coiaf: complexity of infection estimation with allele frequencies

Aris Paschalidis[1], Oliver J. Watson[1,2], Robert Verity[2], Jeffrey A. Bailey[1]

[1]Bailey Lab, Pathology and Laboratory Medicine Department, Brown University, RI, USA

[2]MRC Centre for Global Infectious Disease Analysis, Department of Infectious Disease Epidemiology, Imperial College London, London, England

Overview

In this work we present two new methods that use easily calculated measures to directly estimate the complexity of infection (COI) from within sample allele frequencies. We incorporate these methods into a software package: *coiaf*.

Background

- Computational models are increasingly used to help guide malaria control policy and are a key component in understanding the spread of malaria [1].
- The complexity of infection (COI) represents the number of genetically distinct malaria genomes or strains that can be identified in a particular individual and is a strong correlate of the level of transmission [2].
- Best existing approaches to estimate the COI rely on computationally intensive probabilistic likelihood and Bayesian models [3].
- **Aim:** Develop a rapid, direct measure to estimate the COI that works effectively on a set of loci, and at the genome-wide level.

Formulation

- For a biallelic SNP, we 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 (less than 50%) in a population.
- We define the population-level minor allele frequency (PLMAF) as an l-dimensional vector p composed of the frequencies of the minor allele at each locus across a population, namely $p = (p_1, \dots, p_l)$, where $p_i \in [0, 0.5]$.
- We define the within-sample minor allele frequency (WSMAF) as the frequency of the PLMAF at each locus for a single individual infection.
- We denote the COI as k.

Methods

- Given sample allele frequencies D: $\{(p_i, w_i), i = \{1, \dots, l\}$, where p_i is the PLMAF at locus i and w_i is the WSMAF at locus i
- We aim to estimate the COI of a sample.
- We define V_i a R.V. which takes the value of 1 if a site is heterozygous and 0 otherwise.
- **Variant Method**: $\mathbb{P}(V_i = 1) = 1 - p_i^k - (1 - p_i)^k$
- **Frequency Method**: $\mathbb{E}[W_i|V_i = 1] = \frac{p_i - p_i^k}{1 - p_i^k - (1-p_i)^k}$
- These methods define a relationship between the COI and both individual and population level metrics.

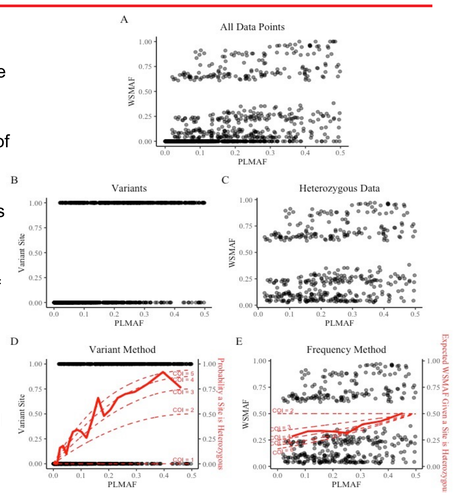


Fig 1. Flowchart of methods. In (A), the relationship between the WSMAF and the PLMAF is shown for an example simulation with a COI of 4. In (B), data have been processed so that loci are deemed variant if they are heterozygous and non-variant otherwise. In (C), homozygous data have been filtered out. Following processing of data, in (D) the theoretical relationships for a COI of 1-5 are shown as dashed red lines. Likewise in (E), theoretical relationships for a COI of 2-6 are shown. The solid red lines in (D) and (E) represent the average of the processed data over each bin of data. The red dashed lines represent our two methods for various values of the COI.

Optimization

- We group our data into N bins based on the PLMAF and compute the midpoint of each bin, $\hat{p}_m$.
- For each bin, we determine the average v_i, denoting the vector of averages as $\hat{t}_{v,m}$and the mean WSMAF for all heterozygous loci, $\hat{t}_{f,m}$.
- Finally, we solve an optimization problem for each method:
- **Variant Method**: $\min_k \left(\sum_{m=1}^{N} |(\hat{t}_{v,m} - f_v(\hat{p}_m)|^q\right)^{1/q}$
- **Frequency Method**: $\min_k \left(\sum_{m=1}^{N} |(\hat{t}_{f,m} - f_f(\hat{p}_m)|^q\right)^{1/q}$
- We note that $q \geq 1, f_v \triangleq \mathbb{P}(V_i = 1)$, and $f_f \triangleq \mathbb{E}[W_i|V_i = 1]$

Results

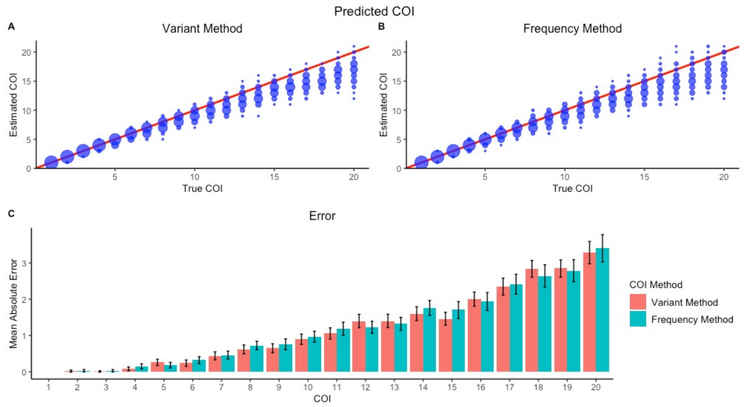
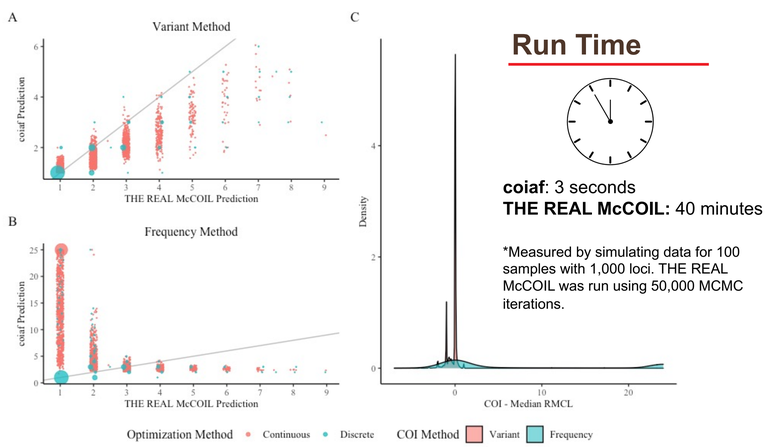


Fig 2. Estimating the COI on simulated data. The performance of the Variant Method (A) and Frequency Method (B) is shown for 100 simulations of a COI of 1-20 with 1,000 loci, a read depth of 100, no error added to the simulations, and no sequencing error assumed. Point size indicates density, with the 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.



Run Time

coiaf: 3 seconds
THE REAL McCOIL: 40 minutes

*Measured by simulating data for 100 samples with 1,000 loci. THE REAL McCOIL was run using 50,000 MCMC iterations.

Fig 3. Comparison Between *THE REAL McCOIL* and Discrete coiaf. The discrete estimation of COI using the a) Variant Method and b) Frequency Method is compared against the *THE REAL McCOIL*. In c) the distribution of differences between each Method and the *THE REAL McCOIL* is shown.

World Map

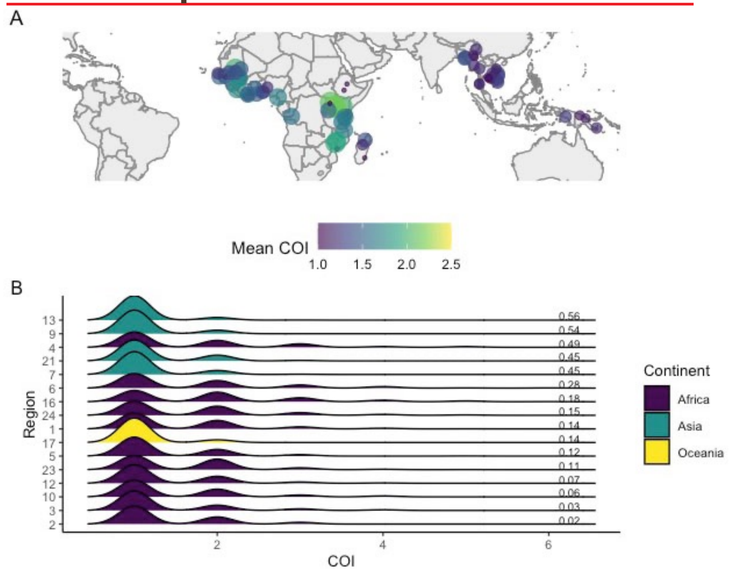


Fig 4. COI Across the Globe. In A we plot the mean COI of all samples in each study location within the 24 regions. The color and size of each point represents the magnitude of the COI. In B, we draw a density plot for each region, sorting the regions by their median prevalence, shown on the right side of the plot. Furthermore, we shade the density plots by the continent data was obtained from.

Conclusions

- 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, which can be used to rapidly estimate the COI of a sample.
- We developed a software package in R, ***coiaf***, which rapidly estimates the COI on a set of loci.
- On simulated data, our methods performed well for low COIs even when the coverage and number of loci was low (1,000 loci), allowing for accurate assessments using targeted and whole genome sequencing data.
- On real data, our methods perform comparably to the current state-of-the-art method, especially for COIs less than 5.
- The biggest advantage of ***coiaf*** is its computational efficiency, ease of use, and potential for further accuracy using read depth and per sample error rates.

References

1. 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. Malar J. 2010;9: 117. doi:10.1186/1475-2875-9-117
2. Daniels RF, Schaffner SF, Wenger EA, Proctor JL, Chang H-H, Wong W, et al. Modeling malaria genomics reveals transmission decline and rebound in Senegal. PNAS. 2015;112: 7067–7072. doi:10.1073/pnas.1505691112
3. Chang H-H, 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: e1005348. doi:10.1371/journal.pcbi.1005348

https://github.com/bailey-lab/coiaf

Research was supported by NIAID R01AI139520.