MOIRE: A software package for the estimation of allele frequencies and effective multiplicity of infection from polyallelic data

Maxwell Murphy * Bryan Greenhouse

October 3, 2023

Abstract

Malaria parasite genetic data can provide insight into parasite phenotypes, evolution, and transmission. However, estimating key parameters such as allele frequencies, multiplicity of infection (MOI), and within-host relatedness from genetic data has been challenging, particularly in the presence of multiple related coinfecting strains. Existing methods often rely on single nucleotide polymorphism (SNP) data and do not account for within-host relatedness. In this study, we introduce a Bayesian approach called MOIRE (Multiplicity Of Infection and allele frequency REcovery), designed to estimate allele frequencies, MOI, and within-host relatedness from genetic data subject to experimental error. Importantly, MOIRE is flexible in accommodating both polyallelic and SNP data, making it adaptable to diverse genotyping panels. We also introduce a novel metric, the effective MOI (eMOI), which integrates MOI and within-host relatedness, providing a robust and interpretable measure of genetic diversity. Using extensive simulations and real-world data from a malaria study in Namibia, we demonstrate the superior performance of MOIRE over naive estimation methods, accurately estimating MOI up to 7 with moderate sized panels of diverse loci (e.g. microhaplotypes). MOIRE also revealed substantial heterogeneity in population mean MOI and mean relatedness across health districts in Namibia, suggesting detectable differences in transmission dynamics. Notably, eMOI emerges as a portable metric of within-host diversity, facilitating meaningful comparisons across settings, even when allele frequencies or genotyping panels are different. MOIRE represents an important addition to the analysis toolkit for malaria population dynamics. Compared to existing software, MOIRE enhances the accuracy of parameter estimation and enables more comprehensive insights into within-host diversity and population structure. Additionally, MOIRE's adaptability to diverse data sources and potential for future improvements make it a valuable asset for research on malaria and other organisms, such as other eukaryotic pathogens. MOIRE is available as an R package at `https://eppicenter.github.io/moire/`.

1 Introduction

Genetic data can be a powerful source of information for understanding malaria parasite phenotype and transmission dynamics, providing insight into population structure and connectivity, and thereby informing control and elimination efforts. However, analysis is complicated in malaria due to the presence of multiple coinfecting, genetically distinct strains. More specifically, genetically distinct strains may share the same alleles at genetic loci, rendering the actual number of strains contributing a particular allele unknown and making it difficult to estimate fundamental statistics such as population allele frequencies and multiplicity of infection (MOI). Standard methods to address this either naively estimate allele frequencies and MOI without considering the total number of strains contributing an allele [1–3], or completely ignore polyclonal samples during analysis. Naive estimation without accounting for strain count contribution results in biased estimates of allele frequencies and MOI, leading to meaningful systematic biases in summary statistics. For example, naive estimation of allele frequencies without consideration of strain composition from polyclonal samples results in a consistent overestimation of heterozygosity, leading to potentially faulty inference about

*Corresponding author: Maxwell Murphy, Department of Biostatistics, School of Public Health, University of California, Berkeley, CA 94720, USA. Email: mm@maxmurphy.dev

population diversity. Additionally, naive estimation offers no principled way to address genotyping error beyond heuristics, further biasing estimates of diversity in ways that depend on choices made during initial interpretation of genotyping data. Alternatively, considering only monoclonal samples is potentially problematic, as this may require a substantial number of samples to be discarded when collected from regions where multiple infection is the rule rather than the exception. Further, the monoclonal subset of samples are fundamentally different from the larger population of interest, as they preclude the possibility of within-host relatedness between strains. This ignores a potentially important source of information about transmission dynamics, as within-host relatedness may be indicative of co-transmission or persistent local transmission [4–6].

To address these issues and make full use of available data, Chang et al. developed a Bayesian approach (*THE REAL McCOIL*) to estimate allele frequencies and MOI in the context of polygenomic infections from single nucleotide polymorphism (SNP) based data [7]. More recently, *coiaf* [8] and *SNP-Slice* [9] have been developed to further improve computational efficiency and resolving power. However, none of these methods directly consider or estimate within-host relatedness. Further, these methods are all tailored to SNP based data and are unable to accommodate more diverse polyallelic loci, such as microsatellites, which have been widely used in population genetic studies [1–3, 10]. Other methods that infer within-host relatedness [11] in contrast rely on whole genome sequencing (WGS) data, however WGS based approaches frequently have poor sensitivity for detecting minority strains and low density infections [12]. In recent years, the declining cost of DNA sequencing and development of high throughput, high diversity, targeted sequencing panels have made polyallelic data even more attractive for genomic based studies of malaria [12–14]. Genetic analysis methods leveraging polyallelic loci have the potential for substantially increased resolving power over their SNP based counterparts, particularly in the context of related polyclonal infections in malaria [15, 16]. Unfortunately, there are limited tools available to analyse these types of data.

We present here a new Bayesian approach, **M**ultiplicity **O**f **I**nfection and allele frequency **RE**covery from noisy polyallelic data (*MOIRE*), that, like *THE REAL McCOIL*, enables the estimation of allele frequencies and MOI from genomic data that are subject to experimental error. In addition, MOIRE estimates and accounts for within-host relatedness of parasites, a common occurrence due to the inbreeding of parasites serially co-transmitted by mosquitoes [5, 17]. Critically, MOIRE takes as input genetic data of arbitrary diversity, allowing for estimation of allele frequencies, MOI, and within-host relatedness from polyallelic as well as biallelic data. MOIRE is able to fully utilize polyallelic data, yielding joint estimates of allele frequencies, sample specific MOIs and within-host relatedness along with probabilistic measures of uncertainty. We demonstrate through simulations and applications to empirical data the ability of MOIRE to leverage a variety of polyallelic markers. Polyallelic markers can greatly improve jointly estimating sample MOI, within-host relatedness, and population allele frequencies, resulting in reduced bias and increased power for understanding population dynamics from genetic data. We also introduce a new metric of diversity, the effective MOI (eMOI), a continuous value that combines estimates of the true MOI and the degree of within-host relatedness in a single sample, providing an interpretable quantity that is comparable across genotyping panels and transmission settings.

2 Methods

2.1 A model of infection and observation

Consider observed genetic data $X = (X_1, \ldots, X_n)$ from n samples indexed by i, where each X_i is a collection of vectors indexed by l of possibly differing length, representing the varying number of alleles possible at each locus, e.g. polyallelic loci. Each vector is binary, with 1 representing the allele was observed or 0 representing the allele went unobserved at locus l for sample i. From this data, we wish to estimate MOI for each individual ($\boldsymbol{\mu} = [\mu_1, \ldots, \mu_n]$), within host relatedness ($\boldsymbol{r} = [r_1, \ldots, r_n]$), defined as the average proportion of the genome that is identical by descent across all strains, individual specific genotyping error rates ($\boldsymbol{\epsilon^+} = [\epsilon_1^+, \ldots, \epsilon_n^+]$ and $\boldsymbol{\epsilon^-} = [\epsilon_1^-, \ldots, \epsilon_n^-]$), and population allele frequencies at each locus ($\boldsymbol{\pi} = [\pi_1, \ldots, \pi_l]$). Similar to [7], we

applied a Bayesian approach and looked to estimate the posterior distribution of $\boldsymbol{\mu}, \boldsymbol{r}, \boldsymbol{\epsilon}^+, \boldsymbol{\epsilon}^-$ and $\boldsymbol{\pi}$ as

$$P(\boldsymbol{\mu}, \boldsymbol{r}, \boldsymbol{\epsilon}^+, \boldsymbol{\epsilon}^-, \boldsymbol{\pi}|X) \propto \prod_{i=1}^{n} P(X_i|\mu_i, r_i, \epsilon_i^+, \epsilon_i^-, \boldsymbol{\pi}) P(\boldsymbol{\mu}, \boldsymbol{r}, \boldsymbol{\epsilon}^+, \boldsymbol{\epsilon}^-, \boldsymbol{\pi}) \tag{1}$$

where we assumed independence between samples.

Given that the observed genetic data is experimentally derived, it is subject to some rate of false positives where an allele is erroneously called as present, and false negatives where an allele is erroneously called as absent. To address this issue, we augmented our model with a latent true genetic state Y, reflecting the true presence or absence of alleles at each locus for each individual. Augmenting our model with this latent state allowed us to incorporate and model the uncertainty around measurement of genetic data separately from the uncertainty around the true genetic state, as expressed in the following factorization:

$$P(X_i|Y_i, \mu_i, r_i, \epsilon_i^+, \epsilon_i^-, \boldsymbol{\pi}) = P(X_i|Y_i, \epsilon_i^+, \epsilon_i^-) P(Y_i|\mu_i, r_i, \boldsymbol{\pi}) \tag{2}$$

We assumed a prior in which the MOI of each individual was independent of the MOI of other individuals, relatedness was independent across individuals, error rates were independent across individuals, and allele frequencies were independent across loci and without linkage disequilibrium, yielding the following factorization:

$$P(\boldsymbol{\mu}, \boldsymbol{r}, \boldsymbol{\epsilon}^+, \boldsymbol{\epsilon}^-, \boldsymbol{\pi}) = \prod_{i=1}^{n} P(\mu_i) P(r_i) P(\epsilon_i^+) P(\epsilon_i^-) \prod_{j=1}^{l} P(\pi_l) \tag{3}$$

2.1.1 Modeling genotyping error

We assumed that each observed sample has an intrinsic rate of false positives ϵ_i^+ and false negatives ϵ_i^-, reflecting the varying quality of samples that are genotyped and factors that impact ability to detect alleles accurately. This rate is divided by the number of alleles possible at each locus (L_j), such that the probability of observing an allele given that it is not present is $\frac{\epsilon_i^+}{L_j}$, and the probability of not observing an allele given that it is present is $\frac{\epsilon_i^-}{L_j}$. The full likelihood of observing the data given the latent true genetic state and error rates for an individual sample is then given by

$$P(X_i|Y_i, \epsilon_i^+, \epsilon_i^-) = \prod_{j=1}^{l} \prod_{a=1}^{L_j} \begin{cases} \frac{\epsilon_i^+}{L_j} & \text{if } X_{ija} = 1 \text{ and } Y_{ija} = 0 \\ 1 - \frac{\epsilon_i^+}{L_j} & \text{if } X_{ija} = 1 \text{ and } Y_{ija} = 1 \\ \frac{\epsilon_i^-}{L_j} & \text{if } X_{ija} = 0 \text{ and } Y_{ija} = 1 \\ 1 - \frac{\epsilon_i^-}{L_j} & \text{if } X_{ija} = 0 \text{ and } Y_{ija} = 0 \end{cases} \tag{4}$$

2.1.2 Modeling latent genetic state

We modeled the latent genetic state as a combination of two processes. For a fixed MOI, the genetic state of each strain at a given locus can be explained either as an independent draw from the background population of parasites characterized by the population allele frequencies $\boldsymbol{\pi}$, or as being related to and explained by another within host strain. We can express this as a mixture model over the possible arrangements of μ_i strains, where each strain is either a within-host related strain at that locus or derived from the background population, giving us:

$$P(Y_i|\mu_i, r_i, \boldsymbol{\pi}) = \prod_{j=1}^{l} \sum_{k=0}^{\mu_i - 1} P(Y_{ij}|m = k, \mu_i, \pi_j) P(m = k|r_i, \mu_i) \tag{5}$$

Here, m is a variable indicating the number of strains that are related to each other within the host, constrained to at most $\mu_i - 1$ related strains, requiring at least one "reference" strain that other strains are related to. We modeled the number of related strains within a host as a binomial random variable with probability of success r_i:

$$P(m = k|r_i, \mu_i) = \binom{\mu_i - 1}{k}(1 - r_i)^k (r_i)^{\mu_i - k - 1} \tag{6}$$

For the remaining $\mu_i - k$ unrelated strains, we modeled the genetic state of each unrelated strain as independent of the genetic state of other within host strains and dependent only on the population allele frequencies $\boldsymbol{\pi}$, allowing us to express the likelihood of the genetic state of each locus in terms of draws from a multinomial distribution:

$$P(Y_{ij}|m = k, \mu_i, \pi_j) = \sum_{y^* \in \mathcal{Y}_{Y_{ij}}} \binom{(\mu_i - k)!}{y_1^*! \cdots y_a^*!} \prod_{a=1}^{L_j} \pi_{ja}^{y_a^*} \tag{7}$$

Here, $\mathcal{Y}_{Y_{ij}}$ is the set of all possible configurations of alleles at a given locus for a given individual that are compatible with the binary vector Y_{ij}, where compatibility is defined as having at least one allele present in each position of the vector that is also present in Y_{ij}. For example, if $Y_{ij} = (1, 0, 1, 0)$ and there are 4 unrelated strains present, then $\mathcal{Y}_{Y_{ij}} = \{(1, 0, 3, 0), (3, 0, 1, 0), (2, 0, 2, 0)\}$. Naively computing this term would require enumerating all possible configurations of alleles at each locus for each individual, which may be computationally challenging for large datasets with high MOIs. We instead used a more efficient method of computing the likelihood of the latent genetic state which we describe in the supplementary material (6.1).

2.1.3 Prior distributions

We assumed that sample MOIs were drawn from a zero-truncated Poisson (ZTP) distribution with rate λ. In general, we do not know the population mean MOI λ, so we used a hyperprior distribution to model our uncertainty about λ. We assumed a gamma distribution hyperprior on λ with user specified shape α_λ and rate β_λ. Within-host relatedness was bounded between 0 and 1 and modeled with a beta distribution prior with user specified shape parameters α_r and β_r. Sample specific false positive and false negative rates were bounded between 0 and user specified ϵ^+_{max} or ϵ^-_{max}, respectively, and were modeled with scaled beta distributions with user specified shape parameters $(\alpha_{\epsilon^+}, \beta_{\epsilon^+})$ and $(\alpha_{\epsilon^-}, \beta_{\epsilon^-})$. We placed a uniform prior on allele frequencies π_j for each locus j.

2.2 Effective MOI

By estimating MOI and within-host relatedness, we can estimate a continuous metric of genetic diversity within a host, the effective MOI (eMOI), which we define as:

$$eMOI = 1 + (1 - r)(\mu - 1) \tag{8}$$

One interpretation of the effective MOI is the expected number of distinct alleles at a locus with infinite diversity, i.e. a locus where heterozygosity is 1 (see Supplementary 6.3 for a formal derivation). In the case of no within-host relatedness, this is simply the MOI. However, when there is within-host relatedness, the effective MOI is the MOI weighted by the probability that a given strain is unrelated to all other strains within the host, and ranges from 1 to μ. This value better reflects the true genetic diversity within a host than the MOI alone, and allows for comparison and differentiation of genetic diversity across hosts with the same MOIs. We also note that eMOI is likely to be more identifiable than MOI or within-host relatedness alone because it is a one-dimensional combination of the two estimated parameters with synergistic properties around precision. As MOI increases, precision around estimates of within-host relatedness also increases as

there are more observations available to inform within-host relatedness. As MOI decreases, precision around the estimate of within-host relatedness decreases, however the contribution to the estimate of eMOI from within-host relatedness also decreases, and thus the overall precision of eMOI is maintained.

2.3 Simulation Procedure

We sampled population allele frequency vectors for each locus from a uniform Dirichlet distribution, or used empirical estimates from regional populations, resulting in an L_j-dimensional simplex for each genetic locus. For each individual, we sampled the MOI for each individual from a ZTP distribution with rate λ, and then independently sampled alleles for each locus from a multinomial distribution parameterized by p_j. We simulated relatedness within-host for individual i by sampling alleles from a randomly chosen existing strain within the host with probability r_i, or from the population allele frequency distribution otherwise. Relatedness r_i was drawn from a beta distribution with $\alpha = 1$ and $\beta = \{.2, .6, 2\}$ for low, moderate, and highly related populations respectively. We then generated observed genetic data by applying a corrupting error process where alleles were false positives or false negatives consistent with our error model. We compared our approach to estimating allele frequencies, MOI, and observed heterozygosity to naive empirical estimators that do not consider within-host relatedness, MOI, or genotyping errors. We also evaluated the ability of our model to recover true relatedness and eMOI of samples.

2.4 Inference and Implementation

We fit our model to the observed genetic data using a Markov Chain Monte Carlo (MCMC) approach using the Metropolis-Hastings algorithm with a variety of update kernels. Details of sampling and implementation are described in the supplementary material (6.2). MOIRE is implemented as an R package and is available with tutorials and usage guidance at `https://eppicenter.github.io/moire/`. All sampling procedures were implemented using `Rcpp` [18] for efficiency. Substantial effort was placed on ease of use and limiting the amount of tuning required by the user by leveraging adaptive sampling methods. We provide uninformative default priors for all parameters and recommend that users only modify priors if they have strong prior knowledge about the parameters, such as experimentally derived estimates of false positive and false negative rates using samples with known parasite compositions and densities. All analysis conducted in this paper was done using MOIRE with default priors and settings, using 40 parallel tempered chains for 5000 burn-in steps, followed by 10,000 samples which were thinned to 1000 total samples.

3 Results

3.1 Estimation of multiplicity of infection, within-host relatedness, and allele frequencies

Using our previously described simulation procedure, we simulated collections of 100 samples under varied combinations of population mean MOI, average within-host relatedness, false positive and false negative rates, and different genotyping panels. Individual MOIs were drawn from ZTP distributions with rate parameters 1, 3, and 5, resulting in mean MOIs of 1.58, 3.16, and 5.03 respectively. Within-host relatedness was simulated from settings with low, moderate, and high relatedness. False positive and false negative rates were varied from 0 to 0.1. We first simulated synthetic genomic loci with prespecified diversity: 100 SNPs, 30 loci with 5 alleles (moderate diversity), 30 loci with 10 alleles (high diversity), and 30 loci with 20 alleles (very high diversity) with frequencies drawn from the uniform Dirichlet distribution. We also assessed potential real world performance of MOIRE by simulating data for 5 currently used genotyping panels from 12 regional populations characterized by the MalariaGEN Pf7 dataset [19] as described in the supplementary material (6.4, Supplementary Fig. 2). Genetic loci were selected according to a 24 SNP panel [20], a 101 SNP panel [21], and 3 recently developed amplicon sequencing panels consisting of 128 [13], 165 [22], and 233 [14] diverse microhaplotypes respectively. These simulations were done at moderate false positive and false negative rates of .01 and .1 respectively. We then ran MOIRE and calculated summary statistics of interest on the sampled posterior distributions.

We estimated allele frequencies, heterozygosity, MOI, within-host relatedness, and eMOI using the mean or median of the posterior distribution output by MOIRE. It should be noted that within-host relatedness is only defined for polyclonal infections, so the posterior distribution of within-host relatedness is conditional on the MOI being greater than 1. We contrasted these with naive estimates of allele frequency and MOI by assuming that an observed allele was contributed by a single strain, and estimated MOI as equal to the second-highest number of alleles observed across loci. We calculated ground truth allele frequencies using the true number of strains contributing each allele.

Under moderate false positive and false negative rates of 0.01 and 0.1 respectively, MOIRE accurately recovered parameters of interest across a range of genotyping panels, population MOI, and within-host relatedness (Fig. 1, Table 1). Allele frequencies estimated by MOIRE were unbiased across genotyping panels, leading to unbiased estimates of heterozygosity. Naive estimation exhibited substantial bias that varied with respect to the true allele frequency. Rare alleles tended to be overestimated and common alleles underestimated, leading to inflated estimates of heterozygosity.

MOI was also well estimated by MOIRE, with accuracy increasing substantially in the presence of more diverse loci. In the context of SNPs, MOIRE recovered MOI accurately up to approximately 4 strains, and then began to exhibit limited ability to resolve. More diverse panels enabled greatly improved resolving power, allowing for the accurate recovery of MOI up to approximately 7 strains. Naive estimation substantially underestimated MOI in comparison, due in part to the limited capacity of low diversity loci to discriminate MOI, as well as the presence of related strains that deflate the observed number of distinct alleles. This bias was particularly prominent for low diversity markers such as SNPs which can only resolve up to 2 strains. These patterns held across varied false positive and false negative rates, with bias for both estimators most prominent when using SNPs and in the presence of increased false negative rates (Supplementary Fig. S1).

MOIRE was generally able to recover within-host relatedness, particularly for moderate and high diversity markers in the context of high relatedness. SNP based panels had difficulty resolving individual level within-host relatedness and were sensitive to the uniform prior. It should be noted that in the circumstance that a monoclonal infection has an inferred MOI greater than 1, MOIRE will likely classify these infections with very high relatedness (Figure 1E). This is due to the presence of false positives that MOIRE will sometimes infer as an infection consisting of highly related strains rather than being explained by observation error. Therefore, within-host relatedness should be interpreted in the context of the probability of the infection being polyclonal. A more robust metric is eMOI, since it is a metric of diversity that integrates MOI and within-host relatedness.

MOIRE recovered eMOI with high accuracy under all conditions using polyallelic panels. SNP panels exhibited a larger degree of bias at higher eMOI, but still performed relatively well for eMOI of up to 4. This demonstrates that while identifiability of MOI or within-host relatedness may be challenging in some situations, eMOI is a reliably identifiable quantity when estimated using highly polymorphic markers.

All simulations were also conducted without any relatedness present. MOIRE was still able to accurately recover allele frequencies, heterozygosity, and MOI, indicating that minimal bias or uncertainty are introduced by attempting to estimate relatedness (Supplementary Figure S2).

3.2 Population inference

MOIRE is a probabilistic approach providing a full posterior distribution over model parameters, allowing estimation of credible intervals for model parameters as well as functions thereof. While sample level parameters estimated by the model are useful, it may also be useful to estimate population level summary statistics for reporting and comparison purposes. We thus calculated the posterior distribution of population level summaries of interest, such as mean MOI, mean within-host relatedness, and mean eMOI. We note that mean within-host relatedness is defined only for samples with MOI greater than 1, therefore the posterior distribution of mean within-host relatedness was calculated across samples with MOI greater than 1 at each iteration of the MCMC algorithm. MOIRE accurately estimated these quantities across a range of conditions (Fig. 2), with the best performance seen for polyallelic data.

Panel	Source	Heterozygosity		Allele Freqs.		MOI		Relatedness	eMOI
		MOIRE	Naive	MOIRE	Naive	MOIRE	Naive	MOIRE	MOIRE
100 SNP	Synthetic	0.01 (.95)	0.05	0.02 (.95)	0.06	1.72 (.85)	3.61	0.20 (.70)	0.37 (.77)
Moderate Div.	Synthetic	0.01 (.99)	0.04	0.01 (.96)	0.02	1.55 (.88)	2.53	0.14 (.77)	0.17 (.91)
High Div.	Synthetic	0.01 (.99)	0.02	0.01 (.91)	0.01	1.29 (.86)	1.87	0.11 (.75)	0.12 (.89)
Very High Div.	Synthetic	0.02 (.60)	0.01	0.01 (.82)	0.01	1.02 (.86)	1.28	0.10 (.77)	0.10 (.85)
24 SNP	[20]	0.01 (.90)	0.04	0.02 (.90)	0.05	1.95 (.81)	3.66	0.21 (.75)	0.45 (.86)
101 SNP	[21]	0.01 (.90)	0.04	0.02 (.95)	0.05	1.77 (.85)	3.62	0.20 (.71)	0.36 (.79)
AMPLseq	[13]	0.01 (.97)	0.05	0.01 (.94)	0.02	1.32 (.88)	1.86	0.12 (.71)	0.14 (.88)
MaD^{4}HatTeR	[22]	0.01 (.98)	0.05	0.01 (.95)	0.03	1.24 (.90)	1.88	0.13 (.68)	0.12 (.88)
AmpliSeq	[14]	0.02 (.94)	0.06	0.01 (.93)	0.02	1.34 (.83)	1.56	0.12 (.67)	0.12 (.88)

Table 1. Mean absolute deviation (MAD) of estimates of MOI, heterozygosity, within-host relatedness, and eMOI across simulations using synthetic (top) and real-world (bottom) genotyping panels. The MAD of estimates of MOI were calculated by taking the mean of the MAD for each stratum of true MOI between 1 and 10. MOI Within-host relatedness accuracy is only considered for samples with a true MOI > 1. Coverage rates of 95% credible intervals are shown in parentheses for estimates by MOIRE.

Population mean MOI was accurately estimated across all panels, with improved precision at lower levels (Fig. 2A, Table 2). Credible interval (CI) coverage in general was poor, likely due to the challenge of identifiability in conjunction with within-host relatedness. SNP panels were largely unable to resolve population level mean within-host relatedness and exhibited poor CI coverage and substantial sensitivity to the uniform prior specification due to the low relative information contained in these markers. Polyallelic panels in contrast had improved precision as more diverse panels were used, although CI coverage was also poor due to persistent sensitivity to the uniform prior as indicated by slightly overestimating within-host relatedness below .5 and underestimating within-host relatedness above .5.

Population mean eMOI was remarkably accurate for low and medium mean MOI when using SNP based panels, with bias only becoming apparent at higher mean MOI (Fig. 2C, Table 2). Polyallelic panels had substantially improved precision across a wide range of values, further demonstrating that while population mean within-host relatedness or mean MOI may be challenging to identify, mean eMOI remains a highly identifiable quantity when genetic markers with sufficient diversity are used.

Panel	Mean MOI (Cov. %)	Mean Relatedness (Cov. %)	Mean eMOI (Cov. %)
100 SNP	1.18 (.17)	0.24 (.05)	0.26 (.23)
Moderate Div.	0.67 (.24)	0.15 (.06)	0.07 (.49)
High Div.	0.38 (.29)	0.10 (.09)	0.04 (.49)
Very High Div.	0.29 (.27)	0.08 (.11)	0.04 (.34)
24 SNP	0.63 (.21)	0.23 (.06)	0.32 (.08)
101 SNP	0.53 (.30)	0.20 (.09)	0.33 (.09)
AMPLseq	0.39 (.23)	0.11 (.15)	0.09 (.23)
MaD^{4}HatTeR	0.42 (.28)	0.11 (.27)	0.09 (.28)
AmpliSeq	0.40 (.33)	0.10 (.24)	0.08 (.33)

Table 2. Mean absolute deviation (MAD) and CI coverage at the 95% interval of estimates of population mean MOI, mean within-host relatedness, and mean eMOI across all simulations using synthetic (top) and real-world (bottom) genotyping panels.

3.3 Metric stability across genetic backgrounds

Population metrics of genetic diversity enable researchers to make comparisons across space and time, and to answer questions relating to differences in transmission dynamics. In order for a metric to be useful for these purposes, it must be sensitive to changes in transmission dynamics while remaining insensitive to other factors that vary and may confound interpretation, such as the genotyping panel used, or the local allele

frequencies for a given panel. For example, if we were to compare two populations that exhibit the same transmission dynamics, we would want the metric to be the same, uninfluenced by differing population allele frequencies. It would be even better if the metric is insensitive to the genotyping panel used, allowing for comparisons across studies that are independent of the technology utilized.

To explore the performance of eMOI across varying transmission settings, we simulated 100 samples with MOI drawn from a ZTP distribution with either $\lambda = 1$ or $\lambda = 3$. For each sample, we then simulated either low or high within-host relatedness. For each individual level simulation, we then observed simulated genetics parameterized by each of the 12 regional populations previously described using the 5 genotyping panels, followed by the previously described observation process with false positive and false negative rates of .01 and .1 respectively. We then fit MOIRE on each simulation independently.

For each simulation, we calculated mean eMOI, mean naive MOI, and the within-host infection fixation index (F_{WS}) [1,23], a frequently used metric of within-host diversity that relates genetic diversity of the individual infection to diversity of the parasite population. Mean MOI was calculated using the second-highest number of observed alleles, and F_{WS} used the observed genetics, assuming all alleles were equifrequent within hosts, and naive estimates of allele frequencies to estimate heterozygosity. For these metrics to be most useful in characterizing transmission dynamics, they should be the same for all simulations with the same degree of within-host relatedness and mean MOI, no matter the panel used nor the genetic background of the population. We found that mean eMOI was stable across all genetic backgrounds using microhaplotype based panels, yielding accurate estimates of mean eMOI despite substantial variability in local diversity of alleles, as shown by heterozygosity, and differing genomic loci (Fig. 3C). Interestingly, while the SNP panels exhibited reduced precision and downward bias as expected, they were consistently biased with respect to the true eMOI, even across different panels. This suggests that SNP panels, while limited in resolving power, may still have utility in estimating relative ordering of eMOI. These results also demonstrate that eMOI may be readily used and compared across transmission settings and is relatively insensitive to other factors such as heterozygosity that may vary across settings. In contrast, mean naive MOI and F_{WS} were sensitive to genetic background and genotyping panel in confounded ways. Mean naive MOI, only useful with polyallelic markers, exhibits an inherent upward bias as mean heterozygosity increases that is most severe at higher mean MOI. This bias also varied with the genotyping panel used, making it difficult to interpret and compare across settings (Fig. 3B). F_{WS} is also sensitive to genetic background and panel used, exhibiting an upward trend as heterozygosity increases and a bias that varies across panels. This is inherent to the construction of the metric, as it is coupled to an estimate of the true heterozygosity of genetic loci being used (Fig. 3A). This simulation demonstrates limitations in the utility of F_{WS} as a metric of within-host diversity for a population as it is inherently uncomparable across settings due to its high sensitivity to varying genetic background and genotyping panel used. Mean eMOI, in contrast, is a stable metric of genetic diversity that is insensitive to genetic background and genotyping panel, and is thus readily comparable across settings.

3.4 Application to a study in Northern Namibia

We next used MOIRE to reanalyse data from a previously conducted study carried out in northeastern Namibia consisting of 2585 samples from 29 health facilities across 4 health districts genotyped at 26 microsatellite loci [2]. We ran MOIRE across samples collected from each of the 4 health districts independently. Running MOIRE in this way implies that we are assuming that all samples from each health district come from a shared population with the same allele frequencies. We then calculated summary statistics of interest on the sampled posterior distributions.

We compared our results to the naive estimation conducted in the original study and found that overall relative ordering of mean MOI was maintained, with Andara and Rundu exhibiting the highest MOI, Zambezi the lowest, and Nyangana in between, consistent with contemporary estimates of transmission intensity [2]. However, similar to our simulations, naive estimation substantially underestimated mean MOI across health districts compared to MOIRE (Fig. 4A and C). Individual within-host relatedness was estimated to be very high across sites (IQR: .61-.91) with no differences between sites (Fig. 4B). This suggests substantial inbreeding which may be indicative of persistent local transmission, consistent with the original findings by Tessema et al. We also found that heterozygosity across loci estimated by MOIRE was generally lower (IQR:

.55 - .85), consistent with the previously described simulations demonstrating that naive estimation overestimates heterozygosity, and that previously detected statistically significant differences in heterozygosity between the Zambezi region and the other three regions may have been an artifact of biased estimation (Fig. 4D).

We also ran MOIRE independently across each of the 29 health facilities, excluding 2 health facilities from the Zambezi region due to low total number of samples (n = 9 in each). Stratifying by health facility revealed substantial heterogeneity in mean MOI, within-host relatedness, and consequently eMOI, also consistent with the findings by Tessema et al. (Fig. 5). Interestingly, Tessema et al. identified Rundu district hospital as having exceptionally high within-host diversity as measured by F_{WS}, which was posited to be due to a large fraction of the patients having traveled or resided in Angola. We found that Rundu district hospital had the highest mean eMOI and greatest spread across observations (mean = 4.3 [95% CI: 4.18 - 4.4], IQR = 4.88). This was mainly driven by much higher mean MOI (7 [95% CI: 6.5-7.5]), and low mean within-host relatedness (.47 [95% CI: .43 - 0.51]). The combination of high MOI and relatively low within-host relatedness, translating into high population mean eMOI, suggests that samples collected here reflect a parasite population experiencing less inbreeding and more superinfection, which may be indicative of higher transmission intensity in tandem with a larger effective population size.

4 Discussion

Translating Plasmodium genetic data from naturally acquired infections into meaningful insights about population genetics or malaria transmission dynamics often begins with estimation of allele frequencies and MOI. We demonstrated through simulation that naive estimation introduces substantial biases, rendering estimation unreliable and uncomparable between settings. In particular, naive estimation systematically overestimates measures of allelic diversity such as heterozygosity and systematically underestimates MOI. State-of-the-art methods previously available to more accurately estimate individual level MOI and population allele frequencies only allow for SNP based data, and fail to directly consider within-host relatedness as an important biologic factor [7–9]. MOIRE fills these important gaps, demonstrating both the power and necessity of polyallelic data to obtain precise estimates of these key parameters for understanding of parasite population structure and dynamics. The R package implementing MOIRE provides a user-friendly interface for researchers to easily leverage SNP and polyallelic data to estimate these individual and population diversity metrics which are fundamental for many downstream analyses and often of direct interest themselves.

By estimating within-host relatedness, we also have introduced a new metric of diversity—eMOI—a continuous metric that integrates within-host relatedness and MOI, providing the first portable metric of within-host diversity. This metric is highly identifiable and robust to varying genetic backgrounds, and thus readily comparable across settings and genotyping technologies. We demonstrated that eMOI is a more stable metric of genetic diversity than naive MOI or F_{WS}, and is insensitive to other factors that may vary across settings such as allele frequencies of given genetic markers. Further, by decomposing the genetic state of an infection into components of within-host relatedness and the number of distinct strains present, we have enabled the characterization of these quantities independently, which may be of interest in their own right. For example, within-host relatedness may be of interest in the context of understanding the role of inbreeding and co-transmission in the parasite population [5, 24], and the number of distinct strains may be of interest in the context of understanding superinfection dynamics.

While we have demonstrated the utility of polyallelic data, MOIRE is still compatible with SNP based data and can offer benefits over other approaches. When using SNP based panels, eMOI is still well characterized up to moderate levels, and while the reduced capacity of SNPs generally results in biased estimates, the estimates recovered reflect changes in within-host relatedness yet are stable across genetic backgrounds. Thus, these data may be useful for comparing relative ordering of eMOI across settings and providing inference. In contrast, existing analytical approaches are likely to be sensitive to model misspecification by not considering within-host relatedness and varying genetic backgrounds, and may be biased in ways that are difficult to interpret and compare across settings.

We also note that while increasing the number of loci genotyped is always beneficial, the largest gains in recovering estimates of interest are through using sufficiently diverse loci. Our simulations demonstrate that, even with a modest number of very diverse loci such as our synthetic simulations using 30 loci, eMOI can be recovered with a high accuracy and precision. Marginal increases in complexity of incorporating several highly diverse loci, for example in the context of drug resistance monitoring, may be outweighed by the substantial insights obtained from jointly understanding transmission dynamics, population structure, and drug resistance through increased accuracy of estimating resistance marker allele frequency. Modern amplicon sequencing panels have been developed precisely for these contexts, combining high diversity targets with comprehensive coverage of known resistance markers [13, 14, 22].

MOIRE provides a powerful tool for leveraging polyallelic data to understand malaria epidemiology, and there are multiple avenues for future work to further improve inference. First, the observation model does not currently fully leverage the information in sequencing based data where the actual number of reads may be available. This may provide additional information, e.g. to inform false positive rates by considering the number of reads attributable to an allele, as well as false negative rates by considering the total number of reads at a locus which may be indicative of sample quality. Second, we currently consider only a single, well mixed, background population parameterized by allele frequencies at each locus. However, it may be the case that there are multiple distinct populations with their own allele frequencies, and that the observed data is a mixture of these populations. This may be particularly relevant in the context of malaria transmission where there may be multiple distinct populations of parasites circulating in a region. Future work may consider a mixture model over allele frequencies, where the number of populations is a priori specified or determined through data adaptive non-parametric Bayesian modeling, and thereby identify population substructure. Alternatively, a spatially explicit approach may be feasible that would model the allele frequencies as a function of geographic location, potentially enabling resolving geographic origin of parasites within observed infections. Third, MOIRE currently assumes independence of loci. In the case of locus dependence where there is some amount of linkage disequilibrium, we would expect estimates of allele frequencies and sample specific eMOI to still be consistent if there is not a systematic bias in loci towards regions of high or low within-host relatedness.

In summary, MOIRE enables the use of polyallelic data to estimate allele frequencies, MOI, and within-host relatedness, and provides a new metric of genetic diversity, the eMOI. We have demonstrated that eMOI has improved utility, interpretability, and stability across simulated transmission settings than existing metrics of within-host diversity such as F_{ws}. Furthermore, we demonstrated the utility of MOIRE through simulation and reanalysis of previously collected data, and have provided an R package to enable researchers to easily leverage polyallelic data to make inferences about malaria population dynamics. MOIRE also serves as a fundamental building block for future work, as it provides a principled approach to jointly estimate allele frequencies, MOI, and within-host relatedness from polyallelic data, which can be used as a basis for more complex modeling of population dynamics. These methods may also be of utility for other pathogens where superinfection is common, such as schistosomiasis or filarial diseases [25, 26].

5 Acknowledgements

We thank Nicholas Hathaway for generating and sharing the regional allele frequency data used in this study. This work was supported by the Bill & Melinda Gates Foundation (INV-024346, INV-019043 and INV-035751). This work was also supported by the National Institutes of Health (K24 AI144048).

6 Supplementary Material

6.1 Supplementary Material: Alternative Latent Genetic State Formulation

While the formulation of latent genetic state as a draw from a multinomial distribution that is then masked into a binary vector indicating presence or absence is intuitive, it is a computationally inefficient approach as we must enumerate all possible arrangements that are compatible with the latent genetic state. We can

instead consider an alternative but equivalent formulation. As before, we are interested in calculating the probability of the latent genetic state, $P(Y_{ij}|m = k, \mu_i, \pi_j)$. Let O_{ij} be the set of indices of alleles that are observed in sample i at locus j. For example, if $Y_{ij} = (0, 1, 1, 0)$ then $O_{ij} = \{2, 3\}$. Let $\pi(O_{ij})$ equal the sum of allele frequencies indexed by O_{ij}, and $n = \mu_i - k$, the number of unrelated strains present. We can then write the probability of the latent genetic state as

$$P(Y_{ij}|m = k, \mu_i, \pi_j) = P(O_{ij}|n, \pi(O_{ij}))\pi(O_{ij})^n \tag{9}$$

where $P(O_{ij}|n, \pi(O_{ij}))$ is the probability that each allele indexed by O_{ij} is present at least once after k draws, conditional on that all draws are from the set of alleles indexed by O_{ij}, and $\pi(O_{ij})^n$ is the probability that all n alleles come from the set of alleles indexed by O_{ij}. The probability of every allele being present at least once is equal to 1 minus the probability that any allele is not present, which can be calculated via the inclusion exclusion principle:

$$P(O_{ij}|n, \pi(O_{ij})) = 1 - \sum_{S \subseteq O_{ij}} (-1)^{|S|-1} (1 - \sum_{a \in S} \frac{\pi_{ja}}{\pi(O_{ij})})^n \tag{10}$$

where S is a subset of O_{ij} and $|S|$ is the cardinality of S. This formulation is computationally more efficient, though less intuitive, by allowing us to calculate the probability of the latent genetic state without directly enumerating all possible arrangements of alleles that are compatible with the latent genetic state.

6.2 Supplementary Material: Sampling and Inference

Error Rates

Sample specific false positive and false negative rates were randomly initialized to a value between 0 and 0.1. False positive and false negative rates were constrained to be between 0 and 2, so sampling was conducted after transforming to an unconstrained space to improve efficiency. Proposals were sampled from a normal distribution centered around the transformed current value with a standard deviation that was adapted during burnin to achieve an acceptance rate of 0.23 with proposals accepted according to a Metropolis-Hastings acceptance probability [27, 28]. Priors were uniform between 0 and 1.

Within-host Relatedness

Within-host relatedness was randomly initialized uniformly between 0 and 1 for each sample and constrained to be between 0 and 1, so sampling was conducted after transforming to an unconstrained space to improve efficiency. Proposals were sampled from a normal distribution centered around the transformed current value with a standard deviation that was adapted during burnin to achieve an acceptance rate of 0.23 with proposals accepted according to a Metropolis-Hastings acceptance probability. Priors were uniform between 0 and 1, reflecting our lack of prior knowledge about within-host relatedness.

MOI and Latent Genotypes

Sample specific MOIs were randomly initialized to a value between the maximum number of observed alleles ± 3 without exceeding specified constraints. Sample specific MOIs were constrained to be between 0 and 40. Latent genotypes were initialized to the observed genotype. Proposals for sample MOIs were generated by sampling from a symmetric distribution centered around the current value. Proposals that exceeded the constraints were rejected. Simultaneously, latent genotypes for each locus were sampled by randomly sampling the number of false positives and false negatives present, while ensuring at least one allele being a true positive, and the total number of true positives not exceeding the proposed MOI. The number of false positives and false negatives were sampled from a binomial distribution with the number of trials selected to satisfy the previously described constraints, and the probability of success equal to the false positive and false negative rates at that step, respectively. The alleles that were false positives or false negatives were

randomly selected from the observed genotype. Proposals were accepted according to a Metropolis-Hastings acceptance probability.

Allele Frequencies

Allele frequencies were randomly initialized on the unit simplex. Allele frequencies were then sampled according to the self adjusting logit transform proposal, also known as the SALT sampler [29] and accepted according to a Metropolis-Hastings acceptance probability.

Mean MOI

The mean MOI was randomly initialized to a random draw from the prior distribution. The mean MOI was constrained to be between 0 and 40, so sampling was conducted after transforming to an unconstrained space to improve efficiency. Proposals were sampled from a normal distribution centered around the transformed current value with a standard deviation that was adapted during burnin to achieve an acceptance rate of 0.23 with proposalas accepted according to a Metropolis-Hastings acceptance probability. The hyperprior on the mean MOI was a gamma distribution with shape and scale parameters of 0.1 and 10 respectively, reflecting our assumptions around low mean MOI.

Metropolis Coupled Markov Chain Monte Carlo

It may be the case that the posterior distribution is multimodal, and that the Markov chain may get stuck in a local maxima, leading to poor mixing. To address this, we use Metropolis Coupled Markov Chain Monte Carlo (MC^3), also sometimes referred to as parallel tempering, or replica exchange MCMC sampling [30]. We provide a fully parallelized implementation, allowing for the full utilization of modern high performance computing clusters or multicore desktop machines. We also leverage a non-reversible algorithm with an adaptive temperature gradient as described by [31] to further improve mixing and reduce tuning required by the user.

6.3 Supplementary Material: Effective MOI

Consider an infection with n distinct unrelated strains drawn from the background population. Let X be the observed genotype at a single locus with L possible alleles and population frequencies π, where every allele has a non-zero frequency in the population, and D be the number of distinct alleles observed at the locus

$$D = \sum_{i=1}^{L} \mathbb{I}(X_i = 1) \tag{11}$$

where $X_i = 1$ if allele i is observed. By linearity of expectation, the expected number of distinct alleles (EDA) is

$$\begin{aligned}
\mathbb{E}[D] &= \sum_{i=1}^{L} \mathbb{E}[\mathbb{I}(X_i = 1)] && (12)\\
&= \sum_{i=1}^{L} \mathbb{P}(X_i = 1) && (13)\\
&= \sum_{i=1}^{L} 1 - (1 - \pi_i)^n && (14)\\
& && (15)
\end{aligned}$$

where $(1-\pi_i)^n$ is the probability that allele i is not observed in any of the n strains.

We note that the EDA is a metric that is dependent on MOI and allele frequencies, and is itself an interesting quantity. In the special case of $n = 2$, it is equivalent to the heterozygosity of the locus after subtracting 1. When $n > 2$, the EDA is no longer necessarily equivalent. Unsurprisingly, the EDA for a fixed n and parameterized by allele frequencies is maximized under the same conditions as the heterozygosity (when all alleles are equally likely to be observed). However, loci with differing allele frequencies (in distribution and cardinality) may have the same heterozygosity but different EDAs. This suggests that heterozygosity may be an imperfect metric of diversity in the context of mixed infections. For a fixed MOI, a higher EDA indicates greater information capacity in the locus, and thus higher precision as a tool for estimating parameters. A locus with higher EDA should be preferred when considering loci for genotyping panel development.

We also note that the EDA of a locus may not be strictly larger than the EDA of another locus across MOI, suggesting that genetic loci are more powerful for statistical purposes under limited ranges. Because of this, EDA should be considered in the context of the population distribution of MOI. In principle, it would be possible to marginalize over an estimate of the population distribution of MOI to obtain a generalized estimate of EDA that is weighted by the probability of observing a given MOI, providing a metric that is targeted to the population of interest.

The EDA also provides a natural approach to defining effective MOI. Consider an idealized locus with a heterozygosity of 1, meaning that every allele drawn from the population is unique. Practically, this is an impossibility, however we may approximate such a locus by considering the limit as the number of alleles goes to infinity and occur with equal probability. Let $L \to \infty$ and $\pi_i = \frac{1}{L}$, then the EDA for an infection with n distinct strains is

$$\lim_{L\to\infty} E[D] = \lim_{L\to\infty} \sum_{i=1}^{L} 1 - (1 - \frac{1}{L})^n \tag{16}$$

$$= \lim_{L\to\infty} L - L(\frac{L-1}{L})^n \tag{17}$$

$$= n \tag{18}$$

As expected, in the limit of an infinitely diverse locus, the EDA is equal to the MOI. We now consider the EDA for an infinitely diverse locus in an infection with within-host relatedness r. While MOI remains a fixed quantity, the number of related strains is now a random variable distributed as a binomial with parameters $n-1$ and r. We define the effective MOI (eMOI) as the expected number of distinct alleles observed in an infection with n distinct strains and within-host relatedness r at a locus with infinite diversity. By construction, only unrelated strains contribute to the EDA, thus the eMOI is simply the EDA marginalized over the distribution of unrelated strains in an infection with n distinct strains and within-host relatedness r

$$\text{eMOI} = \sum_{k=0}^{n-1} \binom{n-1}{k} r^k (1-r)^{n-1-k}(n-k) \tag{19}$$

$$= 1 + (1-r)(n-1) \tag{20}$$

As r goes to zero, this quantity approaches the MOI, and as r goes to one, this quantity approaches 1, recovering a natural measure of within-host diversity that is sensitive to both MOI and within-host relatedness.

6.4 Supplementary Material: Regional Populations

In order to assess potential real world performance of MOIRE, we simulated genotyping data from 12 regional populations parameterized by the MalariaGEN Pf7 dataset [19]. Marginal allele frequencies were estimated within each region for each genotyping panel from whole genome sequencing data available from the Pf7 dataset. Alleles within regions with a frequency below 1% were excluded from microhaplotype panels to avoid excessive computational burden when running MOIRE. Uninformative loci (only 1 allele observed) were removed for each region. Interquartile ranges of the number of alleles observed and the number of loci in each panel are summarized in Table S1. Summaries of panel diversity, evaluated by calculating the expected number of distinct alleles from 2 to 10 possible strains, are provided in figure S3. We note that not all unique alleles included in the simulation may be reliably genotypable, e.g. homopolymers and tandem repeats, depending on the fidelity of amplification and sequencing. Simulations may provide an optimistic estimate of performance.

Region	AMPLseq	MaD^{4}HatTeR	AmpliSeq
Central Africa	3-8 (n=121)	4-7 (n=165)	2-10 (n=158)
Central West Africa	3-8 (n=121)	3-7 (n=164)	2-10 (n=155)
East Africa	3-8 (n=122)	4-8 (n=165)	2-11 (n=156)
Papua New Guinea	2-4 (n=115)	2-5 (n=151)	2-6 (n=148)
South Africa	3-9 (n=125)	4-8 (n=165)	2-7 (n=161)
South America - Central	2-3 (n=106)	2-3 (n=141)	2-4 (n=128)
South America - North	2-3 (n=107)	2-3 (n=136)	2-3 (n=114)
South Asia	3-7 (n=121)	3-7 (n=165)	2-9 (n=155)
South East Africa	3-8 (n=119)	4-7 (n=165)	2-11 (n=153)
South East Asia - East	2-5 (n=106)	2-5 (n=137)	2-7 (n=129)
South East Asia - West	2-5 (n=110)	2-4 (n=147)	2-5 (n=148)
West Africa	2-8 (n=122)	3-7 (n=164)	2-11 (n=149)

Table S1. Interquartile range (IQR) of the number of alleles per locus for each region and genotyping panel. Number of loci included in each panel is indicated in parentheses.

References

1. Roh ME, Tessema SK, Murphy M, Nhlabathi N, Mkhonta N, Vilakati S, et al. High Genetic Diversity of Plasmodium Falciparum in the Low-Transmission Setting of the Kingdom of Eswatini. The Journal of Infectious Diseases. 2019;220(8):1346–1354. doi:10.1093/infdis/jiz305.

2. Tessema S, Wesolowski A, Chen A, Murphy M, Wilheim J, Mupiri AR, et al. Using Parasite Genetic and Human Mobility Data to Infer Local and Cross-Border Malaria Connectivity in Southern Africa. eLife. 2019;8:e43510. doi:10.7554/eLife.43510.

3. Pringle JC, Tessema S, Wesolowski A, Chen A, Murphy M, Carpi G, et al. Genetic Evidence of Focal Plasmodium Falciparum Transmission in a Pre-Elimination Setting in Southern Province, Zambia. The Journal of Infectious Diseases. 2019;219(8):1254–1263. doi:10.1093/infdis/jiy640.

4. Wong W, Griggs AD, Daniels RF, Schaffner SF, Ndiaye D, Bei AK, et al. Genetic Relatedness Analysis Reveals the Cotransmission of Genetically Related Plasmodium Falciparum Parasites in Thiès, Senegal. Genome Medicine. 2017;9(1):5. doi:10.1186/s13073-017-0398-0.

5. Nkhoma S, Trevino SG, Gorena KM, Nair S, Khoswe S, Jett C, et al. Co-Transmission of Related Malaria Parasite Lineages Shapes Within-Host Parasite Diversity. Cell Host & Microbe. 2020;27(1). doi:10.1016/j.chom.2019.12.001.

6. Wong W, Wenger EA, Hartl DL, Wirth DF. Modeling the Genetic Relatedness of Plasmodium Falciparum Parasites Following Meiotic Recombination and Cotransmission. PLoS Computational Biology. 2018;14(1):e1005923. doi:10.1371/journal.pcbi.1005923.

7. Chang HH, Worby CJ, Yeka A, Nankabirwa J, Kamya MR, Staedke SG, et al. THE REAL McCOIL: A Method for the Concurrent Estimation of the Complexity of Infection and SNP Allele Frequency for Malaria Parasites. PLOS Computational Biology. 2017;13(1):e1005348. doi:10.1371/journal.pcbi.1005348.

8. Paschalidis A, Watson OJ, Aydemir O, Verity R, Bailey JA. Coiaf: Directly Estimating Complexity of Infection with Allele Frequencies. PLOS Computational Biology. 2023;19(6):e1010247. doi:10.1371/journal.pcbi.1010247.

9. Ju NP, Liu J, He Q. SNP-Slice: A Bayesian Nonparametric Framework to Resolve SNP Haplotypes in Mixed Infections; 2023.

10. Anderson TJC, Haubold B, Williams JT, Estrada-Franco§ JG, Richardson L, Mollinedo R, et al. Microsatellite Markers Reveal a Spectrum of Population Structures in the Malaria Parasite Plasmodium Falciparum. Molecular Biology and Evolution. 2000;17(10):1467–1482. doi:10.1093/oxfordjournals.molbev.a026247.

11. McVean G, Hendry JA, Almagro-Garcia J, Almagro-Garcia J, Pearson RD, Amato R, et al. The Origins and Relatedness Structure of Mixed Infections Vary with Local Prevalence of P. Falciparum Malaria. eLife. 2019;8. doi:10.7554/elife.40845.

12. Tessema SK, Hathaway NJ, Teyssier NB, Murphy M, Chen A, Aydemir O, et al. Sensitive, Highly Multiplexed Sequencing of Microhaplotypes from the Plasmodium Falciparum Heterozygome. The Journal of Infectious Diseases. 2022;225(7):1227–1237. doi:10.1093/infdis/jiaa527.

13. LaVerriere E, Schwabl P, Philipp Schwabl, Carrasquilla M, Aimee R Taylor, Zachary M Johnson, et al. Design and Implementation of Multiplexed Amplicon Sequencing Panels to Serve Genomic Epidemiology of Infectious Disease: A Malaria Case Study. Molecular Ecology Resources. 2022;doi:10.1111/1755-0998.13622.

14. Kattenberg JH, Fernandez-Miñope C, van Dijk NJ, Llacsahuanga Allcca L, Guetens P, Valdivia HO, et al. Malaria Molecular Surveillance in the Peruvian Amazon with a Novel Highly Multiplexed Plasmodium Falciparum AmpliSeq Assay. Microbiology Spectrum. 2023;11(2):e00960–22. doi:10.1128/spectrum.00960-22.

15. Taylor AR, Jacob PE, Neafsey DE, Buckee CO. Estimating Relatedness Between Malaria Parasites. Genetics. 2019;212(4):1337–1351. doi:10.1534/genetics.119.302120.

16. Inna Gerlovina, Boris Gerlovin, Isabel Rodríguez-Barraquer, Bryan Greenhouse. Dcifer: An IBD-based Method to Calculate Genetic Distance between Polyclonal Infections. Genetics. 2022;doi:10.1093/genetics/iyac126.

17. Nkhoma SC, Nair S, Cheeseman IH, Rohr-Allegrini C, Singlam S, Nosten F, et al. Close Kinship within Multiple-Genotype Malaria Parasite Infections. Proceedings of the Royal Society B: Biological Sciences. 2012;279(1738):2589–2598. doi:10.1098/rspb.2012.0113.

18. Eddelbuettel D, Francois R. Rcpp: Seamless R and C++ Integration. Journal of Statistical Software. 2011;40:1–18. doi:10.18637/jss.v040.i08.

19. MalariaGEN, Abdel Hamid MM, Abdelraheem MH, Acheampong DO, Ahouidi A, Ali M, et al. Pf7: An Open Dataset of Plasmodium Falciparum Genome Variation in 20,000 Worldwide Samples. Wellcome Open Research. 2023;8:22. doi:10.12688/wellcomeopenres.18681.1.

20. Daniels R, Volkman SK, Milner DA, Mahesh N, Neafsey DE, Park DJ, et al. A General SNP-based Molecular Barcode for Plasmodium Falciparum Identification and Tracking. Malaria Journal. 2008;7(1):223. doi:10.1186/1475-2875-7-223.

21. Chang HH, Wesolowski A, Sinha I, Jacob CG, Mahmud A, Uddin D, et al.. Mapping Imported Malaria in Bangladesh Using Parasite Genetic and Human Mobility Data; 2019. https://elifesciences.org/articles/43481/figures.

22. Aranda-Diaz A, Neubauer Vickers E. MAD4HatTeR. protocolsio. 2022;doi:10.17504/protocols.io.14egn779mv5d/v3.

23. Manske M, Miotto O, Campino S, Auburn S, Almagro-Garcia J, Maslen G, et al. Analysis of Plasmodium Falciparum Diversity in Natural Infections by Deep Sequencing. Nature. 2012;487(7407):375–379. doi:10.1038/nature11174.

24. Wong W, Volkman S, Daniels R, Schaffner S, Sy M, Ndiaye YD, et al. RH: A Genetic Metric for Measuring Intrahost Plasmodium Falciparum Relatedness and Distinguishing Cotransmission from Superinfection. PNAS Nexus. 2022;1(4):pgac187. doi:10.1093/pnasnexus/pgac187.

25. Aemero M, Boissier J, Climent D, Moné H, Mouahid G, Berhe N, et al. Genetic Diversity, Multiplicity of Infection and Population Structure of Schistosoma Mansoni Isolates from Human Hosts in Ethiopia. BMC Genetics. 2015;16:137. doi:10.1186/s12863-015-0297-6.

26. Hedtke SM, Kuesel AC, Crawford KE, Graves PM, Boussinesq M, Lau CL, et al. Genomic Epidemiology in Filarial Nematodes: Transforming the Basis for Elimination Program Decisions. Frontiers in Genetics. 2020;10.

27. Gelman A, Gilks WR, Roberts GO. Weak Convergence and Optimal Scaling of Random Walk Metropolis Algorithms. The Annals of Applied Probability. 1997;7(1):110–120. doi:10.1214/aoap/1034625254.

28. Chib S, Greenberg E. Understanding the Metropolis-Hastings Algorithm. The American Statistician. 1995;49(4):327–335. doi:10.2307/2684568.

29. Director HM, Gattiker J, Lawrence E, Vander Wiel S. Efficient Sampling on the Simplex with a Self-Adjusting Logit Transform Proposal. Journal of Statistical Computation and Simulation. 2017;87(18):3521–3536. doi:10.1080/00949655.2017.1376063.

30. Earl DJ, Deem MW. Parallel Tempering: Theory, Applications, and New Perspectives. Physical Chemistry Chemical Physics. 2005;7(23):3910–3916. doi:10.1039/B509983H.

31. Syed S, Bouchard-Côté A, Deligiannidis G, Doucet A. Non-Reversible Parallel Tempering: A Scalable Highly Parallel MCMC Scheme. Journal of the Royal Statistical Society Series B: Statistical Methodology. 2022;84(2):321–350. doi:10.1111/rssb.12464.

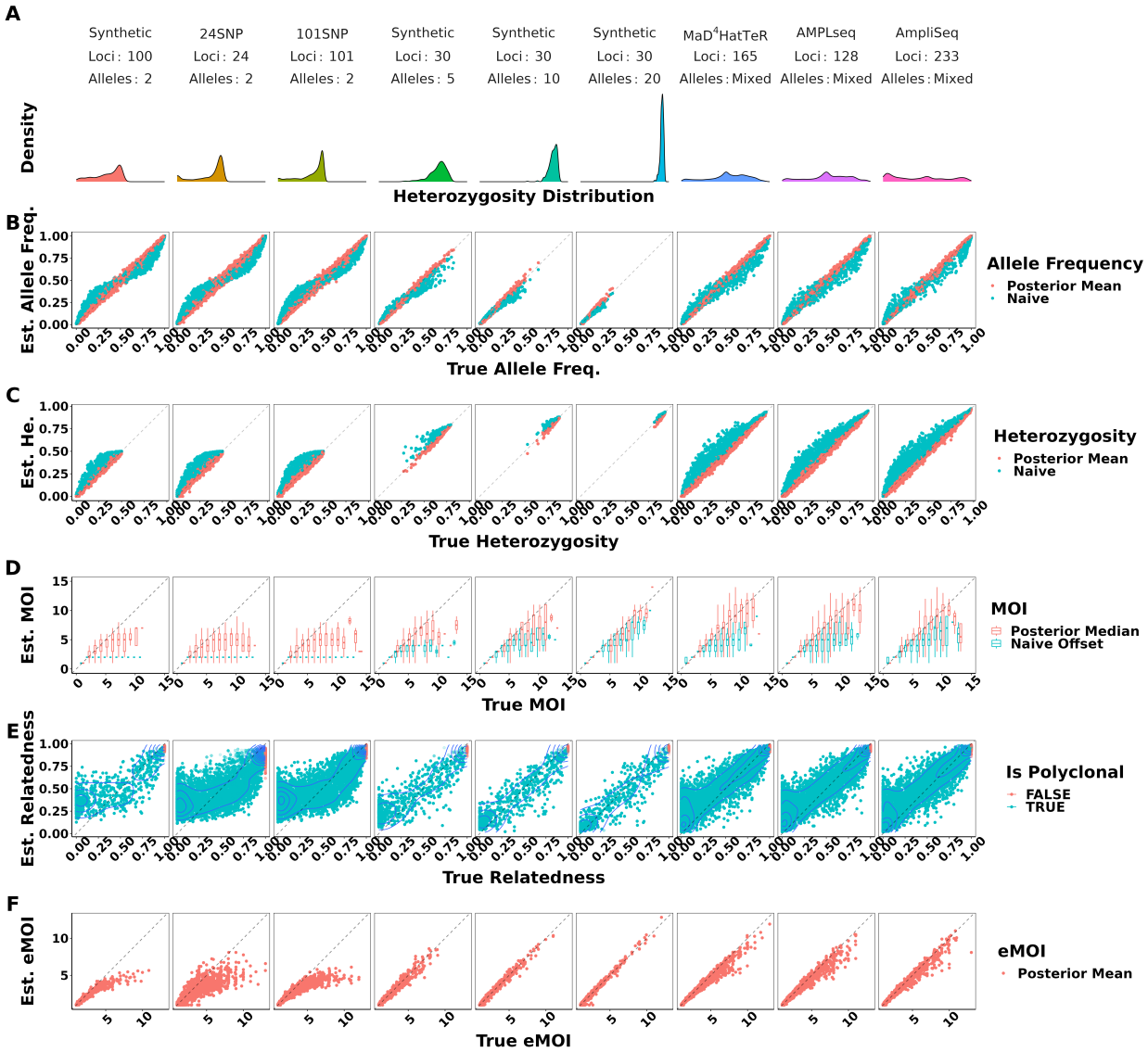

Figure 1. True vs. estimated values of parameters across panels of varying genetic diversity. The top panel summarizes the distribution of heterozygosity across each panel used. Each symbol represents the estimated value of the parameter for a single simulated dataset, with the true value of the parameter on the x-axis and the estimated value on the y-axis. Simulations were pooled across mean MOIs and levels of relatedness. False positive and false negatives rates were fixed to 0.01 and 0.1 respectively. Opacity was set to accommodate overplotting, except in the case of within-host relatedness where opacity reflects the estimated probability that a sample is polyclonal, calculated as the posterior probability of the sample MOI being greater than 1. MOIRE accurately recovered parameters of interest with increasing accuracy as panel diversity increased, while naive estimation exhibited substantial bias where such estimators exist.

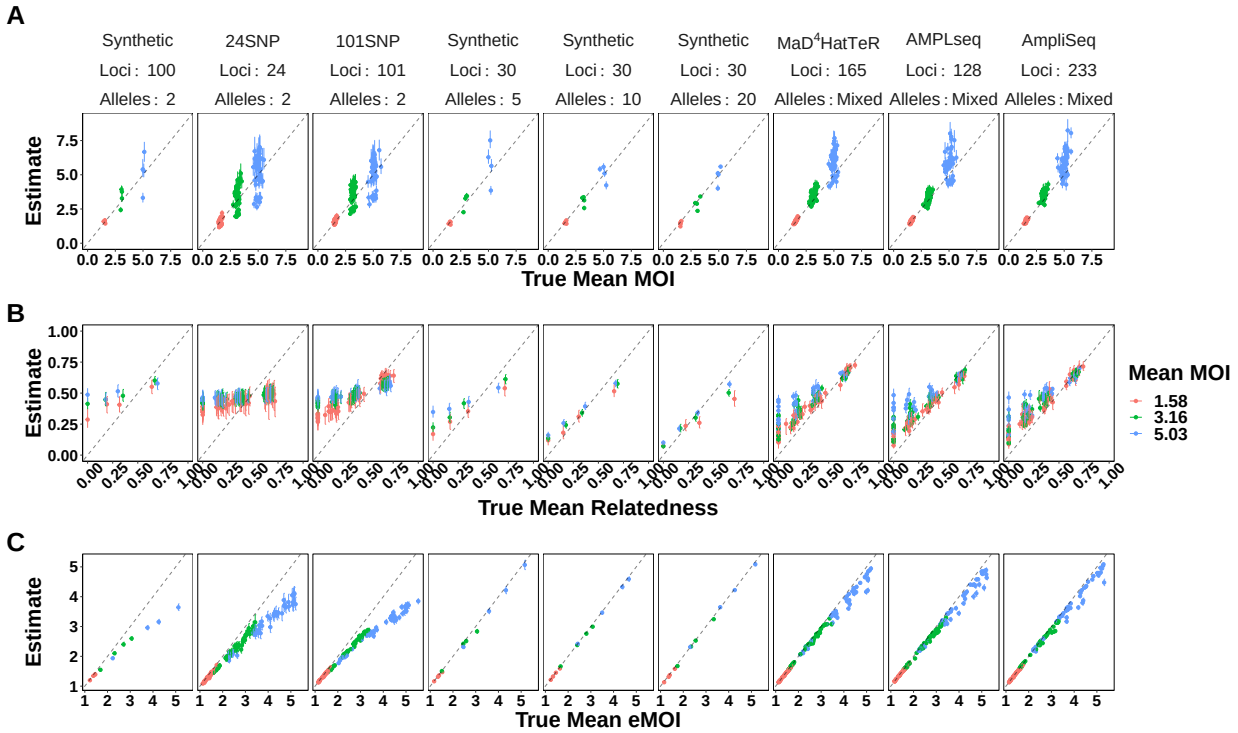

Figure 2. True vs. estimated values of population level parameters across panels of varying genetic diversity. Each symbol represents the estimated value of the population summary parameter for a single simulated dataset, with the true value of the parameter on the x-axis and the estimated value on the y-axis. False positive and false negatives rates were fixed to 0.01 and 0.1 respectively. MOIRE accurately recovered population mean MOI and eMOI, as well as mean relatedness when relatively high. Overestimation of mean relatedness at low true values did not result in a significant bias in eMOI when sufficiently diverse genetic markers were used.

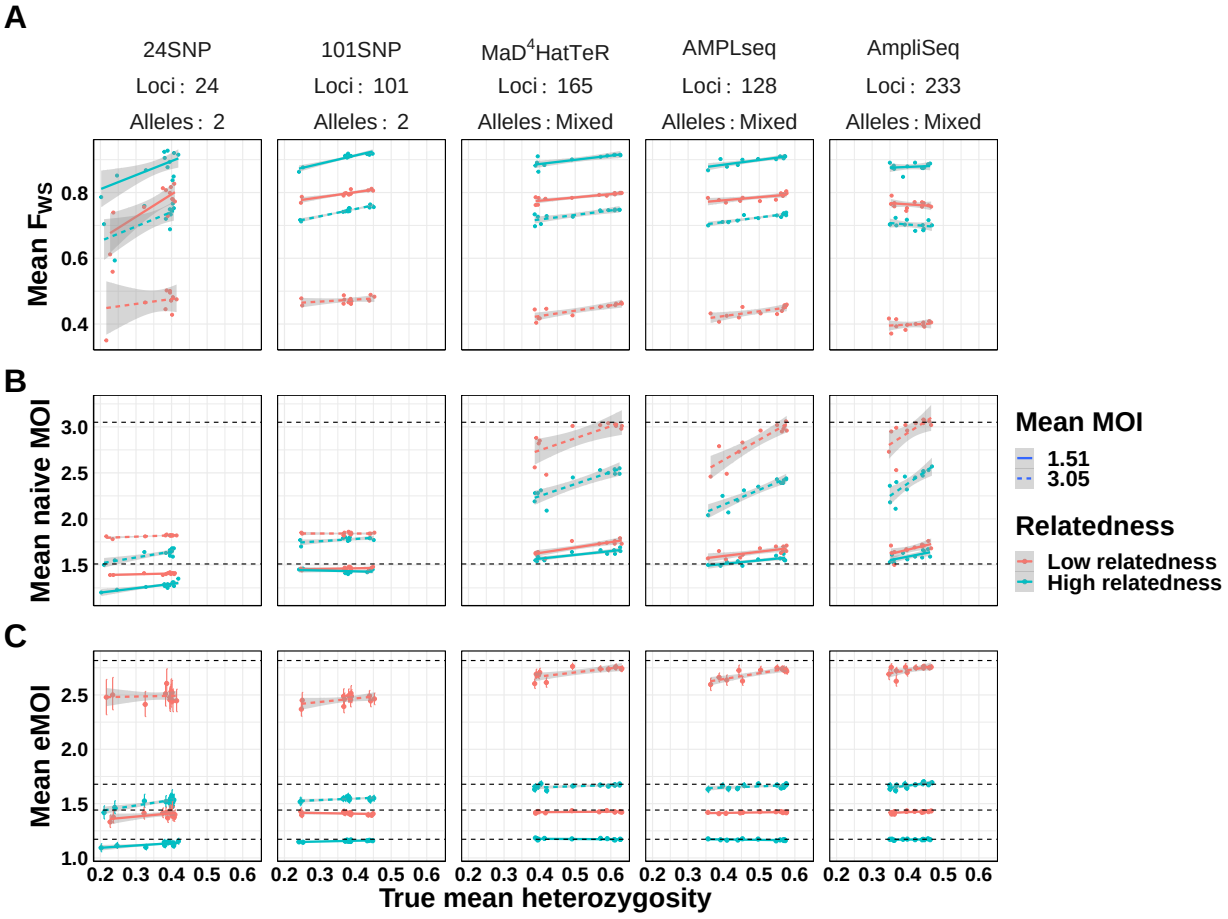

Figure 3. Comparison of mean eMOI to other summary measures of diversity across varying levels of within-host relatedness. For each level of relatedness (low and high), we simulated 100 infections with a mean MOI of 1.51 and 3.16, for a total of 400 infections across 4 conditions. Keeping the MOI and relatedness fixed for each sample, we varied the genetic diversity of the panel used to genotype each sample. We then calculated the mean eMOI from MOIRE, mean MOI using the naive estimator, and mean F_{WS} using a naive estimate of allele frequencies for each simulation to assess the sensitivity of each metric to varying the genetic diversity of the panel. True mean eMOI and mean MOI are fixed values within levels of within-host relatedness and are annotated by dashed lines. Mean F_{WS} is not fixed within levels of within-host relatedness and MOI because it is a function of the genetic diversity of the panel.

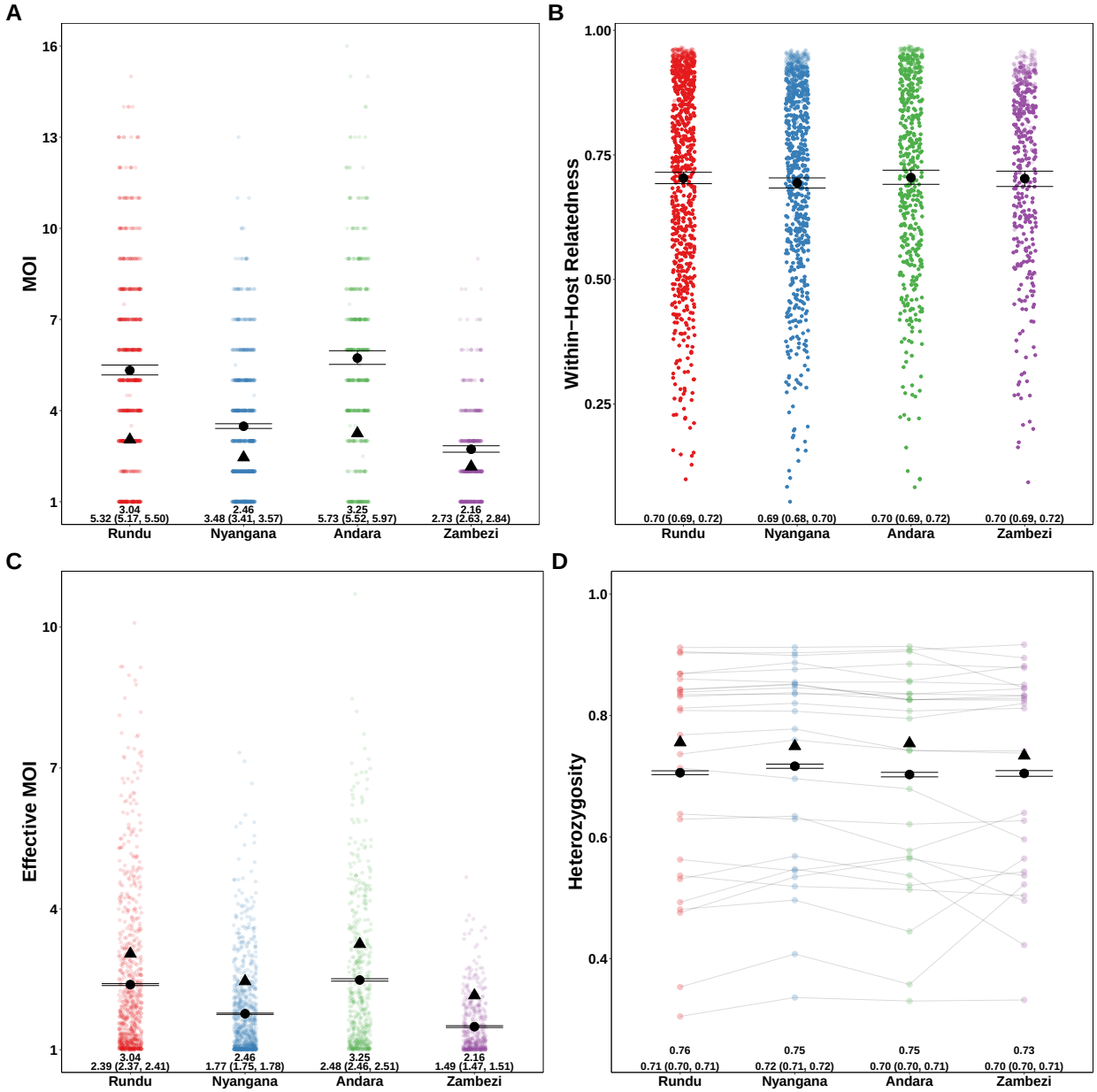

Figure 4. Estimated MOI, relatedness, eMOI and heterozygosity in Northern Namibia. MOIRE was run on data from 2585 samples from 29 clinics genotyped at 26 microsatellite loci, subset across four health districts. Each point represents the posterior mean or median for each sample or locus level parameter. The black circle represents the population mean with 95% credible interval for each health district and the black triangle indicates the naive estimate where applicable. In the case of eMOI (C), the naive estimate is simply the MOI. Opacity was used to accommodate overplotting in A, C and D, however opacity in B is reflective of the posterior probability of a particular sample being polyclonal. This is due to the fact that mean within-host relatedness is only defined for samples with MOI greater than 1, and thus the posterior distribution of within-host relatedness was calculated by taking the mean within-host relatedness across samples with MOI greater than 1 at each iteration of the MCMC algorithm. Therefore, the opacity of each point in B is reflective of the contribution of that sample to the posterior distribution of mean within-host relatedness.

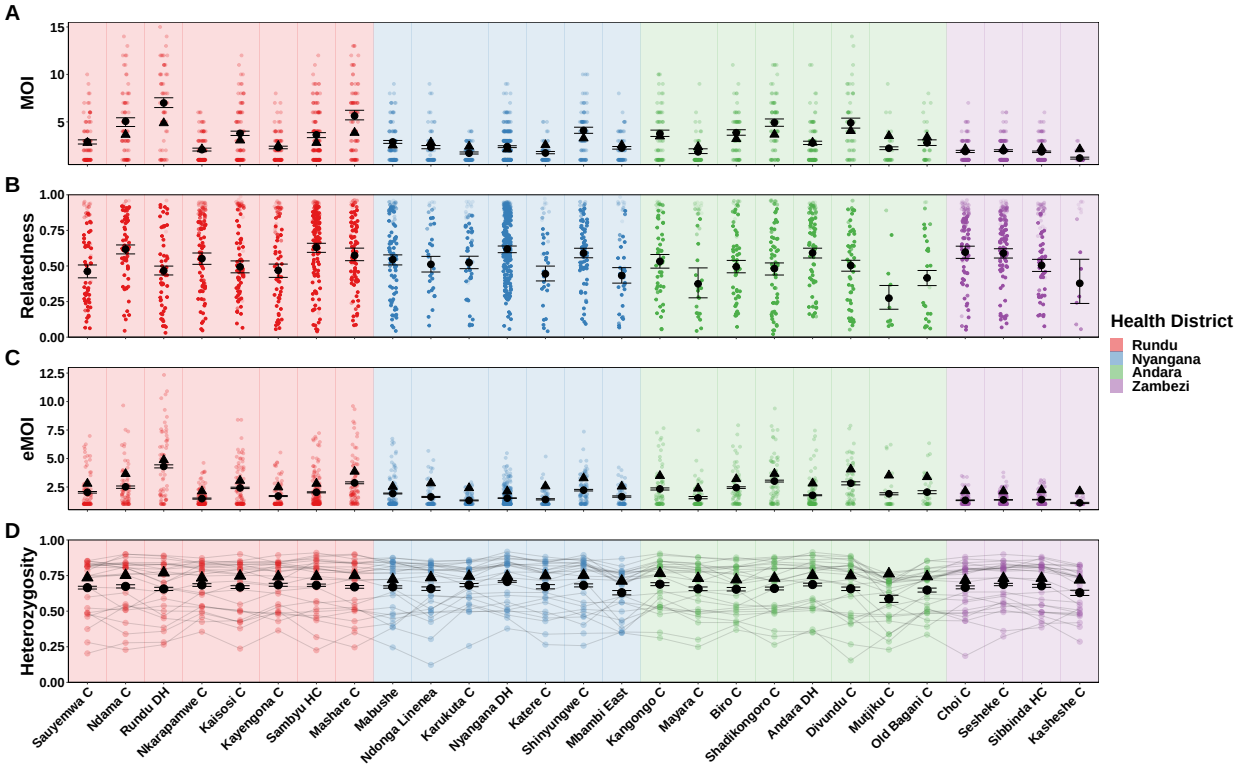

Figure 5. Estimated MOI, relatedness, eMOI and heterozygosity in Northern Namibia, stratified by health facility. MOIRE was run independently on data from each health facility. Two health facilities from the Zambezi region were excluded due to only having 9 samples present in each subset. Health facilities are plotted in geographic order from West to East. Plotting conventions are the same as in Figure 4.

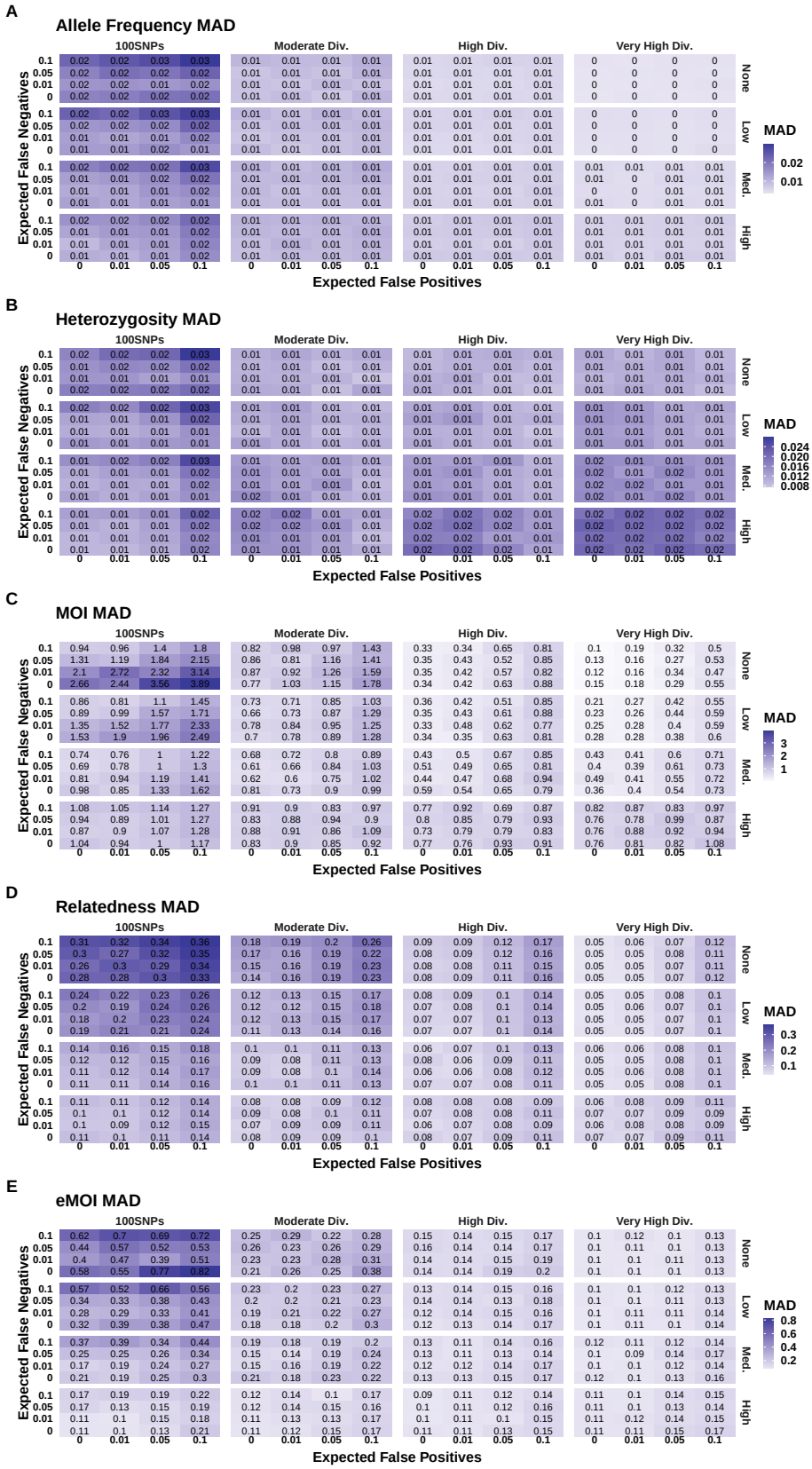

Figure S1. Mean absolute deviation (MAD) of parameter estimates by MOIRE across panels of varying genetic diversity and stratified by population levels of within-host relatedness.

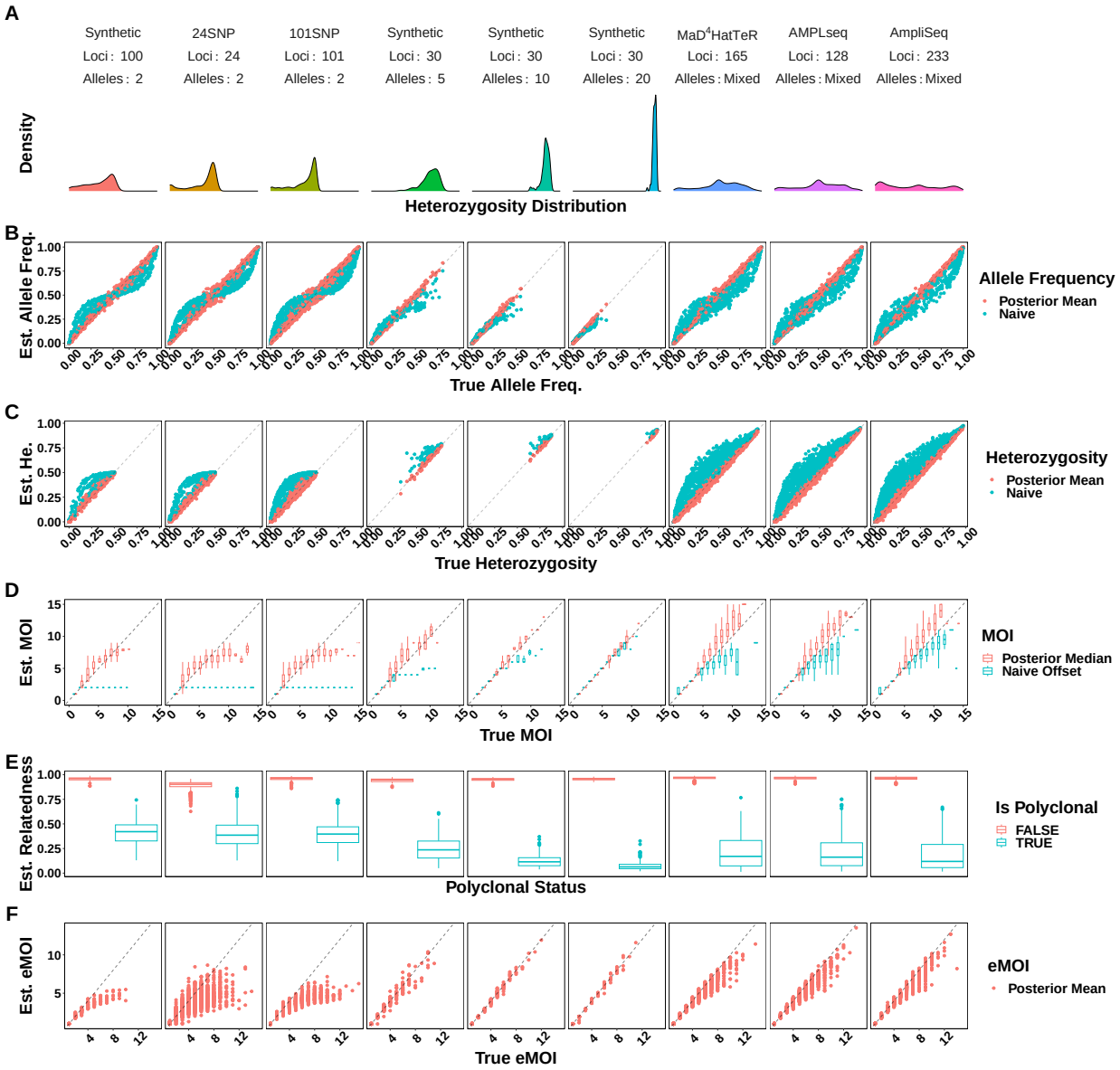

Figure S2. Attempting to estimate relatedness in the absence of relatedness did not introduce bias into parameters of interest such as MOI, heterozygosity, and allele frequencies. Simulations were pooled across mean MOIs. False positive and false negatives rates were fixed to 0.01 and 0.1 respectively.

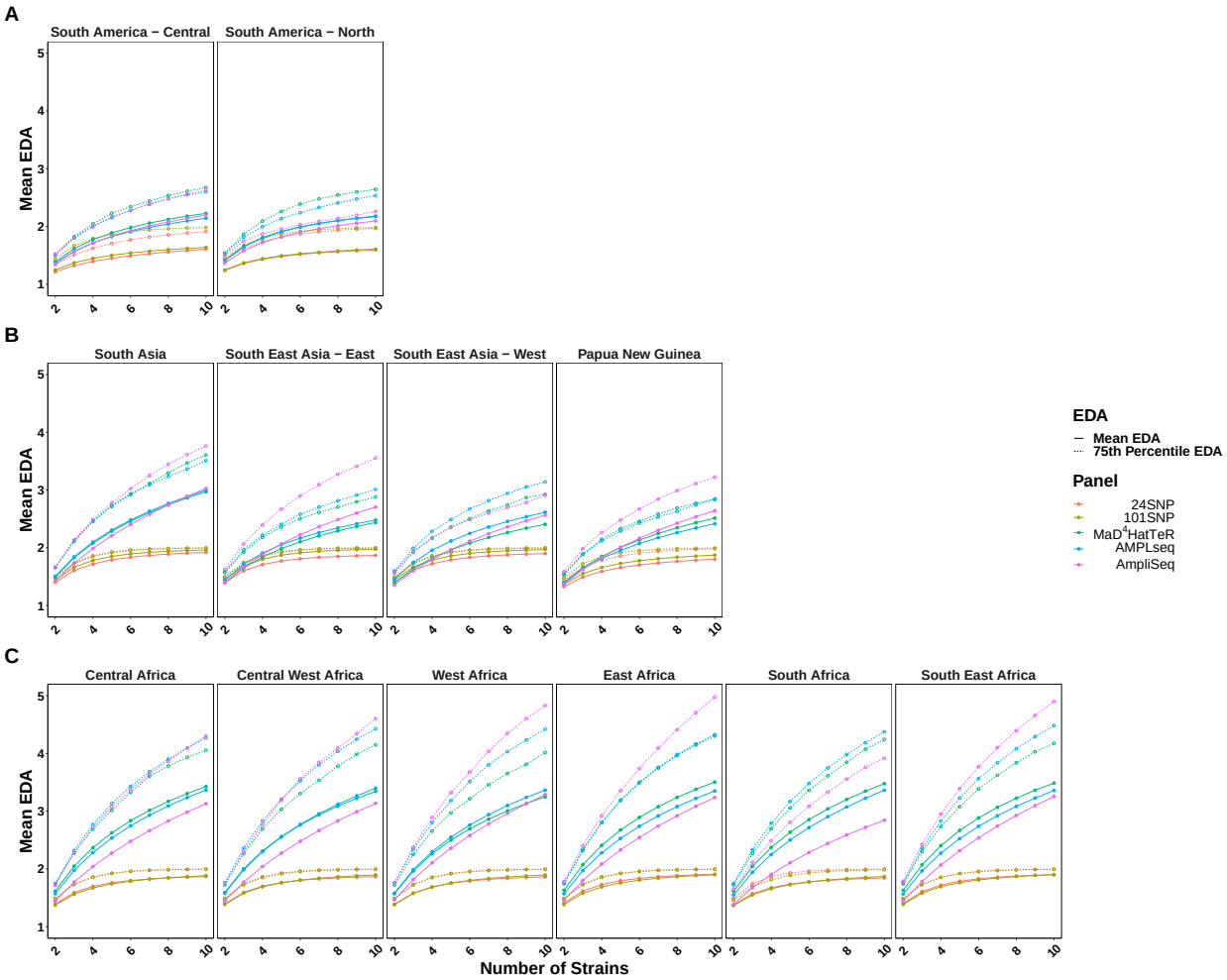

Figure S3. Mean EDA across 12 regional populations for 5 selected genotyping panels. Marginal allele frequencies within each regional population were estimated from the MalariaGEN Pf7 dataset. We then evaluated the expected number of distinct alleles for each locus within each regional population for each genotyping panel. The solid lines indicate the mean EDA across loci within each regional population, and the dashed lines indicate the 75th percentile. We note that heterozygosity for each panel is equal to the EDA minus 1 when the number of strains is 2.