

## coiaf: directly estimating complexity of infection with allele frequencies

Aris Paschalidis<sup>1+\*</sup>, Oliver J. Watson<sup>1,2+</sup>, Ozkan Aydemir<sup>1,3</sup>, Robert Verity<sup>2</sup>, Jeffrey A. Bailey<sup>1</sup>

**1** Department of Pathology and Laboratory Medicine, Brown University, Providence RI, USA

**2** Department of Infectious Disease Epidemiology, Imperial College London, London, UK

**3** Program in Molecular Medicine, University of Massachusetts Medical School, Worcester MA, USA

+ These authors contributed equally to this work.

\* Corresponding author. E-mail: jeffrey\_bailey@brown.edu

## Abstract

In malaria, individuals are often infected with different parasite strains; the complexity of infection (COI) is defined as the number of genetically distinct parasite strains in an individual. Changes in the mean COI in a population have been shown to be informative of changes in transmission intensity with a number of probabilistic likelihood and Bayesian models now developed to estimate the COI. However, rapid, direct measures based on heterozygosity or *FwS* do not properly represent the COI. In this work, we present two new methods that use easily calculated measures to directly estimate the COI from allele frequency data. Using a simulation framework, we show that our methods are computationally efficient and comparably accurate to current methods in the literature. Through a sensitivity analysis, we characterize how the bias and accuracy of our two methods are impacted by the distribution of parasite densities and the assumed sequencing depth and number of sampled loci. We further estimate the COI globally from *Plasmodium falciparum* sequencing data using our developed

methods and compare the results against the literature. We show significant differences in estimated COI globally between continents and a weak relationship between malaria prevalence and COI.

## Author summary

Computational models, used in conjunction with rapidly advancing sequencing technologies, are increasingly being used to help inform surveillance efforts and understand the epidemiological dynamics of malaria. One such important metric, the complexity of infection (COI), indirectly quantifies the level of transmission. Existing “gold-standard” COI measures rely on complex probabilistic likelihood and Bayesian models. As an alternative, we have developed the statistics and software package *coiaf*, which features two rapid, direct measures to estimate of the number of genetically distinct parasite strains in an individual (the COI). Our methods were evaluated using simulated data and subsequently compared to current “state-of-the-art” methods, yielding comparable results. Lastly, we examined the distribution of the COI in several locations across the world, identifying significant differences in COI between continents. *coiaf*, therefore, provides a new, promising framework for rapidly characterizing polyclonal infections.

## Introduction

Malaria remains a leading cause of death worldwide—in 2020, there were an estimated 241 million cases, and 627,000 deaths around the globe [1]. Despite the considerable burden of malaria, these numbers represent the substantial global progress made to control malaria in the last two decades. The WHO reports that 1.5 billion malaria cases and 7.6 million malaria deaths were averted globally from 2000 to 2019 [2]. The majority of these gains reflect an increase in 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) [9–14]. However, evidence indicates that progress has slowed and that there is a need for new approaches to capitalize on the gains already made [2].

One approach is the use of computational methods, which often rely on recent advances in genetic sequencing and provide an increased understanding of malaria biology, to help inform control efforts [15–17]. For example, molecular and genomic epidemiology surveillance tools, which have been developing rapidly, can be used to track drug resistance and understand and evaluate control efforts [18, 19]. Moreover, computational methods have been applied in the identification of polyclonal infections—malaria infections with multiple distinct strains [20, 21]. Such polyclonal infections introduce additional genetic complexity that is often difficult to account for computationally. 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 [22].

Although determining the most informative metrics is an active field of investigation [23], one important metric is the complexity of infection (COI). Sometimes referred to multiplicity of infection, although this is generally reserved for infections within the same cell, the COI represents the number of genetically distinct malaria genomes or strains that can be identified in a particular individual. These polyclonal infections may arise from: (i) a single infectious mosquito feeds on a human host, transferring several genetically distinct parasite strains (often referred to as a co-transmission event [24, 25]) or (ii) two or more infectious mosquitoes with distinct malaria strains feed on an individual (known as a superinfection event [25, 26]). Measures of genetic diversity and the COI are increasingly used for inferring malaria transmission intensity and evaluating malaria control interventions [19]. 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 [27], age, and exposure acquired immunity [28]. Additionally, the COI provides a practical approach for identifying monoclonal infections to simplify genomic analyses.

Traditionally, the COI was measured in one or a few regions of the genome, relying on the enumeration of the maximal number of haplotypes detected through PCR amplification at genes or markers encoding highly diverse length polymorphisms. Two of the most common markers are the merozoite surface proteins 1 and 2 (*msp1* and *msp2*), surface proteins found on the merozoite stage of the malarial parasite [29, 30].

Traditional methods are hindered by limitations on the number of loci examined [31, 32],  
44 and lack the ability to detect parasites at low parasitaemia in samples [33].  
45

Furthermore, sanger sequencing lacks the sensitivity for low parasitemia samples  
46 without employing laborious subcloning [34].  
47

High-throughput sequencing provides more sensitive and specific methods. As a  
48 result, new computational methods have been developed to determine the COI. Two  
49 early proposed methods were the *FwS* metric by Auburn et al., which characterizes  
50 within-host diversity and its relationship to population-level diversity [31], and the  
51 *estMOI* software, which utilizes local phasing information of microhaplotypes within  
52 read pairs to estimate the COI [33]. Unfortunately, the *FwS* metric does not directly  
53 relate to the COI nor have a concrete biological interpretation, and the *estMOI*  
54 software relies on observed local haplotypes and heuristic interpretation. More recently,  
55 new tools been developed to better measure the COI beyond the maximal observed  
56 haplotype. For instance, the *DEploid* software package uses haplotype structure to infer  
57 the number of strains, their relative proportions, and the haplotypes present in a  
58 sample [35]. However, it is known that *DEploid* under predicts COI for high COI  
59 infections [25]. Other methods have been developed to examine the relatedness between  
60 parasite strains [36, 37]. The current “state-of-the-art” method for determining the COI  
61 of a sample is *THE REAL McCOIL*, which is an extension of *COIL* method [32]. *THE*  
62 *REAL McCOIL* employs a Bayesian approach, turning heterozygous data into estimates  
63 of allele frequency using Markov chain Monte Carlo methods and jointly estimating the  
64 likelihood of the COI [38].  
65

Despite various methods for estimating the COI, no rapid, direct measures have  
66 been developed to work effectively on a set of loci or at the genome-wide level. In this  
67 work, we present two new methods that use easily calculable metrics to directly  
68 estimate the COI from allele frequency data. Our two methods closely resemble the  
69 categorical and proportional methods implemented in *THE REAL McCOIL* [38], yet are  
70 geared towards large numbers of loci and providing rapid estimates of COI. Using our  
71 methods, we have developed the software package *coiaf* in the programming language  
72 , a language and environment for statistical computing and graphics [39].  
73

## Materials and methods

### Problem formulation

Current “state-of-the-art” approaches for estimating the COI rely on identifying the number of different parasites present in an infection using high-throughput sequencing. In monoclonal infections, i.e., infections composed of only one parasite strain (COI = 1), all sequence reads should be identical, originating from the same parasite strain. However, in mixed infections, a combination of each parasite strain, proportional to the strain’s abundance, will contribute to the observed sequence reads. At genetic loci containing variation, there is an increased chance of observing multiple alleles as the number of unrelated parasite strains within an infection increases. Therefore, the likelihood of observing multiple alleles at any locus is dependent on the number of parasite strains in an infection and the prevalence of genetic polymorphisms in the population.

We focus on only biallelic SNPs—the vast majority of loci—and define the major allele as the allele that is most prevalent in a population. We note that any multiallelic site can be collapsed to a biallelic site, although information will be lost. Assuming for any individual there are  $l$  biallelic loci, we define the within-sample allele frequency (WSAF) as a vector  $\hat{\mathbf{w}}$  of length  $l$  composed of the frequencies of the reference allele at each locus for a single individual infection and the population-level allele frequency (PLAF) as a vector  $\hat{\mathbf{p}}$  of length  $l$  composed of the frequencies of the reference allele at each locus across a population. The PLAF may be represented as the mean of the WSAF across a population, i.e.,  $\hat{\mathbf{p}} = \frac{1}{l} \sum_{i=1}^l \hat{\mathbf{w}}_i$ .

We further define the population-level minor allele frequency (PLMAF) as a vector  $\mathbf{p}$  of length  $l$  composed of the frequencies of the minor allele at each locus across a population, namely  $\mathbf{p} = (p_1, \dots, p_l)$ , where  $p_i \in [0, 0.5]$ . The within-sample minor allele frequency (WSMAF) is, additionally, defined as a vector  $\mathbf{w}$  of length  $l$  composed of the frequencies of the population-level minor allele at each locus for a single individual infection. For instance, the WSMAF will be equal to one when all sequence reads observed at a given locus are of the population-level minor allele. Given the WSAF and

the PLAF, the WSMAF and the PLMAF may be expressed as follows:

$$(\mathbf{w}, \mathbf{p}) = \begin{cases} (\hat{\mathbf{w}}, \hat{\mathbf{p}}) & \hat{\mathbf{p}} \leq 0.5 \\ 1 - (\hat{\mathbf{w}}, \hat{\mathbf{p}}) & \hat{\mathbf{p}} > 0.5 \end{cases} \quad (1)$$

## Variant and Frequency Methods

Our overall goal is to estimate the COI of the sample, denoted by  $k$ , using the WSMAF and the PLMAF. We do this by comparing observed data to derived expressions that define a relationship between the WSMAF, the PLMAF, and the COI. We present two alternative expressions that we refer to as the Variant Method and the Frequency Method.

In the Variant Method, we examine a set of SNPs and express the probability of SNP  $i$  being heterozygous with respect to the PLMAF and the COI. We define  $V_i$ , a Bernoulli random variable that takes the value of one if a site is heterozygous and zero otherwise. The probability that locus  $i$  is heterozygous, written as  $\mathbb{P}(V_i = 1)$ , will be equal to one minus the probability that a locus is homozygous (see Appendix A in S1 Appendices). We thus write,

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

As the COI increases, the probability of observing a heterozygous locus within an infection also increases (see Fig 1). We note that this is the same expression used within the Categorical method of *THE REAL McCOIL* (Eq (2)) [38].

**Fig 1. Flowchart of methods.** In **A**, the relationship between the WSMAF and the PLMAF is shown for an example simulation with a COI of 4. In **B**, data have been processed so that loci are deemed variant if they are heterozygous and invariant otherwise. In **C**, homozygous data have been filtered out. Following processing of data, Eq (2) and Eq (3) have been plotted for varying COIs from 1 to 4 in **D** and **E**, respectively.

In the Frequency Method, we focus on the expected value of the within-sample minor allele frequency. For the sake of simplicity, the complete derivation has been left to Appendix A in S1 Appendices. Briefly, we determine the probability of a particular strain carrying the minor allele and then determine the expected WSMAF by summing

over the expected contribution of each strain. We represent the expected value of the within-sample minor allele frequency given that a site is heterozygous as follows:

$$\mathbb{E}[W_i|V_i = 1] = \frac{p_i - p_i^k}{1 - p_i^k - (1 - p_i)^k}. \quad (3)$$

## Estimation method

Given data,  $D : \{(p_i, w_i, d_i), \quad i = 1, \dots, l\}$ , where  $p_i$  is the PLMAF at locus  $i$ ,  $w_i$  is the WSMAF at locus  $i$ , and  $d_i$  is the coverage at locus  $i$ , we next explore our methods to approximate the COI of a sample. Data are first processed to account for sequencing error. This process denotes loci at which there was suspected sequencing error as homozygous instead of heterozygous (for additional information, see Appendix C in S1 Appendices).

Following adjustment for sequence error, we consider an arbitrary data point  $(p_i, w_i, d_i)$ . Recall that the Variant Method and the Frequency Method examine different random variables. Specifically, the Variant Method identifies the probability of a locus being heterozygous,  $\mathbb{P}(V_i = 1)$ , and the Frequency Method identifies the expected value of the WSMAF given a site is heterozygous,  $\mathbb{E}[W_i|V_i = 1]$ . In order to determine the COI, we utilize Eq (2) and Eq (3). We solve the following minimization problem for the Variant Method:

$$\min_k \left( \sum_{i=1}^l \left( \frac{(v_i - (1 - p_i^k - (1 - p_i)^k))d_i}{\sum_{j=1}^l d_j} \right)^2 \right), \quad (4)$$

where  $v_i$  is defined as the value of the random variable  $V_i$  and  $q \geq 1$ . Note that  $\sum_{j=1}^l d_j$  is a constant and, therefore, we may remove it from our expression to get our final minimization problem:

$$\min_k \left( \sum_{i=1}^l ((v_i - (1 - p_i^k - (1 - p_i)^k))d_i)^2 \right). \quad (5)$$

Similarly, we solve the following minimization problem for the Frequency Method:

$$\min_k \left( \sum_{i=1}^l \left( \mathbb{1}_{\mathbb{P}(V_i=1)=1} \frac{\left( w_i - \left( \frac{p_i - p_i^k}{1 - p_i^k - (1 - p_i)^k} \right) d_i \right)^2}{\sum_{j=1}^l d_j} \right) \right), \quad (6)$$

where  $q \geq 1$ . Once again, we may remove the constant from our expression to get our final minimization problem:

$$\min_k \left( \sum_{i=1}^l \left( \mathbb{1}_{\mathbb{P}(V_i=1)=1} \left( w_i - \left( \frac{p_i - p_i^k}{1 - p_i^k - (1 - p_i)^k} \right) d_i \right)^2 \right) \right). \quad (7)$$

Note that the estimation methods described minimize the sum of squared residuals between the observed data and the relationships derived in Eq (2) and Eq (3).

## Solution methods

We solve this optimization problem using two methods: (i) assuming discrete values of the COI and (ii) assuming 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. While a continuous value of the COI has no direct biological interpretation, significant departures from discrete values are expected due to parasite relatedness, and a range of other biological phenomenon including, but not limited to, overdispersion in sequencing and parasite densities. Therefore, a continuous value of the COI may provide a more accurate representation of the overall population of samples being studied. Furthermore, as relatedness in mixed infections is common [37], a continuous COI may be able to provide insights into the degree of relatedness between parasite strains in mixed infections and detect highly-related polyclonal infections that may traditionally be categorized as monoclonal.

To solve the discrete versions of the previously defined optimization problems we use a brute force approach, which 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. To solve the continuous versions of the optimization problems, we utilize R's built-in optimization function [39]. In particular, we leverage a quasi-Newton L-BFGS-B approach with box constraints [40]. We set the lower and upper bounds of the COI as 1 and 25, respectively, with the default starting value of the

COI equal to two. Note that in both the discrete and continuous case, the upper bound of the COI is much larger than most COIs seen in the real world [38,41]. In both cases, we also provide the capability to determine the 95% confidence interval for our COI estimates by leverage bootstrapping techniques [42] (see Appendix E in S1 Appendices for more details).

## Data

To evaluate the accuracy and sensitivity of our methods, we created a simulator that generates synthetic sequencing data for a number of individuals in a given population. In overview, each individual is assigned a COI value. The haplotype of each strain is then assigned by sampling from the population level minor allele frequency. Next, we simulate the number of sequence reads mapped to the reference and alternative allele by sampling in proportion to the parasite densities for each strain. After simulating sequence error, the mapped sequence reads are then used to derive the within-sample minor allele frequency. A detailed description of our simulator can be found in Appendix F in S1 Appendices.

In addition to simulated data, we use sequencing data sampled from infected individuals worldwide to compare our methods to the current state-of-the-art COI estimation metric and to investigate the distribution of the COI across the world. We analysed over 7,000 *P. falciparum* samples from 28 malaria-endemic countries in Africa, Asia, South America, and Oceania from 2002 to 2015 from the MalariaGEN *Plasmodium falciparum* Community Project [43]. 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. [43]. We used the provided variant call files (VCFs) generated using a standard analysis pipeline. The median read depth of coverage of the initially sequenced field isolates was 73 across across all samples. After removing replicate samples, mixed-species samples, and samples with a low coverage, suspected contamination or mislabelling, 5,970 samples remained for further analysis. Genomic data were further filtered for high quality biallelic coding and non-coding SNPs as outlined in Zhu et al. [37]. Additionally, data were filtered to sites that are part of the core genome.

To apply our developed methods, we must estimate the population-level frequency of the minor allele. Consequently, we sought to assign the samples to a suitable number of geographic regions such that the number of samples per region was suitable for the reliable estimation of the population-level minor allele frequencies. We used the Partitioning Around Medoids (PAM) algorithm to solve a k-medoids clustering problem [44, 45] to group samples based on the longitude and latitude of sample collection. We next calculated the silhouette information for each clustering of k groups [46], arriving at 24 regions globally (see Appendix K.2 in S1 Appendices for a map of locations). Given these 24 clusters, we filtered SNPs to variants with a population level alternative allele frequency greater than 0.005 in each region. The 0.005 frequency cutoff was chosen as sequence error likely obscures the detection of true variation from parasite strains comprising less than 0.5% of total parasite density. Clusters of data were additionally traced to a specific continent and subregion as defined by the World Development Indicators [47, 48].

## Results

### Performance on simulated data

Using our simulator, we simulated data for 1,000 loci with a read depth of 100 at each locus. Data was simulated with a complexity of infection ranging from 1 to 20. This simulation did not introduce error in order to determine optimal performance based on sampling. Our methods, therefore, accounted for no sequencing error. The results of running the discrete version of both the Variant Method and the Frequency Method are illustrated in Fig 2A and Fig 2B, respectively.

**Fig 2. Estimating the COI on simulated data.** The performance of the Variant Method (A) and Frequency Method (B) is shown for 100 simulations of a COI of 1-20 with 1,000 loci, a read depth of 100, 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 C. The black bars indicate the 95% confidence interval.

The Variant Method and the Frequency Method both perform well for all COIs between 1 and 20. Notably, the lower the true value of the COI, the better our models perform with a mean absolute error close to zero. As the COI increases, our models

exhibit more variability across subsequent iterations and underestimate the true COI. 189  
For example, at a COI of 20, the estimated COI from the Frequency Method ranges 190  
from 14 to 22—the maximum COI our model could output was 25 in these trials. 191  
Nevertheless, the majority of predicted COIs remain close to the true COI as witnessed 192  
by the low mean absolute errors of at most 3.32 in Fig 2C. Comparing our two methods, 193  
we find that the Variant Method and the Frequency Method perform equally with 194  
insignificantly different mean absolute errors (p-value = 0.396) and biases (p-value = 195  
0.661). 196

## Sensitivity analysis

 197

In order to understand the sensitivity of our models to alterations in the parameters 198  
considered, we tested the performance of the discrete and continuous representations of 199  
the Variant Method and the Frequency Method, assessing changes in the accuracy of 200  
our predictions. For each sample, we utilized bootstrapping techniques [49, 50] to 201  
determine the mean absolute error and bias of the predicted COI compared to the true 202  
COI. Furthermore, we ran each algorithm several times to ensure reliable results. A 203  
description of several key parameters perturbed and their default values can be found in 204  
Appendix G in S1 Appendices. Resulting figures can be found in Appendix K.1 in S1 205  
Appendices. 206

Here, we highlight the effect of varying two metrics that can be controlled in the 207  
field: the read depth at each locus and the number of loci sequenced. Sequencing more 208  
loci at larger read depths is preferred as this results in more higher quality data. In 209  
general, as the coverage at each locus increases, the performance of our methods also 210  
improves (see Appendix K.1 Fig 2 and Fig 3 in S1 Appendices). This relationship, 211  
however, is not linear. Rather, a non-monotonic relationship is observed between 212  
sequence coverage and the mean absolute error, with diminishing returns in 213  
performance observed with sequence coverage greater than 100. As was the case for our 214  
coverage data, when the number of loci sequenced is low, around 100 loci, our methods 215  
have a high variability and tend to underpredict the true COI (see Appendix K.1 Fig 4 216  
and Fig 5 in S1 Appendices). However, as the number of loci increases to 1,000, the 217  
performance increases. In addition, increasing the number of loci examined above a 218

certain threshold, in this case 1,000 loci, does not seem to substantially impact the 219 performance of our models. However, note that an increase in the number of loci does 220 reduce the variance of our estimates. 221

## Comparison to “state-of-the-art” methods 222

In this section we compare our novel methods to the current “state-of-the-art” method 223 used to estimate the COI, *THE REAL McCOIL* [38]. As was previously described, we 224 grouped our data into 24 regions around the world. To estimate the COI for each of the 225 5,970 samples, we examined an average of 32,362 (range: 15, 276-40,272) loci in each 226 region (see Appendix I in S1 Appendices). Furthermore, we ran five repetitions of the 227 *THE REAL McCOIL* on each sample, with a burn-in period of 1,000 iterations followed 228 by 5,000 sampling iterations, and using standard methodology to confirm convergence 229 between Monte Carlo Markov chains [51]. For additional information, see Appendix J in 230 S1 Appendices. Fig 3 examines the COI estimation of *THE REAL McCOIL* and *coiaf* 231 for all samples. We note that the relationship between *coiaf*’s estimated COI and *THE* 232 *REAL McCOIL*’s estimated COI for each of the 24 individual regions is shown in 233 Appendix K.3 Fig 1 and Fig 2 in S1 Appendices. 234

**Fig 3. Comparison Between *THE REAL McCOIL* and *coiaf*.** The COI estimation using (A) the Variant Method and (B) the Frequency Method is compared against the *THE REAL McCOIL*. In C the distribution of differences between our estimation and the *THE REAL McCOIL*’s estimation is shown.

We observe that the Variant Method and the Frequency Method are strongly 235 correlated with the estimates from *THE REAL McCOIL*. In (Fig 3A and Fig 3B), When 236 the COI is estimated to be below five, both methods estimate COI values that are close 237 to one another. However, as the estimated COI increases, there is a greater variability 238 in predictions. At these high COI values, our methods tend to estimate the COI within 239 three of the true COI (Fig 3C). As expected, the continuous estimation methods align 240 with the discrete estimation methods. Furthermore, we note that the Frequency Method 241 does not show estimates for when *THE REAL McCOIL* predicts a COI of one. This is 242 because the Frequency Method at a COI of one is undefined; at a COI of one, there 243 would be no heterozygous loci, which are used in the Frequency Method. To quantify 244 the relationship between our novel software and the current “state-of-the-art” method, 245

we fit linear regression models to the data and evaluated the Pearson correlation  
246  
between estimation methods. The results are reported in Table 1 and indicate that each  
247  
of the methods introduced in *coiaf* is highly correlated with *THE REAL McCOIL*.  
248

<i>coiaf Estimation Method</i>	Linear Regression		Correlation
	R <sup>2</sup>	P-value	
Discrete Variant Method	0.840	< 0.001	0.916
Continuous Variant Method	0.883	< 0.001	0.940
Discrete Frequency Method	0.804	< 0.001	0.896
Continuous Frequency Method	0.844	< 0.001	0.919

**Table 1. Relationship Between *coiaf* and *THE REAL McCOIL*.** A linear regression model was fit to the data to evaluate the relationship between *coiaf*'s and *THE REAL McCOIL*'s estimation method. Furthermore, the Pearson correlation between the estimated COI's was computed.

## Mapping COI worldwide

To demonstrate *coiaf*'s utility and better understand global patterns of the COI, we  
250  
examined the distribution of COI in twenty-four different regions. In each region we  
251  
studied an average of 248 (range: 29 to 909) samples and 32,362 loci (range: 15,276 to  
252  
40,272) (see Appendix I in S1 Appendices). Our samples can be traced to four different  
253  
continents, with most originating in either Africa (55.5%) or Asia (41.8%). All other  
254  
samples were sequenced in Oceania (2.03%) or the Americas (0.620%).  
255

Fig 4 highlights the mean and median COI across all samples in each of the 24  
256  
regions outlined previously. We, furthermore, aimed to understand the relationship  
257  
between the complexity of infection and malaria prevalence by leveraging estimates of  
258  
malaria microscopy prevalence in children aged 2–10 years old generated by the Malaria  
259  
Atlas Project [9, 10, 52]. Fig 4C plots the density of the COI for each region sorted by  
260  
the region's malaria prevalence. Table 2 outlines the mean COI in each of the four  
261  
continents and seven subregions analyzed.  
262

**Fig 4. COI Across the Globe.** In A we plot the mean COI of all samples in each  
263  
study location within the 24 regions, in B we plot the mean COI. The color and size of  
each point represents the magnitude of the COI. In C, we draw a density plot for each  
region, where the color of the plot indicates in what subregion the data was sampled.  
The plots are sorted by the median microscopy prevalence in children aged 2–10 years  
old as estimated in the Malaria Atlas Project [9, 10, 52] and indicated to the right of  
each density plot.

Of all the continents, Africa has the highest mean COI of 1.87, followed by Asia with  
263

Continent	Subregion	Number of Samples	Mean COI (SD)
Africa	Eastern Africa	739	1.88 (1.10)
Africa	Middle Africa	579	1.87 (0.948)
Africa	Western Africa	1,996	1.87 (1.05)
Americas	South America	37	1.03 (0.164)
Asia	South-Eastern Asia	2,421	1.25 (0.506)
Asia	Southern Asia	77	1.52 (0.641)
Oceania	Melanesia	121	1.20 (0.440)

**Table 2. Mean COI.** Mean COI across each continent and subregion analyzed.

a mean COI of 1.26, Oceania with a mean COI of 1.20, and the Americas with a mean COI of 1.03. A Nemenyi post-hoc test [53] indicates that while Africa is statistically different than all the other continents (p-value: < 0.001 in all cases), there exist no significant differences between each pairing of the other three continents (Americas vs Asia p-value: 0.204, Americas vs Oceania: p-value: 0.568, p-value: Asia vs Oceania: 0.816). Within each continent, there exist further differences among the subregions. No statistically significant difference was found between each of the three subregions in Africa (Eastern vs Middle p-value: 0.914, Eastern vs Western p-value: > 0.999, Middle vs Western p-value: 0.886). However, in Asia, there was a statistically significant difference between the mean COI in South-Eastern Asia and Southern Asia (p-value: 0.0282).

We found a positive correlation between the COI and the microscopy prevalence (Table 3). In regions with a lower prevalence, there were few samples with a COI larger than two. In fact, in regions with a prevalence less than or equal to 0.01, more than 95% of samples had a COI of one or two. In regions with a higher prevalence, there were more samples with a higher COI—in regions where the prevalence was greater than or equal to 0.1, more than 20% of samples had a COI greater than two.

<i>coiaf</i> Estimation Method	Linear Regression		Correlation
	R <sup>2</sup>	P-value	
Discrete Variant Method	0.0770	< 0.001	0.278
Continuous Variant Method	0.0787	< 0.001	0.281
Discrete Frequency Method	0.0190	< 0.001	0.138
Continuous Frequency Method	0.0218	< 0.001	0.148

**Table 3. Relationship Between *coiaf* and malaria prevalence.** A linear regression model was fit to the data to evaluate the relationship between the COI and prevalence. Furthermore, the Pearson correlation was examined.

## Discussion

Despite advances in sequencing technologies and the development of various methods for estimating the COI, no direct measures have been developed to rapidly estimate the COI. Here we develop two direct measures based on minor allele frequencies that address these needs. Our methods can provide rapid estimates of the COI in under a second. When compared to the current state-of-the-art COI estimation software, *THE REAL McCOIL*, *coiaf* was able to estimate the COI with similar accuracy approximately 250 times faster (see Appendix H in S1 Appendices) and required more than eight times less memory. Moreover, our methods can be scaled to fully use all available SNPs within the genome and provide a continuous measure that can provide insight into relatedness.

Through a number of simulations, we further explored how changing key sequencing variables, such as the number of loci and read depth at each locus, altered our software's performance. We showed that for samples with a low and moderate COI our methods were able to accurately predict the COI even with a low coverage and number of loci, however, as the COI increased, these parameters became more important—a lack of sufficient sequencing corresponded with underprediction of the true COI. Additionally, we demonstrated that there are several important factors that can greatly influence results, such as sequencing error or overdispersion in parasite density. Importantly, we also show that the population mean WSMAF is an unbiased estimator of the PLMAF (see Appendix B in S1 Appendices). This provides further advantages to using allelic read depth for COI estimation rather than haplotype calls, which are known to lead to biased estimates of population allele frequency if the COI of samples is not accounted for [54].

An application of *coiaf* on several thousand *P. falciparum* samples from malaria-endemic countries in four continents from 2002-2015 [43] resulted in a comprehensive map of the complexity of infection worldwide. This study builds on previous reviews of the distribution of the COI globally [55], and is the first study to our knowledge to provide a holistic view of the COI based on allelic read depth as opposed to traditional methods leveraging *msp1* and *msp2* haplotyping.

In general, our results were in agreement with previously reported findings. For instance, in areas with historically lower malaria prevalence, such as South America and

Southern Asia, we estimated lower average COI values. In particular, in the Americas and Asia we report a mean COI of 1.03 and 1.26, respectively. Previous efforts to estimate the COI in these region have also found similarly low values of the COI. While directly comparing our estimates against these would be incorrect as the date and location of sample collection are very different, we are encouraged by the similarity in estimates of the COI. For example, in Brazil, the COI has been previously measured as low as 1.1 in the early 1990s [56]. In Papua New Guinea [41,57,58], Bangladesh [59] and Malaysia [60] previous estimates of COI range between 1.00 and 2.12, 1.22 and 1.58 and 1.20 and 1.37 respectively. In Africa, on the other hand, we were surprised to find little to no difference in average COI estimates between the three subregions we studied: Eastern Africa, Middle Africa, and Western Africa, despite large differences in average malaria prevalence in these regions. In contrast, previous studies of the COI in African settings have found both higher average COI values in some regions and, in general, greater variability. For example, in Cameroon large mean COIs between 2.33 to 3.82 have been reported [61–63]. Conversely, in Ghana the mean COI has been reported to be between 1.13 and 1.91 in 2012-2013 [64].

Much of the work surrounding the complexity of infection is motivated by the fact that the COI has been proposed as an indicator of transmission intensity. Unfortunately, the relationship between the COI and malaria prevalence remains an area of much debate, with many individual studies finding different relationships between the COI and malaria prevalence [65,66]. Lopez and Koepfli report in their review article that across the 153 studies examined there was a weak correlation between mean COI and prevalence, an observation that agrees with our findings [55]. As previously noted, there are multiple patient-level factors (e.g., age and clinical status) which may affect the relationship between malaria prevalence and the COI. Additionally, Karl et al. suggested that this weak correlation may be attributed to spatial effects and the existence of geographic “hotspots” where transmission may be much higher than in surrounding areas, causing individuals to have a greater COI [67]. Moreover, multiple studies have highlighted that seasonality affects observed COI [68,69]. Consequently, while there is undoubtedly a relationship between transmission intensity and COI, it is important to be aware of how many factors (age, clinical status, seasonality, spatial effects, parasite density, time since last infection, methods of detecting multiple

infections) may impact this relationship. For example, while analysing the data from the MalariaGEN *Plasmodium falciparum* Community Project [43], metadata is only available regarding the year and location of sample collection. Without being able to account for the other factors known to impact COI, we are unable to make more meaningful interpretations of our analysis of COI patterns. Lastly, it is worth recalling that malaria prevalence is not directly related to transmission intensity. For example, two regions with the same malaria prevalence will likely have different transmission intensities if intervention coverage differs between the region. Therefore, the variance observed between malaria prevalence and the COI may reflect that malaria prevalence is itself an imperfect predictor of transmission intensity.

Our work is not without limitations. In part, limitations stem from the fact that our methods rely on certain biological assumptions, which may not be met in the real world. Additionally, the accuracy of our algorithms is impacted by sequence error. While this was not an issue in our analysis of the MalariaGEN *Plasmodium falciparum* Community Project (see Appendix K.3 Fig 5 in S1 Appendices), high levels of sequence error need to be monitored and accounted for. While our software package does allow users to account for this by providing a level of suspected sequencing error, sequence error is unlikely to be constant across the genome and accurate inference of sequencing error is itself an active research challenge [70]. Furthermore, while our methods do account for the coverage at each locus, if there is a low overall coverage for a sample, our results may underpredict the true COI. Lastly, our methods assume that the population level minor allele frequency (PLMAF) is well captured by the samples provided. Sampling bias or undersampling in a region may result in an inaccurate PLMAF estimation, which may influence our estimated COI.

In conclusion, we developed two direct measures for estimating the COI given the within sample allele frequency of a sample and the population level allele frequency. Our methods were able to estimate the COI for samples in less than a second and were shown to be accurate when compared to simulated data and current COI estimation techniques. Our software will aid in the estimation of the complexity of infection, an increasingly important population genetic metric for inferring malaria transmission intensity and evaluating malaria control interventions [19, 71].

## Supporting information

**S1 Appendices.** Appendices. Includes the complete problem formulation including derivations of the Variant Method and the Frequency Method, additional information about our methods, and all omitted figures.

375

376

377

378

## Acknowledgments

379

We thank the MalariaGEN *Plasmodium falciparum* Community Project for maintaining a large collection of sequencing data and variant calls.

380

381

## References

1. Organization WH. World malaria report 2021. World Health Organization; 2021.
2. Organization WH, others. World malaria report 2020: 20 years of global progress and challenges. World Health Organization; 2020.
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. 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.
10. 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.
11. 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>.
12. 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.
13. Prevention CCfDCa. CDC - Malaria - Diagnosis & Treatment (United States) - Diagnosis (U.S.); 2019. Available from: [https://www.cdc.gov/malaria/diagnosis\\_treatment/diagnosis.html](https://www.cdc.gov/malaria/diagnosis_treatment/diagnosis.html).
14. 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.
15. 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.
16. 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.

17. 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.
18. Kümporns K, Kochakarn T, Chookajorn T. The resistome and genomic reconnaissance in the age of malaria elimination. *Disease Models & Mechanisms*. 2019;12(12). doi:10.1242/dmm.040717.
19. 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.
20. 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.
21. 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.
22. 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.
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. 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.

25. 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.
26. 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.
27. 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.
28. 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.
29. 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.
30. Konaté L, Zwetyenga J, Rogier C, Bischoff E, Fontenille D, Tall A, et al. 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.
31. 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.
32. 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.

33. 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.
34. Hu T, Chitnis N, Monos D, Dinh A. Next-generation sequencing technologies: An overview. *Human Immunology*. 2021;82(11):801–811. doi:10.1016/j.humimm.2021.02.012.
35. 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.
36. 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.
37. 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.
38. 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.
39. 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/>.
40. 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.
41. Fola AA, Harrison GLA, Hazairin MH, Barnadas C, Hetzel MW, Iga J, et al. Higher Complexity of Infection and Genetic Diversity of *Plasmodium vivax* Than

Plasmodium falciparum across all Malaria Transmission Zones of Papua New Guinea. *The American Journal of Tropical Medicine and Hygiene*. 2017;96(3):630–641. doi:10.4269/ajtmh.16-0716.

42. Davison AC, Hinkley DV. *Bootstrap methods and their applications*. Cambridge: Cambridge University Press; 1997. Available from: <http://statwww.epfl.ch/davison/BMA/>.
43. MalariaGEN, Ahoudi A, Ali M, Almagro-Garcia J, Amambua-Ngwa A, Amaralunga 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.
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. Bank W. *World Development Indicators*. The World Bank;.
48. Arel-Bundock V, Enevoldsen N, Yetman C. *countrycode: An R package to convert country names and country codes*. *Journal of Open Source Software*. 2018;3(28):848. doi:10.21105/joss.00848.
49. Mooney CZ, Mooney CF, Mooney CL, Duval RD, Duvall R. *Bootstrapping: A Nonparametric Approach to Statistical Inference*. SAGE; 1993.
50. DiCiccio TJ, Efron B. Bootstrap confidence intervals. *Statistical Science*. 1996;11(3):189–228. doi:10.1214/ss/1032280214.
51. Gelman A, Rubin DB. Markov chain Monte Carlo methods in biostatistics. *Statistical Methods in Medical Research*. 1996;5(4):339–355. doi:10.1177/096228029600500402.

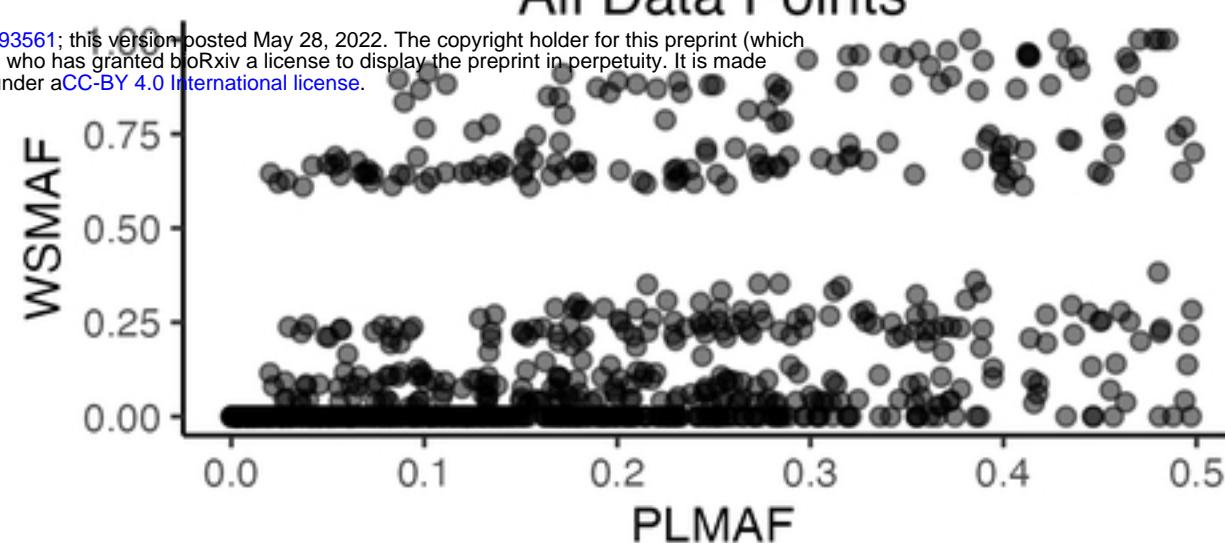
52. Weiss DJ, Lucas TCD, Nguyen M, Nandi AK, Bisanzio D, Battle KE, et al. Mapping the global prevalence, incidence, and mortality of *Plasmodium falciparum*, 2000–17: a spatial and temporal modelling study. *The Lancet*. 2019;394(10195):322–331. doi:10.1016/S0140-6736(19)31097-9.
53. Nemenyi PB. Distribution-free multiple comparisons. Princeton University; 1963.
54. Hastings IM, Nsanzabana C, Smith TA. A Comparison of Methods to Detect and Quantify the Markers of Antimalarial Drug Resistance. *The American Journal of Tropical Medicine and Hygiene*. 2010;83(3):489–495. doi:10.4269/ajtmh.2010.10-0072.
55. Lopez L, Koepfli C. Systematic review of *Plasmodium falciparum* and *Plasmodium vivax* polyclonal infections: Impact of prevalence, study population characteristics, and laboratory procedures. *PLoS ONE*. 2021;16(6):e0249382. doi:10.1371/journal.pone.0249382.
56. Sallenave-Sales S, Daubersies P, Mercereau-Puijalon O, Rahimalala L, Contamin H, Druilhe P, et al. *Plasmodium falciparum*: a comparative analysis of the genetic diversity in malaria-mesoendemic areas of Brazil and Madagascar. *Parasitology Research*. 2000;86(8):692–698. doi:10.1007/PL00008554.
57. Mita T, Hombhanje F, Takahashi N, Sekihara M, Yamauchi M, Tsukahara T, et al. Rapid selection of sulphadoxine-resistant *Plasmodium falciparum* and its effect on within-population genetic diversity in Papua New Guinea. *Scientific Reports*. 2018;8:5565. doi:10.1038/s41598-018-23811-7.
58. Barry AE, Schultz L, Senn N, Nale J, Kiniboro B, Siba PM, et al. High Levels of Genetic Diversity of *Plasmodium falciparum* Populations in Papua New Guinea despite Variable Infection Prevalence. *The American Journal of Tropical Medicine and Hygiene*. 2013;88(4):718–725. doi:10.4269/ajtmh.12-0056.
59. Alam MS, Elahi R, Mohon AN, Al-Amin HM, Kibria MG, Khan WA, et al. *Plasmodium falciparum* Genetic Diversity in Bangladesh Does Not Suggest a Hypoendemic Population Structure. *The American Journal of Tropical Medicine and Hygiene*. 2016;94(6):1245–1250. doi:10.4269/ajtmh.15-0446.

60. Atroosh WM, Al-Mekhlafi HM, Mahdy MA, Saif-Ali R, Al-Mekhlafi AM, Surin J. Genetic diversity of *Plasmodium falciparum* isolates from Pahang, Malaysia based on MSP-1 and MSP-2 genes. *Parasites & Vectors*. 2011;4(1):233. doi:10.1186/1756-3305-4-233.
61. Apinjoh TO, Tata RB, Anchang-Kimbi JK, Chi HF, Fon EM, Mugri RN, et al. *Plasmodium falciparum* merozoite surface protein 1 block 2 gene polymorphism in field isolates along the slope of mount Cameroon: a cross – sectional study. *BMC Infectious Diseases*. 2015;15(1):309. doi:10.1186/s12879-015-1066-x.
62. Roman DNR, Anne NNR, Singh V, Luther KMM, Chantal NEM, Albert MS. Role of genetic factors and ethnicity on the multiplicity of *Plasmodium falciparum* infection in children with asymptomatic malaria in Yaoundé, Cameroon. *Heliyon*. 2018;4(8):e00760. doi:10.1016/j.heliyon.2018.e00760.
63. Metoh TN, Chen JH, Fon-Gah P, Zhou X, Moyou-Somo R, Zhou XN. Genetic diversity of *Plasmodium falciparum* and genetic profile in children affected by uncomplicated malaria in Cameroon. *Malaria Journal*. 2020;19(1):115. doi:10.1186/s12936-020-03161-4.
64. Duah NO, Matrevi SA, Quashie NB, Abuaku B, Koram KA. Genetic diversity of *Plasmodium falciparum* isolates from uncomplicated malaria cases in Ghana over a decade. *Parasites & Vectors*. 2016;9(1):416. doi:10.1186/s13071-016-1692-1.
65. Bei AK, Niang M, Deme AB, Daniels RF, Sarr FD, Sokhna C, et al. Dramatic Changes in Malaria Population Genetic Complexity in Dielmo and Ndlop, Senegal, Revealed Using Genomic Surveillance. *The Journal of Infectious Diseases*. 2018;217(4):622–627. doi:10.1093/infdis/jix580.
66. Pacheco MA, Forero-Peña DA, Schneider KA, Chavero M, Gamardo A, Figuera L, et al. Malaria in Venezuela: changes in the complexity of infection reflects the increment in transmission intensity. *Malaria Journal*. 2020;19(1):176. doi:10.1186/s12936-020-03247-z.
67. Karl S, White MT, Milne GJ, Gurarie D, Hay SI, Barry AE, et al. Spatial Effects on the Multiplicity of *Plasmodium falciparum* Infections. *PLOS ONE*. 2016;11(10):e0164054. doi:10.1371/journal.pone.0164054.

68. Watson OJ, Verity R, Ghani AC, Garske T, Cunningham J, Tshefu A, et al. Impact of seasonal variations in *Plasmodium falciparum* malaria transmission on the surveillance of pfhrp2 gene deletions. *eLife*. 2019;8:e40339. doi:10.7554/eLife.40339.
69. Sondo P, Derra K, Rouamba T, Nakanabo Diallo S, Taconet P, Kazienga A, et al. Determinants of *Plasmodium falciparum* multiplicity of infection and genetic diversity in Burkina Faso. *Parasites & Vectors*. 2020;13:427. doi:10.1186/s13071-020-04302-z.
70. 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.
71. Zhong D, Koepfli C, Cui L, Yan G. Molecular approaches to determine the multiplicity of *Plasmodium* infections. *Malaria Journal*. 2018;17(1):172. doi:10.1186/s12936-018-2322-5.

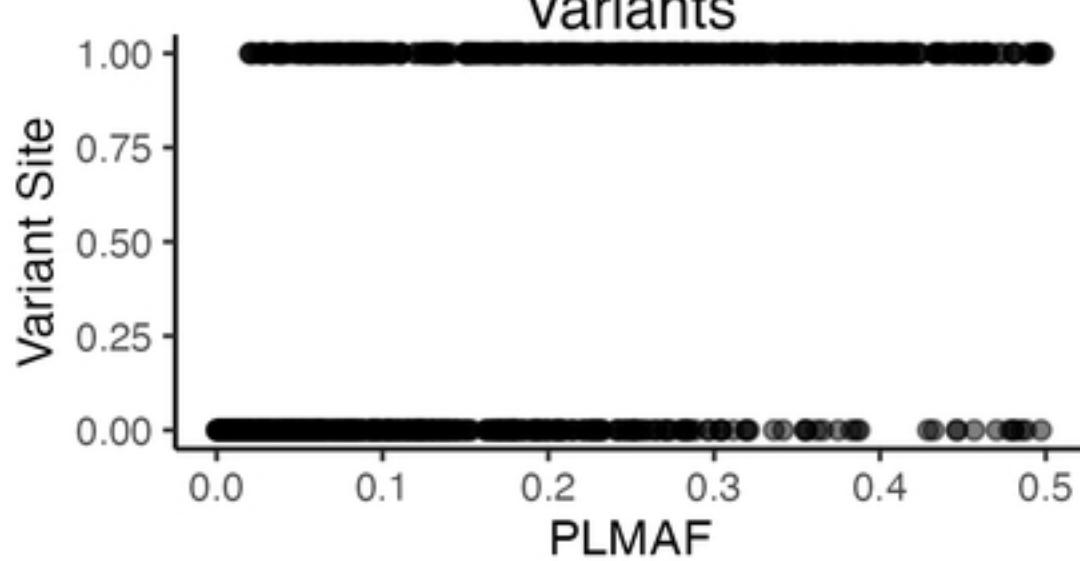
**A**

All Data Points



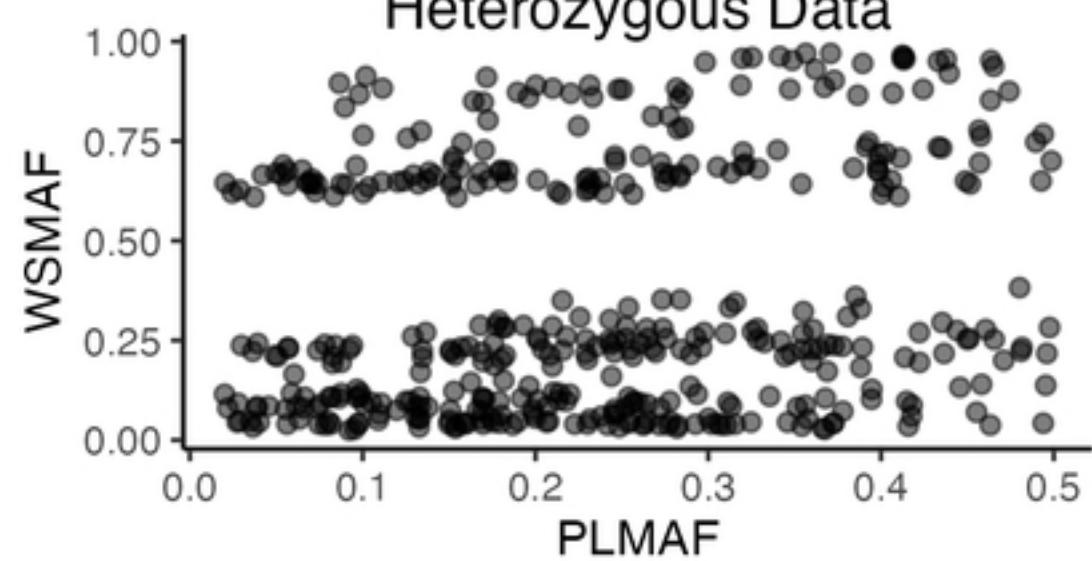
**B**

Variants



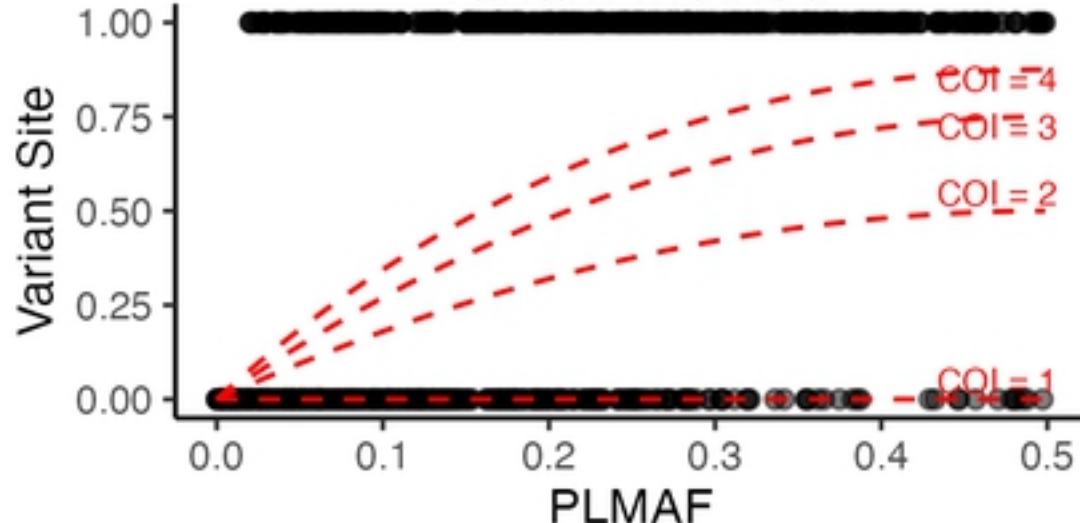
**C**

Heterozygous Data



**D**

Variant Method



**E**

Frequency Method

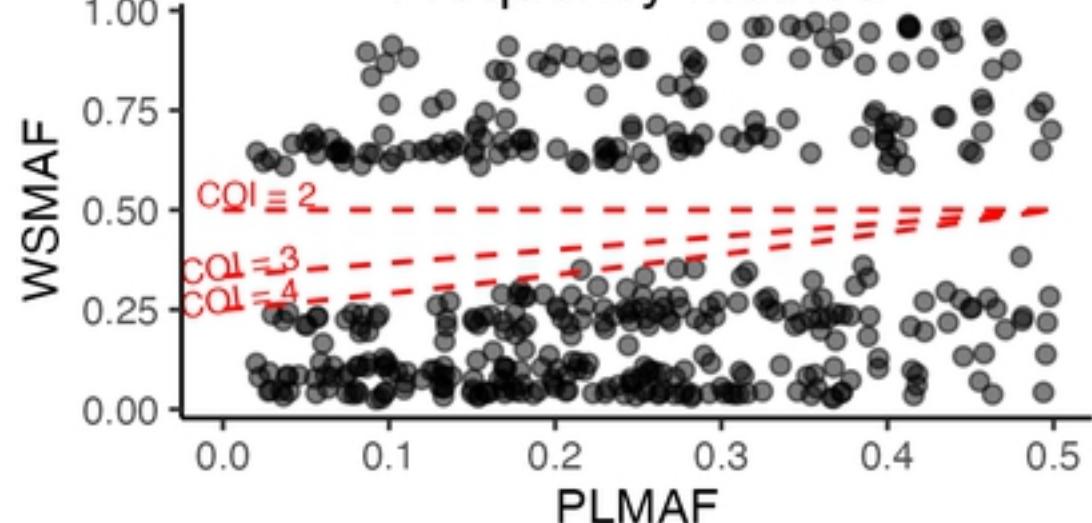
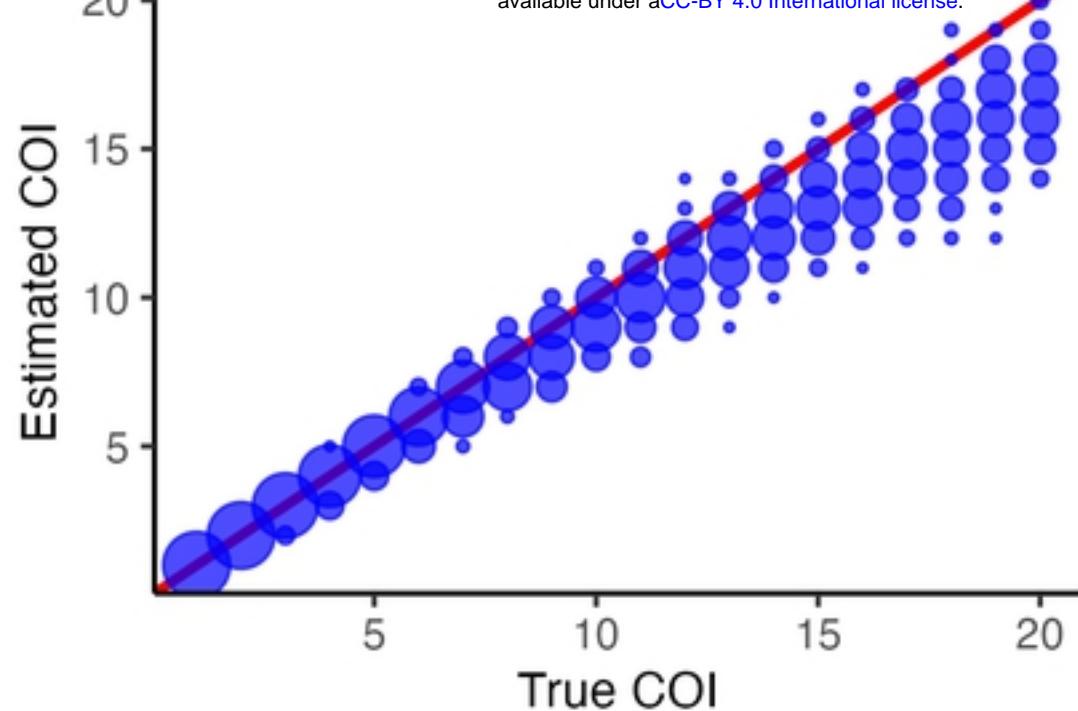
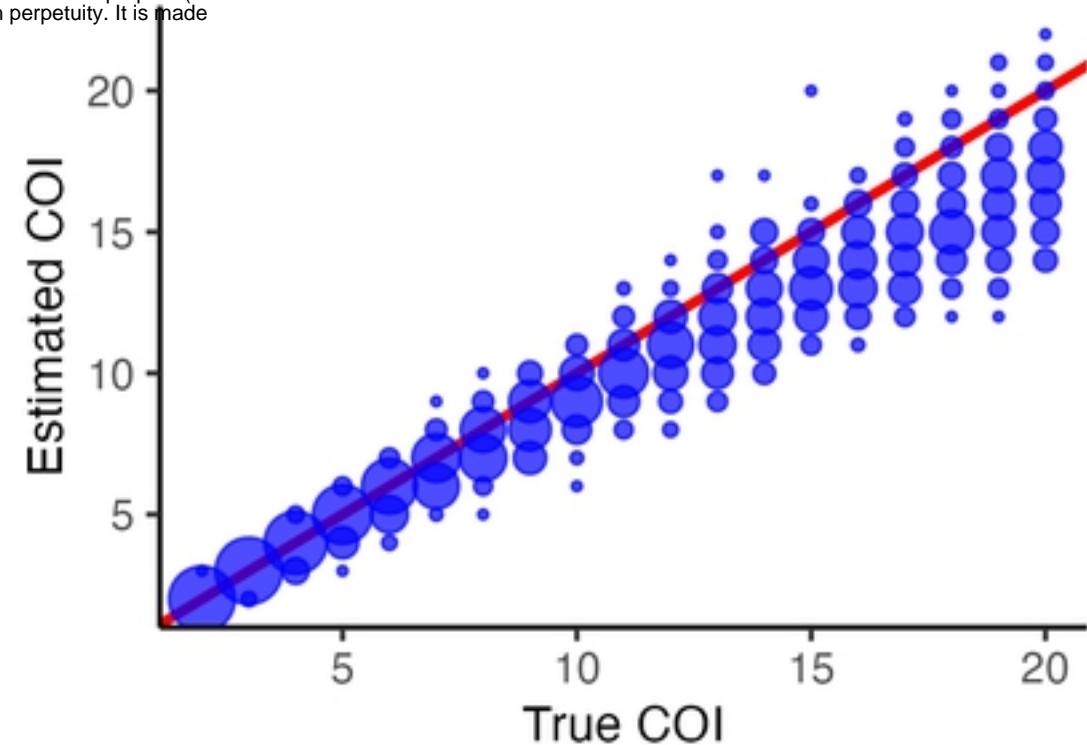
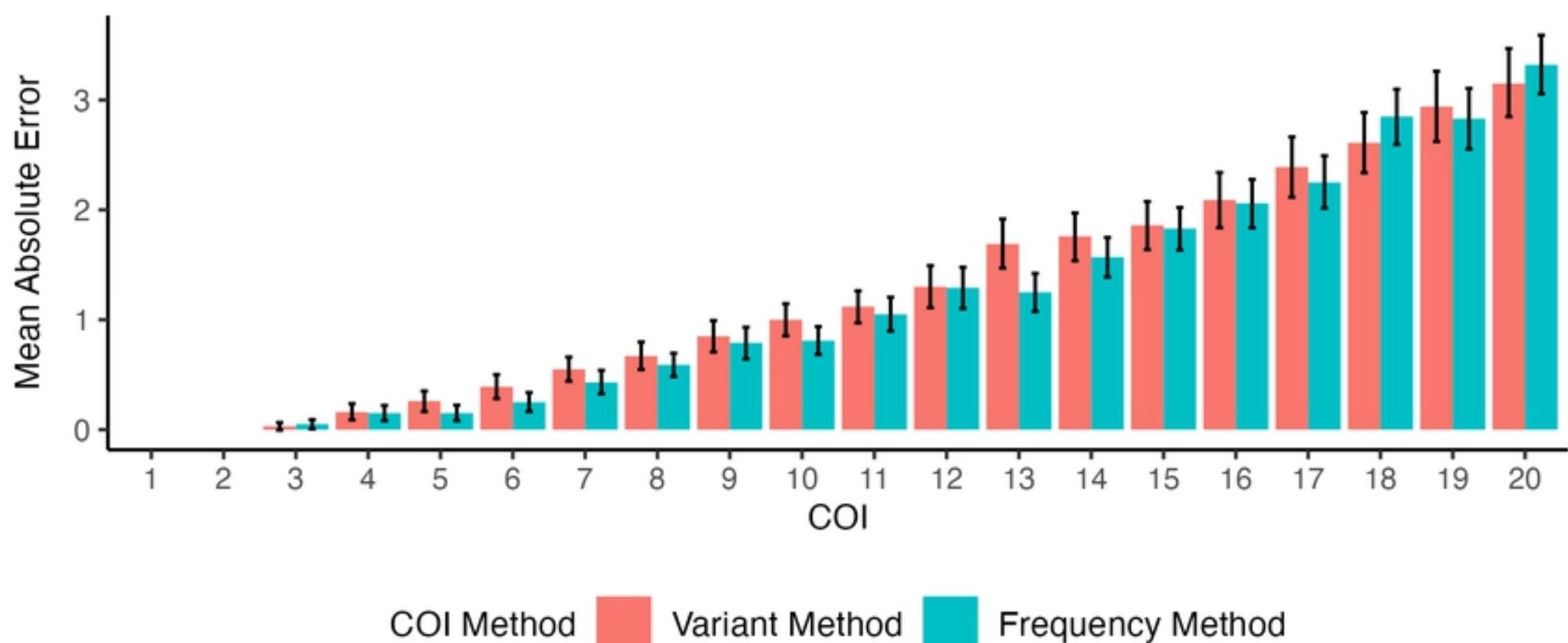


Figure 1

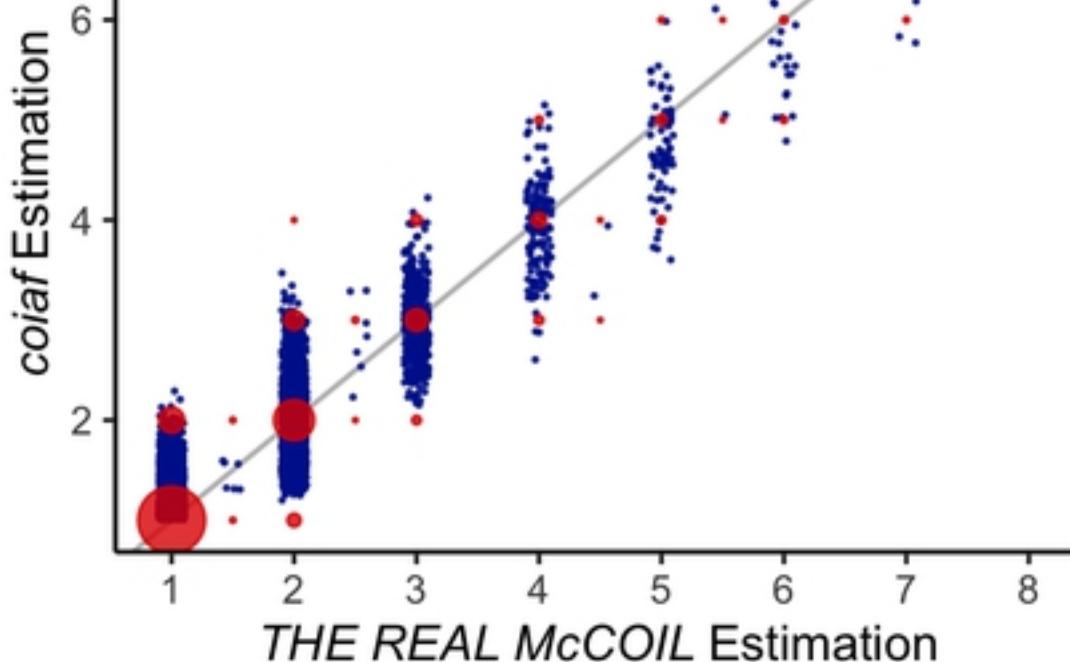
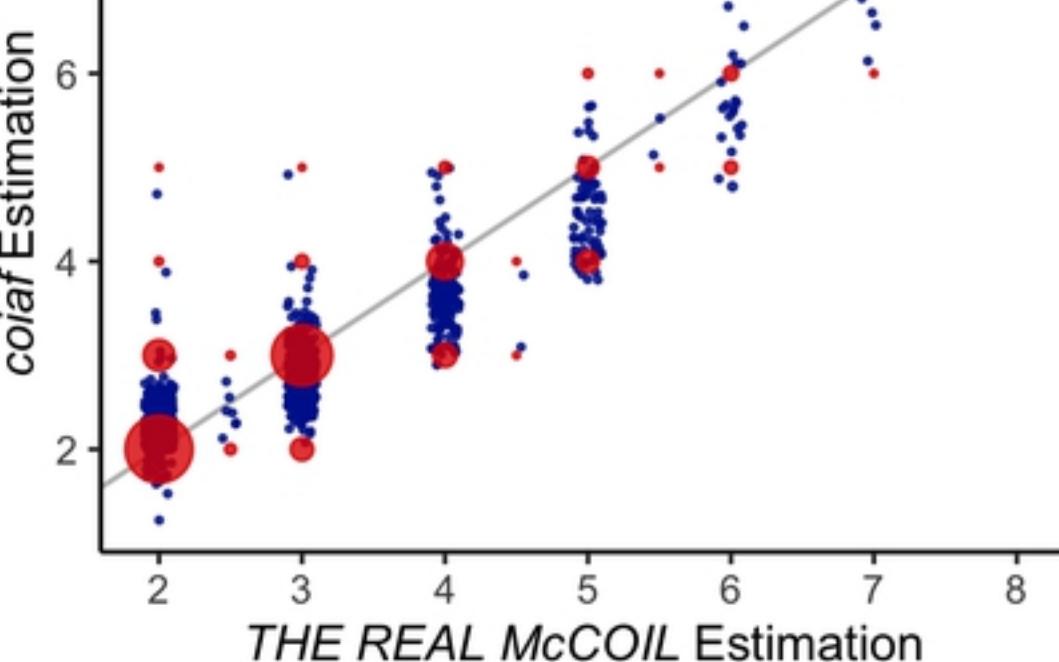
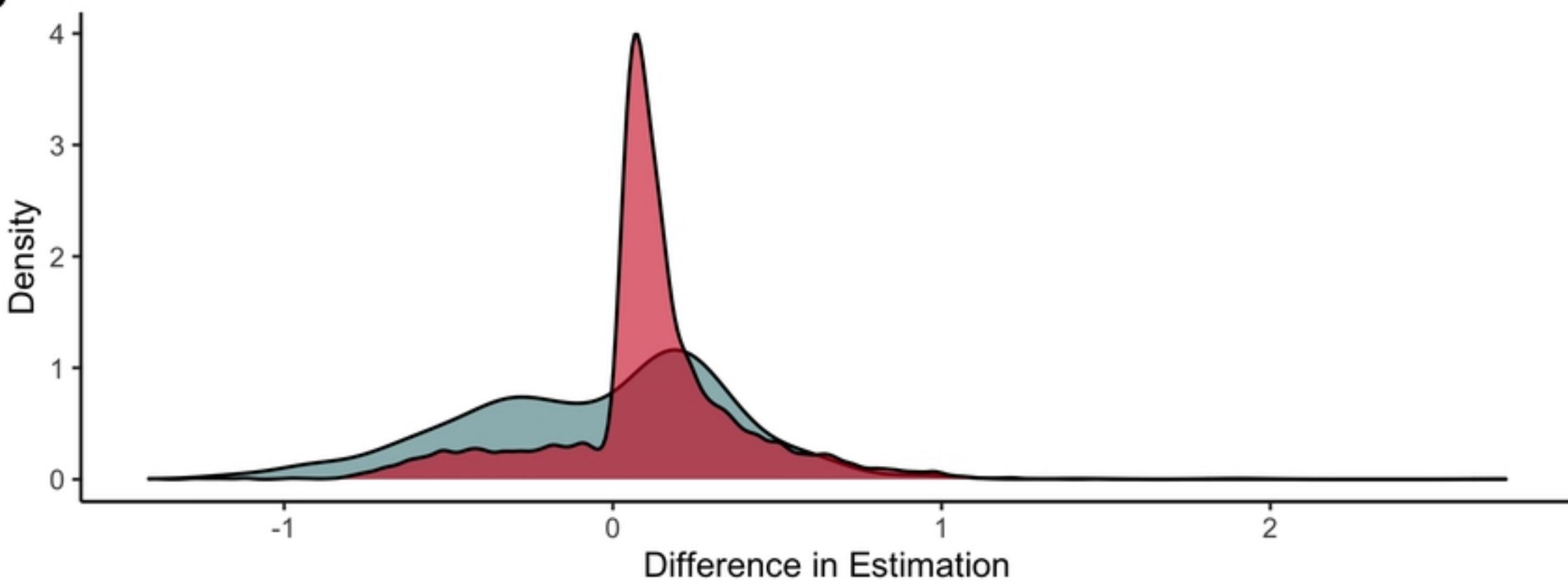
**A****Variant Method**

bioRxiv preprint doi: <https://doi.org/10.1101/2022.05.26.493561>; this version posted May 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

**B****Frequency Method****C****Error****Figure 2**

**A****Variant Method**

bioRxiv preprint doi: <https://doi.org/10.1101/2022.05.26.493561>; this version posted May 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

**B****Frequency Method****C**

Optimization Method

● Continuous

● Discrete

COI Method

Variant

Frequency

**Figure 3**

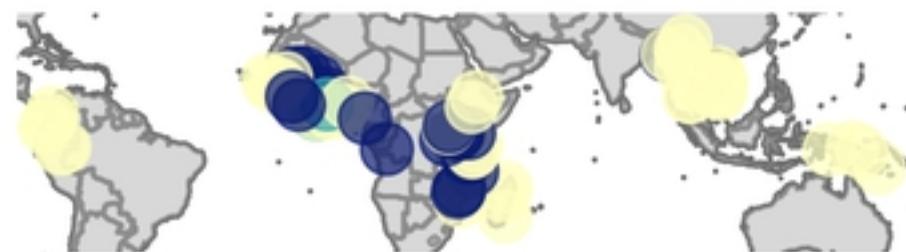
**A**

bioRxiv preprint doi: <https://doi.org/10.1101/2022.05.26.493561>; this version posted May 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



Mean COI

1.0 1.5 2.0 2.5

**B**

Median COI

1.0 1.5 2.0

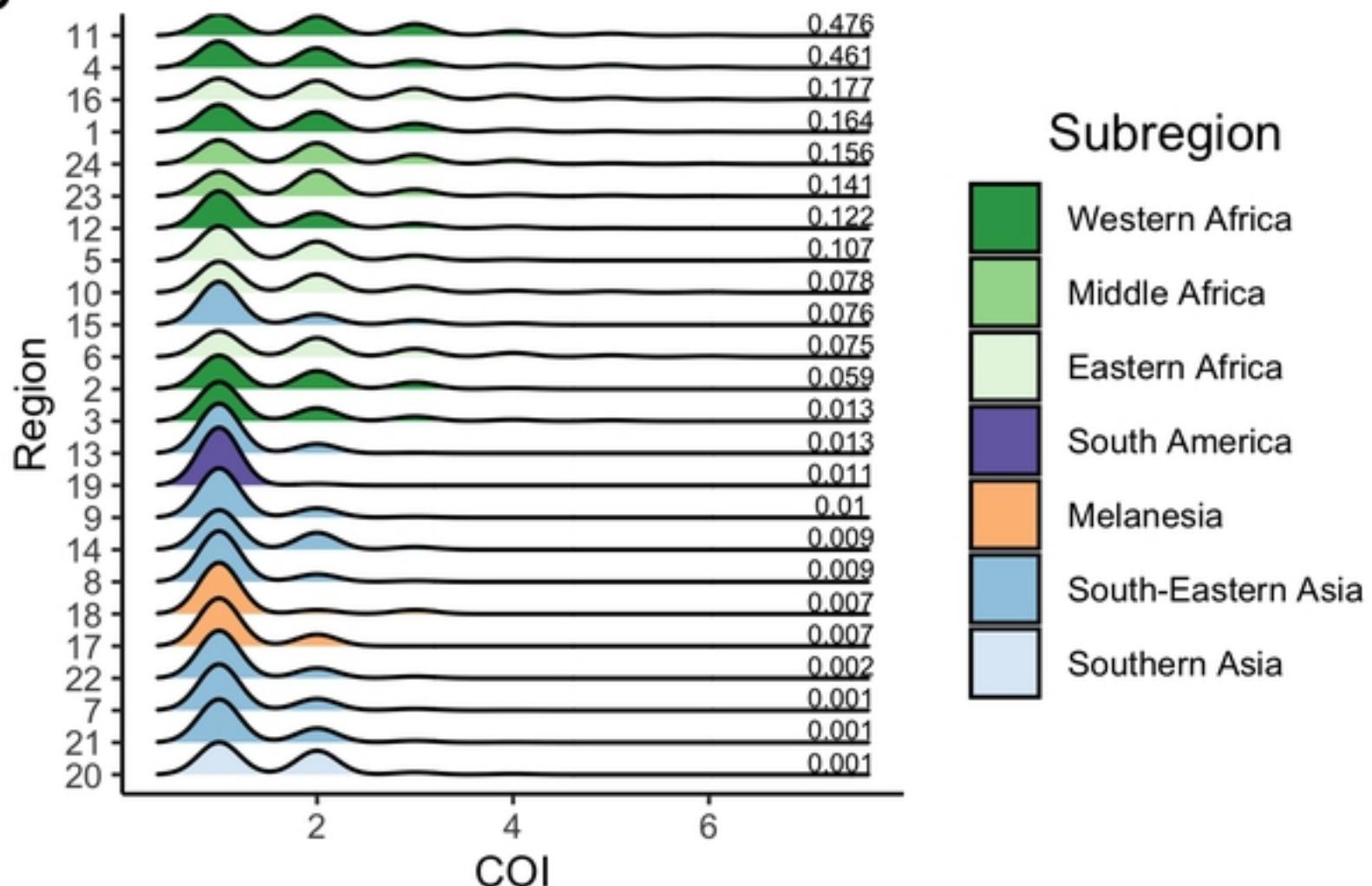
**C**

Figure 4