Classification of outcomes in antimalarial therapeutic efficacy studies with Aster

Inna Gerlovina[1,*], Sophie Berube[2], Jessica Briggs[1], Kathryn Murie[1], Maxwell Murphy[1], Amy Wesolowski[3], and Bryan Greenhouse[1]

[1]EPPIcenter research program, Division of HIV, ID and Global Medicine, Department of Medicine, University of California, San Francisco, CA, USA
[2]Dept of Biostatistics, University of Florida, Gainesville, FL, USA
[3]Dept of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA
[*]Corresponding author: inna.gerlovina@ucsf.edu

Abstract

Reliable assessment of antimalarial drug efficacy is crucial for effective response to emerging drug resistance, and therapeutic efficacy studies (TES) are the primary means of estimating *in vivo* efficacy. Accuracy of such estimates rests on correctly classifying recurrent infections developed during follow-up as recrudescences or new infections. Genotyping is used to guide classification, but polyclonal infections and alleles matching by chance make classification challenging, especially in high transmission settings. Match-counting algorithms currently recommended by World Health Organization are unreliable and produce biased results, necessitating development of principled statistical approaches. Modern genotyping methods such as multiplexed amplicon sequencing hold great potential for resolving recurrences and motivate the need for corresponding statistical methods able to utilize the rich data they provide. We propose an Adaptive Statistical framework for Therapeutic Efficacy and Recrudescence (Aster) that delivers accurate and consistent results by explicitly incorporating complexity of infection (COI), population allele frequencies, and imperfect detection of alleles in minority strains. Using an identity by descent approach, Aster accounts for alleles matching by chance and for a background infection relatedness structure that can otherwise lead to misclassification. The flexible framework can also use external information, such as parasite density and performance characteristics of a genotyping panel. Using simulations, we show that Aster dramatically outperforms match-counting algorithms in a wide variety of transmission settings and demonstrates consistently balanced performance that improves with more informative genotyping panels. Aster is implemented in a fast, fully scalable, and user-friendly R software package *asterTES* and provides accurate estimates of treatment failure for TES with any type of genotyping data, facilitating reliable evaluation of drug efficacy and effective management of malaria.

1 Introduction

Widespread availability of highly effective artemisinin combination therapies (ACTs) has greatly improved treatment for *Plasmodium falciparum* infections, the most fatal cause of malaria, and has contributed to a global reduction in malaria burden over the last few decades [1]. Unfortunately, progress in malaria control has recently stalled, with over 600,000 malaria deaths estimated in 2023 [2]. One of the largest threats to malaria control right now is the emergence of artemisinin partial resistance in sub-Saharan Africa, where $> 95\%$ of malaria cases and deaths occur [3, 4]. Treatment with a failing drug often results in partial clearance of infecting parasites, which then recrudesce,

resulting in increased morbidity, mortality, and ongoing transmission of resistant parasites. Antimalarial efficacy is primarily evaluated through therapeutic efficacy studies (TES). A TES assesses *in vivo* efficacy by enrolling patients with symptomatic malaria infections, administering directly observed drug therapy, and actively following these individuals to assess parasite clearance. Since recrudescence often occurs several weeks after therapy, extended follow-up times are required to assess for recurrent parasitemia. Importantly, the World Health Organization (WHO) recommends changing therapy if genotype-corrected failure (recrudescence) rates exceed 10%, making TES results a vital source of information to determine treatment policy [5].

In malaria endemic areas, TES participants can be reinfected during the long follow-up period. Measuring efficacy thus requires determining whether detection of asexual parasites during follow-up is due to an entirely new infection, which is not an indication of drug failure, or a recrudescence (possibly in combination with a new infection), suggesting true drug failure [5, 6]. In areas with high malaria transmission, including most of sub-Saharan Africa where artemisinin partial resistance is now emerging, people can be infected hundreds of times per year, and over 50% of study participants can develop new infections during follow-up [7, 8]. Results of TES in these areas – and by extension treatment policy – can therefore be dramatically affected by the accuracy of outcome classification. [9, 10, 11, 12]. The genotypes of parasites present in the blood prior to treatment (day zero [D0]) are compared with those present when recurrent parasitemia is first detected by microscopy (day of recurrence [DR]); matching parasite genotypes between D0 and DR suggest recrudescence, while the DR genotypes that are different from the ones on D0 suggest that a new infection has occurred [6]. Since *P. falciparum* populations in sub-Saharan Africa exhibit high levels of genetic diversity, it is unlikely that an individual will, by chance, be reinfected with a genetically identical parasite [6, 13, 14] .

In practice, biological and technical factors complicate the comparison, compromising the accuracy of TES results [4, 15]. *P. falciparum* infections frequently contain multiple genetically distinct strains, i.e. have a complexity of infection (COI) greater than 1 [9, 16]. While genotyping can usually detect multiple alleles at each locus, it is not currently feasible to assign alleles across loci to underlying strains (i.e. the genotyping data are unphased). Since a recurrent infection would constitute a recrudescence event if it contains at least one persistent strain, regardless of the presence or absence of newly infecting strains, any matching alleles between D0 and DR at a given locus provide potential evidence of recrudescence. However, alleles from genetically distinct parasites can also match by chance, potentially resulting in misclassifying a new infection (i.e. an infection with no persistent pre-treatment strains) as a recrudescence; polyclonality exacerbates the issue by increasing the chance of matching [9, 10]. Technical issues pose additional challenges for classification: misidentification of genotyping artifacts as additional alleles (i.e. false positive alleles) can produce non-existent matches, while alleles from strains at low frequencies in the blood might go undetected (i.e. false negative alleles), potentially masking matches that are due to recrudescence [17, 18, 12].

To address these challenges, the WHO recommends genotyping 3 diverse loci and assessing allele matches between D0 and DR samples [19]. If these genotypes have at least one matching allele at all 3 loci, the recurrence is considered a recrudescence ("3/3 algorithm") [20], otherwise it is classified as a new infection. With this algorithm, however, undetected alleles in recrudescent strains could result in misclassifying such recurrences as new infections. To allow for imperfect sensitivity of genotyping methods, the "2/3 algorithm" has been proposed as an alternative, requiring only 2 of the 3 loci to have matching alleles. These match-counting algorithms are easy to implement, but they do not explicitly account for factors that can have a large effect on the presence of matching

alleles not due to recrudescence, such as allele frequency, COI, and underlying relatedness structure in the local parasite population. Resulting misclassifications lead to estimates of drug failure that may be considerably biased by factors such as transmission intensity or the genotyping method used [10, 15, 11, 4].

In light of the limitations of match-counting algorithms, the need for statistical approaches to estimate therapeutic failure rate have become apparent. Such approaches can account for biological and technical factors in a principled way, reducing bias and providing measures of uncertainty. The first likelihood-based method, CDC Bayesian algorithm [12], uses a Monte Carlo Markov chain approach and provides recurrence classification and posterior probability of recrudescence. More recently, another likelihood-based method, *PfRecur*, has been proposed [21]; it provides a posterior probability of a recrudescence event and a posterior expected proportion of recrudescent clones. Both methods are designed for genotyping data based on length polymorphisms and account for incorrectly identified and undetected alleles, assigning the same probability of detection to every allele [12] or parasite clone [21]. A different, nonparametric statistical approach [22] uses a null distribution of a genetic dissimilarity statistic between two non-recrudescent infections, approximated by empirical distribution of pairwise dissimilarity between D0 infections, to classify recurrences. This method can be used with a dissimilarity measure of the user's preference.

Continuing development of amplicon sequencing methods, which likely have advantages over length polymorphisms in accuracy, reproducibility, accessibility, throughput, and cost [23, 24, 25, 26, 27, 28] offers potential for more reliable assessment of therapeutic failure rates. Higher numbers of genotyped loci and more reliable detection of alleles from minority strains in complex infections can improve the accuracy of estimation. Increasing generation of amplicon sequencing data motivates the need for corresponding statistically rigorous approaches that are flexible, fully scalable, able to utilize various types of information, and are easy to use. Here, we propose a novel statistical approach, Adaptive Statistical framework for Therapeutic Efficacy and Recrudescence (Aster) that performs recurrence classification along with failure rate estimation, providing inference based on an empirical Bayes approach. Aster accounts for population allele frequencies, COI, imperfect allele detection, and an underlying relatedness structure in the local parasite population (here referred to as "background relatedness"). It is further able to incorporate external information and adapt to developing genotyping methodology. The model approaches recrudescence through the concept of identity by descent (IBD), which allows us to track recrudescent and minor clones across loci. It also provides a natural way to account for the possibility of being reinfected with clones that are genetically related to the ones in a D0 sample. Aster is implemented in a software package *asterTES* [29], which is fast, numerically stable, and user-friendly, using efficient combinatorial algorithms to process unphased data. Like the framework, it is easily extendable to include new features and accomodate new inputs.

First we present the overall framework, then evaluate Aster performance against match-counting algorithms across a variety of genotyping methods and epidemiological settings using simulations. We conclude by discussing practical applications of the method.

2 Methods

2.1 Framework

Aster is designed to infer drug failure rates in antimalarial therapeutic efficacy studies. The adaptive framework consists of two main components: 1) the genetic data generating model and 2) the detection model. We assume independence between the corresponding random processes. Each of the models can be adapted to different settings and can utilize external information, allowing for flexibility in practical applications. Overlaying these models is an observation mechanism, which reflects the fact that phased genotype data are not usually directly observed; this mechanism is represented by a deterministic function. At the top level, the framework addresses recrudescence, allowing both strain- and person-specific definitions. Here we present the latter, most common approach where recrudescence is defined as an event where at least one strain from the D0 infection is recrudescent. In this case, therapeutic failure rate is the probability that an individual has a recrudescence event following therapy. First, we describe the framework components and observation mechanism for a single sample, not considering recrudescence, then expand them to a pair of samples from the same individual. The relationship between recrudescence and identity by descent (IBD) is a part of the genetic model and is described in Section 2.1.2.

2.1.1 Single sample

Genetic data generating model

Let $\mathbf{X}$ be a $n \times L$ matrix of random variables, where n is COI, L is the number of loci, rows $X_{i\cdot} = (X_{i1}, \ldots, X_{iL})$ represent distinct parasite genotypes, and columns $X_{\cdot l} = (X_{1l}, \ldots, X_{nl})$ represent alleles at each locus (Figure 1). $\mathbf{X}$ can be thought of as phased genotype data for a single infection. Each X_{il} has a categorical distribution with values in a set $A_l = \{a_{l1}, \ldots, a_{lK_l}\}$ of possible alleles at locus l and corresponding probabilities - population allele frequencies $\pi(a_{lk})$. X_{il} are assumed to be independent, which implies no linkage disequilibrium, perfect mixing of parasite strains, and no intrahost relatedness.

$$P(\mathbf{X}) = \prod_{i=1}^{n} \prod_{l=1}^{L} \pi(X_{il})$$

Detection model

In a polyclonal infection, parasite strains that are present in a sample in smaller proportions (minor strains) might be harder to detect during genotyping, and for these strains, alleles at some loci might not be detected. Let $\mathbf{D}$ be a binary $n \times L$ matrix that represents detection: if $D_{il} = 1$, an allele in strain $X_{i\cdot}$ at locus l has been detected, and if $D_{il} = 0$, it has not been (Figure 1). D_{il} are independent random variables, $D_{il} \sim Bernoulli(p_{il})$, where p_{il} is a probability of detection.

$$P(\mathbf{D}) = \prod_{i=1}^{n} \prod_{l=1}^{L} p_{il}^{D_{il}} \left(1 - p_{il}\right)^{1-D_{il}}$$

In addition to strain proportion, p_{il} can depend on panel characteristics, overall parasite density, COI, and other factors. In many cases, it is reasonable to assume the same probability for alleles from a strain to be detected at any locus ($p_{i1} = \ldots = p_{iL}$), but it is possible that some markers can be more sensitive, in which case these probabilities will vary from locus to locus. Major strains

can be assumed to be fully detected ($p_{jl} = 1\ \forall\, l$ for some major strain j). There might also be completely undetected strains that are not present in the sample; this special case of $p_{kl} = 0\ \forall\, l$ for an undetected strain k can be important for classification if that strain is also recrudescent. While the Aster framework formally allows such a scenario, incorporating it would require careful consideration (and possibly extending the framework) to avoid classification bias and/or extremely high uncertainty (see Supplementary Section S.4).

Observation mechanism

When genetic data for polyclonal infections are unphased, observed data consist of unordered lists of unique alleles detected at each locus. Let $\boldsymbol{U} = (U_1, \ldots, U_L)$ be a sequence of sets that represents the observed data. $\boldsymbol{U}$ is fully determined by matrices $\mathbf{X}$ and $\mathbf{D}$, and the observation mechanism can be represented by a deterministic function G_u that maps haplotype and detection matrices to a sequence of sets of detected unique alleles (Figure 1). To define G_u, we first consider a single locus. Let $g_u\big((a_i)_{i=1}^n, (b_i)_{i=1}^n)\big)$ be a function that maps two sequences of the same length, the second of which is binary, to a set $U = \{a_i \colon b_i = 1\}$ ($(b_i)_{i=1}^n$ represents detection and indicates which elements of $(a_i)_{i=1}^n$ are included in U). Applying g_u to paired columns of $\mathbf{X}$ and $\mathbf{D}$, we get $\boldsymbol{U} = \big(g_u(X_{\cdot 1}, D_{\cdot 1}), \ldots, g_u(X_{\cdot L}, D_{\cdot L})\big)$, full observed data for a single sample. Thus G_u is defined as $G_u(\mathbf{X}, \mathbf{D}) \equiv \big(g_u(X_{\cdot l}, D_{\cdot l})\big)_{l=1}^L$. Note that if an allele from a minor strain i at locus l has not been detected ($D_{il} = 0$), it can still belong to U_l if it is present and has been detected in other strains.

Probability of observed genotyping data

Since $X_{il} \perp D_{jm}\ \forall\, i, j, l, m$, the probability of observed data $\boldsymbol{U}$ for an infection with n distinct clones can be written simply as

$$P(\boldsymbol{U}) = \sum_{\mathbf{X},\, \mathbf{D} \colon G_u(\mathbf{X},\, \mathbf{D}) = \boldsymbol{U}} P(\mathbf{X})\, P(\mathbf{D}) \tag{1}$$

The summation in Equation (1) is taken over all the combinations of full genetic and detection matrices that are compatible with observed data, thus making this direct approach computationally unfeasible for most cases and generally unscalable; including a second sample combined with recrudescence adds another layer of complexity. While Aster uses other approaches described below to efficiently calculate the likelihood, it is a useful reference to what exactly needs to be calculated. Equation (1) does not explicitly exclude matrices with repeated rows, and there is a non-zero probability that two unrelated strains are identical by state at all the loci in a sequencing panel. However, if the panel is informative (i.e. has a sufficient number of diverse loci), this probability is small; in addition, if $|U_l| = n$ for at least one locus l, which is often the case in practice as most COI estimation methods for multiallelic data would yield estimates that are no greater than $\max(|U_1|, \ldots, |U_L|)$, no matrices with repeated rows will be compatible with $\boldsymbol{U}$.

Using a single sample, we introduce some of the concepts of Aster implementation. First, since loci can be considered separately (while keeping in mind that for both recrudescence and detection, strains/rows need to be tracked across the loci by keeping their indices fixed), we can switch the order of the sum and the product, greatly reducing the number of combinations:

$$P(\boldsymbol{U}) = \sum_{\substack{\mathbf{X},\, \mathbf{D}\colon \\ G_u(\mathbf{X},\, \mathbf{D}) = \boldsymbol{U}}} P(\mathbf{X})\, P(\mathbf{D}) = \sum_{\substack{\mathbf{X},\, \mathbf{D}\colon \\ G_u(\mathbf{X},\, \mathbf{D}) = \boldsymbol{U}}} \left[\prod_{l=1}^{L} P(X_{\cdot l})\right] \left[\prod_{l=1}^{L} P(D_{\cdot l})\right] = \prod_{l=1}^{L} \sum_{\substack{X_{\cdot l},\, D_{\cdot l}\colon \\ g_u(X_{\cdot l},\, D_{\cdot l}) = U_l}} P(X_{\cdot l})\, P(D_{\cdot l})$$

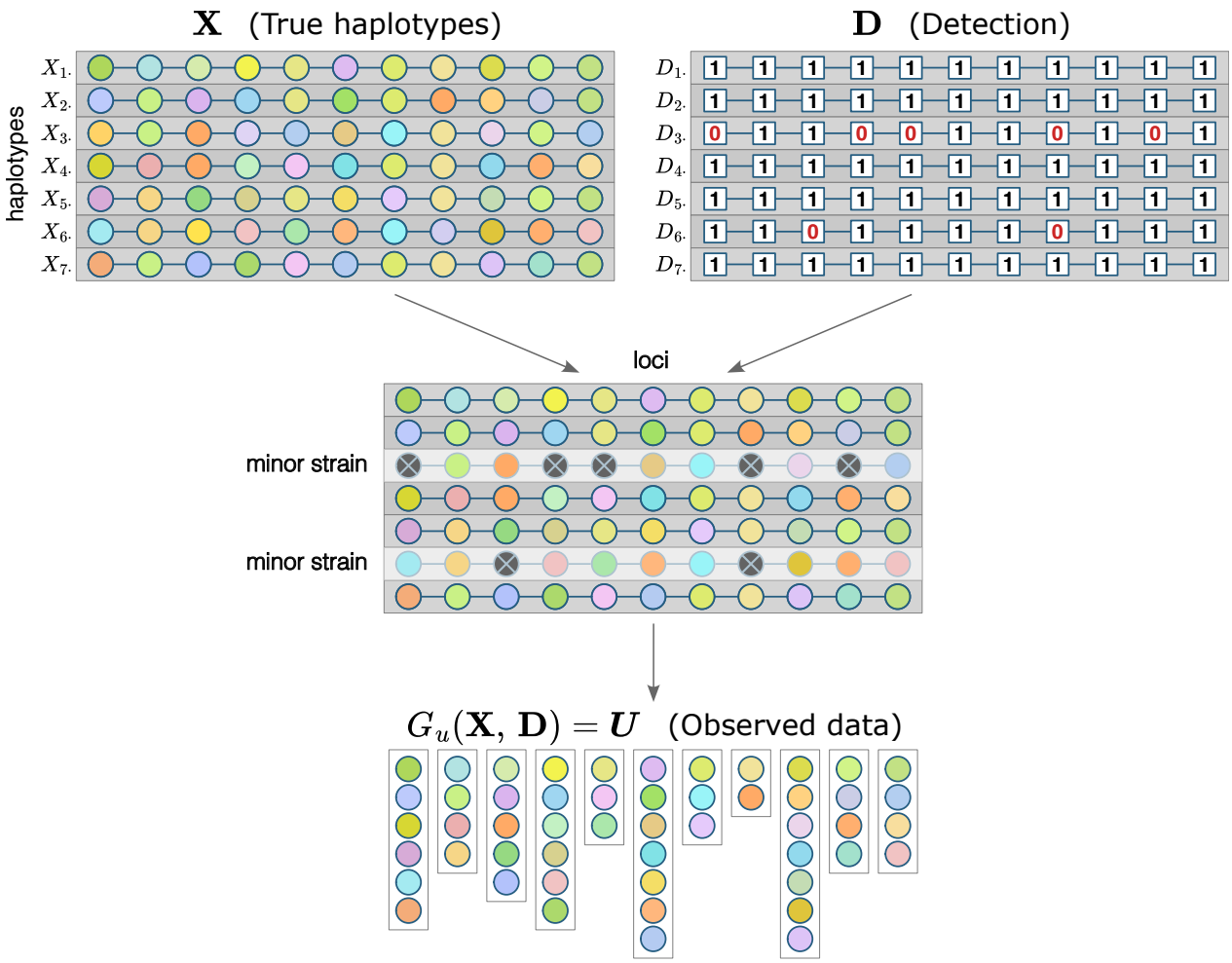


Figure 1: Framework components for a single sample. Each row of **X** and **D** matrices represents a strain, each column - a locus. Colors inside circles represent alleles: for each locus, the same color indicates the same allele; multiple strains can have the same allele at a given locus. Minor strains are not fully detected, which is indicated with crossed out dark gray circles, but an allele that is undetected in one strain can still be detected at a locus if it is present in another strain. Observed data for each locus are a set of detected unique alleles.

Next, we define a function that allows us to calculate the probability of a set of unique alleles in a number of strains at a given locus without explicitly allocating these alleles to specific strains. Let $P_u(U, n)$ be a function that calculates a probability that a sequence of length n has all the elements of set U with corresponding probabilities and only elements from set U: for a sequence $Z_1, \ldots, Z_n$ of independent identically distributed categorical random variables,

$$P_u(U, n) = P(Z_i \in U \; \forall i \text{ and } \forall a \in U \; \exists j\colon Z_j = a). \tag{2}$$

Set $P_u(\varnothing, 0) = 1$ (no elements in a zero-length sequence); also note that $P_u(U, n) = 0$ if $|U| > n$ and $P_u(\varnothing, n) = 0$ if $n > 0$. In Aster, elements of U are alleles, their corresponding probabilities are population allele frequencies, and n is the number of strains to which alleles in U belong. Supplementary section S.1 provides two efficient ways to calculate $P_u(U, n)$.

To incorporate missingness, recall that $P(D_{il} = 1) = p_{il}$. If $D_{il} = 0$, i.e. an allele in strain i at

locus l is not detected, X_{il} is unknown and can be any allele $a \in A_l$; consequently, if alleles from all the other strains $j \neq i$ are detected, $P(U_l \,|\, n, D_{il} = 0) = P_u(U_l,\, n-1)$ (see Supplementary Section S.2 for details). Then, for a sample with COI of n and a single minor strain i,

$$\begin{aligned} P(U_l \,|\, n) &= P(U_l \,|\, n, D_{il} = 1)\, P(D_{il} = 1) + P(U_l \,|\, n, D_{il} = 0)\, P(D_{il} = 0) \\ &= p_{il}\, P_u(U_l,\, n) + (1 - p_{il})\, P_u(U_l,\, n-1) \end{aligned} \tag{3}$$

for locus l.

In general, for any set of minor strains,

$$P(U_l \,|\, n) = \sum_{D_{\cdot l} \in \mathcal{B}^{(n)}} \left[\prod_{i:\, D_{il}=1} p_{il} \right] \left[\prod_{j:\, D_{jl}=0} (1 - p_{jl}) \right] P_u\left(U_l,\, \sum_{i=1}^{n} D_{il} \right) \tag{4}$$

where $\mathcal{B}^{(n)}$ is a set of all the possible binary sequences of length n (note that the terms with $D_{\cdot l}$ such that $\sum_{i=1}^{n} D_{il} < |U_l|$ are equal to 0).

2.1.2 Pair of samples and recrudescence

Recrudescence can be viewed at a person level (i.e. an individual in a study) or, alternatively, considering that many infections are polyclonal, at a strain level. Commonly, when therapeutic failure rates are estimated, a failure is defined on a person level - even if multiple strains in a single infection have recrudesced. Adopting this approach, we consider a case where at least one of the D0 strains has not completely cleared and is present in the DR sample (regardless of the presence or absence of newly accquired strains) to be a recrudescence event. Let R_j be a binary random variable representing recrudescence for individual j where $R_j = 1$ indicates a recrudescence event, and let θ be a population-level failure rate. In TES, individual responses to therapy are commonly considered to be independent, so we assume $R_j \sim Bernoulli(\theta)$, $j = 1, \ldots, N$, where N is the number of individuals in the study, to be independent and identically distributed.

Genetic data generating model

Let $\mathbf{X}$ and $\mathbf{Y}$ be $n_x \times L$ and $n_y \times L$ matrices representing D0 and DR infections from the same individual. We assume that newly infecting clones are acquired independently from recrudescence, and consequently that recurrence of parasitemia can be a result of a new infection, a recrudescence, or both. If $\mathbf{Y}$ has recrudescent strains, the rows representing them are identical to the corresponding rows in $\mathbf{X}$, while entries Y_{jl} in the rows of $\mathbf{Y}$ that represent newly infecting strains (if any) are drawn from the same distribution as X_{il}. Since the order of the strains within an infection does not matter, we order them in such a way that recrudescent strains occupy the top rows of both matrices and their indices match in $\mathbf{X}$ and $\mathbf{Y}$, e.g. if an individual has two recrudescent strains, then $X_{1l} = Y_{1l}\ \forall\, l$ and $X_{2l} = Y_{2l}\ \forall\, l$.

We approach modeling alleles in recrudescent strains by considering identity by descent. A recrudescent strain i is the same in both samples, and it can be stated that X_{il} and Y_{il} are IBD at all loci. Let IBD_{il} be an indicator that X_{il} and Y_{il} are IBD. Then, using population allele frequencies $\pi(a_{lk})$,

$$P(X_{il},\, Y_{il} \,|\, IBD_{il} = 1) = \begin{cases} \pi(X_{il}) & \text{if } X_{il} = Y_{il} \\ 0 & \text{otherwise} \end{cases}$$

$$P(X_{il},\, Y_{il} \,|\, IBD_{il} = 0) = \pi(X_{il})\, \pi(Y_{il})$$

For a pair of strains $X_{i\cdot}$, $Y_{i\cdot}$, let R'_i be an indicator that strain i is recrudescent (we want to distinguish a strain specific variable R'_i from R, a person-level recrudescence). If $R'_i = 1$, $IBD_{il} = 1\ \forall\, l$; for the reverse to be true, we would need to assume that a person cannot be reinfected with strains that are related to the ones present at D0. That is not necessarily the case, and a population can have a non-zero level of background relatedness r_{bg}, which we define as a probability that a newly infecting strain and an originally present strain are IBD at a given locus. Then

$$P(IBD_{il} = 1 \,|\, R'_i = 1) = 1$$
$$P(IBD_{il} = 1 \,|\, R'_i = 0) = r_{bg}$$

Detection model

Let $\mathbf{D}_x$ and $\mathbf{D}_y$ be $n_x \times L$ and $n_y \times L$ binary matrices representing detection for $\mathbf{X}$ and $\mathbf{Y}$, with corresponding probabilities of detection $P(D_{x,il} = 1) = p_{x,il}$ and $P(D_{y,il} = 1) = p_{y,il}$. In addition to independence between full haplotype data and detection, we assume independence between detection in two samples $D_{x,il} \perp D_{y,jm} \quad \forall\ i, j, l, m$ as well as independence between detection and recrudescence/IBD status: $D_{x,il} \perp IBD_{il}$ and $D_{y,il} \perp IBD_{il}$.

Observation mechanism

We denote observed data for samples from D0 and DR by $\boldsymbol{U}_x = (U_{x,1},\, \ldots,\, U_{x,L}) = G_u(\mathbf{X},\, \mathbf{D}_x)$ and $\boldsymbol{U}_y = (U_{y,1},\, \ldots,\, U_{y,L}) = G_u(\mathbf{Y},\, \mathbf{D}_y)$ respectively.

Likelihood

In general, for a failure rate θ,

$$L(\theta;\, \boldsymbol{U}_x, \boldsymbol{U}_y) = \theta\, P(\boldsymbol{U}_x, \boldsymbol{U}_y \,|\, R = 1) + (1 - \theta)\, P(\boldsymbol{U}_x, \boldsymbol{U}_y \,|\, R = 0) \tag{5}$$

$P(\boldsymbol{U}_x, \boldsymbol{U}_y \,|\, R = 0) = P(\boldsymbol{U}_x)\, P(\boldsymbol{U}_y)$, but to derive $P(\boldsymbol{U}_x, \boldsymbol{U}_y \,|\, R = 1)$, we need to consider strain-level recrudescence and look at probabilities with given sets of indices for minor and recrudescent strains (recall that recrudescent strain indices should match in $\mathbf{X}$ and $\mathbf{Y}$). We start with a base case example where a possibly recrudescent strain is also a minor strain in both samples, then proceed to the general case; for derivations and special cases, see Supplementary Section S.2. Suppose there is a single strain $i = 1$ that might be recrudescent, a single minor strain $X_{1\cdot}$ in the D0 sample and a single minor strain $Y_{1\cdot}$ in the DR sample; also assume that strains $Y_{k\cdot}$, $k > 1$, if any, are not related to any strains in the D0 sample. Then

$$\begin{aligned} P(\boldsymbol{U}_x, \boldsymbol{U}_y) = \quad & \theta \prod_{l=1}^{L} P(U_{x,l}, U_{y,l} \,|\, IBD_{1l} = 1) \\ & + (1 - \theta) \prod_{l=1}^{L} \Big[P(U_{x,l}, U_{y,l} \,|\, IBD_{1l} = 1)\, r_{bg} \\ & \qquad + P(U_{x,l}, U_{y,l} \,|\, IBD_{1l} = 0)\, (1 - r_{bg}) \Big] \end{aligned}$$

(see Supplementary Section S.3).

To find $P(U_{x,l}, U_{y,l} \,|\, IBD_{1l})$, we need to account for all four combinations of $D_{x,1l}$ and $D_{y,1l}$ with corresponding conditional probabilities $P(U_{x,l}, U_{y,l} \,|\, D_{x,1l}, D_{y,1l}, IBD_{1l})$:

$$\begin{aligned} P(U_{x,l}, U_{y,l} \,|\, n_x, n_y, IBD_{1l} = 1) = \quad & p_{x,1l} \quad p_{y,1l} \; P^{(1)}_{xy}(U_{x,l}, U_{y,l},\, n_x, n_y) \\ + \quad & p_{x,1l} \; (1-p_{y,1l}) \, P_u(U_{x,l},\, n_x) \quad P_u(U_{y,l},\, n_y - 1) \\ + \, & (1-p_{x,1l}) \quad p_{y,1l} \; P_u(U_{x,l},\, n_x - 1) \; P_u(U_{y,l},\, n_y) \\ + \, & (1-p_{x,1l})(1-p_{y,1l}) \, P_u(U_{x,l},\, n_x - 1) \; P_u(U_{y,l},\, n_y - 1), \end{aligned}$$

where

$$\begin{aligned} P^{(1)}_{xy}(U_x, U_y,\, n_x, n_y) = \sum_{a \in U_{xy}} & \pi(a) \big[P_u(U_x,\, n_x - 1) + P_u(U_x \backslash \{a\},\, n_x - 1) \big] \\ & \times \big[P_u(U_y,\, n_y - 1) + P_u(U_y \backslash \{a\},\, n_y - 1) \big], \end{aligned} \tag{6}$$

$U_{xy} = U_x \cap U_y$, and a set difference $U \backslash \{a\}$ is a set U without an element a. $P^{(1)}_{xy}(U_x, U_y,\, n_x, n_y)$ is a function that accounts for a single recrudescent strain between two samples using Equation (2). When $IBD_{1l} = 0$,

$$P(U_{x,l}, U_{y,l} \,|\, n_x, n_y, IBD_{1l} = 0) = P(U_{x,l} \,|\, n_x) \, P(U_{y,l} \,|\, n_y),$$

where $P(U_{x,l} \,|\, n_x)$ and $P(U_{y,l} \,|\, n_y)$ are calculated as in Equation (3).

In general, to accommodate any number and combination of possibly recrudescent strains, IBD pairs of strains, and minor strains, whether recrudescent or not, in both samples, we combine Equation (4) with the generalized conditional probability, which is the key component of the likelihood:

$$P(U_{x,l}, U_{y,l} \,|\, D_{x,\cdot l}, D_{y,\cdot l}, IBD_{\cdot l}) = P^{(m)}_{xy} \left(U_{x,l}, U_{y,l},\, \sum_{i=1}^{n_x} D_{x,il}, \sum_{j=1}^{n_y} D_{y,jl} \right),$$

where $IBD_{\cdot l}$ is a sequence $(IBD_{il})_{i=1}^{min(n_x,n_y)}$, $m = \sum_{i=1}^{min(n_x,n_y)} D_{x,il} D_{y,il} IBD_{il}$, and

$$P^{(m)}_{xy}(U_{x,l}, U_{y,l},\, n_x, n_y) = \sum_{U' \subseteq U_{xy}} P_u(U',\, m) \sum_{U'' \subseteq U'} P_u(U_{x,l} \backslash U'',\, n_x - m) \sum_{U''' \subseteq U'} P_u(U_{y,l} \backslash U''',\, n_y - m) \tag{7}$$

Terms with $|U'| > m$ in Equation (7) are equal to 0 since $P(U,\, n) = 0$ for $|U| > n$. Note that

$$P^{(0)}_{xy}(U_{x,l}, U_{y,l},\, n_x, n_y) = P_u(U_{x,l},\, n_x) \, P_u(U_{y,l},\, n_y). \tag{8}$$

Alternatively, multisets can be used to calculate $P(U_{x,l}, U_{y,l} \,|\, D_{x,\cdot l}, D_{y,\cdot l}, IBD_{\cdot l})$; this approach can be more efficient in some situations. A mixed radix system algorithm [30, 31] is used for implementation of both approaches.

Equation (8) provides guidance for the treatment of missing data (no alleles detected at some loci in one or both samples). The proof that excluding such loci from the analysis does not introduce any bias can be found in Supplementary Section S.4.

2.1.3 Adaptivity

The framework can incorporate different types of available information and can be combined with other methods. Depending on the amount of external information and prior knowledge, additional assumptions and constraints can be used. Finally, it can accommodate different parameters of interest and produce a corresponding output, e.g., strain-level recrudescence.

Aster explicitly accounts for population allele frequencies and COI. Allele frequencies can be estimated from D0 samples, from a combination of D0 and new infections in recurrent samples, or can be provided from larger sample estimates with information borrowed from other studies; COI can be estimated using genetic data from a single sample or a collection of samples [32]. Similarly, background relatedness can be estimated from D0 samples or its distribution can be estimated from other sources; other information (e.g., household location, travel history, or use of interventions such as bed nets) can be available to estimate the probability of being reinfected with a related strain for a particular individual. The detection model has a flexible number of parameters, where probability of detection can vary between samples, loci, and strains. They can depend on sample parasite density, COI, and characterics of the genotyping panels such as sensitivity and diversity of the markers. Prior information on therapeutic failure rates can be accommodated (and might differ for different drug combinations) with a Bayesian approach; different treatments could be compared using different priors.

Instead of a fixed value or point estimate, a distribution of r_{bg} might be available and can be incorporated into analysis. Let $f_{bg}(r)$ be a probability density function (pdf) of r_{bg}. Then, for the likelihood of a failure rate θ,

$$L(\theta;\, \boldsymbol{U}_x, \boldsymbol{U}_y) = \int_0^1 L(\theta;\, \boldsymbol{U}_x, \boldsymbol{U}_y, r_{bg})\, f_{bg}(r_{bg})\, dr_{bg}$$

Similarly, distributions of detection probabilities can be incorporated. As an example, let the D0 sample have a single minor strain i with $p_{x,i1} = \cdots = p_{x,iL} = p_x$ and the DR sample have a minor strain j with $p_{y,j1} = \cdots = p_{y,jL} = p_y$; let $f_x(p)$ and $f_y(p)$ be pdf's of p_x and p_y respectively. Then

$$L(\theta;\, \boldsymbol{U}_x, \boldsymbol{U}_y) = \int_0^1 \int_0^1 L(\theta;\, \boldsymbol{U}_x, \boldsymbol{U}_y,\, p_x, p_y)\, f_x(p_x)\, f_y(p_y)\, dp_y\, dp_x$$

Since true recrudescence status for recurrent samples is inferred and not known for certain, there are no simple conjugate priors for the Bayesian approach. However, the posterior distribution of θ can be calculated numerically:

$$f_{\theta|\boldsymbol{U}_x,\boldsymbol{U}_y}(s) = \frac{L(s;\, \boldsymbol{U}_x, \boldsymbol{U}_y)\, f_\theta(s)}{\int_0^1 L(u;\, \boldsymbol{U}_x, \boldsymbol{U}_y)\, f_\theta(u) du}, \tag{9}$$

where $f_\theta(u)$ and $f_{\theta|\boldsymbol{U}_x,\boldsymbol{U}_y}(u)$ are prior and posterior pdf's of θ.

2.2 Classification and inference

To classify a recurrence as a recrudescence or a new infection, we use the likelihood in Equation (5). The likelihood is linear, and classification can be based on the sign of the slope of the likelihood, or, in other words, on comparing $P(\boldsymbol{U}_x, \boldsymbol{U}_y \,|\, R = 1)$ and $P(\boldsymbol{U}_x, \boldsymbol{U}_y \,|\, R = 0)$. Formally, for a person-level recrudescence,

$$P(\boldsymbol{U}_x, \boldsymbol{U}_y \,|\, R = 1) = P(\boldsymbol{U}_x, \boldsymbol{U}_y \,|\, M = 1)P(M = 1) + \cdots + P(\boldsymbol{U}_x, \boldsymbol{U}_y \,|\, M = n_{xy})P(M = n_{xy})$$

where $n_{xy} = \min(n_x, n_y)$ and $M = \sum_{i=1}^{n_{xy}} R'_i$ is the number of recrudescent strains. In most cases, however, it is sufficient to only consider a case with $M = 1$: if multiple strains are recrudescent while $M = 1$ is assumed, this only strengthens the evidence that $R = 1$, and if there is more evidence for $M = 0$ than for $M = 1$, it is unlikely that multiple recrudescent strains are present. While it is theoretically possible for the slopes of the likelihood functions with and without $M = 1$ assumption to have different signs (e.g. if $P(M = 1) \ll P(M > 1)$), such a case is unlikely to represent a realistic scenario.

For classification inference, a useful measure would quantify the uncertainty of the individual classification, for which a posterior probability $P(R = 1 \,|\, \boldsymbol{U}_x, \boldsymbol{U}_y)$ is a natural choice [12]. Let $A_1 \equiv P(\boldsymbol{U}_x, \boldsymbol{U}_y \,|\, R = 1)$ and $A_0 \equiv P(\boldsymbol{U}_x, \boldsymbol{U}_y \,|\, R = 0)$. Then

$$P(R = 1 \,|\, \boldsymbol{U}_x, \boldsymbol{U}_y) = \frac{A_1 P(R = 1)}{A_1 P(R = 1) + A_0 P(R = 0)}$$

Using an empirical Bayes approach, we borrow information from all the individuals in the study by first estimating $1 - \theta$ with Kaplan-Meier survival estimator $\widehat{S}(t_{end})$ [33], where t_{end} is the end of the follow-up period, and then using that estimate as a prior probability:

$$P(R = 1 \,|\, \boldsymbol{U}_x, \boldsymbol{U}_y) = \frac{A_1 \big(1 - \widehat{S}(t_{end})\big)}{A_1 \big(1 - \widehat{S}(t_{end})\big) + A_0 \, \widehat{S}(t_{end})}$$

In effect, this approach provides a formal way to compare A_0 and A_1, weighted by prior probabilities. If we assume a prior probability $P(R = 1) = 0.5$, $P(R = 1 \,|\, \boldsymbol{U}_x, \boldsymbol{U}_y) = A_1/(A_1 + A_0)$. The procedure used in this approach can be iterated: recurrence classification can be updated based on the posterior probability (e.g., adopting a 0.5 probability cut-off for the classification), leading to an updated failure rate estimate. For most practical applications, a single update will be sufficient for convergence.

In addition, available external information that would contribute to prior probability or distribution can be similarly incorporated (e.g. as shown in Equation (9)).

2.3 Simulations

We used simulations to compare performances of Aster and match-counting algorithms for various genotyping panels, transmission intensities, genotyping error scenarios, and background relatedness levels. Panel comparisons were based on the set of 3 length polymorphisms (LP) currently recommended by the WHO (MSP-1, MSP-2, and a single microsatellite) [20], and a contemporary amplicon sequencing panel containing diverse microhaplotypes (MAD^{4}HatTeR) [24]. For the amplicon sequencing panel, subsets of loci for simulations were chosen based on their heterozygosity ranking using population allele frequencies obtained from the previously analyzed datasets from high transmission areas of sub-Saharan Africa. To reflect higher genotyping error rates of LP panels, we increased missingness (1 - probability of detection) twice and false positive rates 5 times compared to the amplicon sequencing panels [34, 17]. Transmission intensity levels were reflected in sample COI and the number of new infections where applicable. COI were drawn from zero-truncated Poisson (ZTP) distributions, with D0 means of 1.2, 3, and 5 for low, moderate, and high transmission settings respectively; since COI for recurrent infections are usually lower, DR means were respectively decreased by a factor of 1.8 (bounded below by 1). Genotyping errors were included with a simple "per strain" detection model for missing (false negative) alleles and a

"split" model for false positive alleles, where each correctly detected allele can give rise to additional falsely detected ones ([30]).

We used three types of simulation schemes with increasing complexity: simulated pairs of samples to evaluate classification performance, simulated studies with fixed proportions of recrudescences and newly acquired infections, and studies with simulated responses to treatment combinations using an established pharmacokinetic/pharmacodynamic (PK/PD) model [35, 36, 37, 10]. With the first simulation scheme, groups of three samples were generated: a D0 sample and two DR samples, with and without recrudescence. D0-DR pairs with recrudescence were used to obtain sensitivity, while D0-DR pairs without recrudescence were used for specificity. We simulated 10,000 such triads for each combination of transmission intensity, allele detection probability, background relatedness, and genotyping panel. With the second scheme, 1000 studies were simulated for each combination. While the proportions of individuals with recrudescent strains and with newly infecting ones were fixed to remove variability due to the randomness of recrudescence events, the individuals to whom these strains were assigned were random, creating a random number of overlaps; thus the numbers of total recurrences varied. The number of recrudescent strains for each individual was drawn from a ZTP distribution with a mean of 1/3 of their D0 COI. For PK/PD simulations, parasite density profiles were generated using model parameters suggested in [37] with adjusted half-maximal inhibitory concentrations and maximum kill rates to achieve therapeutic failure rates at or near 0.1 [38, 39]. We used homogeneous Poisson processes to introduce new infection times with incidence per year of 6 for moderate and 12 for high transmission; the numbers of coinfecting strains were drawn from ZTP distributions with means dependent on transmission intensity (2.5 for moderate, 4 for high transmission). To mimic a TES protocol, we measured parasite density at routine time points (days 1-7, 14, 21, 28); once a detection limit of 10^8 parasites/body had been reached, a recurrence was confirmed.

Aster analysis was conducted using *asterTES* R package [29]. Where applicable, we used naive COI estimation with a locus rank and a probabilistic COI-adjusted method ([30]) to estimate population allele frequencies for Aster. Allele detection probabilities in PK/PD simulations were estimated from sample parasite density using empirical data from mixed-strain controls for model fitting. While false positive alleles are not formally included in the Aster framework, its implementation allows for processing genetic data with false positives. When the number of alleles detected at a locus is greater than or equal to the COI, we allow for a possibility that some alleles might be false positives and therefore some might still be undetected by including a tuning parameter regulating the probability of this scenario. A small value for such probability would not affect classification results unless there is overwhelming evidence of recrudescence at other loci, in which case this evidence would be weighed against that value, and recrudescence would not be ruled out. For classification using match-counting algorithms, we calculated a proportion of loci with matches to apply these algorithms to simulated data from panels with more than 3 loci. Matches in all the loci were required for recrudescence with the 3/3 algorithm and matches in at least two thirds of the loci with the 2/3 algorithm; loci with no detected alleles in at least one of the samples were excluded from these calculations.

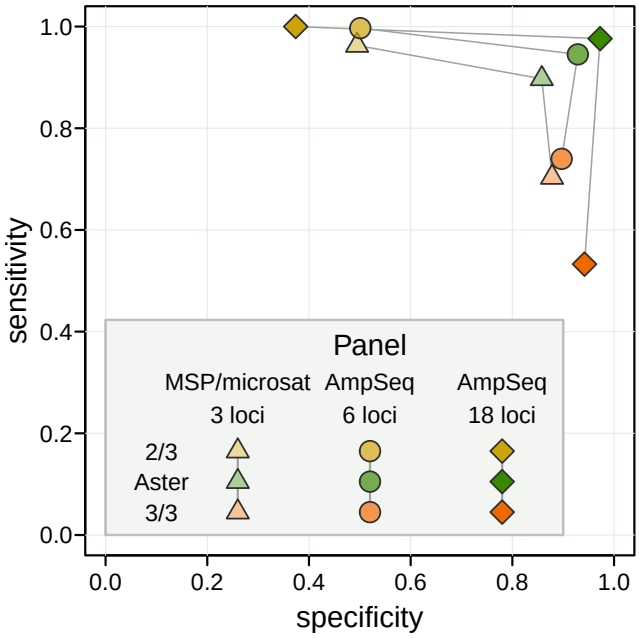


Figure 2: Sensitivity vs specificity for Aster, 3/3 algorithm, and 2/3 algorithm with three simulated genotyping panels: 3-locus length polymorphism (merozoite surface protein (MSP) 1, MSP-2, and a microsatellite) and 6 and 18-locus amplicon sequencing panels. A high transmission setting was used for these simulations, with a single recrudescent strain for each pair of samples and a partially detected recrudescent strain in a D0 sample; the probability of detection for that strain was set to 0.9 for all the loci in each panel. When applying match-counting algorithms to panels with > 3 loci, the proportion of loci with allele matches among all loci with data in both sampes was calculated; the value of 1 was required to classify a recurrence as a recrudescence by the 3/3 algorithm and of 2/3 or greater by the 2/3 algorithm.

3 Results

3.1 Classification performance

Aster uses a probabilistic approach for recurrence classification that explicitly accounts for biological and technical factors, including COI, allele frequency, background relatedness between infections, and imperfect detection of alleles. To evaluate its performance in identifying whether a DR infection is a recrudescence or a new infection, we first simulated pairs of infections with and without recrudescences. Using different genotyping panels and a range of transmission settings, allele detection probabilities, and background relatedness levels, we compared the sensitivity and specificity of Aster to those of the 3/3 and 2/3 match-counting algorithms.

In a simulated high transmission setting, Aster demonstrated balanced performance, exceeding 0.85 specificity and 0.9 sensitivity for all simulated genotyping panels and reaching near perfect classification with an 18-microhaplotype amplicon sequencing panel (Figure 2). In contrast, the 3/3 match-counting algorithm had high specificity but much lower sensitivity, classifying many recurrences as new infections; the opposite was true for the 2/3 algorithm. Performance of match-counting algorithms worsened with increasing numbers of loci, reflecting the limiting nature of a mechanistic identity by state approach that does not benefit from additional information.

Across the full range of simulations, classification was more accurate, as expected, when alleles were more likely to be detected and when transmission was lower, due to lower COI (Figure 3). However, we found that Aster had consistently higher overall and more balanced sensitivity and specificity than either match-counting algorithm. Aster performance improved with larger, more informative panels, with 6 or more loci yielding excellent accuracy in moderate transmission and 12 or more loci in most high transmission settings. Even in simulations where recurrences were the most challenging to classify, such as high transmission settings with poor allele detection, the sensitivity and specificity of Aster remained high for the 18-locus panel and improved further

with 48 loci (see Figure 3). In contrast, while match-counting algorithms performed well in low transmission settings (with mostly monoclonal infections), the differences between sensitivity and specificity became pronounced at moderate and high transmission settings. This imbalance could lead to considerable overestimation or underestimation of overall failure rates, as explored below. Supplementary Figure S.1 provides similar comparisons but with varying background relatedness of infections and detection probability fixed at 0.9; as expected, higher relatedness led to decreased performance for match-counting algorithms. The performance of Aster was more robust, particularly when using larger panels.

In some cases, it may be more appropriate to include a distribution of background relatedness instead of a single value (e.g., when relatedness seems overdispersed or bimodal). As a real-world example of this type of distribution, we used the empirical distribution of pairwise relatedness obtained from D0 samples from a TES recently performed in Asayita, Ethiopia (Supplementary Figure S.2). We simulated pairs of infections with background relatedness drawn from this distribution and then classified them using Aster and match-counting algorithms. Aster provided accurate and balanced results, particularly when using more informative genotyping panels with larger numbers of diverse microhaplotype loci (Supplementary Figure S.3). In contrast, match-counting algorithms gave poor results in moderate and high transmission settings for all genotyping panels.

3.2 Study-level estimation accuracy

The results of a TES are interpreted at the study level, so we next simulated studies to evaluate the effect of classification accuracy on these results. First, we present the summary of the estimates obtained from studies simulated with different proportions of individuals with recrudescent and newly infecting strains. Each study had 100 individuals, a typical size for a TES. Proportions of individuals with new strains were set to 0.1, 0.3, and 0.5 for simulated settings of low, moderate, and high transmission intensity respectively. For each combination of a genotyping panel, treatment failure rate, transmission intensity level, and probability of detection, 1000 studies were simulated, and a failure rate was estimated as a proportion of individuals with recrudescences.

As expected from classification results presented earlier, failure estimates based on Aster were the most accurate and were increasingly concentrated around the true proportion of recrudescence with greater numbers of loci (Figure 4). For example, in the most challenging high transmission setting, when the true failure rate was 0.15, Aster produced results between 0.14 and 0.18 95% of the time with an 18-locus panel. In contrast, match-counting algorithm estimates were more biased, with the extent and direction of bias depending on transmission setting, true failure rate, genotyping panel, and algorithm. In general, higher transmission resulted in greater estimate variability, more bias, and greater differences between the estimates obtained with the two algorithms; notably, more informative panels led to greater bias. The 2/3 algorithm tended to overestimate failure, with dramatic overestimation in high transmission simulations, e.g. producing estimates between 0.28 and 0.41 95% of the time for a true failure rate of 0.05 when using 18 loci. The 3/3 algorithm underestimated the failure in moderate transmission, with more biased results for higher failure rates, e.g. with 95% of the estimates between 0.06 and 0.15 for a true rate of 0.2 with an 18-locus panel; the estimates were expectedly higher for high transmission, where the bias went in either direction. To evaluate the effect of imperfect allele detection on study results, we next fixed the proportion of recrudescences at 15% and considered a range of detection probabilities. Expectedly, all the methods had excellent performance when allele detection was perfect in low transmission settings. In more challenging settings, Aster results remained close to the true values (particularly

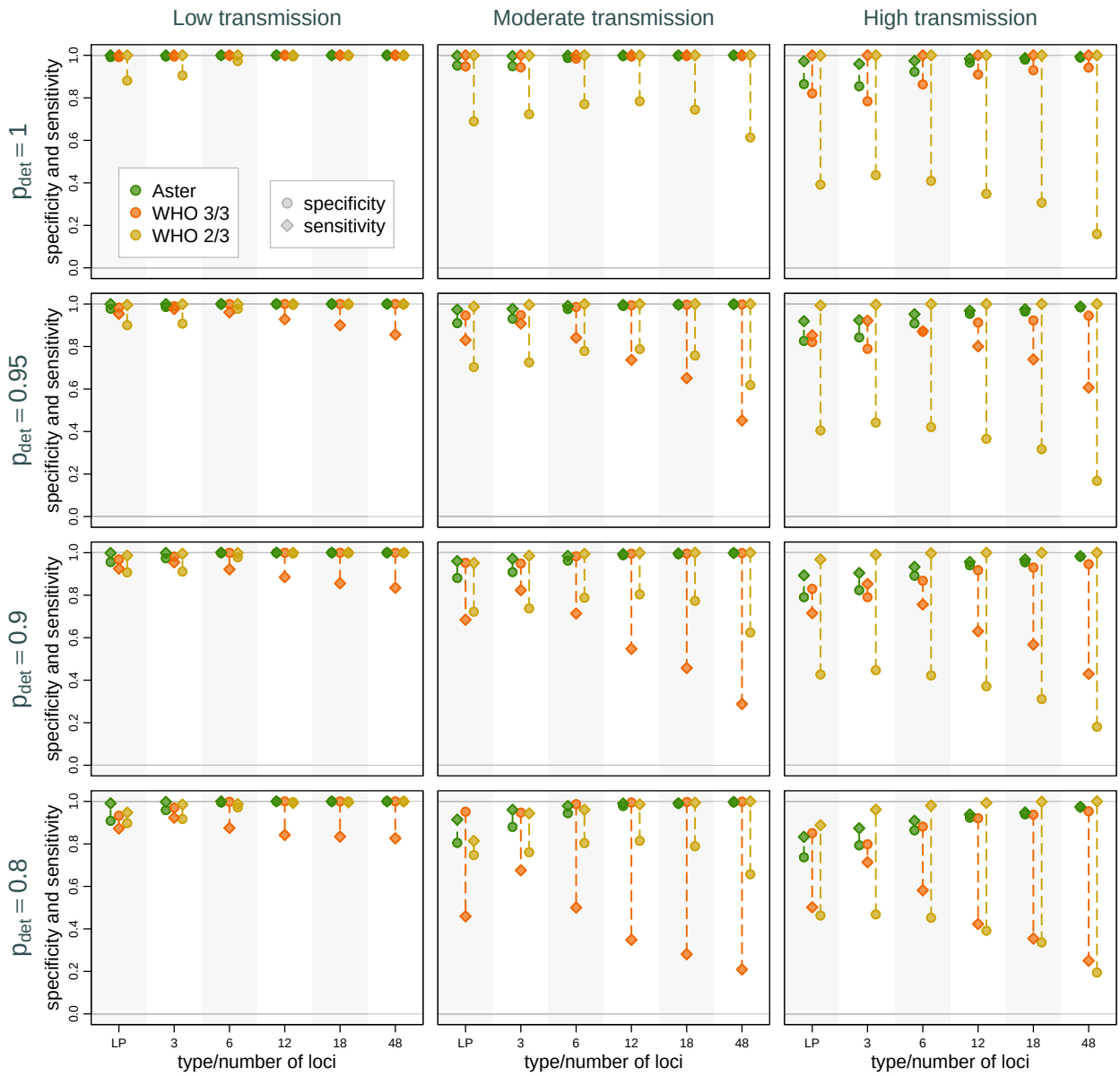


Figure 3: Sensitivity and specificity of Aster and match-counting algorithms across transmission intensity levels, genotyping panels, and detection probabilities (p_{det}). The panels presented are length polymorphisms (LP, 3 loci) and MAD^{4}HatTeR (varying number of loci). Dotted lines mark the difference between sensitivity and specificity for a specific method in a setting; longer lines imply greater imbalance and potentially greater bias in failure rate estimation. Background relatedness was set to 0.125.

for larger microhaplotype panels) for all levels of missingness, while match-counting estimates shifted downward with lower detection probabilities as more true allele matches were missed (Supplementary Figure S.4).

To better approximate the underlying biological and ascertainment processes of a TES, we used a previously developed PK/PD model to simulate the dynamics of parasite clones present in D0 infections, new infections acquired during the 28 day follow-up period, and the effect of antimalarial

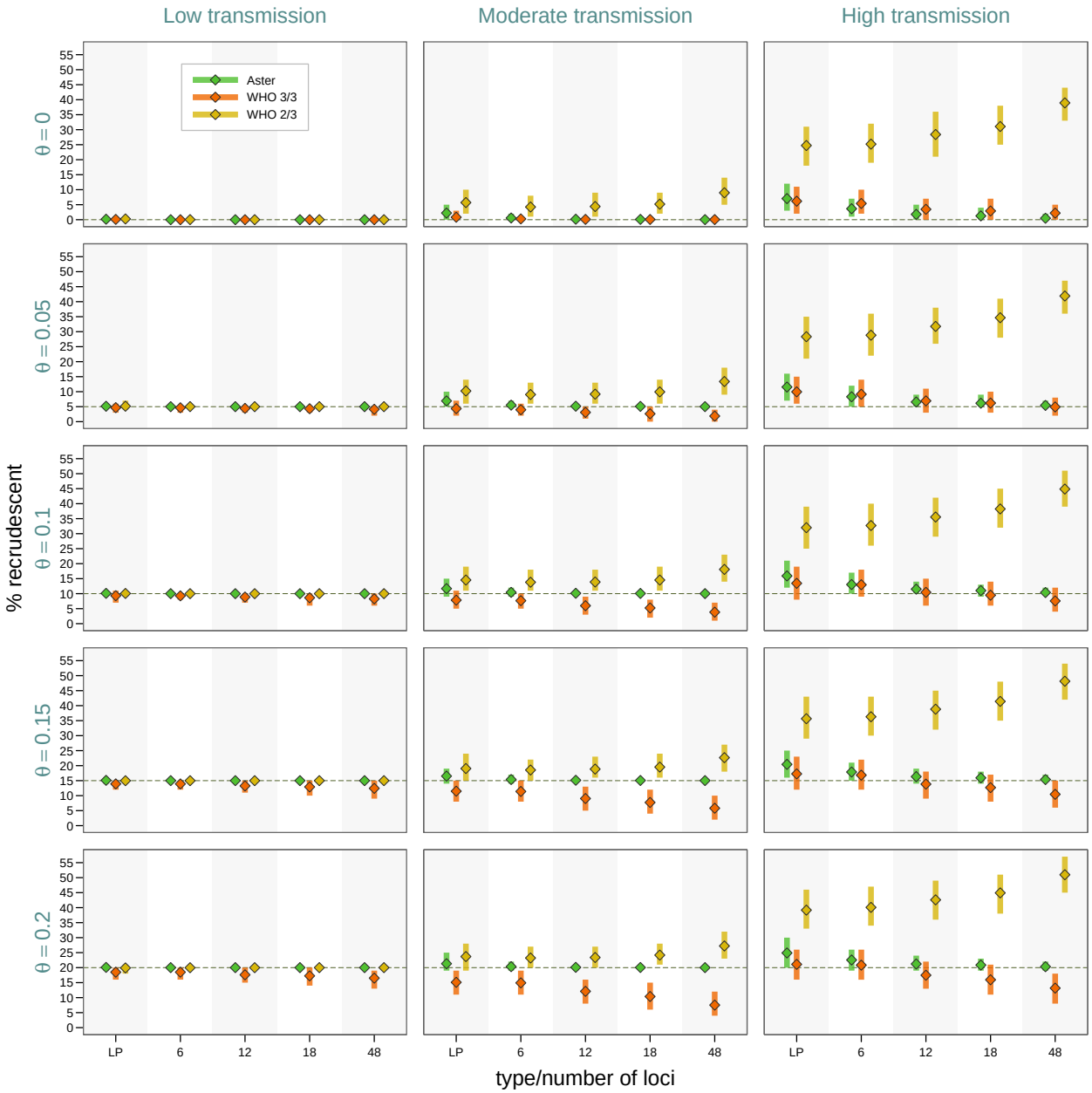
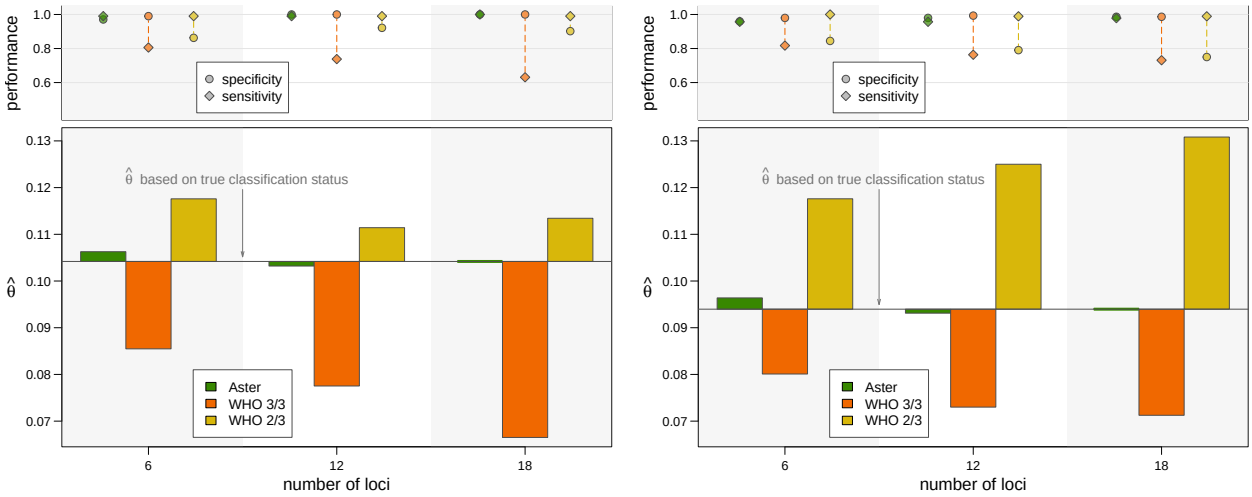


Figure 4: Study-level estimates across transmission intensity levels, genotyping panels, and true proportions of recrudescence events. The vertical bars represent a 95% range (0.025 to 0.975 quantiles) of the results, the diamond symbols represents a mean. Detection probability was fixed at 0.9, and background relatedness was 0.

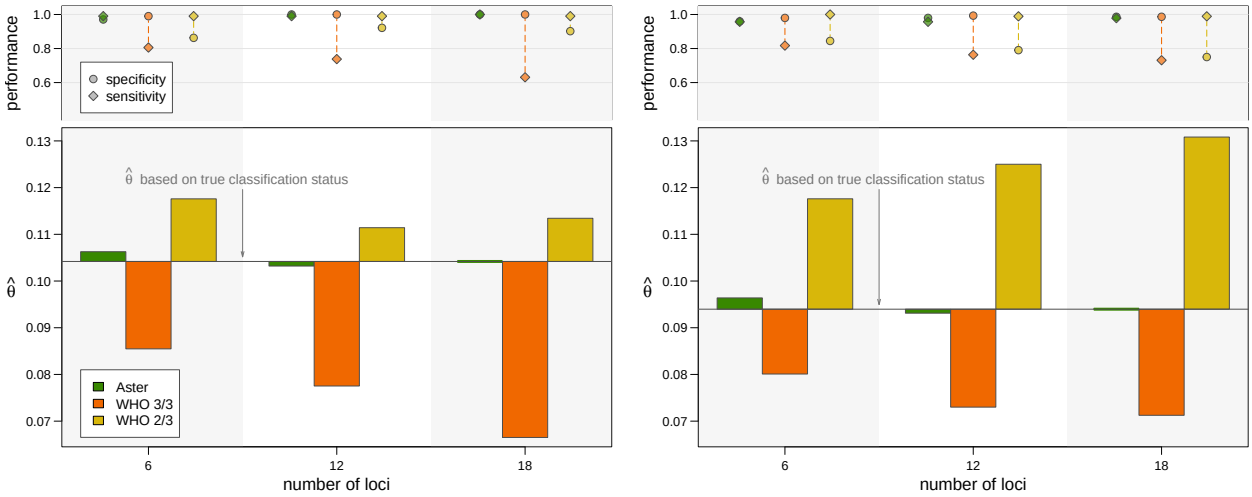
treatment with artemether-lumefantrine (Supplementary Figure S.5). For each simulated infection where recurrence would be detected, genotyping data were generated with detection based on relative and absolute parasite density (Supplementary Figure S.6). We then applied genotyping classification algorithms to these data and calculated failure rates using the Kaplan-Meier survival estimator. The results are presented for two simulations with moderate and high transmission settings respectively, each with 1000 individuals. Comparison results were qualitatively similar to the simple simulations presented above where estimates were accurate for Aster but biased for both

match-counting algorithms (Figure 5). In both simulations, the 3/3 method consistently underestimated while the 2/3 method consistently overestimated failure rates. Underestimation with the 3/3 method was more pronounced in the moderate transmission setting (producing an estimate of < 0.07 using 18 loci when the true failure rate was > 0.10), and overestimation with the 2/3 method was more pronounced in the high transmission setting (producing an estimate of > 0.13 using 18 loci when the true failure rate was < 0.10). This trend is consistent with previous results and is a likely consequence of more alleles matching by chance in the high transmission setting due to higher COI. Also consistent with previous simulation results, Aster outperforms match-counting algorithms in individual recurrence classifications (reflected in the top panels of Figure 5).



(a) Moderate transmission

(b) High transmission



(c) Moderate transmission

(d) High transmission

Figure 5: Therapeutic failure rate estimates with PK/PD model simulations using an amplicon sequencing panel with 6, 12, or 18 loci. For each figure, the bars in the lower panel represent deviation from the estimate based on true classification status; the upper panel displays performance measures. **(a)** Simulated dataset from moderate transmission: 1000 individuals with mean COI of 3.5 for D0, incidence of 6 infections/year, and an average of 2.5 coinfecting strains per new infection. **(b)** Simulated dataset from moderate transmission: 1000 individuals with mean COI of 5 for D0, incidence of 12 infections/year, and an average of 4 coinfecting strains per new infection.

3.3 Input estimation, parameter misspecification, and assumption violations

Since COI, population allele frequencies, detection probabilities, and other Aster model parameters are not usually known in practice and are estimated, inaccuracies of such estimates might affect Aster results. Therefore, we explored Aster's sensitivity to misspecifications by testing the method with different scenarios. Estimating COI and allele frequencies from simulated data had a minimal effect on Aster results (Supplementary Figure S.7, rows 2, 4). Simulation of false positive alleles, though not formally included in the estimation process, had a minimal effect on amplicon sequencing results but did decrease accuracy for length polymorphisms, which were simulated with a higher false positive rate (Supplementary Figure S.7, rows 3, 4).

Sensitivity to misspecification of allele detection probability is of particular importance, since perfect detection is unlikely and missingness is generally difficult to estimate in practice. To assess the effect of misspecification, we simulated studies of 100 individuals and looked at the distribution of failure rate estimates obtained with different detection probabilities in a recrudescent strain in a D0 sample (Supplementary Figure S.8). Assuming perfect detection resulted in considerable underestimation when true detection probability was less than 1, but the effect of other misspecified values was much smaller. For example, at a true value of 0.7 allele detection (fairly poor sensitivity), assumed values between 0.8 and 0.9 still resulted in estimates of treatment failure within 0.012 and 0.018 of the true value on average; assuming perfect detection shifted the values by nearly 0.15. Thus, in all the cases when true detection was not perfect, specifying a reasonable estimate of detection probability resulted in more accurate classification than assuming no missingness at all.

4 Discussion

Results from TES inform national and regional antimalarial drug policies. For these results to be reliable, genotype-corrected outcomes should be consistently accurate across transmission intensities and choice of genotyping assay. We found that by explicitly accounting for population allele frequencies, COI, background relatedness, and allele detection probability within a flexible statistical framework, Aster identifies recrudescent events with high sensitivity and specificity, exceeding the performance of WHO recommended match-counting algorithms. Improvements in accuracy were particularly salient in data simulated from high transmission areas - such as those where partial artemisinin resistance is now spreading - due to higher rates of acquiring new infections and higher COI, making outcome classification more important and more challenging. In addition, as more informative data from larger genotyping panels were included, Aster estimates became more accurate – a feature that will become increasingly useful as genotyping methods continue to evolve [40, 41, 42, 43, 24].

Since Aster incorporates important factors that meaningfully impact estimates of treatment failure, we consider practical implications of connecting upstream data and external information to Aster methodology and using them as Aster inputs. The COI of infections varies with malaria transmission intensity, affecting the number of alleles matching by chance and consequently the results of traditional match-counting algorithms [9, 12, 10]. While the true COI of samples are unknown, they can be readily estimated from empirical data and incorporated into Aster to minimize bias in outcome classification. Similarly, allele frequencies will vary based on the loci genotyped and the local parasite population, affecting the probability of alleles matching. These

too can be readily estimated from empirical data, for example using all or a random subset of D0 samples. Information on background relatedness between infections is explicitly included in Aster since underestimating the relatedness in a local population can lead to misclassifying a new infection as a recrudescence. As with allele frequencies, an estimate of background relatedness can be obtained from D0 samples, for example by estimating their pairwise IBD proportion, and supplied as an input to Aster. Imperfect detection of alleles may also affect results but is more difficult to estimate, potentially varying with genotyping method, COI, and parasite density. Fortunately, we show that Aster is robust to misspecification of detection probability as long as detection is not assumed to be perfect. Thus, including imperfect detection in the framework calibrates missingness in a principled way, allowing Aster to correctly classify recurrence as a recrudescence when recrudescent strains are not fully detected and would likely be missed by the 3/3 algorithm. Another type of genotyping error, false positive alleles, are not formally included in the current framework. However, while false positives can have an effect on the accuracy of match-counting algorithms, Aster is based on an IBD approach, which we demonstrate minimizes the effect of false positives on classification accuracy.

A formal framework allows Aster to approach the issue of missing data (no detected alleles at a locus, missing samples, or failed genotyping) in a systematic way. In turn, this can inform a principled treatment of right censoring and missingness in a failure rate estimation, regardless of the estimator used. For example, a recurrence with no genotyping data needs to be treated differently than a voluntary withdrawal from the study at the same time point: an individual with a recurrent infection would be more likely to have experienced a recrudescence by the end of the follow-up had they completed the study than an individual with loss of follow-up without a recurrence. In addition, the framework can be extended to include components beyond genotyping correction, for example those that deal with completely undetectable strains or persistence of gametocytes. These could affect the estimates and especially the uncertainty, more realistically reflecting the amount of unknowable information.

Reliable assessment of antimalarial efficacy is more pressing than ever, with resistance spreading in high transmission areas where accurate genotyping correction is the most challenging. Our results suggest that accuracy can be greatly improved through application of genotyping tools that provide rich genetic data, such as sensitive, diverse amplicon sequencing panels [24, 44], combined with analytical methods that take full advantage of these data and deliver consistent results across transmission levels. Fortunately, access to next generation sequencing has dramatically improved; over 80% of national public health labs in sub-Saharan Africa now have technology in place to generate high quality genotyping data (Africa CDC, personal communication, and [45]). Taking advantage of these data requires principled statistical approaches in place of simple matching-counting algorithms that can provide inconsistent results which vary by genotyping panel and epidemiological setting. Aster provides a structured statistical framework that is able to accommodate various types of information and can be amended with additional features in concordance with evolving laboratory methods and bioinformatic pipelines. The *asterTES* package, which currently provides recurrence classification with corresponding posterior probabilities and a failure rate estimate using Kaplan-Meier survival estimator, features a high level of flexibility for the user and can be extended along with the framework. With its fast and user-friendly implementation, Aster can contribute to accurate assessment of antimalarial TES, aiding stakeholders in timely determination of effective management of malaria.

5 Acknowledgements

We thank Ian Hastings and Sam Jones for their help and insight with PK/PD model simulations and parameter adjustments. We also thank Fitsum Girma for sharing unpublished data that provided a real life example of background relatedness from a TES. We thank Mateusz Plucinski and Andrés Aranda-Díaz for productive discussions and feedback.

This work was supported by NIH/NIAID (U01AI184646 and K24AI144048) and the Gates Foundation (INV-081860, INV-067310, and INV-048214).

References

[1] Samir Bhatt, DJ Weiss, Ewan Cameron, D Bisanzio, B Mappin, U Dalrymple, KE Battle, CL Moyes, A Henry, PA Eckhoff, et al. The effect of malaria control on plasmodium falciparum in africa between 2000 and 2015. *Nature*, 526(7572):207–211, 2015.

[2] World Health Organization. World malaria report 2023, 2023.

[3] Melissa D Conrad, Victor Asua, Shreeya Garg, David Giesbrecht, Karamoko Niaré, Sawyer Smith, Jane F Namuganga, Thomas Katairo, Jennifer Legac, Rebecca M Crudale, et al. Evolution of partial resistance to artemisinins in malaria parasites in Uganda. *New England Journal of Medicine*, 389(8):722–732, 2023.

[4] Philip J Rosenthal, Victor Asua, Jeffrey A Bailey, Melissa D Conrad, Deus S Ishengoma, Moses R Kamya, Charlotte Rasmussen, Fitsum G Tadesse, Aline Uwimana, and David A Fidock. The emergence of artemisinin partial resistance in Africa: how do we respond? *The Lancet Infectious Diseases*, 24(9):e591–e600, 2024.

[5] World Health Organization. Methods for surveillance of antimalarial drug efficacy, 2009.

[6] G Snounou and HP Beck. The use of PCR genotyping in the assessment of recrudescence or reinfection after antimalarial drug treatment. *Parasitology today*, 14(11):462–467, 1998.

[7] Moses R Kamya, Emmanuel Arinaitwe, Humphrey Wanzira, Agaba Katureebe, Chris Barusya, Simon P Kigozi, Maxwell Kilama, Andrew J Tatem, Philip J Rosenthal, Chris Drakeley, et al. Malaria transmission, infection, and disease at three sites with varied transmission intensity in Uganda: implications for malaria control. *The American journal of tropical medicine and hygiene*, 92(5):903, 2015.

[8] Adoke Yeka, Erika Wallender, Ronald Mulebeke, Afizi Kibuuka, Ruth Kigozi, Agaba Bosco, Paul Kyambadde, Jimmy Opigo, Simeon Kalyesubula, Joseph Senzoga, et al. Comparative efficacy of artemether-lumefantrine and dihydroartemisinin-piperaquine for the treatment of uncomplicated malaria in Ugandan children. *The Journal of infectious diseases*, 219(7):1112–1120, 2019.

[9] Bryan Greenhouse, Christian Dokomajilar, Alan Hubbard, Philip J Rosenthal, and Grant Dorsey. Impact of transmission intensity on the accuracy of genotyping to distinguish recrudescence from new infection in antimalarial clinical trials. *Antimicrobial agents and chemotherapy*, 51(9):3096–3103, 2007.

[10] Sam Jones, Mateusz Plucinski, Katherine Kay, Eva Maria Hodel, and Ian M Hastings. A computer modelling approach to evaluate the accuracy of microsatellite markers for classification of recurrent infections during routine monitoring of antimalarial drug efficacy. *Antimicrobial Agents and Chemotherapy*, 64(4):10–1128, 2020.

[11] Ian M Hastings and Ingrid Felger. WHO antimalarial trial guidelines: good science, bad news? *Trends in parasitology*, 38(11):933–941, 2022.

[12] Mateusz M Plucinski, Lindsay Morton, Mary Bushman, Pedro Rafael Dimbu, and Venkatachalam Udhayakumar. Robust algorithm for systematic classification of malaria late treatment failures as recrudescence or reinfection using microsatellite genotyping. *Antimicrobial agents and chemotherapy*, 59(10):6096–6100, 2015.

[13] Muzamil Mahdi Abdel Hamid, Mohamed Hassan Abdelraheem, Desmond Omane Acheampong, Ambroise Ahouidi, Mozam Ali, Jacob Almagro-Garcia, Alfred Amambua-Ngwa, Chanaki Amaratunga, Lucas Amenga-Etego, Ben Andagalu, et al. Pf7: an open dataset of plasmodium falciparum genome variation in 20,000 worldwide samples. *Wellcome open research*, 8:22, 2023.

[14] Timothy JC Anderson, Bernhard Haubold, Jeff T Williams, Jose G Estrada-Franco§, Lynne Richardson, Rene Mollinedo, Moses Bockarie, John Mokili, Sungano Mharakurwa, Neil French, et al. Microsatellite markers reveal a spectrum of population structures in the malaria parasite plasmodium falciparum. *Molecular biology and evolution*, 17(10):1467–1482, 2000.

[15] Mateusz M Plucinski, Ian M Hastings, Leah F Moriarty, Meera Venkatesan, Ingrid Felger, and Eric S Halsey. Variation in calculating and reporting antimalarial efficacy against *Plasmodium falciparum* in sub-Saharan Africa: a systematic review of published reports. *The American Journal of Tropical Medicine and Hygiene*, 104(5):1820, 2021.

[16] Flavia Camponovo, Caroline O Buckee, and Aimee R Taylor. Measurably recombining malaria parasites. *Trends in parasitology*, 39(1):17–25, 2023.

[17] Bryan Greenhouse, Alissa Myrick, Christian Dokomajilar, Jonathan M Woo, Elaine J Carlson, Philip J Rosenthal, and Grant Dorsey. Validation of microsatellite markers for use in genotyping polyclonal plasmodium falciparum infections. *The American journal of tropical medicine and hygiene*, 75(5):836, 2006.

[18] Isaie J Reuling, Lisanne A Van De Schans, Luc E Coffeng, Kjerstin Lanke, Lisette Meerstein-Kessel, Wouter Graumans, Geert-Jan van Gemert, Karina Teelen, Rianne Siebelink-Stoter, Marga van de Vegte-Bolmer, et al. A randomized feasibility trial comparing four antimalarial drug regimens to induce plasmodium falciparum gametocytemia in the controlled human malaria infection model. *Elife*, 7:e31549, 2018.

[19] World Health Organization. Methods and techniques for clinical trials on antimalarial drug efficacy: Genotyping to identify parasite populations. Informal consultation organized by the Medicines for Malaria Venture and cosponsored by the World Health Organization, 29-31 May 2007, Amsterdam, The Netherlands, 2008.

[20] World Health Organization. Informal consultation on methodology to distinguish reinfection from recrudescence in high malaria transmission areas. Report of a virtual meeting, 17–18 May 2021, 2021.

[21] Somya Mehra, Aimee R Taylor, Mallika Imwong, Nicholas J White, and James A Watson. Probabilistic classification of late treatment failure in uncomplicated malaria. *medRxiv*, pages 2025–01, 2025.

[22] Mateusz M Plucinski and Joel LN Barratt. Nonparametric binary classification to distinguish closely related versus unrelated plasmodium falciparum parasites. *The American Journal of Tropical Medicine and Hygiene*, 104(5):1830, 2021.

[23] Annina Schnoz, Carla Beuret, Maura Concu, Salome Hosch, Liliana K Rutaihwa, Monica Golumbeanu, and Christian Nsanzabana. Genotyping methods to distinguish *Plasmodium falciparum* recrudescence from new infection for the assessment of antimalarial drug efficacy: an observational, single-centre, comparison study. *The Lancet Microbe*, 5(11), 2024.

[24] Andrés Aranda-Díaz, Eric Neubauer Vickers, Kathryn Murie, Brian Palmer, Nicholas Hathaway, Inna Gerlovina, Simone Boene, Manuel García-Ulloa, Pau Cisteró, Thomas Katairo, Francis Ddumba Semakuba, Bienvenu Nsengimaana, Hazel Gwarinda, Carla García-Fernández, William Louie, Endashaw Esayas, Clemente Da Silva, Debayan Datta, Shahiid Kiyaga, Innocent Wiringilimaana, Sindew Mekasha Feleke, Adam Bennett, Jennifer L. Smith, Endalamaw Gadisa, Jonathan B. Parr, Melissa D. Conrad, Jaishree Raman, Stephen Tukwasibwe, Isaac Ssewanyana, Eduard Rovira-Vallbona, Cristina M. Tato, Jessica Briggs, Alfredo Mayor, and Bryan Greenhouse. Sensitive and modular amplicon sequencing of *Plasmodium falciparum* diversity and resistance for research and public health. *Scientific Reports*, 15(1):10737, 2025.

[25] Sam Jones, Katherine Kay, Eva Maria Hodel, Maria Gruenberg, Anita Lerch, Ingrid Felger, and Ian Hastings. Should deep-sequenced amplicons become the new gold standard for analyzing malaria drug clinical trials? *Antimicrobial Agents and Chemotherapy*, 65(10):10–1128, 2021.

[26] Camilla Messerli, Natalie E Hofmann, Hans-Peter Beck, and Ingrid Felger. Critical evaluation of molecular monitoring in malaria drug efficacy trials and pitfalls of length-polymorphic markers. *Antimicrobial agents and chemotherapy*, 61(1):10–1128, 2017.

[27] Johanna Helena Kattenberg, Norbert J Van Dijk, Carlos A Fernández-Miñope, Pieter Guetens, Mathijs Mutsaers, Dionicia Gamboa, and Anna Rosanas-Urgell. Molecular surveillance of malaria using the pf ampliseq custom assay for plasmodium falciparum parasites from dried blood spot dna isolates from peru. *Bio-protocol*, 13(5):e4621, 2023.

[28] Thomas Katairo, Victor Asua, Bienvenu Nsengimaana, Stephen Tukwasibwe, Francis D Semakuba, Innocent Wiringilimaana, Brian A Kagurusi, Caroline Mwubaha, Jackie Nakasaanya, Shreeya Garg, et al. Performance of molecular inversion probe dr23k and paragon mad4hatter amplicon sequencing panels for detection of plasmodium falciparum mutations associated with antimalarial drug resistance. *Malaria Journal*, 24(1):188, 2025.

[29] I. Gerlovina. astertes: Adaptive statistical framework for therapeutic efficacy and recrudescence. r package. `https://github.com/EPPIcenter/asterTES`, 2025.

[30] Inna Gerlovina, Boris Gerlovin, Isabel Rodríguez-Barraquer, and Bryan Greenhouse. Dcifer: an IBD-based method to calculate genetic distance between polyclonal infections. *Genetics*, 222(2):iyac126, 2022.

[31] Inna Gerlovina. MIRSA. https://doi.org/10.5281/zenodo.6438679, 2022.

[32] Maxwell Murphy and Bryan Greenhouse. Moire: A software package for the estimation of allele frequencies and effective multiplicity of infection from polyallelic data. *Bioinformatics*, 40(10):btae619, 2024.

[33] Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

[34] Anne Liljander, Lisa Wiklund, Nicole Falk, Margaret Kweku, Andreas Mårtensson, Ingrid Felger, and Anna Färnert. Optimization and validation of multi-coloured capillary electrophoresis for genotyping of plasmodium falciparum merozoite surface proteins (msp1 and 2). *Malaria journal*, 8(1):78, 2009.

[35] Katherine Winter and Ian M Hastings. Development, evaluation, and application of an *in silico* model for antimalarial drug treatment and failure. *Antimicrobial agents and chemotherapy*, 55(7):3380–3392, 2011.

[36] Ian M Hastings, Eva Maria Hodel, and Katherine Kay. Quantifying the pharmacology of antimalarial drug combination therapy. *Scientific Reports*, 6(1):32762, 2016.

[37] Sam Jones, K Kay, EvaMaria Hodel, S Chy, A Mbituyumuremyi, A Uwimana, D Menard, I Felger, and Ian Hastings. Improving methods for analyzing antimalarial drug efficacy trials: molecular correction based on length-polymorphic markers *msp-1*, *msp-2*, and *glurp*. *Antimicrobial agents and chemotherapy*, 63(9):10–1128, 2019.

[38] Eva Maria Hodel, Katherine Kay, and Ian M Hastings. Incorporating stage-specific drug action into pharmacological modeling of antimalarial drug treatment. *Antimicrobial agents and chemotherapy*, 60(5):2747–2756, 2016.

[39] Katherine Kay and Ian M Hastings. Measuring windows of selection for anti-malarial drug treatments. *Malaria journal*, 14:1–10, 2015.

[40] Maria Gruenberg, Anita Lerch, Hans-Peter Beck, and Ingrid Felger. Amplicon deep sequencing improves plasmodium falciparum genotyping in clinical trials of antimalarial drugs. *Scientific Reports*, 9(1):17790, 2019.

[41] Mathilde Gondard, Meredith Lane, Joel Barratt, Eldin Talundzic, and Yvonne Qvarnstrom. Simultaneous targeted amplicon deep sequencing and library preparation for a time and cost-effective universal parasite diagnostic sequencing approach. *Parasitology research*, 122(12):3243–3256, 2023.

[42] Mariateresa de Cesare, Mulenga Mwenda, Anna E Jeffreys, Jacob Chirwa, Chris Drakeley, Kammerle Schneider, Brenda Mambwe, Karolina Glanz, Christina Ntalla, Manuela Carrasquilla, et al. Flexible and cost-effective genomic surveillance of p. falciparum malaria with targeted nanopore sequencing. *Nature Communications*, 15(1):1413, 2024.

[43] Aurel Holzschuh, Anita Lerch, and Christian Nsanzabana. Multiplexed nanopore amplicon sequencing to distinguish recrudescence from new infection in antimalarial drug trials. *bioRxiv*, pages 2024–09, 2024.

[44] Emily LaVerriere, Philipp Schwabl, Manuela Carrasquilla, Aimee R Taylor, Zachary M Johnson, Meg Shieh, Ruchit Panchal, Timothy J Straub, Rebecca Kuzma, Sean Watson, et al. Design and implementation of multiplexed amplicon sequencing panels to serve genomic epidemiology of infectious disease: a malaria case study. *Molecular Ecology Resources*, 22(6):2285–2303, 2022.

[45] Gerald Mboowa, Sofonias K Tessema, Alan Christoffels, Nicaise Ndembi, Yenew Kebede Tebeje, and Jean Kaseya. Africa in the era of pathogen genomics: Unlocking data barriers. *Cell*, 187(19):5146–5150, 2024.

Supplementary Materials

S.1 Calculation of $P_u(U, n)$

Here we describe two approaches to calculate $P_u(U, n)$ that are used in implementation of Aster:

1. For $U = \{a_1, \ldots, a_m\}$ and probabilities $\{\pi(a_1), \ldots, \pi(a_m)\}$, calculate the probability of all the sequences with elements in U, which is $\left(\sum_{i=1}^{m} \pi(a_i)\right)^n$, then subtract probabilities of the sequences that do not contain all of the elements in U using the inclusion-exclusion principal:

$$\begin{aligned} P_u(U, n) &= \left[\sum_{i=1}^{m} \pi(a_i)\right]^n - \sum_{U^{(m-1)} \subset U} \left[\sum_{a_k \in U^{(m-1)}} \pi(a_k)\right]^n + \cdots + (-1)^{m-1} \sum_{i=1}^{m} \pi(a_i)^n \\ &= \sum_{i=1}^{m} (-1)^{m-i} \sum_{U^{(i)} \subseteq U} \left[\sum_{a_k \in U^{(i)}} \pi(a_k)\right]^n \\ &= \sum_{U' \subseteq U} (-1)^{m-|U'|} \left(\sum_{a_k \in U'} \pi(a_k)\right)^n, \end{aligned}$$

   where $U^{(k)}$ denotes a set of cardinality k.

2. Generate all the multisets S of cardinality n and support U: $|S| = n$, $Supp(S) = U$. $P_u(U, n)$ is the sum of all $P(S)$, which are calculated using multinomial probabilites.

The first approach is more computationally efficient for lower values of m and higher values of n. It gets slower with higher m and might encounter numerical issues (depending on a combinatorial implementation) if the orders of terms with alternating signs are much greater than the end result. The second approach is fully scalable and more efficient when m is close to n. Ultimately, the greatest efficiency is achieved with a rule prescribing which method to use in which situation, e.g. as a function of n and m that compares the number of combinations required for each method. Multisets and combinations are generated with a combinatorial mixed radix system algorithm (MIRSA) [30, 31].

$P_u(U, n)$ can also be defined recursively:

$$P_u(U, n) = \sum_{a \in U} \pi(a) \big[P_u(U, n-1) + P_u(U \backslash \{a\}, n-1)\big]. \tag{S.1}$$

This definition follows from Equation (2) and provides an essential simplification step that will allow us to derive a conditional probability of observed data given detection and recrudescence for a pair of samples and generalize it to any combination and number of undetected alleles and recrudescent strains.

S.2 Likelihood derivation

To translate observed unique alleles into full data and to detangle missingness and recrudescence, we need to combine combinatorics with tracking specific strains across loci (rows in $\mathbf{X}$, $\mathbf{Y}$) by fixing their indices. Starting with a single sample, let $\mathbf{D}$ be a random binary detection matrix such that

$P(D_{il} = 1) = 1 \; \forall i \neq k$ for a fixed index k of a single minor strain in the sample, $p_{kl} < 1 \; \forall l$. D_{il}, $i \neq k$, have a degenerate distribution and take a value of 1. k'th row of $\mathbf{D}$ can take values in $\{0, 1\}$. If $D_{kl} = 1$ for some locus l, alleles in U_l can be spread out across all n strains. If $D_{kl} = 0$, those alleles have to be in the other $n-1$ strains, and an unobserved realization of X_{kl} could be any allele $a \in A_l$ with probability $\pi(a)$. Then

$$P(U_l \,|\, n, D_{kl} = 0) = \sum_{a \in A_l} \pi(a) P_u(U_l,\, n-1) = P_u(U_l,\, n-1) \underbrace{\sum_{a \in A_l} \pi(a)}_{1} = P_u(U_l,\, n-1)$$

In general, for any number of minor strains, $P(U_l \,|\, n, D_{\cdot l}) = P_u(U_l,\, \sum_{i=1}^{n} D_{il})$.

Next, we consider a pair of samples with a single recrudescent strain $i = 1$ (the strains in $\mathbf{X}$ and $\mathbf{Y}$ are ordered in such a way that recrudescent strains are in the top rows and their indices are matching in $\mathbf{X}$ and $\mathbf{Y}$), $IBD_{1l} = 1 \; \forall l$. For some locus l assume that all the alleles have been detected in both samples and that allele $a \in U_{xy}$, $U_{xy} = U_x \cap U_y$ is in a recrudescent strain. That means that remaining $n_x - 1$ strains in $\mathbf{X}$ contain either all of the alleles in $U_{x,l}$ or all of the alleles except allele a; the same goes for $n_y - 1$ strains in $\mathbf{Y}$. Therefore, when there is no missingness,

$$\begin{aligned} P(U_{x,l}, U_{y,l} \,|\, n_x, n_y, IBD_{1l} = 1) &= \sum_{a \in U_{xy}} \pi(a) \big[P_u(U_{x,l},\, n_x - 1) + P_u(U_{x,l} \backslash \{a\},\, n_x - 1) \big] \\ &\qquad \times \big[P_u(U_{y,l},\, n_y - 1) + P_u(U_{y,l} \backslash \{a\},\, n_y - 1) \big] \\ &= P_{xy}^{(1)}(U_{x,l}, U_{y,l},\, n_x, n_y), \end{aligned} \tag{S.2}$$

using $P_{xy}^{(1)}$ function defined in Equation (6). If $|U_{x,l}| = n_x$, an allele in a recrudescent strain cannot be in any other strains; in that case $P_u(U_{x,l},\, n_x - 1) = 0$. There is a similarity between Equations (S.1) (one sample) and (6) (pair of samples); Equation (6) however cannot be simplified further as the summation goes over the elements of U_{xy} only and not over all the elements of $U_{x,l}$ or $U_{y,l}$.

Adding missingness to this scenario, we first consider a single minor strain in each sample. At a locus, there could be no alleles missing in either sample, one sample with an undetected allele and another with detected, or both undetected. For minor strains j in the D0 sample and k in the DR sample (omitting conditioning on n_x, n_y in notation for brevity),

$$\begin{aligned} P(U_{x,l}, U_{y,l} \,|\, IBD_{1l} = 1) = \; & p_{x,jl} \; p_{y,kl} \; P(U_{x,l}, U_{y,l} \,|\, D_{x,jl} = 1, D_{y,kl} = 1, IBD_{1l} = 1) \\ + \; & p_{x,jl} \; (1 - p_{y,kl}) \; P(U_{x,l}, U_{y,l} \,|\, D_{x,jl} = 1, D_{y,kl} = 0, IBD_{1l} = 1) \\ + \; & (1 - p_{x,jl}) \; p_{y,kl} \; P(U_{x,l}, U_{y,l} \,|\, D_{x,jl} = 0, D_{y,kl} = 1, IBD_{1l} = 1) \\ + \; & (1 - p_{x,jl}) \; (1 - p_{y,kl}) \; P(U_{x,l}, U_{y,l} \,|\, D_{x,jl} = 0, D_{y,kl} = 0, IBD_{1l} = 1) \end{aligned}$$

To find $P(U_{x,l}, U_{y,l} \,|\, D_{x,jl}, D_{y,kl}, IBD_{1l})$, there are three main cases and six combinations altogether that need to be covered:

1. alleles detected in both samples;
2. allele detected in one sample, not detected in another:
   (a) an undetected allele is in a recrudescent strain;
   (b) an undetected allele is not in a recrudescent strain;

3. alleles undetected in both samples:

   (a) undetected alleles in both samples are in a recrudescent strain;

   (b) an undetected allele is in a recrudescent strain in one sample and not in a recrudescent strain in another;

   (c) undetected alleles in both samples are not in a recrudescent strain;

The first case with both alleles detected is equivalent to the one with no minor strains:

$$P(U_{x,l}, U_{y,l} \,|\, D_{x,1l} = 1, D_{y,1l} = 1, IBD_{1l} = 1) = P^{(1)}_{xy}(U_{x,l}, U_{y,l},\, n_x, n_y)$$

For the second case, suppose the allele in recrudescent strain $i = 1$ is missing in the DR sample. Using Equation (S.1),

$$\begin{aligned} P(U_{x,l}, U_{y,l} \,|\, D_{x,jl} = 1, D_{y,1l} = 0, IBD_{1l} = 1) &= \sum_{a \in U_{x,l}} \pi(a) \big[\, P_u(U_{x,l},\, n_x - 1) + P_u(U_{x,l} \backslash \{a\},\, n_x - 1) \big] \\ &\qquad \times P_u(U_{y,l},\, n_y - 1) \\ &= P_u(U_{x,l},\, n_x)\, P_u(U_{y,l},\, n_y - 1) \end{aligned} \tag{S.3}$$

When the minor strain with an undetected allele in the DR sample is not the recrudescent one ($k \neq 1$),

$$\begin{aligned} P(U_{x,l}, U_{y,l} \,|\, D_{x,jl} = 1, \underset{k>1}{D_{y,kl}} = 0, IBD_{1l} = 1) &= \sum_{a \in U_{xy}} \pi(a) \big[P_u(U_{x,l},\, n_x - 1) + P_u(U_{x,l} \backslash \{a\},\, n_x - 1) \big] \\ &\qquad \times \Bigg[P_u(U_{y,l},\, n_y - 2) \underbrace{\sum_{b \in A_l} \pi(b)}_{1} + P_u(U_{y,l} \backslash \{a\},\, n_y - 2) \underbrace{\sum_{b \in A_l} \pi(b)}_{1} \Bigg] \\ &= P^{(1)}_{xy}(U_{x,l}, U_{y,l},\, n_x, n_y - 1) \end{aligned} \tag{S.4}$$

using Equation (6). Note that the index k is known and fixed across the loci, so when an allele Y_{kl} is undetected, that position needs to be excluded from combinatorial calculations in $P^{(1)}_{xy}$ function (hence $n_y - 2$ in the RHS of Equation (S.4)).

For the third case, when undetected alleles in both samples are in the recrudescent strain ($j = 1$, $k = 1$),

$$\begin{aligned} P(U_{x,l}, U_{y,l} \,|\, D_{x,1l} = 0, D_{y,1l} = 0, IBD_{1l} = 1) &= \sum_{a \in A_l} \pi(a) P_u(U_{x,l},\, n_x - 1)\, P_u(U_{y,l},\, n_y - 1) \\ &= P_u(U_{x,l},\, n_x - 1)\, P_u(U_{y,l},\, n_y - 1) \underbrace{\sum_{a \in A_l} \pi(a)}_{1} \\ &= P_u(U_{x,l},\, n_x - 1)\, P_u(U_{y,l},\, n_y - 1) \end{aligned} \tag{S.5}$$

When the minor strains with undetected alleles are the recrudescent strain in the D0 sample ($j = 1$) but not the recrudescent one in the DR sample ($k \neq 1$),

$$\begin{aligned} P(U_{x,l}, U_{y,l} \,|\, D_{x,1l} = 0, \underset{k>1}{D_{y,kl}} = 0, IBD_{1l} = 1) &= \sum_{a \in U_{y,l}} \pi(a) P_u(U_{x,l},\, n_x - 1) \\ &\qquad \times \Bigg[P_u(U_{y,l},\, n_y - 2) \underbrace{\sum_{b \in A_l} \pi(b)}_{1} + P_u(U_{y,l} \backslash \{a\},\, n_y - 2) \underbrace{\sum_{b \in A_l} \pi(b)}_{1} \Bigg] \\ &= P_u(U_{x,l},\, n_x - 1)\, P_u(U_{y,l},\, n_y - 1) \end{aligned} \tag{S.6}$$

Finally, when undetected alleles are not in the recrudescent strain in both samples ($j \neq 1$, $k \neq 1$),

$$P(U_{x,l}, U_{y,l} \,|\, \underset{j>1}{D_{x,jl}} = 0, \underset{k>1}{D_{y,kl}} = 0, IBD_{1l} = 1) = \sum_{a \in U_{xy}} \pi(a) \Bigg[P_u(U_{x,l},\, n_x - 2) \underbrace{\sum_{b \in A_l} \pi(b)}_{1} + P_u(U_{x,l} \backslash \{a\},\, n_x - 2) \underbrace{\sum_{b \in A_l} \pi(b)}_{1} \Bigg]$$
$$\times \Bigg[P_u(U_{y,l},\, n_y - 2) \underbrace{\sum_{b \in A_l} \pi(b)}_{1} + P_u(U_{y,l} \backslash \{a\},\, n_y - 2) \underbrace{\sum_{b \in A_l} \pi(b)}_{1} \Bigg]$$
$$= P^{(1)}_{xy}(U_{x,l},\, U_{y,l},\, n_x - 1, n_y - 1) \tag{S.7}$$

So far, we have only considered a single minor strain in each sample and a single recrudescent strain (or, more precisely, a single pair of strains that are IBD at a locus since alleles can also be IBD as a result of a reinfection with a related strain). For a general case, i.e. a conditional probability $P(U_{x,l}, U_{y,l} \,|\, D_{x,\cdot l}, D_{y,\cdot l}, IBD_{\cdot l})$ where $D_{x,\cdot l}$, $D_{y,\cdot l}$, and $IBD_{\cdot l}$ are any binary sequences of lengths n_x, n_y, and $\min(n_x, n_y)$ respectively, note that the only strains that contribute to possible dependence between $U_{x,l}$ and $U_{y,l}$ are the ones where alleles are simultaneously IBD and detected in both samples; let $m \equiv \sum_{i1}^{\min(n_x,n_y)} D_{x,il}\, D_{y,il}\, IBD_{il}$ be the number of such strains. Undetected alleles, either IBD or not, factor into the conditional probability only through the second argument of $P_u(U,\, n)$ function, which makes intuitive sense (reducing the number of strains to which elements of U can be allocated) and is formally shown in Equations (S.3), (S.4), (S.5), (S.6), and (S.7). Therefore, the conditional probability only depends on m and the number of detected alleles in each sample, which can be summarized as follows:

$$P(U_{x,l}, U_{y,l} \,|\, D_{x,\cdot l}, D_{y,\cdot l}, IBD_{\cdot l}) = P^{(m)}_{xy}\left(U_{x,l}, U_{y,l}, \sum_{i=1}^{n_x} D_{x,il}, \sum_{j=1}^{n_y} D_{y,jl} \right)$$

To derive $P^{(m)}_{xy}(U_x, U_y,\, n_x, n_y)$, we divide strain indices in $\mathbf{X}$ and $\mathbf{Y}$ into three groups: matching indices for IBD and detected strains and remaining indices in each sample. In m strains of the first group, all alleles can be different, some can be the same, or all can be the same; they also do not need to include all the elements of U_{xy}, some of which could be matching by chance. The other two groups should contain all the remaining alleles for each sample not allocated to the first group and may or may not also contain alleles already present in the first group:

$$P^{(m)}_{xy}(U_{x,l}, U_{y,l},\, n_x, n_y) = \sum_{U' \subseteq U_{xy}} P_u(U',\, m) \sum_{U'' \subseteq U'} P_u(U_{x,l} \backslash U'',\, n_x - m) \sum_{U''' \subseteq U'} P_u(U_{y,l} \backslash U''',\, n_y - m)$$

Some $U' \subseteq U_{xy}$ in the first sum can produce terms that are equal to 0: if $|U'| > m$, $P_u(U',\, m) = 0$; an empty set, which is an element of a power set of U_{xy} will also result in a term of 0 if $m > 0$ since $P(\varnothing,\, m > 0) = 0$. Another constraint on the cardinality of $|U'|$ is imposed by the cardinality of its complements with respect to $U_{x,l}$ and $U_{y,l}$ as the number of alleles not included in U' should not exceed $n_x - m$ or $n_y - m$; if it does, all of the terms in the second or the third sum will be equal to 0. The cases with $U'' = \varnothing$ or $U''' = \varnothing$ however represent scenarios where all the IBD alleles are also present in the other strains and are not generally equal to 0. Using empty sets, we obtain a special case of $m = 0$:

$$P^{(0)}_{xy}(U_{x,l}, U_{y,l},\, n_x, n_y) = \underbrace{P_u(\varnothing,\, 0)}_{1}\, P_u(U_{x,l} \backslash \varnothing,\, n_x - 0)\, P_u(U_{y,l} \backslash \varnothing,\, n_y - 0)$$
$$= P_u(U_{x,l},\, n_x)\, P_u(U_{y,l},\, n_y)$$

Another special case, $m = 1$, is covered in Equation (S.2) when there is no missingness; the same logic applies when alleles in some strains are undetected.

S.3 Background relatedness

Consider a base case where at most one strain $i = 1$ can be recrudescent: $P(R'_j = 1) = 0$ for $j = 2, \ldots, \min(n_x, n_y)$ and consequently $P(R = 1) = P(R'_1 = 1) = \theta$. Then

$$
\begin{aligned}
P(\boldsymbol{U}_x, \boldsymbol{U}_y) = \; & P(R'_1 = 1)\, P(\boldsymbol{U}_x,\ \boldsymbol{U}_y \mid R'_1 = 1) \\
& + P(R'_1 = 0)\, P(\boldsymbol{U}_x,\ \boldsymbol{U}_y \mid R'_1 = 0) \\
= \; & P(R'_1 = 1) \prod_{l=1}^{L} \Big[\; P(U_{x,l}, U_{y,l} \mid IBD_{1l} = 1)\, \underbrace{P(IBD_{1l} = 1 \mid R'_1 = 1)}_{1} \\
& \qquad + P(U_{x,l}, U_{y,l} \mid IBD_{1l} = 0)\, \underbrace{P(IBD_{1l} = 0 \mid R'_1 = 1)}_{0} \Big] \\
& + P(R'_1 = 0) \prod_{l=1}^{L} \Big[\; P(U_{x,l}, U_{y,l} \mid IBD_{1l} = 1)\, \underbrace{P(IBD_{l} = 1 \mid R'_1 = 0)}_{r_{bg}} \\
& \qquad + P(U_{x,l}, U_{y,l} \mid IBD_{1l} = 0)\, \underbrace{P(IBD_{il} = 0 \mid R'_1 = 0)}_{1 - r_{bg}} \Big] \\
= \; & \theta \prod_{l=1}^{L} \; P(U_{x,l}, U_{y,l} \mid IBD_{1l} = 1) \\
& + (1 - \theta) \prod_{l=1}^{L} \Big[\; P(U_{x,l}, U_{y,l} \mid IBD_{il} = 1)\, r_{bg} \\
& \qquad + P(U_{x,l}, U_{y,l} \mid IBD_{1l} = 0)\, (1 - r_{bg}) \Big],
\end{aligned}
$$

using the fact that observed data depends on recrudescence through IBD only, i.e. $P(U_{x,l}, U_{y,l} \mid IBD_{il}, R'_i) = P(U_{x,l}, U_{y,l} \mid IBD_{il})$.

In Section 2.1.2 we defined r_{bg} on a strain level, which implies that every newly infecting strain can be related to an originally present strain with r_{bg} IBD on average. If multiple pairs of non-recrudescent strains are related between a pair of the D0 and DR samples, all the IBD combinations at a locus need to be accounted for. In addition, an assumption of independence of all interhost IBD variables at a locus would be required to avoid transitive dependencies. In practice, however, relatedness for a pair of samples is often estimated on an infection level. This interpretation can be easily accommodated in Aster by restricting interhost relatedness to a single pair of strains, which would allow infection-level estimates to be used for r_{bg} directly while decreasing the number of combinations required to calculate the likelihood.

S.4 Missing data

Missing data $U_{x,l} = \varnothing$ or $U_{y,l} = \varnothing$, i.e. no detected alleles at locus l in a sample, is a special case of realization of $D_{x,\cdot l}$ or $D_{y,\cdot l}$ that is actually observed (a sequence of 0's); another special observed case being a sequence of 1's, which would be inferred when $|U_{x,l}| = n_x$ (assuming no false positive

alleles). Intuitively, such locus provides no information as to recrudescence regardless of detected alleles in the other sample; we show this formally. Since $\sum_{i=1}^{n_{xy}} D_{x,il} D_{y,il} IBD_{il} = 0$ regardless of $\sum_{i=1}^{n_{xy}} IBD_{il}$, where $n_{xy} = \min(n_x, n_y)$, we use Equation (8) and $P(\varnothing,\, 0) = 1$ to get

$$
\begin{aligned}
P\big(U_{x,l} = \varnothing, U_{y,l} \,|\, n_x, n_y, IBD_{\cdot l}\big) &= P\Big(\sum_{j=1}^{n_x} D_{x,jl} = 0\Big) \sum_{D_{y,\cdot l} \in \mathcal{B}^{(n_y)}} P(D_{y,\cdot l})\, P_{xy}^{(0)}\Big(\varnothing, U_{y,l},\, 0, \sum_{i=1}^{n_y} D_{y,il}\Big) \\
&= \Bigg[\prod_{j=1}^{n_x} (1 - p_{x,jl})\Bigg] \sum_{D_{y,\cdot l} \in \mathcal{B}^{(n_y)}} P(D_{y,\cdot l})\, P_u(\varnothing,\, 0)\, P_u\Big(U_{y,l}, \sum_{i=1}^{n_y} D_{y,il}\Big) \\
&= \Bigg[\prod_{j=1}^{n_x} (1 - p_{x,jl})\Bigg] \sum_{D_{y,\cdot l} \in \mathcal{B}^{(n_y)}} P(D_{y,\cdot l})\, P(U_{y,l} \,|\, D_{y,\cdot l}) \\
&= \Bigg[\prod_{j=1}^{n_x} (1 - p_{x,jl})\Bigg] P(U_{y,l} \,|\, n_y),
\end{aligned}
$$

where $\mathcal{B}^{(n)}$ is a set of all the binary sequences of length n. Similarly,

$$
\begin{aligned}
P\big(U_{x,l}, U_{y,l} = \varnothing \,|\, n_x, n_y, IBD_{\cdot l}\big) &= \Bigg[\prod_{k=1}^{n_y} (1 - p_{y,kl})\Bigg] P(U_{x,l} \,|\, n_x) \qquad \text{and} \\
P\big(U_{x,l} = \varnothing, U_{y,l} = \varnothing \,|\, n_x, n_y, IBD_{\cdot l}\big) &= \Bigg[\prod_{j=1}^{n_x} (1 - p_{x,jl})\Bigg] \Bigg[\prod_{k=1}^{n_y} (1 - p_{y,kl})\Bigg],
\end{aligned}
$$

the latter being a special case when no alleles are detected at locus l in either sample. The fact that $P\big(U_{x,l} = \varnothing, U_{y,l} \,|\, n_x, n_y, \sum_{i=1}^{n_{xy}} IBD_{il} > 0\big) = P\big(U_{x,l} = \varnothing, U_{y,l} \,|\, n_x, n_y, \sum_{i=1}^{n_{xy}} IBD_{il} = 0\big)$ means that recrudescence and no recrudescence are given the same weight, which can be factored out of the likelihood (of the form $\theta\, C_1 + (1-\theta)\, C_0$) and thus does not affect the shape of the log-likelihood function. This confirms that loci with no data in at least one of the samples can be ignored without introducing any bias.

Using similar logic, consider another special case that warrants attention and might be potentially problematic - completely undetected strains if they are recrudescent. Let strain $i = 1$ be possibly recrudescent and let $p_{x,1l} = 0\ \forall\, l$; assume no other strains can be recrudescent or IBD. Then $D_{x,1l} = 0$, $\sum_{i=1}^{n_{xy}} D_{x,il} D_{y,il} IBD_{il} = 0\ \forall\, l$, and

$$
P(U_{x,l}, U_{y,l} \,|\, n_x, n_y, IBD_{1l}) = P(U_{x,l} \,|\, n_x)\, P(U_{y,l} \,|\, n_y)
$$

regardless of the value of IBD_{1l}. Since that is the case for all the loci, $P(\boldsymbol{U}_x, \boldsymbol{U}_y \,|\, R = 1) = P(\boldsymbol{U}_x, \boldsymbol{U}_y \,|\, R = 0)$, which implies that recurrence classification cannot be performed with any degree of certainty, and the individual essentially does not contribute to failure rate estimation (apart from censoring). This situation presents a potential source of bias: incorrectly assuming no recrudescent strains to be completely missing might lead to underestimation of drug failure, and always assuming that there is an undetected recrudescent strain will never allow a recurrence to be classified as a new infection. Fortunately, this is not likely to occur commonly; if available, an estimate of the probability of this occuring could be incorporated into the framework.

S.5 Supplementary Figures

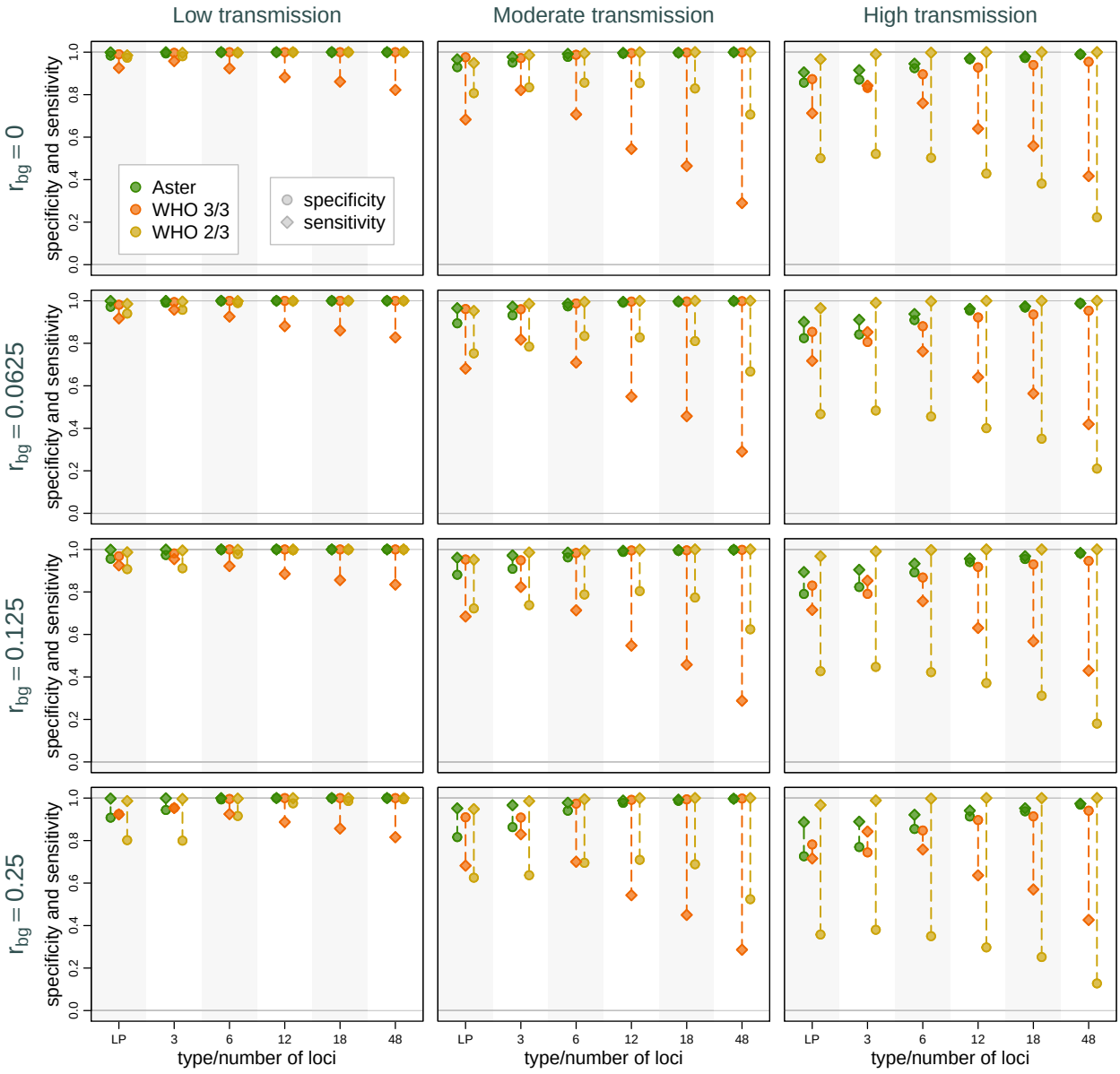


Figure S.1: Sensitivity and specificity of Aster and match-counting algorithms across trasmission intensity levels, genotyping panels, and background relatedness (r_{bg}) levels. Dotted lines mark the difference between sensitivity and specificity for a specific method in a setting; longer lines imply greater imbalance and potentially greater bias in failure rate estimation. Detection probability was set to 0.9.

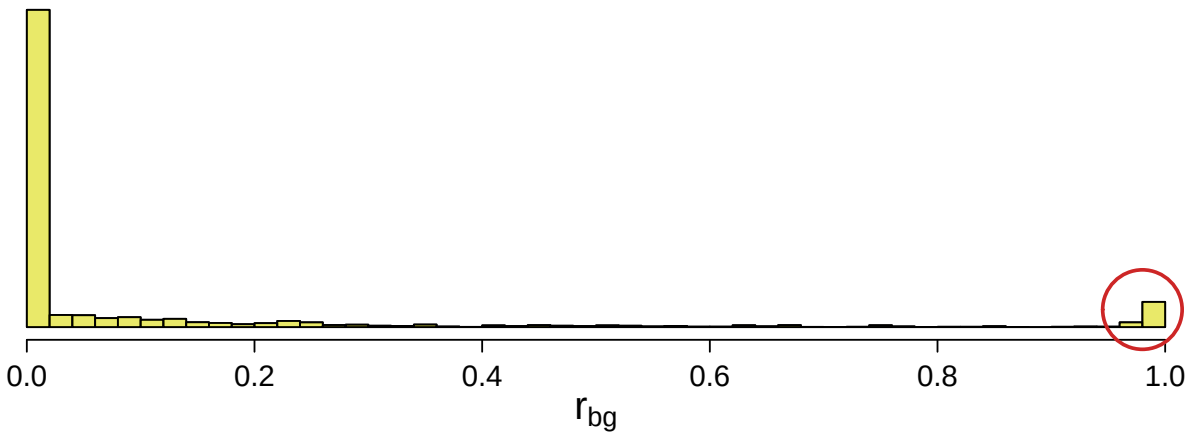
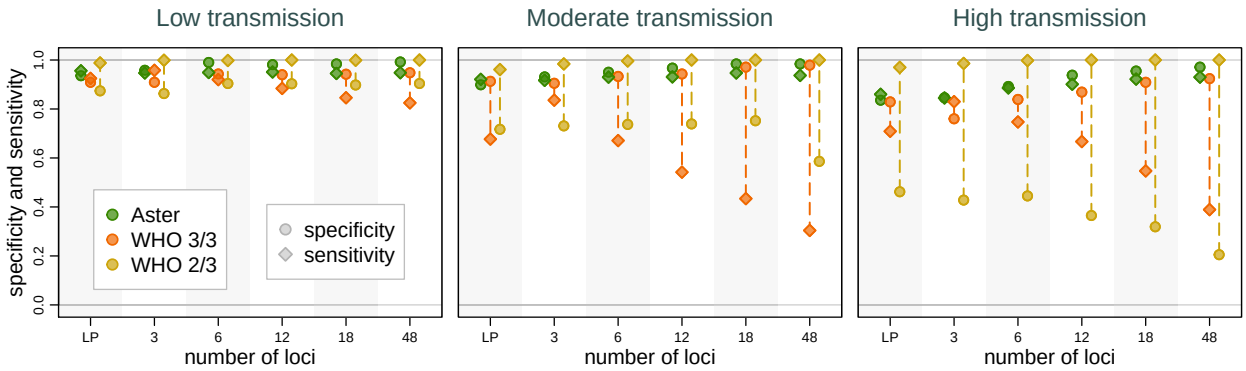
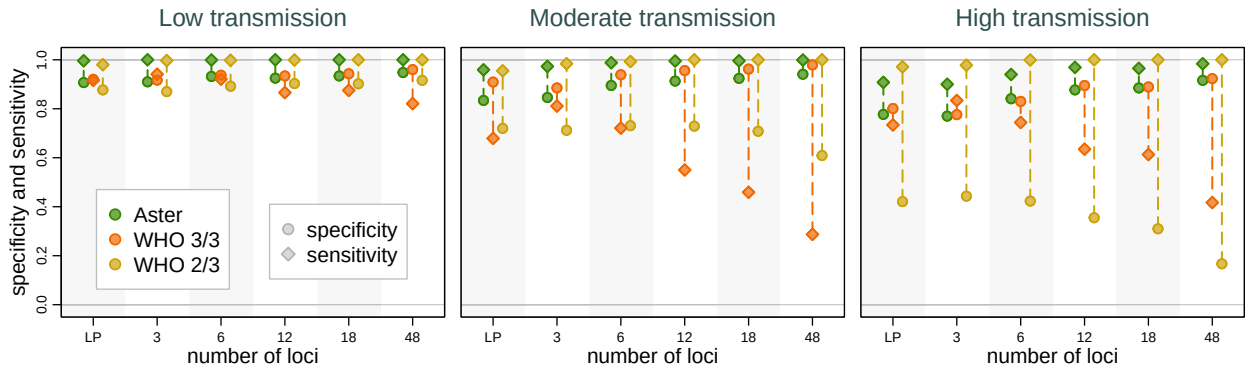


Figure S.2: Histogram of pairwise sample relatedness in Asayita, Ethiopia, estimated using Dcifer [30]. A red circle highlights an unusually high proportion of sample pairs that were highly related.



(a) True value of r_{bg} (usually unknown in practice) used as input to Aster for each pair of samples



(b) A distribution of r_{bg} used as input to Aster for each pair of samples

Figure S.3: Classification performance of Aster and match-counting algorithms across genotyping panels and transmission intensities from simulations using an empirical distribution of r_{bg} from Asayita, Ethiopia. Detection probability was set to 0.9.

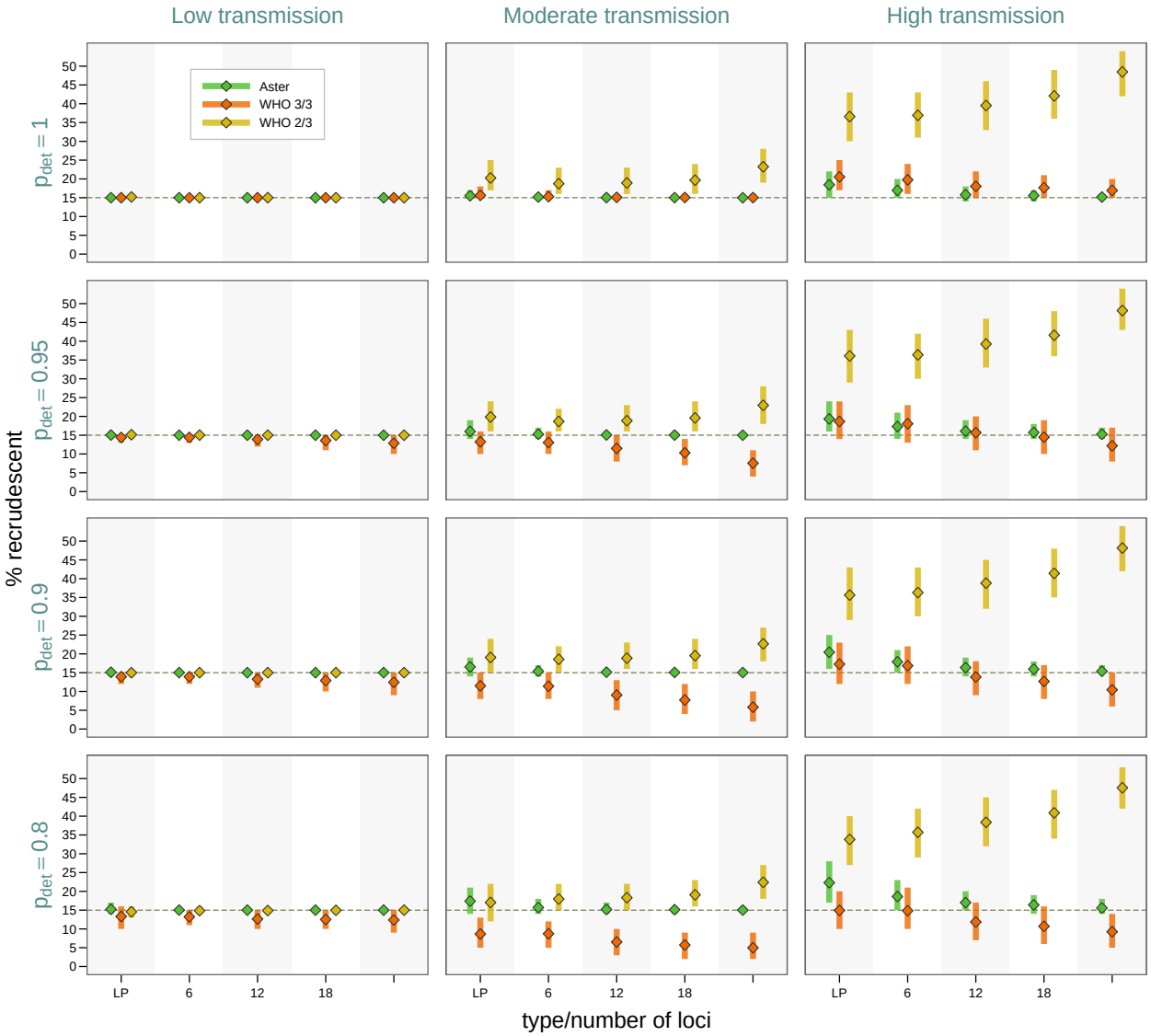
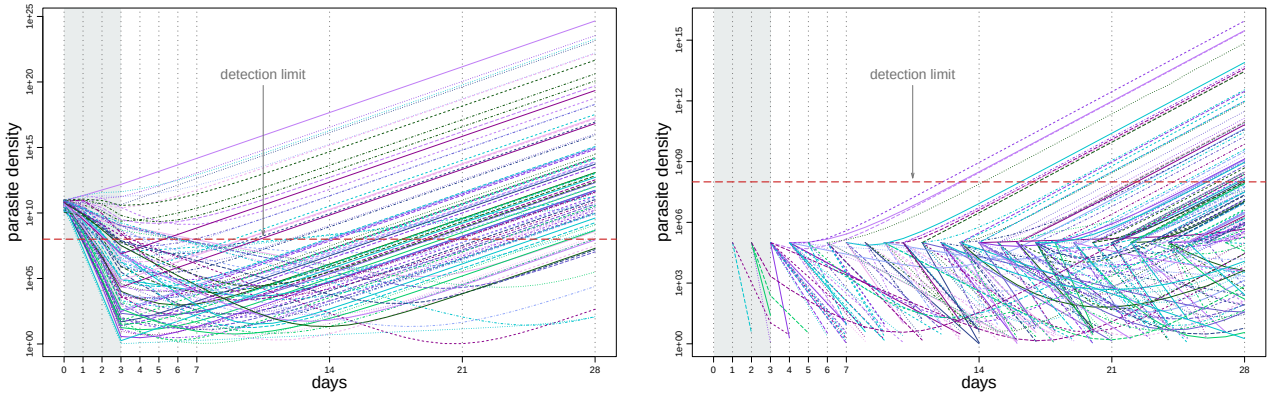


Figure S.4: Study-level estimates across transmission intensity levels, genotyping panels, and detection probabilities (p_{det}). The vertical bars represent a 95% range (0.025 to 0.975 quantiles) of the results, the diamond symbols represents a mean. Proportion of recrudescent events was fixed at 0.15 and background relatedness was 0.



(a) Recrudescent strains

(b) Newly acquired strains (introduced at 10^5 density)

Figure S.5: Simulated per-strain parasite density dynamics from PK/PD model. Vertical dotted lines represent time points at which the sample parasite density is measured; shaded area represents a treatment period. A dotted red horizontal line indicates a per-sample detection limit.

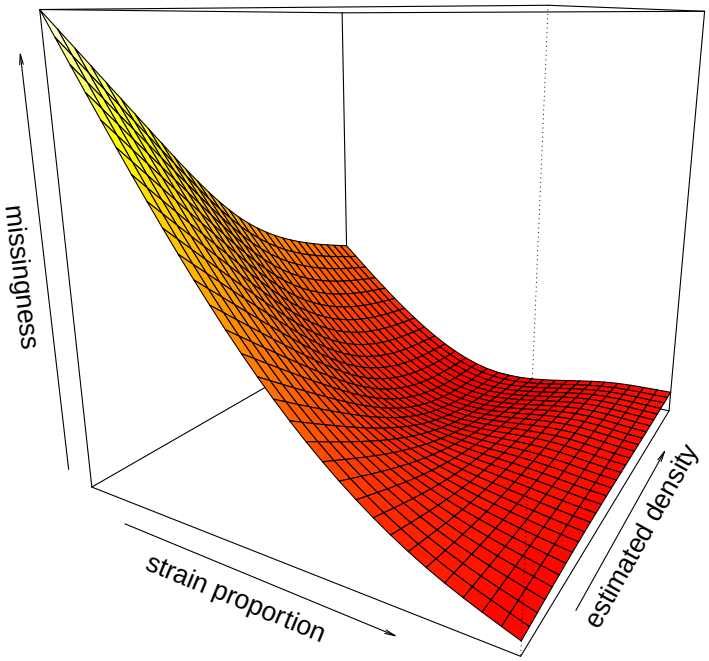


Figure S.6: A model fit to predict detection probability from sample parasite density and within-sample strain proportion using empirical data from mixed-strain controls taken from [24]. A general additive model (GAM) with full tensor product smooth was used for the fit, with parasite density and proportion on the log scale. Smaller within-host proportion and overall parasite density resulted in more undetected alleles.

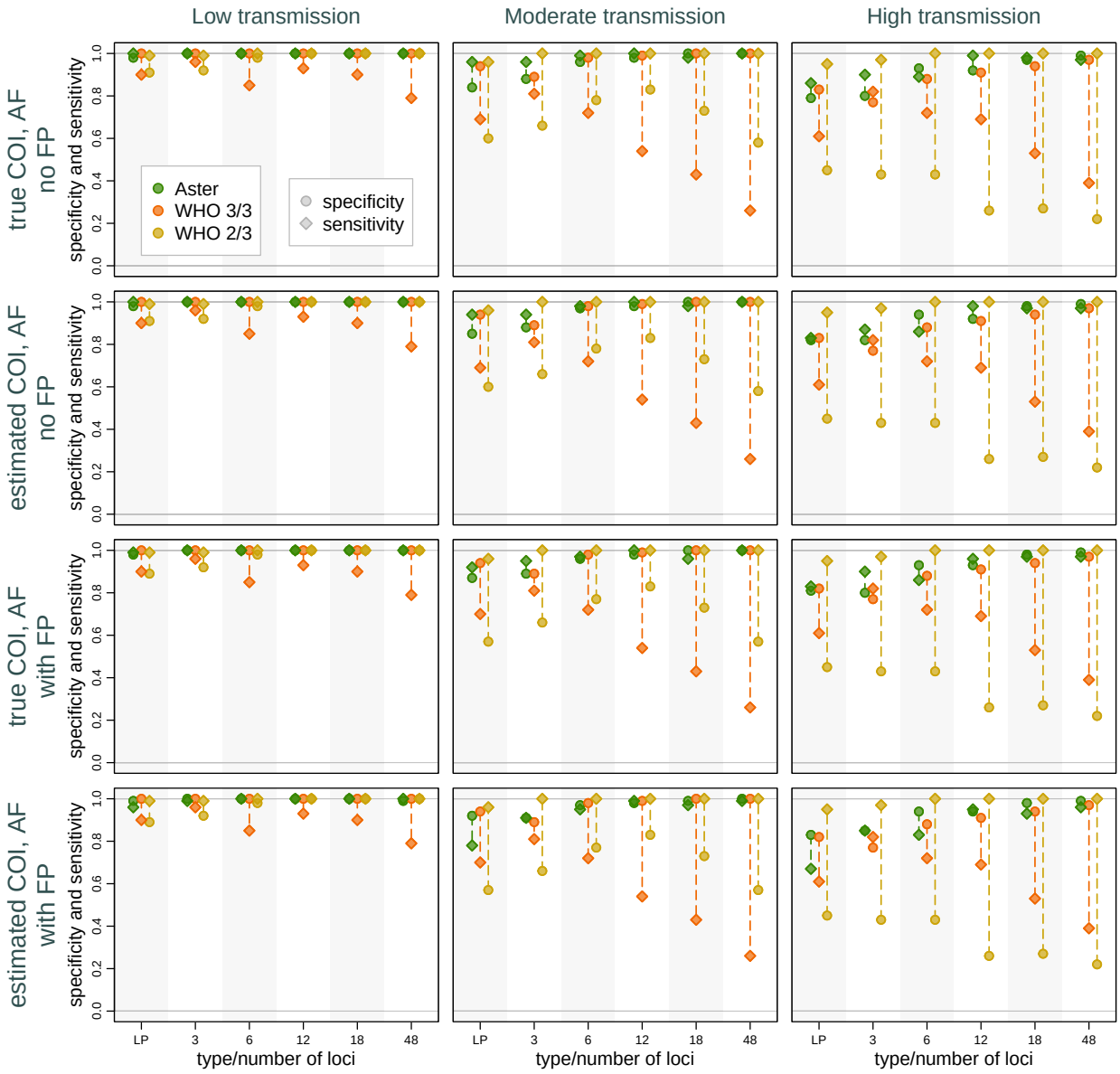


Figure S.7: Sensitivity and specificity of Aster and match-counting algorithms exploring the effect of using estimated vs true COI and population allele frequencies (AF) and of adding simulated false positive alleles (FP). Allele frequencies were estimated from 100 D0 samples. False positive rates were 0.1 for the length polymorphism (LP) panel and 0.02 for the amplicon sequencing panels.

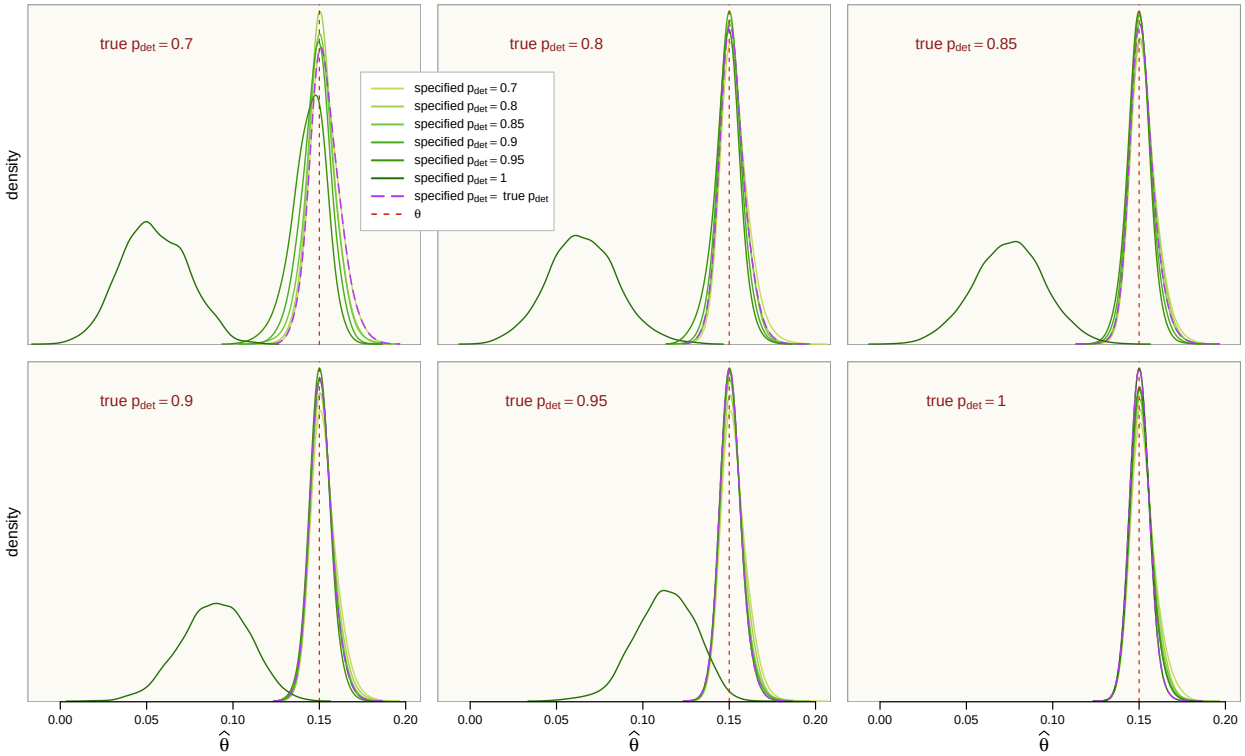


Figure S.8: Distribution of failure rate estimates with Aster when using misspecified allele detection probability (p_{det}). Each panel features one value of true detection probability evaluated with a range of specified values, one of which matched the true probability (purple dotted line marking the corresponding density). Study-level simulation settings were moderate transmission intensity, failure rate (θ) of 0.15 (vertical red line), and 0 background relatedness. A 12-locus MAD^{4}HatTeR panel was used for the simulations.