coiaf:
Complexity Of Infection Estimation with Allele Frequencies

A thesis submitted in partial fulfillment of the requirements for the degree of Bachelor of Science with Honors in Computational Biology

Aris Paschalidis
Brown University
April 2, 2021

CENTER FOR
Computational
Molecular Biology

coiaf: Complexity Of Infection Estimation with Allele Frequencies

Aris Paschalidis [*]

April 2, 2021

Abstract

In malaria, individuals are often infected with different parasite strains. These multi-clonal infections occur either because an individual received repeated mosquito bites from infectious mosquitoes or because they received a single bite from an infectious mosquito harboring multiple parasites. Such mixed infections may be composed of genetically unrelated strains or parasites that are related. The number of genetically different parasite strains in an individual, known as the complexity of infection (COI), provides important insight into the severity of infection and the parasite population. Probabilistic likelihood and Bayesian models have been developed to estimate the COI, but rapid direct measures such as heterozygosity or *FwS* have not been directly related to the COI. Moreover, current methods have utilized statistics to determine the relative COI that rely on the assumption that parasites in mixed infections are unrelated. In this paper we present two new methods that use easily computable measures to directly estimate the COI from sequence read depth data. Our methods are computationally efficient and are proven to be comparably accurate to current methods in literature.

Keywords: Malaria; *P. falciparum*; Complexity of Infection (COI); Allele Frequency; Direct Measure; Optimization.

[*] In collaboration with and advised by Dr. Oliver Watson, Department of Infectious Disease Epidemiology, Imperial College London, London, UK, and Dr. Jeffrey Bailey, Department of Pathology and Laboratory Medicine, Brown University, Providence, RI, USA

Contents

1 Introduction

In 2015, the World Health Organization (WHO) released its global technical strategy for malaria 2016-2030, the newest in a set of reports outlining a strategy to reduce the malaria burden in the world [1]. The report set several ambitious goals for the year 2030 including the reduction of malaria mortality by 90% compared to the rates in 2015. To reach this goal, the authors set a milestone of reducing mortality by 40% by 2020. Unfortunately, as the year 2020 has come and gone, this milestone has not yet been reached [2].

Malaria remains a leading cause of death worldwide—in 2019, there were an estimated 229 million cases and 409,000 deaths around the globe [2]. Despite the considerable burden malaria causes today, these numbers represent substantial progress that has been made globally to control malaria in the last two decades. In fact, the WHO reports that 1.5 billion malaria cases and 7.6 million malaria deaths have been averted globally in the period of 2000–2019 [2]. The majority of these gains reflect the increase of vector control initiatives [3–5], the development of highly efficacious antimalarial combination therapies [6–8], and improved case management through the deployment of rapid diagnostic tests (RDTs) [9–13].

However, there is evidence that progress has slowed and that there is a need for new approaches to capitalize on the gains made [2]. One approach is the use of computational methods to help inform eradication efforts [14–16], which often rely on recent advances in genetic sequencing and an increased understanding of malaria biology. For example, the use of computational methods can been used in the identification of polyclonal infections—malarial infections with multiple distinct strains [17, 18]. Such polyclonal infections introduce additional genetic complexity that is often difficult to computationally account for. As a result, many of the population genetic tools often applied for the study of other organisms are unsuitable for studying malaria and researchers often rely on limiting genetic analyses to individuals who are monoclonally infected [19].

In order to account for these limitations, a new field of work has emerged focusing on the complexity of infection (COI). The COI represents the number of genetically distinct malaria strains that can be identified in a particular individual. Polyclonal infections arise from one of two ways: (*i*) a single infectious mosquito feeds on a human host transferring several parasite genomes (often referred to as a co-transmission event [20, 21]) or (*ii*) an individual is infected by two or more infectious mosquitoes with distinct malarial genotypes (known as a superinfection event [21, 22]). Transmission intensity has been shown to impact the contribution of each event towards the generation of within-host parasite genetic diversity [23]. Superinfection is modulated by the host's current infections [24] as well as their age and exposure acquired immunity [25]. For these reasons, measures of genetic diversity and COI are increasingly used for inferring malaria transmission intensity and evaluating malaria control interventions [26].

2 Related work

Traditionally, the COI has been referred to as the multiplicity of infection (MOI) and has been estimated using PCR amplification—a method to selectively amplify DNA or RNA targets. In particular, the presence of different alleles at genes encoding the merozoite surface protein 1 and 2 (MSP1 and MSP2), surface proteins found on the merozoite stage of the malarial parasite [27, 28], can provide an estimate of the COI. However, such methods are hindered by limitations on the number of loci examined [29, 30], and lack the ability to detect parasites at low parasitaemia in samples [31].

With advances in sequencing technologies, new methods have been developed to address these concerns. Two early proposed methods: the *FwS* metric suggested by Auberun *et al.*, which characterizes within-host diversity and its relationship to population-level diversity [29], and the *estMOI* software, which utilizes phasing information of alleles to estimate the COI [31], are a step forward. However, the *FwS* metric does not provide a direct estimate of the COI and the *estMOI* software relies on deep whole genome shotgun sequence data, which is costly to generate at a large scale [30]. Alternative solutions to estimating the COI include the *DEploid* software package which uses haplotype structure to infer the number of strains, their relative proportions, and the haplotypes present in a sample [19]. Other methods have been developed to examine the relatedness between parasite strains [32, 33]. The current "gold-standard" method for determining the COI of a sample is *THE REAL McCOIL*, which was built on the initially published *COIL* method [30]. *THE REAL McCOIL* employs a Bayesian approach, turning heterozygous data into estimates of allele frequency using Markov chain Monte Carlo [34]. Furthermore, it estimates the COI of a sample using likelihood [34]. However, this approach is computationally expensive.

Despite various methods for estimating the COI, no rapid direct measures have been developed that directly relate allelic read depth data to the COI, allowing for the easy estimation of the COI. In this thesis, we present two new methods that use easily calculable measures to directly estimate the COI from sequence read depth data. Our methods are computationally efficient and are proven to be comparably accurate to current methods in literature. Furthermore, our methods provide a framework for gaining insight into the relatedness of strains in mixed infections. Using our methods, we have developed the software package *coiaf* in the programming language R, a language and environment for statistical computing and graphics [35].

The remainder of this thesis is organized as follows. In Section 3 we develop the formulation of the complexity of infection problem. Section 4 presents the methodologies we use to solve the problem. Section 5 describes the data we use to test and investigate our models. Section 6 presents our findings on simulated data and Section 7 illustrates our findings on real data. Furthermore, in Section 8 we present new findings indicating the COI across the globe. Additional mathematical formulations and figures are delegated to the Appendices.

3 Problem Formulation

Current gold-standard approaches for estimating COI rely on identifying the number of different parasites present in an infection using high-throughput sequencing, equivalently referred to as next-generation sequencing. In monoclonal infections, composed of only one parasite strain, all sequence reads will be from the same parasite strain. However, in mixed infections, a combination of each parasite strain will contribute to observed sequence reads. At genetic loci with prevalent single-nucleotide polymorphisms (SNPs), there is an increased chance of observing multiple alleles. The likelihood of observing a mix of nucleotides at any locus is dependent on both the number of parasite strains in an infection as well as how prevalent polymorphisms are. This is the basis on which the methods *COIL* [30] and *THE REAL McCOIL* [34] are derived and is also where we start our formulation for a direct estimate of the COI.

For the purposes of our formulation, we focus on biallelic loci—loci at which only two nucleotides are observed. In doing so, we may define the major allele as the allele that is most prevalent in a population, and the minor allele as the allele that is least prevalent in a population. Assuming for any individual there are l biallelic loci, we define the population-level allele frequency (PLAF) as an l-dimensional vector $\mathbf{p}$ composed of the frequencies of the minor allele at each locus across a population, namely $\mathbf{p} = (p_1, \dots, p_l)$, where $p_i \in [0, 0.5]$. Furthermore, we define the within-sample minor allele frequency (WSAF) as the frequency of the minor allele at each locus for a single individual, where the minor allele is defined as the minor allele in the population. For example, the WSAF will be equal to 1 when all sequence reads observed at a given locus are the minor allele defined for the population.

3.1 Number of Strains Containing the Minor Allele

We first derive an expression for the number of strains that have the minor allele. At any given locus i, let N_i be the random variable corresponding to the number of strains with the minor allele such that $N_i \in \{0, \ldots, k\}$, where k is the COI. Assuming that parasite strains in a mixture are unrelated, alleles in a sample are drawn at random from the PLAF. As a result, N_i follows a binomial distribution:

$$\mathbb{P}(N_i = n_i) = \binom{k}{n_i} p_i^{n_i}(1 - p_i)^{k-n_i}, \quad n_i \in \{0, \ldots, k\}. \tag{1}$$

3.2 Method 1: Probability a Locus is Heterozygous

In Method 1, we examine all loci and wish to identify the probability of locus i being heterozygous. Recall that heterozygous sites contain two distinct alleles: the major and minor alleles. We denote by V_i, a random variable which takes the value of 1 if a site is heterozygous and 0 otherwise. Therefore, the probability that locus i is heterozygous, written as $\mathbb{P}(V_i = 1)$, is equal to 1 minus the probability that a locus is homozygous, i.e., all strains contribute to the minor allele or all strains contribute to the major allele. Using Equation (1), we thus write,

$$\mathbb{P}(V_i = 1) = 1 - p_i^k - (1 - p_i)^k, \tag{2}$$

where p_i is the population-level allele frequency of the minor allele, $1 - p_i$ is the population frequency of the major allele, and k is the COI. As the COI increases, the probability of observing a heterozygous locus also increases, as described by Equation (2) and shown in Figure 1A.

3.3 Method 2. Expected within-sample Minor Allele Frequency

As defined earlier, let p_i be the PLAF of the minor allele at locus i. Furthermore, let k be the COI of the sample. We wish to examine all heterozygous loci. If locus i is heterozygous, i.e., $V_i = 1$, the number of copies of the minor allele at i must be $1 \leq n_i \leq k - 1$.

Next, suppose we have k malarial parasite strains. At each locus, we denote the number of times each strain was sampled with a vector of random variables $\mathbf{Y}_i = (Y_{i_1}, \ldots, Y_{i_k})$. We can model the number of samples of each strain using a Multinomial distribution. Thus the expected number of samples of strain j at locus i can be represented as

$$\mathbb{E}[Y_{i_j}] = m_i s_{i_j}, \tag{3}$$

where m_i is the number of samples drawn at locus i and s_{i_j} is the probability of drawing strain j at locus i such that $s_{i_j} \geq 0, \forall j$ and $\sum_{j=1}^{k} s_{i_j} = 1$. We can further treat the proportion of each strain as a random variable, i.e., $\mathbf{S}_i = (S_{i_1}, \ldots, S_{i_k})$ such that $S_{i_j} \geq 0, \forall j$ and $\sum_{j=1}^{k} S_{i_j} = 1$, described by a Dirichlet distribution. Combining our two distributions, we obtain a Dirichlet-multinomial distribution. In essence, at each locus, we draw from the k strains using a Multinomial distribution, where the probability of sampling each strain is given by the Dirichlet distribution. Using the law of iterated expectations, we represent the expected value of this compound distribution as

$$\mathbb{E}[Y_{i_j}] = m_i \mathbb{E}[S_{i_j}]. \tag{4}$$

Recall that the expected value of S_{i_j} can be given by the expected value of the Dirichlet distribution:

$$\mathbb{E}[S_{i_j}] = \frac{\alpha_{i_j}}{\sum_k \alpha_{i_k}}, \tag{5}$$

where α_{i_j} is the concentration or scaling parameter of the Dirichlet distribution of strain j at locus i. Combining Equation (4) and Equation (5), we obtain the expected value of our compound distribution:

$$\mathbb{E}[Y_{i_j}] = m_i \frac{\alpha_{i_j}}{\sum_k \alpha_{i_k}}. \tag{6}$$

Given the expected number of samples of strain j at locus i, we can find the expected WSAF of the minor allele by measuring the expectation over all strains. Furthermore, we are interested in strains that contribute to the minor allele and as a result, the number of samples will be n_i at locus i. We denote the WSAF at locus i as W_i and can define the expected WSAF given n_i strains contribute to the minor allele at locus i as

$$\begin{aligned}\mathbb{E}[W_i|N_i = n_i] &= \frac{1}{k}\sum_{j=1}^{k} n_i \frac{\alpha_{i_j}}{\sum_k \alpha_{i_k}} \\ \mathbb{E}[W_i|N_i = n_i] &= \frac{n_i}{k}. \end{aligned} \tag{7}$$

In order to find the expected value of the WSAF given that locus i is heterozygous, we can use the law of total expectation:

$$\begin{aligned}\mathbb{E}[W_i|V_i = 1] &= \mathbb{E}[\mathbb{E}[W_i|V_i = 1, N_i = n_i]] \\ &= \sum_{n_i=0}^{k} \mathbb{E}[W_i|V_i = 1, N_i = n_i]\mathbb{P}(N_i = n_i) \\ &= \sum_{n_i=0}^{k} w_i \mathbb{P}(W_i|V_i = 1, N_i = n_i)\mathbb{P}(N_i = n_i) \\ &= \sum_{n_i=0}^{k} w_i \mathbb{P}(W_i, V_i = 1|N_i = n_i)\frac{\mathbb{P}(N_i = n_i)}{\mathbb{P}(V_i = 1)}. \end{aligned} \tag{8}$$

We note that $\mathbb{P}[W_i, V_i = 1|N_i = n_i] = 0$ if $n_i = 0$ or $n_i = k$. Therefore,

$$\begin{aligned}\mathbb{E}[W_i|V_i = 1] &= \sum_{n_i=1}^{k-1} w_i \mathbb{P}(W_i, V_i = 1|N_i = n_i)\frac{\mathbb{P}(N_i = n_i)}{\mathbb{P}(V_i = 1)} \\ &= \sum_{n_i=1}^{k-1} \mathbb{E}[W_i|N_i = n_i]\frac{\mathbb{P}(N_i = n_i)}{\mathbb{P}(V_i = 1)} \\ &= \sum_{n_i=1}^{k-1} \frac{n_i}{k} \frac{\binom{k}{n_i} p_i^{n_i}(1-p_i)^{k-n_i}}{1 - p_i^k - (1-p_i)^k} \\ &= \frac{k^{-1}}{1 - p_i^k - (1-p_i)^k} \sum_{n_i=1}^{k-1} n_i \binom{k}{n_i} p_i^{n_i}(1-p_i)^{k-n_i}. \end{aligned} \tag{9}$$

We note that the sum in Equation (9) represents the expected value of the binomial distribution less the case where $n = 0$ or $n = k$. Thus,

$$\begin{aligned}
\sum_{n_i=1}^{k-1} n_i \binom{k}{n_i} p_i^{n_i} (1-p_i)^{k-n_i} &= \sum_{n_i=0}^{k-1} n_i \binom{k}{n_i} p_i^{n_i} (1-p_i)^{k-n_i} \\
&= \sum_{n_i=0}^{k} n_i \binom{k}{n_i} p_i^{n_i} (1-p_i)^{k-n_i} - kp_i^k \\
&= kp_i - kp_i^k. \qquad (10)
\end{aligned}$$

Substituting Equation (10) into Equation (9), we obtain

$$\mathbb{E}[W_i|V_i = 1] = \frac{p_i - p_i^k}{1 - p_i^k - (1-p_i)^k}. \qquad (11)$$

Note that Method 1 and Method 2 both depend on the COI and the population-level allele frequency. However, Method 1 identifies the probability of a locus being heterozygous, denoted as $\mathbb{P}(V_i = 1)$, whereas Method 2 identifies the expected value of the within-sample allele frequency given a site is heterozygous, denoted as $\mathbb{E}[W_i|V_i = 1]$.

4 Estimation Method

Given data, $D : \{(p_i, w_i), \quad i = 1, \ldots, l\}$, where p_i is the PLAF at locus i and w_i is the WSAF at locus i, we next explore our methods to approximate the COI of a sample.

4.1 Data Processing

Data are first processed to account for sequencing errors. The amount of sequence error assumed may either be provided by the user or inferred directly from the distribution of the data. For additional information, see Appendix A.

Following adjustment for sequence error, consider an arbitrary data point (p_i, w_i). We define a partition over the range of p_i constituent of sets $\mathcal{P}_1, \ldots, \mathcal{P}_N$ such that $p_i \in \mathcal{P}_m$ if $p_i \in [a_m, b_m), \quad a_m < b_m, \ i = 1, \ldots, l$. In doing so, we group our data into N bins. Furthermore, we define $\hat{p}_m$ as

$$\hat{p}_m = \frac{b_m - a_m}{2}. \qquad (12)$$

$\hat{p}_m$ can therefore be thought of as the midpoint of each group of data points. In order to ensure the presence of sufficient data in each set of our partition, we adjust the values of a_m and b_m such that the number of data points in $\mathcal{P}_m$ is greater than c, i.e., $|i : p_i \in \mathcal{P}_m| > c$, where c is a predefined size.

Recall that Method 1 and Method 2 examine different random variables. Specifically, Method 1 identifies the probability of a locus being heterozygous, $\mathbb{P}(V_i = 1)$ and Method 2 identifies the expected value of the WSAF given a site is heterozgyous, $\mathbb{E}[W_i|V_i = 1]$. When approximating the COI using Method 1, we define $\hat{t}_{1,m}$ as the average v_i over $\mathcal{P}_m$, using the subscript 1 to indicate Method 1.

$$\hat{t}_{1,m} = \frac{\sum_{i=1}^{l} \mathbb{1}_{\{p_i \in \mathcal{P}_m\}} v_i}{|i : p_i \in \mathcal{P}_m|}, \qquad (13)$$

where $v_i = 1$ if locus i is heterozygous and $\mathbb{1}$ is the indicator function, such that $\mathbb{1}_{\{p_i \in \mathcal{P}_m\}} = 1$ when locus i is in set $\mathcal{P}_m$. $\hat{t}_{1,m}$ therefore represents the mean number of loci with PLAF less than b_m and greater than or equal to a_m, i.e., within the set $\mathcal{P}_m$, that are heterozygous.

For Method 2, we define $\hat{t}_{2,m}$ as the mean WSAF for all heterozygous loci with PLAF less than b_m and greater than or equal to a_m, i.e., within the set $\mathcal{P}_m$:

$$\hat{t}_{2,m} = \frac{\sum_{i=1}^{l} \mathbb{1}_{\{p_i \in \mathcal{P}_m\}} w_i}{|i : p_i \in \mathcal{P}_m|}, \tag{14}$$

where w_i is the WSAF at locus i.

4.2 Optimization Problem

In order to determine the COI, we utilize the expressions Equation (2) and Equation (11). For the sake of simplicity, we define Equation (2) as $f_1(\mathbf{p})$ and Equation (11) as $f_2(\mathbf{p})$:

$$f_1(\mathbf{p}) \triangleq \mathbb{P}(V_i = 1) = 1 - p_i^k - (1 - p_i)^k \tag{2}$$

$$f_2(\mathbf{p}) \triangleq \mathbb{E}[W_i | V_i = 1] = \frac{p_i - p_i^k}{1 - p_i^k - (1 - p_i)^k}. \tag{11}$$

To estimate the COI, we next solve the following minimization problem for Method 1:

$$\min_k \left(\sum_{m=1}^{N} |(\hat{t}_{1,m} - f_1(\hat{p}_m)|^q \right)^{1/q}, \tag{15}$$

and the following minimization problem for Method 2:

$$\min_k \left(\sum_{m=1}^{N} |(\hat{t}_{2,m} - f_2(\hat{p}_m)|^q \right)^{1/q}, \tag{16}$$

where $q \geq 1$. In our implementation, the default value of q is 2, corresponding to the ℓ_2 norm.

4.3 Solution Methods

We solve this optimization problem using two methods: (*i*) we solve the optimization problem using only discrete values of the COI and (*ii*) we solve the optimization problem using continuous values of the COI. Recall that the COI is defined as the number of genetically distinct malaria parasite strains an individual is infected with. As such, a continuous value for the COI does not have biological meaning—only discrete values for the COI are applicable. However, in the real world, deviations from our assumptions may be present. When this is the case, our discrete estimate will not be accurate. A continuous estimate, although not biologically interpretable as the COI, may provide more information on potential causes of the deviation from our assumptions. For example, our model assumes that parasite strains are unrelated. However, in the real world, relatedness in mixed infections is common [33]. A continuous COI may be able to provide insight into the level of relatedness between malarial strains.

4.3.1 Discrete Optimization

The simplest method to solve the discrete versions of the aforementioned optimization problems is through the use of a brute force approach—for our optimization problems, this involves computing the objective function for each COI considered. As brute force approaches can be computationally inefficient, we limit the range of values of the COI. In particular, we set the default range of COIs to examine from 1 till 25. In addition to the estimated COI, our discrete optimization method also returns a measure of the confidence of the result. Let us assume that we optimize over k COI values. For any given COI, i, let o_i be the result of our objective function. As we aim to reduce the value of our objective function, we first compute the reciprocal of each o_i: $1/o_i$. We next normalize $1/o_i$ by dividing by the sum of the vector $\mathbf{o}$, which represents the value of the objective function at all k COIs. In doing so, we generate a probability density function, which can provide insight into the confidence of our results.

4.3.2 Continuous Optimization

To solve the continuous versions of the optimization problems, we utilize R's built in optimization function [35]. This function provides general-purpose optimization based on several optimization algorithms. For our purposes, we utilize an L-BFGS-B approach to solve our optimization problem [36]. This method builds on the BFGS quasi-Newton algorithm for solving optimization problems [37–40]. L-BFGS-B provides an added benefit over BFGS as it allows for box constraints, i.e., each variable can be given an upper and lower bound. In using this method, we may thus restrict our estimated COI to a certain range. As when solving the discrete versions of our optimization problems, we set the default range of COIs to examine from 1 to 25.

5 Data

To evaluate the performance of our methods, we utilize simulated data, which allow us to evaluate the accuracy and sensitivity of our methods, and sequencing data sampled from infected individuals collected worldwide, which aids in the comparison to the current state-of-the-art metric. Furthermore, using data sampled from individuals worldwide, we investigate the distriibution of COI across the world.

5.1 Simulator Details

In order to test our mathematical methods, we created a simulator that generates synthetic sequencing data for a number of individuals in a given population. Each individual is assigned a COI value, which is used to simulate the number of sequence reads mapped to the reference and alternative allele for the biallelic SNPs considered. These are then used to derive the within-sample frequency of the population-level minor allele, w_i, at each locus.

We first define the number of loci being simulated, l. We assume that the distribution of reference allele frequencies for locus i, R_i, is described by a Beta distribution with shape parameters α and β:

$$\mathbb{P}(R_i = r_i) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} r_i^{(\alpha-1)}(1-r_i)^{(\beta-1)},$$

where Γ is the Gamma function. In our simulations, we assume $\alpha = 1$ and $\beta = 5$ and draw l values of R_i. In Equation (2) or Equation (11), we have defined relationships with respect to the frequency of the minor allele, p_i. Recall that $p_i \in [0, 0.5]$. Thus, we define

$$p_i = \begin{cases} r_i & r_i \leq 0.5 \\ 1 - r_i & r_i > 0.5. \end{cases}$$

To simulate w_i for each individual, we first assign the COI, k. Next, we simulate the genotype at locus i, drawing k values from a Binomial distribution with probability p_i, i.e., the probability of having the minor allele at locus i is equal to the population-level frequency of the minor allele. This yields the matrix, $\mathbb{G} \in \mathbb{R}^{k \times l}$, with k rows and l columns defining the phased haplotype of each strain, where $\mathbb{G}_{i,j} = 1$ indicates that strain j at locus i has the minor allele.

To determine the number of sequence reads related to each strain, we first draw the proportion of each strain, $\mathbf{S} = (S_1, \ldots, S_k)$, in an individual with k strains, which we assume is described by a Dirichlet distribution with concentration parameters $\alpha_1, \ldots, \alpha_k$. The "true" WSAF of the minor allele at locus i, denoted τ_i, is thus given by the sum of the strain proportions for strains with the minor allele. We can express this using matrix multiplication:

$$\boldsymbol{\tau} = \mathbf{S} \times \mathbb{G}_i,$$

where $\boldsymbol{\tau} \in \mathbb{R}^{1 \times l}$, $\mathbf{S} \in \mathbb{R}^{1 \times k}$, and $\mathbb{G} \in \mathbb{R}^{k \times l}$.

In simulations with no assumed sequence error, the number of sequence reads with the minor allele at each locus is given by sampling from a Binomial distribution c times with probability $\boldsymbol{\tau}$, where c is the assumed number of sequence reads, i.e., the read depth or coverage. However, in simulations with sequence error, we perturb the "true" WSAF by assuming that a number of sequence reads are incorrectly sampled. An incorrectly sampled locus with the major allele will yield the minor allele and vice versa. In our simulations, we assume a fixed sequence error, ϵ, such that the probability of correctly sampling the minor allele is $1 - \epsilon$, and the probability of incorrectly sampling the minor allele is ϵ. Therefore, we can represent the probability of sampling the minor allele, now denoted as $\boldsymbol{\tau}_\epsilon$ to indicate the added error, as

$$\boldsymbol{\tau}_\epsilon = \boldsymbol{\tau}(1 - \epsilon) + (1 - \boldsymbol{\tau})\epsilon.$$

Lastly, we may account for overdispersion—additional unexpected variability in our data—and sample the number of sequence reads with the minor allele at each locus from a Beta-binomial distribution rather than a Binomial distribution. Dividing the number of instances of the minor allele by the coverage at each locus results in the simulated WSAF.

5.2 Real Data

To apply our developed methods to real whole genome sequence data, we analysed samples from the MalariaGEN *Plasmodium falciparum* Community Project [41]. The MalariaGEN *Plasmodium falciparum* Community Project provides genomic data from over 7,000 *P. falciparum* samples from 28 malaria-endemic countries in Africa, Asia, South America, and Oceania from 2002-2015 [41]. Detailed information about the data release including brief descriptions of contributing partner studies and study locations is available in the supplementary of MalariaGEN *et al.* [41].

5.2.1 Pre-processing

Samples collected from field isolates were initially sequenced. The median depth of coverage was 73 sequence reads averaged across the whole genome and across all samples. Human-derived reads were removed, samples were mapped to the *P. falciparum* 3D7 reference genome, and variants called using GATK (Genome Analysis Toolkit) best practices [42, 43]. After removing replicate samples as well as samples with low coverage, suspected contamination or mislabelling, and mixed-species samples, 5,970 samples remained for further analysis.

After selecting the 5,970 samples, we first filtered the genomic data for high quality biallelic coding and non-coding SNPs as outlined in [33]. We next filtered to variants with an alternative allele frequency of greater than 0.05—sequence error likely complicates the detection of true variation stemming from multiple strains for variants at lower frequencies.

To apply our developed methods for estimating COI, we need to estimate the population-level frequency of the minor allele. Consequently, we sought to assign the samples to a suitable number of geographic regions that would be likely to have similar genetic diversity, while ensuring a sufficient number of samples per region for reliable estimation of the population-level allele frequencies. We used the Partitioning Around Medoids (PAM) algorithm to solve a k-medoids clustering problem [44,45] in order to group samples based on their longitude and latitude of sample collection. We next calculated the silhouette information for each clustering of k groups [46], arriving at 24 locations globally (see Appendix B for a map of locations).

5.2.2 *THE REAL McCOIL* Runs

To compare our method against the "gold-standard" for estimating COI, *THE REAL McCOIL*, we first had to estimate the COI using *THE REAL McCOIL* [34]. The total number of variants across the genome is computationally untractable for *THE REAL McCOIL* and is unsuitable for the algorithm given that many SNPs will be in linkage disequilibrium (LD), the non-random association of alleles at different loci in a given population, with each other. In response, we first identified a common set of SNPs such that the minor allele frequency of variants was greater than 0.05 globally and within each of the 24 identified regions. From this set of SNPs we selected subsets of SNPs that were not in LD by first calculating the genetic autocorrelation [47] among SNPs by genetic distance and filtering to sets of SNPs with autocorrelation less than 0.8. For each of the 24 regions, 5 unique sets of LD filtered variants were created. The COI was estimated for each set using both the methods developed in this analysis as well as using *THE REAL McCOIL* (v2) with details outlined in [48]. In overview, the *THE REAL McCOIL* categorical method was used, with heterozygous loci called using the genotype called during variant calling. Default priors were assigned for each parameter used by *THE REAL McCOIL*, with a maximum observable COI equal to 25 and sequencing measurement error estimated along with COI and allele frequencies. COI estimates were compared between regions using 100,000 repetitions of a non-parametric bootstrap [49, 50] to estimate the 95% confidence interval from the bootstrapped COI density.

6 Performance on Simulated Data

6.1 Overall Performance

We initially present the results of our findings on data sequenced from 1,000 loci with a read depth of 400 at each locus. Data was simulated ranging from a complexity of infection of 1 to 25. No error was introduced to our simulator and as a result, our methods accounted for no sequencing error. The results of running the discrete version of both Method 1 and Method 2 are illustrated in Figure 1C and Figure 1D, respectively.

Method 1 and Method 2 perform well for all COIs between 1 and 25. Notably, the lower the true value of the COI, the better our models perform with a mean absolute error close to 0. As the

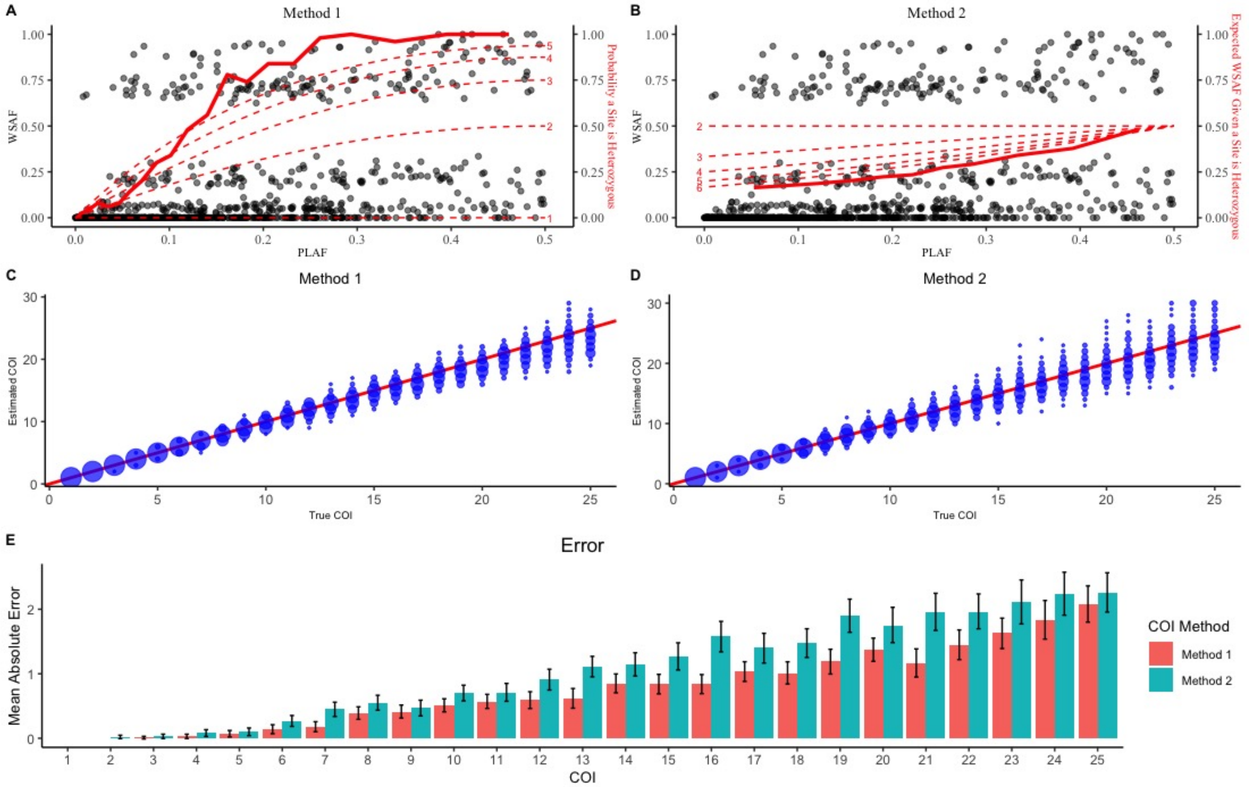
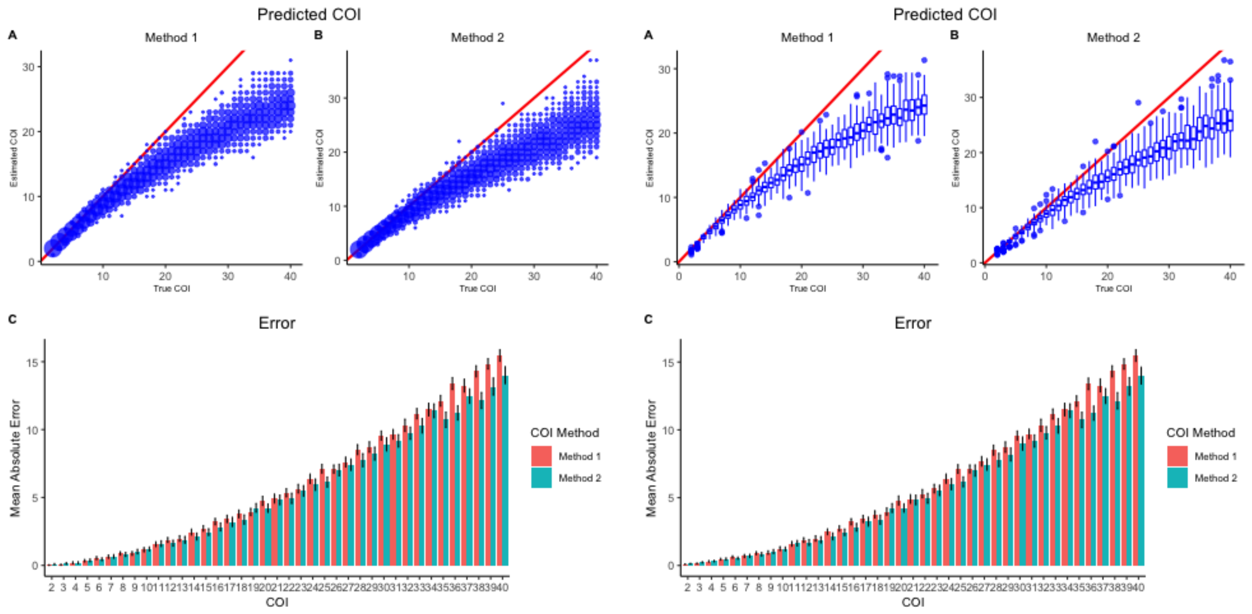


Figure 1. Estimating the COI on simulated data. In (A) and (B) the relationship between the within-sample minor allele frequency (WSAF) and the population-level allele frequency (PLAF) is shown for an example simulation with a COI of 4. In (A), all loci are shown whereas in (B), only heterozygous loci are plotted. The theoretical relationships for a COI of 1-5 (A) and 2-6 (B) are shown as dashed red lines for Method 1 (A) and Method 2 (B). The solid red line represents the vectors we examine in our optimization problems, namely $\mathbf{t}_{1,m}$ and $\mathbf{t}_{2,m}$. In other words, in (A), the solid red line indicates the mean number of loci that are heterozygous in each PLAF bin and in (B), the solid red line indicates the mean WSAF for all heterozygous loci in each PLAF bin. The performance of Method 1 (C) and Method 2 (D) is shown for 100 simulations of a COI of 1-25 with 1,000 loci, a read depth of 400, no error added to the simulations, and no sequencing error assumed. Point size indicates density, with the red line representing the line $y = x$. The mean absolute error for each method is shown in (E). The black bars indicate the 95% confidence interval.

COI increases, our models exhibit more variability across subsequent iterations. For example, at a COI of 25, the estimated COI from Method 2 ranges from 20 to 30, the maximum COI our model could output in these trials (see Figure 1D). Nevertheless, the majority of predicted COIs remain close to the true COI as witnessed by the low mean absolute errors of at most 2 (see Figure 1E).

Comparing our two methods, Method 1 appears to outperform Method 2. In particular, Method 1 exhibits less variability than Method 2. Furthermore, Method 1 tends to have a slightly lower mean absolute error than Method 2. However, the mean absolute error's 95% confidence intervals overlap and as such the difference in the performance of the two methods is not statistically significant.

The difference in performance between Method 1 and Method 2 is further amplified when data is simulated for a larger range of COI values. Figure 2 illustrates the performance of Method 1 and Method 2 for COI values from 1 to 40. Although a COI of 40 is very unrealistic and not likely to be seen in a real sample, by examining the performance of our algorithms in extreme situations, we can better understand the strengths and weaknesses of our models.



(a) Discrete Method

(b) Continuous Method

Figure 2. Performance with Changing COI. The performance of Method 1 and 2 is compared across a COI of 1-40. Data was simulated 100 times from 1,000 loci with a coverage of 200. No error was introduced into our simulations and a sequence error of 1% was assumed. In (a), the discrete estimation is shown and in (b), the continuous estimation is shown. In both cases, the distribution of estimated COI for Method 1 (A) and Method 2 (B) is plotted. In (a), dot plots are used where point size indicates density. In (b), a boxplot shows the interquartile range, with whiskers indicating the 95% quantile range and outliers indicated with dots. The mean absolute error is shown in (C) for both methods, with the 95% quantile computed from 1,000 bootstrap repetitions at each COI indicated with error bars.

Again visible is the fact that Method 1 tends to outperform Method 2, especially at high values of the COI. Interestingly, it is also apparent that at high COI values, both Method 1 and Method 2 tend to underpredict the true value of the COI, with mean absolute errors around 15 when the true value of the COI is 40. Looking at Figure 1A and Figure 1B, this underprediction is expected. In particular, recall that the PLAF is bounded by $[0, 0.5]$ and that the output of Method 1 and Method 2 is bounded by $[0, 1]$. As the COI increases, the value of the objective function for subsequent COIs becomes more similar. This occurs for both Method 1 and Method 2. Visually speaking, this behavior can be witnessed by the dashed red lines in Figures 1A and 1B appearing closer. As a result, at larger COI values, it is harder for our algorithm to solve our optimization problems and consequently, we are less confident in our estimate of the COI. Additionally, at higher COIs it becomes increasingly likely that each strain does not contribute to the observed sequence reads. This will be increasingly likely as the sequence coverage decreases. For example, in extreme cases when the COI is greater than the sequence coverage, it is impossible for each strain to have been sequenced at any given locus. As a result, it becomes increasingly combinatorially probable that some strains will be underrepresented in the sequencing data and thus will lead to lower inferred COIs. This will be more pronounced if the proportions of each strain are heavily skewed.

6.2 Sensitivity Analysis

In order to understand the sensitivity of our models to alterations in the parameters considered, we rigorously tested the performance of the discrete and continuous representations of Method 1 and Method 2, assessing changes in the accuracy of our predictions. For each sample, we utilized bootstrapping techniques to determine the mean absolute error of the predicted COI compared to the true COI. Furthermore, we ran each sample several times to ensure reliable results. A description of several key parameters perturbed and their default values can be found in Table 1. Unless otherwise specified, simulations used the default values presented. Additional figures may be found in Appendix C.

Parameter	Description	Default Value
COI	The complexity of infection	Estimated
Coverage	The coverage at each locus	200
Loci	The number of loci examined across the genome	1,000
Alpha	The alpha parameter used to generate strain proportions from a Dirchlet distribution	1
Overdispersion	The extent to which counts are overdispersed relative to the Binomial distribution when simulating data	0
Epsilon	The probability of a single read being mis-called as the other allele	0
Sequence Error	The level of sequencing error that is assumed	1%

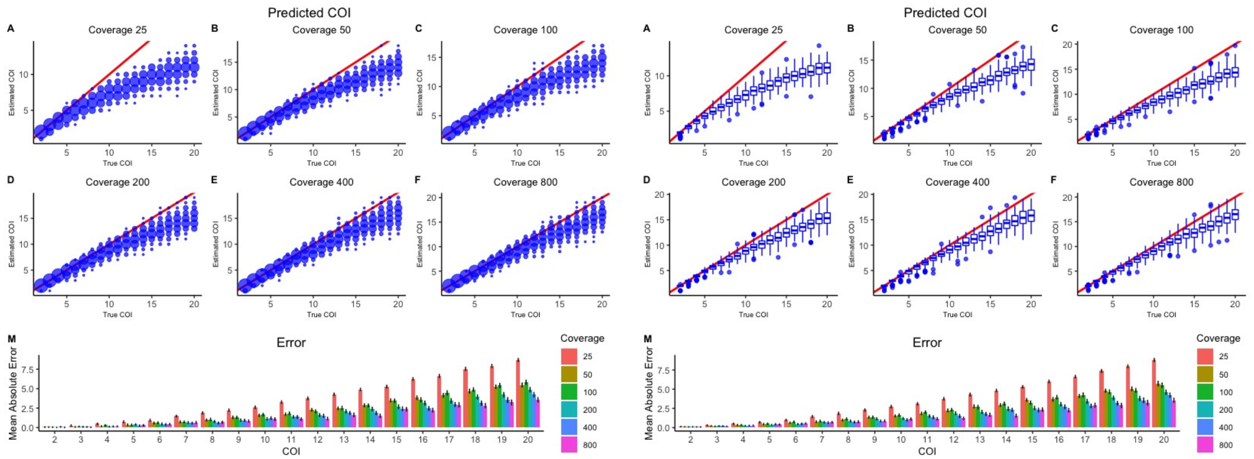
Table 1. Description of Select Parameters

6.3 Controllable Sequencing Conditions

We next examine the effect of varying two metrics that can be controlled in the field: the read depth at each locus and the number of loci sequenced. Sequencing more loci at larger read depths is generally preferred. However, this can be technically difficult and relies on high-quality sample collection and often larger volumes of infected blood being drawn rather than the use of dried red blood spots. However, analyzing larger numbers of data might prove computationally expensive. Therefore, the trade-off between additional data and computational efficiency must be carefully weighed.

Figure 3 demonstrates the performance of Method 1 as the coverage is varied from as low as 25 to as high as 800 reads. The analogous figure for Method 2 can be found in Appendix C. In general, as the coverage at each locus increases, the performance of our methods also increases. Interestingly, however, the relationship is not linear, rather, a non-monotonic relationship is observed between sequence coverage and the mean absolute error. At the lowest coverage tested, 25 reads per locus, our models exhibit a high mean absolute error—at such a low coverage, low prevalence malarial strains may be underrepresented and, as such, we expect that our models will underpredict the COI. An increase to a coverage of 50 greatly improves the performance. From a coverage of above 50 there are further improvements in the accuracy, however, it is dependent on the COI being estimated. At low COIs (< 10), non-significant decreases in mean absolute error are observed. However, at higher COIs, there is a significant benefit to higher sequence coverage. However, for all COI explored, the performance remains relatively constant with coverage greater than 200, indicating that more than 200 reads may not provide a substantial improvement in COI estimation.

Figure 4 indicates the performance of Method 1 as the number of loci sampled changes for the discrete and continuous versions of our optimization problems. Results for Method 2 are similar and can be found in Appendix C. As was the case for our coverage data, when the number of loci sequenced is low, around 100 loci, our methods have a very high variability and tend to underpredict.



(a) Discrete Method

(b) Continuous Method

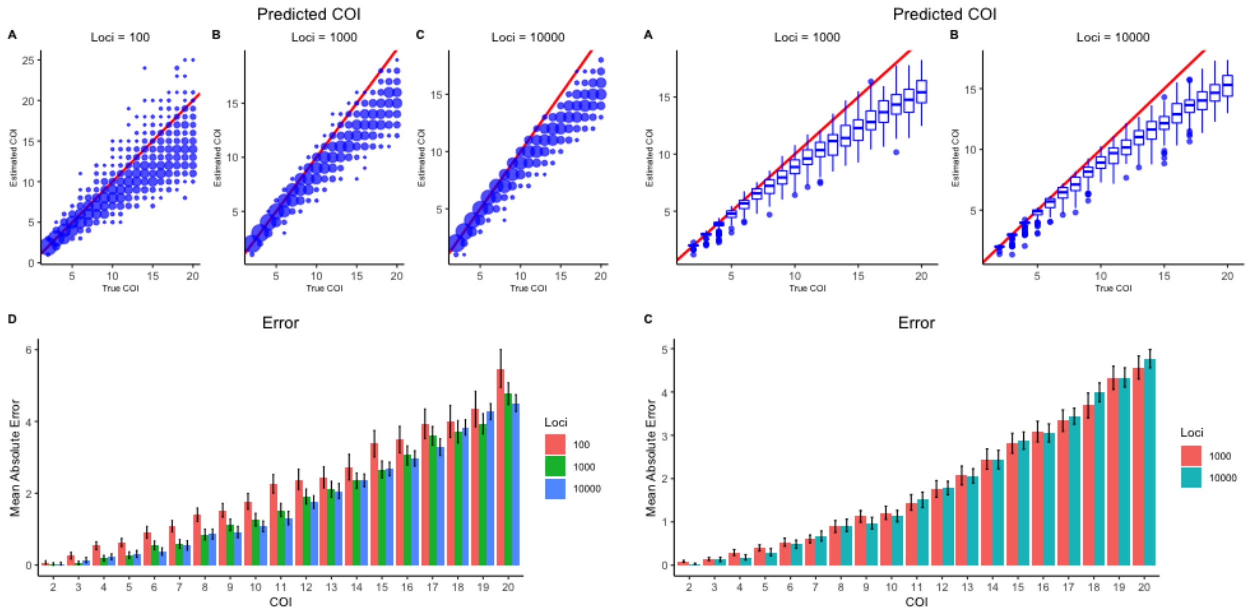
Figure 3. Performance of Method 1 with Changing Coverage. The performance of Method 1 is compared across a COI of 1-20. Data was simulated 100 times with a coverage of 200. No error was introduced into our simulations and a sequence error of 1% was assumed. In (a), the discrete estimation is shown and in (b), the continuous estimation is shown. In both cases, the distribution of estimated COI as the number of loci examined changes from 100 (A) to 1,000 (B) and 10,000 (C) is plotted. In (a), dot plots are used where point size indicates density. In (b), a boxplot shows the interquartile range, with whiskers indicating the 95% quantile range and outliers indicated with dots. The mean absolute error is shown in (C) for both methods, with the 95% quantile computed from 1,000 bootstrap repetitions at each COI indicated with error bars.

However, as the number of loci increases to 1,000, the performance increases. As was the case in examining the read depth of loci, increasing the number of loci examined above a certain threshold, in this case 1,000 loci, does not seem to substantially impact the performance of our models. We also note that the increase in the number of loci does reduce the variance of our estimates.

6.4 Malarial Biology

We next examine the effect of varying inputs to our simulator. In particular, we explore how perturbing the value of alpha, which is the concentration parameter used to determine the proportion of each strain from a Dirichlet distribution changes the performance of our methods. A larger alpha value decreases the variance in strain proportions and consequently the strains are more likely to be at comparable parasite density and more likely to equally contribute to sequence reads, which, in turn, would be expected to improve the accuracy of our methods. This is more likely to occur during cotransmission events when parasites are acquired from the same infectious bite [21]. However, note this will likely increase the relatedness of parasites, which will invalidate our assumptions and decrease the accuracy of our methods. Furthermore, we examine the effect of introducing overdispersion in sequence coverage such that the probability of sampling a minor allele at each locus with each sequence read is not equal to the true WSAF, but randomly drawn from a Beta-Binomial distribution. With smaller overdispersion assumed, the sampling process increasingly tends towards being described by the Binomial distribution. Modelling sequencing in this way is chosen to model scenarios in which parasite strains, despite a high parasite density, may be underrepresented in sequencing. For example, this could be due to sequestration during part of the parasite's life cycle [51].

Figure 5 demonstrates the effect of changing alpha for the discrete and continuous versions of



(a) Discrete Method

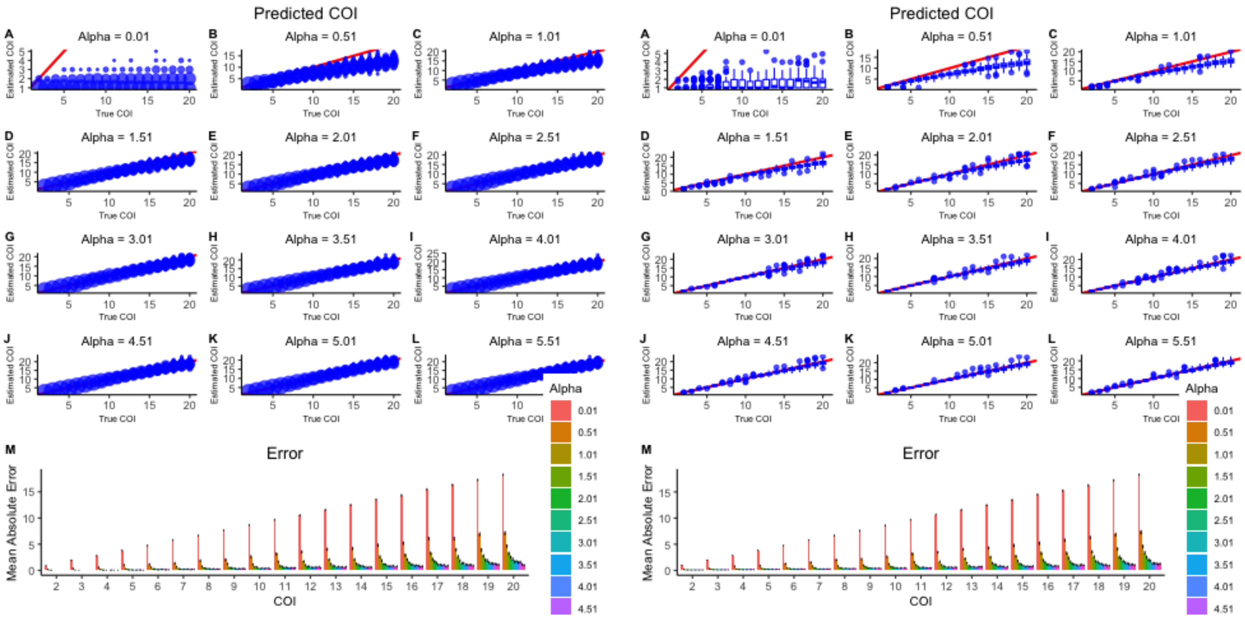
(b) Continuous Method

Figure 4. Performance of Method 1 with Changing Number of Loci. The performance of Method 1 is compared across a COI of 1-20. Data was simulated 100 times with a coverage of 200. No error was introduced into our simulations and a sequence error of 1% was assumed. In (a), the discrete estimation is shown and in (b), the continuous estimation is shown. In both cases, the distribution of estimated COI as the number of loci examined changes is plotted. In (a), dot plots are used where point size indicates density. In (b), a boxplot shows the interquartile range, with whiskers indicating the 95% quantile range and outliers indicated with dots. The mean absolute error is shown in (C) for both methods, with the 95% quantile computed from 1,000 bootstrap repetitions at each COI indicated with error bars.

Method 1. The analogous figure for Method 2 can be found in Appendix C. Figure 6 illustrates the effect of changing the overdispersion for both Method 1 and Method 2.

We note that as the value of alpha increases, the performance of our methods also increases, as was expected. Furthermore, the smaller the overdispersion present in our simulated data, the better our models perform. This makes intuitive sense as larger overdispersion values leads to deviation from our assumed distributions.

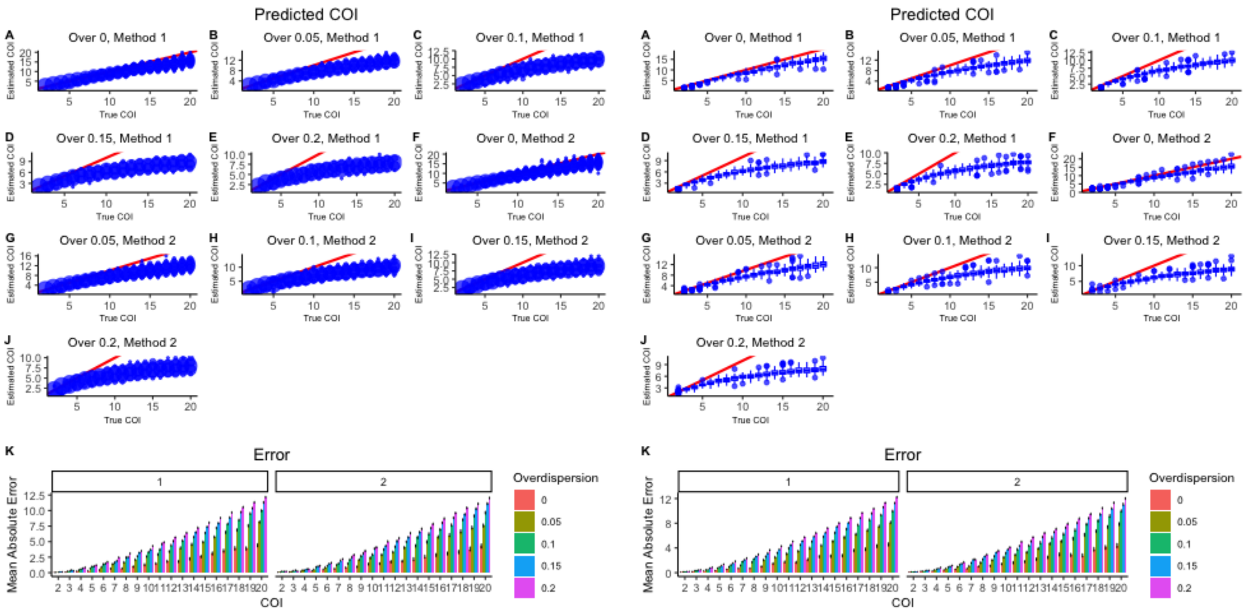
Lastly, we examine the effect of introducing error into our simulated data. In particular, we set the error rate to be 0.1, i.e., the probability that a single read is miscalled as the other allele is 1%. We then evaluate the impact of introducing sequencing error on the performance of our methods. Figure 7 illustrates the results when we utilize Method 1. An analogous figure for Method 2 can be found in Appendix C. When we introduce error into our simulations, we see that our methods overpredict the COI, especially when the level of sequence error assumed is low. However, as the level of sequence error assumed increases, our methods' performance increases. Too much assumed sequence error causes our methods to underpredict the COI, indicating that there is fine line between too little and too much sequence error. These results highlight the importance of accurately detecting the level of sequence error of samples.



(a) Discrete Method

(b) Continuous Method

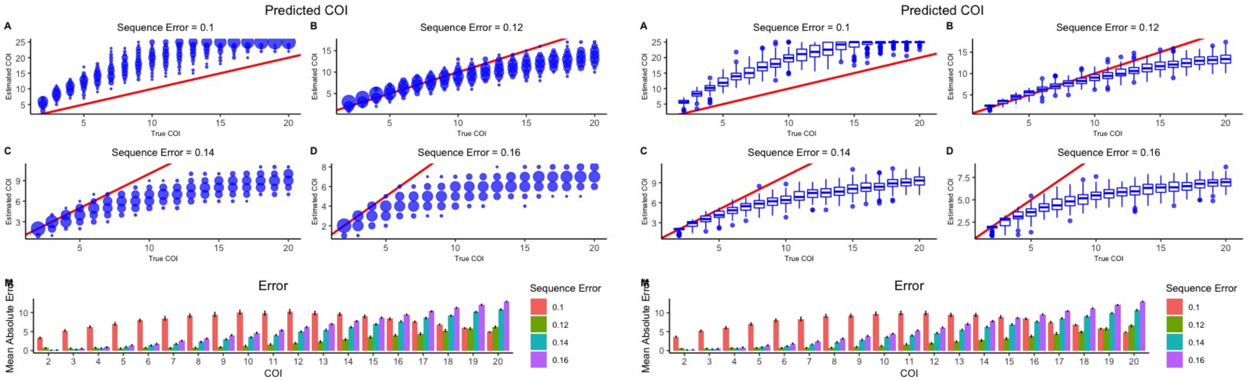
Figure 5. Performance of Method 1 with Changing Alpha. The performance of Method 1 is compared across a COI of 2-20. Data was simulated 100 times from 1,000 loci with a coverage of 200. No error was introduced into our simulations and a sequence error of 1% was assumed. In (a), the discrete estimation is shown and in (b), the continuous estimation is shown. In both cases, the distribution of estimated COI as alpha changes is plotted. In (a), dot plots are used where point size indicates density. In (b), a boxplot shows the interquartile range, with whiskers indicating the 95% quantile range and outliers indicated with dots. The mean absolute error is shown in (C) for both methods, with the 95% quantile computed from 1,000 bootstrap repetitions at each COI indicated with error bars.



(a) Discrete Method

(b) Continuous Method

Figure 6. Performance with Changing Overdispersion. The performance of Method 1 and Method 2 is compared across a COI of 2-20. Data was simulated 100 times from 1,000 loci with a coverage of 200. No error was introduced into our simulations and a sequence error of 1% was assumed. In (a), the discrete estimation is shown and in (b), the continuous estimation is shown. In both cases, the distribution of estimated COI as the overdispersion changes is plotted. In (a), dot plots are used where point size indicates density. In (b), a boxplot shows the interquartile range, with whiskers indicating the 95% quantile range and outliers indicated with dots. The mean absolute error is shown in (C) for both methods, with the 95% quantile computed from 1,000 bootstrap repetitions at each COI indicated with error bars. The panel with a "1" indicates the error for Method 1 and the panel with a "2" indicates the error for Method 2.



(a) Discrete Method

(b) Continuous Method

Figure 7. Performance of Method 1 with Changing Sequence Error. The performance of Method 1 is compared across a COI of 2-20. Data was simulated 100 times from 1,000 loci with a coverage of 200. An error rate of 0.1 was introduced into our simulations, i.e., the probability that a single read is miscalled as the other allele is 1%. In (a), the discrete estimation is shown and in (b), the continuous estimation is shown. In both cases, the distribution of estimated COI as the level of sequence error assumed changes is plotted. In (a), dot plots are used where point size indicates density. In (b), a boxplot shows the interquartile range, with whiskers indicating the 95% quantile range and outliers indicated with dots. The mean absolute error is shown in (C) for both methods, with the 95% quantile computed from 1,000 bootstrap repetitions at each COI indicated with error bars.

7 Comparison to Gold Standard

In this section we compare our novel methods to estimate the COI to the current "gold-standard" method to estimate the COI, *THE REAL McCOIL*. As was previously described, we identified 24 regions around the world and estimated the COI for each sample in each region. For each of the 5,970 samples, we examined an average of $5,537$ loci. We ran *THE REAL McCOIL* 5 times on each sample and ran our models 10 times on each sample, reporting the median estimated COI. Figure 8 and Figure 9 examine our estimated COI as well as *THE REAL McCOIL*'s estimated COI stratified by the 24 regions data was sampled from. Figure 19 compares the predicted COI over all regions and can be found in Appendix C. In each figure, the estimated COI using *THE REAL McCOIL* is plotted on the x-axis. For all samples of a specific COI value, we plot the results of our methods on the y-axis. If our methods and *THE REAL McCOIL* predict the same COI, we expect that data points will fall along the line $y = x$.

Predicting the COI by solving the discrete version of our optimization problems, we observe that when the COI is low, Method 1 performs well and frequently estimates the same value as *THE REAL McCOIL*. More specifically, if *THE REAL McCOIL* predicts the COI is 3 or less, Method 1 predicts the same value more than 90% of the time. Although there are some exceptions to this observation, these results are encouraging. For larger values of the COI, i.e., when *THE REAL McCOIL* predicts that the COI is more than 4, Method 1 predicts lower values than *THE REAL McCOIL*. Such is apparent in Regions 14 and 24, for example. Method 2 follows the same general trends, but its predictions vary more. When *THE REAL McCOIL* predicts the COI is 1, Method 2 agrees with the majority of the samples. However, there are several samples for which Method 2 greatly overpredicts the COI. For instance, in Region 1, Method 2 predicts a COI as high as 25, the highest our algorithm was allowed to predict. Several reasons for why this might be the case will be explored in the next subsection. Interestingly, Method 2 also overpredicts when *THE REAL McCOIL* predicts the COI is 2. However, the same trend is not seen for when *THE REAL McCOIL* estimates a COI of 3. Rather, for a COI of 3 and higher, Method 2 begins to exhibit trends similar to Method 1, underestimating the COI.

When we estimate the COI by solving the continuous version of our optimization problems, our results support the observations of the discrete case. Namely, when *THE REAL McCOIL* estimates the COI to be low, our methods also estimate the COI to be low. As was the case in the discrete case, although Method 2 predicts high COI values for several samples, the majority of its predictions are the same as *THE REAL McCOIL*'s. Interestingly, in many cases, our predictions appear to be closer to *THE REAL McCOIL*'s prediction, i.e., the points in our figure are closer to the line $y = x$. This result is expected. If we compare the values estimated by our discrete methods and continuous methods, we notice that the discrete COI often represents the rounded continuous COI. In other cases, the discrete COI is the floor or ceiling of our continuous estimate. As a result, in cases where we estimated the continuous value of the COI was, for example, 1.2, the discrete value of the COI would be 1. If *THE REAL McCOIL* predicted that the COI was 2, our continuous estimation would appear more similar to *THE REAL McCOIL*'s prediction. As noted previously, a continuous value of the COI has no biological relevance—it is not possible to be infected with a non-integer number of malarial parasite. However, a continuous value of the COI may indicate additional information about the parasite dynamics of an infection. In particular, a COI of 1.7 may indicate that the patient is infected with 2 related parasite strains. Following the same logic, a COI of 5.5 may mean that a patient is infected with 6 parasite strains where some are related. However, it could also mean that the patient is infected with 7, 8, or even more strains that are highly related. As a result, by assuming unrelatedness in our methods it becomes increasingly difficult to accurately estimate high COI values. In particular, at high COI values, it is likely that a number of parasites will be related, having arisen from cotransmission events. A study of mixed infections in Malawi using single cell sequencing showed that the highest COI samples arose from the majority of strains being acquired from cotransmission events and being highly related [21].

Another reason why our predictions may differ from *THE REAL McCOIL*'s predictions is due to the way in which the two methods treat the PLAF. In particular, our methods assume that the PLAF is given as an input whereas *THE REAL McCOIL* estimates the PLAF concurrently to estimating the COI. In regions with few samples, our methods may struggle to accurately predict the COI as the input PLAF may not be representative of the population. *THE REAL McCOIL*, however, may be better suited to correctly estimate the COI, given it estimates the PLAF. Although *THE REAL McCOIL* may be a better option to estimate the COI in some circumstances, it is a more computationally intensive estimation method, often taking over a day when analysing 100s of samples with 1,000 loci. Our methods, on the other hand, can estimate the COI of a sample in less than 1 second.

To further test the accuracy of our models on sequenced data, we further compared our methods to the *FwS* metric [29], a measure of the within-host genetic diversity. We found a strong negative correlation between our estimated COI using the discrete version of Method 1 and the *FwS* metric, as indicated in Figure 10. A similar relationship was found between *THE REAL McCOIL* and the *FwS* metric indicating that our results are on par with the *THE REAL McCOIL*. These results are expected as a low *FwS* reflects a high within-host diversity [29].

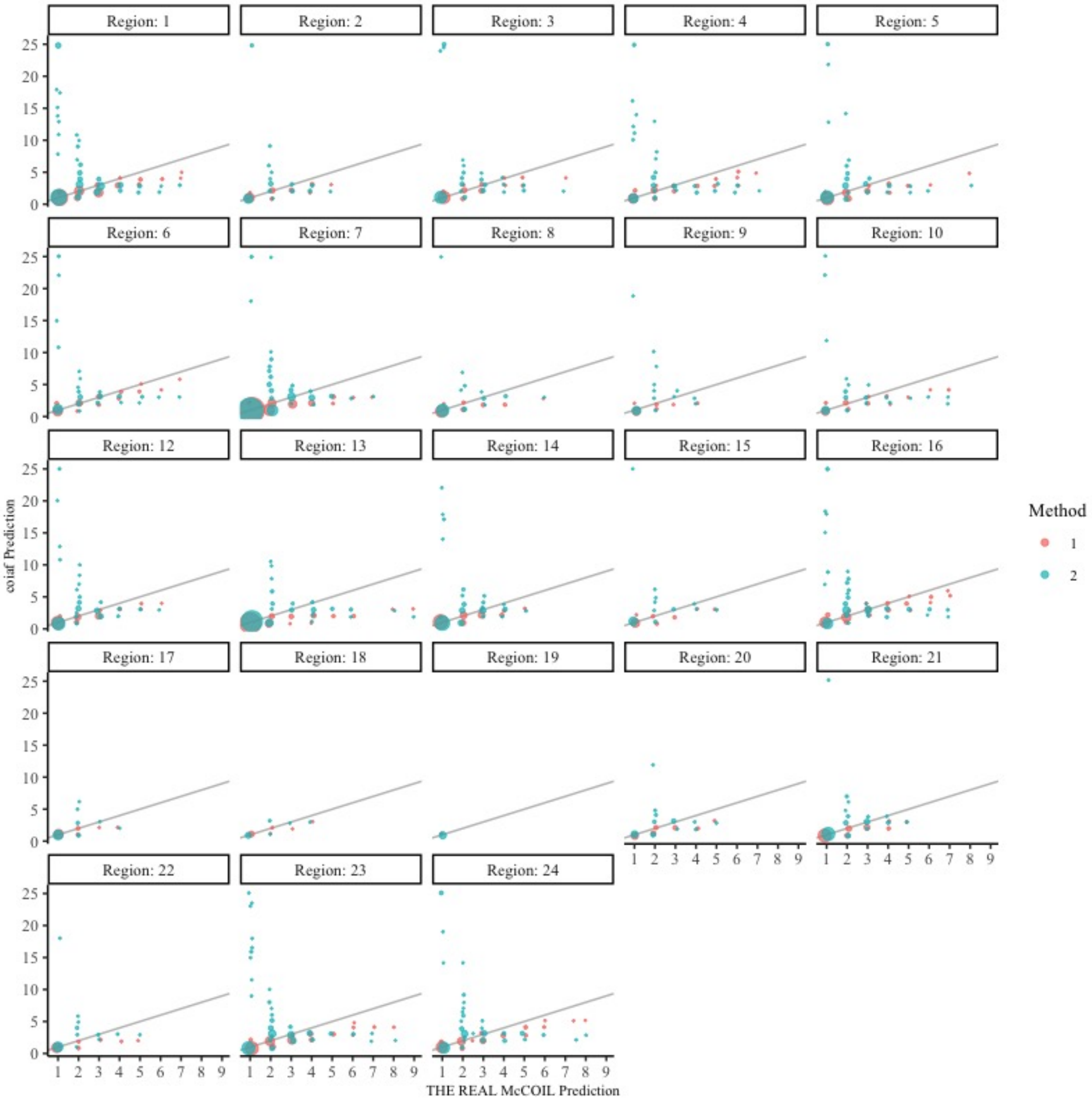


Figure 8. Comparison Between RMCL and Discrete coiaf Across all Regions. We estimated the COI for 5,970 samples grouped in 24 regions around the world. For each sample, we ran *THE REAL McCOIL* 5 times and ran our models 10 times, reporting the median estimated COI. For each region we plot the discrete estimation of the COI for our models vs the results of *THE REAL McCOIL*. A boxplot shows the interquartile range, with whiskers indicating the 95% quantile range and outliers indicated with dots.

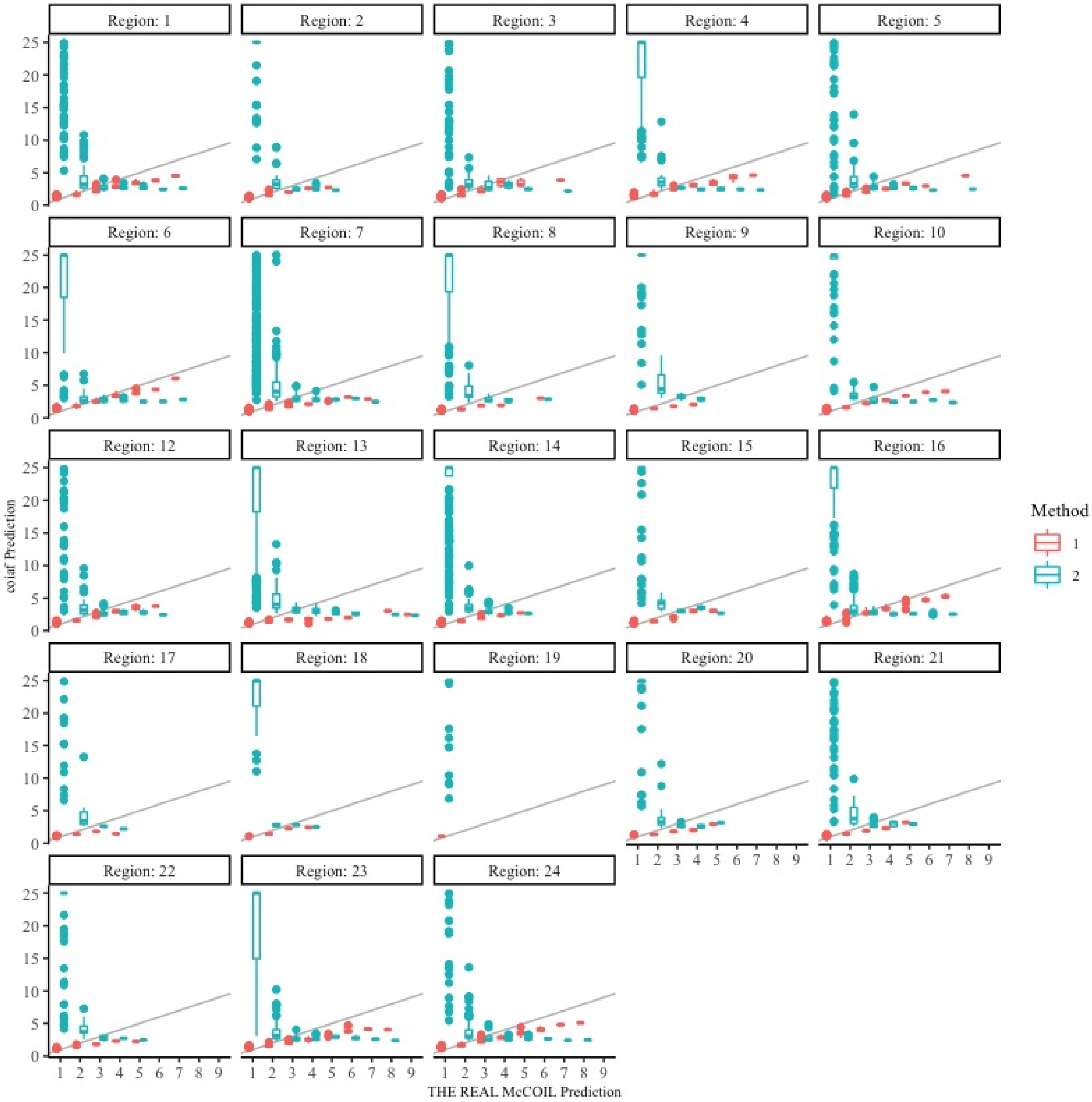


Figure 9. Comparison Between RMCL and Continuous coiaf Across all Regions. We estimated the COI for 5,970 samples grouped in 24 regions around the world. For each sample, we ran *THE REAL McCOIL* 5 times and ran our models 10 times, reporting the median estimated COI. For each region we plot the continuous estimation of the COI for our models vs the results of *THE REAL McCOIL*. Point size indicates density.

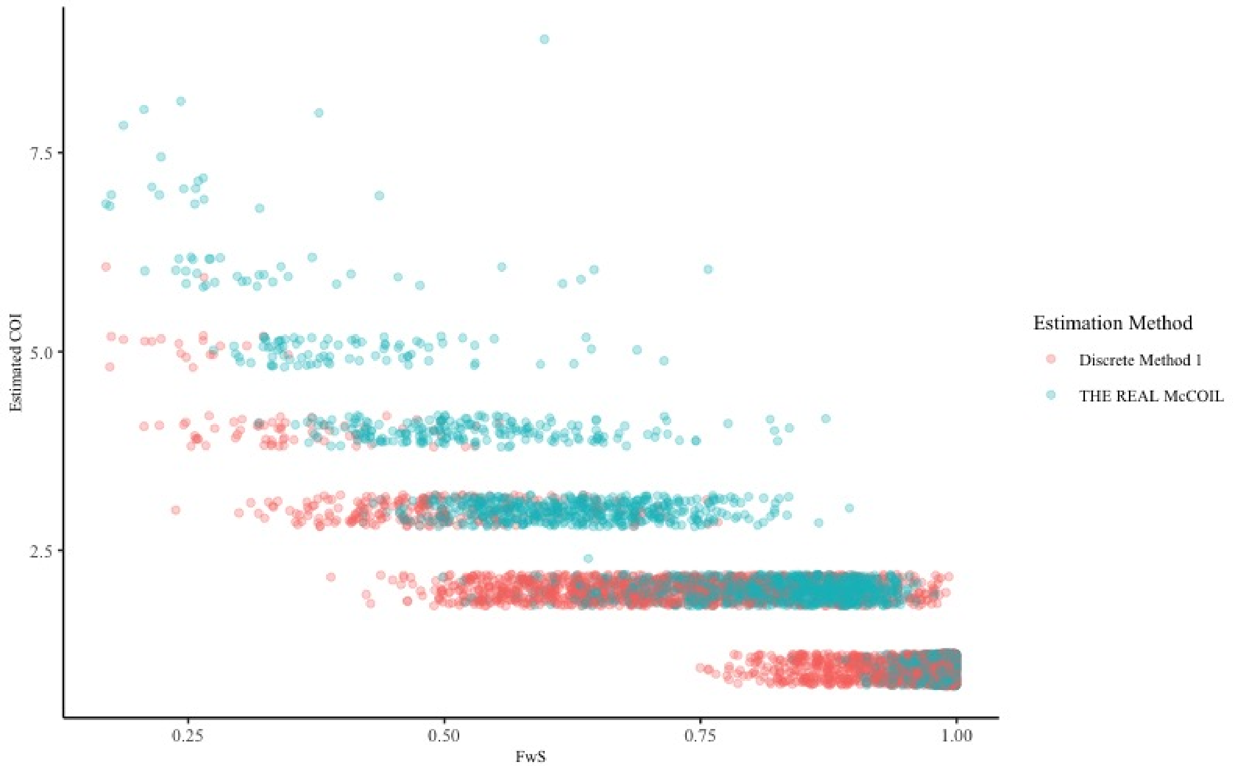
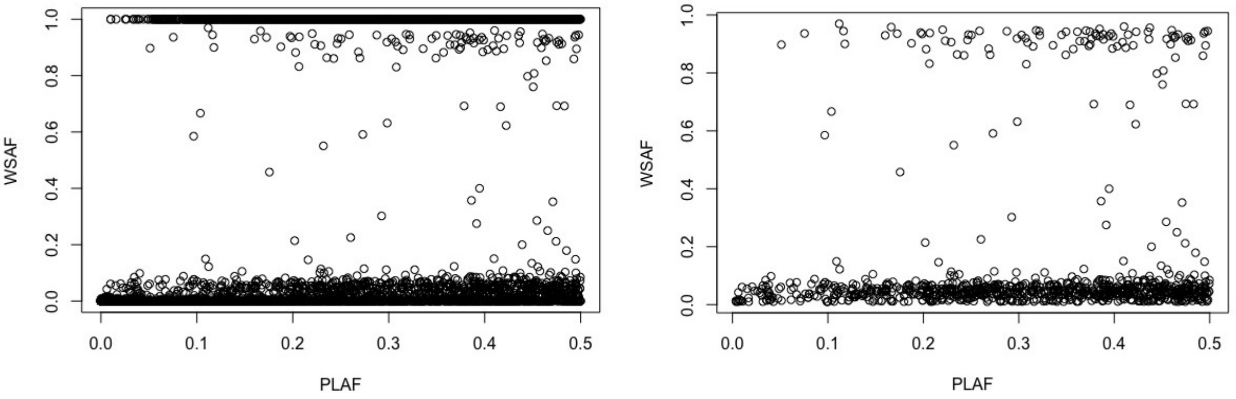


Figure 10. Comparison Between Estimation Methods and *FwS*. We estimated the COI for 5,970 samples grouped in 24 regions around the world. For each sample, we ran *THE REAL McCOIL* 5 times and ran our models 10 times, reporting the median estimated COI. The x-axis indicates the *FwS*, a measure of the within-host genetic diversity, where a lower *FwS* indicates a higher within-host diversity. The y-axis indicates the estimated COI when using either the discrete version of Method 1 or *THE REAL McCOIL*. The color of the points reflects whether the COI was estimated using the discrete version of Method 1 or *THE REAL McCOIL*. Note that each band of data corresponds to integer value of the COI—noise in the y-direction has been added for interpretability.

7.1 A Case Study: Why Method 2 Exhibits Increased Variance

Among our two methods presented, Method 2 clearly struggles the most. In order to better understand some of the reasons why Method 2 underperforms, we consider the sample with PLAF and WSAF as indicated in Figure 11 Panel (a). For this sample, our discrete version of Method 1 predicts a COI of 2 and the *THE REAL McCOIL* predicts a COI of 1. The discrete version of Method 2, on the other hand, predicts a COI of 25, the maximum COI our model could return for this particular sample.



(a) All Data

(b) Subset of Data

Figure 11. PLAF vs WSAF of a Sample. The PLAF and WSAF are plotted for a sample. In (a) all loci are plotted whereas in (b) a subset of data is plotted. In particular, data are first filtered to account for sequencing error and then only heterozygous loci are plotted.

When this data is provided to Method 2, we first account for any potential sequence error and then subset our data to only focus on heterozygous data. In this sample, the first difficulty stems from determining the level of sequencing error. In particular, it is difficult to detect if the large band of data at a WSAF close to 0 is sequencing error or not. If we assume that this data is sequencing error and we filter it out, we are left with very few loci to examine. For example, if we assume a sequencing error of 20%, then only data with a WSAF of greater than 0.2 or less than 0.8 will remain. This makes it difficult to confidently solve our minimization problems and as a result, our estimate may be incorrect.

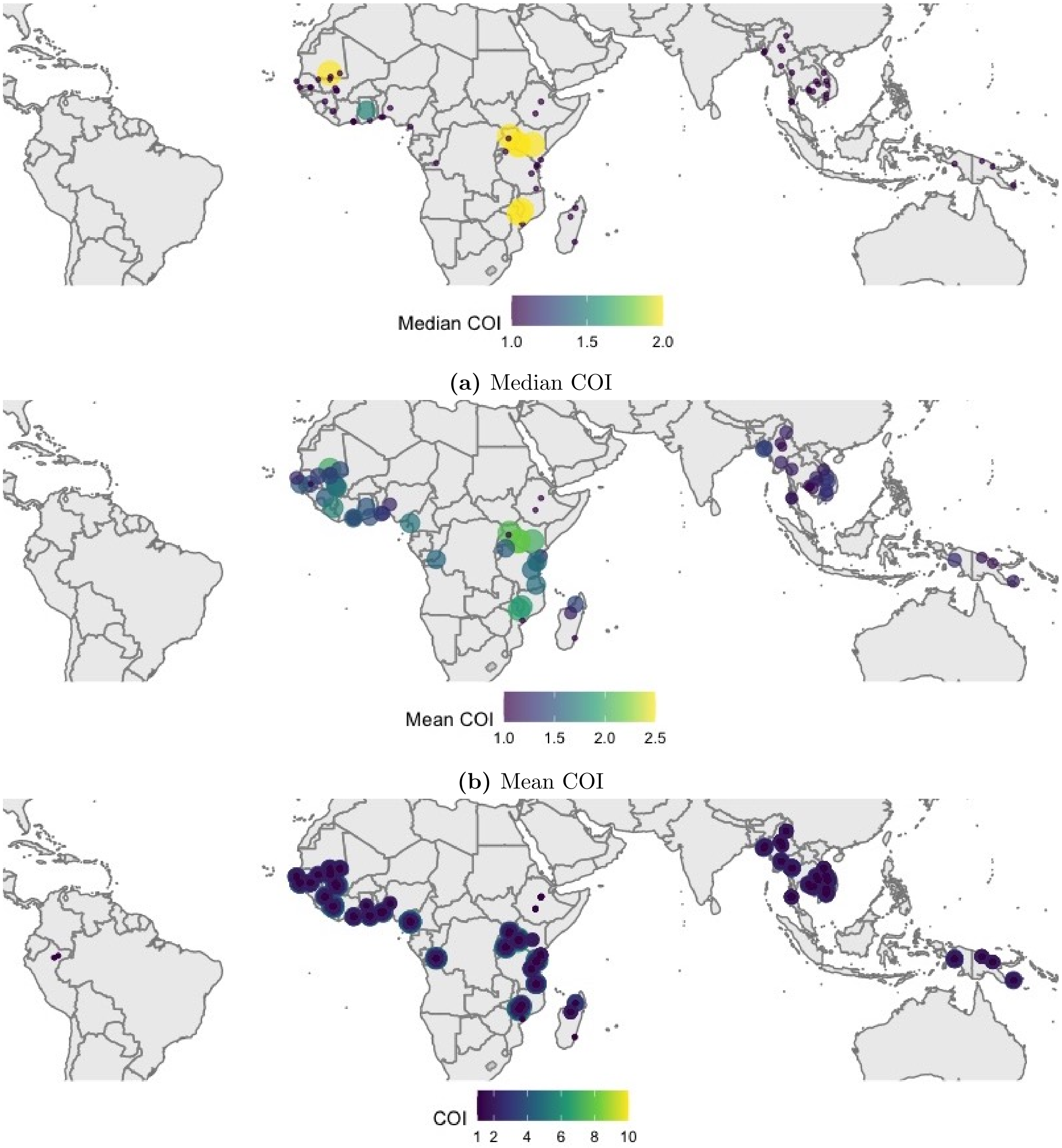
On the other hand, if we assume a low level of sequence error, e.g., 1%, then other difficulties in estimating the COI arise. Assuming a low level of sequencing error after filtering out homozygous data, we are left with the data pictured in Figure 11 Panel (b). Note that the number of data points at a low WSAF is substantially higher than the number of data points at a high WSAF. As a result, when we split our data into bins and find the mean WSAF for each bucket, the mean WSAF remains low, resulting in Method 2 estimating a very high COI.

In order to deal with noisy data, *THE REAL McCOIL* infers the sequence error across the population of samples, specifying the same level of sequencing error for each sample. However, as the number of noisy samples is low, in assigning the same level of sequencing error to all samples, *THE REAL McCOIL* may be inadvertently assigning unreasonably high levels of sequence error to samples. Our models, on the other hand, infer the sequencing error on a per-sample basis. Therefore, for some samples that *THE REAL McCOIL* assigns a high sequence error and we assume a low level of sequence error, it may very well be that our estimated COI values differ. We believe that inferring the level of sequence error on a per-sample basis will yield more accurate results. However, estimating sequencing error is complicated and is itself an active area of research [52] and one that we may need to further refine in our analysis in future work.

8 Mapping COI Worldwide

In the next section we present several figures using our novel methods to estimate the COI around the world. We estimate the COI using the discrete version of Method 1 on the same data used in Section 7. After determining the COI for each sample, we compute the median and mean COI for each study location in each region and plot them on a world map. Figure 12 Panel (a) indicates the median COI in each region and Panel (b) indicates the mean COI in each region. Panel (c) indicates the aggregate COI, where we plot the COI of each sample in each region.

Our samples were sequenced primarily in Africa with several samples originating from South East Asia. Examining the median COI of each region, we observe that the COI tends to be higher in Africa. Interestingly, regions of high median COI tend to cluster together. For instance, there are several regions on the Eastern coast of Africa where the median COI is 2. However, in some cases there are also low COI regions next to the higher COI regions. Further examining the prevalence of malaria in these regions may provide additional insight into malaria dynamics. The mean COI tends to be higher than the median COI in each region, indicating that the majority of patients sampled have a low COI, but that there are a few patients with large COI values, which drive the mean to be large.



(a) Median COI

(b) Mean COI

(c) Aggregate COI

Figure 12. COI Across the Globe. We estimated the COI for 5,970 samples grouped in 24 regions around the world. For each sample, we ran our models 10 times and took the median COI. In (a) we plot the median COI of all samples in each study location within the 24 regions, in (b) we plot the mean COI, and in (c) we plot the COI of each sample. The color and size of each point represents the magnitude of the COI.

9 Conclusion

In this thesis, we described our novel software package *coiaf* which can be used to predict the complexity of infection of a sample. We derived two different methods to estimate the COI, one which identifies the probability that a locus is heterozygous, and the other which identifies the expected value of the within-sample allele frequency given a site is heterozygous. We analyzed our methods on simulated data and explored how changing key parameters influenced our predictive accuracy, concluding that our methods accurately estimate the COI for low COIs, but that as the COI increased, our methods exhibit more variability. Furthermore, we observed that Method 2 tended to underperform compared to Method 1. We then compared our methods to the current state-of-the-art method and discovered that for low COIs our methods are comparable. However, differences arose when the COI was higher than 3. Once again, we observed that Method 2 struggled to predict the COI, in some instances greatly overestimating the COI. We explored several reasons for why this might be the case. Lastly, we examined the COI in 24 regions across the world.

Future work can be performed to address some of the limitations of Method 2. In particular, refining the method we use to infer the sequence error of samples on a per-sample basis may enhance performance. Furthermore, as we have suggested, comparing the continuous and the discrete versions of the optimization problems may provide insight into parasite relatedness. Moreover, comparing estimates from Method 1 and Method 2 may provide insight into parasite relatedness. Additionally, we can conduct a more rigorous comparison of our algorithms and *THE REAL McCOIL*'s comparing computational efficiency, as well as memory usage. Lastly, we can compare our methods to prevalence data from the regions we sampled.

10 Acknowledgments

First and foremost, I would like to extend my deepest gratitude to Dr. Jeffrey Bailey and Dr. Oliver Watson. Without their assistance and dedicated involvement in every step of my long process, this thesis would not have been possible. Dr. Bailey welcomed me into his lab over a year ago, taking a chance on a junior who had limited experience with malaria. I am deeply appreciate of him and Bailey Lab for teaching me so much. I would also like to thank Dr. Watson who, despite being half a world away in England, has continued to provide me with irreplaceable support and has taught me so much about programming. I look forward to the day, if it will ever come, when I can say I taught him a new programming trick. I would also like to thank my advisor, Dr. Daniel Weinreich, who introduced me to the field of infectious disease and whose classes have been instrumental in motivating my interests.

In addition to the academic support I have received, I have my family and friends to thank. My roommates, especially, have been unwavering in their personal support throughout the time I have spent at Brown and I cannot express my appreciation for their friendship. I would like to thank my grandmother, godmother, and extended family for calling me to offer words of encouragement when I was stressed. To my parents and two brothers, you mean the world to me. You inspire me to be better, to always strive for excellence, and have always been there for me. This work is dedicated to you.

References

1. WHO | Malaria rapid diagnostic test performance. Results of WHO product testing of malaria RDTs: round 8 (2016-2018);. Available from: http://www.who.int/malaria/publications/atoz/9789241514965/en/.

2. Organization WH. World malaria report 2020: 20 years of global progress and challenges. 2020; p. 151.

3. Benelli G, Beier JC. Current vector control challenges in the fight against malaria. Acta Tropica. 2017;174:91–96. doi:10.1016/j.actatropica.2017.06.028.

4. Raghavendra K, Barik TK, Reddy BPN, Sharma P, Dash AP. Malaria vector control: from past to future. Parasitology Research. 2011;108(4):757–779. doi:10.1007/s00436-010-2232-0.

5. Takken W, Knols BGJ. Malaria vector control: current and future strategies. Trends in Parasitology. 2009;25(3):101–104. doi:10.1016/j.pt.2008.12.002.

6. Mutabingwa TK. Artemisinin-based combination therapies (ACTs): best hope for malaria treatment but inaccessible to the needy! Acta Tropica. 2005;95(3):305–315. doi:10.1016/j.actatropica.2005.06.009.

7. White NJ. Qinghaosu (Artemisinin): The Price of Success. Science. 2008;320(5874):330–334. doi:10.1126/science.1155165.

8. Lin JT, Juliano JJ, Wongsrichanalai C. Drug-Resistant Malaria: The Era of ACT. Current infectious disease reports. 2010;12(3):165–173. doi:10.1007/s11908-010-0099-y.

9. The Malaria Atlas Project;. Available from: https://www.tki-dev.malariaatlas.org/.

10. Murray CK, Bell D, Gasser RA, Wongsrichanalai C. Rapid diagnostic testing for malaria. Tropical Medicine & International Health. 2003;8(10):876–883. doi:https://doi.org/10.1046/j.1365-3156.2003.01115.x.

11. Mouatcho JC, Goldring JPD. Malaria rapid diagnostic tests: challenges and prospects. Journal of Medical Microbiology. 2013;62(10):1491–1505. doi:10.1099/jmm.0.052506-0.

12. Prevention CCfDCa. CDC - Malaria - Diagnosis & Treatment (United States) - Diagnosis (U.S.); 2019. Available from: https://www.cdc.gov/malaria/diagnosis_treatment/diagnosis.html.

13. Watson OJ, Slater HC, Verity R, Parr JB, Mwandagalirwa MK, Tshefu A, et al. Modelling the drivers of the spread of Plasmodium falciparum hrp2 gene deletions in sub-Saharan Africa. eLife. 2017;6:e25008. doi:10.7554/eLife.25008.

14. Andrade BB, Reis-Filho A, Barros AM, Souza-Neto SM, Nogueira LL, Fukutani KF, et al. Towards a precise test for malaria diagnosis in the Brazilian Amazon: comparison among field microscopy, a rapid diagnostic test, nested PCR, and a computational expert system based on artificial neural networks. Malaria Journal. 2010;9(1):117. doi:10.1186/1475-2875-9-117.

15. Band G, Le QS, Clarke GM, Kivinen K, Hubbart C, Jeffreys AE, et al. Insights into malaria susceptibility using genome-wide data on 17,000 individuals from Africa, Asia and Oceania. Nature Communications. 2019;10(1):5732. doi:10.1038/s41467-019-13480-z.

16. Timmann C, Thye T, Vens M, Evans J, May J, Ehmen C, et al. Genome-wide association study indicates two novel resistance loci for severe malaria. Nature. 2012;489(7416):443–446. doi:10.1038/nature11334.

17. Bushman M, Morton L, Duah N, Quashie N, Abuaku B, Koram KA, et al. Within-host competition and drug resistance in the human malaria parasite Plasmodium falciparum. Proceedings of the Royal Society B: Biological Sciences. 2016;283(1826):20153038. doi:10.1098/rspb.2015.3038.

18. Birger RB, Kouyos RD, Cohen T, Griffiths EC, Huijben S, Mina MJ, et al. The potential impact of coinfection on antimicrobial chemotherapy and drug resistance. Trends in Microbiology. 2015;23(9):537–544. doi:10.1016/j.tim.2015.05.002.

19. Zhu SJ, Almagro-Garcia J, McVean G. Deconvolution of multiple infections in Plasmodium falciparum from high throughput sequencing data. Bioinformatics. 2018;34(1):9–15. doi:10.1093/bioinformatics/btx530.

20. Wong W, Griggs AD, Daniels RF, Schaffner SF, Ndiaye D, Bei AK, et al. Genetic relatedness analysis reveals the cotransmission of genetically related Plasmodium falciparum parasites in Thiès, Senegal. Genome Medicine. 2017;9. doi:10.1186/s13073-017-0398-0.

21. Nkhoma SC, Trevino SG, Gorena KM, Nair S, Khoswe S, Jett C, et al. Co-transmission of Related Malaria Parasite Lineages Shapes Within-Host Parasite Diversity. Cell Host & Microbe. 2020;27(1):93–103.e4. doi:10.1016/j.chom.2019.12.001.

22. Portugal S, Drakesmith H, Mota MM. Superinfection in malaria: Plasmodium shows its iron will. EMBO Reports. 2011;12(12):1233–1242. doi:10.1038/embor.2011.213.

23. Watson OJ, Okell LC, Hellewell J, Slater HC, Unwin HJT, Omedo I, et al. Evaluating the Performance of Malaria Genetics for Inferring Changes in Transmission Intensity Using Transmission Modeling. Molecular Biology and Evolution. 2021;38(1):274–289. doi:10.1093/molbev/msaa225.

24. Portugal S, Carret C, Recker M, Armitage AE, Gonçalves LA, Epiphanio S, et al. Host mediated regulation of superinfection in malaria. Nature medicine. 2011;17(6):732–737. doi:10.1038/nm.2368.

25. Rodriguez-Barraquer I, Arinaitwe E, Jagannathan P, Kamya MR, Rosenthal PJ, Rek J, et al. Quantification of anti-parasite and anti-disease immunity to malaria as a function of age and exposure. eLife. 2018;7:e35832. doi:10.7554/eLife.35832.

26. Daniels RF, Schaffner SF, Wenger EA, Proctor JL, Chang HH, Wong W, et al. Modeling malaria genomics reveals transmission decline and rebound in Senegal. Proceedings of the National Academy of Sciences. 2015;112(22):7067–7072. doi:10.1073/pnas.1505691112.

27. Snounou G, Beck HP. The Use of PCR Genotyping in the Assessment of Recrudescence or Reinfection after Antimalarial Drug Treatment. Parasitology Today. 1998;14(11):462–467. doi:10.1016/S0169-4758(98)01340-4.

28. Konaté L, Zwetyenga J, Rogier C, Bischoff E, Fontenille D, Tall A, et al. 5. Variation of Plasmodium falciparum msp1 block 2 and msp2 allele prevalence and of infection complexity

in two neighbouring Senegalese villages with different transmission conditions. Transactions of The Royal Society of Tropical Medicine and Hygiene. 1999;93(Supplement_1):21–28. doi:10.1016/S0035-9203(99)90323-1.

29. Auburn S, Campino S, Miotto O, Djimde AA, Zongo I, Manske M, et al. Characterization of Within-Host Plasmodium falciparum Diversity Using Next-Generation Sequence Data. PLOS ONE. 2012;7(2):e32891. doi:10.1371/journal.pone.0032891.

30. Galinsky K, Valim C, Salmier A, de Thoisy B, Musset L, Legrand E, et al. COIL: a methodology for evaluating malarial complexity of infection using likelihood from single nucleotide polymorphism data. Malaria Journal. 2015;14(1):4. doi:10.1186/1475-2875-14-4.

31. Assefa SA, Preston MD, Campino S, Ocholla H, Sutherland CJ, Clark TG. estMOI: estimating multiplicity of infection using parasite deep sequencing data. Bioinformatics. 2014;30(9):1292–1294. doi:10.1093/bioinformatics/btu005.

32. Wong W, Wenger EA, Hartl DL, Wirth DF. Modeling the genetic relatedness of Plasmodium falciparum parasites following meiotic recombination and cotransmission. PLOS Computational Biology. 2018;14(1):e1005923. doi:10.1371/journal.pcbi.1005923.

33. Zhu SJ, Hendry JA, Almagro-Garcia J, Pearson RD, Amato R, Miles A, et al. The origins and relatedness structure of mixed infections vary with local prevalence of P. falciparum malaria. eLife. 2019;8:e40845. doi:10.7554/eLife.40845.

34. Chang HH, Worby CJ, Yeka A, Nankabirwa J, Kamya MR, Staedke SG, et al. THE REAL McCOIL: A method for the concurrent estimation of the complexity of infection and SNP allele frequency for malaria parasites. PLOS Computational Biology. 2017;13(1):e1005348. doi:10.1371/journal.pcbi.1005348.

35. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2020. Available from: https://www.R-project.org/.

36. Byrd RH, Lu P, Nocedal J, Zhu C. A Limited Memory Algorithm for Bound Constrained Optimization. SIAM Journal on Scientific Computing. 1995;16(5):1190–1208. doi:10.1137/0916069.

37. BROYDEN CG. The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations. IMA Journal of Applied Mathematics. 1970;6(1):76–90. doi:10.1093/imamat/6.1.76.

38. Fletcher R. A new approach to variable metric algorithms. The Computer Journal. 1970;13(3):317–322. doi:10.1093/comjnl/13.3.317.

39. Goldfarb D. A family of variable-metric methods derived by variational means. Mathematics of Computation. 1970;24(109):23–26. doi:10.1090/S0025-5718-1970-0258249-6.

40. Shanno DF. Conditioning of quasi-Newton methods for function minimization. Mathematics of Computation. 1970;24(111):647–656. doi:10.1090/S0025-5718-1970-0274029-X.

41. MalariaGEN, Ahouidi A, Ali M, Almagro-Garcia J, Amambua-Ngwa A, Amaratunga C, et al. An open dataset of Plasmodium falciparum genome variation in 7,000 worldwide samples. Wellcome Open Research. 2021;6:42. doi:10.12688/wellcomeopenres.16168.1.

42. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics. 2011;43(5):491–498. doi:10.1038/ng.806.

43. Auwera GAVd, O'Connor BD. Genomics in the Cloud: Using Docker, GATK, and WDL in Terra. 1st ed. O'Reilly Media; 2020.

44. Kaufman L, Rousseeuw PJ. Finding Groups in Data: An Introduction to Cluster Analysis. 1st ed. Hoboken, N.J: Wiley-Interscience; 2005.

45. Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K. cluster: Cluster Analysis Basics and Extensions; 2021. Available from: `https://CRAN.R-project.org/package=cluster`.

46. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. 1987;20:53–65. doi:10.1016/0377-0427(87)90125-7.

47. Barbujani G. Autocorrelation of Gene Frequencies under Isolation by Distance. Genetics. 1987;117(4):777–782.

48. Verity R, Aydemir O, Brazeau NF, Watson OJ, Hathaway NJ, Mwandagalirwa MK, et al. The impact of antimalarial resistance on the genetic structure of Plasmodium falciparum in the DRC. Nature Communications. 2020;11(1):2107. doi:10.1038/s41467-020-15779-8.

49. DiCiccio TJ, Efron B. Bootstrap confidence intervals. Statistical Science. 1996;11(3):189–228. doi:10.1214/ss/1032280214.

50. Mooney CZ, Mooney CF, Mooney CL, Duval RD, Duvall R. Bootstrapping: A Nonparametric Approach to Statistical Inference. SAGE; 1993.

51. Koepfli C, Schoepflin S, Bretscher M, Lin E, Kiniboro B, Zimmerman PA, et al. How Much Remains Undetected? Probability of Molecular Detection of Human Plasmodia in the Field. PLoS ONE. 2011;6(4). doi:10.1371/journal.pone.0019010.

52. Zhu X, Wang J, Peng B, Shete S. Empirical estimation of sequencing error rates using smoothing splines. BMC Bioinformatics. 2016;17. doi:10.1186/s12859-016-1052-3.

53. Hay SI, Snow RW. The Malaria Atlas Project: Developing Global Maps of Malaria Risk. PLOS Medicine. 2006;3(12):e473. doi:10.1371/journal.pmed.0030473.

54. Guerra CA, Hay SI, Lucioparedes LS, Gikandi PW, Tatem AJ, Noor AM, et al. Assembling a global database of malaria parasite prevalence for the Malaria Atlas Project. Malaria Journal. 2007;6(1):17. doi:10.1186/1475-2875-6-17.

55. Pfeffer D, Lucas T, May D, Harris J, Rozier J, Twohig K, et al. malariaAtlas: an R interface to global malariometric data hosted by the Malaria Atlas Project. Malaria Journal. 2018;17(1):352. doi:10.1186/s12936-018-2500-5.

Appendices

A Accounting for Sequence Error

During sequencing, error may arise for several reasons. In order to address this error, we allow the user to choose between two options. The first option is the simplest, and allows the user to have full control over the level of error assumed. The user may specify the amount of error believed to be present in their samples. For instance, there may be a 5% sequence error.

The second option does not require any user input. Rather the sequence error is inferred from the distribution of the data. In particular, we focus on data that has a low PLAF and compare the number of points with a WSAF greater than zero to the expected number of points with a WSAF greater than zero. Consider an arbitrary data point (p_i, w_i). We define a partition over the range of p_i constituent of sets $\mathcal{P}_1, \ldots, \mathcal{P}_N$ such that $p_i \in \mathcal{P}_m$ if $p_i \in [a_m, b_m), \quad a_m < b_m,\ i = 1, \ldots, l$. In doing so, we group our data into N bins. We first count the number of loci in the first bin. We define the PLAF for this partition as indicated by Equation (12).

Given the number of loci in the first partition and the PLAF, we can next compute the expected number of loci with the minor allele. Recall that sampling the minor allele from a site can be expressed using a Binomial distribution with 1 trial and probability of success (drawing the minor allele) equal to the PLAF at that locus. Therefore, to find the expected number of loci with the minor allele, we can multiply the number of loci with the probability of a locus being the variant allele. At a PLAF of zero, we expect very few sites to have the minor allele. Therefore, if we remove the expected number of loci from the true number of loci with a minor allele, all the loci that remain are likely the result of sequencing error. Therefore, by examining the distribution of the WSAF of these points, we can estimate the amount of sequence error present in our sample.

B Clustering Real Data Samples

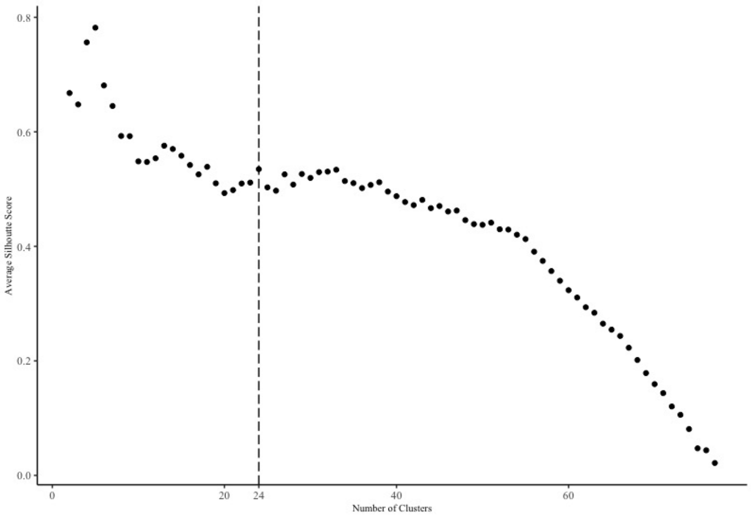


Figure 13. Silhouette Plot. The x-axis represents the number of clusters the data was split into. The y-axis indicates the average silhouette score, a measure of cluster validity and strength. The silhouette score can provide a measure to select the optimal number of clusters [46].

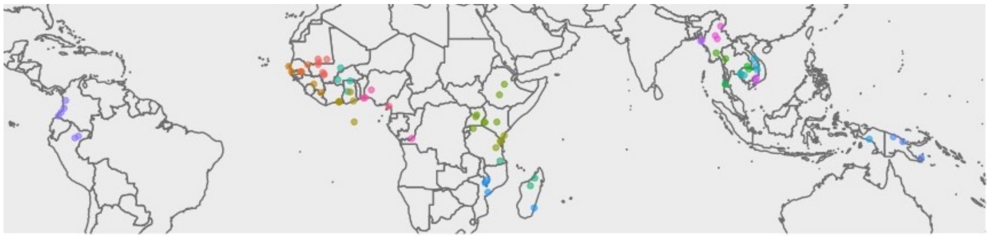
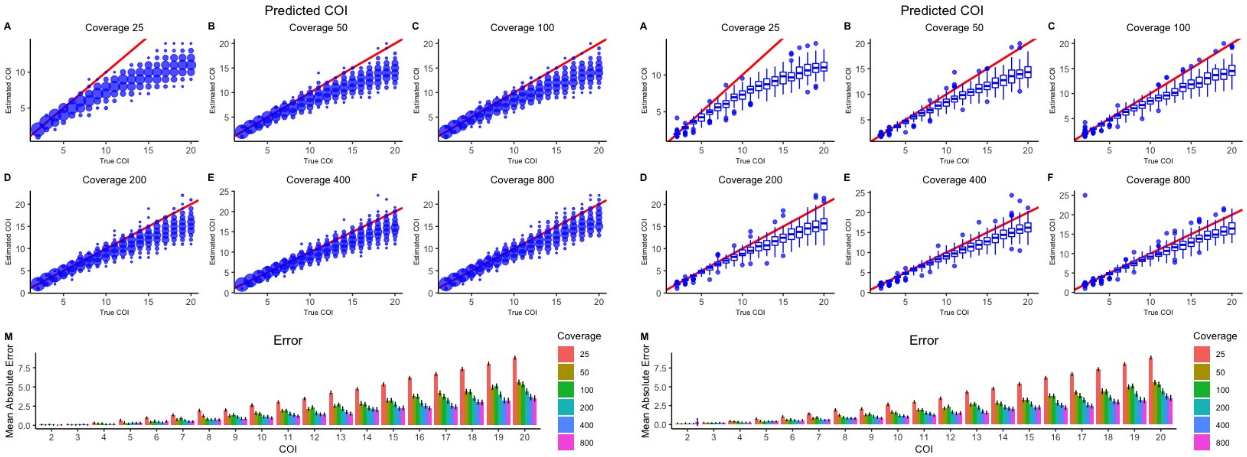


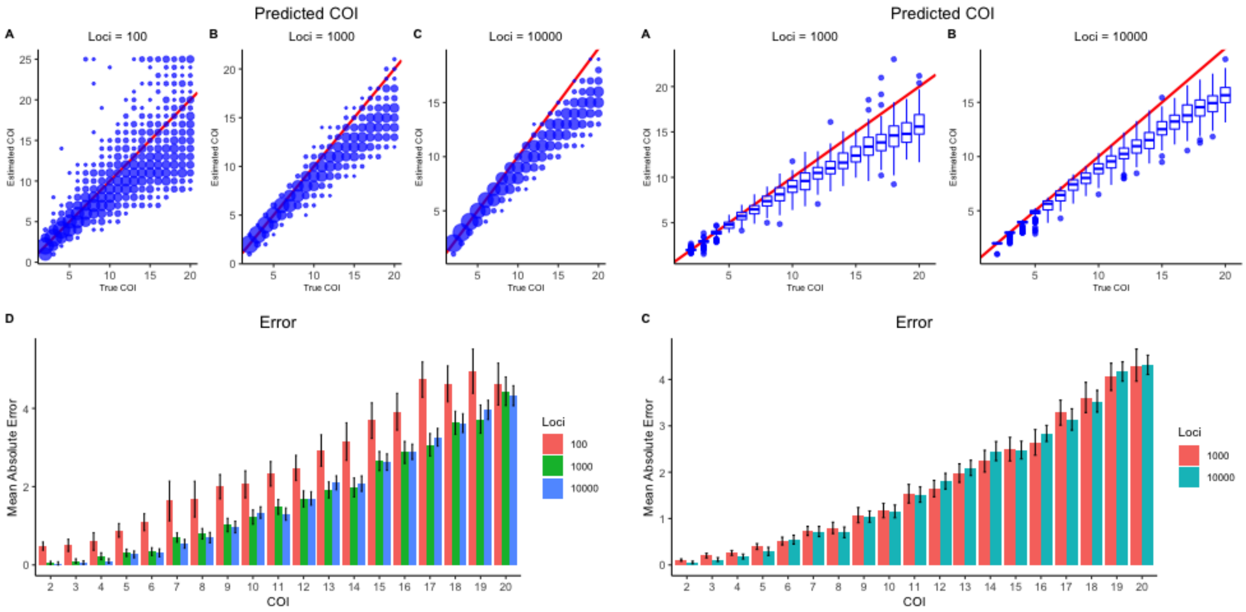
Figure 14. Location of 24 Clusters. Plot of each individual location data was sampled from. The color of the point indicates which of the 24 clusters data was assigned to.

C Omitted Figures



(a) Discrete Method **(b)** Continuous Method

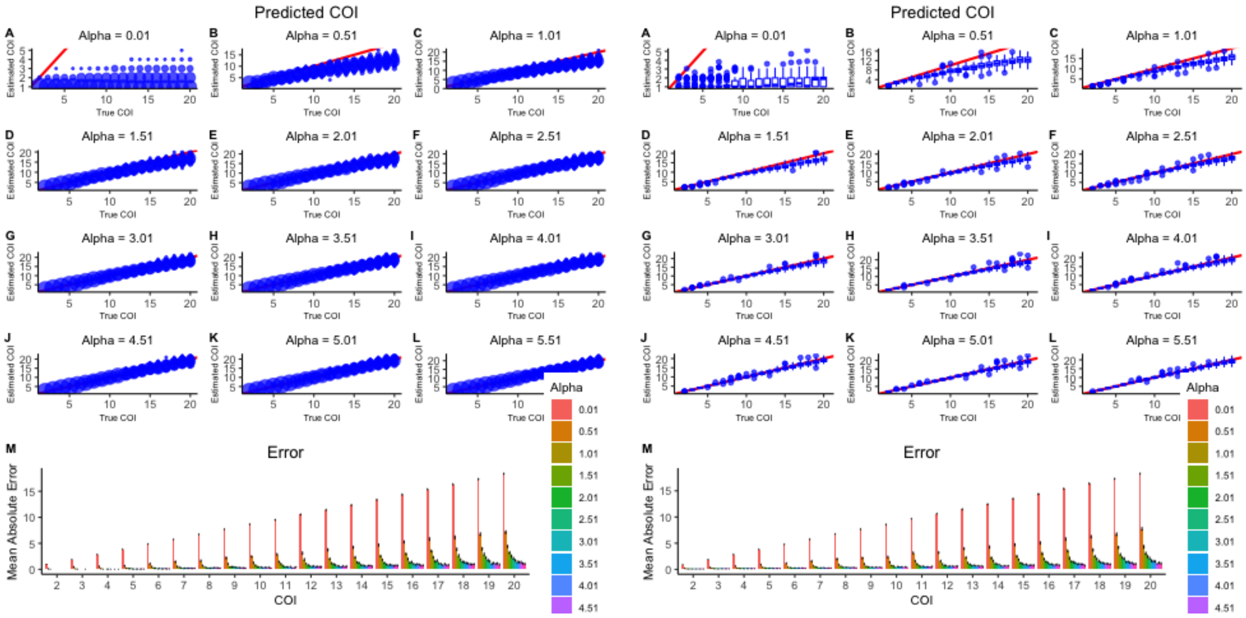
Figure 15. Performance of Method 2 with Changing Coverage. The performance of Method 2 is compared across a COI of 2-20. Data was simulated 100 times with a coverage of 200. No error was introduced into our simulations and a sequence error of 1% was assumed. In (a), the discrete estimation is shown and in (b), the continuous estimation is shown. In both cases, the distribution of estimated COI as the number of loci examined changes from 100 (A) to 1,000 (B) and 10,000 (C) is plotted. In (a), dot plots are used where point size indicates density. In (b), a boxplot shows the interquartile range, with whiskers indicating the 95% quantile range and outliers indicated with dots. The mean absolute error is shown in (C) for both methods, with the 95% quantile computed from 1,000 bootstrap repetitions at each COI indicated with error bars.



(a) Discrete Method

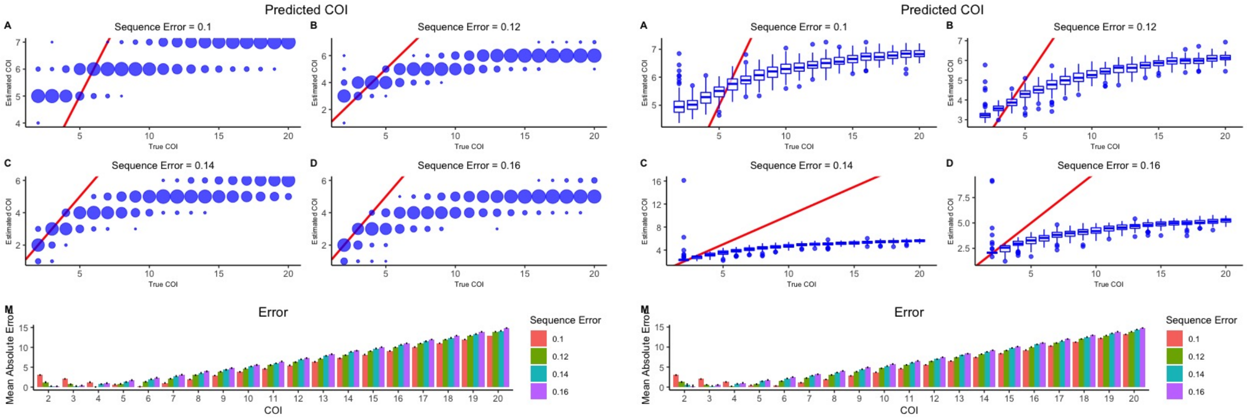
(b) Continuous Method

Figure 16. Performance of Method 2 with Changing Number of Loci. The performance of Method 2 is compared across a COI of 2-20. Data was simulated 100 times with a coverage of 200. No error was introduced into our simulations and a sequence error of 1% was assumed. In (a), the discrete estimation is shown and in (b), the continuous estimation is shown. In both cases, the distribution of estimated COI as the number of loci examined changes is plotted. In (a), dot plots are used where point size indicates density. In (b), a boxplot shows the interquartile range, with whiskers indicating the 95% quantile range and outliers indicated with dots. The mean absolute error is shown in (C) for both methods, with the 95% quantile computed from 1,000 bootstrap repetitions at each COI indicated with error bars.



(a) Discrete Method **(b)** Continuous Method

Figure 17. Performance of Method 2 with Changing Alpha. The performance of Method 2 is compared across a COI of 2-20. Data was simulated 100 times from 1,000 loci with a coverage of 200. No error was introduced into our simulations and a sequence error of 1% was assumed. In (a), the discrete estimation is shown and in (b), the continuous estimation is shown. In both cases, the distribution of estimated COI as alpha changes is plotted. In (a), dot plots are used where point size indicates density. In (b), a boxplot shows the interquartile range, with whiskers indicating the 95% quantile range and outliers indicated with dots. The mean absolute error is shown in (C) for both methods, with the 95% quantile computed from 1,000 bootstrap repetitions at each COI indicated with error bars.



(a) Discrete Method **(b)** Continuous Method

Figure 18. Performance of Method 2 with Changing Sequence Error. The performance of Method 2 is compared across a COI of 2-20. Data was simulated 100 times from 1,000 loci with a coverage of 200. An error rate of 0.1 was introduced into our simulations, i.e., the probability that a single read is miscalled as the other allele is 1%. In (a), the discrete estimation is shown and in (b), the continuous estimation is shown. In both cases, the distribution of estimated COI as the level of sequence error assumed changes is plotted. In (a), dot plots are used where point size indicates density. In (b), a boxplot shows the interquartile range, with whiskers indicating the 95% quantile range and outliers indicated with dots. The mean absolute error is shown in (C) for both methods, with the 95% quantile computed from 1,000 bootstrap repetitions at each COI indicated with error bars.

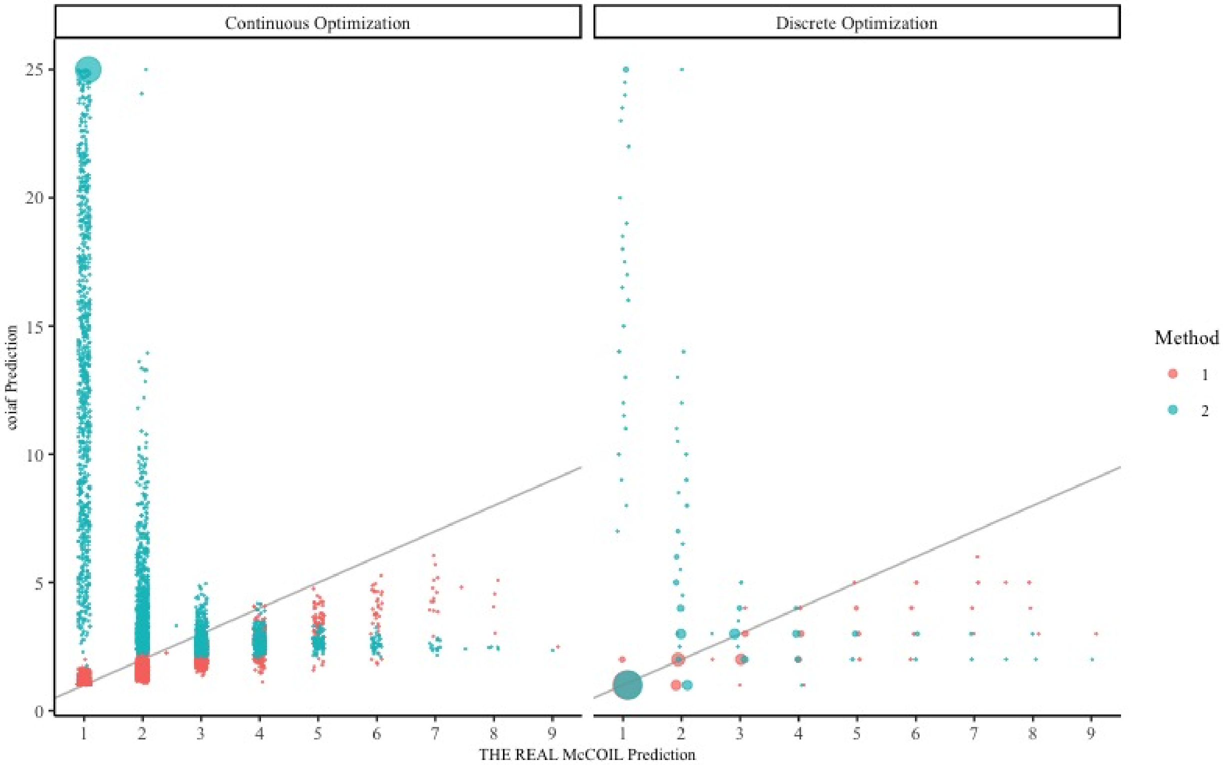


Figure 19. Comparison Between RMCL and coiaf. We estimated the COI for 5,970 samples grouped in 24 regions around the world. For each sample, we ran *THE REAL McCOIL* 5 times and ran our models 10 times, reporting the median estimated COI. We plot the discrete and continuous estimations of the COI for all points, regardless of their region. The color of the point indicates which Method was used and the size of the point indicates density.

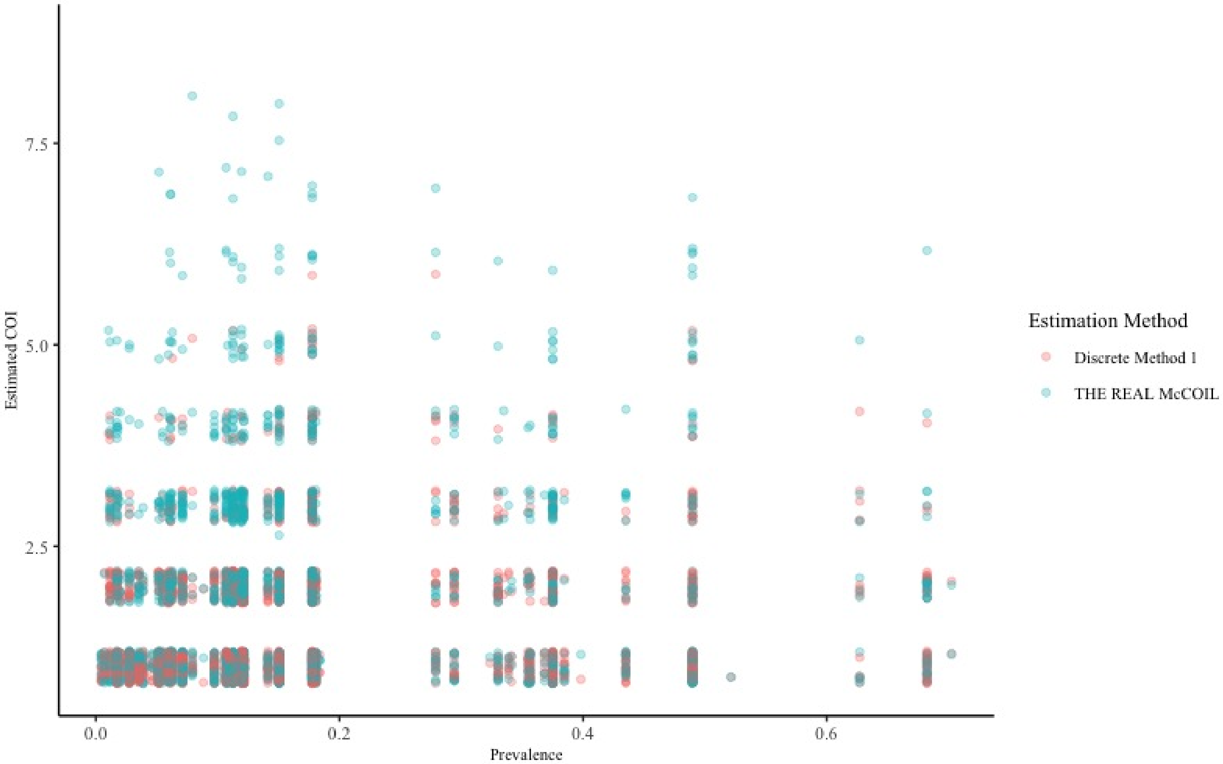


Figure 20. Comparison Between Estimation Methods and Malaria Prevalence. We estimated the COI for 5,970 samples grouped in 24 regions around the world. For each sample, we ran *THE REAL McCOIL* 5 times and ran our models 10 times, reporting the median estimated COI. The x-axis indicates the prevalence of malaria, which was obtained from the the Malaria Atlas Project [53–55]. The y-axis indicates the estimated COI when using either the discrete version of Method 1 or *THE REAL McCOIL*. The color of the points reflects whether the COI was estimated using the discrete version of Method 1 or *THE REAL McCOIL*. Note that each band of data corresponds to integer value of the COI—noise in the y-direction has been added for interpretability.