boldsymbol{r}}$, and $\tilde{r}_{\text{total}}$ that use an iterative procedure with underlying calculation of $\hat{\boldsymbol{r}}$ at each step. The first step is to set $M = 1$ and calculate $\hat{\boldsymbol{r}}$, then at each consecutive step increment M and recalculate $\hat{\boldsymbol{r}}$ until it contains one zero ($\sum_{i=1}^{M} \mathbb{1}(\hat{r}_i = 0) = 1$) or until $M = \min(n_x, n_y)$. Accept $\hat{\boldsymbol{r}}$ obtained at the final step as $\tilde{\boldsymbol{r}}$, with $\tilde{r}_{\text{total}} = \sum_{i=1}^{M} \tilde{r}_i$ and $\tilde{M}' = \sum_{i=1}^{M} \mathbb{1}(\tilde{r}_i > 0)$. If r_{total}, rather than $(r_1, \ldots, r_M)$, is of main interest, the computation time can be cut even further by assuming $r_1 = \cdots = r_M = r$ and using [Equation (3)](#) to calculate the likelihood. In this case, we propose yet another set of estimators $\tilde{M}'_{eq}$ and $\tilde{r}_{\text{total},eq}$, where $\hat{r}$ is calculated for all $M = 1, \ldots, \min(n_x, n_y)$, and $\tilde{M}'_{eq}$ is the value of M that produced the highest maximum likelihood. Then $\tilde{r}_{\text{total},eq} = \tilde{M}'_{eq} \hat{r}'$, where $\hat{r}'$ is an MLE at $\tilde{M}'_{eq}$.

Simulations and comparison with an IBS metric

Our simulations are based on previously published SNP and microhaplotype panels ([Daniels *et al.* 2008](#); [Tessema *et al.* 2022](#); [Jacob *et al.* 2021](#)); allele frequencies for these panels were obtained from previously analyzed empiric datasets. The SNP panels have 23 and 101 loci with 2 alleles per locus, and the microhaplotype panel has 91 loci with the number of alleles at each locus ranging between 3 and 95. To assess the performance of Dcifer with varying number of multiallelic loci, we used allele frequencies for the 91-loci panel and repeated them for consistency (thus creating synthetic 182-loci, 273-loci panels, and so on).

To include genotyping errors in simulations, we devised a "miss-and-split" model with parameters ϵ and λ:

1) False negatives: one of k present alleles (drawn with probabilities $1/k$) has zero probability of being missed; the remaining $k - 1$ alleles can be missed with probability ϵ. Let K be a number of alleles remaining present after this step; then $E(K) = 1 + (k - 1)(1 - \epsilon)$.

2) False positives: draw a number $N_{\text{add}} \sim Pois(\lambda)$ of added alleles ("splitting" event) for each nonmissing allele; subsequently draw N_{add} alleles from $K - 1$ alleles with replacement. In the final "observed" data, an allele is considered present if selected by at least one of the splitting events.

Note that $P(N_{\text{add}} \geq 2)$ is very small for reasonably small λ's.

For analysis procedures that involve estimating COI and population allele frequencies prior to Dcifer, we used naïve COI estimation with a locus rank c that depended on the number of loci (COI determined by a locus with c'th greatest number of detected alleles) for multiallelic panels, and THE REAL McCOIL method for biallelic SNP panels ([Chang *et al.* 2017](#)). Population allele frequencies were estimated from simulated datasets of 400 samples, where relatedness was induced in 10% of the pairs and COI for the rest of the samples was generated with truncated Poisson distribution with parameter $\lambda = 3$. We used COI-adjusted estimation (see [Supplementary Section 5 in File_1](#)), which is important for polyclonal infections; failure to adjust for COI can lead to overestimating heterozygosity and, consequently, relatedness parameters. Most simulations consisted of 10,000 pairs of related infections for a given COI and $\boldsymbol{r}$ combination.

To compare Dcifer performance with an IBS approach, we used the Jaccard similarity coefficient ([Jaccard 1912](#)) as an example since it is a commonly used statistic, which is conceptually simple and fast, and which performs similarly to the other IBS measures we have considered. For this calculation, loci data for a sample (each locus represented by a binary sequence with elements indicating if an allele has been detected or not) were concatenated into a single binary sequence of length $\kappa = \sum_{t=1}^{T} K_t$, where K_t is the number of possible alleles at a locus t. Then Jaccard similarity $J(\boldsymbol{v}, \boldsymbol{w})$ between 2 sequences $\boldsymbol{v} = (v_1, \ldots, v_\kappa)$ and $\boldsymbol{w} = (w_1, \ldots, w_\kappa)$ that represent 2 infections was calculated as

$$J(\boldsymbol{v}, \boldsymbol{w}) = \frac{\sum_{i=1}^{\kappa} \mathbb{1}(v_i + w_i = 2)}{\sum_{i=1}^{\kappa} \mathbb{1}(v_i + w_i > 0)}.$$

Results

The main goal of Dcifer is to estimate parameters describing relatedness between infections, and this estimation requires values of the other parameters in the model. These external parameters represent COI (n_x, n_y) and population allele frequencies ($\boldsymbol{\pi}$), which can be known (e.g. in simulations), estimated from data, or otherwise specified. Dcifer is implemented in a software package that takes raw data on the alleles detected at each locus (biallelic or multiallelic) in *Plasmodium* infections, allowing for missing data, along with COI and population allele frequencies. In simulations, we assess the performance of Dcifer when data have no

genotyping error and COI and π are known, as well as in the presence of genotyping error with COI and π estimated from these data. We also evaluate how sensitive Dcifer is to misspecification of these external quantities and to assumption violations. We start with a case when only one pair of strains can be related ($M = 1$), since it can be used to quickly identify related infections in a large dataset, and later proceed to the general case. Finally, we apply Dcifer to analysis of real data, where COI and allele frequencies are estimated from the data.

Dcifer produces accurate and interpretable estimates of relatedness

Unlike IBS metrics that simply measure similarity between infections comparing detected alleles, Dcifer aims to produce more interpretable results by estimating parameters that represent IBD and thus separating shared ancestry and chance as underlying reasons for alleles matching between 2 infections. To evaluate the performance of this method in comparison to IBS approach (we used Jaccard similarity coefficient as an example), we simulated genetic data for infection pairs with different degrees of relatedness (induced on a single pair of strains between 2 infections) and COI, based on previously

published SNP and microhaplotype panels (Daniels *et al.* 2008; Tessema *et al.* 2022; Jacob *et al.* 2021). Across various values of COI, Dcifer estimates were concentrated around the true values of the parameter while IBS results were not (Fig. 2, shown for a panel of 91 microhaplotypes with no genotyping error and known COI and allele frequencies). As COI increased, Dcifer estimates became more variable but remained centered around the true values and maintained some degree of separation, whereas IBS results shifted and overlapped considerably more. For these simulations, separation between results for completely unrelated ($r = 0$) and related infections, quantified in receiver operating characteristic (ROC) curves, indicated considerable gain in accuracy by Dcifer compared to the IBS metric across the range of COI, especially for lower degrees of relatedness (Supplementary Fig. 1 in File_1). For example, Dcifer estimates of sibling-level relatedness ($r = 0.5$) remained readily distinguishable from those of unrelated infections even for fairly high COI.

When the task of detecting related infections is approached in practice, there are additional issues to be considered because neither COI nor population allele frequencies are known, genetic data often contain genotyping errors, and the extent of these
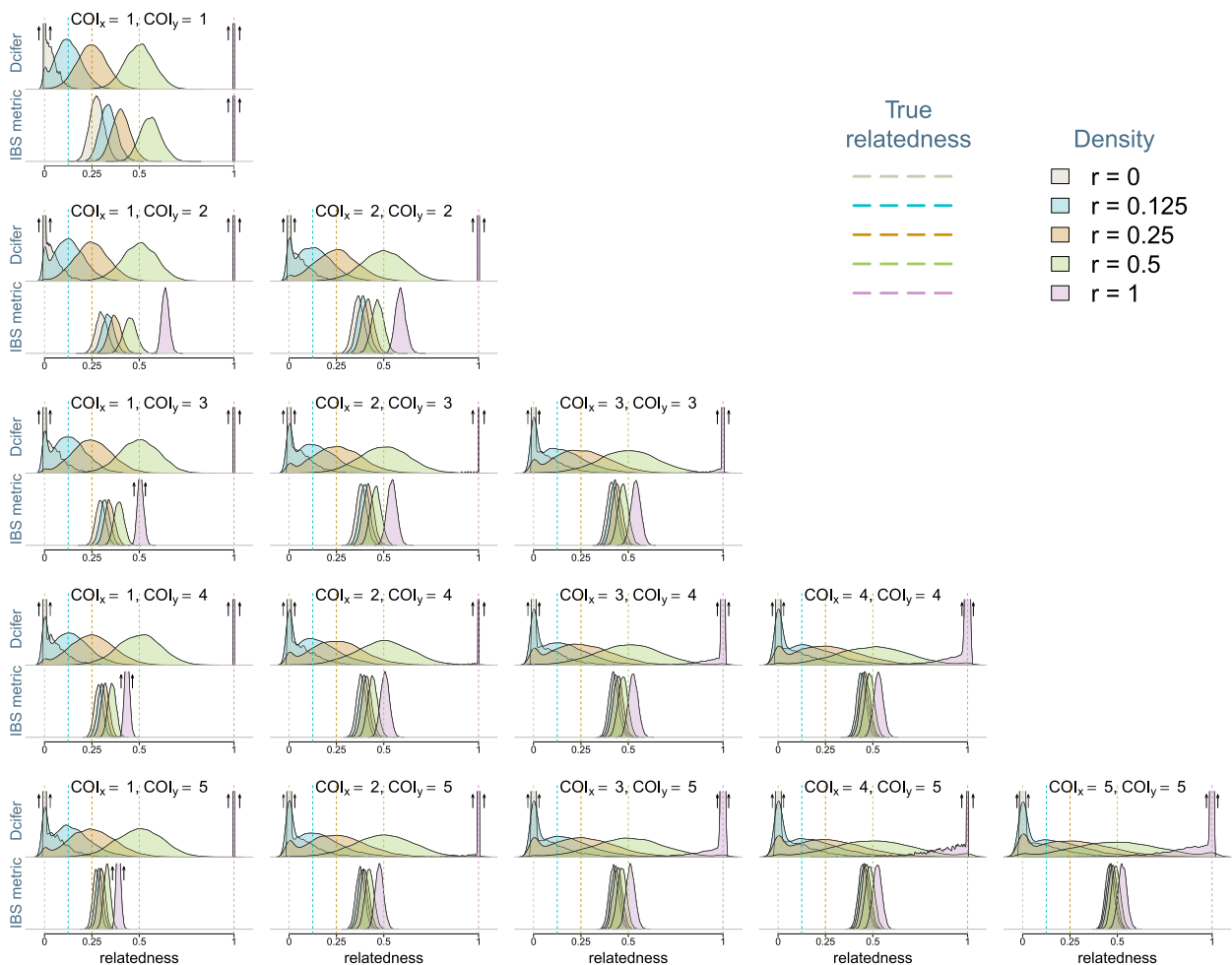
Fig. 2. Densities of Dcifer relatedness estimator $\hat{r}$ and IBS similarity metric results obtained from data simulated using a panel of 91 microhaplotypes (Tessema *et al.* 2022). Simulations were performed for 5 values of r and for COI combinations ranging between 1 and 5; true values of COI and population allele frequencies were used for Dcifer. Upward arrows indicate highly concentrated distributions with density values extending above the plot range (note: y-axis scales are different for the 2 methods).

errors is unknown. Supplementary Fig. 2 in File_1 illustrates how results changed when genotyping error was included in the simulations, and estimates of COI and population allele frequencies, and not their true values, were used as inputs to Dcifer. IBS results shifted to the left more or less uniformly; since distributions for different values of *r* were so tightly concentrated and close together, apparently small shifts were significant compared to the differences between these distributions. Dcifer estimates also shifted to the left, but, relative to differences in sampling distributions, the shifts were smaller than those for IBS. The fact that the shifts were more pronounced for larger values of *r* is explained by genotyping errors breaking up some of the relatedness between infections.

Greater power of hypothesis tests using Dcifer vs. IBS

One way of detecting related infections along with a measure of uncertainty (e.g. *P*-value) is to compare Dcifer relatedness estimates or IBS similarity results with their corresponding null distributions ($H_0 : r = 0$), which can in theory be obtained by simulating a large number of unrelated infections. To evaluate the performances of Dcifer and the IBS metric, we calculated false-positive rates (FPR) and power of tests with significance level $\alpha = 0.05$ for different types of genetic data across a range of COI. Genetic data were simulated with genotyping errors; they were incorporated into simulated "null" distributions as well. Distributions for $r = 0$ are different for different COI and therefore a separate null distribution was generated for each COI pair combination; the effect of COI on such distributions was substantial for IBS (Supplementary Fig. 3 in File_1). Relatedness estimates for each pair of infections were then compared to a rejection cut-off determined by a null distribution corresponding to their estimated COI, and FPR and statistical power were subsequently calculated. In addition to the complexity and computational costs associated with generating a null distribution, this approach relies on a number of assumptions such as COI, allele frequencies, and the error model and its parameters, which are all subject to misspecification.

As a welcome alternative, Dcifer offers another inferential approach based on the likelihood ratio, which does not require any additional information (i.e. does not require generating a null distribution) and has essentially no computational overhead. Figure 3 compares hypothesis testing results for IBS, using simulated reference distributions, and Dcifer, using likelihood-ratio *P*-values adjusted for 1-sided tests. For both methods, FPR was mostly at or below the nominal significance level α across different simulations of COI and genotyping panels, with Dcifer close to α. Statistical power, however, varied considerably. As expected, higher values of relatedness were detected with greater power, increasing the number or diversity of loci increased power, and higher COI led to lower power. Across all simulations, Dcifer consistently demonstrated greater power to detect related infections than the IBS metric, with differences particularly notable for polyclonal infections. For example, with a 91 microhaplotype panel, the power to detect half-siblings ($r = 0.25$) in a pair of infections with COI of 2 was 0.81 for Dcifer and 0.43 for the IBS metric; with 455 microhaplotypes and COI of 5 that power was 0.88 and 0.22, respectively. Results for an alternative scenario when α is a function of COI, which might be useful if error rate control on the scale of parasite strain pairs rather than infection pairs is desired, are presented in Supplementary Fig. 4 in File_1. While Supplementary Figs. 1 and 2 in File_1 would suggest that there is still some separation between distributions for different

values of *r* for the IBS metric results, which would be expected to improve with increasing the number of loci, its performance was remarkably poor, having very low power for larger COI and $r < 0.5$ even with highly informative panels. This reflects the fact that for tightly concentrated distributions of IBS results, the difference between cut-offs associated with different assumed null distributions is critical, and consequently, misspecification of COI or an error process had a deleterious effect on either FPR or power (Supplementary Fig. 3 in File_1). The likelihood-ratio-based approach performed very similarly to the one based on null distributions for Dcifer, evidencing this as a preferred approach for the reasons described above (Supplementary Fig. 5 in File_1).

Dcifer provides likelihood-ratio-based CI

The Dcifer likelihood-ratio-based approach allows for calculating *M*-dimensional confidence regions (where *M* is the number of related pairs)—or, in a case when only one pair of strains is assumed to be related between 2 infections, CI. Figure 4 shows CI's for a range of true *r* values and COI. Infections were simulated using microhaplotype panels with various number of loci. As expected, CI's were narrower for panels with more loci. In general, the intervals were narrower near endpoints ($r = 0$ and $r = 1$) and wider in the midrange. Interestingly, the least COI in the pair ($\min(n_x, n_y)$) had a greater effect on the CI than the sum of COI ($n_x + n_y$); this can be seen in more rapid widening of the intervals from left to right than from top to bottom of the figure. With large numbers of diverse loci, CI stayed narrow even for higher COI. Coverage for these CI was around $1 - \alpha = 0.95$, and consistently higher for endpoints, indicating that CI for these endpoints were conservative, even taking into account the 1-sided nature of such intervals (Supplementary Fig. 6 in File_1; also demonstrated by FPR in Fig. 3).

Allowing for multiple pairs of strains to be related

So far we have only presented results for a single related pair of strains between 2 infections ($M = 1$) regardless of COI. When we allow that multiple pairs of strains may be related, Dcifer produces a corresponding number of estimates—one for each pair. To accurately estimate multiple relatedness parameters without any additional assumptions, a large number of diverse loci is needed; otherwise, there is a lot of variation in the individual estimates (see an example in Supplementary Fig. 7 in File_1). However, while the estimation of individual relatedness parameters was challenging, their sum r_{total} was estimated more accurately even with a lower number of loci (which can be seen in the contour plots of Supplementary Fig. 7 in File_1). Supplementary Fig. 8 in File_1 shows likelihood surfaces for 2-dimensional parameters ($M = 2$), where we can clearly see the ridge along $\hat{r}_1 + \hat{r}_2$; that value is close to the true sum even when the individual estimates are further away from (r_1, r_2).

If the goal is to estimate overall relatedness between 2 infections, we suggest r_{total} as a more identifiable and useful quantity than ($r_1, \ldots, r_M$). To estimate r_{total} and the number of positively related strain pairs M', we used the procedure described in *Estimating the Number of Related Strain Pairs and r_{total}* and compared the estimates obtained from 2 approaches: (1) with "equality assumption" $r_1 = \cdots = r_M$ (estimators $\tilde{M}'_{eq}, \tilde{r}_{total,eq}$) and (2) without it (estimators $\tilde{M}', \tilde{r}_{total}$). First, we validated the stopping rule for the second approach (when r_i are not assumed to be equal), confirming that incrementing *M* past the iteration that estimates one of r_i's to be 0 only appended additional 0's to MLE in most cases. Next, we compared the 2 approaches and assessed the accuracy
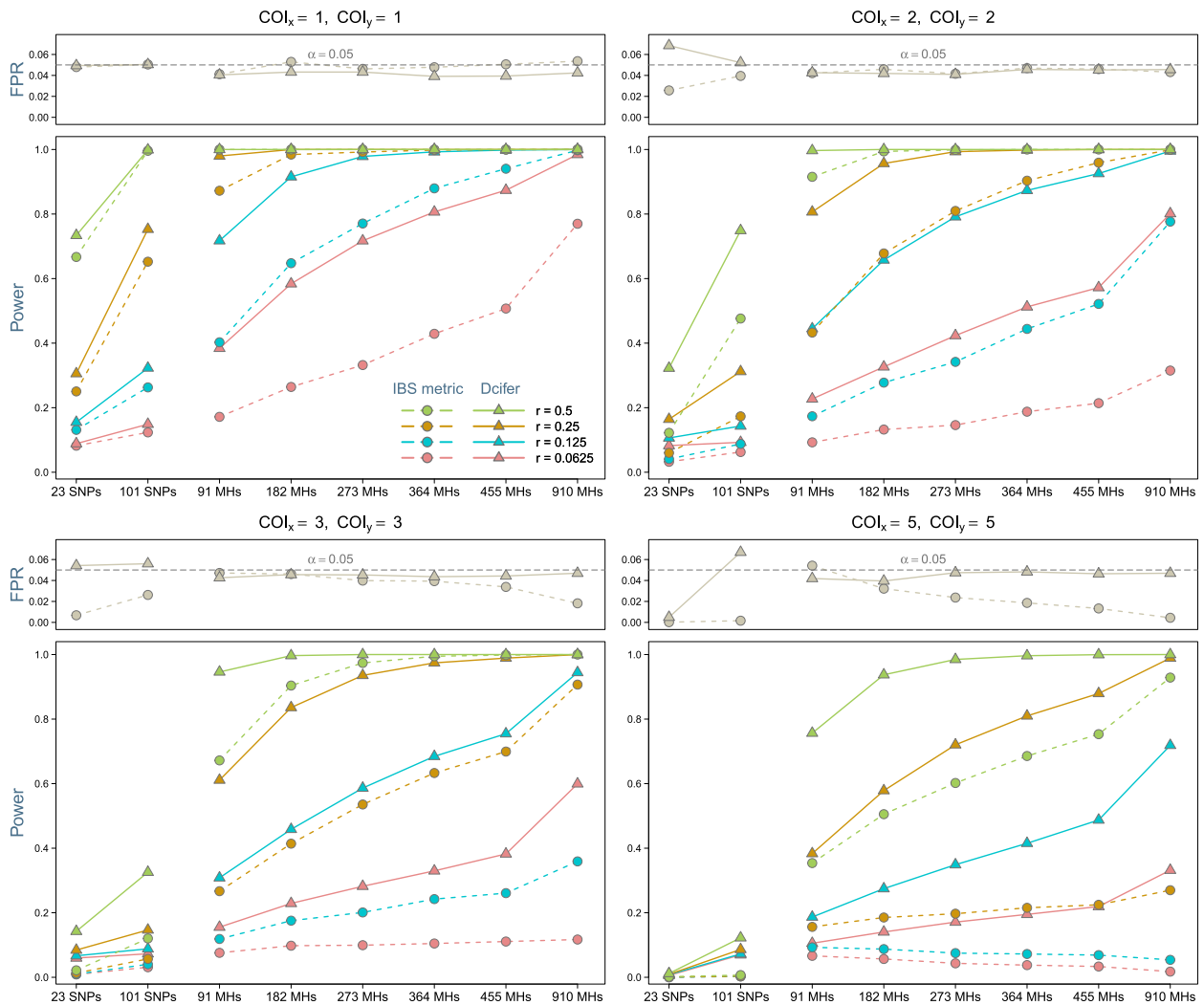
Fig. 3. Detecting related infections. FPR and statistical power of a test $H_0 : r = 0$ at significance level $\alpha = 0.05$ are shown. Simulations included genotyping error with fixed error model parameters; COIs were estimated from these data. For simulated null distributions, we varied error model parameters since they would not be normally known. SNP and microhaplotype (MH) panels were used as a basis for simulations.

of the corresponding estimators. For each simulated pair of infections, we first randomly generated $r_1, \ldots, r_{M'}$, $\sum_{i=1}^{M'} r_i = r_{total}$ for given r_{total} and M'. The estimates were compared across a grid of COI, r_{total}, and M'. Figure 5 shows illustrative examples of these comparisons: in 5(a), M' is changed while COI and r_{total} are fixed, in 5(b) r_{total} is changed, and in 5(c) COI is changed. Distributions of $\tilde{r}_{total}$ and $\tilde{r}_{total,eq}$ were quite similar, so the equality constraint had a very limited effect on the overall relatedness estimates. There was more difference between $\tilde{M}'$ and $\tilde{M}'_{eq}$, but, importantly, these differences did not significantly affect r_{total} estimates. An effect of varying M' on the distributions of $\tilde{r}_{total}$ and $\tilde{r}_{total,eq}$ was small, while lower COI resulted in more accurate estimates. Higher r_{total} made for more accurate estimation of M', as it eliminated lower values incompatible with r_{total} estimates. It is worth noting that simply increasing dimensionality of the grid of relatedness values to evaluate over can become unfeasible for larger M, so the grid would have to be coarsened to accommodate, which in turn would affect precision. No such limitation exists for the fast "equal r_i" approach as it estimates a single parameter.

Misspecifications and assumption violations

In data analysis, COI and population allele frequencies are usually unknown and need to be estimated from data. Allele frequencies can often be estimated from sufficiently large datasets (e.g. over 100 samples) and as such, their estimates are often fairly stable; some implications of their misspecifications are discussed in Supplementary Section 3 in File_1. COI estimation, however, relies on a smaller amount of information, resulting in greater variability of the estimates and more frequent misspecifications. Fortunately, Dcifer appeared to be relatively robust to COI misspecifications, especially for less complex infections (Supplementary Fig. 9 in File_1). Even for higher COI, relatedness estimates were fairly close to the true value in the neighborhood of the correct COI.

Next, we address our working model and its defining assumption of no intrahost relatedness. To assess how violating this assumption affects interhost relatedness estimation, we compared 5 scenarios: one with no intrahost relatedness and 4 where different strains within the samples are siblings ($r = 0.5$)—see the diagrams in Supplementary Fig. 10 in File_1. Note that in scenario 2
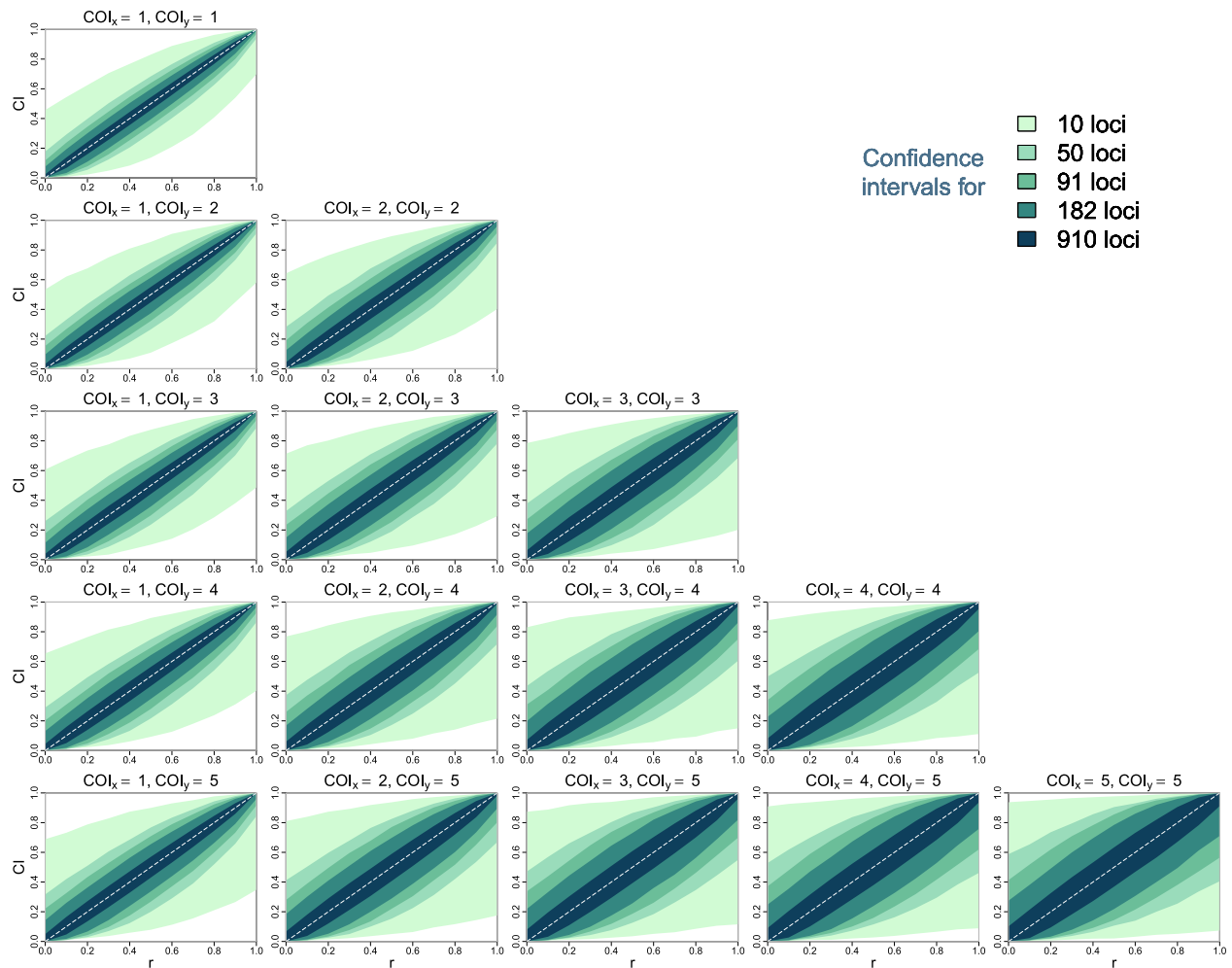
Fig. 4. 95% CI for relatedness estimates using the likelihood ratio produced by Dcifer.

(Supplementary Fig. 10b), there are 2 interhost pairs that are related (strains X_1-Y_1 and X_2-Y_1) but that "extra" relatedness is only a consequence of the induced X_1-X_2 sibship and does not add anything to our quantity of interest. The same is true for more complex scenarios 4 (Supplementary Fig. 10d, with related interhost pairs X_1-Y_1, X_2-Y_1, X_1-Y_2, and X_2-Y_2) and 5 (Supplementary Fig. 10e, with related interhost pairs X_1-Y_1, X_2-Y_1, and X_3-Y_1). All simulations included genotyping error, and processing involved estimating COI and population allele frequencies. Relatedness estimates for all 3 scenarios were very similar (Supplementary Figs. 11 and 12 in File_1), confirming that in many common cases of intrahost relatedness, the working model estimates interhost relatedness without significant biases.

Applications to empirical data

We applied Dcifer to a small dataset that has 87 microhaplotypes and consists of samples obtained from patients presenting with malaria from 2 health facilities in Maputo and Inhambane provinces of Mozambique (Tessema *et al.* 2022). There were 52 samples overall with 26 from each clinic; only samples with data for at least 75 loci were considered for the analysis (Supplementary Data_1). From these samples, naïve COI estimates (60% polyclonal samples with maximum COI of 6) and subsequently estimates of population allele frequencies adjusted for COI were calculated (Supplementary Data_2). We initially set $M = 1$ and

used likelihood-ratio statistics to test a null hypothesis $H_0 : r = 0$ at significance level $\alpha = 0.05$ (with the procedure adjusted for a 1-sided test). For comparison, Jaccard similarity was used as an IBS metric; Fig. 6 displays results from both methods. Dcifer results indicated that the majority of samples were unrelated and that related samples were mostly from the same clinic. Note that significant pairs with relatively low relatedness estimates usually shared a rare allele at one or two loci. The IBS metric also picked up very highly related pairs, but, apart from those, it was more difficult to distinguish related samples from background. Some samples appeared to be less related to all the other ones in IBS results, and some—more (stripe-like patterns in the lower triangle); these single-sample relatedness levels correlated with estimated COI (e.g. all the lighter "stripes" corresponded to monoclonal samples) highlighting the fact that IBS similarity is strongly influenced by COI, which obscures contribution of descent. For related samples, we also estimated M' (the number of related strain pairs) and r_{total} (overall relatedness). For these samples, $\hat{r}_{total}$ ranged between 0.105 and 1.97 and there were 4 pairs of samples (all in Maputo), for which $\hat{r}_{total}$ exceeded 1, with estimated COI of 2 in all samples in these pairs and $\tilde{M}' = 2$ for all such pairs.

We also reevaluated microsatellite data from a previously published dataset, which contained 2,585 samples from 29 clinics in 4 districts in Namibia (Tessema *et al.* 2019), using
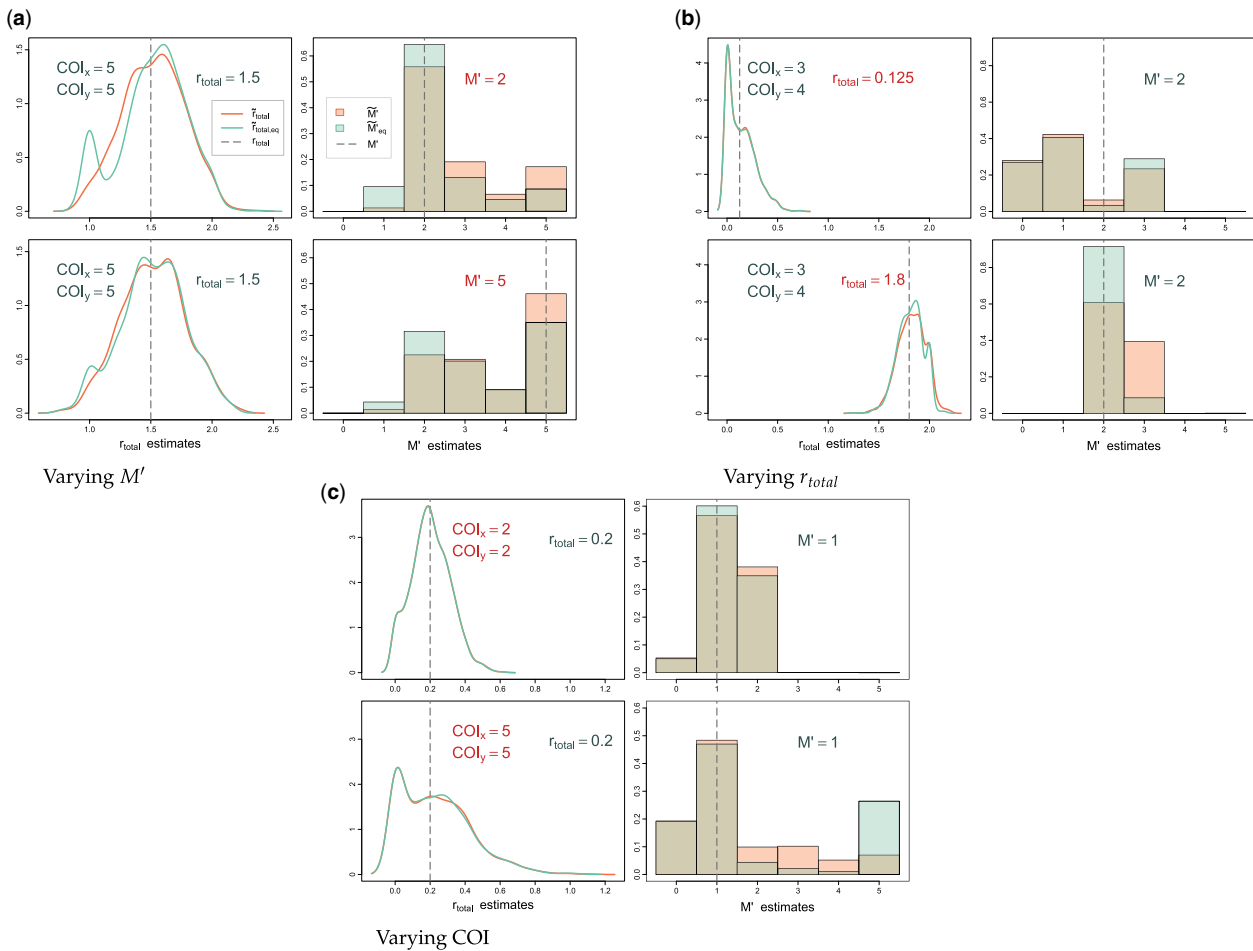
Fig. 5. Estimation of r_{total} and M' with and without equality assumption $r_1 = \cdots = r_M$. Densities of $\tilde{r}_{total}$ (no assumption) and $\tilde{r}_{total,eq}$ (with assumption) are shown on the left, and probabilities for the values of $\tilde{M}'$ and $\tilde{M}'_{eq}$ are shown on the right. Quantities highlighted in red indicate those varied between the top and bottom simulations within each panel. Simulations were performed using a panel of 91 microhaplotypes. a) Varying M'. b) Varying r_{total}. c) Varying COI.

Dcifer to estimate relatedness (Supplementary Data_3 contains estimated allele frequencies). These data had 26 loci, and 77% of the samples were polyclonal. The average of relatedness estimates between monoclonal samples (which can be taken to represent relatedness between individual parasites) was similar to the average overall relatedness ($\hat{r}_{total}$) between polyclonal infections scaled by the minimum COI for each pair. Minimum COI represents the number of interhost strain pairs in 2 infections that can be related within the Dcifer working model, so scaled $\hat{r}_{total}$ can be viewed as an average estimate of relatedness between individual strains belonging to 2 different infections. We assessed the proportion of related pairs of samples within and between clinics ($\alpha = 0.05$), then performed a permutation test to determine which clinic combinations had more related samples than expected by chance. Most of the within-clinic entries had significantly large numbers of related samples (Fig. 7). In addition, clinics with geographical proximity had significantly more related between-clinic infections, as illustrated by clusters of darker circles along the diagonal. Rundu DH is a large referral hospital, which could explain relative genetic closeness between samples from this and more geographically distant clinics. Rundu, Nyanganna, and Andara districts are adjacent to each other and Zambezi district is distant from them, which is reflected in the relative lack of relatedness between Zambezi and other districts.

Computational efficiency

Because of various simplifications of the likelihood expression for important special cases, there are 3 main versions of the estimation procedure (in the order of increasing computational complexity): (1) $M = 1$, using Newton's method adapted for a bounded parameter space to find MLE or calculating likelihood for a 1-dimensional grid of parameter values if a CI is requested, (2) $M > 1$ with $r_1 = \cdots = r_M$ assumption involving likelihood calculation for a 1-dimensional grid, and (3) $M > 1$ without $r_1 = \cdots = r_M$ assumption involving likelihood calculation for an M-dimensional grid. Since we expect that most sample pairs in a typical dataset will be unrelated, an initial analysis step will likely attempt to identify significantly related (or unrelated) ones by evaluating all pairs assuming $M = 1$. Therefore, execution times for this special case are the most important ones. Times for "grid" versions depend heavily on the resolution of the grid, while the Newton-based procedure usually converges in 3–4 iterations even with low tolerance. Table 1 shows execution times for calculating $\hat{r}$ for 1,000 sample pairs (100 loci, each with 2–20 alleles) with $M = 1$ and various COI combinations for both Newton's method and a grid; times for all but the highest COI combinations ($COI_x + COI_y \leq 15$ for Newton's method and ≤ 14 for the grid) were <5 s. Simple vectorized computation of the IBS metric does not depend on COI and averages 0.055 s for the same data (1,000 pairs). Calculation times are essentially linear in the number of loci, so times for
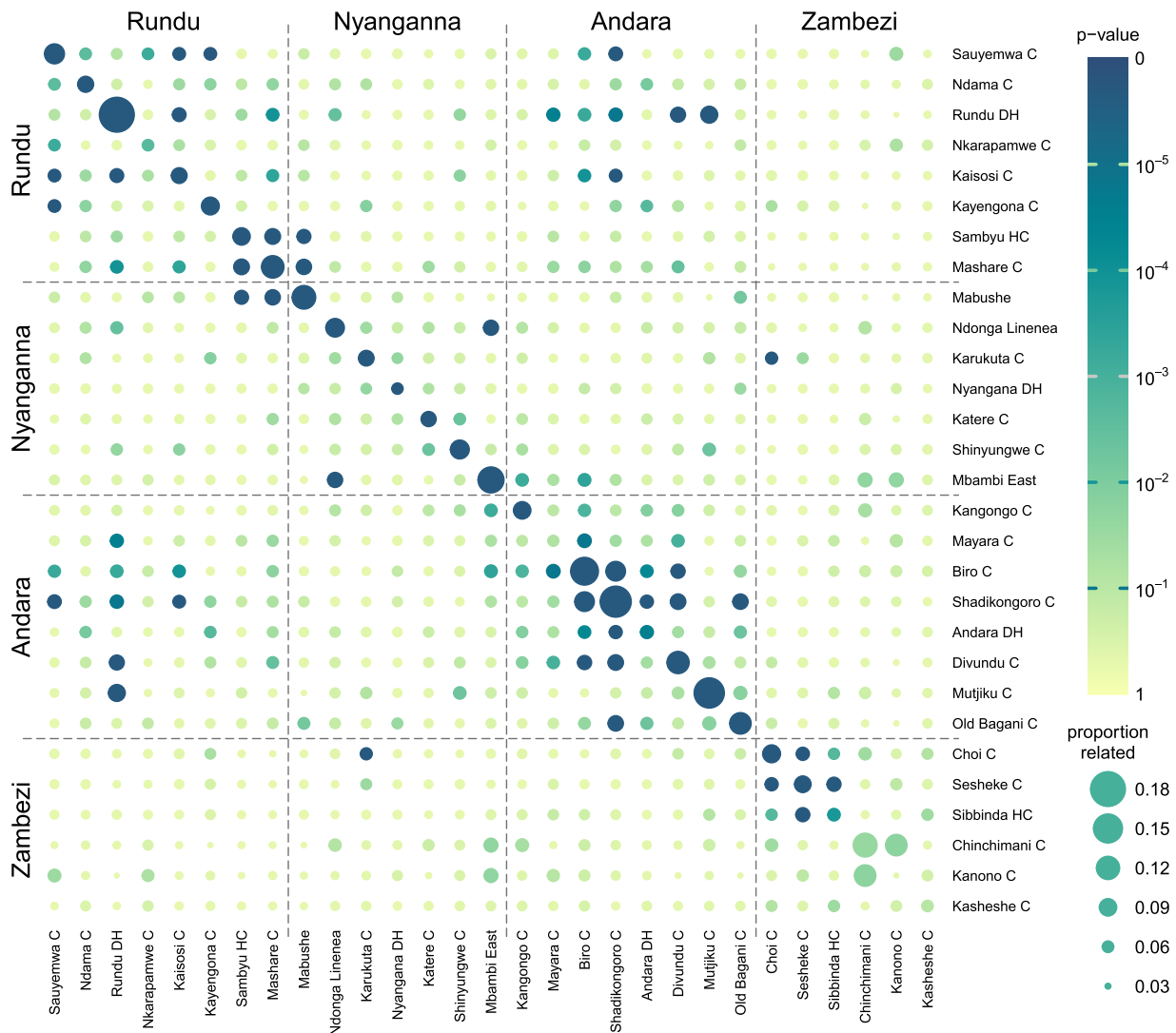
Fig. 6. Relatedness between samples from 2 clinics: one from Inhambane province and one from Maputo province of Mozambique. Lower triangular matrix displays IBS metric results and upper triangular matrix—Dcifer estimates. The color of each matrix entry represents an estimate for 2 corresponding samples; pairs for which H_0 has been rejected using the Dcifer likelihood ratio test are outlined in red. Estimated COIs are displayed on the margins.

larger panels can easily be estimated. When $M > 1$, calculation times are roughly exponential in M for when r_i equality is not assumed and increase only slightly with M when it is; these times are shown in Table 2.

For data analysis presented in *Applications to empirical data*, Dcifer execution times were 2.7 s for Mozambique (pairwise distances for 52 samples, or 1,326 comparisons) and 32 min for Namibia (2,585 samples, 3,339,820 comparisons). Consequently, we can estimate that it would take approximately 17 min to calculate all pairwise distances for 1,000 samples of Mozambique-type data (87 microhaplotypes). Calculations were performed on MacBook Pro, 2019, 2.3 GHz Intel Core i9 and were not parallelized.

Discussion

The ability to infer genetic distance between infections is a critical step in translating pathogen genetic data into insight regarding transmission. Despite this, there is a lack of established methods available to infer genetic distance between malaria infections containing multiple parasites, which are the majority in many endemic areas. Options are even more limited when

using multiallelic loci, which offer more resolution than biallelic SNPs. The lack of any formal approach has left the community with only ad hoc calculations such as IBS, which yield ambiguous results and require extensive efforts to guide any attempt at meaningful inference. In contrast, Dcifer provides relatedness estimates that are based on IBD, interpretable quantities with consistent meaning across studies—regardless of genotyping methods used—with clear implications for ancestry. Importantly, we also show that Dcifer's statistical power to detect related infections consistently surpasses that of IBS. The method produces reliable measures of uncertainty, and inference obtained from Dcifer vs IBS is more robust to misspecifications of estimated quantities such as COI or population allele frequencies. The R software package implementing Dcifer provides a fast, convenient, and flexible tool that can be easily incorporated into the analysis stream of a wide range of genotyping data to understand transmission.

While Dcifer is designed to work with many types of genotyping data, e.g. biallelic SNPs, multiallelic loci such as microsatellites and microhaplotypes, and any combination thereof, we show that including multiallelic loci results in substantial gains in power to detect related infections. These benefits become

Fig. 7. Namibia clinic-level relatedness and permutation test results. Each circle represents a single clinic (on a diagonal) or a 2-clinic combination (off-diagonal), with clinics ordered geographically and divided into districts. The color of the circle corresponds to the permutation test's *P*-values, and the diameter—to the proportion of related samples within or between the clinics. Permutation distributions for clinics with smaller numbers of samples have larger variance leading to larger *P*-values—e.g. Chinchimani and Kanono with only 9 samples each have relatively high proportions of related infections but their *P*-values are not small.

more dramatic as COI increases. The result makes intuitive sense, as multiallelic panels provide more within-host strain differentiation and consequently can allow information pertaining to descent to be more easily detected, benefiting relatedness estimation. For example, where 2 infections with high COI may look similar with biallelic genotyping panels regardless of their level of relatedness (both alleles present at most loci with some diversity), having multiallelic data provides the opportunity to compare these infections more meaningfully. Fortunately, the greater availability of methods to obtain multiallelic data from across the genome makes it feasible to generate these data efficiently and in a high-throughput manner (Aydemir *et al.* 2018; Tessema *et al.* 2022; LaVerriere *et al.* 2022).

The concept of relatedness for individual parasites does not extend trivially to polyclonal infections, where strains within and between infections can be related. Dcifer offers an approach that focuses on relatedness between infections as this information is very relevant to transmission. Simulations with imposed intra-host relatedness indicate that the working model achieves its stated goal of capturing interhost relatedness by implicitly

downweighting the independent contribution of related strains within a host to comparisons between hosts. When multiple pairs of strains are related between 2 infections, degrees of relatedness between these strains could potentially provide an insight into more nuanced aspects of transmission. However, these strain-level comparisons may be difficult to estimate without accurate phasing, which remains a challenging problem. Here, we proposed a more easily identifiable summary, r_{total}, which provides a single measure of the overall degree of relatedness between 2 infections. This summary reliably encapsulates information encoded by the number of related strains and degree of relatedness between them even when individual r values are difficult to identify from realistic data—which was the case in our analyses except in situations when a large number of highly informative loci were used. For example, an r_{total} of 1.5 could indicate, amongst other scenarios, that 3 pairs of siblings or one clonal pair and one sibling pair are present between 2 infections. Either way, multiple closely related parasites exist between these 2 infections and they are likely to be closely linked by transmission events, e.g. via a single cotransmission event or multiple

Table 1. Execution times (in seconds) for Dcifer processing 1,000 pairs of samples with COI combinations ranging between 1 and 10.

	1	2	3	4	5	6	7	8	9	10
1	2.24									
	3.16									
2	2.31	2.42								
	3.24	3.24								
3	2.31	2.35	2.39							
	3.29	3.30	3.29							
4	2.20	2.30	2.43	2.43						
	3.47	3.44	3.40	3.32						
5	2.21	2.26	2.28	2.35	2.36					
	3.41	3.55	3.51	3.53	3.66					
6	2.26	2.34	2.32	2.36	2.48	2.58				
	3.38	3.47	3.47	3.57	3.69	3.79				
7	2.26	2.32	2.37	2.49	2.61	2.86	3.46			
	3.52	3.51	3.65	3.68	3.90	4.07	4.60			
8	2.30	2.40	2.47	2.60	2.85	3.38	4.46	6.53		
	3.51	3.68	3.65	3.82	4.09	4.68	5.70	7.99		
9	2.40	2.52	2.61	2.88	3.37	4.38	6.50	10.81	19.31	
	3.75	3.61	3.78	4.01	4.62	5.60	7.78	12.18	21.03	
10	2.67	2.77	2.95	3.45	4.43	6.65	11.28	20.03	37.02	72.40
	3.87	3.98	4.15	4.61	5.72	7.69	12.16	21.04	38.43	72.37
COI	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**

The data were simulated from 100 multiallelic loci (the number of alleles at each locus distributed uniformly between 2 and 20). $M = 1$, each COI combination displays times for using Newton's method (unshaded) and calculating likelihood for a grid (shaded). Precision was set at 0.001 for both methods.

Table 2. Execution times (in seconds) for a pair of infections with COI of 6 and 9.

M	1	2	3	4	5	6
$r_1 = \cdots = r_M$ constraint	0.01	0.67	0.79	0.86	0.89	0.92
No constraint	0.01	1.32	2.17	3.72	6.65	11.54

M ranges between 1 and 6, and the likelihood is calculated over a grid of 1,000 $r_1, \ldots, r_M$ combinations (or 1,000 r values when $r_1 = \cdots = r_M = r$ is assumed).

independent transmissions. We anticipate that the interpretation of this summary and its potential derivatives, as well as its incorporation into downstream analyses, will evolve as it is evaluated in more sophisticated population-level simulations and estimates from empirical data sets. Another conceptual issue concerns population allele frequencies, which can affect Dcifer estimates. The foremost question is what constitutes the relevant source population in regards to relatedness between samples and consequently from which data the frequencies should be estimated. If 2 infections are from communities with different within-community allele frequencies, what are the implications for descent? Dcifer currently assumes the same allele frequencies for both samples but further exploration might be warranted depending on the question of interest. Questions concerning population and scope of the analysis might also arise in regards to potential multiple testing procedures when many pairwise relatedness hypotheses are tested simultaneously. The fact that these hypotheses are not independent should be taken into account when such procedures are considered.

The Dcifer model does not account for linkage disequilibrium and assumes independence of loci. As the malaria genome has relatively short linkage disequilibrium segments, loci independence can be assumed up to a reasonably large number of loci for a correspondingly designed genotyping panel. If, however, the panel has loci that are likely to be linked, e.g. those selected to be in close proximity or for a large number of loci, the independence assumption would no longer hold, which could result in anticonservative inference. Another limitation is that the model currently does not account for genotyping errors. Future modification could explicitly incorporate the error process via an appropriate model or assess how a specific error process affects the estimates and inference beyond the explorations we have performed here. Another potential venue for further work is developing an MLE estimator for r_{total} directly as this might become a commonly used summary. A direct estimator might be more efficient, would have the properties of MLE, and would require less processing time. In addition, scaling r_{total} by some function of COI could provide a useful way of comparing all pairwise relatedness estimates with standardized values in a [0,1] interval. Other future directions could explore alternative inferential approaches, including a nonparametric bootstrap, where loci data would be sampled with replacement. In that case, the fact that variables associated with different loci are not identically distributed, and therefore loci might not be equally informative, would need to be addressed.

With potential to facilitate understanding of relatedness structure from unphased genetic data, including multiallelic loci, Dcifer can provide a vital link in the analytical process leading to better understanding of malaria transmission dynamics. While we have demonstrated the utility of this method for *Plasmodium* infections here, Dcifer may be useful in analyses of other organisms that undergo sexual recombination and where polyclonal infections are encountered, such as shistosomiasis, filarial disease, and soil transmitted helminths (Brouwer *et al.* 2001; Churcher *et al.* 2008). With the ability to incorporate most types of genetic data, rapid computation, and readily available inference, Dcifer may prove to be an important tool in the analytical toolbox for obtaining epidemiologic insight from pathogen genetics.

Data availability

Supplementary Data_1 contain microhaplotype data from Mozambique. Microsatellite data from Namibia are publicly available at https://elifesciences.org/articles/43510/figures#supp1 (accessed 2022 Aug 25). Supplementary Data_2 and Data_3 contain population allele frequencies estimated from Mozambique and Namibia datasets, respectively.

Supplemental material is available at *GENETICS* online.

submission. Funding for this project also came from the National Institutes of Health (K24 AI44048).

Conflicts of interest

None declared.

Literature cited

Atuh NI, Anong DN, Jerome FC, Oriero E, Mohammed NI, D'Alessandro U, Amambua-Ngwa A. High genetic complexity but low relatedness in *Plasmodium falciparum* infections from Western Savannah Highlands and coastal equatorial Lowlands of Cameroon. Pathog Glob Health. 2021;1–10.

Aydemir O, Janko M, Hathaway NJ, Verity R, Mwandagalirwa MK, Tshefu AK, Tessema SK, Marsh PW, Tran A, Reimonn T, *et al.* Drug-resistance and population structure of *Plasmodium falciparum* across the Democratic Republic of Congo using high-throughput molecular inversion probes. J Infect Dis. 2018;218(6): 946–955.

Beale E. Confidence regions in non-linear estimation. J R Stat Soc B (Methodol). 1960;22(1):41–76.

Brouwer KC, Ndhlovu P, Munatsi A, Shiff CJ. Genetic diversity of a population of *Schistosoma haematobium* derived from schoolchildren in east central Zimbabwe. J Parasitol. 2001;87(4):762–769.

Chang HH, Wesolowski A, Sinha I, Jacob CG, Mahmud A, Uddin D, Zaman SI, Hossain MA, Faiz MA, Ghose A, *et al.* Mapping imported malaria in Bangladesh using parasite genetic and human mobility data. Elife. 2019;8:e43481.

Chang HH, Worby CJ, Yeka A, Nankabirwa J, Kamya MR, Staedke SG, Dorsey G, Murphy M, Neafsey DE, Jeffreys AE, *et al.* THE REAL McCOIL: a method for the concurrent estimation of the complexity of infection and SNP allele frequency for malaria parasites. PLoS Comput Biol. 2017;13(1):e1005348.

Churcher TS, Schwab AE, Prichard RK, Basáñez MG. An analysis of genetic diversity and inbreeding in *Wuchereria bancrofti*: implications for the spread and detection of drug resistance. PLoS Negl Trop Dis. 2008;2(4):e211.

Cook RD, Weisberg S. Confidence curves in nonlinear regression. J Am Stat Assoc. 1990;85(410):544–551.

Cox DR, Hinkley DV. Theoretical Statistics. Oxfordshire (UK): CRC Press; 1979.

Daniels R, Volkman SK, Milner DA, Mahesh N, Neafsey DE, Park DJ, Rosen D, Angelino E, Sabeti PC, Wirth DF, *et al.* A general SNP-based molecular barcode for *Plasmodium falciparum* identification and tracking. Malar J. 2008;7:223–211.

Efron B, Hinkley DV. Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information. Biometrika. 1978;65(3):457–483.

Gerlovina I. dcifer: Genetic Relatedness Between Polyclonal Infections. R package version 1.1.1. 2022. [Accessed 2022 Aug 25]. https://cran.r-project.org/web/packages/dcifer/index.html

Henden L, Lee S, Mueller I, Barry A, Bahlo M. Identity-by-descent analyses for measuring population dynamics and selection in recombining pathogens. PLoS Genet. 2018;14(5):e1007279.

Jaccard P. The distribution of the flora in the alpine zone. New Phytol. 1912;11(2):37–50.

Jacob CG, Thuy-Nhien N, Mayxay M, Maude RJ, Quang HH, Hongvanthong B, Vanisaveth V, Duc TN, Rekol H, van der Pluijm R, *et al.* Genetic surveillance in the Greater Mekong subregion and South Asia to support malaria control and elimination. Elife. 2021;10:e62997.

Knuth DE. The Art of Computer Programming, Volume 4A: Combinatorial Algorithms, Part 1. Delhi (India): Pearson Education India; 2011.

Koepfli C, Ross A, Kiniboro B, Smith TA, Zimmerman PA, Siba P, Mueller I, Felger I. Multiplicity and diversity of *Plasmodium vivax* infections in a highly endemic region in Papua New Guinea. PLoS Negl Trop Dis. 2011;5(12):e1424.

LaVerriere E, Schwabl P, Carrasquilla M, Taylor AR, Johnson ZM, Shieh M, Panchal R, Straub TJ, Kuzma R, Watson S, *et al.* Design and implementation of multiplexed amplicon sequencing panels to serve genomic epidemiology of infectious disease: a malaria case study. Mol Ecol Resour. 2022;22(6):2285–2303.

Lerch A, Koepfli C, Hofmann NE, Messerli C, Wilcox S, Kattenberg JH, Betuela I, O'Connor L, Mueller I, Felger I. Development of amplicon deep sequencing markers and data analysis pipeline for genotyping multi-clonal malaria infections. BMC Genomics. 2017; 18(1):864.

Meeker WQ, Escobar LA. Teaching about approximate confidence regions based on maximum likelihood estimation. Am Stat. 1995; 49(1):48–53.

Neyman J, Pearson ES. IX. On the problem of the most efficient tests of statistical hypotheses. Philos Trans R Soc Lond A. 1933;231: 289–337.

Pringle JC, Tessema S, Wesolowski A, Chen A, Murphy M, Carpi G, Shields TM, Hamapumbu H, Searle KM, Kobayashi T, *et al.* Genetic evidence of focal *Plasmodium falciparum* transmission in a pre-elimination setting in southern province, Zambia. J Infect Dis. 2019;219(8):1254–1263.

Roh ME, Tessema SK, Murphy M, Nhlabathi N, Mkhonta N, Vilakati S, Ntshalintshali N, Saini M, Maphalala G, Chen A, *et al.* High genetic diversity of *Plasmodium falciparum* in the low-transmission setting of the Kingdom of Eswatini. J Infect Dis. 2019;220(8): 1346–1354.

Schaffner SF, Taylor AR, Wong W, Wirth DF, Neafsey DE. hmmIBD: software to infer pairwise identity by descent between haploid genotypes. Malar J. 2018;17(1):4.

Taylor AR, Jacob PE, Neafsey DE, Buckee CO. Estimating relatedness between malaria parasites. Genetics. 2019;212(4):1337–1351.

Taylor AR, Schaffner SF, Cerqueira GC, Nkhoma SC, Anderson TJ, Sriprawat K, Pyae Phyo A, Nosten F, Neafsey DE, Buckee CO. Quantifying connectivity between local *Plasmodium falciparum* malaria parasite populations using identity by descent. PLoS Genet. 2017;13(10):e1007065.

Tessema S, Wesolowski A, Chen A, Murphy M, Wilheim J, Mupiri AR, Ruktanonchai NW, Alegana VA, Tatem AJ, Tambo M, *et al.* Using parasite genetic and human mobility data to infer local and cross-border malaria connectivity in Southern Africa. Elife. 2019; 8:e43510.

Tessema SK, Hathaway NJ, Teyssier NB, Murphy M, Chen A, Aydemir O, Duarte EM, Simone W, Colborn J, Saute F, *et al.* Sensitive, highly multiplexed sequencing of microhaplotypes from the *Plasmodium falciparum* heterozygome. J Infect Dis. 2022;225(7):1227–1237.

Vander Wiel SA, Meeker WQ. Accuracy of approx confidence bounds using censored Weibull regression data from accelerated life tests. IEEE Trans Rel. 1990;39(3):346–351.

Wald A. Tests of statistical hypotheses concerning several parameters when the number of observations is large. Trans Amer Math Soc. 1943;54(3):426–482.

Wesolowski A, Taylor AR, Chang HH, Verity R, Tessema S, Bailey JA, T Perkins A, Neafsey DE, Greenhouse B, Buckee CO. Mapping

malaria by combining parasite genomic and epidemiologic data. BMC Med. 2018;16(1):1–8.

White MT, Karl S, Koepfli C, Longley RJ, Hofmann NE, Wampfler R, Felger I, Smith T, Nguitragool W, Sattabongkot J, *et al. Plasmodium vivax* and *Plasmodium falciparum* infection dynamics: reinfections, recrudescences and relapses. Malar J. 2018;17(1):1–15.

Wilks SS. The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann Math Statist. 1938;9(1): 60–62.

World Health Organization. Meeting report of the technical consultation on the role of parasite and anopheline genetics in malaria surveillance, 5–7 June 2019, Geneva, Switzerland. Technical report. World Health Organization. 2019.

Zhu SJ, Hendry JA, Almagro-Garcia J, Pearson RD, Amato R, Miles A, Weiss DJ, Lucas TC, Nguyen M, Gething PW, *et al.*; for the Pf3k Project. The origins and relatedness structure of mixed infections vary with local prevalence of *P. falciparum* malaria. Elife. 2019;8:e40845.

Communicating editor: S. Browning