OXFORD
G3 Genes | Genomes | Genetics

G3, 2025, **15(5),** jkaf018
https://doi.org/10.1093/g3journal/jkaf018
Advance Access Publication Date: 30 January 2025
Investigation

Systematic bias in malaria parasite relatedness estimation

Somya Mehra ,[1,2,3,4,*] Daniel E. Neafsey ,[1,2] Michael White ,[5] Aimee R. Taylor [5,*]

[1]Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA
[2]Infectious Disease and Microbiome Program, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
[3]School of Mathematics and Statistics, The University of Melbourne, Parkville 3010, Australia
[4]Present address: Mahidol Oxford Tropical Medicine Research Unit, Mahidol University, Bangkok 10400, Thailand
[5]Infectious Disease Epidemiology and Analytics G5 Unit, Institut Pasteur, Université Paris Cité, Paris 75015, France

*Corresponding author: Mahidol Oxford Tropical Medicine Research Unit, Mahidol University, Bangkok 10400, Thailand. Email: somya@tropmedres.ac; *Corresponding author: Infectious Disease Epidemiology and Analytics G5 Unit, Institut Pasteur, Université Paris Cité, Paris 75015, France. Email: ataylor@pasteur.fr

Genetic studies of *Plasmodium* parasites increasingly feature relatedness estimates. However, various aspects of malaria parasite relatedness estimation are not fully understood. For example, relatedness estimates based on whole-genome-sequence (WGS) data often exceed those based on sparser data types. Systematic bias in relatedness estimation is well documented in the literature geared towards diploid organisms, but largely unknown within the malaria community. We characterize systematic bias in malaria parasite relatedness estimation using three complementary approaches: theoretically, under a non-ancestral statistical model of pairwise relatedness; numerically, under a simulation model of ancestry; and empirically, using data on parasites sampled from Guyana and Colombia. We show that allele frequency estimates encode, locus-by-locus, relatedness averaged over the set of sampled parasites used to compute them. Plugging sample allele frequencies into models of pairwise relatedness can lead to systematic underestimation. However, systematic underestimation can be viewed as population-relatedness calibration, i.e., a way of generating measures of relative relatedness. Systematic underestimation is unavoidable when relatedness is estimated assuming independence between genetic markers. It is mitigated when relatedness is estimated using WGS data under a hidden Markov model (HMM) that exploits linkage between proximal markers. The extent of mitigation is unknowable when a HMM is fit to sparser data, but downstream analyses that use high relatedness thresholds are relatively robust regardless. In summary, practitioners can either resolve to use relative relatedness estimated under independence, or try to estimate absolute relatedness under a HMM. We propose various tools to help practitioners evaluate their situation on a case-by-case basis.

Keywords: malaria; relatedness; identity by descent; hidden Markov model; independence model; bias

Introduction

Relatedness is a genome-wide measure of identity-by-descent (IBD); i.e., identity due to common ancestry (Weir *et al.* 2006; Speed and Balding 2015). IBD is defined relative to founders, or to a previous time point, without accounting for more distant coancestry (Thompson 2013). It is a useful concept when studying malaria-causing *Plasmodium* parasites because they sexually recombine (Baton and Ranford-Cartwright 2005). Although obligate, recombination is only effective when genetically distinct parasites recombine. The probability that genetically distinct parasites recombine is context-specific: it depends on parasite diversity, on the prevalence of infected hosts, and on the within-host diversity and prevalence of polyclonal infections among infected hosts (Camponovo *et al.* 2022). As such, relatedness analyses reflect the recent, context-specific history of malaria parasites, generating epidemiologically relevant insight. For example, analyses of relatedness have been used to evaluate and inform efforts to reduce transmission (Shetty *et al.* 2019; Daniels *et al.* 2020; Morgan *et al.* 2020), to elucidate population connectivity on granular spatio-temporal scales (Omedo *et al.* 2017; Taylor *et al.* 2017, 2020; Fola *et al.* 2023; Kebede *et al.* 2023), to characterize the structure of inbred populations (de Oliveira *et al.* 2020; Carrasquilla *et al.* 2022), to resolve transmission heterogeneity (Schaffner *et al.* 2023), and to identify regions of the parasite genome subject to recent selective pressure (Henden *et al.* 2018; Amambua-Ngwa *et al.* 2019; Carrasquilla *et al.* 2022). Relatedness has further applications in clinical trials of antimalarial drugs, i.e., in the classification of *Plasmodium falciparum* reinfection and recrudescence (Plucinski *et al.* 2015) and of *Plasmodium vivax* reinfection, recrudescence, and relapse (Cowell *et al.* 2018; Taylor, Watson, *et al.* 2019).

IBD is not observable. As such, relatedness must be estimated under a statistical model. Various software exist for estimating relatedness between malaria parasites (Henden *et al.* 2018; Schaffner *et al.* 2018; Zhu *et al.* 2019; Gerlovina *et al.* 2022; Taylor 2022a). However, not all aspects of malaria parasite relatedness estimation are fully understood, especially those pertaining to systematic bias. For example, relatedness estimates based on dense whole-genome-sequence (WGS) data often exceed those based on sparse data, with striking zero-inflation of the sparse-data estimates (e.g. see Fig. S2Q of Taylor *et al.* 2017). As another example, consider the allele frequencies that are typically plugged into the models used to estimate relatedness between paired parasites. They are estimated from a sample of parasites assuming inter-parasite independence—an assumption

Received on 26 August 2024; accepted on 22 November 2024
© The Author(s) 2025. Published by Oxford University Press on behalf of The Genetics Society of America.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

that contradicts pairwise relatedness estimation. They are thus liable to systematically bias relatedness estimates (Rousset 2002; Wang J 2002, 2011, 2014, 2022; Yu et al. 2006; Bink et al. 2008; Kang et al. 2010; Weir and Goudet 2017). Weir and Goudet (2017) and others have addressed the consequent notion that relatedness is estimated relative to the average relatedness within the sample from which allele frequency estimates are computed. For diploid organisms, systematic bias is well characterized in the literature (see, for example, Anderson and Weir 2007; Bink et al. 2008; Wang J 2011, 2014, 2022) and alternative approaches have been proposed; for example, the KING-robust estimator, which is designed for dense data and predicated on pairwise genotype counts rather than sample allele frequencies (Manichaikul et al. 2010). Implications of inbreeding and "cryptic relatedness" have been addressed in the context of forensic typing (Weir 1994), and case-control studies to characterize genetic determinants of human disease (Devlin and Roeder 1999; Newman et al. 2001; Voight and Pritchard 2005). However, for practitioners of malaria genomic epidemiology, practical tools, and guidelines that address these biases are, to the best of our knowledge, unavailable.

Here, we characterize systematic biases in malaria parasite relatedness estimation using three complementary approaches. First, we analyze theoretically non-ancestral models of pairwise relatedness, characterizing the ramifications of the use of sample allele frequencies, shedding light on the aforementioned zero-inflation, and establishing common ground with Weir and Goudet (2017). Second, we simulate malaria parasite ancestries over successive generations of inbreeding, verifying independently our theoretical results, and elucidating systematic differences in relatedness estimates that are generated under models that assume inter-marker independence vs marker linkage due to proximity. Our numerical results help to explain differences between estimates based on sparse and dense data. Both our theoretical and numerical results assume parasites are sampled from a single population. Using data on *P. falciparum* parasites from different populations in Guyana (Vanhove et al. 2024) and Colombia (Carrasquilla et al. 2022), we illustrate how our results based on theory and simulation translate empirically. Our empirical results are context-specific. Beyond malaria genomic epidemiology, our results generalize to systems of largely haploid recombining eukaryotes (Fisher et al. 2002; Stauber et al. 2022; Wang et al. 2022; Huang et al. 2023; Sandler et al. 2023) or highly inbred diploid populations for which the haploid model of Leutenegger et al. (2003) is applicable.

Methods

We characterize systematic biases in malaria parasite relatedness by analysing statistical models of pairwise relatedness, simulated data with known parasite ancestries, and *P. falciparum* data from Guyana and Colombia. A summary of notation used throughout our main text is provided in Table 1. A detailed description of the methods described below is available in Appendices A and B, while a glossary of terms is provided in Appendix C.

Statistical models of pairwise relatedness

We derive theoretical results and estimate pairwise relatedness using models that couple a non-ancestral model of latent IBD and non-IBD (nIBD) states describing a pair of parasites along a sequence of marker loci and a locuswise observation model.

The latent state model either assumes independence between markers or accounts for marker linkage under the intuition that IBD segments are fragmented at randomly-distributed recombination breakpoints over successive generations, with a genome-wide-constant recombination rate (Stam 1980; Leutenegger et al. 2003; Taylor, Jacob, et al. 2019). Of key interest is the pairwise relatedness parameter r, which describes the marginal probability that a given marker locus is IBD for a parasite pair, i.e., if we consider a single locus in isolation under the model then, $\mathbb{P}(\text{IBD}) = r$.

Table 1. Summary of terms and notation.

Terms/notation	Working definition
$\overline{\text{IBS}}_\ell$	Locuswise average IBS sharing: the proportion of pairs in a sample (including self–self comparisons) that are IBS at a given marker ℓ
$\overline{\text{IBD}}_\ell$	Locuswise average IBD sharing: the proportion of pairs in a sample (including self–self comparisons) that are IBD at a given marker ℓ
$\text{mean}(\overline{\text{IBD}}_\ell)$	Locuswise average IBD sharing averaged over all markers $\ell = 1, \ldots$
r	Relatedness parameter governing the marginal probability of IBD at each marker under non-ancestral models of pairwise relatedness
$\hat{r}$	Maximum likelihood estimate of the parameter r; an estimate of pairwise relatedness
realized relatedness	Fraction of polymorphic markers across the genome that are IBD for a parasite pair

The locuswise observation model provides an explicit link between latent and observable states. Observable states are either pairs of alleles or descriptives of identity-by-state (IBS); i.e., IBS for a pair of identical alleles and non-IBS (nIBS) otherwise. For IBS descriptives, the observational model takes the form $\mathbb{P}_{\text{obs}}(\text{IBS} \mid \text{IBD}) = 1$ in the absence of genotyping error (since IBD necessarily implies IBS) while $\mathbb{P}_{\text{obs}}(\text{IBS} \mid \text{nIBD})$ must be defined appropriately.

Under the coupled model of relatedness, the marginal likelihood of IBS at a given marker can be written

$$\mathbb{P}(\text{IBS} \mid r) = r + \mathbb{P}_{\text{obs}}(\text{IBS} \mid \text{nIBD}) \cdot (1 - r). \tag{1}$$

Here, we focus on maximum likelihood estimates $\hat{r}$ as an estimate of relatedness between a pair of malaria parasites. We do not address the estimation or identification of IBD segments.

It is standard practice to estimate allele frequencies from data on a set of sampled parasites (either before or after removing replicates of apparent clones) and plug them into observation models. We refer to these frequency estimates as sample allele frequencies and to the observation models into which they are plugged as standard practice observation models. Other terms we use include (n)IBD to refer to IBD and nIBD collectively, (n)IBS to refer to IBS and nIBS collectively, hidden Markov model (HMM) to refer to models of relatedness that allow for linkage between proximal markers, and independence model to refer to models of relatedness that assume marker loci are independent. We use the word model to refer to models of relatedness, the models within the models of relatedness (i.e., (n)IBD models and (n)IBD-to-observation models) and the models within the (n)IBD models and (n)IBD-to-observation models (e.g., IBD-to-allele model). We use the term fraction to indicate pairwise averages over loci across the genome, and proportion to indicate locuswise averages over parasites within a population.

Here, we are primarily concerned with bias stemming from standard practice observation models. As a theoretical comparator, we adopt a corrected model of (n)IBS descriptives with conceptual parallels to Weir and Goudet (2017). The corrected model is not available practically, but illustrates theoretically the partial encoding of population relatedness in the sums of squares of sample allele frequencies. Further details of the observation models are provided in the results section. For clarity, genotyping error is not modeled and theoretical derivations are restricted to the independence model of relatedness.

Simulation model of parasite ancestries

While our theoretical results are derived under a non-ancestral pairwise framework, we perform numerical analyses under an ancestrally-informed simulation of a single parasite population. Although our simulation framework is highly simplified and does not fully recapitulate epidemiological reality, it enables independent verification of our theoretical results and allows us to examine the consequences of marker linkage.

A detailed description of the simulation model is provided in Appendix B. Under our simulation framework, IBD is defined relative to founders that pre-date parasites in generation zero. Generation zero is a key time point at which we fix the parasite population size and thus initiate inbreeding (e.g., the start of some control strategy), and the time point against which background and recent relatedness is defined: IBD present in generation zero is low-level and spread maximally across parasite genotypes; recent IBD that post-dates generation zero can range from low to high level and is concentrated on subsets of parasite genotypes. The main assumptions of the simulation model are as follows:

- Sufficient temporal separation between "generation zero" and a bygone outbred founder population to break down marker linkage, whereby the ancestry of each individual in generation zero is sampled uniformly over the set of all possible founder mosaics; this construction distributes ancient low-level background relatedness broadly across the generation zero population.
- Discrete, non-overlapping generations of inbreeding from generation zero onwards, modeled using a transitive relationship graph with sibling/clonal/stranger edges as per Taylor, Watson, *et al.* 2019; relationship graphs are obtained by amalgamating uniformly sampled subgraphs of a fixed size (Taylor 2022b), biasing the distribution towards small, balanced clusters of siblings and clones.
- Stochastic drift arising from the random crossing of parental genotypes from generation $(k-1)$ to yield filial genotypes in generation k, with an augmented probability of selfing and sibling–sibling crosses serving as a proxy for monoclonal mosquito infection and serial cotransmission (Wong *et al.* 2018).
- The absence of immigration, mutation, selection, and population structure under a small, fixed parasite population size.
- A genome-wide-constant recombination rate, with markers treated as nominal point polymorphisms.

When we analyze simulated data, the ground truth against which we characterize systematic bias in pairwise relatedness estimates is the fraction of polymorphic marker loci that are IBD for a given parasite pair; otherwise known as realized relatedness (Speed and Balding 2015). Akin to the Wright–Fischer model, in which fixation necessarily occurs, the number of polymorphic markers tends to decrease over successive generations of inbreeding under our simulation model. Because they are predicated on data from polymorphic markers only, pairwise relatedness estimates are liable to increased variability as the set of informative markers is thinned over successive generations. For comparability, we compute realized relatedness over polymorphic markers. There is a conceptual distinction between realized relatedness and the relatedness parameter r of the non-ancestral pairwise model: given a finite number of markers, a fixed value of r gives rise to a *distribution* of realized relatedness (Speed and Balding 2015; Taylor, Jacob, *et al.* 2019). However, since we simulate recombination from first principles, our simulation framework does not yield a direct analogue for the non-ancestral relatedness parameter r.

Case study of *P. falciparum* data from Guyana and Colombia

Using a case study, we show our theoretical and numerical results are practically relevant. Specifically, we analyze a monoclonal subset of high-quality, whole-genome-sequenced *P. falciparum* isolates from passively sampled symptomatic patients in Guyana between 2016 and 2020 (Vanhove *et al.* 2024) and Colombia between 1993 and 2017 (Carrasquilla *et al.* 2022). Variants were called in accordance with best practices stipulated by GATK and the MalariaGEN Pf3k consortium (Carrasquilla *et al.* 2022; Vanhove *et al.* 2024). We restrict our attention to biallelic SNPs that are in core regions of the nuclear genome (Miles *et al.* 2016) and are polymorphic among the isolates, masking any variant calls that are heteroallelic or have read support below five (based on the DP tag). We remove SNPs with missingness >30% across isolates and isolates with missingness >30% across SNPs to yield a WGS dataset comprising $n = 306$ isolates ($n = 278$ from Guyana, $n = 28$ from Colombia) and $n = 30,694$ SNPs. Because our theoretical and numerical results assume parasites are sampled from a single population, results in the main text are restricted to the $n = 278$ isolates from Guyana, among which $n = 16,115$ SNPs are polymorphic. Sparse datasets are generated by downsampling SNPs uniformly at random without replacement.

Data analyses

All data analyses are done in R (R Core Team 2021). Relatedness estimates based on allelic states are generated using the R package `paneljudge` (Taylor 2022a), which implements the HMM with allelic observations and its independent counterpart, and employs a maximum likelihood estimation scheme. The model of independent (n)IBD states with (n)IBS observations has been custom coded in R. We do not estimate relatedness using a HMM with (n)IBS observations. All code and a minimum analysis dataset for the empirical case study is available in an accompanying GitHub repository (Mehra *et al.* 2024): https://doi.org/10.5281/zenodo.14176553

Results

A detailed description of the results below can be found in Appendix A, which starts with an overview of the support (theoretical, numerical, and/or empirical) per result (Table A1). Our theoretical and numerical results are conditioned on a single parasite population; an analysis of empirical data in the presence of population structure is deferred to Appendix A4.

Theoretical results

Standard (n)IBD-to-observation models are twice misspecified

In standard practice, (n)IBD-to-observation models are informed by sample allele frequencies (Wang J 2014; Taylor *et al.* 2017, 2020; Henden *et al.* 2018; Taylor, Jacob, *et al.* 2019; Zhu *et al.* 2019). Taylor, Jacob, *et al.* (2019) allude to potential misspecification arising from this construction: while relatedness estimation seeks to estimate dependence between individuals, the sample allele frequencies, and thus any observation models into which they are plugged, are constructed under the implicit assumption of independence between individuals. Implications for diploid organisms have been addressed extensively by Wang J (2014) and others. Weir and Goudet (2017) have articulated the notion that relatedness is consequently estimated relative to the average relatedness within the sample.

Here, we argue that standard (n)IBD-to-observation models are twice misspecified:

1) The standard IBD-to-allele model, whereby the probability of an IBD pair exhibiting allele q is given by the sample frequency of allele q, is implicitly predicated on the independence of allelic and IBD states which may not hold in reality (Equation (A10)), particularly in the presence of selection.
2) The standard nIBD-to-IBS model, under which the probability of an IBS observation given a nIBD latent state is equal to the proportion of IBS pairs in the set of sampled parasites (calculated by taking the sums of squares of sample allele frequencies), is inflated by the encoding of average locuswise relatedness (Equation (A12)). Weir and Goudet (2017) similarly exploit IBS descriptives rather than allelic states for conceptual clarity.

Removing clonal replicates will not circumvent IBD-to-allele misspecification although it seems reasonable to expect it might mitigate it; otherwise, consequences of the IBD-to-allele misspecification are case-specific and beyond the scope of our current study. For the remainder of this study, we focus on nIBD-to-IBS misspecification because it is more pervasive and has systematic consequences.

An illustration of nIBD-to-IBS misspecification is shown in Box 1, echoing the work of Weir and Goudet (2017). We can summarize this form of misspecification mathematically as follows. Denote by $\overline{\text{IBS}}$ and $\overline{\text{IBD}}$ the proportion of pairs of sampled parasites (including self–self comparisons) that are IBS and IBD, respectively, at a given locus (Table 1, but with locus identifiers dropped for notational convenience). Under the standard nIBD-to-IBS model, we set

$$\mathbb{P}_{\text{standard}}(\text{IBS} \mid \text{nIBD}) = \overline{\text{IBS}}. \tag{2}$$

A proportion of IBS sharing in $\overline{\text{IBS}}$, however, is attributable to IBD. To adjust for locuswise relateness in the set of sampled parasites, we would need to average IBS sharing over nIBD pairs only, that is,

$$\mathbb{P}_{\text{corrected}}(\text{IBS} \mid \text{nIBD}) = \frac{\overline{\text{IBS}} - \overline{\text{IBD}}}{1 - \overline{\text{IBD}}}. \tag{3}$$

The correction (Equation (3)) cannot be implemented in practice because (n)IBD states are unobservable. Nonetheless, it provides a theoretical basis for understanding the misspecification of the standard nIBD-to-IBS model. To see how relatedness structure, in the form of locuswise average relatedness $\overline{\text{IBD}}$, is implicitly embedded in the standard nIBD-to-IBS model, we rearrange Equation (3) to yield

$$\begin{aligned}\mathbb{P}_{\text{standard}}(\text{IBS} \mid \text{nIBD}) &= \overline{\text{IBD}} + \left(1 - \overline{\text{IBD}}\right) \cdot \mathbb{P}_{\text{corrected}}(\text{IBS} \mid \text{nIBD}),\\ &> \mathbb{P}_{\text{corrected}}(\text{IBS} \mid \text{nIBD}),\end{aligned} \tag{4}$$

where the strict inequality is just a technicality. Since we have included self–self comparisons (i.e. comparisons with replacement) for consistency with standard nIBD-to-allele models (Henden et al. 2018; Schaffner et al. 2018), $\overline{\text{IBD}} = 1/n > 0$ for an outbred sample of size n. The finite sample adjustment of Purcell et al. (2007), which considers pairwise comparisons without replacement, yields $\overline{\text{IBD}} = 0$ for an outbred sample. However, under either construction, the form of the theoretical correction to the nIBD-to-IBS model is identical.

In summary, the probability of IBS sharing for nIBD pairs is systematically overestimated under the standard nIBD-to-IBS model, with a particularly pronounced effect in inbred populations with large $\overline{\text{IBD}}$.

Misspecification: relative and zero-valued relatedness estimates

Here, we examine systematic bias in maximum likelihood estimates $\hat{r}_{\text{standard}}$ generated under the standard nIBD-to-IBS model (Equation (2)), relative to hypothetical estimates $\hat{r}_{\text{corrected}}$ generated under the corrected nIBD-to-IBS model (Equation (3)). For conceptual clarity, we enforce the assumption of marker independence (consequences of marker linkage are explored in our numerical analyses).

We can intuit that misspecification of the standard nIBD-to-IBS model will lead to systematic underestimation of pairwise relatedness, that is, $\hat{r}_{\text{standard}} < \hat{r}_{\text{corrected}}$ (Box 1). Two additional consequences for which we have theoretical support are as follows:

1) The standard pairwise relatedness parameter r_{standard} can be reinterpreted as a relative measure of deviation from population-averaged locuswise relatedness (Appendix A.2.2.1). The marginal likelihood of IBS sharing at a given marker ℓ under the standard model with relatedness parameter r_{standard} is equivalent to the likelihood under the corrected model with relatedness parameter $r_{\text{corrected}} = r_{\text{standard}} + \overline{\text{IBD}}_\ell[1 - r_{\text{standard}}]$, where $\overline{\text{IBD}}_\ell$ is the proportion of sampled parasites that are IBD at locus ℓ, that is,

$$\begin{aligned}&\mathbb{P}_{\text{standard}}(\text{IBS}_\ell \mid r_{\text{standard}})\\ &\quad = \mathbb{P}_{\text{corrected}}\left(\text{IBS}_\ell \,\middle|\, r_{\text{corrected}} = \left(r_{\text{standard}} + \overline{\text{IBD}}_\ell[1 - r_{\text{standard}}]\right)\right).\end{aligned}$$

In practice, the average locuswise relatedness $\overline{\text{IBD}}_\ell$ will vary across loci ℓ. In a hypothetical population with identical average locuswise relatedness over all loci, $\overline{\text{IBD}}_\ell = \overline{\text{IBD}}_{\text{constant}}$, we obtain the functional relationship

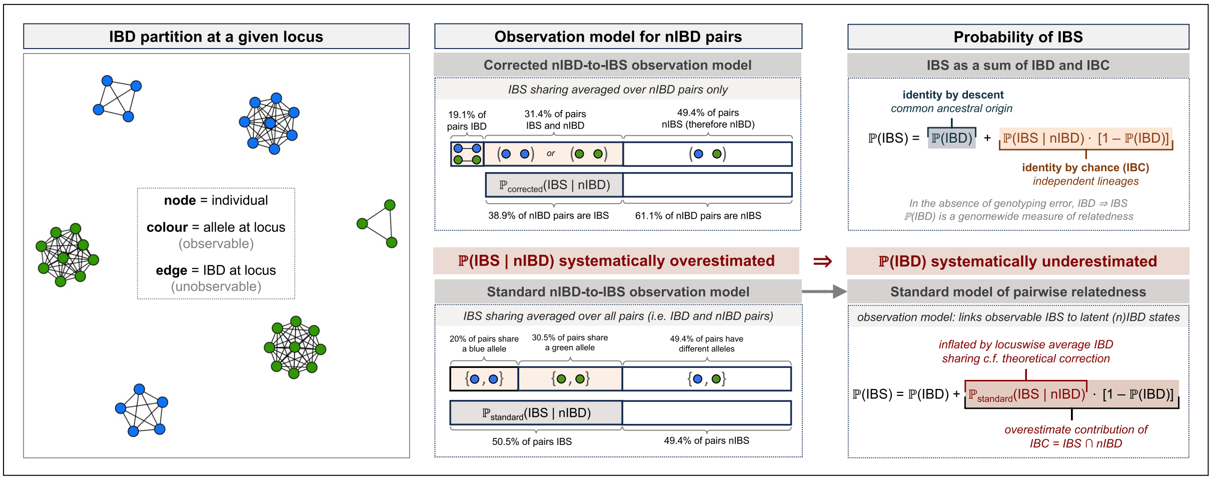
$$r_{\text{standard}} = \frac{r_{\text{corrected}} - \overline{\text{IBD}}_{\text{constant}}}{1 - \overline{\text{IBD}}_{\text{constant}}}. \tag{5}$$

This interpretation of relative relatedness echoes the work of Weir and Goudet (2017).
2) Relatedness estimates can be stratified by average sample relatedness: zero for parasite pairs with relatedness below the sample average, positive otherwise (Appendix A.2.2.2). To understand why this is the case, using Equations (1) and (4), we observe that

$$\begin{aligned}&\mathbb{P}_{\text{standard}}(\text{IBS}_\ell \mid r_{\text{standard}} = 0)\\ &\quad = \overline{\text{IBD}}_\ell + \left(1 - \overline{\text{IBD}}_\ell\right)\mathbb{P}_{\text{corrected}}(\text{IBS}_\ell \mid \text{nIBD}_\ell)\\ &\quad = \mathbb{P}_{\text{corrected}}(\text{IBS}_\ell \mid r_{\text{corrected}} = \overline{\text{IBD}}_\ell).\end{aligned} \tag{6}$$

That is, plugging $r_{\text{standard}} = 0$ into the standard model likelihood of IBS at a given marker ℓ is equivalent to plugging in the average locuswise relatedness $r_{\text{corrected}} = \overline{\text{IBD}}_\ell$ into the corrected model likelihood. In other words, population-averaged locuswise relatedness is implicitly encoded in the standard model even in the case $r_{\text{standard}} = 0$, and parasite pairs with less IBS sharing than predicted under population-averaged relatedness (given explicitly by the threshold (A25)) are assigned zero estimates $\hat{r}_{\text{standard}} = 0$. Similar observations have been made previously by Hall et al. (2012), Weir and Goudet (2017)

Box 1: Illustrating misspecification of the standard nIBD-to-IBS model


IBD partition at a given locus
node = individual
colour = allele at locus
(observable)
edge = IBD at locus
(unobservable)
Observation model for nIBD pairs
Corrected nIBD-to-IBS observation model
IBS sharing averaged over nIBD pairs only
19.1% of pairs IBD
31.4% of pairs IBS and nIBD
49.4% of pairs nIBS (therefore nIBD)
or
ℙcorrected(IBS | nIBD)
38.9% of nIBD pairs are IBS
61.1% of nIBD pairs are nIBS
ℙ(IBS | nIBD) systematically overestimated
⇒
Standard nIBD-to-IBS observation model
IBS sharing averaged over all pairs (i.e. IBD and nIBD pairs)
20% of pairs share a blue allele
30.5% of pairs share a green allele
49.4% of pairs have different alleles
ℙstandard(IBS | nIBD)
50.5% of pairs IBS
49.4% of pairs nIBS
Probability of IBS
IBS as a sum of IBD and IBC
identity by descent
common ancestral origin
ℙ(IBS) = ℙ(IBD) + ℙ(IBS | nIBD) · [1 – ℙ(IBD)]
identity by chance (IBC)
independent lineages
In the absence of genotyping error, IBD ⇒ IBS
ℙ(IBD) is a genomewide measure of relatedness
ℙ(IBD) systematically underestimated
Standard model of pairwise relatedness
observation model: links observable IBS to latent (n)IBD states
inflated by locuswise average IBD sharing c.f. theoretical correction
ℙ(IBS) = ℙ(IBD) + ℙstandard(IBS | nIBD) · [1 – ℙ(IBD)]
overestimate contribution of IBC = IBS ∩ nIBD

Standard nIBD-to-IBS model

At a given locus, $b = 17$ blue nodes and $g = 21$ green nodes are observed. The proportion of pairs that are IBS, which we denote $\overline{\mathrm{IBS}}$, can be written as the sum of squares of the blue $b/(b+g)$ and green $g/(b+g)$ allele frequencies:

$$\overline{\mathrm{IBS}} = (b^2 + g^2)/(b+g)^2 = 0.506.$$

Under the standard nIBD-to-IBS model the probability of an IBS pair given an nIBD pair is equal to the overall proportion of pairs that are IBS:

$$\mathbb{P}_{\text{standard}}(\mathrm{IBS} \mid \mathrm{nIBD}) = \overline{\mathrm{IBS}} = 0.506.$$

Corrected nIBD-to-IBS model

There are three green clusters in the IBD partition, of size $c_1 = 9$, $c_2 = 9$, and $c_3 = 3$. Out of a total of $(b+g)^2 = 1,444$ possible pairwise comparisons, this means that

$$(c_1^2 + c_2^2 + c_3^2)/(b+g)^2 = 0.118 \text{ of pairs are IBD } \textit{and} \text{ share green alleles}$$

Likewise, since there are three blue clusters of size $s_1 = 8$, $s_2 = 5$, and $s_3 = 4$

$$(s_1^2 + s_2^2 + s_3^2)/(b+g)^2 = 0.073 \text{ of pairs are IBD } \textit{and} \text{ share blue alleles}$$

The proportion of pairs which are IBD is then

$$\overline{\mathrm{IBD}} = 0.118 + 0.073 = 0.191,$$

while the proportion of pairs that are IBS and nIBD is

$$\overline{\mathrm{IBS}} - \overline{\mathrm{IBD}} = 0.506 - 0.191 = 0.314.$$

To construct a model of IBS specifically for nIBD pairs, we would ideally focus on IBS sharing in the proportion of $(1 - \overline{\mathrm{IBD}}) = 0.809$ pairs that are nIBD, that is,

$$\mathbb{P}_{\text{corrected}}(\mathrm{IBS} \mid \mathrm{nIBD}) = \frac{\overline{\mathrm{IBS}} - \overline{\mathrm{IBD}}}{1 - \overline{\mathrm{IBD}}} = 0.389.$$

Misspecification of the standard nIBD-to-IBS model

In failing to adjust for IBD sharing under the standard nIBD-to-IBS model, we have overestimated the probability of IBS sharing for nIBD pairs:

$$\mathbb{P}_{\text{standard}}(\mathrm{IBS} \mid \mathrm{nIBD}) = 0.506 > 0.389 = \mathbb{P}_{\text{corrected}}(\mathrm{IBS} \mid \mathrm{nIBD}).$$

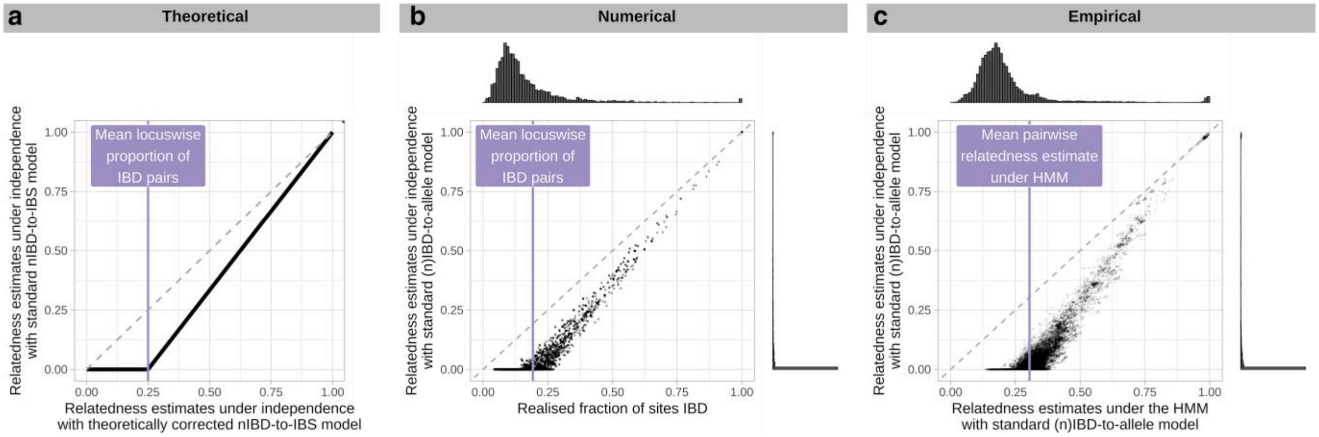


Fig. 1. Characterization of systematic bias in pairwise relatedness under (n)IBD independence. We characterize bias against three comparators: a) $\hat{r}_{\text{corrected}}$ estimates computed theoretically; b) realized relatedness computed using simulated data; and c) $\hat{r}_{\text{standard}}$ estimates computed using the HMM fit to WGS *P. falciparum* data from Guyana. In cases (a) and (b), we recover an elbow-like characteristic with change point near the mean locuswise proportion of IBD pairs $\text{mean}(\overline{\text{IBD}}_\ell)$.

and others. Moment estimators, which permit negative estimates unlike the maximum likelihood estimators, may be more informative for parasite pairs with relatedness below the sample average (Hall *et al.* 2012).

Accounting for variability in $\overline{\text{IBD}}_\ell$ across loci ℓ, we predict estimates $\hat{r}_{\text{standard}}$ against the theoretical comparator $\hat{r}_{\text{corrected}}$ to exhibit a fuzzy elbow-like characteristic, with a change point in the vicinity of the mean locuswise average IBD sharing, $\text{mean}(\overline{\text{IBD}}_\ell)$ (Fig. 1a). If the distribution of $\hat{r}_{\text{corrected}}$ is unimodal and positively skewed, we expect pronounced zero inflation in estimates of r_{standard} under the standard nIBD-to-IBS model.

Numerical results

Theoretical validation

Estimates generated under the independence model of relatedness using simulated data support theoretical results as follows. In line with standard nIBD-to-IBS model misspecification, systematic bias in estimates of r_{standard} increases as $\text{mean}(\overline{\text{IBD}}_\ell)$ increases (Appendix A.3.1). In line with the proposed nIBD-to-IBS model correction, $\hat{r}_{\text{corrected}}$ are largely unbiased (Appendix A.3.2, Fig. 2a vs 2c). Bias-mitigation under (n)IBD independence supports the notion that systematic bias of $\hat{r}_{\text{standard}}$ can be attributed to the partial encoding of sample relatedness in standard observation models. In line with expected zero-valued estimates, $\hat{r}_{\text{standard}} = 0$ for simulated pairs that exhibit a smaller IBS fraction than that which is expected for $r_{\text{standard}} = 0$ under the standard nIBD-to-observation model (Appendix A.3.4). In line with Fig. 1a, plots of $\hat{r}_{\text{standard}}$ against realized relatedness yield an elbow-like characteristic, branching approximately at $\text{mean}(\overline{\text{IBD}}_\ell)$ (Fig. 1b).

Exploiting marker linkage can mitigate bias

While systematic underestimation given (n)IBD independence persists irrespective of marker density (Fig. 2a), it is mitigated as a function of marker density when data are analyzed under the HMM (Fig. 2b). We attribute this trend to the exploitation of increasingly detailed linkage information under the HMM, reducing the reliance of r_{standard} estimation on the standard observation model, which is misspecified. We thus propose a dense data diagnostic that leverages increased precision under the HMM (Appendix A.3.5): for dense data, comparison of $\hat{r}_{\text{standard}}$ estimates under (n)IBD independence vs the HMM is expected to yield an elbow-like characteristic (analogous to Fig. 1b), which can be used to ascertain both the severity of underestimation and approximate $\text{mean}(\overline{\text{IBD}}_\ell)$.

For sparse data that do not encode linkage information, analyses of simulated data under the independence model of relatedness confirm that bias-mitigation would be possible if model-adjustment were available, and that model-adjustment would be sufficient (Fig. 2c).

Relative vs absolute relatedness

Across successive generations of inbreeding, realized relatedness between siblings (gray shading, Fig. 3) is systematically enriched above 0.5. For dense data, we can interpret $\hat{r}_{\text{standard}}$ under the HMM (navy blue, Fig. 3) as a measure of absolute relatedness which recapitulates this enrichment. In contrast, $\hat{r}_{\text{standard}}$ for siblings under (n)IBD independence (orange, Fig. 3) is generally centred around 0.5, supporting interpretation as a measure of relative relatedness. This points towards two choices for practitioners: the analysis of absolute relatedness under a HMM, which requires sufficiently dense genotypic data; or the use of relative relatedness under (n)IBD independence, which warrants careful interpretation in cross-population comparisons (the degree of bias may vary across transmission settings, but relative relatedness may have utility, for instance, as a relationship proxy).

Empirical results

Case study: inbred parasite population from Guyana

To illustrate the practical implications of our findings, we analyze WGS data from $n = 278$ high-quality *P. falciparum* isolates (deemed to be monoclonal) sampled from patients in Guyana in 2016–2020 (Vanhove *et al.* 2024).

Based on our numerical results (specifically, Fig. 2b), estimates of relatedness generated under the HMM using WGS data are expected to exhibit relatively little bias and therefore serve as a pragmatic gold-standard. To gauge the severity of systematic underestimation due to standard observation model misspecification, we draw on the proposed dense data diagnostic: a comparative plot of dense data $\hat{r}_{\text{standard}}$ estimates generated under the independence model and the HMM (Fig. 1c). The position of the change point suggests that the mean locuswise proportion of IBD pairs is in the vicinity

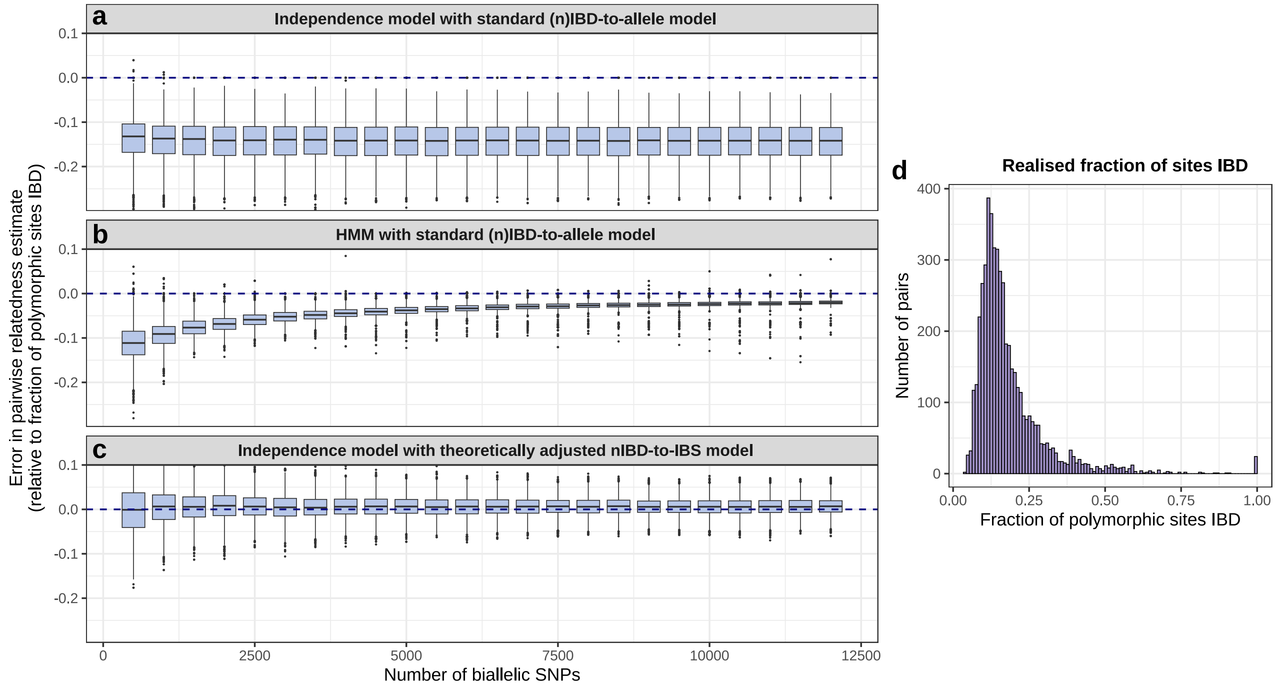


Fig. 2. Summary of pairwise relatedness for simulated data after 10 generations of inbreeding. Systematic bias is shown as a function of marker density, with the fraction of (polymorphic) marker loci IBD taken as the ground truth, for a) $\hat{r}_{\text{standard}}$ under (n)IBD independence and the standard (n)IBD-to-allele model; b) $\hat{r}_{\text{standard}}$ computed using the HMM and the standard (n)IBD-to-allele model; c) $\hat{r}_{\text{corrected}}$ under (n)IBD independence and the corrected nIBD-to-IBS model. d) Histogram of realized relatedness (fraction of polymorphic sites that are IBD for simulated parasites).

of mean($\overline{\text{IBD}}_\ell$) ≈ 30%. Across parasite pairs, the average estimate $\hat{r}_{\text{standard}}$ under the HMM using WGS data is 30.4% (vertical line, Fig. 1c).

In addition to systematic bias stemming from model misspecification, uncertainty due to marker sparsity may become significant for sparse data (Taylor *et al.* 2020). We expect the point at which uncertainty/variance obfuscates the elbow-like characteristic to be dependent on the average relatedness within the set of sampled parasites. A sparse-dense data diagnostic, comprising comparative plots of estimates generated under (n)IBD independence using sparse data vs estimates generated under the HMM using dense data, can be used to elucidate this trade-off because under (n)IBD independence, increasing marker density does not mitigate systematic bias but does reduce uncertainty in pairwise relatedness estimates. For $n = 278$ *P. falciparum* isolates from Guyana, Fig. 4a suggests bias due to elevated population relatedness dominates uncertainty due to marker sparsity: even at low marker densities, systematic underestimation is apparent.

A possible strategy for offsetting systematic bias is to exploit linkage information under the HMM regardless of marker density. However, systematic underestimation is only partially mitigated when the HMM is fit to sparse data (Fig. 4b). Unlike estimates under the independence model—which can be interpreted as relative measures—sparse data HMM estimates do not have a clear interpretation.

The error structure in sparse data HMM estimates is heteroskedastic: underestimation is most pronounced for parasite pairs with low relatedness. Genomic epidemiology applications typically focus on highly related parasite pairs, often using a threshold-based approach (Taylor *et al.* 2017; Henden *et al.* 2018; Miotto *et al.* 2020; Fola *et al.* 2023; Carrasquilla *et al.* 2022; Harrison *et al.* 2023). Since relatedness is systematically underestimated using sparse data, false positives (pairs with low levels of relatedness that appear highly related) are relatively rare. The sensitivity of sparse data HMM estimation in identifying pairs with relatedness above a threshold is shown in Fig. 4c. When thresholds are high, sparse data HMM estimates suffice; sensitive classification for low thresholds, however, requires higher resolution data (Harrison *et al.* 2023). We are reluctant to posit marker density thresholds necessary to sensitively identify highly related parasite pairs in generality, because the degree of linkage structure within a set of sampled parasites depends on the distribution of shared IBD segment lengths—which, in turn, is driven by demographic processes that we neither fully understand, nor control. Downsampling WGS data provides insight on a case-by-case basis.

Population structure

Since our theoretical and numerical results assume a single parasite population, we additionally analyze jointly the $n = 278$ isolates from Guyana (Vanhove *et al.* 2024) and $n = 28$ isolates from Colombia (Carrasquilla *et al.* 2022). Principal coordinates analysis (PCoA) of pairwise fractions of IBS markers yields two distinct clusters, stratified by country (Fig. A.12). A multimodal distribution of IBS fractions, and systematic differences between relatedness estimates under the standard nIBD-to-IBS vs (n)IBD-to-allele model corroborate the presence of population structure (Appendix A.4.1), and aid in explaining deviations from our theoretical and numerical results because they are conceptually aligned with the present framework.

In this setting, the dense data diagnostic is characterized by multiple elbows, corresponding to different within- and between-population comparisons (Fig. A.15), and the severity of zero-inflation under (n)IBD independence varies for the major/minor subpopulation depending on the use of allelic states or IBS descriptives (Fig. A.16). Given sample allele frequencies constitute a weighted average across subpopulations, the interpretation of zero inflation and branch points in the dense data diagnostic is

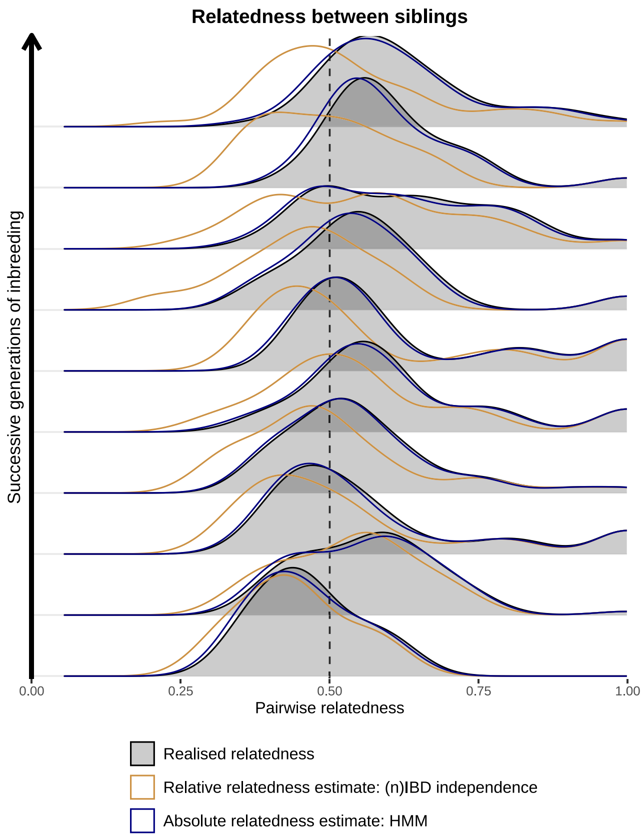


Fig. 3. Summary of pairwise relatedness for simulated siblings across successive generations of inbreeding. We compare realized relatedness (i.e., the fraction of (polymorphic) sites that are IBD for a given parasite pair) against $\hat{r}_{\text{standard}}$ predicated on the standard (n)IBD-to-allele model with (n)IBD independence (relative relatedness estimates) vs the HMM (absolute relatedness estimates).

unclear; an observation model predicated on subpopulation-stratified allele frequencies (Schaffner *et al.* 2018), may yield more interpretable results. Further discussion is provided in Appendix A.4.

Discussion

Summary and interpretation of results

We characterize systematic bias in malaria parasite relatedness estimation, providing theoretical, numerical and/or empirical support for our results (Table A1). In the context of diploid organisms, biases in relatedness estimation are well established (Anderson and Weir 2007; Bink *et al.* 2008; Wang J 2014), and estimators that are more robust to inbreeding are available (Manichaikul *et al.* 2010; Wang J 2011, 2022). In the context of malaria parasites, which are haploid and can self, biases in relatedness estimation are not broadly recognized. Our results are not limited to malaria parasites, and extend to any system concerned with pairwise relatedness of predominately haploid recombining eukaryotes (e.g. *Cryptosporidium hominis* (Huang *et al.* 2023) and *Cryptosporidium parvum* (Wang *et al.* 2022), leading causes of human and zoonotic cryptosporidiosis respectively; *Coccidioides* species which give rise to human coccidioidomycosis (Fisher *et al.* 2002); *Cryphonectria parasitica*, the pathogenic agent responsible for Chestnut blight (Stauber *et al.* 2022) and *Marchantia polymorpha*, a model species of liverwort (Sandler *et al.* 2023)) or highly inbred populations of diploid organisms for which pairwise relatedness can be interrogated using a haploid model (Leutenegger *et al.* 2003).

Relatedness estimates with relatively low root mean squared error can be generated using moderate marker counts under a standard model of pairwise relatedness when it is well specified (Taylor, Jacob, *et al.* 2019). Misspecification of standard (n)IBD-to-observation models, which are predicated on sample allele frequencies, constitutes our conceptual starting point (Rousset 2002; Wang J 2014; Weir and Goudet 2017; Taylor, Jacob, *et al.* 2019). Theoretically, we show via the nIBD-to-IBS model that the implicit embedding of population-averaged locuswise relatedness in standard observation models is pervasive (Weir and Goudet 2017) and can lead to the systematic underestimation of pairwise relatedness, especially when background relatedness is high. Larremore (2019) characterize conceptually concordant systematic underestimation of pairwise overlap between subsampled *P. falciparum var* gene repertoires. Guo *et al.* (2024) have shown that positive selection can also bias malaria parasite relatedness estimates, but selection-bias is minor when background relatedness is high. Beyond malaria, our theoretical results bear strong conceptual similarity to the work of Weir and Goudet (2017). Both theoretical and numerical analyses support the reinterpretation of pairwise relatedness under the independence model as a relative measure: non-zero estimates are calibrated intrinsically for relatedness averaged over the set of sampled parasites; zero estimates flag parasite pairs with below-average relatedness, but are otherwise uninformative (Hall *et al.* 2012; Weir and Goudet 2017). Numerically, we show that exploiting linkage structure using a HMM can mitigate bias in absolute relatedness estimates when genotypic data are sufficiently dense, and propose a dense data diagnostic to assess the severity of systematic bias. Using WGS *P. falciparum* data, we illustrate the use of the dense data diagnostic and characterize consequences of marker sparsity. Sparse data HMM estimates are difficult to interpret because systematic underestimation is only partially mitigated; nevertheless, threshold-based classification of HMM sparse data estimates is relatively robust.

The relevance of our findings is context-specific: the demand for absolute vs relative estimates varies on a case-by-case basis; the severity of systematic bias in absolute estimates depends on the average relatedness of sampled parasites, and thus on the sampling scheme and the parasite population; similarly, mitigation of bias under the HMM requires dense data and depends on linkage structure, which is population specific. In many malaria studies, data on most isolates are sparse, because sparse data are often of greater practical utility (Apinjoh *et al.* 2019; Noviyanti *et al.* 2020). Some studies complement sparse data analyses using WGS data on a subset of isolates. Others complement sparse data analyses using published WGS data from MalariaGEN, an invaluable community resource (Ahouidi *et al.* 2021). When all isolates have WGS data, practitioners can generate estimates at will, knowing bias under the HMM fit to WGS data is limited. For sparse data in isolation, HMM estimates can be categorized using high relatedness thresholds, or estimates can be generated under (n)IBD independence and interpreted relatively. When most data are sparse but there are some complementary (or preliminary) WGS data that are representative of the population of interest, the following three plots can help to decide between generating sparse-data estimates under (n)IBD independence or the HMM. Plot WGS (n)IBD independence estimates against WGS HMM estimates (dense data diagnostic) to assess whether caution around sparse-data HMM

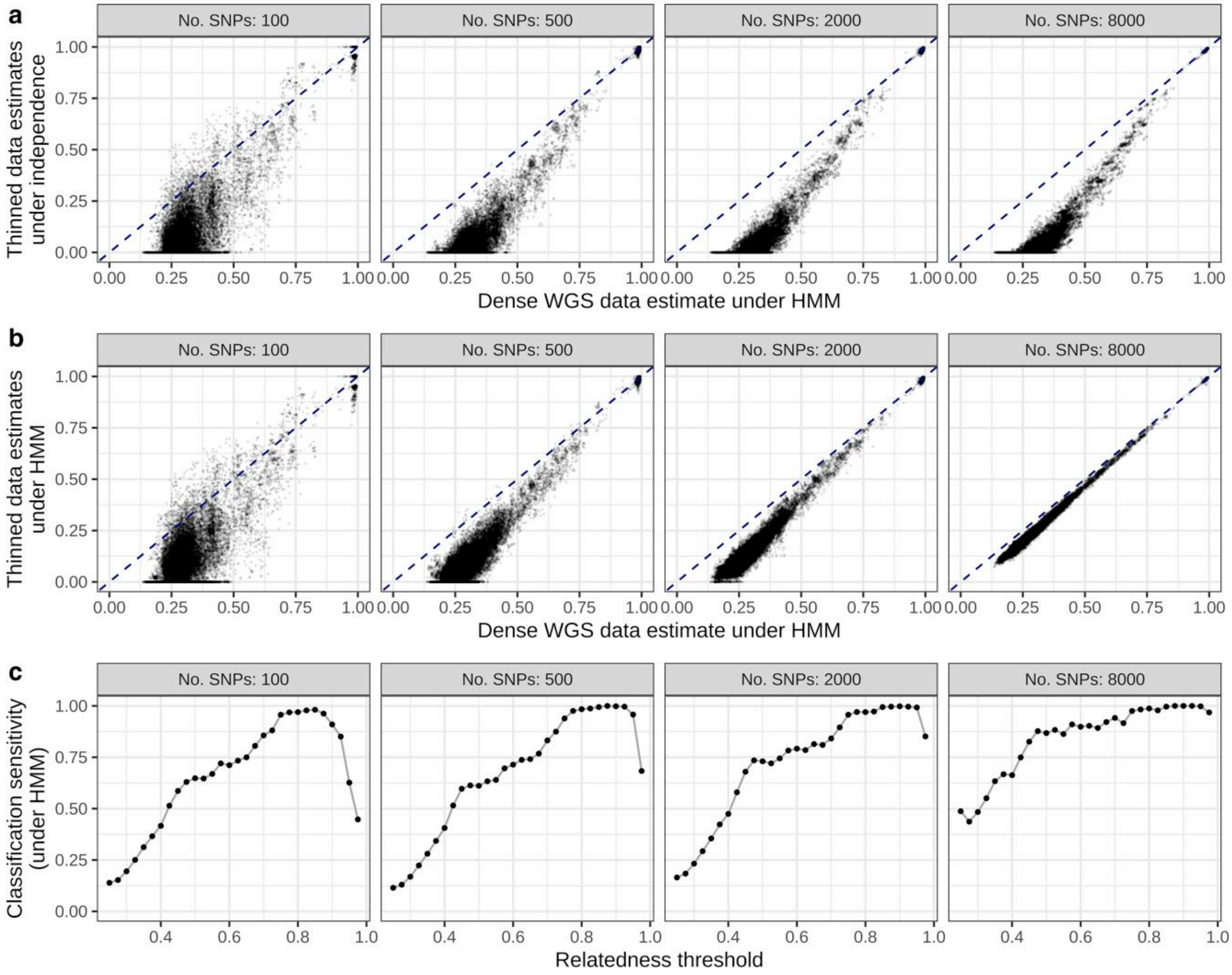


Fig. 4. Consequences of marker sparsity on pairwise relatedness estimates for 278 *P. falciparum* isolates from Guyana. Thinned data are generated by downsampling SNPs in the dense data (complete WGS data) uniformly at random without replacement. a) Relatedness estimated under (n)IBD independence with thinned data vs the HMM with dense data. b) Relatedness estimated under the HMM with thinned data vs dense data. c) Related-pair classification sensitivity of relatedness estimated under the HMM fit to thinned data. For each threshold, we record a true positive (TP) if estimates generated under the HMM for both thinned and dense data exceed the threshold; and a false negative (FN) if only the HMM estimate for dense data exceeds the threshold; sensitivity is given by the ratio TP/(TP+FN). All estimates are generated using the standard (n)IBD-to-allele model.

estimates is warranted. Plot downsampled (n)IBD independence estimates against WGS HMM estimates to assess the relative importance of bias vs uncertainty in sparse-data HMM estimates. Plot downsampled HMM estimates against WGS HMM estimates to identify thresholds above which highly-related parasite pairs can be identified sensitively.

Limitations

IBS descriptives

The theoretical reinterpretation of pairwise relatedness under independence as a relative measure is predicated on IBS descriptives, as in Weir and Goudet (2017). In practice, pairwise relatedness is estimated using allelic states rather than IBS descriptives (Henden *et al.* 2018; Schaffner *et al.* 2018; Taylor, Jacob, *et al.* 2019; Zhu *et al.* 2019; Taylor *et al.* 2020; Taylor 2022a). Misspecification of the IBD-to-allele model—arising from the non-independence of IBD and allelic states—is sensitive to the unobservable relatedness structure at each locus. It is not necessarily rectified by removing replicates of clonal parasites, and may be particularly pronounced in the presence of selection (Guo *et al.* 2024). As such, there may be systematic differences between pairwise relatedness estimates based on IBS descriptives vs allelic states. A rigorous examination of the latter would potentially require joint estimation of relatedness across a set of sampled parasites, which is beyond the scope of the current manuscript.

Demographic processes

Demographic processes have a large effect on real data but are neither explored theoretically nor numerically, to avoid overburdening the manuscript. The effects of selection and immigration on the underlying relatedness structure of a set of sampled parasites are not considered. Our ancestral simulation model is very simple and does not recapitulate epidemiological reality. Transmission dynamics—which govern the propensity for selfing vs inbreeding vs outbreeding—are not explicitly modeled. We do not account for fluctuations in parasite population sizes, for instance, due to bottlenecks or control interventions, which may be particularly relevant in settings with high background relatedness.

Population structure

We did not explore population structure theoretically or numerically. In the presence of population structure, non-stratified sample allele frequencies can introduce further biases in relatedness estimates (Anderson and Weir 2007; Manichaikul *et al.* 2010; Rohlfs *et al.* 2012; Morrison 2013). Given variable levels of relatedness within and between populations, we posit the emergence of multiple elbows in the dense data diagnostic (see Appendix A.4 for an empirical example); however, we have not explored population structure thoroughly enough to know if a non-inbred but structured population could yield an elbow-like characteristic. Zero-inflation may not be present for cross-population comparisons with lower relatedness than the cross-population average. Our simulation model could be used to simulate different populations that unite and mix in order to interrogate the consequences of population structure. The simulation model of Guo *et al.* (2024), which accounts for population structure, could be used as an alternative or complementary approach. The framework of Weir and Goudet (2017), which jointly characterizes relatedness and population structure, could also be drawn on. Ancestry-informed estimation of sample allele frequencies can aid in mitigating biases due to population structure (Thornton *et al.* 2012; Moltke and Albrechtsen 2014). The cross-population model of hmmIBD implements a (n)IBD-to-allele model predicated on sample allele frequencies stratified by user-defined populations (Schaffner *et al.* 2018). It generates interpretable relatedness estimates in the presence of population structure, and could be analyzed to yield theoretical insight.

Multi-allelic markers

We derived results for multi-allelic markers in the absence of genotyping error, under the assumption that there are no systematic biological differences between multi-allelic and biallelic markers. That is to say, both are treated as nominal point polymorphisms, whose alleles can be modeled as categorical random variables (Taylor, Jacob, *et al.* 2019). We do not report our results on multi-allelic markers, because they are not meaningfully different to biallelic markers under these simplistic assumptions. In reality, the ancestral processes governing biallelic markers and multi-allelic markers likely differ. Multi-allelic markers additionally possess an ordinal genotyping error structure that is overlooked in current methods commonly used to estimate relatedness between malaria parasites (Henden *et al.* 2018; Schaffner *et al.* 2018). The practical ramifications could be explored using empirical data: qualitatively compare graphs of relatedness (as in Henden *et al.* 2018), where relatedness is estimated using multi-allelic markers (e.g. microhaplotypes) vs biallelic markers (e.g. SNPs) of equal informativeness (i.e. markers sets whose composite score of average effective cardinality multiplied by marker count is the same; see Taylor, Jacob, *et al.* 2019).

Data sparsity under simulation

We explore data sparsity using real data only. Our simulation model does not lend itself to the exploration of the effect of data sparsity: we intentionally elevate the recombination rate to compensate for computational constraints (small population size, few generations, single chromosome) and simplifying assumptions (no immigration, no mutation). Elevating the recombination rate generates simulated data whose estimates of relatedness resemble real data; it also leads to more IBD segments among simulated data which offsets the effect of data sparsity.

Future work

Generating a community resource

The dense data diagnostic requires WGS data, which are often financially prohibitive (Apinjoh *et al.* 2019; Noviyanti *et al.* 2020). As a community resource, a catalogue of dense data diagnostic plots could be generated for published WGS datasets (Ahouidi *et al.* 2021). Doing so would first require partitioning samples into sets for which population structure is not a complicating factor.

Recurrent infection classification

Systematically elevated population relatedness likely impairs the resolution of recurrent infections during therapeutic efficacy studies. How best to deal with the problem is not yet understood. One existing approach involves embedding a relatedness inflation factor into a classification model (Taylor, Watson, *et al.* 2019). Our results suggest inflation factors in classification models that use sample allele frequencies are obsolete because sample allele frequencies already partially encode population-averaged locuswise relatedness.

Analyses using confidence intervals

A key consideration for sparse genotypic data is the significance of systematic bias in the presence of uncertainty attributable to marker sparsity. While we suggest downsampling loci to evaluate this trade-off heuristically, computing confidence intervals, as recommended in Taylor, Jacob, *et al.* (2019), may yield a more statistically principled approach.

Algorithmic correction of misspecification

Unbiased relatedness estimation using sparse data may be possible with joint inference of relatedness and the probability of allele sharing for nIBD parasites (Taylor, Jacob, *et al.* 2019). For instance, Hall *et al.* (2012) and Wang J (2022) adopt an iterative construction to jointly estimate IBD coefficients and allele frequencies, whereby these quantities are alternately readjusted across successive iterations; these methods could be adapted to the present context.

Conclusion

Based on our results practitioners have two options: resolve to use relative relatedness estimated under independence or try to estimate absolute relatedness under a HMM. Because relative estimates are intrinsically adjusted, caution is required when differences across transmission settings are sought after. If relatedness estimates are used as relationship-indicators, relative values are arguably preferable across transmission settings. Caution should be exercised when estimating absolute relatedness using sparse data under a HMM because the extent to which underestimation is mitigated is unknowable, but classification of absolute values using high relatedness thresholds is relatively robust. We are reluctant to prescribe decision thresholds given most use cases are likely to deviate from any contrived examples. Instead, we provide tools to help practitioners evaluate their individual situations on a case-by-case basis.

Data availability

Sequencing data for *P. falciparum* isolates from both Guyana (Vanhove *et al.* 2024) and Colombia (Carrasquilla *et al.* 2022) are available in the NCBI Sequencing Read Archive; accession numbers are provided in supplementary Table S1. A minimum analysis dataset for the present project, comprising a genotype matrix, is available on the accompanying GitHub repository (Mehra *et al.* 2024): https://github.com/somyamehra/PlasmodiumRelatednessBias https://doi.org/10.5281/zenodo.14176553.

Supplemental material available at G3 online.

Acknowledgments

We thank our collaborators Dr Reza Niles-Robin, Dr Horace Cox, and Dr Collete Clementson from the Ministry of Health, Guyana; Dr Vladimir Corredor from Universidad Nacional de Colombia; and Dr Julian C. Rayner from University of Cambridge. We thank all study participants and technicians who contributed to the data analyzed in this project.

Funding

D.E.N. is supported by the Bill & Melinda Gates Foundation (investment ID INV-009416). Under the grant conditions of the Foundation, a Creative Commons Attribution 4.0 Generic License has already been assigned to the Author Accepted Manuscript version that might arise from this submission. D.E.N. is also supported 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.

M.W. is supported by the French Government's Laboratoire d'Excellence "Integrative Biology of Emerging Infectious Diseases" (Investissement d'Avenir grant no. ANR-10-LABX-62-IBEID), and INCEPTION programs (Investissement d'Avenir grant no. ANR-16-CONV-0005). M.W. and A.R.T. are supported by the P. vivax Serology for Elimination Partnership (Investment ID INV-024368) funded by The Bill & Melinda Gates Foundation.

A.R.T. is now funded by the European Union (project number 101110393). Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union or European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

S.M. gratefully acknowledges a Fulbright Future Scholarship (funded by The Kinghorn Foundation), and the Australian–American Fulbright Commission.

Conflicts of interest

The authors have no conflicts of interest to declare.

Literature cited

Ahouidi A, Ali M, Almagro-Garcia J, Amambua-Ngwa A, Amaratunga C, Amato R, Amenga-Etego L, Andagalu B, Anderson TJ, Andrianaranjaka V. 2021. An open dataset of plasmodium falciparum genome variation in 7,000 worldwide samples. Wellcome Open Res. 42. https://doi.org/10.12688/wellcomeopenres.16168.2.

Amambua-Ngwa A, Amenga-Etego L, Kamau E, Amato R, Ghansah A, Golassa L, Randrianarivelojosia M, Ishengoma D, Apinjoh T, Maïga-Ascofaré O, *et al.* 2019. Major subpopulations of Plasmodium falciparum in sub-Saharan Africa. Science. 365(6455): 813–816. https://doi.org/10.1126/science.aav5427.

Anderson AD, Weir BS. 2007. A maximum-likelihood method for the estimation of pairwise relatedness in structured populations. GENETICS. 176(1):421–440. https://doi.org/10.1534/genetics.106.063149.

Apinjoh TO, Ouattara A, Titanji VP, Djimde A, Amambua-Ngwa A. 2019. Genetic diversity and drug resistance surveillance of Plasmodium falciparum for malaria elimination: is there an ideal tool for resource-limited sub-Saharan Africa? Malar J. 18(1):1–12. https://doi.org/10.1186/s12936-019-2844-5.

Baton LA, Ranford-Cartwright LC. 2005. Spreading the seeds of million-murdering death*: Metamorphoses of malaria in the mosquito. Trends Parasitol. 21(12):573–580. https://doi.org/10.1016/j.pt.2005.09.012.

Bink MC, Anderson AD, van de Weg WE, Thompson EA. 2008. Comparison of marker-based pairwise relatedness estimators on a pedigreed plant population. Theor Appl Genet. 117(6):843–855. https://doi.org/10.1007/s00122-008-0824-1.

Camponovo F, Buckee CO, Taylor AR. 2022. Measurably recombining malaria parasites. Trends Parasitol. 39(1):17–25. https://doi.org/10.1016/j.pt.2022.11.002.

Carrasquilla M, Early AM, Taylor AR, Knudson Ospina A, Echeverry DF, Anderson TJ, Mancilla E, Aponte S, Cárdenas P, Buckee CO, *et al.* 2022. Resolving drug selection and migration in an inbred South American Plasmodium falciparum population with identity-by-descent analysis. PLoS Pathog. 18(12):e1010993. https://doi.org/10.1371/journal.ppat.1010993.

Cowell AN, Valdivia HO, Bishop DK, Winzeler EA. 2018. Exploration of plasmodium vivax transmission dynamics and recurrent infections in the peruvian Amazon using whole genome sequencing. Genome Med. 10(1):1–12. https://doi.org/10.1186/s13073-018-0563-0.

Daniels RF, Schaffner SF, Dieye Y, Dieng G, Hainsworth M, Fall FB, Diouf CN, Ndiop M, Cisse M, Gueye AB, *et al.* 2020. Genetic evidence for imported malaria and local transmission in Richard Toll, Senegal. Malar J. 19(1):1–8. https://doi.org/10.1186/s12936-020-03346-x.

de Oliveira TC, Corder RM, Early A, Rodrigues PT, Ladeia-Andrade S, Alves JMP, Neafsey DE, Ferreira MU. 2020. Population genomics reveals the expansion of highly inbred plasmodium vivax lineages in the main malaria hotspot of Brazil. PLoS Negl Trop Dis. 14(10): e0008808. https://doi.org/10.1371/journal.pntd.0008808.

Devlin B, Roeder K. 1999. Genomic control for association studies. Biometrics. 55(4):997–1004. https://doi.org/10.1111/biom.1999.55.issue-4.

Etheridge A. 2011. Some Mathematical Models from Population Genetics: Ecole D'Ete de Probabilites de Saint-Flour XXXIX-2009. Vol. 2012. Springer Science & Business Media.

Fisher MC, Rannala B, Chaturvedi V, Taylor JW. 2002. Disease surveillance in recombining pathogens: Multilocus genotypes identify sources of human coccidioides infections. Proc Natl Acad Sci U S A. 99(13):9067–9071. https://doi.org/10.1073/pnas.132178099.

Fola AA, Moser KA, Aydemir O, Hennelly C, Kobayashi T, Shields T, Hamapumbu H, Musonda M, Katowa B, Matoba J, *et al.* 2023. Temporal and spatial analysis of Plasmodium falciparum genomics reveals patterns of parasite connectivity in a low-transmission district in Southern Province, Zambia. Malar J. 22(1):208. https://doi.org/10.1186/s12936-023-04637-9.

Gerlovina I, Gerlovin B, Rodríguez-Barraquer I, Greenhouse B. 2022. Dcifer: An IBD-based method to calculate genetic distance between polyclonal infections. GENETICS. 222(2):iyac126. https://doi.org/10.1093/genetics/iyac126.

Guo B, Borda V, Laboulaye R, Spring MD, Wojnarski M, Vesely BA, Silva JC, Waters NC, O'Connor TD, Takala-Harrison S. 2024. Strong positive selection biases identity-by-descent-based inferences of recent demography and population structure in Plasmodium falciparum. Nat Commun. 15(1):1–14. https://doi.org/10.1038/s41467-024-46659-0.

Hall N, Mercer L, Phillips D, Shaw J, Anderson AD. 2012. Maximum likelihood estimation of individual inbreeding coefficients and null allele frequencies. Genet Res (Camb). 94(3):151–161. https://doi.org/10.1017/S0016672312000341.

Harrison GA, Mehra S, Razook Z, Tessier N, Lee S, Hetzel MW, Tavul L, Laman M, Amato R, Miotto O. 2023. Defining malaria parasite population subdivisions, transmission dynamics and infection origins using SNP barcodes. medRxiv 23294444. https://doi.org/10.1101/2023.09.04.23294444.

Henden L, Lee S, Mueller I, Barry A, Bahlo M. 2018. Identity-by-descent analyses for measuring population dynamics and selection in recombining pathogens. PLoS Genet. 14(5):e1007279. https://doi.org/10.1371/journal.pgen.1007279.

Huang W, Guo Y, Lysen C, Wang Y, Tang K, Seabolt MH, Yang F, Cebelinski E, Gonzalez-Moreno O, Hou T, *et al.* 2023. Multiple introductions and recombination events underlie the emergence of a hyper-transmissible cryptosporidium hominis subtype in the USA. Cell Host Microbe. 31(1):112–123. https://doi.org/10.1016/j.chom.2022.11.013.

Jiang H, Li N, Gopalan V, Zilversmit MM, Varma S, Nagarajan V, Li J, Mu J, Hayton K, Henschen B, *et al.* 2011. High recombination rates and hotspots in a Plasmodium falciparum genetic cross. Genome Biol. 12(4):1–15.

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S, Freimer NB, Sabatti C, Eskin E. 2010. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 42(4):348–354. https://doi.org/10.1038/ng.548.

Kebede AM, Sutanto E, Trimarsanto H, Benavente ED, Barnes M, Pearson RD, Siegel SV, Erko B, Assefa A, Getachew S, *et al.* 2023. Genomic analysis of plasmodium vivax describes patterns of connectivity and putative drivers of adaptation in Ethiopia. Sci Rep. 13(1):20788. https://doi.org/10.1038/s41598-023-47889-w.

Larremore DB. 2019. Bayes-optimal estimation of overlap between populations of fixed size. PLoS Comput Biol. 15(3):e1006898. https://doi.org/10.1371/journal.pcbi.1006898.

Leutenegger AL, Prum B, Génin E, Verny C, Lemainque A, Clerget-Darpoux F, Thompson EA. 2003. Estimation of the inbreeding coefficient through use of genomic data. Am J Hum Genet. 73(3):516–523. https://doi.org/10.1086/378207.

Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. 2010. Robust relationship inference in genome-wide association studies. Bioinformatics. 26(22):2867–2873. https://doi.org/10.1093/bioinformatics/btq559.

Mehra S, Neafsey DE, White M, Taylor AR. 2024. Somyamehra/PlasmodiumRelatednessBias: Systematic bias in malaria parasite relatedness estimation (version v1) Zenodo. https://doi.org/10.5281/zenodo.14176553.

Miles A, Iqbal Z, Vauterin P, Pearson R, Campino S, Theron M, Gould K, Mead D, Drury E, O'Brien J, *et al.* 2016. Indels, structural variation, and recombination drive genomic diversity in Plasmodium falciparum. Genome Res. 26(9):1288–1299. https://doi.org/10.1101/gr.203711.115.

Miotto O, Sekihara M, Tachibana SI, Yamauchi M, Pearson RD, Amato R, Gonçalves S, Mehra S, Noviyanti R, Marfurt J, *et al.* 2020. Emergence of artemisinin-resistant Plasmodium falciparum with kelch13 C580Y mutations on the island of New Guinea. PLoS Pathog. 16(12):e1009133. https://doi.org/10.1371/journal.ppat.1009133.

Moltke I, Albrechtsen A. 2014. Relateadmix: A software tool for estimating relatedness between admixed individuals. Bioinformatics. 30(7):1027–1028. https://doi.org/10.1093/bioinformatics/btt652.

Morgan AP, Brazeau NF, Ngasala B, Mhamilawa LE, Denton M, Msellem M, Morris U, Filer DL, Aydemir O, Bailey JA, *et al.* 2020. Falciparum malaria from coastal tanzania and zanzibar remains highly connected despite effective control efforts on the archipelago. Malar J. 19(1):1–14. https://doi.org/10.1186/s12936-020-3137-8.

Morrison J. 2013. Characterization and correction of error in genome-wide IBD estimation for samples with population structure. Genet Epidemiol. 37(6):635–641. https://doi.org/10.1002/gepi.2013.37.issue-6.

Moser KA, Aydemir O, Hennelly C, Kobayashi T, Shields T, Hamapumbu H, Musonda M, Katowa B, Matoba J, Stevenson JC, *et al.* 2021. Temporal and spatial analysis of Plasmodium falciparum genomics reveals patterns of connectivity in a low-transmission district in southern province, Zambia. medRxiv 21264576. https://doi.org/10.1101/2021.10.14.21264576.

Munford A. 1977. A note on the uniformity assumption in the birthday problem. Am Stat. 31(3):119–119.

Nei M. 1973. Analysis of gene diversity in subdivided populations. Proc Natl Acad Sci U S A. 70(12):3321–3323.

Newman DL, Abney M, McPeek MS, Ober C, Cox NJ. 2001. The importance of genealogy in determining genetic associations with complex traits. Am J Hum Genet. 69(5):1146–1148. https://doi.org/10.1086/323659.

Nkhoma SC, Trevino SG, Gorena KM, Nair S, Khoswe S, Jett C, Garcia R, Daniel B, Dia A, Terlouw DJ, *et al.* 2020. Co-transmission of related malaria parasite lineages shapes within-host parasite diversity. Cell Host Microbe. 27(1):93–103.

Noviyanti R, Miotto O, Barry A, Marfurt J, Siegel S, Thuy-Nhien N, Quang HH, Anggraeni ND, Laihad F, Liu Y, *et al.* 2020. Implementing parasite genotyping into national surveillance frameworks: Feedback from control programmes and researchers in the Asia–Pacific region. Malar J. 19(1):271. https://doi.org/10.1186/s12936-020-03330-5.

Olivella S, Shiraito Y. 2017. poisbinom: a faster implementation of the Poisson-binomial distribution. R package version 1.0.1.

Omedo I, Mogeni P, Bousema T, Rockett K, Amambua-Ngwa A, Oyier I, Stevenson JC, Baidjoe AY, De Villiers EP, Fegan G. 2017. Micro-epidemiological structuring of Plasmodium falciparum parasite populations in regions with varying transmission intensities in Africa. Wellcome Open Res. 29. https://doi.org/10.12688/wellcomeopenres.11228.2.

Plucinski MM, Barratt JL. 2021. Nonparametric binary classification to distinguish closely related versus unrelated Plasmodium falciparum parasites. Am J Trop Med Hyg. 104(5):1830–1835.

Plucinski MM, Morton L, Bushman M, Dimbu PR, Udhayakumar V. 2015. Robust algorithm for systematic classification of malaria late treatment failures as recrudescence or reinfection using microsatellite genotyping. Antimicrob Agents Chemother. 59(10):6096–6100. https://doi.org/10.1128/AAC.00072-15.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ, *et al.* 2007. Plink: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 81(3):559–575. https://doi.org/10.1086/519795.

R Core Team. 2021. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Rohlfs RV, Fullerton SM, Weir BS. 2012. Familial identification: Population structure and relationship distinguishability. PLoS Genet. 8(2):e1002469. https://doi.org/10.1371/journal.pgen.1002469.

Rousset F. 2002. Inbreeding and relatedness coefficients: What do they measure? Heredity (Edinb). 88(5):371–380. https://doi.org/10.1038/sj.hdy.6800065.

Sandler G, Agrawal AF, Wright SI. 2023. Population genomics of the facultatively sexual liverwort marchantia polymorpha. Genome Biol Evol. 15(11):evad196. https://doi.org/10.1093/gbe/evad196.

Schaffner SF, Badiane A, Khorgade A, Ndiop M, Gomis J, Wong W, Ndiaye YD, Diedhiou Y, Thwing J, Seck MC, *et al.* 2023. Malaria surveillance reveals parasite relatedness, signatures of selection, and correlates of transmission across senegal. Nat Commun. 14(1):7268. https://doi.org/10.1038/s41467-023-43087-4.

Schaffner SF, Taylor AR, Wong W, Wirth DF, Neafsey DE. 2018. hmmibd: Software to infer pairwise identity by descent between haploid genotypes. Malar J. 17:1–4. https://doi.org/10.1186/s12936-018-2349-7.

Shetty AC, Jacob CG, Huang F, Li Y, Agrawal S, Saunders DL, Lon C, Fukuda MM, Ringwald P, Ashley EA, *et al.* 2019. Genomic structure and diversity of Plasmodium falciparum in southeast Asia reveal recent parasite migration patterns. Nat Commun. 10(1):2665. https://doi.org/10.1038/s41467-019-10121-3.

Smith BR, Herbinger CM, Merry HR. 2001. Accurate partition of individuals into full-sib families from genetic data without parental information. Genetics. 158(3):1329–1338.

Speed D, Balding DJ. 2015. Relatedness in the post-genomic era: Is it still useful? Nat Rev Genet. 16(1):33–44. https://doi.org/10.1038/nrg3821.

Speed T. 1997. Genetic map functions. In: Encyclopedia of Biostatistics.

Stam P. 1980. The distribution of the fraction of the genome identical by descent in finite random mating populations. Genet Res (Camb). 35(2):131–155. https://doi.org/10.1017/S0016672300014002.

Stauber L, Croll D, Prospero S. 2022. Temporal changes in pathogen diversity in a perennial plant–pathogen–hyperparasite system. Mol Ecol. 31(7):2073–2088. https://doi.org/10.1111/mec.v31.7.

Taylor AR. 2022a. paneljudge. https://github.com/aimeertaylor/paneljudge/.

Taylor AR. 2022b. Pv3rs. https://github.com/aimeertaylor/Pv3Rs/.

Taylor AR, Echeverry DF, Anderson TJ, Neafsey DE, Buckee CO. 2020. Identity-by-descent with uncertainty characterises connectivity of Plasmodium falciparum populations on the colombian-pacific coast. PLoS Genet. 16(11):e1009101. https://doi.org/10.1371/journal.pgen.1009101.

Taylor AR, Jacob PE, Neafsey DE, Buckee CO. 2019. Estimating relatedness between malaria parasites. GENETICS. 212(4):1337–1351. https://doi.org/10.1534/genetics.119.302120.

Taylor AR, Schaffner SF, Cerqueira GC, Nkhoma SC, Anderson TJ, Sriprawat K, Pyae Phyo A, Nosten F, Neafsey DE, Buckee CO. 2017. Quantifying connectivity between local Plasmodium falciparum malaria parasite populations using identity by descent. PLoS Genet. 13(10):e1007065. https://doi.org/10.1371/journal.pgen.1007065.

Taylor AR, Watson JA, Chu CS, Puaprasert K, Duanguppama J, Day NP, Nosten F, Neafsey DE, Buckee CO, Imwong M, *et al.* 2019. Resolving the cause of recurrent Plasmodium vivax malaria probabilistically. Nat Commun. 10(1):1–11. https://doi.org/10.1038/s41467-019-13412-x.

Thomas SC, Hill WG. 2000. Estimating quantitative genetic parameters using sibships reconstructed from marker data. Genetics. 155(4):1961–1972.

Thompson EA. 2013. Identity by descent: Variation in meiosis, across genomes, and in populations. GENETICS. 194(2):301–326. https://doi.org/10.1534/genetics.112.148825.

Thornton T, Tang H, Hoffmann TJ, Ochs-Balcom HM, Caan BJ, RischN. 2012. Estimating kinship in admixed populations. Am J Hum Genet. 91(1):122–138. https://doi.org/10.1016/j.ajhg.2012.05.024.

Vanhove M, Schwabl P, Clementson C, Early AM, Laws M, Anthony F, Florimond C, Mathieu L, James K, Knox C, *et al.* 2024. Temporal and spatial dynamics of Plasmodium falciparum clonal lineages in Guyana. PLoS Pathog. 20(6):e1012013. https://doi.org/10.1371/journal.ppat.1012013.

Voight BF, Pritchard JK. 2005. Confounding from cryptic relatedness in case-control association studies. PLoS Genet. 1(3):e32. https://doi.org/10.1371/journal.pgen.0010032.

Wang J. 2002. An estimator for pairwise relatedness using molecular markers. GENETICS. 160(3):1203–1215. https://doi.org/10.1093/genetics/160.3.1203.

Wang J. 2004. Sibship reconstruction from genetic data with typing errors. Genetics. 166(4):1963–1979.

Wang J. 2011. Unbiased relatedness estimation in structured populations. GENETICS. 187(3):887–901. https://doi.org/10.1534/genetics.110.124438.

Wang J. 2014. Marker-based estimates of relatedness and inbreeding coefficients: An assessment of current methods. J Evol Biol. 27(3): 518–530. https://doi.org/10.1111/jeb.2014.27.issue-3.

Wang J. 2022. A joint likelihood estimator of relatedness and allele frequencies from a small sample of individuals. Methods Ecol Evol. 13(11):2443–2462. https://doi.org/10.1111/mee3.v13.11.

Wang T, Guo Y, Roellig DM, Li N, Santín M, Lombard J, Kváč M, Naguib D, Zhang Z, Feng Y, *et al.* 2022. Sympatric recombination in zoonotic cryptosporidium leads to emergence of populations with modified host preference. Mol Biol Evol. 39(7):msac150. https://doi.org/10.1093/molbev/msac150.

Weir BS. 1994. The effects of inbreeding on forensic calculations. Annu Rev Genet. 28(1):597–622. https://doi.org/10.1146/genet.1994.28.issue-1.

Weir BS, Anderson AD, Hepler AB. 2006. Genetic relatedness analysis: Modern data and new challenges. Nat Rev Genet. 7(10): 771–780. https://doi.org/10.1038/nrg1960.

Weir BS, Goudet J. 2017. A unified characterization of population structure and relatedness. GENETICS. 206(4):2085–2103. https://doi.org/10.1534/genetics.116.198424.

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

Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, *et al.* 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet. 38(2):203–208. https://doi.org/10.1038/ng1702.

Zhu SJ, Hendry JA, Almagro-Garcia J, Pearson RD, Amato R, Miles A, Weiss DJ, Lucas TC, Nguyen M, Gething PW, *et al.* 2019. The origins and relatedness structure of mixed infections vary with local prevalence of P. falciparum malaria. Elife. 8:e40845. https://doi.org/10.7554/eLife.40845.

Appendix

Appendix A: Detailed methods and results

Table A1. Overview of results, with theoretic, numerical, and/or empirical support.

	Theory	Numerical	Empirical
Sample allele frequencies partially encode relatedness structure *Misspecification of standard (n)IBD-to-observation models*	✓		
Relatedness is systematically underestimated under the independence model *Re-interpretation as a relative measure, intrinsically adjusted for relatedness averaged over the parasite sample; zero-inflation*	✓	✓	
Exploiting linkage structure using the HMM of relatedness can mitigate underestimation for dense datasets *Reduced sensitivity to misspecified (n)IBD-to-observation models*		✓	
A dense data diagnostic: gauging the severity of underestimation *Comparison of dense data estimates under the independence model vs HMM can elucidate average population-level relatedness*		✓	✓
Analysing sparse data under the HMM yields an intermediary regime *Partial but incomplete mitigation of underestimation; may be sufficient to identify highly related parasite pairs*			✓
Additional manifestations of population structure *Multimodal empirical IBS distributions and systematic differences in relatedness estimates with IBS descriptives vs allelic states indicate population structure*			✓

A.1. Methods

Our approach for characterizing systematic biases in malaria parasite relatedness estimation is 3-fold. We begin by deriving results under a theoretical framework. We then construct a simulation model under which we verify our theoretical results and design practical diagnostics. We conclude with a case study of *P. falciparum* data. Throughout, individual is used to refer to a parasite genotype drawn from an infected host, whereas sample is used to refer to a collection of $k = 1, \ldots, n$ parasite genotypes drawn from many infected hosts. See Table A2 for a complete list of notation used throughout this appendix. For brevity and clarity of exposition, this notation differs from the main text: specifically, the sample proportion of pairs that are IBD (IBS) at locus i is denoted d_i (s_i) here, compared with $\overline{\text{IBD}}_i$ ($\overline{\text{IBS}}_i$) in the main text.

A.1.1 Theoretical framework. While allelic concordance or identity-by-state (IBS) is observable, it can be attributed to one of two latent states: identity-by-descent (IBD), reflecting a common ancestral origin; identity-by-chance (IBC), otherwise. Hereafter, we use nIBD as shorthand for "not IBD". Note that IBC ≠ nIBD; rather, IBC = IBS ∩ nIBD. We use (n)IBD as shorthand for both IBD and nIBD, likewise for (n)IBS.

Before proceeding, let us introduce some notation. Given a sample of n not necessarily distinct parasite genotypes, we index each genotype $k = 1, \ldots, n$. We allow each locus $i = 1, \ldots, m$ to harbor an allele in the set $\{1, \ldots, y\}$, where y denotes the maximal cardinality. We denote by $A_i^{(k)}$ the allele observed at locus i in genotype k. The sample frequency of allele q at locus i

$$f_i(q) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}\{A_i^{(k)} = q\}, \tag{A1}$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function. For a given parasite pair (k, ℓ), we further define

- $S_i^{(k,\ell)} = 1$ if locus i is IBS for the pair (k, ℓ) and zero otherwise,

$$S_i^{(k,\ell)} = \mathbb{1}\{A_i^{(k)} = A_i^{(\ell)}\} = \sum_{q=1}^{y} \mathbb{1}\{A_i^{(k)} = q\}\mathbb{1}\{A_i^{(\ell)} = q\}. \tag{A2}$$

- $D_i^{(k,\ell)} = 1$ if locus i is IBD for the pair (k, ℓ) and zero otherwise.

Throughout, all pairwise comparisons include self–self comparisons.

Table A2. Summary of notation and key quantities for Appendix A only.

Quantity	Interpretation	Equation
$A_i^{(k)}$	Allelic state for individual k at locus i (observable)	–
$S_i^{(k,\ell)}$	IBS state for individuals k, ℓ at locus i (observable)	(A2)
$D_i^{(k,\ell)}$	IBD state for individuals k, ℓ at locus i (unobservable)	–
$f_i(q)$	Sample frequency of allele q at locus i (observable)	(A1)
d_i	Sample proportion of pairs IBD at locus i (unobservable)	(A8)
s_i	Sample proportion of pairs IBS at locus i (observable)	(A12)
c_i	Sample proportion of pairs IBC at locus i (unobservable)	(A13)
$r^{(k,\ell)}$	Pairwise relatedness parameter for parasites k, ℓ	(A19)
$\hat{r}^{(k,\ell)}$	MLE of pairwise relatedness parameter for parasites k, ℓ	–

The default for comparative variables includes self–self comparisons (that is, entails sampling with replacement).

Relatedness structure within a parasite population is governed by some unknown ancestral stochastic process, replete with demographic complexity. The characterization and analysis of this ancestral process is beyond our scope. Instead, we seek to analyze a sample of parasite genotypes at a given point in time using a non-ancestral model. In particular, given $\{\mathbf{A}^{(k)}\}_{k=1}^{n}$ for a sample of n individuals, we seek to estimate $\{\mathbf{D}^{(k,\ell)}\}_{k=1,\ell<k}^{n}$ (Fig. A1).

A.1.1.1 Joint model of malaria parasite relatedness. In an ideal setting, we would perform joint inference over the sample of n individuals (Wang 2004; Taylor, Jacob, *et al.* 2019). The construction of a joint model, however, is nontrivial. The principal complication lies in the nonindependence of IBD states $\mathbf{D}^{(k,\ell)}$ between pairs of individuals. Since IBD is a transitive property, rather than considering parasite pairs in isolation, we would need to reconstruct the complete relatedness structure of a population: a system of correlated, transitive graphs—each representing IBD sharing at a given locus (Taylor, Watson, *et al.* 2019)—with the number of allowable configurations per graph growing as Bell numbers as a

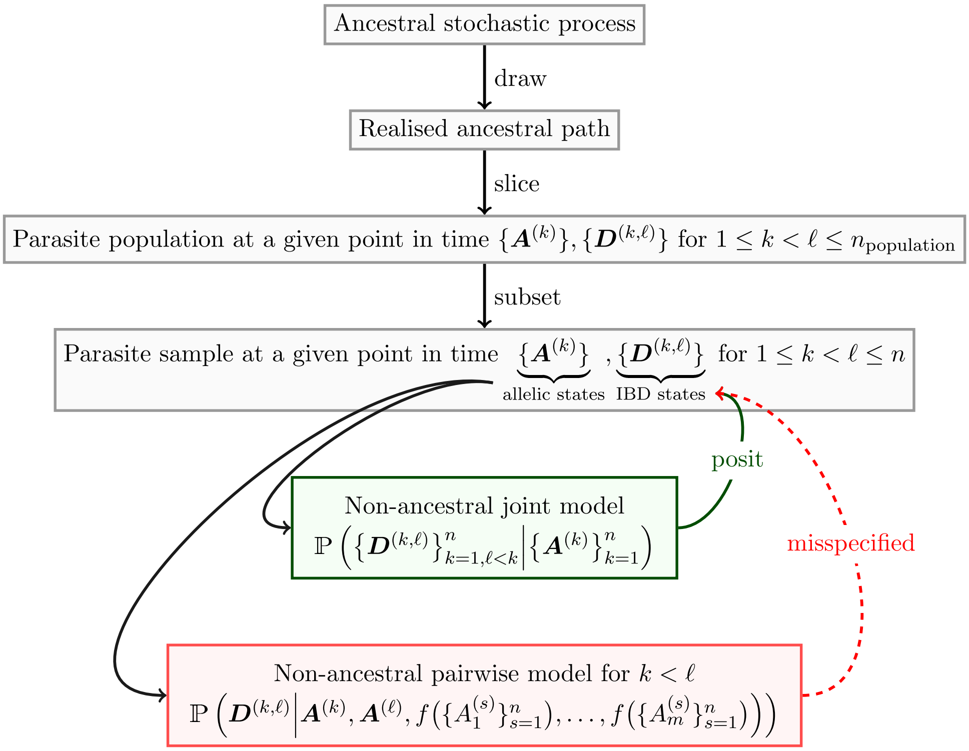

Ancestral stochastic process
draw
Realised ancestral path
slice
Parasite population at a given point in time $\{\boldsymbol{A}^{(k)}\}, \{\boldsymbol{D}^{(k,\ell)}\}$ for $1 \leq k < \ell \leq n_{\text{population}}$
subset
Parasite sample at a given point in time $\underbrace{\{\boldsymbol{A}^{(k)}\}}_{\text{allelic states}}, \underbrace{\{\boldsymbol{D}^{(k,\ell)}\}}_{\text{IBD states}}$ for $1 \leq k < \ell \leq n$
posit
Non-ancestral joint model
$\mathbb{P}\left(\{\boldsymbol{D}^{(k,\ell)}\}^n_{k=1,\ell<k} \middle| \{\boldsymbol{A}^{(k)}\}^n_{k=1}\right)$
misspecified
Non-ancestral pairwise model for $k < \ell$
$\mathbb{P}\left(\boldsymbol{D}^{(k,\ell)} \middle| \boldsymbol{A}^{(k)}, \boldsymbol{A}^{(\ell)}, f(\{A_1^{(s)}\}^n_{s=1}), \ldots, f(\{A_m^{(s)}\}^n_{s=1})\right)$

Fig. A1. Conceptual overview, where $f(\{A_i^{(s)}\}_{s=1}^n)$ denotes the vector of sample allele frequencies for alleles $\{1, \ldots, y\}$ for each locus $i = 1, \ldots, m$.

function of the parasite population size. This poses a significant combinatorial problem, quickly rendering brute force approaches computationally intractable (Taylor, Watson, *et al.* 2019).

A.1.1.2 Non-ancestral pairwise Markov model of (n)IBD states. In practice, relatedness is estimated pairwise using a non-ancestral model, comprising an alternating Poisson process (APP) (Stam 1980). For a given parasite pair, we conceptualize the genome as a mosaic of alternating (n)IBD segments, with the intuition that IBD tracts are fragmented by randomly distributed recombination breakpoints over successive generations (Stam 1980).

While the genome is treated to be continuous under the APP model, we sample a finite number of discrete loci $i = 1, \ldots, m$. To describe the sequence of IBD states

$$\mathbf{D} = (D_1, \ldots, D_m) \in \{0, 1\}^m$$

across loci $i = 1, \ldots, m$ (where parasite identifiers k, ℓ are dropped for notational convenience), Leutenegger *et al.* (2003) construct a discrete-time Markov process. Linkage between successive markers is parameterized by:

- the genomic distance δ_i in units of base pairs (bp) between loci i and $(i + 1)$;
- a constant recombination rate ρ in units of Morgans per base pair (M/bp), typically ascertained from genetic cross experiments (Jiang *et al.* 2011; Miles *et al.* 2016); and
- a (n)IBD latent state switching rate κ that is unobservable and must be inferred.

We assume that locus 1 is IBD with probability r, that is,

$$\mathbb{P}(D_1 = 1) = r, \quad \mathbb{P}(D_1 = 0) = 1 - r.$$

Under the Markov property, D_{i+1} is dependent only on D_i. The complete sequence of IBD states $\mathbf{D}$ is governed by the transition matrix

$$\begin{pmatrix} \mathbb{P}(D_{i+1}(t) = 0 \mid D_i(t) = 0) & \mathbb{P}(D_{i+1}(t) = 1 \mid D_i(t) = 0) \\ \mathbb{P}(D_{i+1}(t) = 0 \mid D_i(t) = 1) & \mathbb{P}(D_{i+1}(t) = 1 \mid D_i(t) = 1) \end{pmatrix} := \begin{pmatrix} 1 - r(1 - e^{-\kappa\rho\delta_i}) & r(1 - e^{-\kappa\rho\delta_i}) \\ (1 - r)(1 - e^{-\kappa\rho\delta_i}) & 1 - (1 - r)(1 - e^{-\kappa\rho\delta_i}) \end{pmatrix}, \tag{A3}$$

as per Taylor, Jacob, *et al.* (2019), where k was used instead of κ for the switch rate parameter. The degree of dependence between successive loci is governed by the product $\kappa\rho\delta_i$: to weaken the dependence between successive loci, we can either increase the switching rate κ, or the genomic distance δ_i between successive loci.

A.1.1.3 Non-ancestral pairwise independence model of (n)IBD states. Under the above-mentioned non-ancestral Markov model in the limit $\kappa \to \infty$ (Taylor, Jacob, *et al.* 2019) or $\delta_i \to \infty$ (Taylor *et al.* 2017), we recover the independence model

$$\mathbb{P}(\mathbf{D}) = \prod_{i=1}^{m} r^{D_i}(1 - r)^{1-D_i}, \tag{A4}$$

where locuswise (n)IBD states are described by independent and identically distributed Bernoulli random variables with success probability r.

A.1.1.4 Observation models. Since (n)IBD states are unobservable, estimation of the pairwise relatedness parameter r necessitates coupling the model of hidden (n)IBD states to a model of observations conditional on (n)IBD states. Herein, observations are either alleles or (n)IBS descriptives (Box A1). We do not model genotyping errors. Elsewhere, where genotyping errors are taken into account (e.g. Taylor, Jacob, et al. 2019), the observation model integrates over latent nonerroneous alleles and is thus made of modules: an (n)IBD-to-latent-allele observation model, and a model capturing genotyping error, e.g.

$$\mathbb{P}\left(\boldsymbol{A}^{(k)}_{\text{obs}}, \boldsymbol{A}^{(\ell)}_{\text{obs}} \mid \boldsymbol{D}^{(k,\ell)}\right) = \sum_{\boldsymbol{A}^{(k)},\boldsymbol{A}^{(\ell)}} \underbrace{\mathbb{P}\left(\boldsymbol{A}^{(k)}, \boldsymbol{A}^{(\ell)} \mid \boldsymbol{D}^{(k,\ell)}\right)}_{\text{(n)IBD−to−latent−allele model}} \times \underbrace{\mathbb{P}\left(\boldsymbol{A}^{(k)}_{\text{obs}}, \boldsymbol{A}^{(\ell)}_{\text{obs}} \mid \boldsymbol{A}^{(k)}, \boldsymbol{A}^{(\ell)}\right)}_{\text{error model}}.$$

Further details of the observations models are provided in the results section (Appendix A.2). They include standard practice observation models (models into which sample allele frequencies are plugged, where sample allele frequencies may be computed before or after removing replicates of seemingly clonal parasites); and a corrected model of independent (n)IBS states, which is not practically available, but useful for validating theory when applied to simulated data.

Box A1: Observations are either alleles or (n)IBS descriptives

Consider the following data on a pair of individual genotypes

genotype 1: A T C G
genotype 2: A C C G

The observations on the pair of genotypes 1 and 2 are

(A, A)	(T, C)	(C, C)	(G, G)	using alleles
IBS	nIBS	IBS	IBS	using (n)IBS descriptives

A.1.2 Simulated data. The model under which simulated data are generated is described in detail in Appendix B. Assumed parameter values, and their respective interpretations, are detailed in Table B1. The model's purpose is not to recapitulate epidemiological reality, but to generate data that can be used to verify theoretical results and facilitate the design of practical diagnostics. To ensure verification is independent, data are simulated under a model built on ancestral principles; they are not simulated under the non-ancestral models used to estimate relatedness.

Under the simulation model, there is a dichotomy between ancient low-level background relatedness stemming from generation zero, and very recent relatedness arising from recent breeding under a small fixed population size. To balance the distribution of IBD segment lengths in spite of this dichotomy, we use a recombination rate and genome size that substantially exceeds that of *P. falciparum* (Miles et al. 2016).

Estimates of pairwise relatedness generated under the non-ancestral relatedness models are compared with true pairwise relatedness values, which are known for simulated data. For a given pair of simulated individuals, we use an approximation of realized relatedness as our truth value, where realized relatedness is the fraction of loci that are IBD (Speed and Balding 2015). Under our simulation model where markers are equidistant, realized relatedness is approximated by the fraction of polymorphic markers that are IBD: IBD is ascertained by comparing ancestral founder mosaics for each pair of simulated individuals; polymorphic markers are those at which there is some variation within the sample, noting that the number of polymorphic markers for a given sample tends to decrease over generations due to loss of diversity. We compute realized relatedness using polymorphic markers because pairwise relatedness estimates are predicated on data from polymorphic markers only, and there may be increasing variability as the set of polymorphic markers is thinned over generations.

A.1.3 P. falciparum data. We illustrate the practical consequences of our theoretical and numerical findings through a case study. We focus on a set of highly quality isolates from passively sampled symptomatic patients in Guyana in 2016–2020 (Vanhove et al. 2024), as well as from Colombia in 1993–2017 (Carrasquilla et al. 2022), that are deemed to be monoclonal. WGS data for these isolates were processed previously in accordance with GATK best practices to yield a genomewide set of variants as described in Carrasquilla et al. (2022) and Vanhove et al. (2024); additional filtration criteria are detailed in the main text. This yields a WGS SNP dataset comprising $n = 306$ isolates ($n = 278$ from Guyana, $n = 28$ from Colombia) and $n = 30,694$ polymorphic biallelic SNPs. Using the dataset, we elucidate the consequences of population structure, which is omitted from our theoretical and numerical analyses. We also illustrate the implications of marker sparsity on pairwise relatedness estimates for $n = 278$ isolates from Guyana ($n = 16{,}115$ polymorphic biallelic SNPs), generating sparse marker panels by down-sampling SNPs sites uniformly at random without replacement.

A.2 Theoretical results

Standard (n)IBD-to-observation models are constructed using sample allele frequencies. When relatedness structure is significant, overlooking it in allele frequency estimation can introduce significant biases (Rousset 2002; Wang J 2002, 2011, 2014, 2022; Wang 2004; Yu et al. 2006; Bink et al. 2008; Kang et al. 2010; Weir and Goudet 2017), with practical consequences including the systematic underestimation of relatedness (Bink et al. 2008) (Case A, Box A2). Removing replicates of seemingly clonal parasites from the sample before computing allele frequencies re-weights allele frequencies in a case-specific manner. It does not account for relatedness between the remaining sample members, however. As such, systematic underestimation of relatedness is liable to persist (Case A, Box A2). An alternative naive approach, evoking the assumption of equifrequent alleles, may yield systematic overestimates of pairwise relatedness (Bink et al. 2008) (Case B, Box A2).

Box A2: Some consequences of misspecified observation models

Likelihood of pairwise IBS sharing for a given locus assuming no error

$$\begin{aligned}\mathbb{P}(\text{IBS}) &= \mathbb{P}(\text{IBS} \mid \text{IBD})\mathbb{P}(\text{IBD}) + \mathbb{P}(\text{IBS} \mid \text{nIBD})\mathbb{P}(\text{nIBD}) \\ &= \mathbb{P}(\text{IBD}) + \mathbb{P}(\text{IBS} \mid \text{nIBD})\big(1 - \mathbb{P}(\text{IBD})\big)\end{aligned}$$

since $\mathbb{P}(\text{IBS} \mid \text{IBD}) = 1$ (in the absence of genotyping error), and where $\mathbb{P}(\text{IBD})$ is a genomewide measure of relatedness.

Implications for relatedness estimates

- $\mathbb{P}(\text{IBS} \mid \text{nIBD})$ overestimated $\Rightarrow \mathbb{P}(\text{IBD})$ underestimated
- $\mathbb{P}(\text{IBS} \mid \text{nIBD})$ underestimated $\Rightarrow \mathbb{P}(\text{IBD})$ overestimated

Observation models

(A) $\mathbb{P}(\text{IBS} \mid \text{nIBD}) \approx \overline{\text{IBS}}$, the observed proportion of IBS pairs in the sample[a]

Calculation:	sums of squares of sample allele frequencies
Possible issue:	IBS due to IBD is not attributed to IBD
Implication:	ℙ(IBS ∣ nIBD) may be overestimated

(B) ℙ(IBS ∣ nIBD) ≈ 1/n where n is locus cardinality

Calculation:	reciprocal of observed allele count
Possible issue:	alleles are not equifrequent
Implication:	ℙ(IBS ∣ nIBD) may be underestimated (Munford 1977)

[a] Replicates of clonal parasites may have been removed from the sample or not

Starting with the standard (n)IBD-to-allele model, and its corrected IBD-to-allele counterpart (Appendix A.2.1), we show that standard observation models can be misspecified given both IBD (Appendix A.2.1.1) and nIBD (Appendices A.2.1.2, A.2.2.1, and A.2.2.2), and reason why removing replicates of seemingly clonal parasites before computing allele frequencies rectifies neither case. More specifically, our theoretical results show that, given IBD, standard observation models are potentially misspecified due to assumed independence between IBD and allelic states, which does not always hold (Appendix A.2.1.1); given nIBD they are misspecified due to partial encoding of the locuswise proportion of IBD pairs within the sums of squares of sample allele frequencies (Weir and Goudet 2017) (Appendix A.2.1.2). Systematic underestimation of pairwise relatedness may be associated with this misspecification. Under the independence model of relatedness with IBS descriptives, we explore additional consequences of this misspecification (Appendix A.2.2). Firstly, echoing the work of Weir and Goudet (2017) (who also adopt IBS descriptives), we re-interpret the pairwise relatedness parameter r as a relative measure, capturing deviation from average relatedness (Appendix A.2.2.1). Secondly, we show pairwise relatedness estimates are stratified by average relatedness, with zero-valued estimates below and positive estimates above (Appendix A.2.2.2), likewise echoing Weir and Goudet (2017). While our theoretical results are restricted to the independence model, the implications of marker density and linkage are explored through analyses of simulated and empirical data (Appendices A.3 and A.4, respectively).

A.2.1 The standard (n)IBD-to-allele model and its corrected IBD-to-allele counterpart. The construction of the standard (n)IBD-to-allele model (Henden *et al.* 2018; Schaffner *et al.* 2018; Taylor, Jacob, *et al.* 2019; Taylor *et al.* 2020) is two-fold:

1) Given a pair of parasites (k, ℓ) is IBD at locus i, the probability that both individuals have allele q is the sample frequency of allele q at locus i:

$$\mathbb{P}_{\text{standard}}(A_i^{(k)} = A_i^{(\ell)} = q \mid D_i^{(k,\ell)} = 1) = f_i(q). \tag{A5}$$

2) Given a pair of parasites (k, ℓ) is nIBD at locus i, the probability of observing alleles (q, u) is the product of the sample frequencies of alleles q and u at locus i:

$$\mathbb{P}_{\text{standard}}(A_i^{(k)} = q,\ A_i^{(k)} = u \mid D_i^{(k,\ell)} = 0) = f_i(q) \cdot f_i(u), \tag{A6}$$

where f_i is a function of 1, …, n individuals (Equation (A1)).

In other words, it is assumed that allele frequencies, when restricted to IBD pairs, are equivalent to sample allele frequencies (Equation (A5)); and the likelihood of each pairwise allelic state, when restricted to non-IBD pairs, is equivalent to the sample incidence of that pairwise allelic state (Equation (A6)).
This description applies when allele frequencies are computed using samples from which replicates of seemingly clonal parasites have been removed or not. As such, all results that follow apply to either case. The only difference between the cases is the specific values of $f_i(q)$, s_i, d_i etc., which will differ on a case-by-case basis not amenable to generalization.

To highlight the misspecification of standard observation models, it is instructive to introduce some unobservable quantities of interest:

- The sample proportion of pairs IBD at locus i harboring allele q

$$d_i(q) = \frac{1}{n^2}\sum_{k=1}^{n}\sum_{\ell=1}^{n} \mathbb{1}\{A_i^{(\ell)} = q\} \cdot D_i^{(k,\ell)} \tag{A7}$$

  in which the indicator features only once since $D_i^{(k,\ell)} = 1$ implies $A_i^{(k)} = A_i^{(l)}$.
- The sample proportion of pairs IBD at locus i, i.e. the locuswise average relatedness

$$d_i = \sum_{q=1}^{y} d_i(q) = \frac{1}{n^2}\sum_{k=1}^{n}\sum_{\ell=1}^{n} D_i^{(k,\ell)}. \tag{A8}$$

The sample proportion of individuals that are IBD at locus i and harbor allele q is $d_i(q)/d_i$. If it were observable, $d_i(q)/d_i$ would be used to approximate the true probability of an IBD pair harboring allele q under an IBD-to-allele model, since we would expect the sample proportion to converge to the true probability as the sample size n goes to infinity, at least under the conventional law of large numbers. As such, we claim that

$$\mathbb{P}_{\text{corrected}}(A_i^{(k)} = A_i^{(\ell)} = q \mid D_i^{(k,\ell)} = 1) \approx \frac{d_i(q)}{d_i}. \tag{A9}$$

A.2.1.1 The standard IBD-to-allele model assumes independence between IBD and allelic states. When IBD and allelic states are independent, $d_i(q) = f_i(q) \cdot d_i$, rendering

$$\mathbb{P}_{\text{corrected}}(A_i^{(k)} = A_i^{(\ell)} = q \mid D_i^{(k,\ell)} = 1) \approx f_i(q). \tag{A10}$$

Otherwise stated, the standard IBD-to-allele model (Equation (A5)) assumes independence between allelic and IBD states. If independence between allelic and IBD states does not hold, then the sample allele frequency $f_i(q)$ is not necessarily a good approximation of the true probability of an IBD pair harboring allele q, and the standard IBD-to-allele model is misspecified.

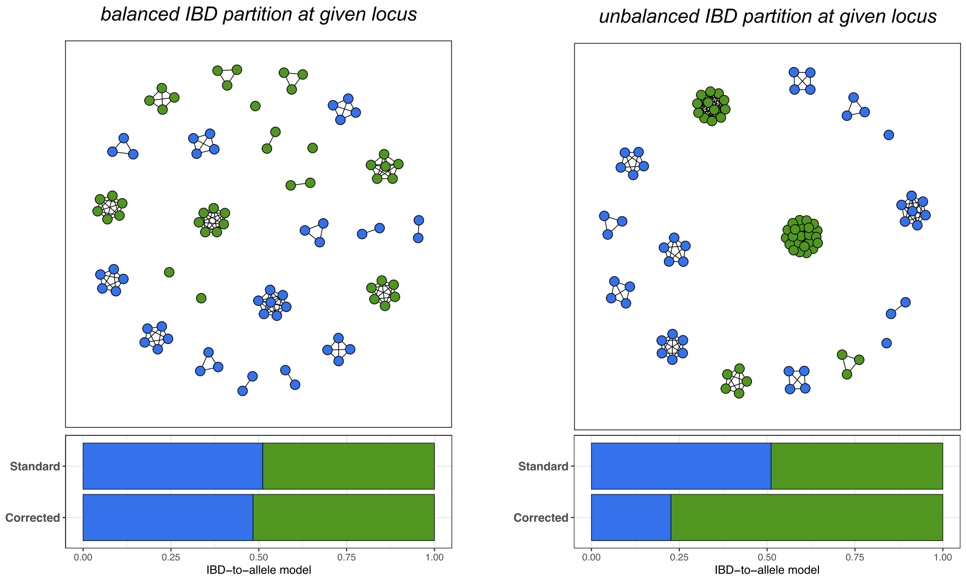


Fig. A2. Schematic of misspecification under the standard IBD-to-allele model, relative to the proposed correction (A9), based on an underlying IBD network for a single locus. Each node represents an individual, colored by allelic state. Individuals that are IBD at that locus are connected by an edge. Since IBD is a transitive property (Taylor, Watson, *et al.* 2019), all components are fully connected. In the absence of genotyping error, edges can only be placed between nodes of the same color (that is, IBD necessarily implies IBS).

The presence of relatedness structure within the sample can violate the assumption of independence between allelic and IBD states. By relatedness structure, we refer to the locuswise partitioning of individuals into IBD clusters (Taylor, Watson, *et al.* 2019). If subcluster sizes are equal, then Equation (A5) will hold. Unbalanced subcluster sizes, however, can lead to substantial divergence between the sample proportion of IBD pairs harboring each allelic state and the likelihood under the standard IBD-to-allele model (Equation (A5)) (e.g. Fig. A2); this may occur if a particular allele is under strong positive selection. In the context of sibship reconstruction, which likewise involves partitioning samples by ancestry, the correction of sample allele frequencies is posited to be particularly pertinent in small populations with variable family/partition sizes (Thomas and Hill 2000; Smith *et al.* 2001); an analogous prescription applies here. Unlike family size, relatedness structure within a parasite population is unobservable and poses significant combinatorial challenges for inference: given dependence between parasite pairs, inference of relatedness, and allele frequencies should be performed jointly over parasite population. Without due consideration of underlying demographic processes and an associated mechanistic model of ancestry, it is difficult to gauge the expected relatedness structure within a population. Diagnosing the degree of misspecification under the model likelihood (Equation (A5)) is therefore nontrivial in practice.

Besides in the unrealistic scenario where the parasite population is split into unrelated clonal clusters, removing replicates of seemingly clonal parasites will not remove replicates within locuswise IBD partitions because individuals can be IBD at a given locus without being IBD at all loci (Fig. A3). Otherwise stated, removing replicates of seemingly clonal parasites from a sample does not remove relatedness between the remaining sample members.

A.2.1.2 The standard nIBD-to-IBS model encodes locuswise average relatedness. The observation model of allelic states for nIBD loci (Equation (A6)) is misspecified in the conflation of IBC and IBS: an unobservable and unaccounted proportion of IBS is driven by IBD. Similarly to Weir and Goudet (2017), we shift to IBS descriptives hereafter to:

1) Circumvent misspecification arising from the assumed independence of allelic and IBD states implicit in Equation (A5).
2) Highlight the confounding effects of relatedness structure in the construction of observation models for nIBD pairs (Equation (A6)).

Condensing Equation (A6) to consider only IBS descriptives yields the following nIBD-to-IBS model:

$$\mathbb{P}_{\text{standard}}(S_i^{(k,\ell)} = 1 \mid D_i^{(k,\ell)} = 0) = \sum_{q=1}^{y} f_i(q)^2 =: s_i. \tag{A11}$$

We note that Equation (A11) is precisely Nei's gene identity metric (Nei 1973; Taylor, Jacob, *et al.* 2019). We can alternatively interpret s_i as an estimate of the probability that two identical alleles are selected from a pool of n alleles $A_i^{(k)}$, $k = 1, \ldots, n$ *with* replacement by rearranging Equation (A11) to obtain the expression

$$\begin{aligned} s_i &:= \sum_{q=1}^{y} \left(\frac{1}{n} \sum_{k=1}^{n} \mathbb{1}\{A_i^{(k)} = q\} \right)^2 \\ &= \frac{1}{n^2} \sum_{k=1}^{n} \sum_{\ell=1}^{n} \left(\sum_{q=1}^{y} \mathbb{1}\{A_i^{(k)} = A_i^{(\ell)} = q\} \right) \\ &= \frac{1}{n^2} \sum_{k=1}^{n} \sum_{\ell=1}^{n} S_i^{(k,\ell)}, \end{aligned} \tag{A12}$$

where we have substituted Equations (A1) and (A2) and interchanged the order of summation and products.

In the absence of genotyping error, such that locus i can only be IBD if it is IBS, that is, $S_i^{(k,\ell)} \geq D_i^{(k,\ell)}$, the sample

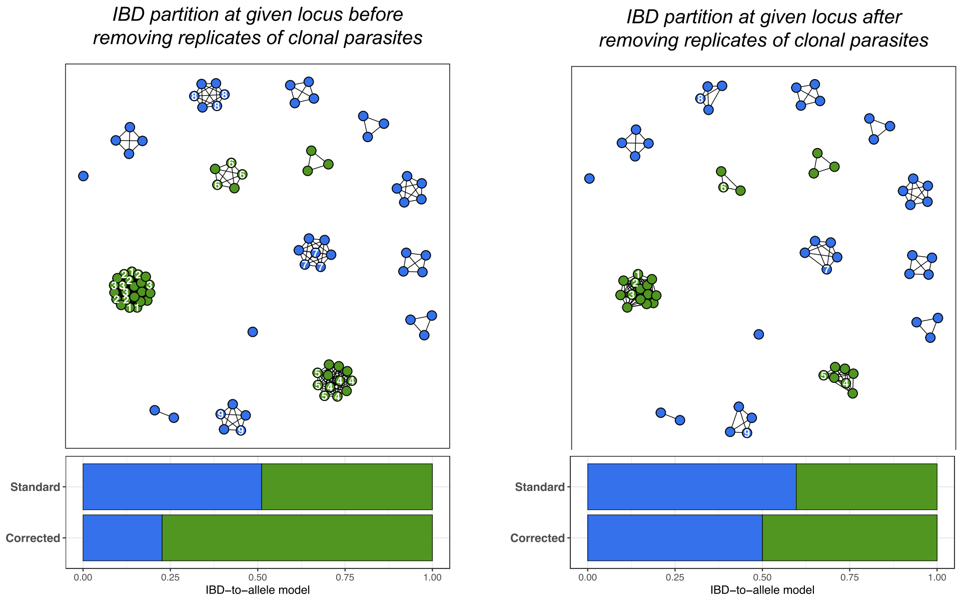


Fig. A3. Unbalanced IBD partitions at a given locus before and after removing replicates of clonal parasites. Groups of clonal parasites are numbered 1–9.

proportion of nIBD individuals that are IBS at locus i can be written

$$c_i := \frac{\sum_{k=1}^{n} \sum_{\ell=1}^{n} [S_i^{(k,\ell)} - D_i^{(k,\ell)}]}{\sum_{k=1}^{n} \sum_{\ell=1}^{n} [1 - D_i^{(k,\ell)}]} = \frac{s_i - d_i}{1 - d_i}, \tag{A13}$$

where we have used Equations (A8) and (A12).

If c_i were observable, the probability of IBS sharing under an nIBD-to-IBS model would be approximated by it, with analogous reasoning to Appendix A.2.1.1: in the infinite sample size limit, that is, as n approaches infinity, c_i would be expected to converge to the true conditional probability of IBS given nIBD (at least in the context of the standard law of large numbers). As such, we purport

$$\mathbb{P}_{\text{corrected}}(S_i^{(k,\ell)} = 1 \mid D_i^{(k,\ell)} = 0) \approx c_i. \tag{A14}$$

This construction allows us to capitulate the misspecification inherent in the standard nIBD-to-IBS model. In particular, rearranging Equation (A13), we see that the sample proportion of IBS pairs s_i—the standard nIBD-to-IBS model likelihood—can be written as a function of the unobservable proportion of IBD pairs d_i and the unobservable sample proportion of nIBD individuals that are IBS c_i as follows:

$$\text{Standard nIBD-to-IBS model} = s_i = c_i + (1 - c_i)d_i \neq c_i. \tag{A15}$$

Otherwise stated, d_i partially encodes relatedness structure in the sample proportion of IBS pairs s_i and thus the standard nIBD-to-IBS model.

Since pairwise comparisons are performed with replacement, i.e. self–self comparisons are included, it is necessarily the case that $d_i \geq 1/n$, whereby $s_i \geq c_i$. As such, the standard nIBD-to-IBS model overestimates the probability of IBS for nIBD pairs at locus i. The degree of overestimation grows linearly as a function of the proportion of pairs d_i that are IBD at locus i—and may therefore be particularly problematic in populations with high levels of relatedness.

Aside: Purcell *et al.* (2007) propose a correction for finite sample bias that is predicated on the probability that two identical alleles are selected from a pool of n alleles *without* replacement, i.e. excluding self–self comparisons. Implementing this correction yields the adjusted quantity

$$\text{Adjusted nIBD-to-IBS model} = s_i' := \frac{1}{\binom{n}{2}} \sum_{k=1}^{n-1} \sum_{\ell=k+1}^{n} S_i^{(k,\ell)} = \frac{s_i - \frac{1}{n}}{1 - \frac{1}{n}}, \tag{A16}$$

where we have substituted Equation (A12). For sufficiently large sample sizes n, both formulations $s_i \approx s_i'$ are approximately equal. This adjustment is sufficient in an outbred setting where all pairs of distinct parasites are unrelated, but falls short for inbred parasite samples.

Aside: While we have adopted the convention of pairwise comparisons with replacement, in line with standard nIBD-to-observation models (Henden *et al.* 2018; Schaffner *et al.* 2018; Taylor, Jacob, *et al.* 2019), this decision does not affect the theoretical correction c_i. That is, if we define

$$d_i' := \frac{1}{\binom{n}{2}} \sum_{k=1}^{n-1} \sum_{\ell=k+1}^{n} D_i^{(k,\ell)}$$

to be the proportion of distinct pairs that are IBD at locus i, then analogous reasoning yields the corrected nIBD-to-IBS model

$$c_i' = \frac{\sum_{k=1}^{n-1} \sum_{\ell=k+1}^{n} [S_i^{(k,\ell)} - D_i^{(k,\ell)}]}{\sum_{k=1}^{n-1} \sum_{\ell=k+1}^{n} [1 - D_i^{(k,\ell)}]} = \frac{s_i' - d_i'}{1 - d_i'}. \tag{A17}$$

Inspection of Equations (A13) and (A17), however, reveals that $c_i = c_i'$.

Aside: As for misspecification given IBD (Appendix A.2.1.1), besides in the unrealistic scenario where the parasite population is split into unrelated clonal clusters—in which case the finite-sample correction (A16) of Purcell *et al.* (2007) is directly applicable—removing replicates of clonal parasite from a parasite sample will not render d_i equal zero, because individuals that are IBD at the ith locus are not necessarily IBD at all loci.

A.2.2 Consequences of the standard nIBD-to-IBS model under marker independence. The misspecification of the standard nIBD-to-IBS model leads to double-counting relatedness in the likelihood of IBS:

$$\mathbb{P}(\text{IBS}) = \mathbb{P}(\text{IBD}) + \underbrace{\overbrace{\mathbb{P}_{\text{standard}}(\text{IBS} \mid \text{nIBD})}^{\substack{\text{obs. model : inflated by} \\ \text{locuswise proportion IBD}}} \cdot [1 - \mathbb{P}(\text{IBD})]}_{\text{likelihood of IBC=IBS} \cap \text{nIBD}} ;$$

intuitively, we expect this to lead to the systematic underestimation of pairwise relatedness (Case A, Box A2). To characterize more thoroughly the consequences of this misspecification, we restrict our attention to the loci-independence model of pairwise relatedness (Taylor, Jacob, *et al.* 2019) with IBS descriptives. For the parasite pair (k, ℓ), the likelihood of observing the sequence of IBS states

$$\mathbf{S}^{(k,\ell)} = (S_1^{(k,\ell)}, \ldots, S_m^{(k,\ell)})$$

across a set of loci $i = 1, \ldots m$, conditional on pairwise relatedness $r^{(k,\ell)}$, is given by

$$\begin{aligned}&\mathbb{P}_{\text{standard}}(\mathbf{S}^{(k,\ell)} \mid r^{(k,\ell)}) \\ &= \prod_{i=1}^{m} (r^{(k,\ell)} + s_i(1 - r^{(k,\ell)}))^{S_i^{(k,\ell)}} (1 - r^{(k,\ell)} - s_i(1 - r^{(k,\ell)}))^{1 - S_i^{(k,\ell)}} .\end{aligned} \tag{A18}$$

The results in the present section bear strong conceptual similarity to the work of Weir and Goudet (2017); however, we consider MLEs whilst Weir and Goudet (2017) addressed method of moments estimators of relatedness.

A.2.2.1 The pairwise relatedness parameter captures deviation from average relatedness. To aid reinterpretation by inspection, we rewrite the likelihood (A18) in the following form:

$$\begin{aligned}\mathbb{P}_{\text{standard}}(\mathbf{S}^{(k,\ell)} \mid r^{(k,\ell)}) &= \prod_{i=1}^{m} [1 - (1 - s_i)(1 - r^{(k,\ell)})]^{S_i^{(k,\ell)}} [(1 - s_i)(1 - r^{(k,\ell)})]^{1 - S_i^{(k,\ell)}}, \\ &= \prod_{i=1}^{m} \{[1 - (1 - c_i)\{1 - d_i - r^{(k,\ell)}(1 - d_i)\}]^{S_i^{(k,\ell)}} \\ &\qquad [(1 - c_i)\{1 - d_i - r^{(k,\ell)}(1 - d_i)\}]^{1 - S_i^{(k,\ell)}}\},\end{aligned} \tag{A19}$$

where $s_i = c_i + (1 - c_i)d_i$ (Equation (A15)) has been substituted into the first line.

Under the corrected nIBD-to-IBS model, c_i (Equation (A13)), the probability of $\mathbf{S}^{(k,\ell)}$ is analogously given by

$$\begin{aligned}\mathbb{P}_{\text{corrected}}(\mathbf{S}^{(k,\ell)} \mid r_{\text{corrected}}^{(k,\ell)}) = \prod_{i=1}^{m} & [1 - (1 - c_i)(1 - r_{\text{corrected}}^{(k,\ell)})]^{S_i^{(k,\ell)}} \\ & \times [(1 - c_i)(1 - r_{\text{corrected}}^{(k,\ell)})]^{1 - S_i^{(k,\ell)}}\end{aligned} . \tag{A20}$$

From inspection of Equations (A19) and (A20), the probability of IBD sharing at locus i under the corrected model can be written in the form

$$\begin{aligned}&\mathbb{P}_{\text{corrected}}(D_i^{(k,\ell)} = 1) \\ &= \underbrace{d_i}_{\substack{\text{average relatedness} \\ \text{in parasite sample}}} + \underbrace{r^{(k,\ell)}(1 - d_i)}_{\substack{\text{deviation from average} \\ \text{relatedness in parasite sample}}} .\end{aligned} \tag{A21}$$

Aside: We do not equate $r_{\text{corrected}}^{(k,\ell)}$ to $\mathbb{P}_{\text{corrected}}(D_i^{(k,\ell)} = 1)$ because d_i are not necessarily the same for all i.

We suggest that Equation (A21) captures the "true" likelihood of IBD sharing at locus i, and can be decomposed into two components:

1) the sample proportion of IBD pairs d_i reflecting the average relatedness at locus i in the parasite sample; and
2) the term $r^{(k,\ell)}(1 - d_i)$ quantifying pairwise IBD sharing at locus i that can not be explained by the average locuswise relatedness in the parasite sample alone.

In practice, inference is performed on the parameter $r^{(k,\ell)}$. Rearranging Equation (A21), we can write

$$r^{(k,\ell)} = \frac{\mathbb{P}_{\text{corrected}}(D_i^{(k,\ell)} = 1) - d_i}{1 - d_i}. \tag{A22}$$

Equation (A22) suggests that the parameter $r^{(k,\ell)}$ is intrinsically adjusted for locuswise average relatedness in the parasite sample, d_i, which varies across loci. In particular, we suggest that $r^{(k,\ell)}$ should be interpreted as a relative measure, quantifying how strongly relatedness between the parasite pair (k, ℓ) deviates from the locuswise average relatedness d_i in the parasite sample. This mirrors the notion of relative relatedness articulated by Weir and Goudet (2017).

A.2.2.2 Relatedness estimates are stratified by average relatedness. In practice, the pairwise relatedness parameter $r^{(k,\ell)}$ is taken to have range [0, 1]. From Equation (A21), however, we see that the constraint $r^{(k,\ell)} \geq 0$ introduces a lower bound

$$\mathbb{P}_{\text{corrected}}(D_i^{(k,\ell)} = 1) \geq d_i$$

on the "corrected" model likelihood of IBD sharing at locus i for the parasite pair (k, ℓ). In particular, setting $r^{(k,\ell)} = 0$ is equivalent to approximating the locuswise probability of pairwise IBD by the locuswise sample average relatedness d_i: plugging $r^{(k,\ell)} = 0$ into Equation (A21) and evoking the assumed independence across markers, we obtain

$$\mathbb{P}_{\text{standard}}(\mathbf{D}^{(k,\ell)} \mid r^{(k,\ell)} = 0) = \prod_{i=1}^{m} d_i^{D_i}(1 - d_i)^{1 - D_i},$$

which can be interpreted as the expected distribution of pairwise relatedness for the parasite genotype sample. This observation echoes Weir and Goudet (2017), who additionally allude to the concept of negative relatedness.

The notion of ostensibly "unrelated" parasite pairs $r^{(k,\ell)} = 0$ warrants attention. To explore this notion, it is instructive to consider an MLE scheme for the parameter $r_{k,\ell}$. From Equation (A19), we verify that the derivative of the log-likelihood for the sequence of observed IBS states $\mathbf{S}^{(k,\ell)}$

$$\frac{d\log \mathbb{P}_{\text{standard}}(\mathbf{S}^{(k,\ell)} \mid r^{(k,\ell)})}{dr^{(k,\ell)}} = \sum_{i=1}^{m}\left\{\frac{(1-s_i)S_i^{(k,\ell)}}{s_i+(1-s_i)r^{(k,\ell)}} - \frac{1-S_i^{(k,\ell)}}{1-r^{(k,\ell)}}\right\} \tag{A23}$$

is a monotonically decreasing function of $r^{(k,\ell)}$. Equation (A23) allows us to identify a sufficient and necessary condition to recover a zero MLE $\hat{r}^{(k,\ell)} = 0$:

$$\left.\frac{d\log \mathbb{P}_{\text{standard}}(\mathbf{S}^{(k,\ell)} \mid r^{(k,\ell)})}{dr^{(k,\ell)}}\right|_{r^{(k,\ell)}=0} \leq 0 \Leftrightarrow \frac{1}{m}\sum_{i=1}^{m}\frac{S_i^{(k,\ell)}}{s_i} \leq 1. \tag{A24}$$

We now examine this threshold in greater detail. Recall from Equation (A15) that

$$s_i = d_i + (1-d_i)c_i,$$

where c_i constitutes the corrected nIBD-to-IBS model. We can therefore interpret s_i as the expected IBS sharing at locus i, predicated on the average relatedness d_i at locus i in the parasite sample. As such, the observable threshold

$$\frac{1}{m}\sum_{i=1}^{m}\frac{S_i^{(k,\ell)}}{s_i} = 1 \tag{A25}$$

can be thought to stratify pairs with below-average vs above-average relatedness. We interpret the LHS of Equation (A25) as a weighted measure of pairwise IBS sharing, with pairwise IBS at loci with a limited sample proportion of IBS pairs (i.e. small s_i) weighted more heavily.

Our interpretation of the MLE $\hat{r}^{(k,\ell)}$, mirroring Hall *et al.* (2012) and Weir and Goudet (2017), is thus two-fold:

- $\hat{r}^{(k,\ell)} = 0$ indicates that the parasite pair (k, ℓ) has below-average relatedness relative to the parasite sample, but is otherwise uninformative (i.e. does not quantify the extent to which pairwise relatedness is below the average relatedness in the parasite sample). Permitting negative relatedness estimates, which arise naturally using moment estimators, may be informative for parasite pairs with below-average relatedness relative to the parasite sample (Hall *et al.* 2012).
- $\hat{r}^{(k,\ell)} > 0$ quantifies the extent to which relatedness between the parasite pair (k, ℓ) exceeds the average relatedness in the parasite sample, or the amount of relatedness between the parasite pair (k, ℓ) that cannot be explained by average relatedness alone.

The adjustment for "average relatedness" in the parasite sample is predicated on the locuswise sample proportion of IBD pairs d_i. If d_i is taken to be constant across all loci $i = 1, \ldots, m$, then from Equations (A20) and (A22), a *nonzero* MLE $\hat{r}^{(k,\ell)}$ can be written in the form

$$\hat{r}^{(k,\ell)} = \frac{\hat{r}^{(k,\ell)}_{\text{corrected}} - d}{1-d},$$

where $\hat{r}^{(k,\ell)}_{\text{corrected}}$ is the MLE obtained under the corrected nIBD-to-IBS model c_i and the index i is dropped on account of all d_i being equal. A schematic of the relationship between $\hat{r}^{(k,\ell)}_{\text{corrected}}$ and $\hat{r}^{(k,\ell)}$ is shown in Fig. A4. Accounting for the variability of d_i across loci, we predict, would yield a fuzzy elbow-like characteristic, with a change point near the average sample relatedness

$$\bar{d} = \frac{1}{m}\sum_{i=1}^{m} d_i.$$

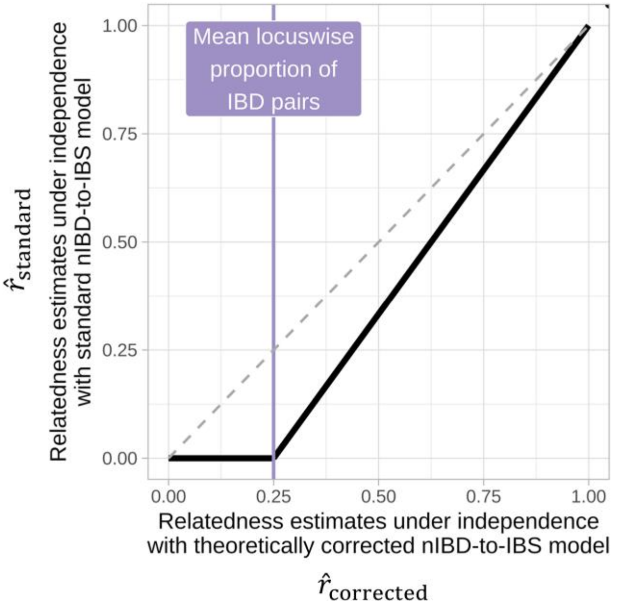


Fig. A4. Schematic of MLEs of pairwise relatedness under the model of (n)IBD independence coupled with the standard (misspecified) model $\hat{r}_{\text{standard}}$ vs corrected nIBD-to-IBS model $\hat{r}_{\text{corrected}}$.

For a sample of individuals with a unimodal positively skewed distribution of pairwise relatedness, we would expect the majority of parasite pairs to exhibit "below-average" relatedness, manifest in pronounced zero inflation of the pairwise MLEs $\hat{r}^{(k,\ell)}$.

A.3 *Numerical results*

Here, we present various numerical results that complement our theoretical findings (Appendices A.3.1 and A.3.2). We additionally show that exploiting linkage structure within dense data using the HMM of relatedness mitigates underestimation driven by standard observation models (Appendix A.3.3). Zero-inflation in relatedness estimates under (n)IBD independence, arising from an over-reliance on the underlying (n)IBD-to-observational model, can be explained through comparative plots of IBS sharing (Appendix A.3.4). Finally, we propose a diagnostic to gauge the average locuswise relatedness $\bar{d}$ in a sample and approximate the severity of underestimation (Appendix A.3.5).

A.3.1 Relatedness is systematically underestimated using standard models. Relatedness is systematically underestimated using standard models (Fig. A5). This holds whether observations are IBS descriptives (plots C1, D1) or alleles (plots C2–C3, D2–D3), whether the standard model assumes (n)IBD independence (plots C1–C2, D1–D2) or not (plots C3, D3), and whether markers are dense or sparse (plots C and D, respectively). Estimates are severely zero-inflated when data are sparse (plots B2–B4) and when dense data are fit under (n)IBD independence (plots A2–A3). Less severe zero-inflation and underestimation when the HMM is fit to dense data (plots A4 and C3, respectively) is addressed in

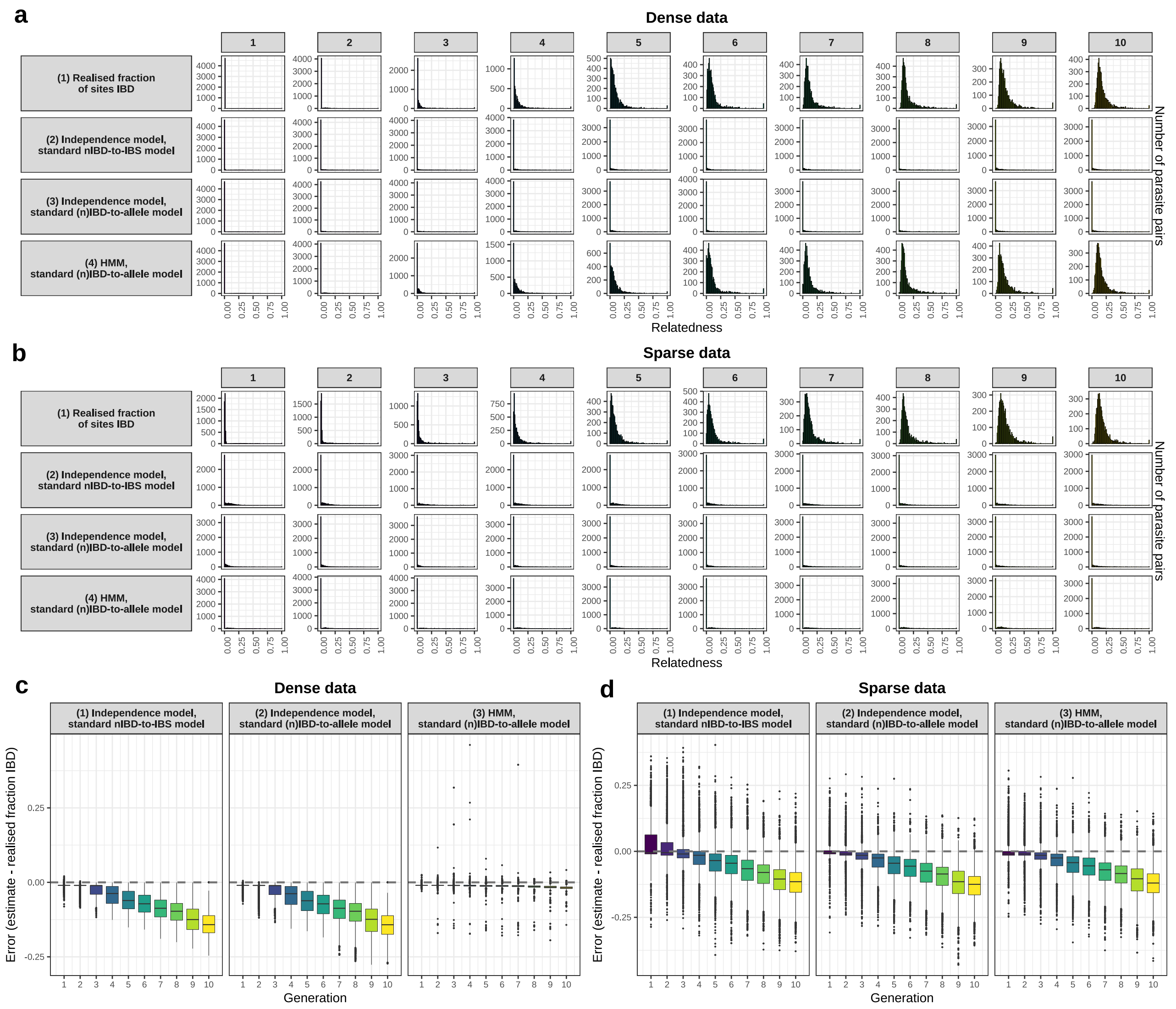


Fig. A5. Pairwise relatedness estimates predicated on standard sample allele frequency-based observation models and either allelic states or IBS descriptives; either the independence model of relatedness or the HMM; and either dense or sparse data. Columns in subplots a) and b) correspond to the number of generations of inbreeding (1–10). Error, where applicable, is measured relative to the realized fraction of polymorphic sites IBD for each pairwise comparison. Sparse data have been obtained by selecting 200 polymorphic markers uniformly at random without replacement from the dense simulated dataset. Results are based a single realization of the simulation model, with parameter values as per Table B1.

Appendices A.3.3 and A.3.4. Zero inflation renders mean($\hat{r}$) under (n)IBD independence a poor approximation of $\bar{d}$.

A.3.2 Underestimation using standard models is due to the partial encoding of population relatedness in sample allele frequencies.

The degree of underestimation of pairwise relatedness using standard models increases with the number of generations of inbreeding because the standard nIBD-to-nIBS model is increasingly misspecified, in that the sums of squares of sample allele frequencies encode more relatedness (Fig. A6). The unbiased nature of relatedness estimates generated under the corrected independence model fit to (n)IBS observations (Fig. A7) also supports, by contrast, the idea that underestimation is due to the use of sample allele frequencies within standard observation models. This is true of both dense and sparse data, although estimates based on sparse data are less precise.

A.3.3 Relatedness is less severely underestimated using the standard HMM fit to dense data.

Consider an inbred population with elevated relatedness (Fig. A8a). Increasing the marker count while accounting for linkage (estimation under the HMM with the standard (n)IBD-to-allele model) generates more precise and less biased estimates (Fig. A8b). Meanwhile, increasing the marker count without accounting for linkage (estimation under (n)IBD independence with either the standard nIBD-to-IBS or (n)IBD-to-allele model) does not remove bias (Fig. A8c, d). We believe this is because estimation under the HMM exploits linkage information, which increases with marker density. Meanwhile, estimation under the independence model does not exploit linkage information, regardless of its extent in the data. Instead, the independence model is entirely reliant on sample allele frequencies, which render the observational model misspecified. Otherwise stated, the standard HMM is less reliant upon sample allele frequencies and thus less susceptible to the misspecification they cause.

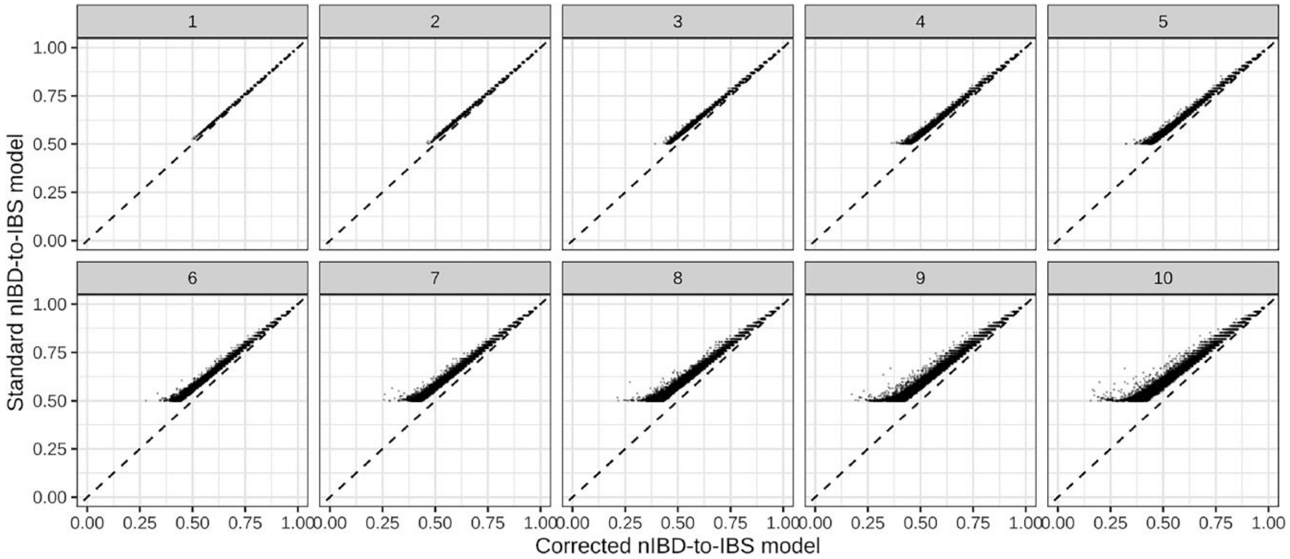


Fig. A6. The locuswise proportion of IBS pairs s_i (i.e. the standard nIBD-to-IBS model) vs the sample proportion of nIBD pairs that are IBS c_i (i.e. the corrected nIBD-to-IBS model) across 10 successive generations of inbreeding (plots 1–10) for a single realization of the simulation model, with parameter values as per Table B1. For biallelic markers, it is necessarily the case that $s_i \geq 0.5$; however, c_i may lie anywhere in the range [0, 1].

Aside: Under the HMM, the extent to which pairwise relatedness is underestimated decreases over the zero to one range granted data are sufficiently dense to encode linkage structure (Fig. A9). This is because, for a given marker count, the extent of linkage depends on the length of shared IBD segments, which is greater for recent relatives, and thus correlated with relatedness. We mention this as an aside only, because linkage due to recent ancestry is an inherent property of a given parasite genotype pair, not something we can control.

A.3.4 Comparative plots of IBS distributions capture the extent of zero-inflation under the independence model. The persistence of zero inflation in dense data relatedness estimates under the independence model can be explained through comparative plots of IBS distributions.

For a given pair of individuals, the fraction of polymorphic sites that are IBS can be computed. The computation can then be repeated for all possible pairs. Hereafter, we refer to this set of values as the empirical IBS distribution for all pairs. As a

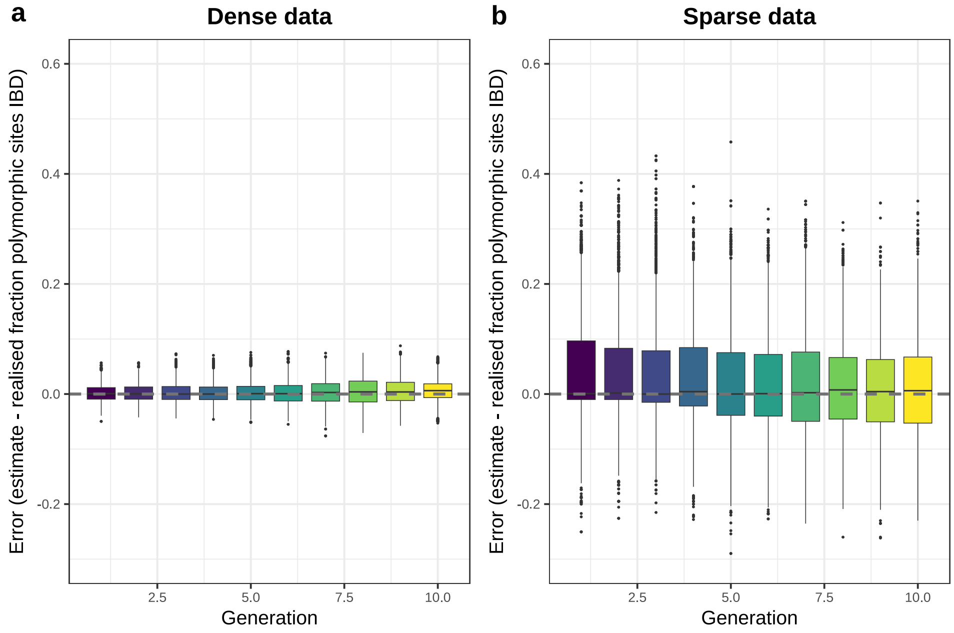


Fig. A7. Relatedness estimates predicated on the independence model and the corrected locuswise nIBD-to-IBS model c_i. Sparse data have been obtained by selecting 200 polymorphic markers uniformly at random without replacement from the dense simulated dataset. Results are based a single realization of the simulation model, with parameter values as per Table B1.

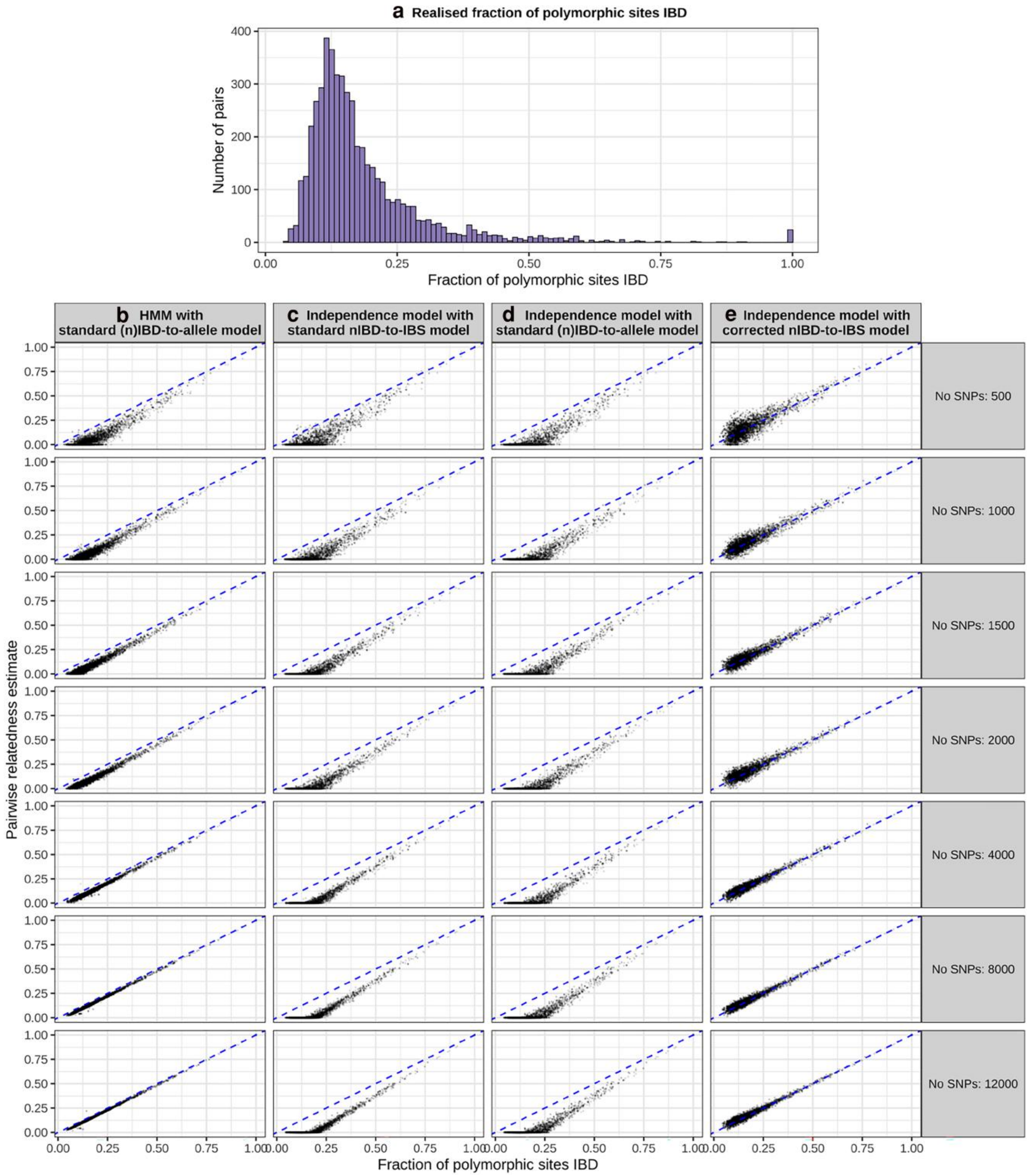


Fig. A8. Summary of pairwise relatedness for a single realization of the simulation model after 10 generations of inbreeding (with parameter values as per Table B1), yielding the distribution of realized relatedness (i.e. the fraction of simulated polymorphic sites that are IBD for each parasite pair) shown in a). For downsampled marker subsets (selected uniformly at random without replacement over the set of polymorphic markers), realized relatedness is calculated over the marker subset and treated as the ground truth. We compare realized relatedness against pairwise relatedness estimates generated under b) the HMM with the standard (n)IBD-to-allele model; c) (n)IBD independence with the standard (n)IBD-to-allele model; d) (n)IBD independence with the standard nIBD-to-IBS model; e) (n)IBD independence with the theoretically corrected nIBD-to-IBS model.

comparator, under the standard nIBD-to-IBS model, i.e. $\mathbf{s} = (s_1, \ldots, s_{n_{\text{markers}}})$, the expected IBS distribution for ostensibly unrelated parasite pairs ($r = 0$) comprises a rescaled Poisson binomial distribution (i.e. a sum of independent, but not necessarily identically distributed Bernoulli random variables):

$$Y_0(\mathbf{s}) = \frac{1}{n_{\text{markers}}} \sum_{i=1}^{n_{\text{markers}}} X_i(s_i)$$

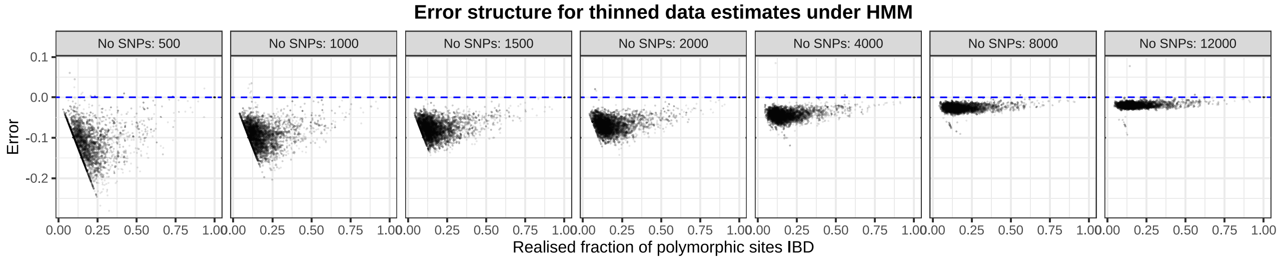


Fig. A9. Error structure for relatedness estimates of downsampled marker subsets (selected uniformly at random without replacement over the set of polymorphic markers) at generation 10 under the HMM of relatedness with the standard (n)IBD-to-allele model, compared with the realized fraction of polymorphic sites IBD for each parasite pair. Results are based on a single realization of the simulation model, with parameter values as per Table B1.

where

$$X_i(s_i) \overset{\text{independent}}{\sim} \text{Bernoulli}(s_i).$$

$Y_0(\boldsymbol{s})$ is also applicable under the HMM, because the HMM collapses down to the independence model for unrelated pairs.

Comparison of the empirical IBS distribution for all pairs and the expected IBS distribution for ostensibly unrelated pairs (Fig. A10a) shows that there is a range of empirical IBS values less than the expected IBS distribution for ostensibly unrelated pairs. The majority of these pairs have zero-valued relatedness estimates under the independence model, because the independence model is heavily reliant on the standard nIBD-to-IBS model (Fig. A10b, orange). Otherwise stated, the extent to which empirical IBS values fall below the expected IBS distribution for unrelated pairs corresponds to the extent of zero-inflation under independence. This result is consistent with the notion that locuswise average relatedness is partially encoded within the locuswise sample proportion of IBS pairs s_i, whereby the expected IBS distribution for ostensibly unrelated pairs ought to be reinterpreted as the expected IBS distribution for pairs with "average" relatedness. Consequently, a range of below-average relatedness values map onto zero-valued relatedness estimates under the independence model. Meanwhile, they have small but nonzero estimates under the HMM model, because the HMM is less reliant on the underlying (n)IBD-to-observation model (Fig. A10b, green).

A.3.5 Dense-dense data plots approximately capture the extent of underestimation. We leverage the precision of estimates generated using dense data to design a practical diagnostic for identifying approximately the corrective value $\bar{d}$. More specifically, although relatedness estimates generated under the HMM using dense marker data from inbred populations are underestimates (Figs. A8 and A9), they are sufficiently less biased downwards compared with those generated under (n)IBD independence to reveal an elbow-like pattern (Fig. A11b), which is otherwise only accessible using either simulation (Fig. A11a) or theory (Fig. A4).

A.4 Empirical results

We devote Appendix A.4 to the empirical effect of population structure, under which our theoretical and numerical results do not necessarily hold. To do so, we pool the $n = 278$ isolates from Guyana (Vanhove *et al.* 2024), analyzed in the main text, with an additional $n = 28$ isolates from Colombia (Carrasquilla *et al.* 2022), yielding $n = 30,694$ biallelic SNPs that are polymorphic among the pooled isolates. Population structure appears pronounced: principal coordinates analysis (PCoA), based on pairwise fractions of IBS markers, yields two distinct clusters, stratified by country (Fig. A12).

A.4.1 Additional manifestations of population structure. We examine two additional plots that indicate population structure whilst linking directly with our conceptual framework. First we examine the empirical distribution of pairwise fractions of IBS markers. In the absence of population structure, we would expect to see a unimodal IBS distribution. The multimodal IBS distribution shown in Fig. A13, in contrast, is indicative of population

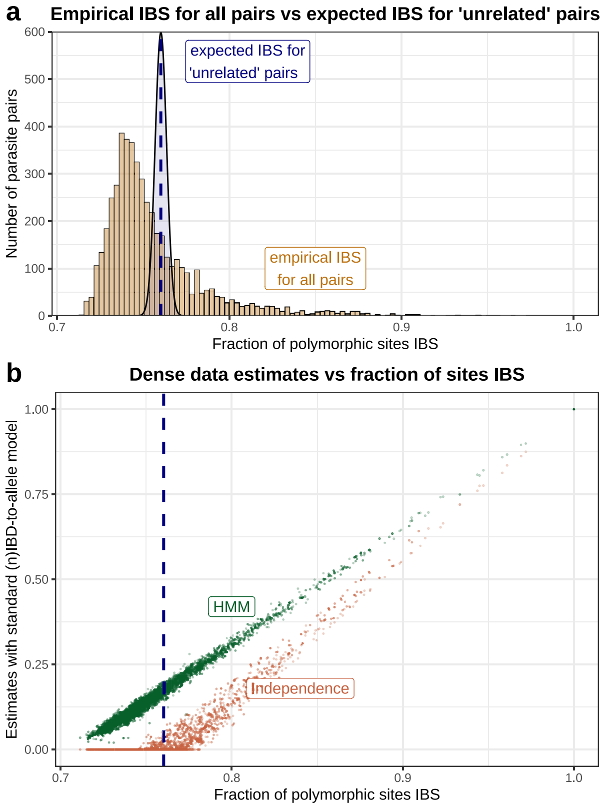


Fig. A10. For dense data at generation 10: a) The empirical pairwise distribution of IBS sharing for all pairs (ochre), defined to be the fraction of polymorphic sites that are IBS, compared with the expected IBS distribution Y_0 for ostensibly unrelated pairs (blue). b) Relatedness estimates generated under the HMM (green) or independence model (orange) using the standard (n)IBD-to-allele model vs the fraction of polymorphic sites IBS. Results are based on a single realization of the simulation model, with parameter values as per Table B1. The probability mass function for the Poisson binomial distribution has been computed using the R package `poisbinom` (Olivella and Shiraito 2017).

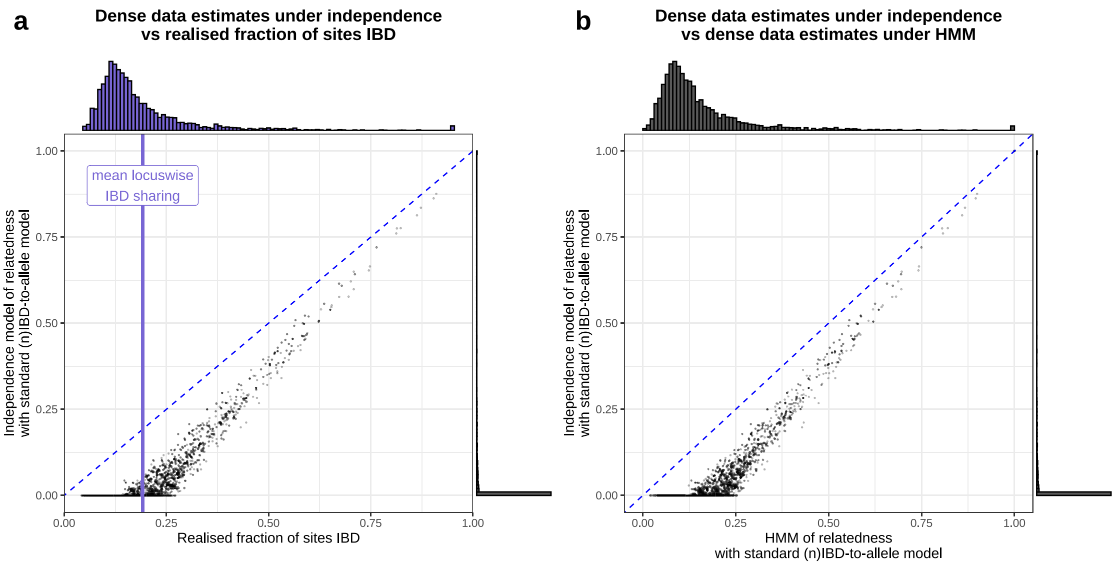


Fig. A11. Relatedness estimates for dense genotypic data at generation 10 under the independence model vs a) the realized fraction of polymorphic sites IBD or b) corresponding estimates under the HMM of relatedness. Here, we use the standard (n)IBD-to-allele model. Results are based on a single realization of the simulation model, with parameter values as per Table B1.

structure (Plucinski and Barratt 2021): we view it as a mixture distribution, with components stratified by within- and between-subpopulation comparisons. The well-separated components suggest pronounced differentiation between allele frequencies in the constituent subpopulations.

An alternative approach, linking back to our characterization of observation models predicated on allelic states vs IBS descriptives in Appendix A.2.1, is the comparison of relatedness estimates under (n)IBD independence predicated on the standard (n)IBD-to-allele model vs the standard nIBD-to-IBS model (Fig. A14). Allelic states are more informative than IBS descriptives. Under the (n)IBD-to-allele model, a shared minor allele points towards IBD more strongly than a shared major allele; this relative weighting is erased when we shift to IBS descriptives. In the absence of population structure, we would expect each pair of individuals to share a mixture of major and minor alleles; a shift from allelic states to IBS descriptives would therefore introduce noise, but we would expect the over/under-weighting of shared major/minor alleles to average out across markers. We would, however, expect systematic differences to emerge in the presence of population structure. Comparisons within the major subpopulation, for instance, systematically yield parasite pairs with a shared major allele, yielding higher relatedness estimates based on IBS descriptives relative to allelic states; the converse applies to the minor subpopulation. Systematic patterns akin to those in Fig. A14 are thus indicative of population structure.

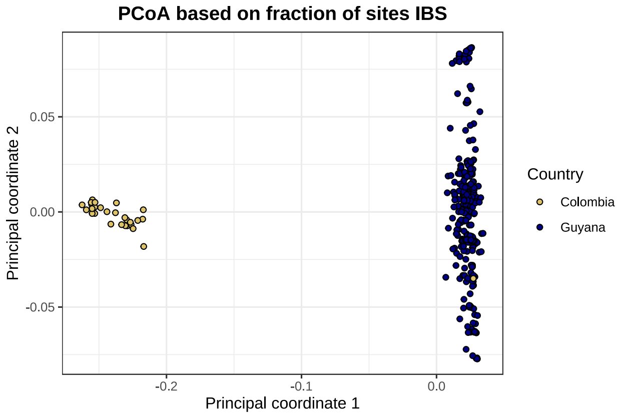


Fig. A12. Principal coordinates analysis, with a distance matrix based on pairwise fractions of IBS markers. PCoA has been performed using the R function `stats::cmdscale` (R Core Team 2021).

A.4.2 Dense data diagnostic in the presence of population structure. We now examine the dense data diagnostic proposed in Appendix A.3.5 in the presence of population structure. In the absence of population structure (main text), we observe a clean elbow-like pattern concordant with our theoretical predictions (Fig. A4) and numerical results (Fig. A11). Population structure, however, yields the emergence of multiple elbows, corresponding to different within- and between-population comparisons, with a different structure depending on whether estimates under (n)IBD independence are predicated on allelic states or IBS descriptives (Fig. A15). Because sample allele frequencies represent a weighted average across subpopulations, interpretation of the branch points is unclear. The apparent absence of bias using allelic states for the inbred minor subpopulation (Colombia) is unclear given this weighted average of sample allele frequencies. However, comparison of estimates using allelic states vs IBS descriptives under (n)IBD independence (Fig. A14) suggests that nonrepresentative sample allele frequencies may introduce an upward bias in relatedness estimates in certain contexts: alleles that are common in the minor subpopulation, but uncommon in the major subpopulation, may inflate relatedness estimates within the minor subpopulation.

A.4.3 Zero inflation and population structure. The standard nIBD-to-IBS model leads to the systematic over/under-weighting of shared major/minor alleles relative to the standard (n)IBD-to-allele model. When sample allele frequencies are averaged over several subpopulations and estimates are generated under (n)IBD independence, allelic states yield zero inflation in comparisons within the major subpopulation; while IBS

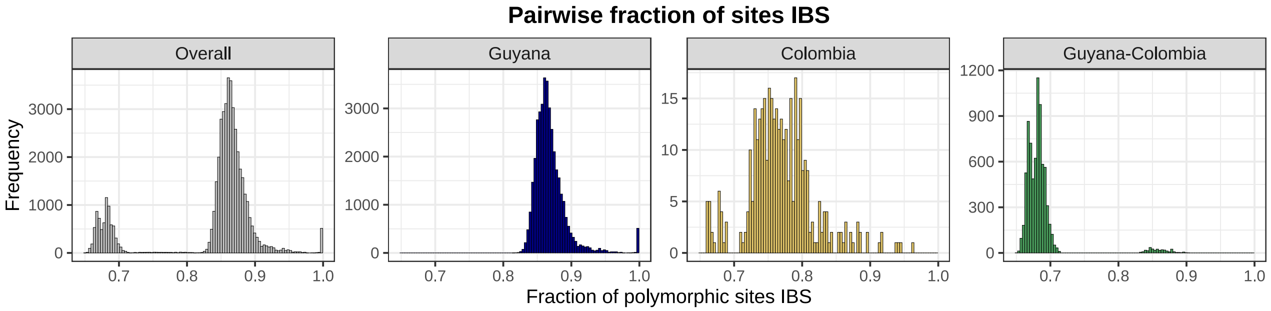


Fig. A13. Multimodality in distributions of pairwise fractions of IBS markers. In the light of missing data, IBS sharing for each pair is defined to be the fraction of polymorphic sites that are not missing data and are IBS. We stratify comparisons within and between countries.

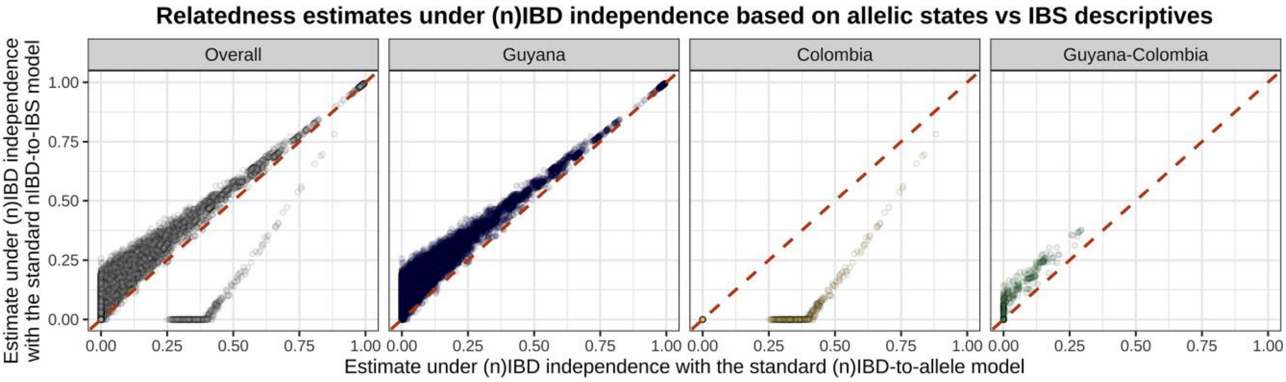


Fig. A14. Comparison of relatedness estimates generated under (n)IBD independence using the standard nIBD-to-IBS model vs the standard (n)IBD-to-allele model as a manifestation of population structure. We stratify comparisons within and between countries.

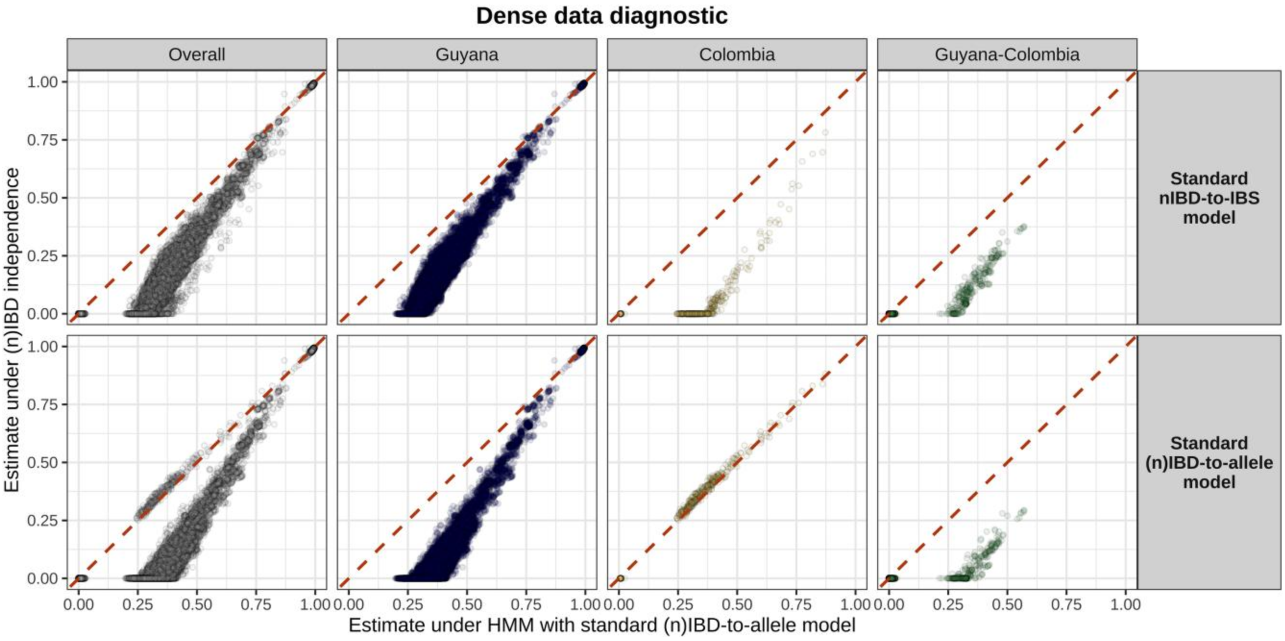


Fig. A15. Dense data diagnostic, comparing relatedness estimates under the HMM and (n)IBD independence, using the standard nIBD-to-IBS or (n)IBD-to-allele model. We stratify comparisons within and between countries.

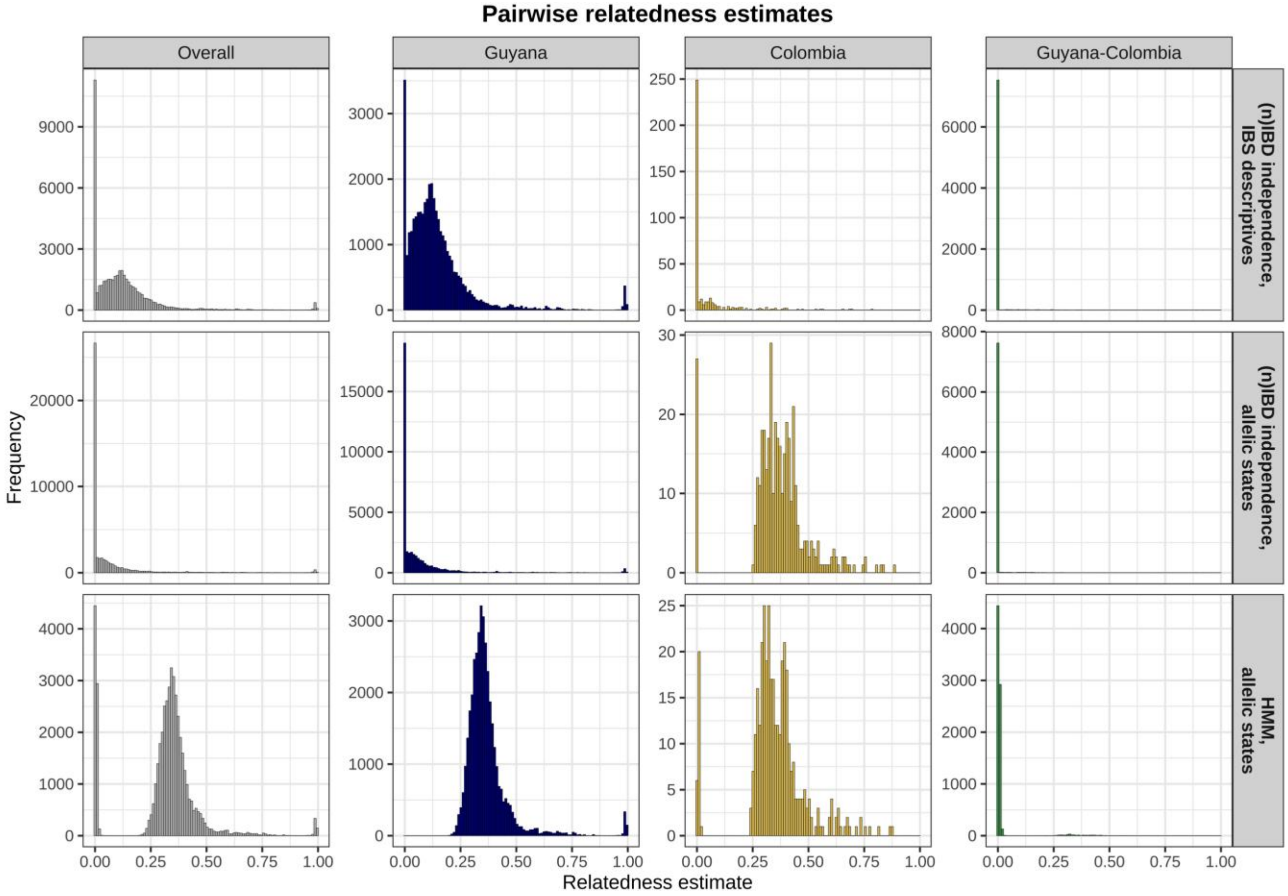


Fig. A16. Pairwise relatedness estimated under (n)IBD independence with the standard nIBD-to-IBS model; (n)IBD independence with the standard (n)IBD-to-allele model; and the HMM with the standard (n)IBD-to-allele model. We stratify comparisons within and between countries.

descriptives yield zero inflation in comparisons within the minor subpopulation (Fig. A16). Theoretically predicted and numerically validated results pertaining to zero inflation, therefore, are sensitive to population structure. Empirically, variable zero inflation under (n)IBD independence and the nIBD-to-IBS model for within- and between-population comparisons can be directly explained by the multimodal distribution of pairwise IBS sharing (Fig. A13) but does not have a direct interpretation.

Appendix B: Simulation model

We describe a simulation model that starts with a population whose ancestry is defined in terms of a bygone outbred founder population and then simulates successive generations of inbreeding. Our model differs from the canonical Wright–Fisher model of stochastic drift, which concerns a single locus (Etheridge 2011), on three grounds:

- We simulate multiple loci and introduce linkage between them through a discrete-time homogeneous Markov chain governing crosses.
- We introduce clonal and sibling substructure through relationship graphs (Taylor, Watson, *et al.* 2019).
- On top of the sibling and clonal substructure, we introduce an increased propensity for selfing and sibling–sibling mating through a single parameter; in doing so, we capture departure from panmixia. In reality, a single parameter cannot represent multiple processes that govern selfing and sibling–sibling mating between malaria parasites: selfing between malaria parasites is always viable theoretically; given a monoclonal mosquito infection, it is inevitable. Mosquito-to-human-to-mosquito cotransmission enhances the probability of sibling–sibling mating (Camponovo *et al.* 2022).

Demographic processes like immigration, mutation, and selection are not accounted for.

Our description is structured as follows. In Appendix B.1, we introduce a model under which offspring are generated as a mosaic of two parents, allowing for linkage between markers which are all equidistant. We then propose a model for our "generation zero" population, whose ancestry is formulated in terms of a bygone outbred founder population (Appendix B.2). In Appendix B.3, we propose a model under which the population in the present generation guides the generation of the population in the next generation. Designed to capture a single generation of inbreeding, our framework is underpinned by the relationship graph model of Taylor, Watson, *et al.* (2019). In Appendix B.4, we detail our approach for assessing misspecification under the standard nIBD-to-IBS model. A summary of simulation parameters, and their respective tradeoffs, is provided in Appendix B.5.

B.1 Generating a single individual with inter-marker linkage

We consider some number, n_{markers}, of multiallelic, *equidistant* markers, spanning a single chromosome, and indexed $i = 1, \ldots, n_{\mathrm{markers}}$, with each marker treated as a point polymorphism (Taylor, Jacob, *et al.* 2019). Each time we simulate a single meiosis, we generate a single offspring as a random n_{markers}-mosaic of parents that are labeled $\{a_1, a_2\}$, respectively. In reality, a single meiosis between two parental genotypes generates four meiotic offspring, which then replicate asexually into thousands of sporozoites; on average, meiotic offspring are related to each other by 1/3 and to each parent by 1/2 if parental genotypes are unrelated (Wong *et al.* 2018; Taylor, Watson, *et al.* 2019).

Denote by

$$\boldsymbol{b} = (b_1, \ldots, b_{n_{\text{markers}}})$$

the mosaic of an offspring of parents $\{a_1, a_2\}$, with $b_i = a_1$ if marker i is inherited from parent a_1 and $b_i = a_2$ otherwise.

Assuming a genomewide constant recombination rate of ρ, $\boldsymbol{b}$ is governed by a discrete-time, homogeneous Markov chain with transition matrix

$$\begin{pmatrix} \mathbb{P}(b_{i+1} = a_1 \mid b_i = a_1) & \mathbb{P}(b_{i+1} = a_1 \mid b_i = a_2) \\ \mathbb{P}(b_{i+1} = a_2 \mid b_i = a_1) & \mathbb{P}(b_{i+1} = a_2 \mid b_i = a_2) \end{pmatrix}$$
$$= \begin{pmatrix} 0.5(1 + e^{-\rho d}) & 0.5(1 - e^{-\rho d}) \\ 0.5(1 - e^{-\rho d}) & 0.5(1 + e^{-\rho d}) \end{pmatrix},$$

equivalent to the Markov model of relatedness (Equation (A3)) in the case $r = 0.5$ with switch rate $\kappa = 1$ where d is measured in base pairs (bp) and the recombination rate ρ has units M/bp.

Due to discrete sampling of the genome, consecutive markers inherited from the same parent may be separated by an even number of recombination breakpoints, which occur between base pairs. For the purposes of simulating ancestry at a fixed set of markers, we do not distinguish whether or not a string of markers inherited from the same parent is separated by recombination breakpoints (Fig. B1), and assume that a Poisson process governs the position of recombination breakpoints (Speed 1997).

Observe that

$$\mathbb{P}(b_{i+1} = b_{i+2} = \ldots b_{i+m-1} = a_1, b_{i+m} = a_2 \mid b_i = a_1)$$
$$= \frac{1}{2^m}(1 - e^{-\rho d})(1 + e^{-\rho d})^{m-1} \quad .$$

If we ignore (temporarily) the finite number of markers n_{markers}, then given $b_i = a_1$, the number of consecutive markers (from i inclusive) inherited from parent a_1 is geometrically distributed with state space $\mathbb{N}$ and mean length

$$\bar{M} := \frac{2}{1 - e^{-\rho d}}. \tag{B1}$$

Stretches of consecutive markers inherited from parent a_2 are identically distributed in length.

We can therefore treat $\boldsymbol{b}$ as a truncated alternating renewal process, where segments alternate from each parent and are independent and identically distributed (i.i.d.) with length

$$L \sim \text{Geometric}\left(\frac{1}{\bar{M}}\right),$$

where the geometric distribution is taken to have support $\mathbb{N}$. The case $\bar{M} = 2$ reduces to the independence model.

To simulate an offspring under this model, we sample a parent a_1 or a_2 with probability 1/2; sample a geometric segment length; select the alternative parent and re-sample a geometric segment length; string together alternating geometric segments from each parent; and then truncate once a length of n_{markers} has been obtained, to account for the finite length that is ignored above.

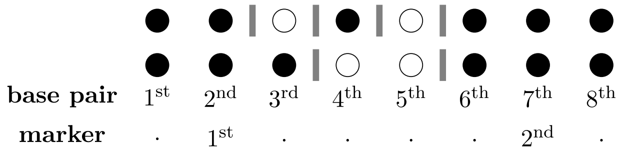


Fig. B1. Two examples of pairwise (n)IBD states across a stretch of 8 bp. Recombination breakpoints (vertical bars) are shown between IBD (black) and nIBD (white) loci. Consecutive IBD markers may be separated by an even number of recombination breakpoints, that may not be detectable due to the discrete sampling of the genome.

B.2 Generating a population of individuals at generation zero

We now propose a model under which to generate a "generation zero" population whose ancestry is constructed in terms of a bygone outbred founder population. This construction allows us to distribute background relatedness across the generation zero population.

Firstly, let us consider a founder population comprising some number of unrelated founders, n_{founders}. At each marker $i = 1, \ldots, n_{\text{markers}}$, we select the allele $q \in \{1, \ldots, y\}$ with probability $p_i(q)$

$$A_{\text{founder}}(f, i) \overset{\text{i.i.d.}}{\sim} \text{Categorical}(p_i(1), \ldots, p_i(y))$$

independently for each founder $f = 1, \ldots, n_{\text{founders}}$. Here, multiallelic markers are treated as point polymorphisms (Taylor, Jacob, *et al.* 2019). For large n_{markers}, each founder is likely to harbor a unique allelic sequence.

We represent individuals within subsequent generations as mosaics of the n_{founders} founders; that is the ancestry of individual j in generation k is represented in the form

$$\mathbf{F}(j, k) = (f_1, \ldots, f_{n_{\text{markers}}}) \in \{1, \ldots, n_{\text{founders}}\}^{n_{\text{markers}}}.$$

Barring genotyping error and mutations, the vector of allelic states observed for individual j in generation k is then given by

$$\mathbf{A}(j, k) = (A_{\text{founder}}(f_1, 1) \ldots, A_{\text{founder}}(f_{n_{\text{markers}}}, n_{\text{markers}})).$$

The ancestry of each individual j in generation zero is independently sampled uniformly at random from the set of all founder sequences, that is,

$$\mathbf{F}(j, 0) \overset{\text{i.i.d.}}{\sim} \text{Uniform}[\{1, , \ldots, n_{\text{founders}}\}^{n_{\text{markers}}}], \tag{B2}$$

and we initialize a population relationship graph for generation zero by placing stranger edges between all individuals/nodes.

Alleles that are derived from the same founder are designated IBD. Alleles that are IBS but not derived from the same founder are designated IBC. Given parameter values listed in Table B1, the vast majority of IBS in generation zero is due to IBC and not IBD. We call this low-level IBD in generation zero background relatedness.

Under this model, low-level background relatedness is maximally spread across parasite pairs in generation zero. Implicit in this construction is the assumption that there is sufficient temporal separation between the founder population and generation zero to break down inter-marker linkage. If we instead modeled generation zero as one generation of mating from the founder population, relatedness would be concentrated on a small subset of pairs with shared parents, consistent with recent rather than background relatedness (Appendix B.5.1), and Equation (B2) would not hold because each individual in generation zero would be constructed as a mosaic of at most two founders.

Aside: The uniform assumption (Equation (B2)) is also compatible with an (n)IBD-to-allele model predicated on founder allele frequencies. Because Equation (B2) holds, the probability of randomly sampling the allelic sequence

$$\mathbf{A} = (q_1, \ldots, q_{n_{\text{markers}}})$$

for each individual in generation zero is proportional to the number of ways of recovering that sequence as a mosaic of founders:

$$\mathbb{P}(\mathbf{A}) = \frac{1}{(n_{\text{founders}})^{n_{\text{markers}}}} \sum_{f_1=1}^{n_{\text{founders}}} \cdots \sum_{f_{n_{\text{markers}}}=1}^{n_{\text{founders}}} \prod_{i=1}^{n_{\text{markers}}} \mathbb{1}\{A_{\text{founder}}(f_i, i) = q_i\}. \tag{B3}$$

Interchanging the summation and product in Equation (B3), we can equivalently write

$$\begin{aligned} \mathbb{P}(\mathbf{A}) &= \frac{1}{(n_{\text{founders}})^{n_{\text{markers}}}} \prod_{i=1}^{n_{\text{markers}}} \sum_{f=1}^{n_{\text{founders}}} \mathbb{1}\{A_{\text{founder}}(f, i) = q_i\} \\ &= \prod_{i=1}^{n_{\text{markers}}} \theta_{\text{founder}}(q, i), \end{aligned} \tag{B4}$$

since the frequency of allele q at locus i in the founder population is

$$\theta_{\text{founder}}(q, i) := \frac{1}{n_{\text{founders}}} \sum_{f=1}^{n_{\text{founders}}} \mathbb{1}\{A_{\text{founder}}(f, i) = q\}. \tag{B5}$$

Equation (B4) is precisely the likelihood of observing the allelic sequence $\mathbf{A}$ that we would obtain by plugging founder allele frequencies (B5) into the standard nIBD-to-allele model under locus independence. In other words, our construction gives rise to an allelic distribution in generation zero that is identical to that which we would obtain by directly drawing alleles with founder allele frequencies, but with the added advantaging of enabling us to explicitly define low-level background relatedness across parasite pairs in generation zero.

B.3 Generating populations of individuals over successive generations of inbreeding

From generation $k = 1$ onwards, we simulate successive discrete, nonoverlapping generations of inbreeding. We capture stochastic drift, with the imposition of additional sibling/clonal substructure in the line with the formulation of Taylor, Watson, *et al.* (2019); but ignore the phenomena of immigration, mutation, and selection.

We assume that the population size $n_{\text{individuals}}$ remains fixed across generations. As stated above, the population relationship graph for generation zero has stranger edges between all individuals. To simulate the ancestry of generation $k \geq 1$, given that of generation $(k - 1)$, we perform the following steps:

1) Simulate a population-level relationship graph for generation k, characterized by clonal, sibling, and stranger edges (Appendix B.3.1).
2) Conditional on the structure of the relationship graphs for generations $(k - 1)$ and k, formulate each individual in generation k as a mosaic of parents from generation $(k - 1)$, allowing for an enriched probability of sibling–sibling crosses and selfing over generations (Appendix B.3.1.1).
3) Recover the ancestry structure (relative to founders $f = 1, \ldots, n_{\text{founders}}$) for each individual in generation k using the encoding $\mathbf{F}(j, k - 1)$.

The simulation structure is informed by the work of Taylor, Watson, *et al.* (2019), and draws on the R package `Pv3Rs` (Taylor 2022b). Below, we describe the first two steps in detail.

B.3.1 Generating relationship graphs. For each generation $k \geq 1$, we simulate a transitive relationship graph with sibling, clonal, and stranger edges, following the framework of Taylor, Watson, *et al.* (2019). Relationship graphs, which are designed to capture one generation of inbreeding, are sampled independently for each generation. These graphs are fully connected, in that each pair of nodes is necessarily connected by either a sibling, clonal, or stranger edge; but transitivity is enforced for sibling and clonal edges.

We generate each population-level relationship graph, comprising $n_{\text{individuals}}$ nodes, by amalgamating subgraphs of fixed size m_{subgraph} as follows. We first simulate $n_{\text{individuals}}/m_{\text{subgraph}}$ constituent subgraphs uniformly at random over the space of transitive graphs with m_{subgraph} labeled nodes and (undirected) sibling/clonal edges. The transitive property dictates that if nodes x and y, and nodes y and z are each connected by a clonal (sibling) edge, then nodes x and z must also be connected by a clonal (sibling or clonal) edge. For example, in the case $m_{\text{subgraph}} = 4$, we can construct 60 possible labeled transitive graphs with sibling/clonal/stranger edges. Removing node labels yields 14 unique configurations, as shown in Fig. B2. Sampling uniformly at random over the 60 possible labeled subgraphs yields the distribution shown in Fig. B2 over the 14 unique graph configurations. Each subgraph is sampled independently using the function `sample_RG`, implemented in the R package `Pv3Rs` (Taylor 2022b). To amalgamate constitutent subgraphs, we ascribe stranger edges for all between-subgraph comparisons.

As a consequence of this construction, which is motivated in part by computational constraints, population-level relationship graphs (for $n_{\text{individuals}}$) are sampled over a nonuniform distribution across the space of transitive graphs with $n_{\text{individuals}}$ labeled vertices and (undirected) sibling/clonal/stranger edges, that is biased towards those containing small, relatively balanced subclusters of clones and siblings: while each subgraph is of fixed size m_{subgraph}, each clonal/sibling component is at most of size m_{subgraph}. When subclusters are balanced, the assumption of conditional independence between allelic and IBD states that underpins the allelic observation model is justified (Fig. A2, Appendix A.2.1.1).

We view simulated individuals within subgraphs as successfully transmitted sporozoites per mosquito bite, and m_{subgraph} as the sporozoite count per bite. Of those sporozoites, all could be clonal, as in a monoclonal inoculation; all could be siblings, as in a multiclonal inoculation from a mosquito in which two genetically distinct parasites recombined, etc. The frequencies of these different scenarios are governed by the distribution over transitive graphs (Fig. B2a), which is uniform for convenience. Cotransmission is modeled through subgraphs derived by sampling parents from a previous subgraph, as captured below through the parameter $p_{\text{cotransmission}}$. Superinfection, which implies independent bites, is modeled through subgraphs derived by sampling parents from the population at large. However, under this construction, bites (and thus genotypes) are not conceptually allocated to human hosts. An additional model is required to allocate bites to humans; it is superfluous for our purposes, but may be relevant for other applications.

B.3.1.1 Generating ancestry and thus genotypes conditional on relationship graphs for the present and previous generation. Given relationship graphs for generations $(k - 1)$ and k, we use Algorithm 1 to simulate the ancestry structure of generation k relative to generation $(k - 1)$; that is, we formulate individuals in generation k as crosses between parents from generation $(k - 1)$. Each sibling/clonal component in generation k is independently assigned two parents from generation $(k - 1)$. Siblings in generation k are modeled as *independent* crosses of the same parental pair from generation $(k - 1)$, and therefore have expected relatedness 1/2 (Wong *et al.* 2018; Taylor, Watson, *et al.* 2019). Clones in generation k are modeled as identical crosses of a given parental pair from generation $(k - 1)$.

For each sibling/clonal component in the relationship graph for generation k, parental pairs are selected under two possible schemes:

- With probability $p_{\text{cotransmission}}$, we sample a constituent subgraph of size m_{subgraph} (that is, one of the $n_{\text{individuals}}/m_{\text{subgraph}}$ subgraphs we generated uniformly at random over the set of transitive relationship graphs of size m_{subgraph}) uniformly at random in generation $(k - 1)$; and then select two parents uniformly at random *without* replacement from that constituent subgraph, yielding an elevated probability of sibling–sibling crosses and selfing. While parents are guaranteed to be sampled from the same subgraph, inbreeding is not guaranteed because a subgraph can contain strangers (Fig. B2a).
- With probability $(1 - p_{\text{cotransmission}})$, we sample two parents from generation $(k - 1)$ uniformly at random *with* replacement, whereby parentals may be derived from the same subgraph but are more likely from different subgraphs.

This sampling scheme is designed to capture stochastic drift, with the enriched probability of selfing and sibling–sibling crosses acting as a proxy for monoclonal mosquito infections and serial cotransmission (Wong *et al.* 2018).

B.4 Simulation model application

We treat the realized fraction of polymorphic markers that are IBD

$$R(j_1, j_2, k) := \frac{\sum_{i=1}^{n_{\text{markers}}} \mathbb{1}\{F_i(j_1, k) = F_i(j_2, k)\} \cdot \mathbb{1}\{| \cup_{j=1}^{n_{\text{individuals}}} A_i(j, k)| > 1\}}{\sum_{i=1}^{n_{\text{markers}}} \mathbb{1}\{| \cup_{j=1}^{n_{\text{individuals}}} A_i(j, k)| > 1\}}$$

as the truth value for the pairwise relatedness of individuals j_1, j_2 in generation k.

To diagnose systematic biases in relatedness estimation, we then generate MLEs of the pairwise relatedness parameter r and, where applicable, the switch rate parameter κ, under both the independence model of relatedness and the HMM using:

- the standard nIBD-to-IBS and (n)IBD-to-allele models predicated on sample allele frequencies;
- the corrected nIBD-to-IBS model based on the sample proportion of nIBD pairs that are IBS at each locus i, that is,

$$c_i := \frac{\sum_{1 \le j_1 < j_2 \le n_{\text{individuals}}} [\mathbb{1}\{A_i(j_1, k) = A_i(j_2, k)\} - \mathbb{1}\{F_i(j_1, k) = F_i(j_2, k)\}]}{\sum_{1 \le j_1 < j_2 \le n_{\text{individuals}}} [1 - \mathbb{1}\{F_i(j_1, k) = F_i(j_2, k)\}]},$$

based on polymorphic markers only. MLEs $(\hat{r}, \hat{\kappa})$ predicated on (n)IBD-to-allele models are generated using the R package `paneljudge` (Taylor 2022a), which implements both the HMM and independence model of relatedness. MLEs for the relatedness parameter r predicated on the nIBD-to-IBS model coupled with the (n)IBD independence model are computed using a custom R script. We do not generate estimates under an HMM with a nIBD-to-IBS observation model.

Under the ancestrally oblivious HMM of relatedness, each parameter set (r, κ) yields a distribution for the realized fraction of IBD

Algorithm 1: Simulate one generation of recombination (based on function `sample_lineages` of R package `Pv3Rs` Taylor 2022b)

```
Input: transitive relationship graph $G_{\text{current}}$ for current generation with sibling/clonal/stranger edges
       possible parents $\ell = 1, \dots, n_{\text{parents}}$
       stratification of parents $L_{\text{parent}}(g) \subset \{1, \dots, n_{\text{parents}}\}$ by parental subgraph $g \in \{1, \dots, n_{\text{individuals}}/m_{\text{subgraph}}\}$
       probability of cotransmission $p_{\text{cotransmission}}$
       number of $n_{\text{markers}}$ at which individuals are genotyped
       size of parasite population $n_{\text{individuals}}$ in subsequent generation
Result ancestry matrix $\mathbf{M} \in \{1, \dots, n_{\text{parents}}\}^{n_{\text{individuals}} \times n_{\text{markers}}}$ for subsequent generation

Delete all stranger edges in $G_{\text{current}}$;
Decompose $G_{\text{current}}$ into fully-connected sibling/clonal subgraphs $S_1, \dots, S_{n_{\text{subgraph}}}$;
for $j \in \{1, \dots, n_{\text{subgraph}}\}$ do
    Delete all sibling edges in $S_j$;
    Decompose $S_j$ into clonal components $C_1^j, \dots, C_{s_j}^j$;
    if $u \sim U(0, 1) < p_{\text{cotransmission}}$ then
        Sample parental subgraph $g \in \{1, \dots, n_{\text{individuals}}/m_{\text{subgraph}}\}$ uniformly at random;
        Sample $\ell_1, \ell_2 \in L_{\text{parent}}(g)$ uniformly at random, without replacement;
    else
        Sample $\ell_1, \ell_2 \in \{1, \dots, n_{\text{parents}}\}$ uniformly at random, with replacement;
    for $C_s^j \in \{C_1^j, \dots, C_{s_j}^j\}$ do
        Simulate one meiosis $\ell_{\text{offspring}} \sim \text{Meiosis}(\ell_1, \ell_2)$ between $\ell_1, \ell_2$
        for individuals $i$ in clonal component $C_s^j$ do
            $\mathbf{M}[i, \,] \leftarrow \ell_{\text{offspring}}$
return $\mathbf{M}$
```

markers (Speed and Balding 2015; Taylor, Jacob, et al. 2019). Convergence of the realized fraction of IBD markers to the relatedness parameter r occurs in the limit where an infinite number of equidistant markers are sampled along an infinitely long genome. While we acknowledge the conceptual difference between the pairwise relatedness parameter r and the realized fraction of sites IBD (Speed and Balding 2015; Taylor, Jacob, et al. 2019), the fact that we model recombination from ancestral principles means that, under our framework, there is no quantity with direct equivalence to r.

B.5 Summary of simulation parameters

We now examine the effects of simulation parameters on distributions of realized IBD/IBS sharing. A summary of simulation parameters is provided in Table B1.

B.5.1 Effect of $n_{\text{individuals}}$ vs n_{founders} on background relatedness. We distinguish two sources of relatedness within our simulation framework:

- "Recent" relatedness, attributed to crosses between shared parents from generation one onwards.
- "Background" relatedness, manifest in generation zero parasites, arising from the finite number of founders n_{founders}.

These two sources of relatedness are different from identity by chance, which stems from limited marker cardinality under our simulation framework.

The population size, $n_{\text{individuals}}$ and the number of founders n_{founders} govern the degree of background relatedness at generation zero. At locus i, each generation zero parasite is assigned ancestry from a founder $f \in \{1, \ldots, n_{\text{founders}}\}$, with independent assignment across both loci and parasites (Appendix B.2). The proportion of pairs IBD at locus i in generation zero, d_i^0, can thus be written

$$d_i^0 = \frac{X_1^2 + \ldots + X_{n_{\text{founders}}}^2}{n_{\text{individuals}}^2}$$

where

$$(X_1, \ldots, X_{n_{\text{founders}}}) \sim \text{Multinomial}(n_{\text{individuals}}, (1/n_{\text{founders}}, \ldots, 1/n_{\text{founders}})),$$

and pairs have been sampled with replacement.

In particular, the expected locuswise relatedness at generation zero is given by

$$\mathbb{E}[d_i^0] = 1 - \left(1 - \frac{1}{n_{\text{individuals}}}\right)\left(1 - \frac{1}{n_{\text{founders}}}\right).$$

The degree of background relatedness exhibited by generation zero parasites influences realized IBD distributions in early generations that follow. It also governs the rate at which sibling and half-sibling pedigrees diverge from the expected relatedness 0.5 and 0.25, respectively. For the chosen parameter values $n_{\text{founders}} = n_{\text{individuals}} = 100$ (Table B1), $\mathbb{E}[d_i^0] \approx 2\%$.

B.5.2 Contribution of sibling/clonal subgraphs: m_{subgraph} vs $n_{\text{individuals}}$. The relative contributions of stranger/sibling/clonal edges in a population-level relationship graph—which are functions of $n_{\text{individuals}}$ and m_{subgraph}, respectively—are important determinants of relatedness. At each generation, we independently sample $n_{\text{individuals}}/m_{\text{subgraph}}$ subgraphs of size m_{subgraph} (uniformly at random, over the set of transitive graphs (Taylor, Watson, et al. 2019; Taylor 2022b)) and knit them together with stranger edges. As such, all between-subgraph comparisons—which constitute proportion $(n_{\text{individuals}} - m_{\text{subgraph}})/n_{\text{individuals}}$ of all parasite pairs (with replacement)—necessarily correspond to stranger edges. Of the remaining proportion of pairwise comparisons, $m_{\text{subgraph}}/n_{\text{individuals}}$, the weighting of sibling/clonal edges is contingent on the underlying uniform distribution over transitive graphs of size m_{subgraph}; the case $m_{\text{subgraph}} = 4$ is shown in Fig. B2a. Population-level relationship graphs enriched for sibling and clonal edges naturally yield rapid growth in relatedness over successive generations of inbreeding. For parasite pairs connected by a stranger edge in generation k, parents are sampled independently from generation $(k - 1)$. Half-sibling, sibling, or clonal relationships between individuals connected by a stranger edge in generation k can therefore arise due to random sampling from the finite population in generation $(k - 1)$, and are a direct function of clonal structure in generation $(k - 1)$. However, we would expect half-sibling, sibling, and clonal relationships in generation k to be driven primarily by sibling/clonal edges within the population-level relatedness graph for generation k itself—particularly in early generations with limited clonal substructure.

Under our framework, we recover a mixture distribution for pairwise IBD sharing $R(j_1, j_2, k)$ in generation k—with a dominant component attributable to parasites separated by at least two generations and smaller components attributable to clones, siblings and half-siblings. The size m_{subgraph} of uniformly sampled transitive graphs with sibling/clonal/stranger edges (Taylor, Watson, et al. 2019; Taylor 2022b), relative to the parasite population size $n_{\text{individuals}}$, governs the relative weightings of the half-sibling, sibling and clonal components to the mixture distribution—and consequently, the rate at which relatedness grows over successive generations.

B.5.3 Breeding between closely related parents: $p_{\text{cotransmission}}$ vs m_{subgraph}. For each sibling/clonal component within the population-graph of generation k, with probability $p_{\text{cotransmission}}$, we have sampled parents without replacement from a subgraph of size m_{subgraph} in the population graph of generation $(k - 1)$. Under our simulation structure, each subgraph is itself sampled uniformly at random over the set of transitive relationship graphs of size m_{subgraph}. The probability of a stranger–stranger cross vs a sibling–sibling cross vs selfing under this regime is therefore a function of m_{subgraph}, with the case $m_{\text{subgraph}} = 4$ illustrated in Fig. B2b. The underlying rationale of this sampling scheme is to yield an enriched probability of sibling–sibling crosses, as a proxy for serial cotransmission (Wong et al. 2018); sampling from the constituent subgraphs *without* replacement augments the contribution of sibling–sibling crosses relative to selfing.

Setting $p_{\text{cotransmission}} = 0$—whereby the sampling of parents is predicated on stochastic drift only (Appendix B.5.2)—yields a "blindspot" in our simulated distributions of realized relatedness: in contrast to empirical estimates, we recover negligibly few *non-clonal* parasite pairs with IBD sharing exceeding 60%, signifying crosses between related parents. To recapitulate the smattering of parasite pairs with IBD sharing in the range 60%–90% seen in empirical estimates, we must inflate the probability of breeding between closely related parental lineages, consistent with evidence of extensive cotransmission (Nkhoma et al. 2020). We modulate $p_{\text{contransmission}} > 0$ and exploit substructure within simulated population-level relatedness graphs to mitigate this blindspot. While we have not explicitly embedded a mechanistic transmission model within our framework, subgraphs derived

Table B1. Summary of simulation parameters.

Feature	Parameter	Interpretation	Value
Markers	n_{markers}	Number of equidistant markers distributed across a single chromosome	24,000
	ρd	Distance between consecutive markers in Morgans	≈ 0.002
	$\bar{M} = 2(1 - e^{-\rho d})^{-1}$	Average number of consecutive markers inherited from the same parent in a cross (reduces to $\bar{M} = 2$ given locus independence)	1,000
Relationship graphs	$n_{\text{individuals}}$	Number of not necessarily distinct genotypes, each representing an equal number of actual parasites	100
	m_{subgraph}	Size of transitive clone/sibling/stranger graphs (Taylor, Watson, *et al.* 2019) sampled uniformly at random; interpretable as a multiplicity of infection upper bound (if subgraphs are viewed as parasite genotypes within a host) or the number of successful, not necessarily distinct genotypes per bite (if subgraphs are viewed as successfully transmitted sporozoites per bite)	4
Founder	n_{founders}	Number of unrelated founders	100
	y	Number of possible alleles at each at marker	2
	$p_i(q)$	Probability of each founder f harboring allele $q \in \{1, \ldots, y\}$ at marker i	(0.1, 0.9)
Inbreeding	$n_{\text{generations}}$	Number of discrete, nonoverlapping generations of inbreeding [exc. 0]	10
	$p_{\text{cotransmission}}$	Probability of sampling parents without replacement from the same constituent subgraph *vs* with replacement from the population as a whole in the previous generation for each sibling/clonal component in a relationship graph	0.4

from parental lineages sampled from the population at large can be thought to encompass superinfection; while subgraphs derived from parental lineages sampled from a previous subgraph can be thought to correspond to cotransmission.

We note, however, that the augmented probability of sibling–sibling crosses under $p_{\text{cotransmission}} > 0$ yields elevated IBD sharing (in the vicinity of 60%–90%) for a relatively small proportion of parasite pairs: it primarily affects within-component comparisons for each sibling/clonal component in the accompanying population-level relationship graph, with an upper bound m_{subgraph} on the size of each sibling/clonal component. Between-cluster comparisons are largely modulated by stochastic drift, as discussed in Appendix B.5.2. Reverting to the mixture distribution interpretation of simulated pairwise IBD delineated in Appendix B.5.2, setting $p_{\text{contransmission}} > 0$ principally has the effect of broadening the sibling component, to include a small proportion of pairs reflecting breeding between recently related parents.

B.5.4 Linkage structure and spanning (n)IBD segments: n_{markers} vs $\bar{M}$. The mean length of consecutive markers $\bar{M}$ inherited from the same parent in a single meiosis governs the degree of dependence between loci, and consequently, the amount of linkage structure exhibited by related parasite pairs. For this linkage structure to emerge, however, the n_{markers} of interest must span sufficiently many segments of (n)IBD markers; and each segment of (n)IBD markers must be sufficiently long. As a function of the pairwise relatedness parameter $r^{(k,\ell)}$, the interplay between $\bar{M}$ and n_{markers} also governs the variance of the realized fraction of (polymorphic) sites IBD—which we treat as our truth value to diagnose biases in relatedness estimation.

We adopt a physical argument to recover $\bar{M}$ as a function of the marker count n_{markers}: for some number of equidistant markers n_{markers}, distributed across a chromosome of length L Morgans, the mean length of consecutive markers inherited from the same parent in a single meiosis is given by

$$\bar{M}(n_{\text{markers}}) = \frac{2}{1 - e^{-L/n_{\text{markers}}}},$$

obtained by plugging the inter-marker distance L/n_{markers} into Equation (B1).

We can derive a related metric for the average length of contiguous (n)IBD markers for independent crosses of a parental pair. For nonmeiotic siblings, loci ℓ and $\ell + 1$ (with intermarker distance L/n_{markers}) share the same (n)IBD status with probability

$$\begin{aligned} &P(d_{\ell+1} = d_\ell) \\ &= \underbrace{[0.5(1 + e^{-L/n_{\text{markers}}})]^2}_{\substack{\text{within each nonmeiotic sibling,} \\ \text{loci } \ell \text{ and } \ell+1 \text{ inherited} \\ \text{from the same parent}}} + \underbrace{[0.5(1 - e^{-L/n_{\text{markers}}})]^2}_{\substack{\text{within each nonmeiotic sibling,} \\ \text{loci } \ell \text{ and } \ell+1 \text{ inherited} \\ \text{from different parents}}} \\ &= 0.5[1 + e^{-2L/n_{\text{markers}}}]. \end{aligned}$$

For equidistant markers, the length of a contiguous segment of (n)IBD markers for nonmeiotic siblings is therefore geometrically distributed, with mean

$$\bar{Q}(n_{\text{markers}}) = \frac{2}{1 - e^{-2L/n_{\text{markers}}}}.$$

B.5.5 Minor allele frequency spectra: $\boldsymbol{p}_i$. At each locus $i = 1, \ldots, n_{\text{markers}}$, the count $n_{\text{founders}} \cdot \theta_{\text{founder}}(q, i)$ of each allele $q \in \{1, \ldots, y\}$ in the founder population (size n_{founders}) follows a multinomial distribution

$$\begin{aligned} &n_{\text{founders}} \cdot (\theta_{\text{founder}}(1, i), \ldots, \theta_{\text{founder}}(y, i)) \\ &\quad \sim \text{Multinomial}\big(n_{\text{founders}}, (p_i(1), \ldots, p_i(y))\big) \end{aligned}$$

where $p_i(q)$ is the probability of assigning allele q to a founder at locus i.

In this manuscript, we consider only biallelic loci, i.e. we fix $y = 2$. The parameter p_i governing founder sample allele frequencies θ_{founder} has a strong bearing on minor allele frequency (MAF) spectra observed across successive generations. In the infinite generation limit, we would expect an allele $q_i \in \{1, 2\}$ to reach fixation at each locus i, akin to the canonical Wright–Fisher model of stochastic drift (Etheridge 2011). As in the Wright–Fisher model, however, it is not necessarily the case that the *major* allele in the founder population reaches fixation. As such, in a transient setting, we can either observe allele frequencies becoming increasingly balanced over successive generations (as the minor allele grows in frequency); or increasingly unbalanced (as the major allele grows in frequency). For the most part, however, we expect a reduction in genetic diversity as crosses are repeatedly assigned major alleles.

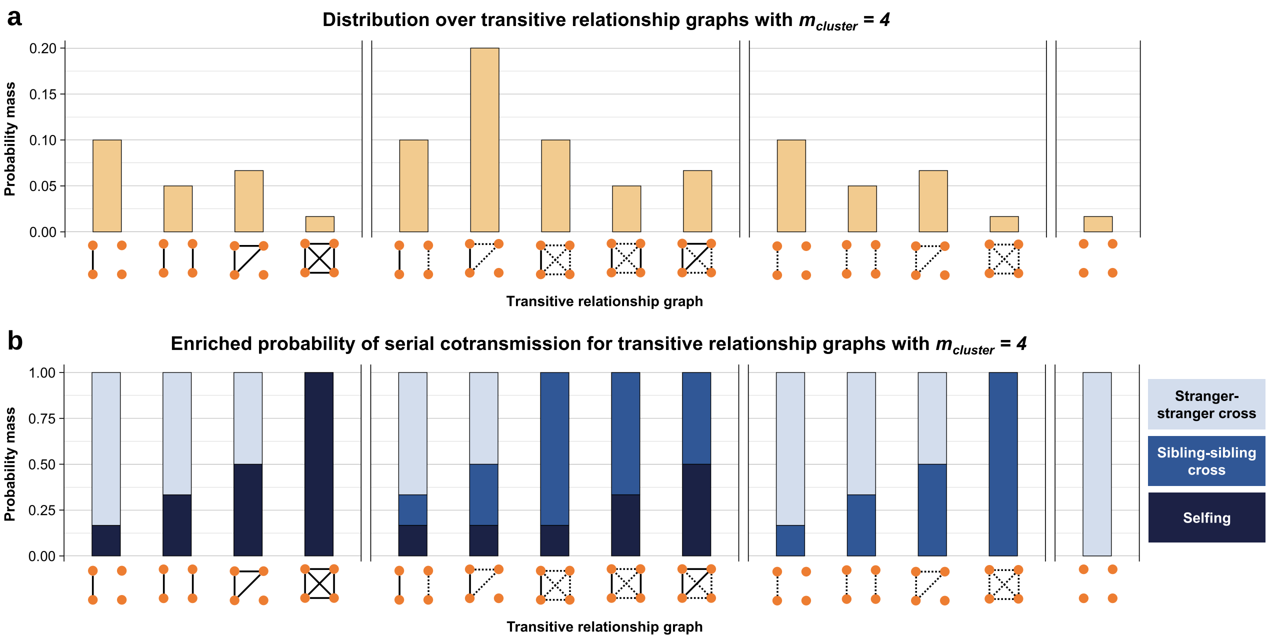


Fig. B2. a) Probability masses for transitive relationship graphs with sibling/clone/stranger edges (Taylor, Watson, *et al.* 2019) in the case $m_{\text{subgraph}} = 4$ (after removing node labels), under a uniform distribution over all allowable graphs (see Appendix B.3.1). Clonal and sibling edges are indicated with solid and dashed lines, respectively. b) Probability of stranger–stranger crosses vs sibling–sibling crosses vs selfing when two parents are sampled uniformly at random *without* replacement from each allowable relationship graph with $m_{\text{subgraph}} = 4$.

To recover positively skewed biallelic MAF spectra with zero mode in the order of 10 generations, we set $\min\{p_i(1), p_i(2)\} \approx 0.1$, translating to unbalanced founder allele frequencies. Under this setting, however, we find that allele frequencies become increasingly balanced over successive generations for a subset of markers.

Appendix C: Glossary of terms

$\overline{\text{IBD}}$ proportion of pairs of sampled parasites (with replacement, i.e. including self–self comparisons) that are IBD at a given locus.

$\overline{\text{IBS}}$ proportion of pairs of sampled parasites (with replacement, i.e. including self–self comparisons) that are IBS at a given locus.

(n)IBD-to-observation model inclusive of nIBD-to-allele, IBD-to-allele and nIBD-to-IBS models, under the assumed absence of genotyping error (whereby IBD necessarily implies IBS, i.e. IBD-to-IBS=1).

background relatedness under the simulation model, relatedness structure in generation zero parasites.

genotype a specific realization of the genome, which is a random variable distributed according to some ancestral process; in this study, we consider the genotype to be a sequence over all polymorphisms in the genome (elsewhere, its definition extends to subsets).

inbreeding recombination between genetically different but related parasites.

individual a parasite genotype.

meiotic siblings siblings derived from the same oocyst, i.e. siblings that are complements of one another and thus not independent.

outbreeding recombination between genetically unrelated parasites.

realised relatedness the fraction of polymorphic markers within the sample that are IBD for a given parasite pair.

recent relatedness under the simulation model, relatedness structure resulting from inbreeding from generation one onwards.

relatedness structure the locuswise partition of individuals in a parasite population or sample into transitive IBD clusters.

sample a collection of individuals, i.e. a collection of parasite genotypes.

selfing recombination between genetically identical parasites.

strangers a pair of individuals separated by two or more generations of recombination.

Editor: M. Sillanpää