bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Estimating relatedness between malaria parasites

Aimee R. Taylor[1,2,†], Pierre E. Jacob[3], Daniel E. Neafsey[1,2], Caroline O. Buckee[1]

[1]Harvard T. H. Chan School of Public Health, Boston, MA, USA; [2]Broad Institute of MIT and Harvard, Cambridge, MA, USA; [3] Department of Statistics, Harvard University, Cambridge, MA, USA

[†]corresponding author, ataylor@hsph.harvard.edu

1 Abstract

Understanding the relatedness of individuals within or between populations is a common goal in biology. Increasingly, relatedness features in genetic epidemiology studies of pathogens. These studies are relatively new compared to those in humans and other organisms, but are important for designing interventions and understanding pathogen transmission. Only recently have researchers begun to routinely apply relatedness to apicomplexan eukaryotic malaria parasites, and to date have used a range of different approaches on an ad hoc basis. It remains unclear how to compare different studies, therefore, and which measures to use. Here, we systematically compare measures based on identity-by-state and identity-by-descent using a globally diverse data set of malaria parasites, *Plasmodium falciparum* and *Plasmodium vivax*, and provide marker requirements for estimates based on identity-by-descent. We formally show that the informativeness of polyallelic markers for relatedness inference is maximised when alleles are equifrequent. Estimates based on identity-by-state are sensitive to allele frequencies, which vary across populations and by experimental design. For portability across studies, we thus recommend estimates based on identity-by-descent. To generate reliable estimates, we recommend approximately 200 biallelic or 100 polyallelic markers. Confidence intervals illuminate inference across studies based on different sets of markers. These marker requirements, unlike many thus far reported, are immediately applicable to haploid malaria parasites and other haploid eukaryotes. This is the first attempt to provide rigorous analysis of the reliability of, and requirements for, relatedness inference in malaria genetic epidemiology, and will provide a basis for statistically informed prospective study design and surveillance strategies.

Keywords: identity-by-state, identity-by-descent, relatedness, independence model, hidden Markov model, malaria, *Plasmodium falciparum*, *Plasmodium vivax*, genetic epidemiology

2 Introduction

Genetic relatedness is a measure of recent shared ancestry (reviewed in [1, 2]). It ranges from zero between two unrelated individuals to one between clones, and in the absence of inbreeding is broken down by recombination [3]. Since the early 20th century, estimates of relatedness have been used across a wide variety of fields: archaeology, agriculture, forensic science, paternity testing, human disease gene mapping, conservation, and ecology [1, 4]. Nevertheless, studies of relatedness

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

are both new and niche in infectious disease molecular epidemiology: new because the field itself is [5], and niche because only a subset of pathogens are eukaryotes, e.g. helmiths and parasitic protoza, which include malaria parasites (reviewed in [6], but without reference to relatedness). Because relatedness is broken down by outbreeding, it can change with each generation [7]. On a population-level, studies of malaria parasite relatedness thus provide a sensitive measure of recent gene flow [8], generating insight on an operationally relevant scale for disease control efforts [9].

Malaria parasites are haploid during the human stage of their life cycle. One measure of relatedness between haploid genotypes is equivalent to the diploid coefficient of inbreeding, defined by Malécot as a probability of identity-by-descent (IBD) [10]. Two alleles are identical by descent (also IBD hereafter) if they are descended from a common ancestor in some ancestral reference population, whose members are assumed unrelated [11, 7, 2].[1] For pedigrees, the reference is the founder population; more generally, it is a population at some arbitrary time-depth (reviewed in [7, 2]). Two alleles that share the same allelic type are identical by state (IBS) and include those that are both IBD and not IBD [11, 1, 12, 7, 13, 2]. identity-by-state (also IBS hereafter) is observed, whereas IBD is hidden. Though hidden, relatedness based on IBD can be inferred from genetic data. Many estimators exist, some assuming independence between hidden IBD states [1, 11], others not (e.g. [14] and subsequent models - see [15] and references therein). Those assuming independence have fewer parameters but impaired power in the presence of dependence [16]. Estimators that do not assume independence are often based on hidden Markov models (HMMs, reviewed in [17]). Measures of relatedness used in studies of malaria include those estimated under HMMs (hmmIBD [18], used in e.g. [19, 20, 21, 22]; isoRelate [23], extension of XIBD [24]; DEploidIBD [25], extension of DEploid [26]). Measures based on IBS (e.g. proportions of alleles shared or counts of allele differences) require only simple calculation and are thus popular also as a proxy for measuring relatedness between malaria parasites (e.g. [27, 28, 19, 29, 30, 31, 32]).

Despite many malaria genetic epidemiology studies using IBD and IBS based analyses, there are few systematic comparisons applicable to malaria studies. Questions also remain about optimal marker requirements for pairwise relatedness inference. To enable comparison between studies of malaria epidemiology, we compare and assess measures based on IBD and IBS using simulated data; various data sets of *Plasmodium falciparum*, the parasite responsible for the most deadly type of human malaria; and a data set of *Plasmodium vivax*, the parasite most commonly responsible for recurrent malaria. We use a model framework encompassing two simple models assuming independence and not. It is an error-modified version of [14] and thus at the core of many probabilistic IBD models (see review in [15]), including those specifically designed for comparison across malaria parasites [18, 23]. To guide future relatedness studies of monoclonal haploid malaria parasite samples and haploid eukaryotes more generally, we explore marker count and number of alleles for relatedness inference with specified error. Simulated data illustrate how IBS is sensitive to marker panels, making conclusions non-portable across studies. Concrete recommendations on marker requirements depend on specified error. Increasing the number of alleles per marker genotyped reduces error, especially when markers are few, but with diminishing returns.

[1] IBD was first defined it terms of mutation: 'pairs of alleles at a locus are mutation-sense IBD if there has been no mutation since their MRCA', where MRCA stands for most recent common ancestor [2]. It can also be interpreted in terms of IBD segments: shared genomic regions unbroken by recombination since their MRCA [7, 2]. This interpretation, referred to as 'recombination-sense' in [2], circumvents the problem of a specifying a reference population but presents the problem of specifying some small segment length [7, 2].

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

3 Methods and Theory

3.1 Relatedness

For the purpose of this study, relatedness r is defined as the probability that, at any locus on the genome, the allele sampled from one individual is IBD to the allele sampled from the other individual. This is referred to as the pointwise pairwise probability of IBD in [15]. We denote by m the number of genotyped markers. Each of them has a locus on the genome. We index these loci by $t = 1, \ldots, m$. We denote by c_t the index of the chromosome of the t-th locus, and by p_t its position on that chromosome. For two indices t_1, t_2 with $t_1 < t_2$, we either have $c_{t_1} < c_{t_2}$, or $c_{t_1} = c_{t_2}$ and $p_{t_1} < p_{t_2}$. For the t-th locus we denote by IBD_t the binary variable indicating whether the two individuals are IBD at that locus; $\mathrm{IBD}_t = 1$ indicates IBD, otherwise $\mathrm{IBD}_t = 0$.[2] We assume that r, the marginal probability that $\mathrm{IBD}_t = 1$, is constant across the genome:

$$\forall t = 1, \ldots, m \quad r = \mathbb{P}(\mathrm{IBD}_t = 1). \tag{3.1}$$

The sequence $(\mathrm{IBD}_t)_{t=1,\ldots,m}$ could be made of independent Bernoulli variables with parameter r, or more generally a Markov chain with a Bernoulli invariant distribution with parameter r. For the Markov chain model, we write the transition probabilities at locus t,

$$A(t) = \begin{pmatrix} a_{00}(t) & a_{01}(t) \\ a_{10}(t) & a_{11}(t) \end{pmatrix} = \begin{pmatrix} 1 - r(1 - \exp(-k\rho d_t)) & r(1 - \exp(-k\rho d_t)) \\ (1-r)(1 - \exp(-k\rho d_t)) & 1 - (1-r)(1 - \exp(-k\rho d_t)) \end{pmatrix}.$$

In the above, $a_{j\ell}(t)$ refers to the probability of $\mathrm{IBD}_t = \ell$ given that $\mathrm{IBD}_{t-1} = j$, d_t denotes a genetic distance in base pairs (bp) between sites $t-1$ and t (i.e. all markers are treated as point polymorphisms). If the locus $t-1$ and t are on different chromosomes ($c_{t-1} \neq c_t$) the distance is set to $+\infty$; in that case the variables IBD_{t-1} and IBD_t are independent. The value $k > 0$ parameterizes the switching rate of the Markov chain and ρ is a constant equal to the recombination rate, assumed known and fixed across both haploid genotypes with value $7.4 \times 10^{-7}\mathrm{M\ bp}^{-1}$ for *P. falciparum* parasites [33].

We now describe how the model connects the quantity of interest r to the data. At each locus t, we assume that the set of possible alleles is denoted by $\mathcal{G}_t = \{g_1, \ldots, g_{K_t}\}$, where $K_t \geq 2$ denotes the cardinality of $\mathcal{G}_t$ (allelic richness of the t-th marker). For individuals i, j in the population and at locus t we observe the pair $Y_t^{(i)}, Y_t^{(j)} \in \mathcal{G}_t$. We assume that alleles occur with frequencies $(f_t(g))_{g\in\mathcal{G}_t}$, with $f_t(g) \geq 0$ for all $g \in \mathcal{G}_t$ and $\sum_{l=1}^{K_t} f_t(g_l) = 1$. The data comprise $Y_t^{(i)}, Y_t^{(j)}$, the distances (d_t) and the frequencies $(f_t(g))_{g\in\mathcal{G}_t}$ at m loci. A simple observation model relates the data to IBD_t by assuming that, if $\mathrm{IBD}_t = 0$, then $Y_t^{(i)}$ and $Y_t^{(j)}$ are independent categorical variables taking values in $\mathcal{G}_t$ with probabilities $(f_t(g))_{g\in\mathcal{G}_t}$. If $\mathrm{IBD}_t = 1$, then $Y_t^{(i)}$ is such a categorical variable and $Y_t^{(j)} = Y_t^{(i)}$ with probability one. A more realistic model accounting for observation error is described in Appendix B.

Combining the Markov model for (IBD_t) with an observation model as above leads to a hidden Markov model (Figure 1) with a likelihood function $(r, k) \mapsto \mathcal{L}_{1:m}(r, k)$. Note that, as mentioned in the introduction, this model is essentially an error-modified version of [14], and thus at the core of the many subsequent probabilistic IBD models (see [15]), including all those specifically designed for comparison across haploid malaria parasites [18, 23]. An independence model can be retrieved by

[2] For reasons outlined at the end of this section (3.1), we purposely omit reference to either an ancestral population or segment unbroken by recombination; IBD_t is simply a binary variable in $\{0, 1\}$.

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

e.g. setting all distances to $+\infty$. In either case we can maximize the likelihood over the parameter space and we denote by $\widehat{r}_m$ the maximum likelihood estimator of r, that can be computed using numerical optimization.

Under some assumptions on the data-generating process, the maximum likelihood estimate $\widehat{r}_m$ could be shown to be consistent for r as the sample size m goes to infinity. However these asymptotic considerations are intricate in the present setting of Mendelian sampling [34]. Indeed, the degree of dependencies between successive observations increases with the sample size m: the more sites are sampled, the closer to one another they become. This departs from the standard asymptotic setting, where the observations are not increasingly dependent as $m \to \infty$ [35, 36, 37]; this is discussed in more details in Appendix B.

Without standard asymptotic results, such as the asymptotic normality of the maximum likelihood estimator, we do not have a simple formula relating the sample size m to the variance of the estimator $\widehat{r}_m$, which would have been useful for sample size determination. The distribution of the $\widehat{r}_m$ can still be approximately normal because the log-likelihood can be approximately quadratic [38]. If that is the case, confidence intervals can be obtained through the second derivative of the log-likelihood at the MLE. The present setting poses an additional difficulty since the MLE can be located on the boundary of the parameter space, e.g. $\widehat{r}_m = 0$ or $\widehat{r}_m = 1$. This suggests that the distribution of the MLE might not always be normal [39]. We therefore rely on the parametric bootstrap [Chapter 9 of 40] to construct confidence intervals around $\widehat{r}_m$ (note that we cannot use the nonparametric bootstrap since we cannot sample positions with replacement). Unless otherwise stated, we use 500 bootstrap draws throughout. Note that even when assuming the absence of genotyping error, if $Y_t^{(i)} \neq Y_t^{(j)}$ for some $t = 1, \ldots, m$, the confidence interval around $\widehat{r}_m > 0$ can contain one, since data simulated with $\widehat{r}_m > 0$ may have identical genotype calls for all $t = 1, \ldots, m$ (especially if m is small), which leads to a bootstrapped estimate of r equal to one.

Under the model framework (Figure 1) no explicit mention is made of ancestors, be they most recent or at some arbitrary time-depth. In the introduction we refer to IBD relative to either some reference population or some small segment length ('recombination-sense' [2]). Akin to many existing IBD models (see catalogue in [15, 41, 42]) and some related imputation methods (see catelogue in [43, 44, 45]), IBD segments can be estimated within the HMM framework (e.g. using the most likely path of hidden states [17], posterior probabilities of IBD at each marker position [42, 46], or a posterior predictive draw of IBD segments [46]). These segments underpin many applications from disease mapping (e.g. [47]) to *P. falciparum* selection detection [23]. They can also be used to generate a recombination-sense IBD estimate [2]. However, like [15] we do not tune the parameter which relates to segment length, k. As such the pointwise estimate $\widehat{r}_m$, averages over all IBD segments, however small, and thus is liable to reflect some antecedent population-level relatedness, i.e. linkage disequilibrium (LD) [48]).

3.2 Fraction IBS

For a pair of samples i and j, we define the fraction IBS as a proportion of m markers that are identical across both samples,

$$\widehat{\mathrm{IBS}}_m = \frac{1}{m} \sum_{t=1}^{m} \mathrm{IBS}_t, \tag{3.2}$$

where $\mathrm{IBS}_t = 1$ if $Y_t^{(i)} = Y_t^{(j)}$ and zero otherwise. The fraction IBS can also be derived from counts of marker differences; e.g. $\sum_{t=1}^{m}(1 - \mathrm{IBS}_t) = m(1 - \widehat{\mathrm{IBS}}_m)$ where $1 - \mathrm{IBS}_t = 1$ denotes

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

a marker difference (equation (3.2) rearranged). Note that $\widehat{\text{IBS}}_m$ is the haploid equivalent of the 'allele-sharing coefficient' referred to in [2].

We can relate the IBS estimator (equation (3.2)) to r by specifying a relationship between IBD_t and IBS_t. For illustration, in the case with no genotyping error, the expectation of IBS_t is the following linear function of relatedness (derivation, equation (A.1)),

$$\mathbb{E}\left[\widehat{\text{IBS}}_m\right] = \bar{h}_m + (1 - \bar{h}_m)r, \tag{3.3}$$

where

$$\bar{h}_m = \frac{1}{m}\sum_{t=1}^{m} h_t \text{ and } h_t = \sum_{l=1}^{K_t} f_t(g_l)^2. \tag{3.4}$$

Here h_t and $1 - h_t$ are equivalent to Nei's 'gene identity' and 'gene diversity', respectively [49]. Considering an outbred diploid, these terms equate to homozygosity and heterozygosity, respectively [49].

Equation (3.3) might suggest that $\widehat{\text{IBS}}_m$ could converge to $\bar{h} + (1 - \bar{h})r$ (where $\bar{h} = \lim_{m\to\infty} \bar{h}_m$) as the number of markers m goes to infinity, under assumptions on the data-generating process such as independent loci; see section A.2. Under this setup, the estimator $\widehat{\text{IBS}}_m$ would not be consistent for r, but could be corrected; see Appendix A. In the Results section we numerically demonstrate how equation (3.3) is a problematic estimator of r using simulated and real data (see details below).

3.3 *Plasmodium* data

Throughout, we illustrate results using *Plasmodium* and simulated data. *P. falciparum* data comprise biallelic (i.e. $K_t = 2\ \forall t = 1, \ldots, m$) single nucleotide polymorphism (SNP) data from monoclonal *P. falciparum* samples (Table 1). All data are published [50, 51, 29, 30, 52, 8]. They were obtained either from sparse genome-wide panels of select markers, called barcodes, or from dense whole genome sequencing (WGS) data sets (reviewed in [53]); full details of sample collection and data generation can be found via the citations above and references therein. Additional steps we took to process the data are as follows.

Besides mapping SNP positions to the *P. falciparum* 3d7 v3 reference genome and recoding heteroallelic calls as missing (since all samples with fewer than 10 heteroallelic SNP calls were classified monoclonal by [50]), we did not post-process the Colombian data in any way. Thailand 93-SNP and WGS samples were used exactly as described in [8]. Data derived from [29, 30] (i.e. all African data) were processed using steps described in 'Sample and SNP cut-off selection criteria' of [29]. In addition, we removed samples with duplicate SNP calls; removed samples classified as not monoclonal using a $\leq 5\%$ heteroallelic SNP call rate to classify monoclonal samples, akin to [51]; and, among monoclonal samples, treated heteroallelic SNP calls as missing and removed monomorphic SNPs.

For each processed data set of monoclonal *P. falciparum* samples, allele frequencies were estimated by simple proportions: $f_t(g_j) = n_*^{-1}\sum_{i=1}^{n_*} \mathbb{1}(Y_t^{(i)} = g_j)$ for all $j = 1, 2$ and each locus t, where $n_* \leq n$ denotes the number of monoclonal parasite samples whose data were not missing at the t-th locus. Minor allele frequencies (i.e. $\min(f_t(g_1), f_t(g_2))\ \forall\ t = 1, \ldots m$) vary considerably by design (i.e. different marker panels) and due to variation among parasite populations over space and time (Figure 2).

In addition to the aforementioned *P. falciparum* data, we generated results for a single *P. vivax* data set, freely available online [54]. The *P. vivax* data were collected between 2010 and 2014 from

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

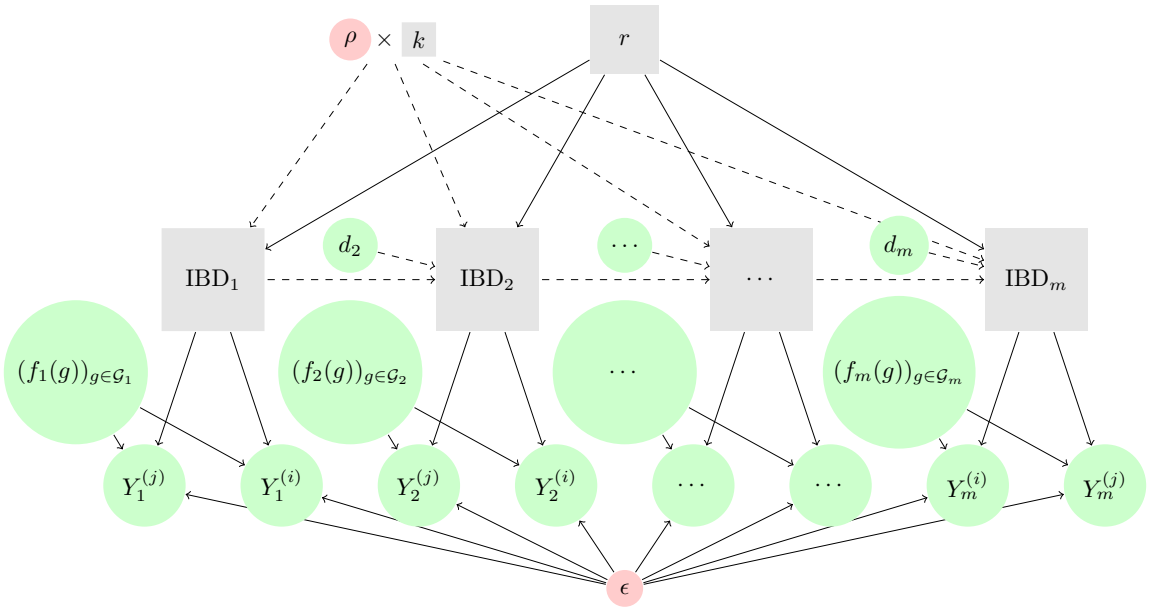


Figure 1: Models relating genetic data to genetic relatedness. Input data are depicted by green circles. They include genotype calls, $Y_1^{(i)}, \ldots, Y_m^{(i)}$ and $Y_1^{(j)}, \ldots, Y_m^{(j)}$, allele frequencies, $(f_1(g))_{g\in\mathcal{G}_1}, \ldots, (f_m(g))_{g\in\mathcal{G}_m}$, and distances between genotyped markers, $d_2, \ldots, d_m$. Parameters considered fixed (the genotyping error, ϵ, and the constant, ρ) are depicted by red circles. Unobserved quantities are depicted by gray squares. They include the hidden IBD states, $\text{IBD}_1, \ldots, \text{IBD}_m$, and the estimands (genetic relatedness, r, and k). Solid arrows depict dependencies under both the independence model and the HMM. Dashed arrows depict dependencies under the HMM only. Distances, ρ and k feature in the HMM only.

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Data set and citation/s	Collection region and years	n	$m_{\max}$	$\bar{h}_{m_{\max}}$	$\bar{K}'_{m_{\max}}$
Colombia [50]	Colombian Pacific region, 1993-2007	325	250	0.66	1.57
Thailand 93-SNP [51, 8]	Thailand-Myanmar border, 2001-2010	1173	93	0.57	1.77
Thailand WGS [52, 8]	Thailand-Myanmar border, 2001-2014	178	34911	0.89	1.17
The Gambia [29]	Kombo coastal districts, 2007–2008	71	31	0.77	1.37
Kilifi [29]	Coastal Kenya, 1998-2010	628	127	0.87	1.19
Western Kenya [30]	Western Kenya, 2008-2010	182	59	0.73	1.43

Table 1: A summary of globally diverse data sets of monoclonal *P. falciparum* samples. All data are published. Full details of sample collection and data generation can be found via the citations above and references therein. Additional steps we took to process the data for use in this study are described in section 3.3. For each processed data set, n denotes the number of monoclonal *P. falciparum* samples; $m_{\max}$ denotes the maximum number of successfully genotyped SNPs per sample; $\bar{h}_{m_{\max}}$ denotes the expected homozygosity (equation (3.4)) averaged over $m_{\max}$ SNPs; and $\bar{K}'_{m_{\max}}$ denotes the average effective cardinality (defined below, equation (3.8)).

two clinical trails on the Thailand-Myanmar border [55, 56]. The data were genotyped at three to nine highly polyallelic microsatellites (MS). In this study, we analyse samples genotyped at nine MSs that have no evidence of multiclonality (detection of two or more alleles at one or more MS). We estimate relatedness between pairs of samples from $n = 204$ different people, selecting one episode per person uniformly at random from all episodes per person. We use the allele frequencies reported in [54]. They have average expected homozygosity $\bar{h}_{m_{\max}} = 0.10$ (equation (3.4)) and average effective cardinality (defined below, equation (3.8)) $\bar{K}'_{m_{\max}} = 13.03$. Since there are only nine markers, we analyse these data under the independence model.

3.4 Simulated data

3.4.1 Biallelic markers:

Unless otherwise stated, biallelic marker data (i.e. data with $K_t = 2\ \forall t = 1, \ldots, m$) were simulated under the HMM with $\varepsilon = 0.001$ using marker loci positions and allele frequency estimates sampled from the Thailand WGS data set. Positions were sampled uniformly at random. Frequencies were sampled separately using one of two approaches: either they were sampled uniformly at random, or, to compensate for the skew towards rare alleles in WGS data set, frequencies were sampled separately with probability proportional to minor allele frequency estimates.

3.4.2 Polyallelic markers:

Polyallelic marker data (i.e. data with $K_t > 2\ \forall t = 1, \ldots, m$) were simulated under the HMM with $\varepsilon = 0.001$ using marker loci positions sampled uniformly at random from the Thailand WGS data set and allele frequency estimates sampled from a Dirichlet distribution. We used a Dirichlet parameter vector equal to $\alpha = (100_1, \ldots, 100_{K_t})$ to generate frequencies such that alleles are approximately equifrequent, and a concentration parameter vector equal to $\alpha = (1_1, \ldots, 1_{K_t})$ to generate frequencies that are uniform over the $K_t - 1$ simplex. The former approach generates ideal frequencies (see below), while the latter generates frequencies that for $K_t > 2$ are increasingly skewed towards rare alleles, thus more representative of real frequency spectra.

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

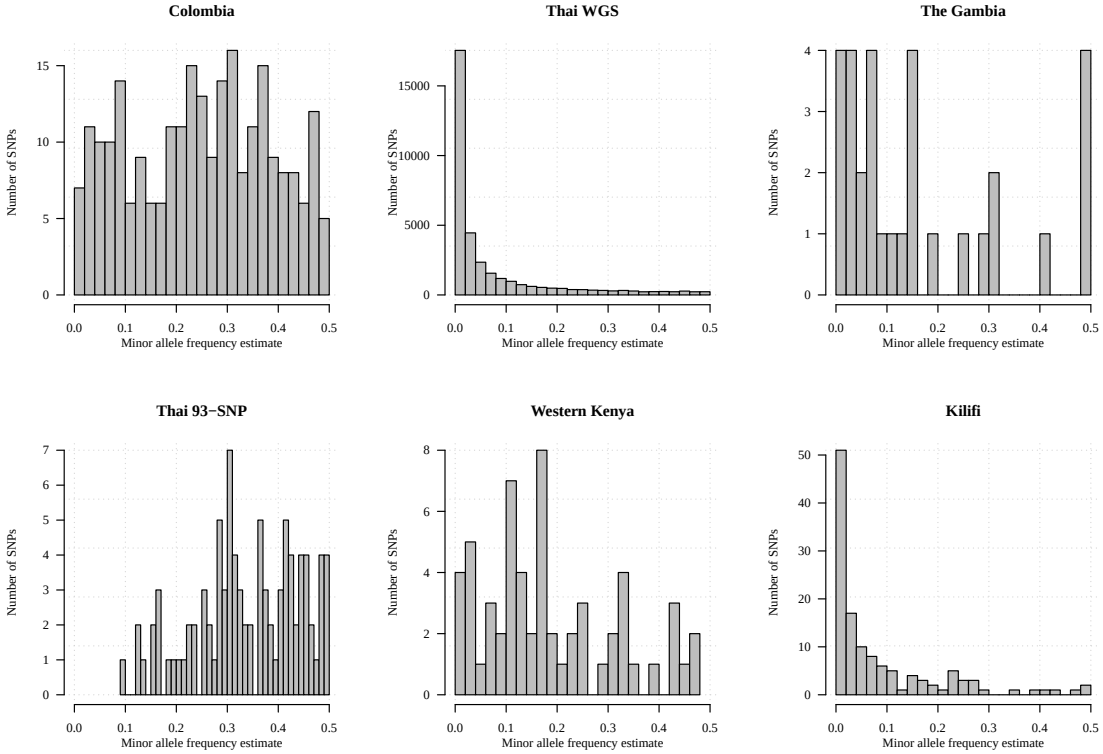


Figure 2: Minor allele frequency estimates from monoclonal *P. falciparum* data sets (Table 1).

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

3.5 Marker requirements for prospective relatedness inference

3.5.1 Biallelic markers:

For a set of parameters (i.e. number of markers m, relatedness r, switch rate parameter k) we simulate 1000 pairs of haploid genotype calls and for each pair compute $\widehat{r}_m$. We compute the root mean squared error (RMSE) by taking the square root of the average of the squared difference between $\widehat{r}_m$ and r. From the RMSEs computed for different sample sizes m, we derive the number of markers required for the RMSE to be below a specified value. Unless otherwise stated we use $k = 12$ where fixed, the mean estimate of k for $\widehat{r}_m \in (0.475, 0.525)$ from the Thailand WGS data set.

Comparison between $\widehat{r}_m$ and r differs from that between $\widehat{r}_m$ and $m^{-1}\sum_{t=1}^{m}$ IBD, which is referred to as the 'realised relatedness' in [2]. The former has the advantage of revealing RMSE due to the finite length of the genome (i.e. Mendelian sampling [34]), while at the same time revealing the excess, and thus theoretically avoidable, error due to marker limitations.

3.5.2 Polyallelic markers:

To explore marker requirements for relatedness inference using polyallelic markers we first consider the impact of increasing K_t beyond two at a single locus. For a given K_t, we measure the informativeness of a set of allele frequencies via the corresponding Fisher information matrix; this can in turn be related to the precision of the maximum likelihood estimator if the log-likelihood is approximately quadratic. We define $\mathrm{FIM}_t = \mathbb{E}[-\nabla_r^2 \log \mathbb{P}(Y_t^{(i)}, Y_t^{(j)}; r)]$, where the expectation is with respect to $Y_t^{(i)}, Y_t^{(j)}$ given r and the allele frequencies, and where we consider an observation model without genotyping error for simplicity; the sign ∇_r^2 stands for the second order derivative with respect to r. The Fisher information matrix FIM_t depends on the allele frequencies $(f(g_j))_{j=1}^{K_t}$ and on r:

$$\mathrm{FIM}_t(f_t(g_1), \ldots, f_t(g_{K_t}), r) = \frac{1}{1-r} + \sum_{j=1}^{K_t} \left\{ \frac{f_t(g_j)(1-f_t(g_j))^2}{r + f_t(g_j)(1-r)} - \frac{f_t(g_j)^2}{1-r} \right\}. \tag{3.5}$$

We show that, for any K_t and r, it is maximized over all $(f(g_j))_{j=1}^{K_t}$ by $f(g_j) = K_t^{-1}$ for all j, i.e. by equifrequent alleles. This is in agreement with the aforementioned long-established result that markers with high minor allele frequencies are preferable for relatedness inference [57]. A proof is provided in Appendix B.3.4. When alleles are equifrequent we obtain

$$\mathrm{FIM}_t(K_t, r) = \frac{1}{1-r} + \frac{(K_t-1)^2}{K_t(1+(K_t-1)r)} - \frac{1}{K_t(1-r)}, \tag{3.6}$$

which is an increasing function of K_t such that $\lim_{K_t\to\infty} \mathrm{FIM}_t(K_t) \to (1-r)^{-1} + r^{-1}$. Equation (3.6) describes the precision of the MLE, assuming that the log-likelihood is approximately quadratic, that K_t is the same at each locus and the allele frequencies are equifrequent.

To explore the relative gain of increasing $K_t > 2$ we calculate the multiplicative increase in $\mathrm{FIM}_t(K_t \geq 2, r)$ relative to $\mathrm{FIM}_t(K_t = 2, r)$ (Figure 3, left). The informativeness of $K_t = 15$ is between approximately two and seven times that of $K_t = 2$, with increasing returns as r approaches zero. However the justification of the FIM as a measure of precision breaks at the boundary of the parameter space. Regardless of r, the biggest increase is obtained upon increasing K_t from 2 to 3

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

with diminishing returns thereafter. The plot on the right of Figure 3 shows multiplicative increase in precision as a function of effective cardinality,

$$K'_t = 1/h_t, \tag{3.7}$$

which can be interpreted as the non-integer number of equifrequent alleles that would give rise to the same h_t as that based on the allele frequencies $(f_t(g))_{g\in\mathcal{G}_t}$ (equation (3.4)). For example, $K'_t = 2$ is the effective cardinality of an ideal biallelic SNP, whereas $K'_t < 2$ is the effective cardinality of a realistic biallelic SNP. Precision increases with K'_t as it does with K_t.

To explore the trade-off between increasing m and increasing K_t for a set of parameters (i.e. various m, K_t, α and r, and $k = 12$), we simulate 1000 pairs of haploid genotype calls, generate $\widehat{r}_m$ for each pair and calculate the RMSE. For simplicity, for a given m, we assume all markers have the same K_t. To compare on a common scale numerical results for makers with and without equifrequent alleles, we use the average effective cardinality:

$$\bar{K}'_m = \frac{1}{m}\sum_{t=1}^{m} K'_t. \tag{3.8}$$

Since $\bar{K}'_m$ is approximately the same for all m, to explore the trade-off between increasing m and increasing K_t, we average the effective cardinality over all $m \in \{24, 96, 192, 288, 384, 480\}$,

$$\bar{K}'_{m_{\text{cum}}} = \frac{1}{m_{\text{cum}}}\sum_{t=1}^{m_{\text{cum}}} K'_t, \tag{3.9}$$

where $m_{\text{cum}} = 24 + 96 + \ldots + 480$.

3.6 Data and code availability

All data used in this study are either simulated or published previously. Additional steps we took to process the data are described in section 3.3. The processed data and code necessary for confirming the conclusions of the article are available at github.com/artaylor85/PlasmodiumRelatedness.

4 Results

This section concerns the estimation of r as defined above and is arranged as follows. First we consider the genomic fraction IBS and show how it is problematic as an estimator of r. Second, we discuss r estimated using *Plasmodium* data, and provide marker requirements based on simulated data with biallelic and polyallelic markers.

4.1 Fraction IBS

Although $\widehat{\text{IBS}}_m$ might not satisfy favorable statistical properties as an estimator of r, its expectation is indeed related to r (equation (3.3)). As such, many studies have recovered meaningful trends in r with respect to epidemiological covariates (e.g. geographic distance) using measures related to $\widehat{\text{IBS}}_m$ [29, 30, 32]. However, since $\bar{h}_m$ is a function of the allele frequencies (equation (3.4)), so too is $\widehat{\text{IBS}}_m$. This is equivalent to the dependence on MAFs of the allele-sharing coefficients reviewed

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

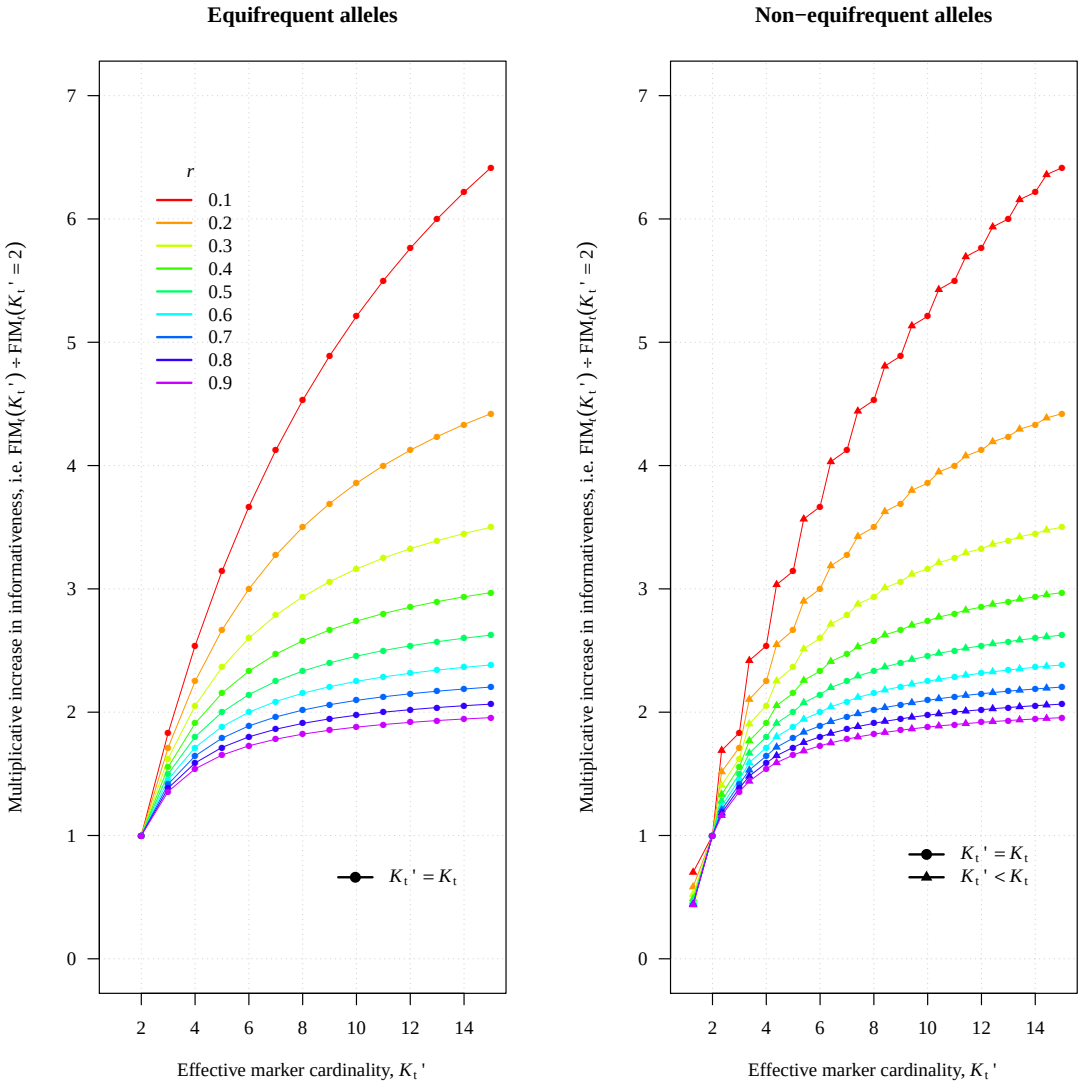


Figure 3: Multiplicative increase in the precision of the MLE with marker cardinality. The left plot shows the multiplicative increase for equifrequent alleles according to equation (3.6). The right plot shows the multiplicative increase with K_t', where precision was calculated according to equation (3.5) with either $f_t(g_i) = {}^1/_{K_t}\ \forall\ i = 1, \ldots, K_t$ (dots) or $f_t(g_1) = {}^{1.75}/_{K_t}$ and $f_t(g_i) = {}^{(1\ -\ f_t(g_1))}/_{(K_t\ -\ 1)}\ \forall\ i = 2, \ldots, K_t$ such that $K_t' < K_t$ (triangles).

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

in [2]. It means that quantitative trends in $\widehat{\text{IBS}}_m$ (e.g. regression coefficients) and absolute values of $\widehat{\text{IBS}}_m$ are only comparable across data on markers whose allele frequencies are the same [32]. This is a limitation given that frequencies almost always differ across data sets (e.g. Figure 2).

To illustrate the effect of differing allele frequencies, we generated $\widehat{\text{IBS}}_m$ for data simulated using allele frequency estimates from published data sets (Figure 4, top). The plot illustrates two notable results. First, for all biallelic marker data sets, the $\widehat{\text{IBS}}_m$ distribution is far from 0.5, the value of r used to simulate the data. The expected difference between $\widehat{\text{IBS}}_m$ and r is a linearly decreasing function of r: $\bar{h}_m + (1 - \bar{h}_m)r - r = \bar{h}_m - \bar{h}_m r$. As such, we would expect to see bigger and smaller distances between $\widehat{\text{IBS}}_m$ distributions and the data generating r were data simulated using $r < 0.5$ and $r > 0.5$, respectively; when $r = 1$ there is no difference. Second, the locations of the $\widehat{\text{IBS}}_m$ distributions vary considerably across data sets, centering around $\bar{h}_m + (1 - \bar{h}_m)r$, which varies due to $\bar{h}_m$, since $r = 0.5$ throughout. This variability, despite all parasite pairs having been simulated with $r = 0.5$, renders absolute values of $\widehat{\text{IBS}}_m$ non-portable across data sets. In contrast, distributions of estimates of relatedness based on IBD, $\widehat{r}_m$, all centre around $r = 0.5$ (Figure 4, bottom). To single out the effect of frequencies, we fixed all parameters besides frequencies across the data sets, including the number of markers ($m = 59$) and their positions; see caption, Figure 4. Results generated using all available data show the same trend, but data sets with many SNPs have tighter distributions.

Figure 5 shows $\widehat{\text{IBS}}_m$ and $\widehat{r}_m$ distributions based on the real sample pairs from published data sets. The location and spread of the $\widehat{\text{IBS}}_m$ distributions vary considerably. As Figure 4 exemplified using simulated data, comparisons of absolute $\widehat{\text{IBS}}_m$ values are non-portable across data sets. It is thus wrong to interpret the left-most centering of the distribution based on the real 93-SNP data set from Thailand as evidence that *P. falciparum* parasites from Thailand are less related than those from Kenya, or that they represent a different population to that represented by the WGS data set also from Thailand. Despite very different absolute values of $\widehat{\text{IBS}}_m$, careful inspection shows all center around $\bar{h}_{m_{\max}}$, which is the expectation of $\widehat{\text{IBS}}_m$ for unrelated parasites pairs whose $r = 0$ (equation 3.3). We thus conclude that many parasite pairs in these real data sets are unrelated. Our conclusion is corroborated by estimates of relatedness based on IBD, $\widehat{r}_m$ (Figure 5, bottom). Though the vast majority of parasite pairs are unrelated, we see some variation in the mean $\widehat{r}_m$. This variation is caused in part by outliers (parasite pairs with high relatedness). It is also caused by variation in $m_{\max} \times \bar{K}'_{m_{\max}}$: estimates of r are more error prone when $m_{\max} \times \bar{K}'_{m_{\max}}$ is small and those close to zero (the vast majority) are liable to upward bias due to boundary effects; see next section. The distribution of $\widehat{\text{IBS}}_m$ based on the *P. vivax* data set (Thailand MS) most closely approximates its partner distribution of $\widehat{r}_m$ due to the highly polymorphic nature of the microsatellite data whose $\bar{K}'_{m_{\max}} = 13.03$.

4.2 Estimating relatedness

Distributions of estimates of relatedness between pairs of *Plasmodium* monoclonal samples are plotted in Figure 5 (bottom plot). For each site, $\widehat{r}_m$ values range from 0 to 1, suggesting presence of unrelated, partially related and clonal parasites across all data sets. The vast majority, however, have $\widehat{r}_m < 0.20$. For a selection of 100 estimates ranging from 0 to 1, Figure 6 shows 95% parametric-bootstrap confidence intervals. In general, confidence intervals are tighter around estimates for data sets with larger $m_{\max} \times \bar{K}'_{m_{\max}}$, a point we shall return to later. Due to the asymmetric nature of confidence intervals near zero and one, estimates of r close to the boundary are liable to be biased. Considering the boundaries, intervals around estimates of r close to one are tighter, in general, than

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

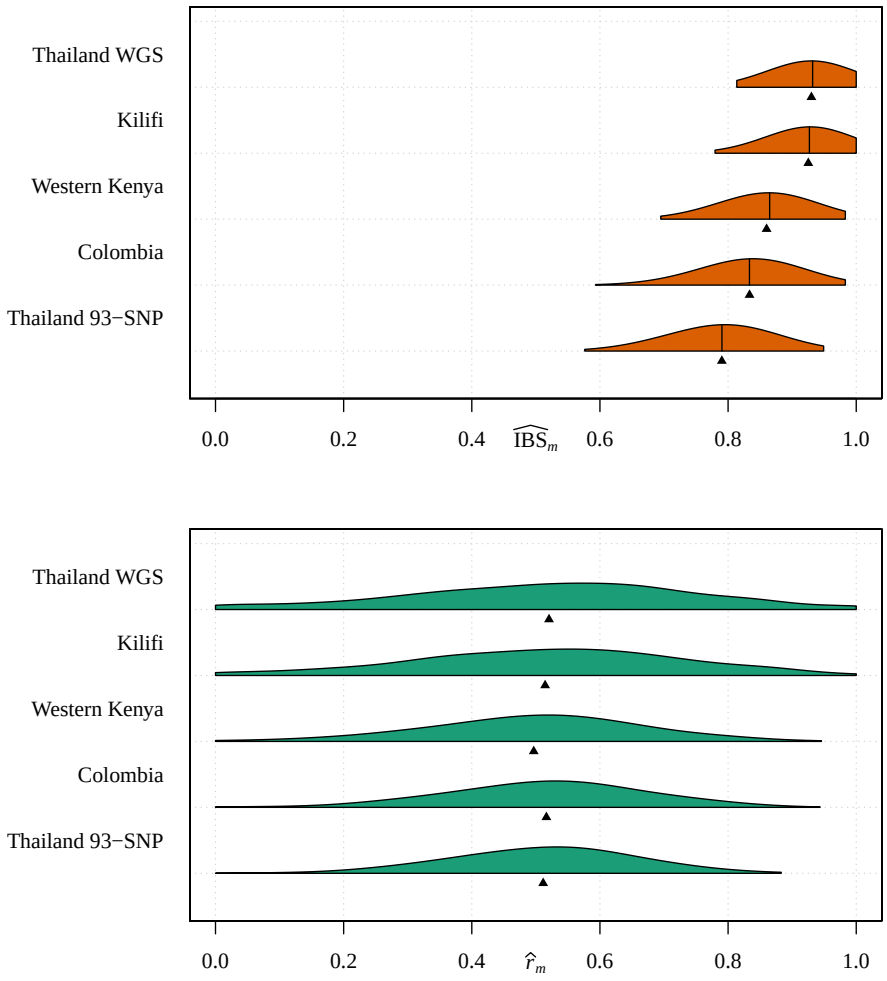


Figure 4: Measures of relatedness: pairs simulated with relatedness 0.5: Violin plots showing distributions of $\widehat{\mathrm{IBS}}_m$ (top) and $\hat{r}_m$ (bottom) each based on 100 pairs simulated using $r = 0.5$ and allele frequency estimates based on *P. falciparum* data sets with at least 59 SNPs (Table 1). To single out the effect of frequencies, we fixed all other parameters across the data sets, including the number of SNPs simulated and their positions. Specifically, we used 59 SNPs whose positions were extracted from the Western Kenyan data set. Allele frequencies were sampled uniformly at random from the full set of allele frequency estimates based on each data set. For each set of 59-SNP allele frequencies, the $\bar{h}_m$ values were 0.86, 0.85, 0.73, 0.67, 0.58 (top to bottom row of each plot, respectively). Data were simulated under the HMM with $\varepsilon = 0.001$, $r = 0.5$ and $k = 1$. Black vertical bars denote $\bar{h}_m + (1 - \bar{h}_m)r$ (top) and triangles denote the mean $\widehat{\mathrm{IBS}}_m$ (top) and mean $\hat{r}_m$ (bottom).

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

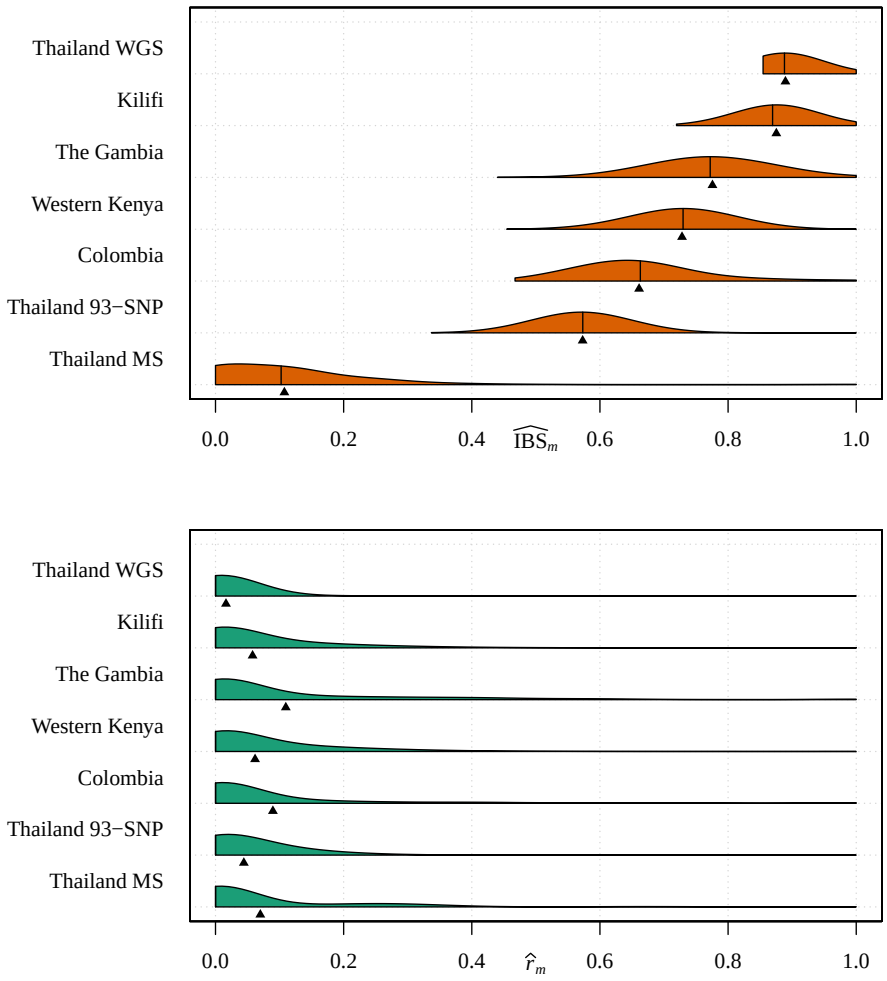


Figure 5: Measures of relatedness: parasite pairs with unknown relatedness. Violin plots showing distributions of $\widehat{\text{IBS}}_m$ (top) and $\widehat{r}_m$ (bottom) based on pairwise comparisons of *Plasmodium* monoclonal samples from six published *P. falciparum* biallelic SNP data sets (Table 1) and a single *P. vivax* microsatellite data set (Thailand MS). Black vertical bars denote $\bar{h}_{m_{\max}}$ (top) and triangles denote the mean $\widehat{\text{IBS}}_m$ (top) and mean $\widehat{r}_m$ (bottom).

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

those for r close to zero.

Considering Figures 5 and 6, we used the HMM to generate $\widehat{r}_m$ for biallelic marker data sets whose $m_{\max}$ is greater than 24 (Table 1), and the independence model to generate $\widehat{r}_m$ for the polyallelic data set whose $m_{\max} = 9$, Thai MS. Based on simulated data, the HMM provides coverage[3] close to 0.95 for $m > 24$, while the independence model provides waning coverage for $m > 24$, especially when k, which parameterizes the switching rate of the Markov chain, is small; for $m = 24$ both the HMM and the independence model provide similar coverage, above or around 0.85 (Figure 7).

For all data sets biallelic and polyallelic, we construct confidence intervals using the parametric bootstrap due to the non-quadratic nature of the log-likelihood of r when $\widehat{r}_m$ is close to either 0 or 1 (e.g. Figure B.3, left top and middle). For $\widehat{r}_m$ away from 0 and 1, the log-likelihood is quadratic (e.g. Figure B.3, bottom left plot) and thus normal-approximation confidence intervals could be constructed.

4.3 Marker requirements for prospective relatedness inference

As Figure 4 exemplified using simulated data, estimates of $\widehat{r}_m$ concentrate around the value of r used to simulate the data but with large variability, in part due to the finite length of the genome and in part due to limited data. We now consider how large m needs to be to estimate r with specified RMSE using different marker types (i.e. considering $K_t = 2\ \forall t = 1, \ldots, m$ and, more generally, $K_t \geq 2$).

4.3.1 Biallelic markers:

First we consider relatedness inference using biallelic markers (e.g. SNPs, the most abundant polymorphic marker type, commonly used for relatedness inference [1]).

Figure 8 shows the RMSE of $\widehat{r}_m$ generated under the HMM given allele frequencies drawn with probability equal to their minor allele frequencies ($\bar{h}_m \approx 0.69$, $\bar{K}'_m \approx 1.53$) versus allele frequencies drawn uniformly at random ($\bar{h}_m \approx 0.89$, $\bar{K}'_m \approx 1.17$). There are three notable results. First, errors obtained using allele frequencies drawn at uniformly at random are smaller (Figure 8, left). This is in agreement with the long-established result that higher minor allele frequencies are preferable for relationship inference [57]. Second, the RMSE is relatively large for 24 markers, decreasing dramatically upon increasing the marker count to 96. Though less dramatic, the decrease in RMSE is appreciable up to 288 markers, with diminishing returns thereafter. RMSE does not tend to zero due to the finite length of the genome. Third, RMSE error decreases with increasing proximity of the data-generating r to either 0 or 1 (especially the latter). As such, biallelic marker requirements for inference of $r = 0.5$ constrain guidelines for inference of r in general (Table 2).

4.3.2 Polyallelic markers:

Highly polyallelic microsatellite length polymorphism markers have long been used for relatedness inference, and there is growing interest in using microhaplotypes (short highly diverse regions of the genome) [1, 58]. Neither microsatellite nor microhaplotypes are point polymorphisms. However, to explore the general utility of polyallelic markers for relatedness inference, we make the simplifying assumption that they are.

[3] Coverage is equal to the fraction of confidence intervals that contain the data generating r.

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

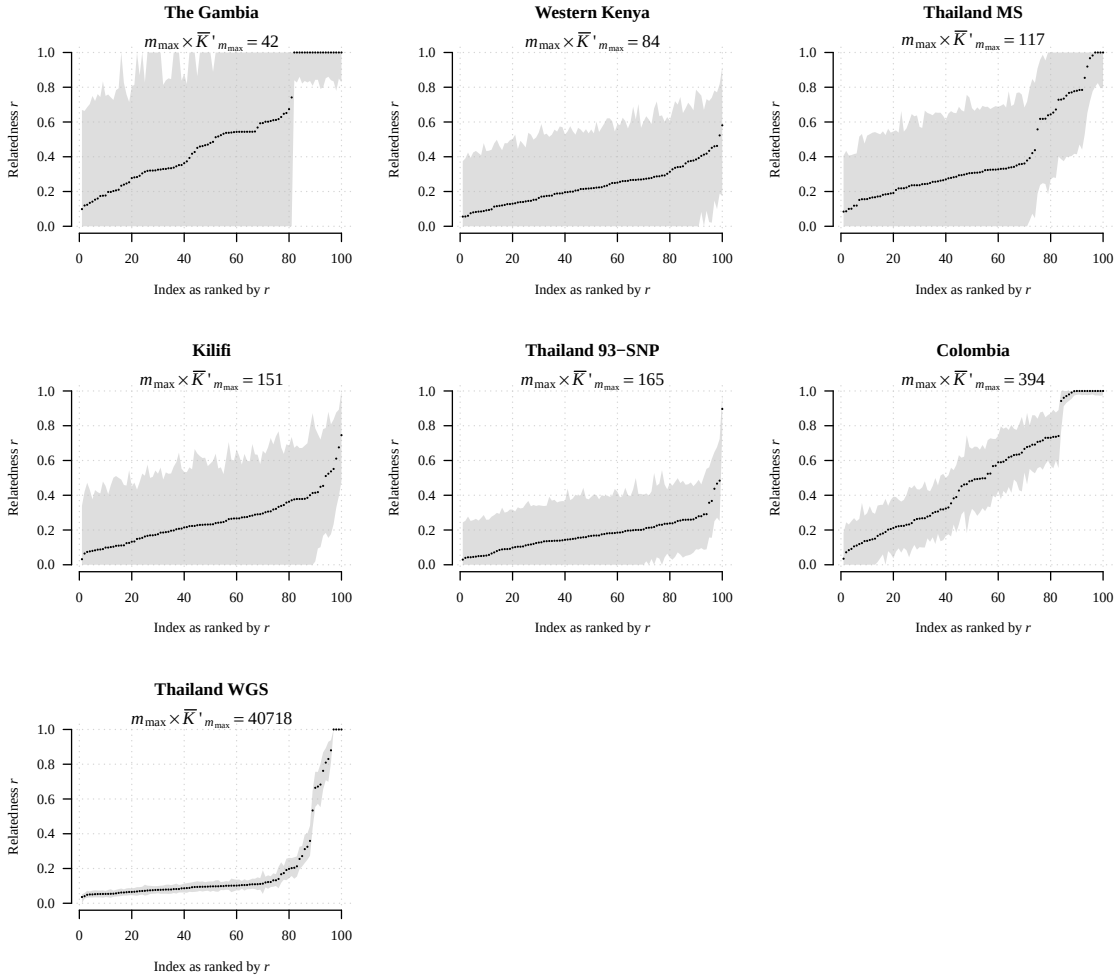


Figure 6: $\widehat{r}_m$ with 95% confidence intervals for 100 select pairwise comparisons of monoclonal *Plasmodium* samples from *P. falciparum* data sets list in Table 1 and a single *P. vivax* data set, Thai MS.

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

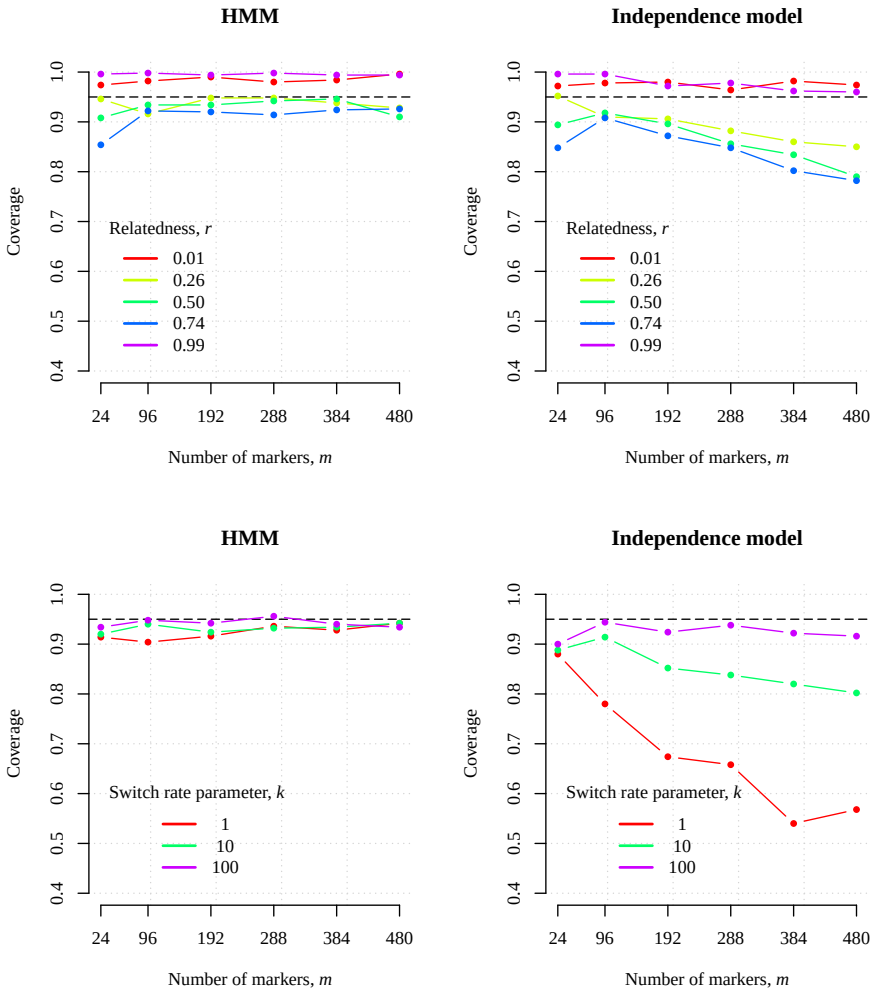


Figure 7: Coverage of 95% parametric bootstrap confidence intervals constructed under the HMM (left) and independence model (right). Coverage is equal to the proportion of 500 $\widehat{r}_m$ whose 95% parametric bootstrap confidence intervals contain the value of r used to simulate the data. It was based on data simulated under the HMM with $\epsilon = 0.001$. Data were simulated for m biallelic markers (i.e. $K_t = 2 \; \forall \; t = 1, \ldots, m$). Plots on the top show coverage for data simulated with different values of r given fixed $k = 12$. Plots on the bottom show coverage for data simulated with different values of k given fixed $r = 0.5$.

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

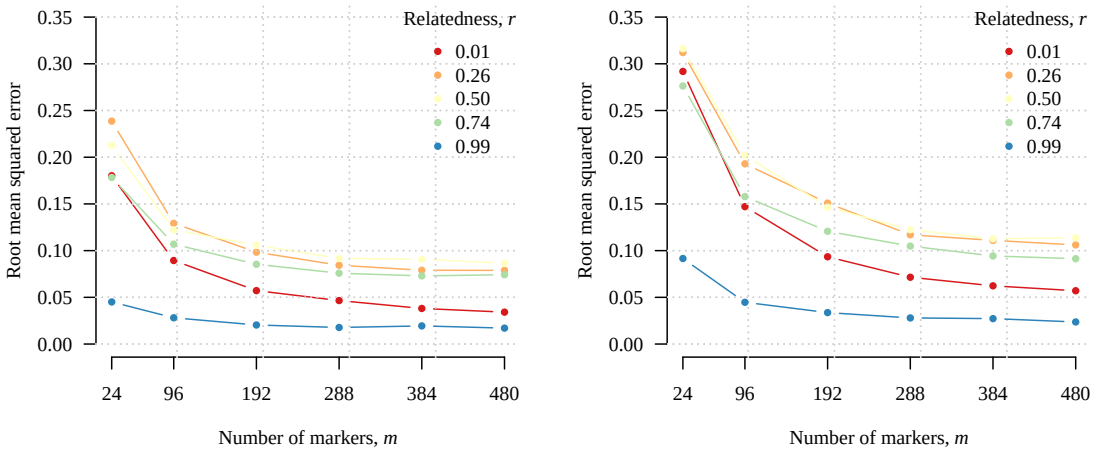
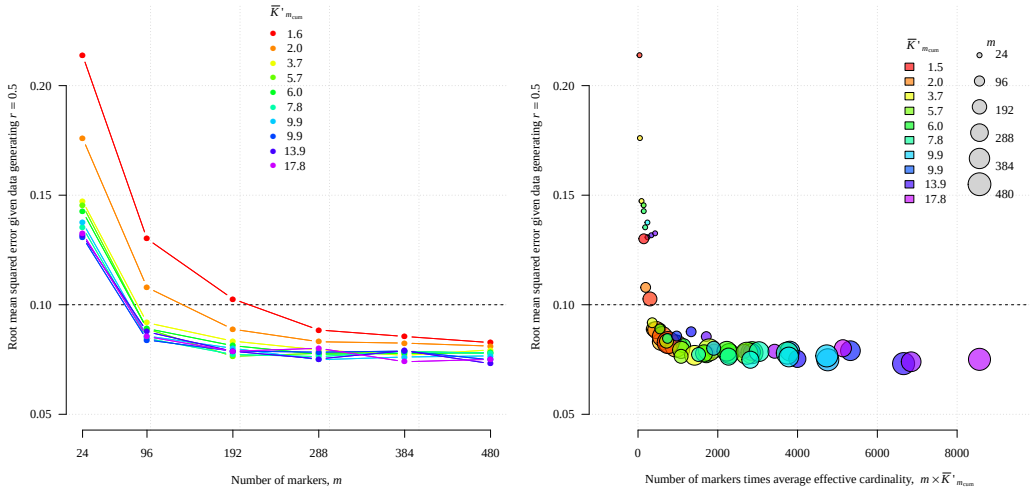


Figure 8: RMSE of $\widehat{r}_m$ generated under the HMM. Data were simulated under the HMM using various r (see legend); allele frequencies with $\bar{h}_m \approx 0.69$, $\bar{K}'_m \approx 1.53$ (left plot) and $\bar{h}_m \approx 0.89$, $\bar{K}'_m \approx 1.17$ (right plot); $\varepsilon = 0.001$, $k = 12$, $K_t = 2 \forall t$.

RMSE	$r = 0.01$	$r = 0.50$	$r = 0.99$	Any $r \in (0, 1)$†
0.00	$> L$	$> L$	$> L$	$> L$
0.05	192-288	>480	<24	>480
0.10	24-96	96-192	<24	192
0.15	24-96	24-96	<24	96
0.20	<24	24-96	<24	96

Table 2: Biallelic marker requirements for specified RMSE around $r \in \{0.01, 0.50, 0.99\}$ and any $r \in (0, 1)$ extracted from Figure 8, left (i.e. given allele frequencies with $\bar{h}_m \approx 0.69$). The length of the genome is denoted by L. †Since $r = 0.5$ has the largest marker requirements in general, inference of any $r \in (0, 1)$ is given by the maximum of the marker requirement interval for $r = 0.5$.

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



(a) As a function of m for various $\bar{K}'_{m_{\text{cum}}}$.

(b) As a function $m \times \bar{K}'_{m_{\text{cum}}}$.

Figure 9: RMSE of $\widehat{r}_m$ around data generating $r = 0.5$ with number of markers, m, and average effective cardinality, $\bar{K}'_{m_{\text{cum}}}$.

Figure 9a shows three notable results. First, if only a small number of markers (e.g. 24) are available, a slight increase in their average effective cardinality markedly reduces RMSE, with diminishing returns as m grows. Second, to obtain RMSE less than some arbitrary amount, there may be an option between increasing m and increasing cardinality. For example, to obtain RMSE < 0.1, our results suggest typing 96 markers with $\bar{K}'_m > 2$ or around 192 markers with $\bar{K}'_m = 1.6$. This latter option agrees roughly with the requirements for $r = 0.5$ in Table 2. Third, within the range of m values explored here, markers with $K_t > 2$ are necessary for optimally low RMSE (i.e. to achieve RMSE comparable with Mendelian sampling and thus negligible RMSE due to marker limitations).

The results shown in Figure 9a are projected onto a single axis in Figure 9b, showing the synergistic effect of increasing both m and K'_t. Clearly, larger $m \times \bar{K}'_{m_{\text{cum}}}$ provides smaller RMSE with diminishing returns beyond $m \times \bar{K}'_{m_{\text{cum}}} \approx 1000$. Informally, this result provides intuition as to why we obtain, in general, tighter confidence intervals around $\widehat{r}_m$ based on *Plasmodium* data sets with larger $m \times \bar{K}'_{m_{\text{max}}}$ (Figure 6). Moreover, it suggests that the confidence intervals around the Thailand WGS estimates are as small as they can be.

5 Discussion

Using a simple model framework, we call attention to properties of estimates of genetic relatedness, r, increasingly used in genetic epidemiology of malaria. These results, though articulated around monoclonal haploid malaria parasites, are applicable more generally to haploid eukaryotes (highly recombining prokaryotes would require a modified model).

The fraction IBS, which does not distinguish between alleles shared due to ancestry versus

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

chance, is not a statistically principled estimator of r. As such, it does not allow calculation of confidence intervals for r, nor marker requirements. Its expectation is a correlate of r, but absolute values and quantitative estimates of trends are not portable across studies due to dependence on allele frequencies, which vary in space and time, and with different marker panels and quality control procedures [2]. On the contrary, measures based on IBS have the advantage of not depending upon potentially problematic allele frequency estimates discussed below [2, 4]. By illustrating how the fraction IBS is expected to change as a function of r and the alleles frequencies, we aid interpretation across studies using measures based on IBS to investigate relatedness.

Model-based relatedness inference allows construction of confidence intervals and marker requirements. Based on the parameters we explored, we recommend successful genotyping of at least 200 biallelic or 100 polyallelic markers for relatedness inference with RMSE less than 0.1 (if markers are highly polyallelic, fewer may be required, as in the Thai MS data set). In practice, a chosen set of makers could combine biallelic SNPs and more polyallelic marker types (e.g. microhaplotypes). Though not directly comparable, our results roughly agree (are of the same order of magnitude), with those reported for diploids and polyploids (Table 3). Relatedness inference for polyploids (e.g. [59, 13]) is comparable to that for polyclonal malaria samples, which arise due co-transmission and superinfection [60]. However, relatedness inference across polyclonal malaria samples is more challenging, since the equivalence of ploidy is unknown and variable. Despite these challenges, methods to infer relatedness within polyclonal malaria samples exist [23, 25], while methods to infer relatedness across polyclonal malaria samples are under development. It will be interesting to see how marker requirements, limited here to monoclonal malaria samples, scale in this more complex setting.

The results presented here are conditional upon an HMM under which various simplifying assumptions were made, the most significant being that of known and fixed allele frequencies. Typically, allele frequencies are estimated using data intended for relatedness inference yet assuming independent and identically distributed samples [61, 62]. These data-derived allele frequencies have been shown to give poor results and lead to underestimation of relatedness since "rare alleles shared by relatives are not recognized as such" [11]. Improving allele frequency estimates could benefit inference more than increasing the number of markers [11]. To better estimate allele frequencies of naturally occurring malaria parasites, for which pedigrees are unattainable, one could jointly model frequencies and relatedness as in [61]. Joint modelling would benefit inference in other ways also. For example, by borrowing information across samples and extending the inference framework, one could theoretically infer the ancestral recombination graph and thus the genetic map (presently assumed uniform across the malaria genome here and in [18, 23, 25]). That said, details specific to malaria (e.g. out-crossing versus selfing and their association with transmission) would present unique challenges (e.g. [63] and references therein). Modular extensions of pairwise methods to perform multi-way relatedness inference (e.g. [64]) have also been shown to outperform pairwise methods.

As formally stated in equation (3.6), we find that a highly polyallelic marker can be several times more informative than a biallelic marker for relatedness inference, comparable to results reported in population assignment [65]. Despite their superior informativeness, microsatellites are being superseded by SNPs due to the relative ease and reliability of typing the latter [1]. Recent interest in microhaplotypes (regions of high SNP diversity, unbroken by recombination) aims to combine the ease of SNPs with the informativeness of polyallelic markers [58]. Microhaplotypes can be defined *in silico*, using a decision theoretic criterion [66, 65], which relates to LD [48]. They can then be captured *in vitro* using amplicon sequencing [58, 67] or molecular inversion probes (MIPs),

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

which can also be used to genotype microsatellites and SNPs [68, 69, 70]. The amplicon and MIP approaches are especially valuable for relatedness inference across multiclonal malaria samples, because amplicons and MIPs can capture within-host densities of different parasite clones as well as the phase of microhaplotypes in polyclonal infections [67, 70]. A model that accurately reflects the fact that microsatellites and microhaplotypes are not point polymorphism, while accounting for their associated mutation and observation error rates, thus merits consideration [71, 72].

Besides motif repeats within microsatellites and SNPs within microhaplotypes (presently overlooked), it is preferable to minimise dependence between markers. For any given r and k, dependence is a function of marker position and LD. As such, marker position is an important design consideration. When considering polyallelic markers, we sampled marker positions uniformly at random from the Thailand WGS data set. For microhaplotypes, a more realistic approach would draw from genomic intervals whose length is amenable to physical phasing and high LD. Doing so presents a trade-off between distance and window-wise effective cardinality. This trade-off is critical if diverse windows are genomically clustered. We do not consider it here, but it can be explored within the current framework and is the topic of future work. On the other hand, LD is a natural phenomena over which investigators have no degree of freedom. Some models commonly used in human genetics account for LD [46, 73] (also see [15]). Those designed to estimate relatedness between malaria parasites account for dependence between IBD states due to their physical proximity but not due to LD [18, 23, 25]. LD reported in malaria parasite populations (e.g. [74, 50, 75]) is generally lower than that reported in human populations [76]. Its incorporation into methods for malaria parasite relatedness inference, both within and between polyallelic markers, warrants further research.

Here and elsewhere marker requirements are based on either down-sampled or simulated data (Table 3). Standard asymptotic theory for HMMs is problematic in the present setting due to the finite length of the genome, and the increasing degree of dependencies between markers as their density grows. Understanding the finite sample properties of the maximum likelihood estimator in this setting remains an open problem. Another open problem beyond the scope of this study, is that of sampling individuals for population-level inference (e.g. how many parasite samples are required to reliably infer gene flow between different geographic locations using relatedness?). Work is ongoing to address these questions, which are very application-specific and dependent on many population factors (e.g. transmission intensity, seasonality, asymptomatic reservoir, etc.).

6 Conclusion

For portability, we recommend estimates of relatedness based on IBD for malaria epidemiology. To generate estimates between monoclonal parasite samples with less than 10% RMSE, approximately 200 biallelic markers or 100 polyallelic markers are required. Where studies inevitably differ in terms of available genetic data, confidence intervals illuminate inference. Together with anticipated work on population-level sampling, we hope this work on genetic-level sampling (and extensions thereof) will aid statistically informed design of prospective molecular epidemiological studies of malaria.

7 Acknowledgements

We thank all the authors of the *Plasmodium* data sets for either sharing their data or making them freely available online for use here and elsewhere. Pierre E. Jacob gratefully acknowledges support

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

by the National Science Foundation through grant DMS-1712872. Aimee Taylor and Caroline Buckee are supported by a Maximizing Investigators' Research Award for Early Stage Investigators, R35GM124715. This project was funded in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under grant number U19AI110818 to the Broad Institute (Daniel Neafsey).

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Study	Result
Relatedness inference for close relatives using poor quality samples [77]	100 SNPs identified individuals and close relatives
Parentage inference in diploids using likelihood ratio test and numeric approximation of false positive and negative rates for different numbers of loci and genotyping error rates [16]	60-100 SNPs sufficient
Parentage and sibship inference in diploids [58] using method of [16]	96 microhaplotype loci
Ancestry assignment and coefficient inference in diploids via inverse expected Fisher information matrix [65]	4-125,000 biallelic SNPs, depending on allele frequencies and required precision
Relatedness inference in diploids using a variety of estimators and sub-sampling of empirical data on 86 microsatellites, each with 2 to 19 alleles [11]	"In this study a set of 34 polymorphic loci seemed to be a good balance between performance of estimators and marker genotyping costs"
Relatedness inference in autopolyploids using a variety of estimators and simulation [13]	Approximately 200 markers, each with 10 alleles, for 95% confidence interval of $r \pm 0.05$ around diploids
Joint parentage and sibship inference of polyploids whose genotypes are transformed into "pseudodiploid-dominant genotypes" to enable application of likelihood methods designed for diploids, using both simulated and empiric data [59]	10-20 microsatellites each having 10 alleles
Connectivity between malaria parasite populations based on relatedness between monoclonal *P. falciparum* parasite samples by subsampling empiric data [8]	96 SNPs sufficient to recover comparable trends to those obtained using WGS
Joint sibship inference in diploids and haplodiploids using maximum likelihood methods and simulation [61]	approx. 6-10 markers each with 10 alleles, or approx. 30-40 biallelic markers, depending on family size and error inclusion
Relatedness inference for zebra finch and pigs reviewed in [2]	More than 771 SNPs (for zebra finch) and 2000 SNPs (for pigs)

Table 3: A non-exhaustive selection of studies in which numbers of loci for relatedness and associated inference are reported. Most of the above studies assume independence between markers because methods that assume dependence are, in general, designed for marker-rich applications where data requirements are not an issue.

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

[1] B. S. Weir, A. D. Anderson, and A. B. Hepler. Genetic relatedness analysis: Modern data and new challenges. *Nature Reviews Genetics*, 7(10):771–780, 2006.

[2] D. Speed and D. J. Balding. Relatedness in the post-genomic era: Is it still useful? *Nature Reviews Genetics*, 16(1):33–44, 2015.

[3] Sewall Wright. Coefficients of inbreeding and relationship. *The American Naturalist*, 56:330–338, 1922.

[4] R. K. Waples, A. Albrechtsen, and I. Moltke. Allele frequency-free inference of close familial relationships from genotypes or low-depth sequencing data. *Molecular Ecology*, 28(1):35–48, 2019.

[5] J. Gardy, N. J. Loman, and A. Rambaut. Real-time digital pathogen surveillance — the time is now. *Genome Biology*, 16(1):155, 2015.

[6] R. E. Blanton. Population genetics and molecular epidemiology of eukaryotes. *Microbiology spectrum*, 6(6), 2018.

[7] E. A. Thompson. Identity by Descent: Variation in Meiosis, Across Genomes, and in Populations. *Genetics*, 194(2):301–326, 2013.

[8] A. R. Taylor, S. F. Schaffner, G. C. Cerqueira, S. C. Nkhoma, T. J. Anderson, K. Sriprawat, A. P. Phyo, F. Nosten, D. E. Neafsey, and C. O. Buckee. Quantifying connectivity between local plasmodium falciparum malaria parasite populations using identity by descent. *PLoS genetics*, 13(10):e1007065, 2017.

[9] A. Wesolowski, A. R. Taylor, H.-H. Chang, R. Verity, S. Tessema, J. Bailey, T. A. Perkins, D. Neafsey, B. Greenhouse, and C. O. Buckee. Mapping malaria by combining parasite genomic and epidemiologic data. *BMC Medicine*, 16(190), 2018.

[10] W. G. Hill. Sewall Wright's 'systems of mating'. *Genetics*, 143(4):1499–1506, 1996.

[11] M. C. Bink, A. D. Anderson, W. E. Van De Weg, and E. A. Thompson. Comparison of marker-based pairwise relatedness estimators on a pedigreed plant population. *Theoretical and Applied Genetics*, 117(6):843–855, 2008.

[12] G. Heckenberg, E. D. O. Roberson, J. D. Baugher, J. Pevsner, E. L. Stevens, and T. J. Downey. Inference of Relationships in Population Data Using Identity-by-Descent and Identity-by-State. *PLoS Genetics*, 7(9):e1002287, 2011.

[13] K. Huang, S. T. Guo, M. R. Shattuck, S. T. Chen, X. G. Qi, P. Zhang, and B. G. Li. A maximum-likelihood estimation of pairwise relatedness for autopolyploids. *Heredity*, 114(2): 133–142, 2015.

[14] A.-L. Leutenegger, B. Prum, E. Génin, C. Verny, A. Lemainque, F. Clerget-Darpoux, and E. A. Thompson. Estimation of the Inbreeding Coefficient through Use of Genomic Data. *The American Journal of Human Genetics*, 73(3):516–523, 2003.

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[15] M. D. Brown, C. G. Glazner, C. Zheng, and E. A. Thompson. Inferring coancestry in population samples in the presence of linkage disequilibrium. *Genetics*, 190(4):1447–1460, 2012.

[16] E. C. Anderson and J. C. Garza. The power of single-nucleotide polymorphisms for large-scale parentage inference. *Genetics*, 172(4):2567–2582, 2006.

[17] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[18] S. F. Schaffner, A. R. Taylor, W. Wong, D. F. Wirth, and D. E. Neafsey. HmmIBD: Software to infer pairwise identity by descent between haploid genotypes. *Malaria Journal*, 17(1):10–13, 2018.

[19] R. F. Daniels, S. F. Schaffner, E. A. Wenger, J. L. Proctor, H.-H. Chang, W. Wong, N. Baro, D. Ndiaye, F. B. Fall, M. Ndiop, et al. Modeling malaria genomics reveals transmission decline and rebound in senegal. *Proceedings of the National Academy of Sciences*, 112(22):7067–7072, 2015.

[20] A. N. Cowell, H. O. Valdivia, D. K. Bishop, and E. A. Winzeler. Exploration of Plasmodium vivax transmission dynamics and recurrent infections in the Peruvian Amazon using whole genome sequencing. *Genome Medicine*, 10(52):1–12, 2018.

[21] S. Auburn, E. D. Benavente, O. Miotto, R. D. Pearson, R. Amato, M. J. Grigg, B. E. Barber, T. William, I. Handayuni, J. Marfurt, H. Trimarsanto, R. Noviyanti, K. Sriprawat, F. Nosten, S. Campino, T. G. Clark, N. M. Anstey, D. P. Kwiatkowski, and R. N. Price. Genomic analysis of a pre-elimination Malaysian Plasmodium vivax population reveals selective pressures and changing transmission dynamics. *Nature Communications*, 9(1):1–12, 2018.

[22] S. Bopp, P. Magistrado, W. Wong, S. F. Schaffner, A. Mukherjee, P. Lim, M. Dhorda, C. Amaratunga, C. J. Woodrow, E. A. Ashley, N. J. White, A. M. Dondorp, R. M. Fairhurst, F. Ariey, D. Menard, D. F. Wirth, and S. K. Volkman. Plasmepsin II-III copy number accounts for bimodal piperaquine resistance among Cambodian Plasmodium falciparum. *Nature Communications*, 9(1), 2018.

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

[24] L. Henden, D. Wakeham, and M. Bahlo. XIBD: software for inferring pairwise identity by descent on the X chromosome. *Bioinformatics*, 32(March):2389–2391, 2016.

[25] S. J. Zhu, J. A. Hendry, J. Almagro-garcia, R. D. Pearson, R. Amato, A. Miles, D. J. Weiss, T. C. D. Lucas, P. W. Gething, D. Kwiatkowski, and G. Mcvean. The origins and relatedness structure of mixed infections vary with local prevalence of P . falciparum malaria. *bioRxiv*, 2018.

[26] S. J. Zhu, J. Almagro-garcia, and G. Mcvean. Deconvoluting multiple infections in Plasmodium falciparum from high throughput sequencing data. *bioRxiv*, 2017.

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[27] P. Orjuela-Snchez, M. Da Silva-Nunes, N. S. Da Silva, K. K. Scopel, R. M. Gonalves, R. S. Malafronte, and M. U. Ferreira. Population dynamics of genetically diverse Plasmodium falciparum lineages: Community-based prospective study in rural Amazonia. *Parasitology*, 136 (10):1097–1105, 2009.

[28] T. J. C. Anderson, J. T. Williams, S. Nair, D. Sudimack, M. Barends, A. Jaidee, R. N. Price, and F. Nosten. Inferred relatedness and heritability in malaria parasites. *Proceedings of the Royal Society of London B: Biological Sciences*, 277(1693):2531–2540, 2010.

[29] I. Omedo, P. Mogeni, T. Bousema, K. Rockett, A. Amambua-Ngwa, I. Oyier, J. C. Stevenson, A. Y. Baidjoe, E. de Villiers, G. Fegan, A. Ross, C. Hubbart, A. Jeffreys, T. N. Williams, D. Kwiatkowski, and P. Bejon. Micro-epidemiological structuring of Plasmodium falciparum parasite populations in regions with varying transmission intensities in Africa. *Wellcome Open Research*, 2(10), 2017.

[30] I. Omedo, P. Mogeni, K. Rockett, A. Kamau, C. Hubbart, A. Jeffreys, E. D. Villiers, A. Noor, B. Snow, D. Kwiatkowski, and P. Bejon. Geographic-genetic analysis of Plasmodium falciparum parasite populations from surveys of primary school children in Western Kenya. *Wellcome Open Research*, 2, 2017.

[31] K. M. Oyebola, O. O. Aina, E. T. Idowu, Y. A. Olukosi, O. S. Ajibaye, O. A. Otubanjo, T. S. Awolola, G. A. Awandare, and A. Amambua-Ngwa. A barcode of multilocus nuclear DNA identifies genetic relatedness in pre- and post-Artemether/Lumefantrine treated Plasmodium falciparum in Nigeria. *BMC infectious diseases*, 18(1):392, 2018.

[32] H.-H. Chang, A. Wesolowski, I. Sinha, C. G. Jacob, A. Mahmud, D. Uddin, S. I. Zaman, M. A. Hossain, M. A. Faiz, A. Ghose, A. A. Sayeed, M. R. Rahman, A. Islam, M. J. Karim, M. K. Rezwan, A. K. M. Shamsuzzaman, S. T. Jhora, M. M. Aktaruzzaman, O. Miotto, K. Engo-Monsen, D. Kwiatkowski, R. J. Maude, and C. O. Buckee. The geography of malaria elimination in bangladesh: combining data layers to estimate the spatial spread of parasites. *bioRxiv*, 2018.

[33] A. Miles, Z. Iqbal, P. Vauterin, R. Pearson, S. Campino, M. Theron, K. Gould, D. Mead, E. Drury, J. O. Brien, V. R. Rubio, B. Macinnis, J. Mwangi, U. Samarakoon, L. Ranford-cartwright, M. Ferdig, K. Hayton, X.-z. Su, T. Wellems, J. Rayner, G. Mcvean, and D. Kwiatkowski. Indels, structural variation and recombination drive genomic diversity in Plasmodium falciparum. *Genome Research*, 26(9):1288–1299, 2016.

[34] W. Hill and B. Weir. Variation in actual relationship as a consequence of mendelian sampling and linkage. *Genetics Research*, 93(1):47–64, 2011.

[35] P. J. Bickel, Y. Ritov, and T. Ryden. Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *The Annals of Statistics*, 26(4):1614–1635, 1998.

[36] R. Douc, E. Moulines, J. Olsson, and R. Van Handel. Consistency of the maximum likelihood estimator for general hidden Markov models. *the Annals of Statistics*, 39(1):474–513, 2011.

[37] R. Douc and E. Moulines. Asymptotic properties of the maximum likelihood estimation in misspecified hidden Markov models. *The Annals of Statistics*, 40(5):2697–2732, 2012.

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[38] C. J. Geyer. Asymptotics of maximum likelihood without the LLN or CLT or sample size going to infinity. In *Advances in Modern Statistical Theory and Applications: A Festschrift in honor of Morris L. Eaton*, pages 1–24. Institute of Mathematical Statistics, 2013.

[39] S. G. Self and K.-Y. Liang. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398):605–610, 1987.

[40] L. Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.

[41] M. D. Ramstetter, T. D. Dyer, D. M. Lehman, J. E. Curran, R. Duggirala, J. Blangero, J. G. Mezey, and A. Williams. Benchmarking relatedness inference methods with genome-wide data from thousands of relatives. *Genetics*, 207(July):75–82, 2017.

[42] T. Druet and M. Gautier. A model-based approach to characterize individual inbreeding at both global and local genomic scales. *Molecular Ecology*, 26(20):5820–5841, 2017.

[43] A. F. Herzig, T. Nutile, M.-C. Babron, M. Ciullo, C. Bellenguez, and A.-L. Leutenegger. Strategies for phasing and imputation in a population isolate. *Genetic epidemiology*, 42(2): 201–213, 2018.

[44] B. L. Browning, Y. Zhou, and S. R. Browning. A one-penny imputed genome from next-generation reference panels. *The American Journal of Human Genetics*, 103(3):338–348, 2018.

[45] S. Das, G. R. Abecasis, and B. L. Browning. Genotype Imputation from Large Reference Panels. *Annual Review of Genomics and Human Genetics*, 19(1):73–96, 2018.

[46] S. R. Browning. Estimation of pairwise identity by descent from dense genetic marker data in a population sample of haplotypes. *Genetics*, 178(4):2123–2132, 2008.

[47] S. R. Browning and E. A. Thompson. Detecting Rare Variant Associations by Identity-by-Descent Mapping in Case-Control Studies. *Genetics*, 190:1521–1531, 2012.

[48] M. Slatkin. Linkage disequilibrium–understanding the evolutionary past and mapping the medical future. *Nature reviews. Genetics*, 9(6):477–85, 2008.

[49] M. Nei. Analysis of Gene Diversity in Subdivided Populations. *Proceedings of the National Academy of Sciences*, 70(12):3321–3323, 1973.

[50] D. F. Echeverry, S. Nair, L. Osorio, S. Menon, C. Murillo, and T. J. C. Anderson. Long term persistence of clonal malaria parasite Plasmodium falciparum lineages in the Colombian Pacific region. *BMC Genetics*, 14(2), 2013.

[51] S. C. Nkhoma, S. Nair, S. Al-Saai, E. Ashley, R. McGready, A. P. Phyo, F. Nosten, and T. J. C. Anderson. Population genetic correlates of declining transmission in a human pathogen. *Molecular Ecology*, 22(2):273–285, 2013.

[52] G. C. Cerqueira, I. H. Cheeseman, S. F. Schaffner, S. Nair, M. McDew-White, A. P. Phyo, E. A. Ashley, A. Melnikov, P. Rogov, B. W. Birren, F. Nosten, T. J. C. Anderson, and D. E. Neafsey. Longitudinal genomic surveillance of Plasmodium falciparum malaria parasites reveals complex genomic architecture of emerging artemisinin resistance. *Genome Biology*, 18(1):78, 2017.

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[53] S. Auburn and A. E. Barry. Dissecting malaria biology and epidemiology using population genetics and genomics. *International Journal for Parasitology*, 47(2-3):77–85, 2017.

[54] A. R. Taylor, J. A. Watson, C. S. Chu, K. Puaprasert, J. Duanguppama, N. P. J. Day, F. Nosten, D. E. Neafsey, C. O. Buckee, M. Imwong, and N. J. White. Estimating the probable cause of recurrence in plasmodium vivax malaria: relapse, reinfection or recrudescence? *bioRxiv*, 2018.

[55] C. S. Chu, A. P. Phyo, K. M. Lwin, H. H. Win, T. San, A. A. Aung, R. Raksapraidee, V. I. Carrara, G. Bancone, J. Watson, K. A. Moore, J. Wiladphaingern, S. Proux, K. Sriprawat, M. Winterberg, P. Y. Cheah, A. L. Chue, J. Tarning, M. Imwong, F. Nosten, and N. J. White. Comparison of the cumulative efficacy and safety of chloroquine, artesunate, and chloroquine-primaquine in plasmodium vivax malaria. *Clinical Infectious Diseases*, 67(10):1543–1549, 2018.

[56] C. S. Chu, A. P. Phyo, C. Turner, H. H. Win, N. P. Poe, W. Yotyingaphiram, S. Thinraow, P. Wilairisak, R. Raksapraidee, V. I. Carrara, et al. Chloroquine versus dihydroartemisinin-piperaquine with standard high-dose primaquine given either for 7 days or 14 days in plasmodium vivax malaria. *Clinical Infectious Diseases*, 2018.

[57] E. A. Thompson. The estimation of pairwise relationships. *Annals of Human Genetics*, 39(2): 173–188, 1975.

[58] D. S. Baetscher, A. J. Clemento, T. C. Ng, E. C. Anderson, J. C. Garza, and C. C. John Garza. Microhaplotypes provide increased power from short-read DNA sequences for relationship inference. *Molecular Ecology Resources*, 18(2):296–305, 2018.

[59] J. Wang and K. T. Scribner. Parentage and sibship inference from markers in polyploids. *Molecular Ecology Resources*, 14(3):541–553, 2014.

[60] S. C. Nkhoma, S. G. Trevino, K. M. Gorena, S. Nair, S. Khoswe, C. Jett, R. Garcia, B. Daniel, A. Dia, D. J. Terlouw, S. A. Ward, T. J. Anderson, and I. H. Cheeseman. Resolving within-host malaria parasite diversity using single-cell sequencing. *bioRxiv*, 2018.

[61] J. Wang. Sibship Reconstruction from Genetic Data with Typing Errors. *Genetics*, 166(4): 1963–1979, 2004.

[62] B. F. Voight and J. K. Pritchard. Confounding from cryptic relatedness in case-control association studies. *PLoS genetics*, 1(3):e(32), 2005.

[63] L. Speidel, M. Forest, S. Shi, and S. R. Myers. A method for genome-wide genealogy estimation for thousands of samples. *bioRxiv*, 2019.

[64] M. D. Ramstetter, S. A. Shenoy, T. D. Dyer, D. M. Lehman, J. E. Curran, R. Duggirala, J. Blangero, J. G. Mezey, and A. L. Williams. Inferring Identical-by-Descent Sharing of Sample Ancestors Promotes High-Resolution Relative Detection. *American Journal of Human Genetics*, 103(1):30–44, 2018.

[65] N. A. Rosenberg, L. M. Li, R. Ward, and J. K. Pritchard. Informativeness of Genetic Markers for Inference of Ancestry *. *Am. J. Hum. Genet*, 73:1402–1422, 2003.

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[66] L. M. Gattepaille and M. Jakobsson. Combining Markers into Haplotypes Can Improve Population Structure Inference. *Genetics*, 190(1):159–174, 2012.

[67] D. E. Neafsey, M. Juraska, T. Bedford, D. Benkeser, C. Valim, A. Griggs, M. Lievens, S. Abdulla, S. Adjei, T. Agbenyega, S. T. Agnandji, P. Aide, S. Anderson, D. Ansong, J. J. Aponte, K. P. Asante, P. Bejon, A. J. Birkett, M. Bruls, K. M. Connolly, U. D'Alessandro, C. Dobaño, S. Gesase, B. Greenwood, J. Grimsby, H. Tinto, M. J. Hamel, I. Hoffman, P. Kamthunzi, S. Kariuki, P. G. Kremsner, A. Leach, B. Lell, N. J. Lennon, J. Lusingu, K. Marsh, F. Martinson, J. T. Molel, E. L. Moss, P. Njuguna, C. F. Ockenhouse, B. R. Ogutu, W. Otieno, L. Otieno, K. Otieno, S. Owusu-Agyei, D. J. Park, K. Pellé, D. Robbins, C. Russ, E. M. Ryan, J. Sacarlal, B. Sogoloff, H. Sorgho, M. Tanner, T. Theander, I. Valea, S. K. Volkman, Q. Yu, D. Lapierre, B. W. Birren, P. B. Gilbert, and D. F. Wirth. Genetic Diversity and Protective Efficacy of the RTS,S/AS01 Malaria Vaccine. *The New England Journal of Medicine*, 373(21): 2025–37, 2015.

[68] J. Mu, R. A. Myers, H. Jiang, S. Liu, S. Ricklefs, M. Waisberg, K. Chotivanich, P. Wilairatana, S. Krudsood, N. J. White, et al. Plasmodium falciparum genome-wide scans for positive selection, recombination hot spots and resistance to antimalarial drugs. *Nature genetics*, 42 (3):268, 2010.

[69] J. B. Hiatt, C. C. Pritchard, S. J. Salipante, B. J. O'Roak, and J. Shendure. Single molecule molecular inversion probes for targeted, high accuracy detection of low frequency variation. *Genome research*, 23(5):843–54, 2013.

[70] O. Aydemir, M. Janko, N. J. Hathaway, R. Verity, M. K. Mwandagalirwa, A. K. Tshefu, S. K. Tessema, P. W. Marsh, A. Tran, T. Reimonn, A. C. Ghani, A. Ghansah, J. J. Juliano, B. R. Greenhouse, M. Emch, S. R. Meshnick, and J. A. Bailey. Drug-Resistance and population structure of plasmodium falciparum across the democratic Republic of Congo using high-Throughput molecular inversion probes. *Journal of Infectious Diseases*, 218(6):946–955, 2018.

[71] M. McDew-White, X. Li, S. C. Nkhoma, S. Nair, I. Cheeseman, and T. J. Anderson. Mode and tempo of microsatellite length change in a malaria parasite mutation accumulation experiment. *bioRxiv*, 2019.

[72] J. Hoffman and W. Amos. Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion. *Molecular Ecology*, 14(2):599–612, 2005.

[73] S. R. Browning and B. L. Browning. High-Resolution Detection of Identity by Descent in Unrelated Individuals. *American Journal of Human Genetics*, 86(4):526–539, 2010.

[74] H. Samad, F. Coll, M. D. Preston, H. Ocholla, R. M. Fairhurst, and T. G. Clark. Imputation-based population genetics analysis of plasmodium falciparum malaria parasites. *PLoS genetics*, 11(4):e1005131, 2015.

[75] D. E. Neafsey, S. F. Schaffner, S. K. Volkman, D. Park, P. Montgomery, D. A. Milner, A. Lukens, D. Rosen, R. Daniels, N. Houde, et al. Genome-wide snp genotyping highlights the role of natural selection in plasmodium falciparum population divergence. *Genome biology*, 9 (12):R171, 2008.

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[76] I. H. Consortium et al. A second generation human haplotype map of over 3.1 million snps. *Nature*, 449(7164):851, 2007.

[77] M. Natesh, R. W. Taylor, N. Truelove, E. A. Hadly, S. Palumbi, U. Ramakrishnan, and D. Petrov. Empowering conservation practice with efficient and economical genotyping from poor quality samples using mPCRseq. *bioRxiv*, 2018.

[78] R. Douc, E. Moulines, and D. Stoffer. *Nonlinear time series: Theory, methods and applications with R examples*. Chapman and Hall/CRC, 2014.

[79] A. Doucet and N. Shephard. Robust inference on parameters via particle filters and sandwich covariance matrices. *University of Oxford, Department of Economics*, (606), 2012.

[80] M. E. R. T. Cappé, O. *Inference in Hidden Markov Models*. Springer, 2005.

[81] M. Bladt and M. Sørensen. Statistical inference for discretely observed Markov jump processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(3):395–410, 2005.

[82] Y. Aït-Sahalia. Maximum likelihood estimation of discretely sampled diffusions: a closed-form approximation approach. *Econometrica*, 70(1):223–262, 2002.

[83] O. E. Barndorff-Nielsen, S. E. Graversen, J. Jacod, and N. Shephard. Limit theorems for bipower variation in financial econometrics. *Econometric Theory*, 22(4):677–719, 2006.

[84] C. Messerli, N. E. Hofmann, H.-P. Beck, and I. Felger. Critical evaluation of molecular monitoring in malaria drug efficacy trials and pitfalls of length-polymorphic markers. *Antimicrobial agents and chemotherapy*, 61(1):e01500–16, 2017.

Appendix A Estimator based on IBS

For clarity of exposition, here we derive results for $\widehat{\text{IBS}}_m$ under a simple model that assumes no genotyping error (a more general result that includes genotyping error can be found in Appendix B, equation (B.9)). The simple model assumes the IBD state at the t-th locus, IBD_t, is Bernoulli with relatedness parameter $r \in [0,1]$. Given $\text{IBD}_t = 0$, we assume that $Y_t^{(i)}$ and $Y_t^{(j)}$ are independent Bernoulli with parameter $(f_t(g))_{g\in\mathcal{G}_t}$. Given $\text{IBD}_t = 1$, we assume that $Y_t^{(i)}$ follows a Bernoulli with parameter $(f_t(g))_{g\in\mathcal{G}_t}$ and that $Y_t^{(j)} = Y_t^{(i)}$ with probability one.

A.1 Expectation of estimator based on IBS

In this section no assumptions are made about dependence between marker loci: equation (A.1) holds under both independence and dependence. The expectation of the estimator $\widehat{\text{IBS}}_m$ conditional

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

on the frequencies $(f_t(g))_{g\in\mathcal{G}_t} \forall t = 1,\ldots,m$ is

$$\begin{aligned}
\mathbb{E}[\widehat{\text{IBS}}_m] &= \frac{1}{m}\sum_{t=1}^{m}\mathbb{E}\left[\text{IBS}_t\right], \\
&= \frac{1}{m}\sum_{t=1}^{m}\mathbb{P}\left(\text{IBS}_t = 1 \mid \text{IBD}_t = 1\right)\mathbb{P}\left(\text{IBD}_t = 1\right) + \mathbb{P}\left(\text{IBS}_t = 1 \mid \text{IBD}_t = 0\right)\mathbb{P}\left(\text{IBD}_t = 0\right), \\
&= \frac{1}{m}\sum_{t=1}^{m}\{r + \sum_{i=1}^{K_t} f_t(g_i)^2(1-r)\}, \\
&= r + \bar{h}_m(1-r), \\
&= \bar{h}_m + (1-\bar{h}_m)r, \qquad \text{(A.1)}
\end{aligned}$$

where $\bar{h}_m = m^{-1}\sum_{t=1}^{m}\sum_{i=1}^{K_t} f_t(g_i)^2$ (equation (3.4)). Under different observation models, we would still obtain $\mathbb{E}[\widehat{\text{IBS}}_m]$ as a linear function of r; see second line above, where $\mathbb{P}\left(\text{IBS}_t = 1 \mid \text{IBD}_t = 1\right)$ and $\mathbb{P}\left(\text{IBS}_t = 1 \mid \text{IBD}_t = 0\right)$ could be anything as long as these expressions do not involve r.

A.2 Convergence of estimator based on IBS

Here we work under the simplest setting: the measurements $(Y_t^{(i)}, Y_t^{(j)})$ are independent across $t = 1,\ldots,m$. In order to discuss convergence we need to imagine an asymptotic regime where $m \to \infty$. We introduce an infinite sequence $(f_t(g_i))_{t\geq 1}, i = 1,\ldots,K_t$, where each $f_t(g)$ is in $(0,1)$, and we introduce $\bar{h} = \lim_{m\to\infty} m^{-1}\sum_{t=1}^{m}\sum_{i=1}^{K_t} f_t(g_i)^2$, assuming the existence of that limit. To show that $\widehat{\text{IBS}}_m$ is not consistent for r, we show that it is consistent for $\bar{h} + (1-\bar{h})r$, which is different to r unless $r = 1$. Thus we show that $\widehat{\text{IBS}}_m$ satisfies,

$$\widehat{\text{IBS}}_m \xrightarrow[m\to\infty]{\mathbb{P}} \bar{h} + (1-\bar{h})r, \qquad \text{(A.2)}$$

where the arrow is interpreted as "convergence in probability". Since $\mathbb{E}[\widehat{\text{IBS}}_m] = \bar{h}_m + (1-\bar{h}_m)r \to \bar{h} + (1-\bar{h})r$ as $m \to \infty$, we can establish (A.2) by showing that for every $\varepsilon > 0$

$$\mathbb{P}\left(\mid \widehat{\text{IBS}}_m - \mathbb{E}\left[\widehat{\text{IBS}}_m\right] \mid > \varepsilon\right) \to 0 \text{ as } m \to \infty. \qquad \text{(A.3)}$$

We show equation (A.3) by use of Hoeffding's inequality (see Chapter 4 in [40]). Since $\widehat{\text{IBS}}_m$ is an average of variables IBS_t, which are bounded ($\text{IBS}_t \in \{0,1\}$) and assumed independent, Hoeffding's inequality yields

$$\mathbb{P}\left(\mid \widehat{\text{IBS}}_m - \mathbb{E}\left[\widehat{\text{IBS}}_m\right] \mid \geq \varepsilon\right) \leq 2\exp\left(-2m\varepsilon^2\right). \qquad \text{(A.4)}$$

Since $2\exp\left(-2m\varepsilon^2\right) \to 0$ as $m \to \infty$, equation (A.4) shows that equation (A.3) holds and therefore that equation (A.2) holds. Note that consistency could also be established in the dependent case, for instance via the application of a version of Hoeffding's inequality for dependent processes.

Plots of $\widehat{\text{IBS}}_m$ for data simulated under the independence model (Figure A.1) numerically show for $r = 0$ and 0.5 that $\widehat{\text{IBS}}_m$ concentrates on its expectation (equation (A.1)) as more and more markers ($m = 24$, 96 and 192) are typed.

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

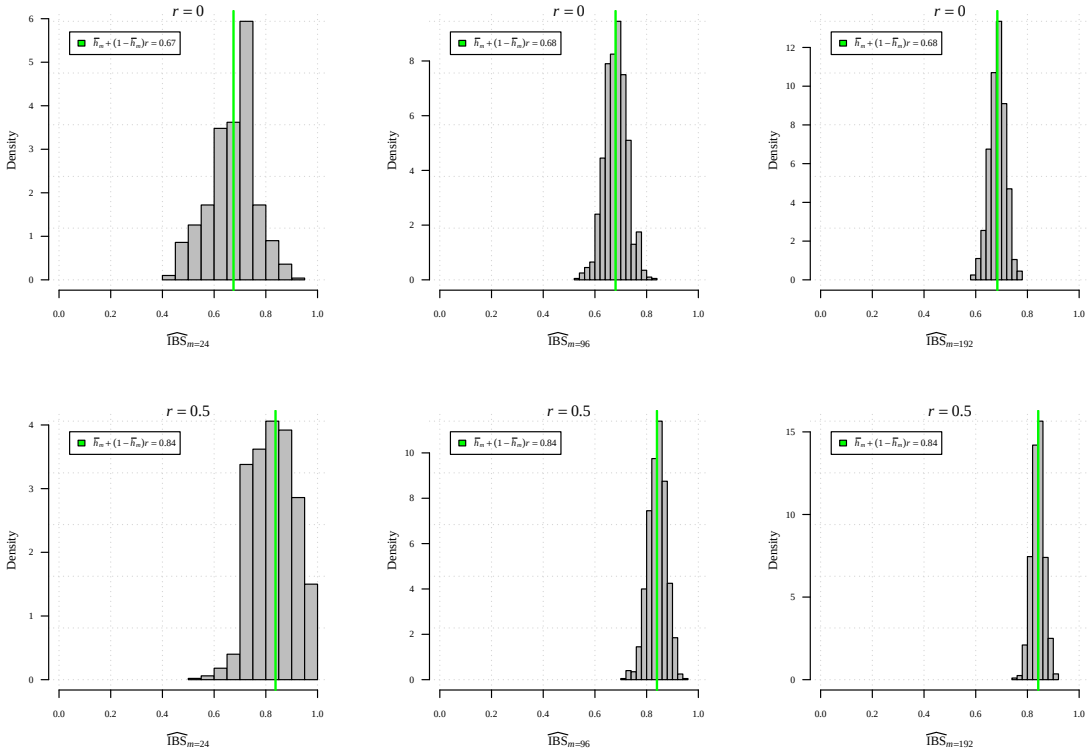


Figure A.1: $\widehat{\text{IBS}}_m$ between pairs of biallelic marker data simulated under the independence model with different numbers of markers, m, and relatedness, r. The green vertical line marks $\bar{h}_m + (1 - \bar{h}_m)r$ which is a function of the allele frequencies (equations (3.3) and (3.4)). Allele frequencies were sampled without replacement from Thai WGS data set with probability proportional to minor allele frequency estimates.

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

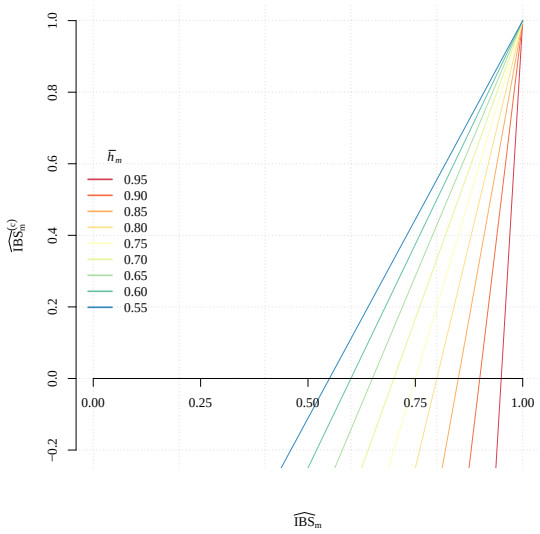


Figure A.2: $\widehat{\mathrm{IBS}}_m^{(c)}$ as a function of $\widehat{\mathrm{IBS}}_m$ for various $\bar{h}_m$ (equation (A.6)).

A.3 Corrected estimator based on IBS

A corrected version of the estimator $\widehat{\mathrm{IBS}}_m$ could be consistent for r (equation (A.7)) and is similar to existing method of moments estimators (reviewed in [11]), which generally underperform compared to maximum likelihood estimators (Chapter 9 of [40]).

By rearranging equation (A.1),

$$r = \frac{1}{(1-\bar{h}_m)}\left(\mathbb{E}\left[\widehat{\mathrm{IBS}}_m\right] - \bar{h}_m\right), \tag{A.5}$$

we can propose the following corrected estimator of r,

$$\widehat{\mathrm{IBS}}_m^{(c)} = \frac{1}{(1-\bar{h}_m)}\left(\widehat{\mathrm{IBS}}_m - \bar{h}_m\right), \tag{A.6}$$

whose expectation is precisely r. The corrected estimator $\widehat{\mathrm{IBS}}_m^{(c)}$ is consistent for r, with the same reasoning as in Appendix A.2 assuming independent observations,

$$\widehat{\mathrm{IBS}}_m^{(c)} = \frac{1}{(1-\bar{h}_m)}\left(\widehat{\mathrm{IBS}}_m - \bar{h}_m\right) \xrightarrow[m\to\infty]{\text{Probability}} \frac{1}{(1-\bar{h})}(\bar{h} + (1-\bar{h})r - \bar{h}) = r. \tag{A.7}$$

Figure A.2 shows a plot of equation (A.6) for different values of $\bar{h}_m \in (0.5, 1)$. The range of $\widehat{\mathrm{IBS}}_m^{(c)}$ includes negative values. Setting negative estimates to zero can considerably improve results [11], but can also introduce bias [13]. For the *Plasmodium* data sets considered in the main text, Figure A.3 shows $\widehat{\mathrm{IBS}}_m^{(c)}$ estimates truncated to [0,1].

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

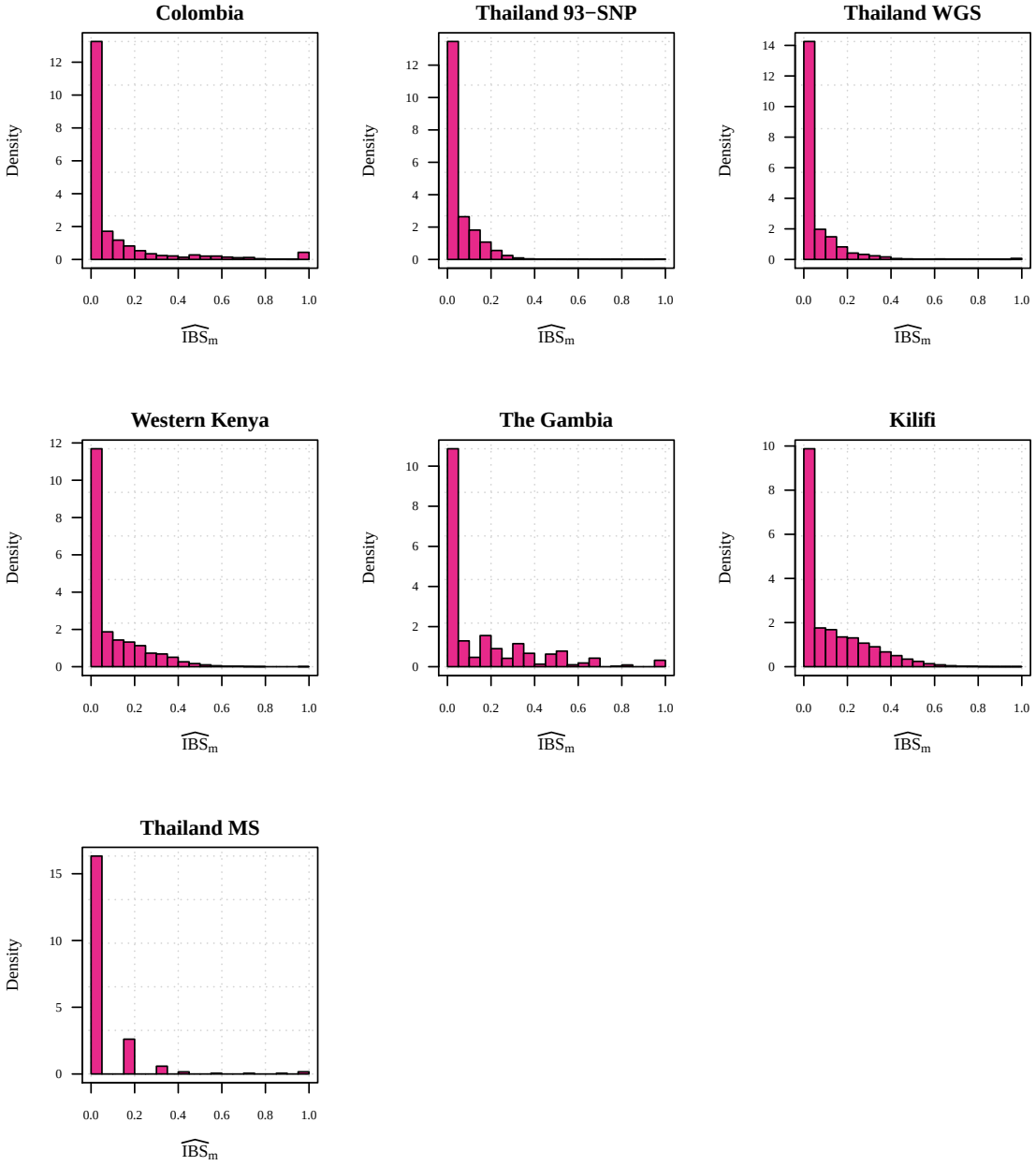


Figure A.3: $\widehat{\text{IBS}}_m^{(c)}$ for several monoclonal *Plasmodium* data sets.

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Appendix B Model-based estimation of relatedness

B.1 Framework

In this section we describe models that relate the available data to the objects of interest, in a self-contained presentation. The data comprise frequencies of alleles denoted by $(f_t(g))_{g\in\mathcal{G}_t}$, and allele indicators $Y_t^{(i)}$, where the index t denotes a locus on the genome, and the superscript (i) refers to the i-th individual. The index t will run from 1 to m, the number of markers genotyped, and we will be particularly interested in the impact of m and K_t on the precision of the estimators. Note that m cannot be larger than L, the total length of the genome, which will create difficulties in making sense of an asymptotic regime where m goes to infinity, as will be discussed below.

We will consider pairs of individuals, i and j, for which we want to estimate the relatedness denoted by r and taking values in the interval $[0, 1]$. The models below might involve other parameters, and overall the vector of parameters is denoted by θ. We will make the first component of θ represent the relatedness r, so that $r = \theta_1$.

For each pair of individuals, we introduce a sequence of latent binary variables denoted by (IBD_t) for identity-by-descent: $\mathrm{IBD}_t = 1$ indicates identity-by-descent at locus t. We view this sequence as a two-state Markov chain. The case of independent variables for (IBD_t) constitutes a particular case. In any case, the relatedness $r \in [0, 1]$ represents the marginal probability that IBD_t is equal to one, assumed to be identical for all t. While we do not observe (IBD_t), we observe $Y_t^{(i)}$ and $Y_t^{(j)}$ that are related to IBD_t at site t via an observation model, which can take into account the presence of genotyping errors. Together, the specification of the latent process (IBD_t) and of the observation model fully describes a hidden Markov model, that can be used to estimate r using the data. Complete model specification is deferred to Appendix B.3, after a description of the general estimation procedure and some specific issues arising in the present case.

The estimation procedure is here based on the maximum likelihood approach. The likelihood function can be written as

$$\mathcal{L}_{1:m}(\theta) = \prod_{t=1}^{m} \mathbb{P}(Y_t^{(i)}, Y_t^{(j)} | \mathcal{Y}_{t-1}, \theta),$$

where $\mathcal{Y}_{t-1}$ represents all the observations from locus 1 to locus $t-1$, with the convention that $\mathcal{Y}_0$ is the empty set. We can further write each "incremental likelihood term" as

$$\mathbb{P}(Y_t^{(i)}, Y_t^{(j)} | \mathcal{Y}_{t-1}, \theta) = \sum_{\mathrm{IBD}_t \in \{0,1\}} \mathbb{P}(Y_t^{(i)}, Y_t^{(j)} | \mathrm{IBD}_t, \theta) \mathbb{P}(\mathrm{IBD}_t | \mathcal{Y}_{t-1}, \theta).$$

Since (IBD_t) is a Markov chain, the forward algorithm [17] can be used to evaluate each incremental likelihood term for $t = 1, \ldots, m$, for a cost of the order of m operations given θ.

We write $\ell_{1:m}(\theta) = \log \mathcal{L}_{1:m}(\theta)$, and $\ell_t(\theta) = \log \mathbb{P}(Y_t^{(i)}, Y_t^{(j)} | \mathcal{Y}_{t-1}, \theta)$. We denote the first and second derivatives of $\ell_t(\theta)$ by $\ell_t'(\theta)$ (a vector) and $\ell_t''(\theta)$ (a matrix) respectively. We will use the maximum likelihood estimator to approximate r, and we define it as

$$\hat{\theta}_m = \mathrm{argmax}_\theta \, \ell_{1:m}(\theta).$$

We next review some asymptotic properties of the maximum likelihood estimator (MLE) and detail how the present setting differs from the one usually considered in asymptotic studies.

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

B.2 Distribution of the MLE

B.2.1 Standard asymptotic theory

We first recall what the usual asymptotic reasoning is for the distribution of the MLE in HMMs [35, 36, 37], in informal terms.

The first step is to imagine that the variables indexed by t (such as IBD_t, $Y_t^{(i)}$, $Y_t^{(j)}$, etc.) are part of infinite sequences of variables indexed by $t \geq 1$. This allows us to consider a regime where the number of sites considered m can go to ∞. In Appendix B.2.2 we will discuss issues arising when applying this asymptotic reasoning in the present context of genetic data.

We observe that the log-likelihood and its derivatives are sums of m terms. Dividing by m yields averages, which might converge to limiting values as m grows large. For instance, the scaled log-likelihood might satisfy

$$\forall \theta \quad m^{-1}\ell_{1:m}(\theta) \xrightarrow[m\to\infty]{\mathbb{P}} \bar{\ell}(\theta),$$

where the arrow is to be interpreted as "convergence in probability", the left hand side of it being random if we consider the data to be random. Under some assumptions, the maximizer $\hat{\theta}_m$ of $\theta \mapsto m^{-1}\ell_{1:m}(\theta)$ converges to the maximizer $\theta^\star$ of the limiting function $\theta \mapsto \bar{\ell}(\theta)$. By the Taylor expansion of $\ell'_{1:m}(\hat{\theta}_m)$ at $\theta^\star$ we have

$$\ell'_{1:m}(\hat{\theta}_m) = \ell'_{1:m}(\theta^\star) + \ell''_{1:m}(\theta^\star)(\hat{\theta}_m - \theta^\star) + \text{rest}. \tag{B.1}$$

At the MLE $\hat{\theta}_m$, the derivative of the log-likelihood cancels: $\ell'_{1:m}(\hat{\theta}_m) = 0$, at least if the MLE is in the interior of the parameter space; extra care is required when the MLE is on the boundary of the parameter space, which occurs in the present setting where $\hat{r}_m$ can be exactly zero or one. Therefore we obtain

$$\begin{aligned} 0 &\approx \ell'_{1:m}(\theta^\star) + \ell''_{1:m}(\theta^\star)(\hat{\theta}_m - \theta^\star), \\ \Leftrightarrow (\hat{\theta}_m - \theta^\star) &\approx -\ell''_{1:m}(\theta^\star)^{-1}\ell'_{1:m}(\theta^\star), \qquad \text{(B.2)} \\ \Leftrightarrow \sqrt{m}(\hat{\theta}_m - \theta^\star) &\approx \left(-m^{-1}\ell''_{1:m}(\theta^\star)\right)^{-1} m^{-1/2}\ell'_{1:m}(\theta^\star), \qquad \text{(B.3)} \end{aligned}$$

where $\Leftrightarrow$ means "equivalently". We will rely on the two following convergence results (see Chapter 13 in [78]),

$$m^{-1/2}\ell'_{1:m}(\theta^\star) \xrightarrow[m\to\infty]{\mathrm{d}} \mathcal{N}\left(0, V^\star\right), \tag{B.4}$$

$$-m^{-1}\ell''_{1:m}(\theta^\star) \xrightarrow[m\to\infty]{\mathbb{P}} J^\star, \tag{B.5}$$

for some matrices $V^\star, J^\star$, assumed to be both semi-definite positive and symmetric. The first line above describes a convergence "in distribution" and can follow from a central limit theorem for the first derivative of the log-likelihood. The second line can follow from a law of large numbers applied to the second derivatives, as in Chapter 13 of [78]. We can combine these two convergence results using Slutsky's lemma to obtain the asymptotic normality of the MLE:

$$\sqrt{m}(\hat{\theta}_m - \theta^\star) \xrightarrow[m\to\infty]{\mathrm{d}} \mathcal{N}\left(0, (J^\star)^{-1}V^\star(J^\star)^{-1}\right). \tag{B.6}$$

This key result can be used for sample size determination and for the construction of confidence intervals, provided that we can approximate $\theta^\star$, $V^\star$ and $J^\star$ based on data. The asymptotic variance

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

$(J^\star)^{-1}V^\star(J^\star)^{-1}$ is sometimes called the sandwich formula, and can be estimated based on samples; see Doucet and Shephard [79] in the setting of hidden Markov models. If we assume that the model is well-specified, i.e. that the data actually are generated from the model with the parameter $\theta^\star$, then it can be shown that $J^\star = V^\star$ under regularity conditions (Chapter 13 of Douc et al. [78]). In this case, the asymptotic variance in (B.6) simplifies to $(J^\star)^{-1}$. The matrix $J^\star$ is often termed the Fisher Information Matrix at $\theta^\star$.

We briefly discuss the numerical obtention of $\hat{\theta}_m = \text{argmax}_\theta \ell_{1:m}(\theta)$. The log-likelihood function $\theta \mapsto \ell_{1:m}(\theta)$ can be plugged in a numerical optimizer, such as that implemented in the `optim` function of `R`. Evaluations of the log-likelihood function require runs of the forward algorithm on the data, for a cost of the order of m operations. Alternatively, one can also run an expectation-maximization algorithm, which involves calculating expectations with respect to the distribution of the latent process (IBD_t) using the forward-backward algorithm [80], also called Baum-Welch in the context of HMMs [17]. If the parameter is small-dimensional, e.g. one or two-dimensional, a simple way of approximating the MLE consists in evaluating the likelihood (using the forward algorithm) on a grid of parameter values, and selecting the parameter associated with the highest likelihood.

The matrix $J^\star$ can be estimated by $-m^{-1}\ell''_{1:m}(\hat{\theta}_m)$, itself computed via numerical differentiation of the log-likelihood function at $\hat{\theta}_m$. The estimation of $V^\star$ is more complicated and has been the topic of a rich literature in time series analysis; see for instance Doucet and Shephard [79] and references therein.

B.2.2 Applicability of the standard asymptotic theory

The law of large numbers and central limit theorems usually employed to carry out the above reasoning, i.e. to establish (B.4) and (B.5) leading to the asymptotic normality of the MLE in (B.6), might not be meaningful in the present context. Indeed they usually apply to stationary processes observed over increasingly long periods of time. In such asymptotic setting, one eventually observes a realization of a stationary stochastic process over an infinitely long time horizon, which is enough to learn the invariant distribution of the process. We refer to this setting as standard asymptotics. Recall that our primary object of interest is the parameter r, which characterizes indeed the invariant distribution of the Markov chain (IBD_t).

In the present setting where data comprise genetic sequences, increasing m means considering more loci on the genome. The m considered loci are located within the genome whose length is, however, fixed. Therefore increasing m amounts to increasing the subsampling frequency at which data are observed. In other words it decreases the distance between successive observed loci. We refer to this as subsampling asymptotics. To see where this differs from standard asymptotics, consider a simpler context where (IBD_t) would not be hidden but directly observed. In the limit $m \to \infty$ in subsampling asymptotics, we would observe a continuous trajectory of (IBD_t), switching from state 0 to state 1 and back again, over a fixed interval. The maximum likelihood estimate of r for such a model would be the proportion of time that the trajectory would spend in state 1 [81]. However this would not be exactly equal to r, even if the trajectory was sampled from the Markov model given r, because the fully-observed realization of (IBD_t) would still be of a finite length; this is well-known, see [34] on the impact of the genome length on relatedness estimates under Mendelian sampling. On the other hand, in the standard asymptotics $m \to \infty$ we would observe an infinitely long trajectory of the Markov chain, for which the maximum likelihood estimator of the transition matrix is consistent. The difference between the two regimes is illustrated in Figure

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

B.1.

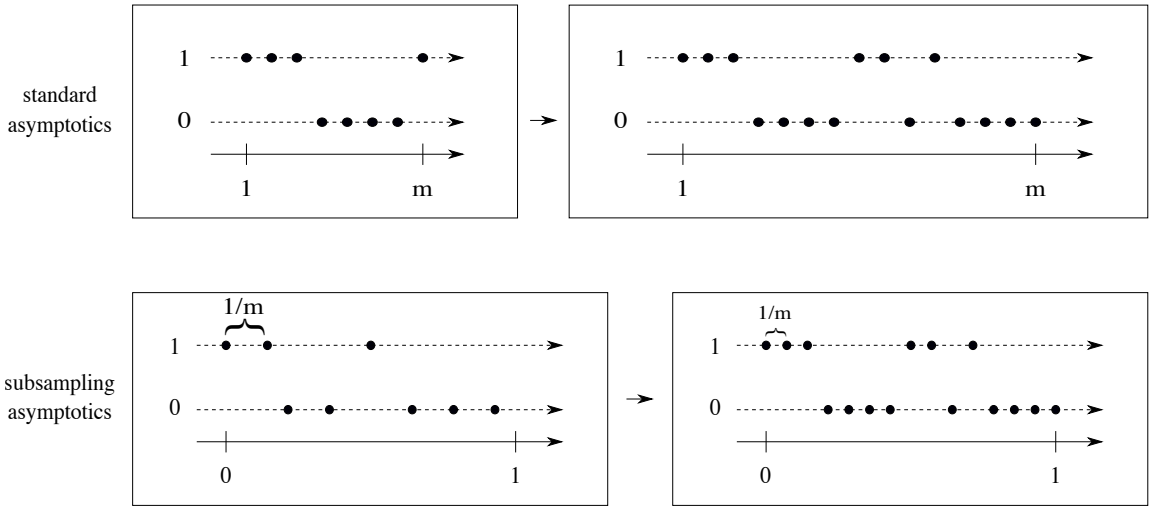


Figure B.1: Two different ways of increasing m: in the top row, m refers to the length of the observation period, while the observations are separated by one unit of time. In the bottom row, the length of the observation interval is fixed to one, and the observations are placed at distance $1/m$ of one another; thus an increase in m means that successive observations are closer to one another, but the length of the observation period is fixed.

The difference in asymptotic regimes has consequences on the estimability of r. In the subsampling asymptotics, it is impossible to arbitrarily decrease the error of $\hat{r}_m$ by increasing m: there is only so much information that can be gathered about r by increasing the number of loci under consideration; hence the distinction between expected IBD and realised IBD in [2]. A result such as the asymptotic normality with a $\sqrt{m}$ rate of convergence, as in (B.6), is in fact unlikely to hold. The numerical experiments indeed suggest that the root mean squared error associated with $\hat{r}_m$ does not decrease beyond a certain point, no matter how large m is. The subsampling asymptotic regime has been formally studied with various applications to financial econometrics [82, 83], but we are not aware of similar results for hidden Markov models such as the ones considered here.

Despite the standard asymptotic results not holding, we do observe that the distribution of $\widehat{r}_m$ is approximately normal for m large enough (Figure B.2). This can be partially explained by the fact that normality of the MLE depends entirely on the log-likelihood being approximately quadratic [38], which itself does not have to follow from standard asymptotic arguments. Since the log-likelihood function is observed to be approximately quadratic providing $\widehat{r}_m$ is not close to the boundaries (Figure B.3), we can still quantify the precision of the MLE by considering the second derivative of the log-likelihood at its maximum. Thus we will rely on the Fisher Information Matrix as a proxy for the precision of the MLE, in particular for the study of the effect of K_t in Appendix B.3.4.

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

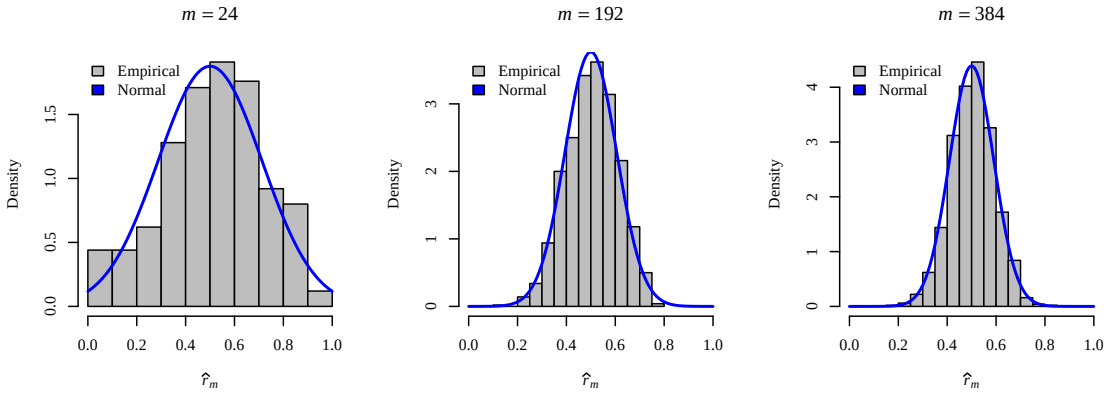


Figure B.2: Empirical distributions of $\widehat{r}_m$ for different numbers of markers, m. Each distribution is based on 1000 estimates of r given data simulated and analyzed under the HMM with $r = 0.5$, $k = 12$, $K_t = 2\ \forall t = 1 \ldots, m$ and $\varepsilon = 0.001$.

B.3 Models

We now describe a Markov chain model for (IBD_t), followed by observation models for $Y_i^{(t)}$ and $Y_j^{(t)}$ given IBD_t.

B.3.1 Hidden Markov model

We write the transition probabilities of (IBD_t) at a locus t,

$$A(t) = \begin{pmatrix} a_{00}(t) & a_{01}(t) \\ a_{10}(t) & a_{11}(t) \end{pmatrix} = \begin{pmatrix} 1 - r(1 - \exp(-k\rho d_t)) & r(1 - \exp(-k\rho d_t)) \\ (1-r)(1 - \exp(-k\rho d_t)) & 1 - (1-r)(1 - \exp(-k\rho d_t)) \end{pmatrix}.$$

In the above, $a_{j\ell}(t)$ refers to the probability of $\mathrm{IBD}_t = \ell$ given that $\mathrm{IBD}_{t-1} = j$.

In the above expression, the relatedness is denoted by r; d_t denotes a genetic distance in base pairs (bp) between sites $t-1$ and t; $k > 0$ parametrizes the switching rate of the Markov chain and ρ is the recombination rate, assumed known and fixed across both haploid genotypes with value $7.4 \times 10^{-7}\mathrm{M\ bp}^{-1}$ for *P. falciparum* parasites [33].

We can check that, if $\mathbb{P}(\mathrm{IBD}_{t-1} = 1) = r$, then

$$\mathbb{P}(\mathrm{IBD}_t = 1) = \mathbb{P}(\mathrm{IBD}_{t-1} = 1)a_{11}(t) + \mathbb{P}(\mathrm{IBD}_{t-1} = 0)a_{01}(t) = r,$$

and thus the invariant marginal distribution of the chain is given by $\mathbb{P}(\mathrm{IBD}_t = 1) = r$.

The above transition probabilities are at the core of many HMMs of relatedness (e.g. [14], where $k \times \rho = a$ and genetic distance $d_t = t_k$ is measured in centi Morgans (cM), plus many subsequent models (see [15]), including [18], where $r = \pi_1$ and $1 - r = \pi_2$.

We can check that, as the distance increases to infinity, the probabilities in $A(t)$ simplify and correspond to the i.i.d. Bernoulli model where IBD_t is equal to one with probability r, independently for each site t. In other words, if sites are distant enough, we expect the HMM and the independence models to give similar results. This will happen in particular when m is small and when the loci under consideration are well-spread across the genome.

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

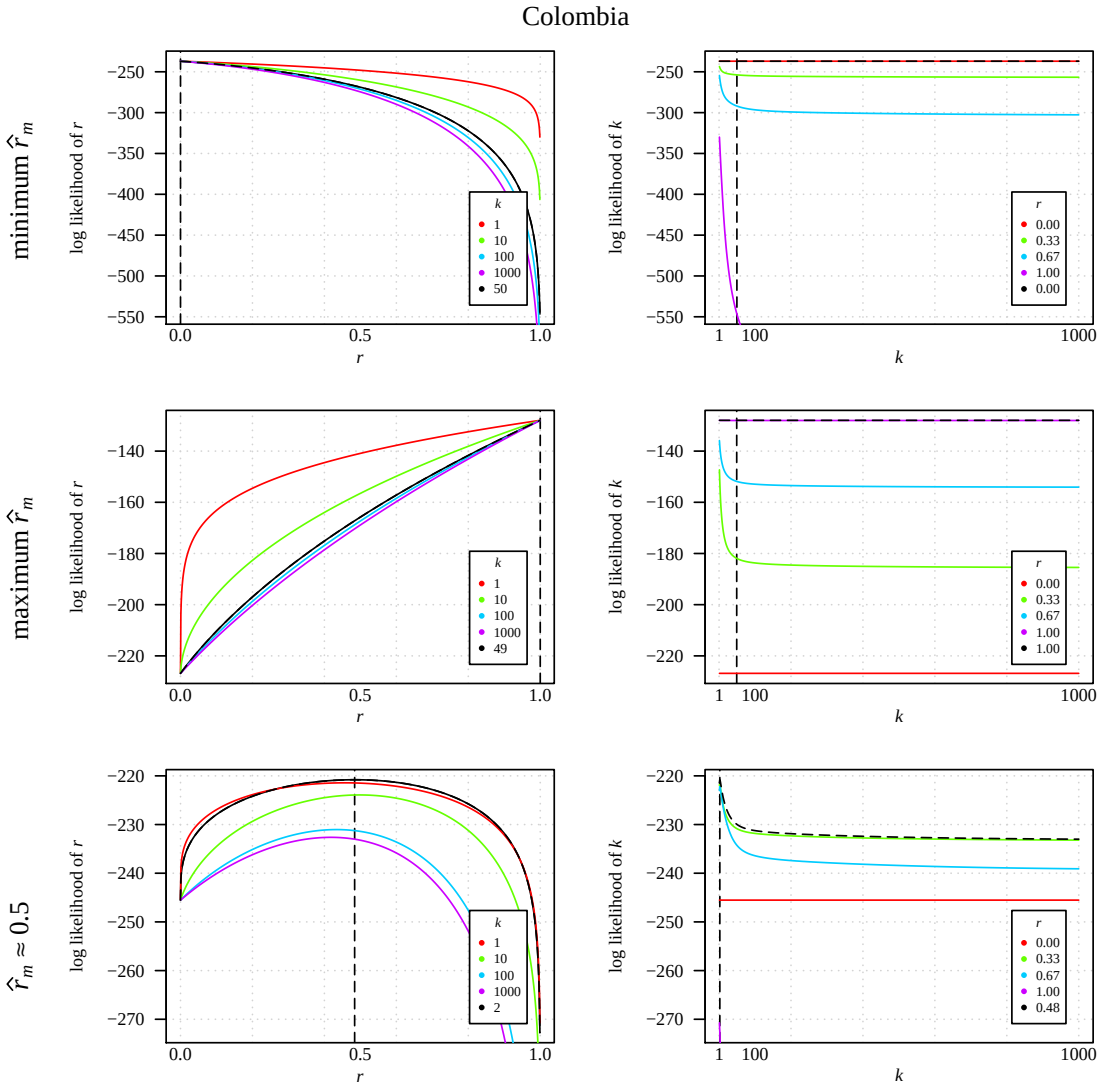


Figure B.3: The log likelihoods of r for different k (left column) and k for different r (right column) for three different example sample pairs from the the Colombian data set: a sample pair with minimum $\widehat{r}_m$ (top row, $m = 248$), a sample pair with maximum $\widehat{r}_m$ (middle row, $m = 246$), and a sample pair with $\widehat{r}_m \approx 0.5$ (bottom row, $m = 245$). Differences in m are due to missing genotype calls in the data. Vertical black dashed lines mark $\widehat{r}_m$ (left column) and $\widehat{k}_m$ (right column). Black dashed function lines show the log likelihood of $\widehat{r}_m$ given $\widehat{k}_m$ (left column) and of $\hat{k}$ given $\widehat{r}_m$ (right column). Coloured function lines show the log likelihood of $\widehat{r}_m$ given values of $k \neq \widehat{k}_m$ (left column) and of $\hat{k}$ given values of $r \neq \widehat{r}_m$ (right column).

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

B.3.2 Observation model

The observations $Y_t^{(i)}, Y_t^{(j)}$ are related to (IBD_t) only through IBD_t at locus t. The observation model introduces some true genotypes $G_t^{(i)}, G_t^{(j)}$ given IBD_t, and then some genotyping error model defining the distribution of $Y_t^{(i)}, Y_t^{(j)}$ given $G_t^{(i)}, G_t^{(j)}$.

First, the variables $G_t^{(i)}, G_t^{(j)}$ given IBD_t are defined as follows. If $\text{IBD}_t = 0$, then $G_t^{(i)}$ is independent of $G_t^{(j)}$ and both follow a Categorical distribution: for a set of values $\mathcal{G} = \{g(1), \dots, g(K_t)\}$ and probabilities $\{f_t(g)\}$ for $g \in \mathcal{G}$, we have $\mathbb{P}(G_t^{(i)} = g) = f_t(g)$, and likewise for $G_t^{(j)}$. If there are only two types (e.g. the case for biallelic SNPs) then it is a Bernoulli distribution. If $\text{IBD}_t = 1$, then $\mathbb{P}(G_t^{(i)} = g) = f_t(g)$ and $G_t^{(j)} = G_t^{(i)}$ with probability one. Overall we can write the model as

$$\mathbb{P}(G_t^{(i)} = g^{(i)}, G_t^{(j)} = g^{(j)} | \text{IBD}_t = 0) = f_t(g^{(i)}) f_t(g^{(j)})$$
$$\mathbb{P}(G_t^{(i)} = g^{(i)}, G_t^{(j)} = g^{(j)} | \text{IBD}_t = 1) = f_t(g^{(i)}) \mathbb{1}(g^{(i)} = g^{(j)}).$$

Next, we assume that genotyping errors occur independently for both individuals. This differs to the typical 'all-or-none' diploid setting (e.g. [14, 15]), since haploid genotypes in monoclonal parasite samples are genotyped separately. If they occur, we do not observe $Y_t^{(i)} = G_t^{(i)}$ but instead we observe another genotype taken uniformly among the other possible values (by assumption); and likewise for the other individual j. This can be written

$$\mathbb{P}(Y_t^{(i)} = g^{(i)} | G_t^{(i)} = g) = \begin{cases} 1 - (K_t - 1)\varepsilon & \text{if } g^{(i)} = g, \\ \varepsilon & \text{if } g^{(i)} \neq g. \end{cases}$$

In the above expression K_t refers to the number of possibilities, which could be different for different sites t, and ε refers to a parameter such that the error rate is $(K_t - 1)\varepsilon$. This is suited to microsatellites in the sense that the error rate scales with K_t [72]. For biallelic SNPs, it amounts to a simple miscall.

Overall we can thus think of the observation model as the combination of a model for $(Y_t^{(i)}, Y_t^{(j)})$ given $(G_t^{(i)}, G_t^{(j)})$ and a model for $(G_t^{(i)}, G_t^{(j)})$ given IBD_t. We can integrate $G_t^{(i)}, G_t^{(j)}$ out to obtain directly the probabilities of $(Y_t^{(i)}, Y_t^{(j)})$ given IBD_t:

$$\mathbb{P}(Y_t^{(i)} = g^{(i)}, Y_t^{(j)} = g^{(j)} | \text{IBD}_t) \tag{B.7}$$
$$= \sum_{g, g' \in \mathcal{G}} \mathbb{P}(Y_t^{(i)} = g^{(i)} | G_t^{(i)} = g) \mathbb{P}(Y_t^{(j)} = g^{(j)} | G_t^{(j)} = g') \mathbb{P}(G_t^{(i)} = g, G_t^{(j)} = g' | \text{IBD}_t). \tag{B.8}$$

The cost of evaluating this expression is quadratic in the cardinality of $\mathcal{G}$.

This observation model is the same (besides notation) as that for within-population samples under the HMM of hmmIBD [18] and, if $K_t = 2$, the same as that of the HMMs of isoRelate [23]. Mutations do not feature in it. However, any that do occur can be absorbed as errors, as they are considered to be in [61]. That said, it does not take into account microsatellite mutations in the sense that they scale with both motif size and repeat number [71], nor the inherent ordinal nature of microsatellites or the bias with regards to their amplification [84]. Bespoke adaptations could be made for specific data types.

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Digression: expection of fraction IBS considering error Equation (B.8) means that

$$\mathbb{P}(Y_t^{(i)} = Y_t^{(j)} | \text{IBD}_t = 1) = (1-(K_t-1)\varepsilon)^2 + \varepsilon^2(K_t-1),$$
$$\mathbb{P}(Y_t^{(i)} = Y_t^{(j)} | \text{IBD}_t = 0) = (1-(K_t-1)\varepsilon)^2 h_t + \varepsilon^2(K_t - 2 + h_t) + 2\varepsilon(1-(K_t-1)\varepsilon)(1-h_t),$$

where $h_t = \sum_{g\in\mathcal{G}} f_t(g)^2$. Consequently, under the present observation model,

$$\begin{aligned}
\mathbb{E}[\widehat{\text{IBS}}_m] = \frac{1}{m}\sum_{t=1}^{m} \big\{ & r\left((1-(K_t-1)\varepsilon)^2 + \varepsilon^2(K_t-1)\right) + \\
& (1-r)\left((1-(K_t-1)\varepsilon)^2 h_t + \varepsilon^2(K_t-2+h_t) + 2\varepsilon(1-(K_t-1)\varepsilon)(1-h_t)\right)\big\}, \\
= r & \left((1-(K_t-1)\varepsilon)^2 + \varepsilon^2(K_t-1)\right) + \\
& (1-r)\left((1-(K_t-1)\varepsilon)^2 \bar{h}_m + \varepsilon^2(K_r-2+\bar{h}_m) + 2\varepsilon(1-(K_t-1)\varepsilon)(1-\bar{h}_m)\right), \quad \text{(B.9)}
\end{aligned}$$

where $\bar{h}_m = \frac{1}{m}\sum_{t=1}^{m} h_t$. Equation (B.9) reduces to (A.1) when $\varepsilon = 0$.

B.3.3 The likelihood under the independence model

This model assumes independent random variables IBD_t across loci $t \in \{1,\dots,m\}$. It is a particular case of the above HMM when all $d_t = \infty$. Given a relatedness parameter $r \in [0,1]$, IBD_t is assumed Bernoulli with parameter r. Next, we define an observation model: given $\text{IBD}_t = 0$, we assume that $Y_t^{(i)}$ and $Y_t^{(j)}$ are independent Bernoulli with parameter $f_t(g)$. Given $\text{IBD}_t = 1$, we assume that $Y_t^{(i)}$ follows a Bernoulli with parameter $f_t(g)$ and that $Y_t^{(j)} = Y_t^{(i)}$ with probability one. This defines the observation model. The associated likelihood at site t is

$$\mathbb{P}\left(Y_t^{(i)} = g^{(i)}, Y_t^{(j)} = g^{(j)} | r\right) = \sum_{\text{IBD}_t=0}^{1} \mathbb{P}\left(Y_t^{(i)} = g^{(i)}, Y_t^{(j)} = g^{(j)} | \text{IBD}_t\right) \mathbb{P}\left(\text{IBD}_t \mid r\right).$$

At this point we can define, for all t,

$$\begin{aligned}
a_t = \sum_{g,g'\in\mathcal{G}} & \left\{\mathbb{1}(g^{(i)} = g)(1-(K_t-1)\varepsilon) + \mathbb{1}(g^{(i)} \neq g)\varepsilon\right\} \times \\
& \left\{\mathbb{1}(g^{(j)} = g')(1-(K_t-1)\varepsilon) + \mathbb{1}(g^{(j)} \neq g')\varepsilon\right\} \times \\
& \left\{f_t(g)\mathbb{1}(g = g')\right\}, \\
b_t = \sum_{g,g'\in\mathcal{G}} & \left\{\mathbb{1}(g^{(i)} = g)(1-(K_t-1)\varepsilon) + \mathbb{1}(g^{(i)} \neq g)\varepsilon\right\} \times \\
& \left\{\mathbb{1}(g^{(j)} = g')(1-(K_t-1)\varepsilon) + \mathbb{1}(g^{(j)} \neq g')\varepsilon\right\} \times \\
& \left\{f_t(g) f_t(g)(g')\right\},
\end{aligned}$$

so that the likelihood reads $\mathcal{L}_t(r) = a_t r + b_t(1-r)$. The full log-likelihood can be simply written as

$$\ell_{1:m}(r) = \sum_{t=1}^{m} \ell_t(r) = \sum_{t=1}^{m} \log\left\{a_t r + b_t(1-r)\right\}.$$

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The gradient of the log-likelihood looks like

$$\ell'_{1:m}(r) = \sum_{t=1}^{m} \ell'_t(r) = \sum_{t=1}^{m} \left\{ \frac{a_t - b_t}{a_t r + b_t(1-r)} \right\}.$$

The second-order derivative of the log-likelihood looks like

$$\ell''_{1:m}(r) = \sum_{t=1}^{m} \ell''_t(r) = -\sum_{t=1}^{m} \left\{ \frac{(a_t - b_t)^2}{(a_t r + b_t(1-r))^2} \right\}.$$

Since both numerator and denominator of each term are positive, $\ell''_{1:m}(r)$ is strictly negative for all $r \in (0,1)$, and thus the function $r \mapsto \ell_{1:m}(r)$ is concave on $(0,1)$.

For the HMM model, the form of the likelihood is less explicit; we do not have an explicit formula giving the likelihood as a function of r and of the data. However this is not a real problem as we can still numerically evaluate the likelihood, using what is usually called the forward algorithm [17]. Being able to numerically evaluate the likelihood leads to being able to optimize it to get the MLE, and to numerically differentiate it as well.

B.3.4 Maximizing Fisher information

We focus on a single site t, which we suppress from the notation. Let us denote the log-likelihood by ℓ and recall the formula

$$\log \mathbb{P}\left(Y^{(i)}, Y^{(j)}; r\right) = \ell(r) = \log\left(ar + b(1-r)\right), \quad \ell''(r) = -\frac{(a-b)^2}{\left(ar + b(1-r)\right)^2}.$$

Assume there is no genotyping error for simplicity. Then $a = f(Y^{(i)})\mathbb{1}(Y^{(i)} = Y^{(j)})$ and $b = f(Y^{(i)})f(Y^{(j)})$. From there the Fisher Information Matrix (FIM) is obtained as

$$\text{FIM} = \mathbb{E}\left[-\ell''(r)\right] = \sum_{y^{(i)}, y^{(j)}} \frac{\left(f(y^{(i)})\mathbb{1}(y^{(i)} = y^{(j)}) - f(y^{(i)})f(y^{(j)})\right)^2}{f(y^{(i)})\mathbb{1}(y^{(i)} = y^{(j)})r + f(y^{(i)})f(y^{(j)})(1-r)}.$$

It is a function of r and of the allele frequencies. The FIM is proportional to the inverse of the asymptotic variance of the MLE, thus if we want precise estimators of r, we want a large FIM. This leads to the idea of maximizing FIM with respect to f for all r, to see which allele frequencies allow the best estimation of r. We can split the sum into the case for which $y^{(i)} = y^{(j)}$ and the case for which $y^{(i)} \neq y^{(j)}$; for simplicity we denote $f(y^{(i)})$ by f_i, which leads to

$$\begin{aligned}
\text{FIM}(f, r) &= \sum_{i=1}^{K} \frac{f_i^2 (1-f_i)^2}{f_i r + f_i^2(1-r)} + \sum_{i=1}^{K}\sum_{j\neq i}^{K} \frac{f_i^2 f_j^2}{f_i f_j (1-r)}, \\
&= \sum_{i=1}^{K} \frac{f_i (1-f_i)^2}{r + f_i(1-r)} + \sum_{i=1}^{K}\sum_{j\neq i}^{K} \frac{f_i f_j}{(1-r)},
\end{aligned}$$

where we recall that K denotes the number of possible alleles. We note that $\sum_{j\neq i} f_j = 1 - f_i$ because $\sum_{i=1}^{K} f_i = 1$, therefore we obtain

$$\sum_{i=1}^{K}\sum_{j\neq i}^{K} \frac{f_i f_j}{(1-r)} = \sum_{i=1}^{K} \frac{f_i(1-f_i)}{(1-r)} = \frac{1}{1-r} - \frac{\sum_{i=1}^{K} f_i^2}{1-r},$$

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

and thus the simpler form for the FIM:

$$\mathrm{FIM}(f,r) = \frac{1}{1-r} + \sum_{i=1}^{K} \left\{ \frac{f_i\,(1-f_i)^2}{r + f_i(1-r)} - \frac{f_i^2}{1-r} \right\}.$$

The notation $\mathrm{FIM}(f,r)$ reflects our consideration of the FIM as a function of f and r. We now wonder how to maximize FIM over the vector $f = (f_1, \ldots, f_K)$, for any r. This is a constrained and nonlinear optimization problem since f has to be made of non-negative entries and sums to one (thus f is in the simplex of dimension K). We restrict our attention to $r \in (0,1)$, that is $r \neq 0$ and $r \neq 1$, since the interpretation of FIM as a measure of the precision of the maximum likelihood estimator is only valid when r is away from the boundaries of the parameter space $[0,1]$. For $r \in (0,1)$, the function $f \mapsto \mathrm{FIM}(f,r)$ is finite and continuous, on the simplex which is a compact set, thus it attains a maximum according to the extreme value theorem.

After plotting the contours of the function FIM on the simplex and for different values of r (and perhaps noticing that $f \mapsto \mathrm{FIM}(f,r)$ is symmetric with respect to the center of the simplex), we gather that the maximizer might be $f^\star = (K^{-1}, \ldots, K^{-1})$, irrespective of the value of r. We now prove that this is indeed the case. We do so by considering an f such that $f_1 < f_2$. We will see that we can increase $\mathrm{FIM}(f,r)$ by modifying f as follows: define $\tilde{f}$ as $\tilde{f}_1 = f_1 + \epsilon$, $\tilde{f}_2 = f_2 - \epsilon$ and $\tilde{f}_j = f_j$ for all $j \in \{3, \ldots, K\}$ (if $K \geq 3$). We will see that there exists an $\epsilon > 0$ such that $\mathrm{FIM}(\tilde{f}, r) > \mathrm{FIM}(f,r)$. Since this holds for all f with a pair of non-equal entries, we will be able to conclude that the unique maximizer of FIM is $f^\star = (K^{-1}, \ldots, K^{-1})$.

So let us consider f with $f_1 < f_2$. We start by noting that, for all $f \in (0,1)$,

$$\psi(f+\epsilon) := \frac{(f+\epsilon)\,(1-(f+\epsilon))^2}{r + (f+\epsilon)(1-r)} - \frac{(f+\epsilon)^2}{1-r}$$

can be expanded as $\epsilon \to 0$ as

$$\begin{aligned}
&\frac{f\,(1-f)^2}{r+f(1-r)} + \epsilon \left\{ \frac{1-f}{r+(1-r)f} \left(1 - 3f - \frac{(1-r)f(1-f)}{r+(1-r)f} \right) \right\} + \mathcal{O}\left(\epsilon^2\right) - \frac{(f + 2\epsilon f + \epsilon^2)}{1-r} \\
=&\psi(f) + \epsilon \left\{ \frac{1-f}{r+(1-r)f} \left(1 - 3f - \frac{(1-r)f(1-f)}{r+(1-r)f} \right) - \frac{2f}{1-r} \right\} + \mathcal{O}\left(\epsilon^2\right),
\end{aligned}$$

where $\mathcal{O}(\epsilon^2)$ refers to terms which behave as ϵ^2 when $\epsilon \to 0$ and thus are negligible in front of the term in ϵ. From this we deduce that $\psi(f+\epsilon) - \psi(f) = \epsilon h(f) + \mathcal{O}(\epsilon^2)$ with

$$h(f) := \frac{1-f}{r+(1-r)f} \left(1 - 3f - \frac{(1-r)f(1-f)}{r+(1-r)f} \right) - \frac{2f}{1-r}.$$

We now show that $f \mapsto h(f)$ is decreasing in f over $[0,1]$. We do so by bruteforce differentiation, yielding

$$\frac{d}{df} h(f) = -\frac{2r}{(1-r)\,(r+(1-r)f)^3}.$$

We see that the above expression is strictly negative for all r and f so that $f \mapsto h(f)$ is strictly decreasing.

bioRxiv preprint doi: https://doi.org/10.1101/575985; this version posted March 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The fact that $f \mapsto h(f)$ is strictly decreasing allows us to conclude the proof. Indeed, combined with the assumption $f_1 < f_2$, we have $h(f_1) > h(f_2)$. Therefore,

$$\begin{aligned}\mathrm{FIM}(\tilde{f}, r) - \mathrm{FIM}(f, r) &= \psi(f_1 + \epsilon) - \psi(f_1) + \psi(f_2 - \epsilon) - \psi(f_2) \\ &= \epsilon\,(h(f_1) - h(f_2)) + \mathcal{O}(\epsilon^2),\end{aligned}$$

from which we deduce that there is an $\epsilon > 0$ small enough so that $\mathrm{FIM}(\tilde{f}, r) - \mathrm{FIM}(f, r) > 0$. To summarize, if f is such that one of its components is strictly greater than another component, then we can increase the objective function FIM. We deduce that the function $f \mapsto \mathrm{FIM}(f, r)$ is uniquely maximized at $f^\star = (K^{-1}, \ldots, K^{-1})$, for which no component is greater than another one.