Benchmarking and Optimization of Methods for the Detection of Identity-By-Descent in High-Recombining *Plasmodium falciparum* Genomes

Bing Guo[1, 2], Shannon Takala-Harrison[1, ☯, *], Timothy D. O'Connor[2, ☯, *]

1 Center for Vaccine Development and Global Health, University of Maryland School of Medicine, Baltimore, MD USA

2 Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD, USA

☯ These authors jointly supervised this work

* Corresponding authors: Shannon Takala-Harrison, `stakala@som.umaryland.edu`, Timothy D. O'Connor, `timothydoconnor@gmail.com`

Abstract

Genomic surveillance is crucial for identifying at-risk populations for targeted malaria control and elimination. Identity-by-descent (IBD) is increasingly being used in *Plasmodium* population genomics to estimate genetic relatedness, effective population size (N_e), population structure, and signals of positive selection. Despite its potential, a thorough evaluation of IBD segment detection tools for species with high recombination rates, such as *P. falciparum*, remains absent. Here, we perform comprehensive benchmarking of IBD callers – probabilistic (`hmmIBD`, `isoRelate`), identity-by-state-based (`hap-IBD`, `phased IBD`) and others (`Refined IBD`) – using population genetic simulations tailored for high recombination, and IBD quality metrics at both the IBD segment level and the IBD-based downstream inference level. Our results demonstrate that low marker density per genetic unit, related to high recombination relative to mutation, significantly compromises the accuracy of detected IBD segments. In genomes with high recombination rates resembling *P. falciparum*, most IBD callers exhibit high false negative rates for shorter IBD segments, which can be partially mitigated through optimization of IBD caller parameters, especially those related to marker density. Notably, IBD detected with optimized parameters allows for more accurate capture of selection signals and population structure; IBD-based N_e inference is very sensitive to IBD detection errors, with IBD called from `hmmIBD` uniquely providing less biased estimates of N_e in this context. Validation with empirical

data from the MalariaGEN *Pf*7 database, representing different transmission settings, corroborates these findings. We conclude that context-specific evaluation and parameter optimization are essential for accurate IBD detection in high-recombining species and recommend `hmmIBD` for quality-sensitive analysis, such as estimation of N_e in these species. Our optimization and high-level benchmarking methods not only improve IBD segment detection in high-recombining genomes but also enhance overall genomic analysis, paving the way for more accurate genomic surveillance and targeted intervention strategies for malaria.

Keywords: Benchmarking, Optimization, Identity-By-Descent, Recombination rate, Marker density, *Plasmodium falciparum*

Introduction

Malaria is a mosquito-borne disease that poses a significant public health challenge globally, with an estimated 249 million clinical cases and 608,000 deaths occurring in 2022 [1]. Surveillance remains an important component of malaria control and elimination efforts. Advances in sequencing technologies and the scale of resequencing studies now allow for parasite genomic surveillance, which can provide insights into the efficacy of malaria interventions and guide the design of targeted elimination strategies in different transmission settings [2–4].

Identity-by-descent (IBD) is an essential tool in population genomics that has been used to estimate genetic relatedness [5–8], positive selection [6, 9–11], effective population size (N_e) [10, 12], fine-scale population structure [4, 10, 13] and migration patterns [4, 14]. However, the reliability of the IBD-based analysis is highly dependent on the accuracy of the detected IBD segments. Insufficient density of genetic markers, on a local or genome-wide scale, probably contributes to high error rates in the identification of IBD segments [11, 15, 16] and reduced accuracy of IBD-based estimates of population demography [5, 17]. Many IBD detection methods have been designed for human genomes, where the demographic history and evolutionary parameters, including the recombination rate, differ considerably from *Plasmodium falciparum* (*Pf*). The effective population size of humans has increased rapidly in recent history [18], while that of *Pf* is decreasing, particularly in regions such as Southeast Asia and South America [1, 19], due to enhanced malaria elimination efforts. More importantly, *Pf* genomes recombine about 70 times more frequently per unit of physical distance [20–24] than the human genome [25], while sharing a similar mutation rate [2,

26–31] as human genomes [32] on the order of 10^{-8} per base pair per generation. The decreasing population size [10] and the high recombination rate in *Pf* result in a reduced number of common variants, such as single nucleotide polymorphisms, per unit of genetic distance. Large human whole genome sequencing data sets typically provide millions of common biallelic SNP variants [33], while *Pf* data sets only have tens of thousands [34]. Given that the human genome is about twice as large as *Pf* in genetic units, the per-centimorgan (cM) SNP density in *Pf* can be two orders of magnitude lower than in humans, which may not provide sufficient information for IBD detection. Thus, it is critical to understand whether IBD detection methods can still generate accurate IBD segments under low SNP density conditions, considering the specific evolutionary parameters of the *Pf* genome.

Evaluating the quality of the detected IBD segments requires benchmarking with the known ground truth through simulation studies [15, 16, 35, 36]. The performance of IBD detection tools developed for use in the human context is typically measured using simulated genomes reflecting demographic and evolutionary parameters of human genomes [15, 16, 35–37]. Given the differences in SNP density between human and *Pf* genomes, the quality assessment and parameter optimization derived from human studies likely do not apply directly to *Pf*. For tools explicitly designed for malaria parasites, such as `isoRelate` and `hmmIBD`, the evaluation of the quality of IBD was based on parent-offspring [8] or pedigree-based simulations (up to 25 generations) [6] that focused primarily on close relatives, which more likely mirrors low malaria transmission settings than high transmission settings. Furthermore, benchmarking methods and definitions of IBD accuracy are inconsistent across studies [6, 8, 15, 16], making the results of the quality evaluation of IBD difficult to compare. Considering the limitations of existing evaluations of IBD detection methods for *Pf* genomes, a unified benchmarking framework specifically designed for high recombining *Pf* genomes from low- and high-transmission settings is needed. Such a framework will assist researchers in comparing and prioritizing different IBD detection methods for intended downstream analysis.

In the present study, we developed a unified IBD benchmarking framework that reflects the demographic and evolutionary parameters of *Pf*. We evaluated how different recombination rates and marker densities affect the quality of detected IBD segments, performed IBD caller-specific parameter optimization, and benchmarked different IBD detection methods with their optimized parameters at both the IBD segment and downstream inference levels. Furthermore, we validated

our findings from simulation analysis with empirical data sets constructed from subsets of samples from the publicly available whole genome sequencing database MalaraiGEN *Pf*7. Our findings indicate that a high recombination rate (given the same mutation rate) is associated with a low SNP density (per genetic unit), which substantially affects the accuracy of the detected IBD segments. IBD segments called by many of the benchmarked methods generally suffer from a high false negative rate (fraction of true IBD not detected) with the default parameters. IBD caller-specific parameter optimization can reduce errors in detected IBD segments and the resulting bias in IBD-based relatedness estimates. With optimized parameters, IBD inferred by most evaluated IBD callers can capture the main patterns of positive selection and population structure; however, only `hmmIBD` provides an accurate estimate of the trajectory of effective population size. To obtain optimal results, we recommend optimizing the parameters of human-oriented IBD callers before applying to *Pf* genomes, and using `hmmIBD` for IBD quality-sensitive analysis when genetic data from haploid genomes are available.

Results

Low SNP density due to high recombination rate affects the accuracy of IBD calls

Ideal IBD segments are typically defined as shared genomic regions descending unbroken by recombination from a common ancestor [38, 39], which can be derived from the ancestral recombination graph (ARG) [10, 17, 35, 40]. When ARG is unknown, IBD segments are generally inferred from mutational information, such as genotype data, as a profile of the underlying ARG [41, 42]. However, high-recombining species like *Pf* have very low marker densities per genetic unit, and may largely affect the accuracy of inferred IBD segments.

To assess how recombination rates affect marker density per genetic unit and the detection of IBD segments, we simulated genomes with varying recombination rates but a fixed mutation rate. Under a single-population demographic model (see Methods), we found that the density of common biallelic SNPs (minor allele frequency ≥ 0.01) per cM, in selectively neutral scenarios, is inversely correlated with recombination rates ranging from 3×10^{-9} to 10^{-6} per base pair per generation (Fig 1a). For instance, the SNP density of *Pf*-like genomes is 25 SNPs per cM, which is approximately

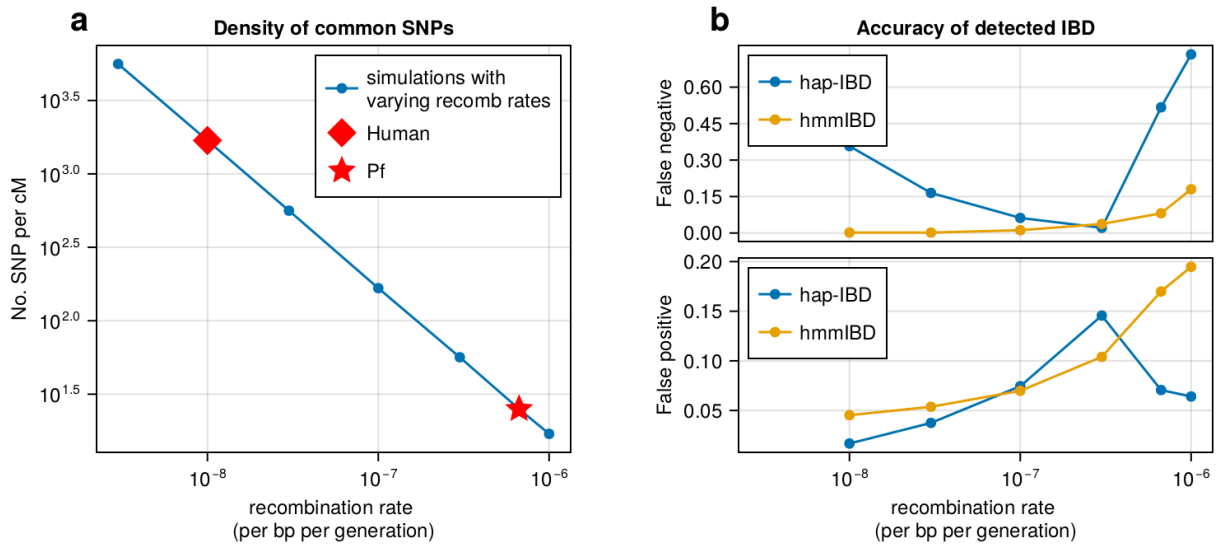

Fig 1. High recombination rates reduce genetic marker density and affect the quality of detected IBD segments. a, The number of common single nucleotide polymorphisms (SNPs) (minor allele frequency $\geq$ 0.01) per genetic unit (centimorgan, cM) in simulated genomes with different recombination rates. In these simulations (blue line), the mutation rates are fixed; the recombination rates vary widely to include the rate for both humans (red diamond) and *Pf* (red star). b, Accuracy of IBD segments detected from genomes simulated with different recombination rates. The accuracy of IBD segments is measured by the false negative rates (top panel) and false positive rates (bottom panel). The plotted error rates reflect the genome-wide rates (defined in Methods) of IBD segments called with default IBD caller parameters unless otherwise specified (see S1) Table. Only error rates for two IBD detection methods, `hmmIBD`, and `hap-IBD`, are included in (b) for simplicity. The error rates for other IBD callers are provided in S2 Fig. For both (a) and (b), the genomes were simulated under the single-population model (see Methods).

1/67 of that of the human-like genomes (1,660 SNPs per cM). We further assessed how low marker densities associated with high recombination rates affect the accuracy of detected IBD segments calculating two metrics (via *ishare/ibdutils*, see Code availability), including the false negative rate (FN), which represents the proportion of a true IBD segment (obtained via `tskibd` [10]) not covered by inferred IBD segments of the same genome pairs, and the false positive rate (FP), which indicates the faction of a true segment not covered by inferred segments (see Methods for detailed definitions, and S1 Fig for method overview). Our analysis showed that as the recombination rate increases, both the genome-wide FN and FP (Fig 1b) increase for IBD inferred from `hmmIBD`. The patterns vary in the other four IBD detection methods, and all suffer elevated FNs and/or FPs as the recombination rate increases (Fig 1b and S2 Fig), except for `isoRelate`, of which the IBD quality tends to be better with a lower marker density. The results suggest that a low SNP density per genetic unit can dramatically affect the reliability of detected IBD segments.

Varying quality of IBD inferred from simulated *Pf* genomes via different IBD callers

Multiple IBD callers have been used for *Pf*, including *Pf*-oriented, Hidden Markov Model-based methods, such as `hmmIBD` [8] and `isoRelate` [6], and those originally designed for human genomes, such as `Refined IBD` [12] and `Beagle` (version 4.1) [4]. We analyzed `hap-IBD` [15] and `phased IBD` [16] IBD callers in addition to `hmmIBD`, `isoRelate` and `Refined IBD`, since the former represents two recent key advancements in the development of IBD detection methods that scale well to large sample sizes and genome sizes.

To evaluate the applicability and accuracy of these IBD detection methods in analyzing *Pf* genomes, we performed benchmarking analyses in simulated genomes (S1 Fig, top panel), mimicking the high recombination rate and the decreasing population size of *Pf* populations. Our analyses include three sets of comparisons: (1) baseline benchmarking, where we mainly used the default parameter values for each IBD caller and compared the performance in *Pf* genomes at the level of IBD segment and their simple statistics; (2) post-optimization benchmarking, where we used parameter values optimized specifically for each IBD caller so that the comparisons are based on the optimal performance of each method; (3) human-like genome benchmarking, where detected IBD segments are expected to have low error rates for human-oriented IBD callers and thus were used as an internal control to validate our benchmarking pipeline (see S1 Table for the IBD caller parameters and Methods for demographic models).

Baseline benchmarking analysis shows that all callers except `hmmIBD` suffer from high FN rates especially for short IBD segments (genome-wide FN/FP rates reflect shorter IBD segments as most segments are short) (Fig 2). Similarly, genetic relatedness metrics based on pairwise total IBD are largely underestimated for most callers (S3 Fig). We found that by default `hmmIBD` has relatively low FN/FP error rates (Fig 2) and is less biased for relatedness estimates (S3 Fig). In contrast, `isoRelate` and human-oriented callers have high false negative rates and varying FP rates (Fig 2). Thus, both *Pf*- and human-oriented IBD callers can suffer high error rates when detecting IBD from genomes with a high recombination rate and a low marker density, with the exent depending on underlying assumptions and methodologies.

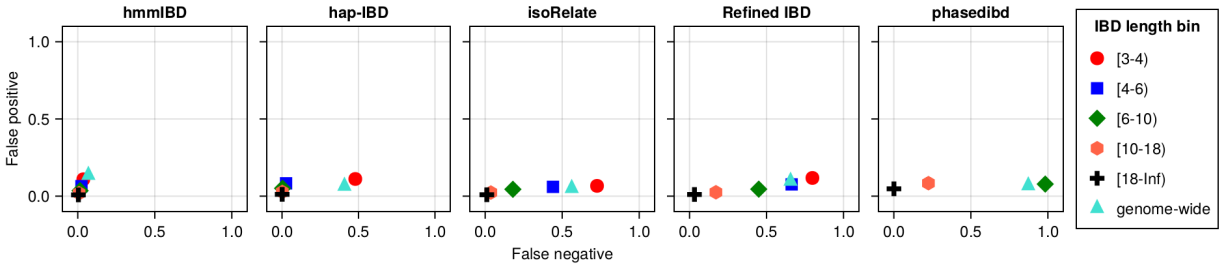

Fig 2. Accuracy of IBD segments detected from *Pf* genomes varies across IBD callers. IBD segments were inferred from genomes simulated under the single-population model with a shrinking population size and a recombination rate compatible with *Pf*. The accuracy of IBD was evaluated using the calculated false positive rate (y axis) and false negative rate (x axis). The rates were calculated for different length bins in centimorgans, including [3-4), [4-6), [6-10), [10-18), [18, inf) centimorgan and at the genome-wide level (defined by overlapping analysis between true IBD segments and inferred IBD segments from each genome pair). The IBD callers analyzed here, from left to right, include `hmmIBD`, `hap-IBD`, `isoRelate`, `Refined IBD`, and `phased IBD`. The results of the simulations under the multiple population model are provided as S1 Data.

IBD-caller-specific parameter optimization for *Pf* improves IBD accuracy

Since these IBD callers are optimized for different species or genotype datasets, we hypothesized that optimization of IBD caller parameters under a unified framework designed for *Pf* genomes can improve the performance of these callers in analyzing malaria parasite data. As searching the entire IBD caller parameter space is inefficient, our optimization focused mainly on parameters potentially affected by or needing adjustment due to differences in marker density between the new target (*Pf*) and previously tested populations (humans). For IBD callers that do not explicitly have marker density-related parameters, such as `hmmIBD`, we explored other parameters that likely affect IBD quality. We performed grid searches to find parameter value(s) that generate inferred IBD with low and balanced error rates (see S2 Table for parameters explored and their corresponding values, and S1 Data for detailed results).

We found that most IBD callers have a key parameter that can dramatically affect their FN/FP rates (S2 Table). For example, the FN rates of IBD called from `hap-IBD` change substantially when the `min-marker` parameter varies (see S1 Data). With a value of 70, the FN rate for short IBD segments dramatically decreases such that the FN and FP error rates become more balanced (Fig 3a). Consistently, genetic relatedness estimates change from being highly underestimated before parameter optimization (Fig 3b, left column) to being more balanced after optimization (Fig 3b, right column). Similar improvements were observed for `Refined IBD` and `phased IBD` (S4

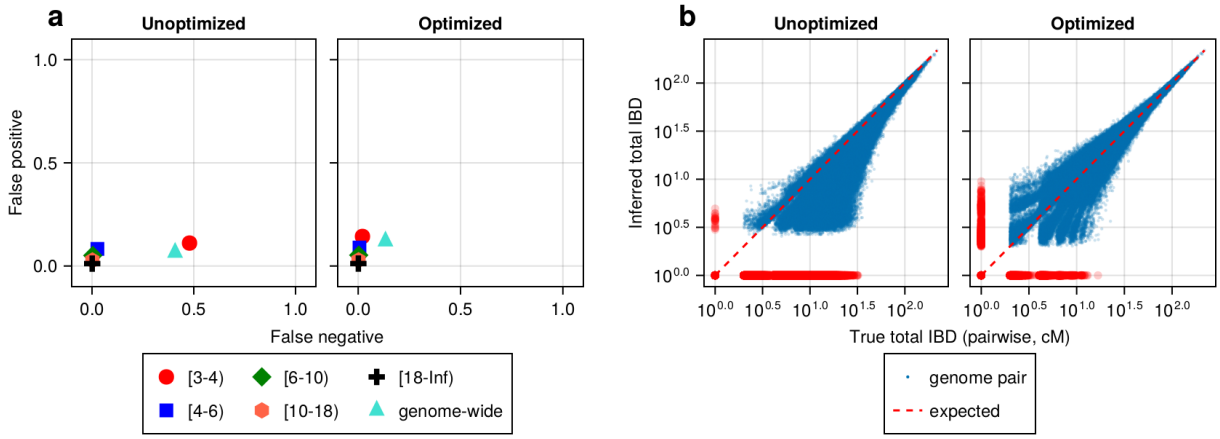

Fig 3. IBD caller-specific parameter optimization can improve the quality of IBD segments inferred from simulated *Pf* genomes (using `hap-IBD` as an example). a, Quality of detected IBD measured by false positive and false negative rates before (left column) and after (right column) `hap-IBD`-specific parameter optimization. As indicated in the axis legend, the error rates were calculated for different length ranges (in centimorgans), including [3-4), [4-6), [6-10), [10-18), [18, inf) and at the genome-wide level. b, Quality of detected IBD measured by total genome pairwise IBD, an estimate of genetic relatedness, before (left column) and after (right column) `hap-IBD` parameter optimization. Each dot represents a pair of genomes with the coordinates x and y being true and inferred total IBD. Note: both the x and y axes in (b) use log scales. In (b), the blue dots are the pairs with nonzero true and inferred total IBD while red dots are pairs with either true total IBD or inferred total IBD being 0; zero-valued total IBD was replaced with 1.0 cM for visualization purposes. The red dotted line of $y = x$ indicates the expected pattern, that is, true total IBD equal to inferred total IBD if the inferred IBD was 100% accurate.

Fig). In contrast, the quality metrics remained largely unchanged during parameter optimization attempts for `hmmIBD` and `isoRelate`, with `hmmIBD` being more accurate and unbiased and `isoRelate` suffering from high false negative rates and an underestimated relatedness (S4 Fig). The observation that human-oriented IBD callers are more optimizable for the simulated *Pf* genome compared to *Pf*-oriented IBD callers confirms that IBD callers originally targeting a non-*Pf* population need adjustment before they can be used for *Pf*.

The human-oriented IBD callers, when not optimized for *Pf*, underperformed `hmmIBD`, especially for `Refined IBD` and `phased IBD` (S4a Fig). To exclude potential problems in our benchmarking pipeline, we simulated genomes with recombination rates and the demographic history consistent with human population in the UK (S5a Fig, left column, and S5b Fig, left column; also see Methods). These callers, evaluated *without* *Pf*-optimized parameters indeed perform much better on human genomes, showing consistently lower FN/FP error rates (S5a Fig, left column) and less biased total IBD-based relatedness estimates (S5b Fig, left column), compared to *Pf*-like genomes (S4a Fig and S4b Fig, left columns). The results support the robustness of our benchmarking pipeline (S5a Fig

and S5b Fig, left columns) and demonstrate challenges in applying human-oriented IBD callers to *Pf*. Furthermore, we found that IBD caller performance varies with demographic configurations (single-population model in S4a Fig and S4b Fig, left columns, versus UK human model in S5a Fig and S5b Fig, right columns) even with the *same* (*Pf*) recombination/mutation rates and default IBD caller parameters, suggesting that the optimization is demography dependent (see details in S1 Data).

Post-optimization benchmarking via downstream inferences

IBD-based downstream analyses, such as estimation of N_e, selection signals, and population structure, are key applications of IBD segments in population genetics, which often rely on the high quality of input IBD segments. With optimized IBD caller-specific parameters tailored for *Pf*-like genomes, we can expect IBD callers to perform at their best, which allows high-level benchmarking by comparing IBD-based downstream estimates.

For IBD-based selection detection, we simulated positive selection on each of the 14 chromosomes via the single-population model, and identified IBD peaks with different callers (see Methods). These peaks, inferred as regions under selection, were true signals if they contain the selected site from simulations. We found that most callers can capture the majority of the simulated signals except `Refined IBD`, which is less sensitive and only detects 3 out of 14 (S6a Fig); `isoRelate` shows an increased level of false positives or low signal-to-noise ratios, evident in IBD coverage curves (S6a Fig).

For IBD-based population structure inference, we simulated *Pf*-like genomes under a selectively neutral condition using the multiple-population demographic model, and performed IBD network community detection via InfoMap [43, 44] to define subpopulations (see Methods). Similarly to IBD-based selection signal detection, we found that IBD inferred from most callers can accurately recapitulate the simulated population structure, comparable to true IBD (S6b Fig). The exception is that `isoRelate` tends to generate many smaller groups, which is likely due to high FN rates for short IBD segments, missing connectivity among distantly related genomes and capturing closely related subgroups of small sizes.

For IBD-based N_e inference, we simulated neutral *Pf* genomes using the single-population model and estiamted N_e from detected IBD via `IBDNe` [17]. We found most of the compared IBD callers

suffer from wild oscillation, which has previously been observed [17], and deviate significantly from the truth for older generations (Fig 4), consistent with the general pattern of high error rates in shorter IBD length bins for these IBD callers (S4a Fig, right column). Meanwhile, the IBD inferred by `hmmIBD` generated highly accurate estimates comparable to true IBD (Fig 4). We explored the mechanisms underlying bias in N_e estimates and found that the strong bias is likely due to underestimation of population-level total IBD for short IBD segments, which is mostly obvious for `hap-IBD`, `isoRelate`, and `phased IBD`, followed by `Refined IBD` (S7 Fig). For `hmmIBD`, both the estimates of N_e (Fig 4) and population total IBD are relatively unbiased (S7 Fig, column 2), consistent with its relatively low and balanced FP/FN rates (Fig 2, leftmost panel). These results suggest that IBD-based N_e estimates are highly sensitive to the quality of input IBD segments, and that `hmmIBD` is more accurate for this analysis.

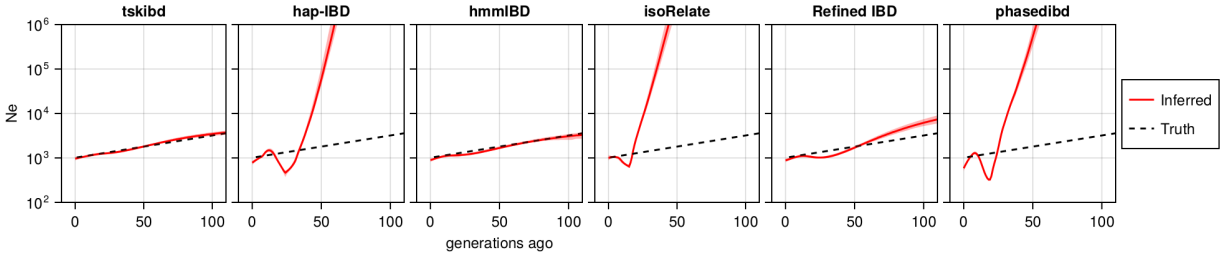

Fig 4. Post-optimization benchmarking of different IBD callers by comparing downstream estimates N_e. With parameters optimized for each IBD caller, the performance of IBD callers was evaluated by comparing the N_e trajectory for the recent 100 generations estimated via `IBDNe` based on true (black dashed line) IBD versus inferred IBD (red solid line). True IBD was calculated from simulated genealogical trees via `tskibd` inferred IBD includes those inferred from `hap-IBD`, `hmmIBD`, `isoRelate`, `Refined IBD`, and, `phased IBD`, with their N_e estimates shown from left to right. The shading areas surrounding the red lines indicate 95% confidence intervals as determined by `IBDNe`. See S10 Fig for pre-optimization results.

Given that N_e estimation in *Pf* is very sensitive to the quality of detected IBD, we explored whether excluding short, error-prone IBD segments ($<$ 4cM) could improve N_e estimates for non-`hmmIBD` callers. The exclusion results in reduced oscillation of the trajectory for some callers, like `hap-IBD`, but wide confidence intervals or underestimation in older generations, in both simulated and emprical data (S8 Fig). We then explored reasons underlying the recent oscillation or drop (around 20 generations ago) commonly observed in the estimated N_e trajectories (S8a Fig, second row and S8b Fig, both rows) [12, 37, 45]. We hypothesized that this oscillation is partially due to IBD segments with TMRCA $<$ 1.5 generations ago being included in `IBDNe` input [17]. We found that removing these segments can greatly mitigate this problem (S8a Fig), especially for `hmmIBD` and

`Refined IBD`. The findings suggest caution when interpreting (1) a recent drop in an estimated N_e trajectory in empirical data sets where TMRCA-based filtering is less practical and (2) extremely large estimates in older generations stemming from high error rates for short IBD segments.

To confirm that parameter optimization improves downstream inferences, we compared post-optimization results (S6 Fig and Fig 4) with pre-optimization estimates (S9 Fig and S10 Fig). We found that parameter optimization indeed increased the power in selection detection (in `isoRelate`, `Refined IBD` and `phased IBD`), improved population structure inference (`phased IBD` and `hap-IBD`), and reduced oscillation on the N_e trajectory (`Refined IBD`).

Validation in empirical data set

We further validated the findings from simulation analysis in emprical datasets, by constructing "single" or "multiple" population data sets based on the MalariaGEN *Pf*7 data [34] (see Methods for details). As true IBD segments are not available here, we focused on high-level benchmarking by evaluating whether IBD-based downstream estimates are consistent with expected patterns, including N_e estimation and selection signal detection with "single" population data sets and InfoMap population structure inference with the "multiple" population data set. (S3 Table, S4 Table, and S5 Table).

With optimized parameters, all callers, except `Refined IBD`, captures most known selection signals from the Southeast Asia data set. These signals include selective sweeps associated with antimalarial drug resistance and sexual commitment, such as dihydrofolate reductase (*dhfr*) [46], multidrug resistance protein 1 (*pfmdr1*) [47], amino acid transporter 1 (*pfaat1*) [23], chloroquine resistance transporter (*pfcrt*) [48], dihydropteroate synthase (*dhps*) [49], Apicomplexan-specific ApiAP2-g(*ap2-g*) [50] and apicoplast ribosomal protein S10 (*arps10*) [51] (Fig 5a). `hmmIBD` detects more peaks but suffer from noise, likely due to the relatively high FP to FN ratios for short (<4 cM) IBD segments (Fig 1 and Fig 2).

Similar to the simulation analysis, IBD detected using most callers resulted in N_e estimates with unrealistic oscillations for the Southeast Asia data set, including extremely large estimates in more distant past (> 20 generations ago) (Fig 5b). The problems are much less severe with the estimates from `hmmIBD`, which mirrored the expected reduction in malaria in this region due to the intense efforts to eliminate malaria in recent decades [1].

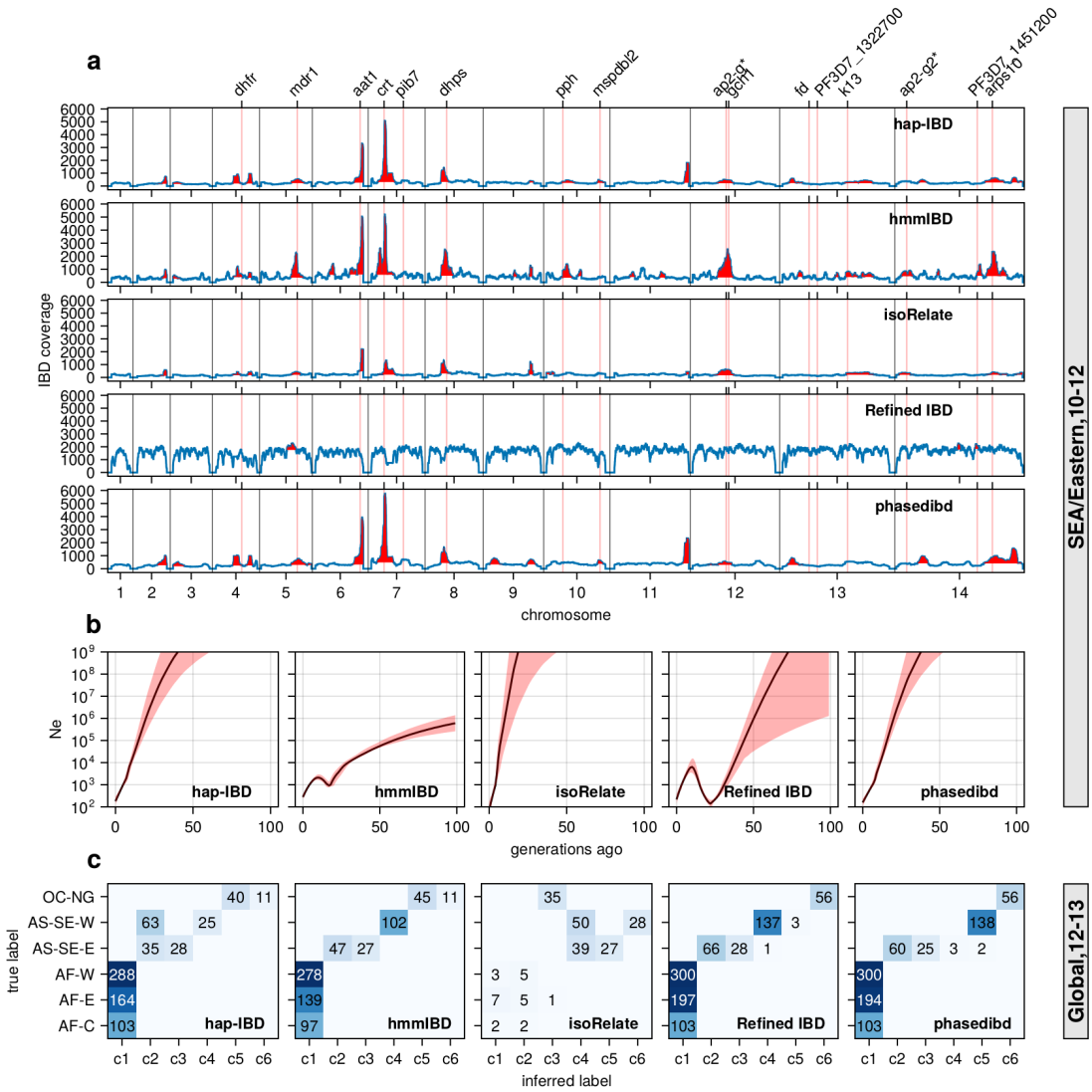

Fig 5. Validation of the performance of IBD callers in empirical data sets by comparing IBD-based downstream analyses. a, IBD coverage and detected selection signals in the SEA data set using different IBD callers (rows 1 to 5). Annotations and corresponding vertical dotted lines at the top indicate the center of known and putative drug resistance genes and genes related to sexual commitment; red shading indicates regions that are inferred to be under positive selection (see Methods for definitions). b, N_e estimates of the SEA data set based on IBD inferred from different callers. Line plots are point estimates; the shading areas around the line plots indicate confidence intervals based on bootstrapping (generated by `IBDNe`). c, Inference of the population structure of the structured data set by the InfoMap community detection algorithm using the IBD inferred from different IBD callers. The rows of the heatmap are geographic regions of isolates, and the columns are the largest, inferred communities, labeled as c1 to c6. The heat map color represents the number of isolates in each block with the given row and column labels. The columns are rearranged so that the diagonal blocks tend to have the largest values per row for better visualization.

InfoMap-based community detection reveals expected continental population structure across most callers: African *Pf* parasites are less structured, and Southeast Asian parasites are more

structured and distinct from Oceanian parasites (Fig 5c), consistent with previous non-IBD-based methods [34]. `isoRelate` IBD estimates generate many small, close groups, likely due to high false negative rates for short IBD segments, especially in high-transmission settings like Africa where parasites have low relatedness and mainly share short IBD segments (Fig 5c).

To further confirm the improvement of IBD-based estimates through parameter optimization, we performed analyses with IBD detected with unoptimized parameters (see S1 Table). We found that the height and number of IBD peaks for `hap-IBD` and `phased IBD` decreased significantly (S11a Fig *versus* Fig 5a). Pre-optimization N_e estimates show more extreme oscillations, especially for human-oriented IBD callers (S11b Fig). Pre-optimization IBD estimates from `hap-IBD` and `phased IBD` fail to reveal the expected population structure, particularly in African parasite populations (S11c Fig). These differences underscore the importance of parameter optimization for *Pf*, especially for IBD callers not validated for *Pf*.

Computation efficiency comparison and improvement

With the decrease in whole genome sequencing cost and the increase in sample availability, it is important to prioritize IBD callers that scale well for large sample sizes, such as MalariaGEN *Pf*7 (n = 20,864) [34]. We compared the IBD inference time and maximum memory usage for different IBD callers with or without parallelization. When using a single thread, probabilistic inference algorithms like `isoRelate`, `Refined IBD`, and `hmmIBD` are about two orders of magnitude slower for those based on identity-by-state-based or positional Burrows-Wheeler transform (PBWT) based algorithms, such as `hap-IBD` and `phased IBD`(Fig 6a). Maximum memory consumption is highest in `Refined IBD`, with `hmmIBD` being around ten times more efficient (S12a Fig). With multithreading, the patterns are similar to single-thread comparison as most allow parallelization (Fig 6b and S12b Fig). The exception is `hmmIBD`, as it currently only supports a single-thread computation. Despite the computation efficiency of `hmmIBD` compared to `isoRelate` and its high accuracy in detected IBD segments from *Pf* genomes, it remains significantly slower than IBS-based methods, highlighting the need for further enhancements for large data sets like MalariaGEN *Pf*7.

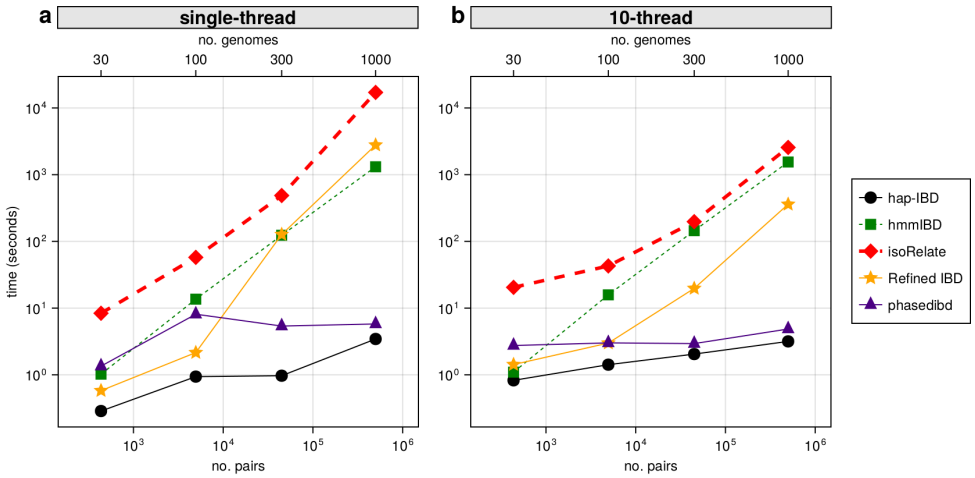

Fig 6. Comparison of computational runtime for IBD calling process for different callers. a, Runtime for different IBD callers to detect IBD from genomes of different sample sizes in single-thread mode. The comparison is based on *Pf* genomes of size of 100 cM simulated under the single population model. The x-axis tick labels include the number of pairs of genomes analyzed (below the plot, on a linear scale) and the sample size (number of haploid genomes, above the plot, on a nonlinear scale) analyzed. The line styles and markers for different callers/tools are provided in the legend box on the far right of the figure, which is shared across the three subplots. b, Runtime in multithreading mode. (b) is organized similarly to (a), except that the IBD calling processes were run in multithreading mode with 10 threads. Also, see S12 Fig for the maximum memory usage for different callers.

Discussion

Our study sought to address important questions regarding the application of identity-by-descent (IBD) in highly recombining species, such as *Plasmodium falciparum*, focusing on the reliability of IBD segment detection and the prioritization of detection methods for accurate IBD-based analyses. We observed that: (1) Low marker density per genetic distance, resulting from high recombination rates relative to mutation rates, significantly affects the accuracy of IBD detection; (2) Optimizing IBD caller-specific parameters can improve the performance of IBD callers for *Pf* genomes, especially for callers designed for human genomes; (3) After parameter optimization, most IBD callers effectively capture expected positive selection signals and population structure patterns; however, only `hmmIBD` can provide plausible N_e estimates for simulated genomes mimicking the *Pf* population; (4) Probabilistic (HMM-based) IBD detection methods are significantly less computationally efficient than IBS-based methods; `hmmIBD` will need enhancement through parallelization and memory optimization, to enable scalability to larger data sets. Consequently, we recommend optimizing human-oriented IBD callers before their application in malaria parasite research and endorse the

use of `hmmIBD` for quality-sensitive IBD-based inferences, for example, the estimation of N_e.

Comparing the performance of multiple IBD detection methods often requires a unified framework, which should include a uniform definition of accuracy and a simulated ground truth that mimics *Pf* genomes. Our benchmarking framework incorporates several novel features. First, it utilizes a consistent definition of IBD length-specific accuracy based on the overlap of IBD segment lengths, closely aligned with metrics used in `hap-IBD` [15], `phased IBD` [16], and `Refined IBD` [52], as detailed in our Methods section. The approach differs from the original evaluation of `hmmIBD`, which assesses the accuracy of IBD based on the fraction of SNPs that share IBD, which could overlook the precise accuracy of IBD length of the detected segments [8]. In particular, the original study of `isoRelate` defined the accuracy (true positive rate) using a less stringent overlap-by-count criterion where a segment is counted as accurate when at least 50% of a detected IBD segment is overlapped by true segments [6]. Second, we generated *Pf*-like genomes via population genetic simulation that reflect a realistic distribution of IBD segment lengths. This contrasts with previous studies, where methods such as `hap-IBD` and `Refined IBD` used human-like data from population genetic simulations [15, 52], whereas `hmmIBD`, `isoRelate`, and phased-ibd relied on simulations based on artificial recombination or pedigree models [6, 8]. These models (no-population-based genetic simulation) often produce long shared IBD segments typical of close relatives, failing to capture the IBD length distribution in population samples predominantly comprising distant relatives. Third, our benchmarking extends beyond the segment-level evaluation of IBD callers and includes downstream inferences of population structure and effective population size, providing a more thorough assessment of their application in real-world analyses. This comprehensive approach seeks to complement and broaden the scope of quality evaluations that are predominantly focused on the IBD segment level in original studies, thus improving our understanding of the performance of IBD detection methods in the context of *Pf* genomes.

The density of markers per genetic unit, as a measure of information enrichment for coancestry inference, plays a crucial role in determining whether and how a candidate IBD segment is accepted in the final IBD calls. IBS-based methods, such as `hap-IBD` and `phased IBD`, first identify long IBS segments ($\geq$ 2 centimorgan) as candidate IBD segments, and subsequently merge short ones separated with small gaps [15], allow a certain number of discordant markers to account for phasing errors [16], or eliminate false positive segments by removing candidate segments supported by only

a small number of markers [15]. Similarly, `Refined IBD`, which combines an IBS-based method with an HMM probabilistic model, uses a LOD score to decide whether a candidate IBD segment should be rejected or accepted [52]. In these studies, default values for threshold parameters related to marker number/density were shown to be effective for genomes like humans. Although the use of different parameter values has been considered for these algorithms to account for marker density differences between sequencing data versus microarray genotype data [15], per-genetic-unit marker density in human data, even for microarray genotype data, is still significantly higher than in *Pf* sequencing data. We evaluated how different levels of per-genetic-unit marker density affect the detection of IBD segments by varying recombination rates in simulations. With simulated *Pf*-like genomes of low marker density and high recombination rate, we found that human-oriented IBD callers suffer high false negative rates. One potential explanation is that the thresholds optimized for human data are too stringent for *Pf* causing excessive rejection of candidate segments. The effect of marker-density on IBD detection is further confirmed by our findings that adjusting the values of marker-density-related parameters using a grid-search approach could significantly reduce IBD error rates and generate more accurate IBD-based downstream estimates. Even though IBD accuracy can be improved by parameter optimization, we found that error rates of the detected IBD segments are still higher in *Pf* genomes than those of human genomes even after IBD caller parameters are optimized. There are several possible explanations for the high error rates of IBD segments detected from low marker density data: (1) Detected IBD segments can only start and end at the genotype marker site, bypassing any non-genotyped sites; (2) A lower marker density is linked with greater uncertainty in the distribution of IBD endpoints [11]; (3) Ancestral relationships including IBD are too difficult to be reliability inferred given limited mutational information [41, 42, 53, 54].

In our benchmark analysis, we used only common biallelic SNPs as markers for inferring IBD, excluding rare variants and indels. The use of this additional information can potentially provide denser genotype information, thus enhancing our understanding of the population's ancestral relationships, a key aspect on which the inference of the IBD segment depends. For instance, large-scale whole genome sequencing studies reveal that rare variants account for the majority of all segregating sites [33, 55], which contain crucial information for deciphering recent evolutionary history. However, rare variants are typically not utilized for two main reasons: (1) rare genotypes are very

sparsely distributed across many sites and are less informative per site [8, 11]; (2) rare genotype calls are more prone to genotyping or phasing errors. As a result, including rare variation cause reduced accuracy in detected IBD segments due to genotyping/phasing errors and may increase IBD inference time due to data sparsity. Our ongoing work aims to improve the computational efficiency of the `hmmIBD` algorithm by parallelization and develop a tabular encoding strategy for efficient rare allele-sharing analyses have the potential to mitigate the computational burden of rare variants. Although our simulation analysis indicates that including rare variants is of little effect or detrimental for IBD detetion (S1 Data), the findings could be skewed due to the absence of genotyping errors, the small sample size simulated, and limited number of cut-off values tested. Further investigation in contexts of large sample sizes and varying levels of genotype errors is needed to inform the usability of rare variants for different IBD detection algorithms. Indels are another significant source of underutilized genetic variations in the *Pf* genome, with abundance on par with that of biallelic SNPs in the MalariaGEN *Pf*7 data set [34]. These indels are, in part, the result of microplasticity related to the high AT content (up to 90% in non-coding regions) in *Pf* genomes [28]. Using these variants could also increase the marker density to infer IBD segments. However, additional research is necessary to determine whether the inclusion of these variants can reduce uncertainty in the inference of IBD for *Pf* or introduce more bias due to challenges such as sequence-read mapping.

Although a substantial portion of this study concentrated on *Pf*, the main findings and methodologies may be relevant to high-recombining species beyond *Pf*. For instance, in regions with intermediate and low malaria transmission, the incidence of *Pf* has markedly decreased, allowing other species, such as *Plasmodium vivax*, to become predominant [1]. Clinical malaria caused by simian *Plasmodium* species, e.g., *Plasmodium knowlesi*, have also increased in some geographic areas where human *Plasmodium* species have declined [56]. We expect that the performance of IBD callers will be similar in other *Plasmodium* species, given their likely comparable high recombination rates [57, 58]; however, generalization would need further exploration as part of future work, considering variations in evolutionary histories and parasite biology [59–61]. For other species that exhibit high recombination, but have markedly distinct evolutionary parameters from *Plasmodium*, the benchmarking framework established in this study can easily be tailored to prioritize and evaluate IBD detection methods in a context-specific manner.

While we have conducted numerous simulation analyses complemented by carefully designed validation studies, our work is subject to a few caveats: (1) Our simulations did not explicitly incorporate inbreeding within the complicated life cycle of the parasite [62], except for the increased inbreeding potential due to the reduced population size in the single population model. Inbreeding can be pervasive, especially in low transmission settings, leading to a change in the length distribution of IBD toward longer segments and a potentially reduced marker density. (2) Our optimization is based on simple accuracy metrics and only focuses on a subset of parameters, to allow faster iteration over different values. Investigating a larger parameter space, including genotyping error rate and higher-level accuracy metrics may generate different optimal values and further improve IBD-based downstream estimates. (3) We assume a constant recombination rate and mutation rate as static genomic/population parameters, rather than traits capable of evolving over time. If this assumption proves to be inaccurate, such as with recombination rates that vary between individuals and populations [63], a more complex benchmarking framework will be required.

In this study, we evaluated the performance of existing IBD segment detection methods in analyzing genomic data of malaria parasite *Pf*, which is characterized with high recombination rates and low marker densities. Our findings underscore that a high recombination rate, relative to the mutation rate, can compromise the accuracy of detected IBD segments when using methods originally calibrated for the human genome, characterized by a significantly lower recombination rate and a higher marker density (per genetic unit). The accuracy of IBD detection can be improved by parameter optimization via grid search techniques. We advocate for a context-specific evaluation of IBD detection methods when applying them to untested species. Specifically for *Pf*, our research indicates that hidden Markov model-based probabilistic methods, such as `hmmIBD` produce less biased IBD estimates leading to more accurate downstream inferences. This is especially important for analyses that heavily rely on the accuracy of detected IBD segments, such as N_e inference.

Methods

Simulation overview

We used population genetic simulations to allow the generation of (1) ground truth, including true IBD segments, true sites under positive selection, true trajectory of population size, and true sub-

population assignments (population structure), and (2) inferred patterns, including IBD inferred from phased genotype data via different IBD callers and IBD-based downstream inferences of N_e, positive selection, and population structure. By comparing inferred patterns with ground truth, we calculated metrics at the IBD segment level and the IBD-based downstream estimate level (high level) for benchmarking and optimizing various IBD detection methods for *Pf* genomes. As described in our accompanying work [10], we combined the flexible forward simulator `SLiM` [64, 65] and the efficient coalescent simulator `msprime` [66] to simulate genomes similar to *Pf* reflecting the high recombination rate, strong positive selection, and population size shrinkage due to malaria reduction (S1 Fig). A detailed implementation of the simulations can be found in a dedicated GitHub repository (`https://github.com/bguo068/bmibdcaller_simulations`).

In these simulations, we assumed constant recombination rates over the genome, such as 6.67×10^{-7} per base pair per generation for *Pf* [23, 67] and 1.0×10^{-8} for human [25], and a mutation rate of 1.0×10^{-8} for both *Pf* [26, 27] and human unless otherwise specified. Parameters for modeling population size changes, population structure, and positive selection are detailed in the following section or our related publication [10].

Simulated demographic models

We used three different demographic models in the simulations, including the single-population model, the multiple-population model, and the UK European human population models [15].

Single- and multiple-population models have been described in our accompanying work [10]. The single population model mimics malaria reduction in settings like Southeast Asia, with a population size decreasing from 10,000 to 1,000 over the last 200 generations. This model was used to benchmark IBD detection methods at the IBD segment level and at the (high) level of downstream estimation, including selection signal detection and N_e estimation. The multiple-population model was mainly used to benchmark IBD calling methods via IBD network-based community detection. Implementation of the two models were provided a GitHub repository (see Code availability).

The UK human demographic model, similar to the one used in [15], simulates a population bottleneck event from a constant size of 10,000 to 3,000 that occurred 5,000 generations ago, followed by growth at rate of 1.4% and 25% per generation beginning 300 and 10 generations ago, respectively. We simulated 14 chromosomes with a size of 60 cM each for a smaller genome size to reduce

simulation time. This demographic model serves as a control to detect IBD segments with human-oriented callers, which can help validate our IBD accuracy evaluation pipeline. We also used this model to test whether demographic models impact the performance of IBD callers by replacing the human recombination rate with that of *Pf*.

To separate the effects of positive selection from that of demographic models and recombination rates, we mostly simulated neutral genomes by setting the selection coefficient s to 0 for the above models, except in the case where we need to benchmark IBD callers for detecting positive selection signals.

Positive selection simulation

To evaluate the performance of IBD callers via IBD-based selection signal detection, we simulated positive selection within the single-population model with selection coefficients s to 0.2 with a single origin starting 80 generations ago. Fourteen chromosomes with a size of 100 cM were simulated independently, each with a selected site at 33.3 centimorgan from their left ends. For selection simulation, we conditioned on the establishment of the selective sweeps, that is, the allele under selection should not be lost at the present-day generation. If lost, the simulation was rerun for a maximum of 100 times until the selective sweep is established.

IBD calling and default parameters

We generated true IBD from simulated genealogical trees in the tree sequences using the `tskibd` algorithm [10]. Briefly, we sample local/marginal trees along a chromosome and tracked changes in the most recent common ancestor (MRCA) for each pair of sample nodes. If MRCA changes, the shared ancestral segment breaks. We report a shared ancestral segment as an IBD segment if it length exceeds a threshold, such as 2 centimorgans.

For inferred IBD, we used phased genotype data as input. Only biallelic sites with a minor allele frequency no less than 0.01 were included, unless otherwise stated. When needed, a genetic map is generated based on the constant recombination rate as specified in the corresponding simulations. For IBD detection methods designed for diploids, we converted each haploid to a pseudohomozygous diploid. The resulting IBD segments of each pair of pseudo-homozygous diploids (A1/A2 and B1/B2) will have redundant information due to the 100% runs of homozygosity. We only keep one

pair A1-B1 and remove other combinations.

As each IBD detection method provides multiple tunable parameters, we detailed values used in S1 Table for both default and optimized scenarios. For the default scenarios, the parameters mostly follow the original documentation. Exceptions are parameters that need consistency across different IBD callers for benchmarking, including the minimum IBD length and the minimum minor allele frequency. The process of obtaining optimized parameter values is described below.

Benchmarking metrics at the IBD segment level

Several metrics were calculated to benchmark IBD methods at the IBD segment level or using their simple aggregates, including the false negative rate and false positive rate, pairwise total IBD and population-level total IBD per length bin.

False positive rate and the false negative rate were obtained following the definition used in the work of Zhou *et al.* [15]. Rates were first calculated per segment via segment-overlapping analysis; then, they were averaged over all segments of the same length bin. The false positive rate per segment is defined as the proportion of a given IBD segment of some genome pair from the inferred set (for example, IBD called via `hmmIBD`) that is not covered by any IBD segment from the truth set (generated by `tskibd`), for the same genome pair. Similarly, the false negative rate per segment is defined as the proportion of a given IBD segment of some genome pair from the truth set that is not covered by any IBD segment from the inferred set. The average false positive rate is calculated as the average per-segment false positive rates for all inferred IBD segments the length of which falls in a certain range (length bin); the average false negative rate is calculated as the average per-segment false positive rates for all true segments of a length bin. The following length bins were used: [3-4), [4-6), [6-10), [10-18), [18, inf) cM, similar to Zhou *et al.*'s method.

Genome-wide FP/FN rates per pair and their averages across all genome pairs were calculated to capture genome-wide bias. The per-pair genome-wide false positive rate is the ratio of two sums: (1) the numerator sum is the total length of parts of all inferred IBD segments of a certain genome pair that are not covered by any true IBD segments of the same genome pair; (2) the denominator sum is the total length of all inferred IBD segments of a certain genome pair. The per-pair genome-wide false negative rate is defined in a similar way as the percentage of pairwise total true IBD that are not overlapped by any inferred segments of the same pair. We then obtain the aggregate

metrics by averaging these rates for all genome pairs.

Pairwise total IBD from truth *versus* inferred set was calculated as it's a useful metric to estimate genetic relatedness and build IBD sharing networks. It was calculated as the sum of the lengths of all inferred or true IBD segments of each genome pair.

Given that the N_e estimator `IBDNe` internally utilizes quantities of population-level total IBD of different length bins, these quantities were calculated here to better examine IBD accuracy for N_e inference. We defined non-overlapping length bins of 0.05 centimorgan width that cover all possible lengths. For each length bin, a poplation total IBD was defined as a sum of length of all IBD segments with segment lengths falling into this bin from any genome pair.

To expedite IBD segment-level analysis and alleviate computational burdens, we developed an open-source tool, *ishare/ibdutils* (available at `https://github.com/bguo068/ishare`), which harnesses algorithms such as interval trees to efficiently calculate metrics like those described above.

IBD caller parameter optimization

We optimized key IBD caller-specific parameters by iterating each parameter over the list of discrete values, or two or more parameters over a grid of discrete values. Many optimized parameters were related to marker density for IBD callers `hap-IBD`, `phased IBD`, `Refined IBD`, and `isoRelate`. Other parameters were searched, if they potentially have a great impact on the quality of the detected IBD. The optimal values for explored parameters are determined by the length-bin-specific or genome-wide error rates (FN and FP) for detected IBD segments as defined above. The parameter value or combination of values that generate lower and generally balanced error rates were selected as optimal values. The parameters searched, the value lists explored and the optimal values selected are summarized in S1 Table and S2 Table. When the optimized values vary across different demographic models, we used the ones optimized for the single-population model for downstream analyses. We provided detailed heatmaps of error rates for all demographic models tested (see S1 Data).

Benchmarking via IBD-based downstream analysis

At a higher level, we benchmarked different IBD callers by comparing downstream estimates based on true IBD sets *versus* inferred IBD sets. These IBD-based estimates include positive selection scan, N_e estimates, and population structure inference via the community detection algorithm InfoMap.

We scanned positive selection signals using the IBD-based thresholding method followed by validation with integrated haplotype score-based statistics X_{iHS} as previously described [10].

We inferred the trajectory of effective population size, N_e, for the last 100 generations using `IBDNe`. As this method uses IBD shared by diploid individuals as input, we converted each haploid genome to pseudo-homozygous diploid individuals. We infer the trajectory N_e using most of the default parameters except for setting the minregion parameter to 10 cM to allow the inclusion of short contigs in the analysis. The final estimates are scaled by 0.25 to compensate for the haploid-to-diploid conversion. By default, we used the value of 2 cM for the mincm parameter to only include IBD segments $\geq$ 2 cM for inferring N_e trajectory; as indicated in the Results section, we also set mincm to 4 to test whether excluding short IBD segments can improve the accuracy of N_e estimates. As the `IBDNe` algorithm does not work well when IBD segments shared by close relatives are included in the input, we followed the procedure described in the original work by Browning *et al.* [17]. For simulated data, we utilized the TRMCA information of the true IBD segments to filter the IBD segments before calling `IBDNe`. For true IBD segments (generated by `tskibd`), we excluded IBD segments with TMRCA < 1.5 ; for inferred IBD segments (called by `hap-IBD`, `hmmIBD`, `isoRelate`, `Refined IBD`, `phased IBD`), we removed (inferred) segments that overlap with any true IBD segment with TMRCA < 1.5 shared by the same genome pair. For empirical data where true IBD and TMRCA are not available, we pruned highly related isolates by iteratively removing the genome that has the highest number of close relatives defined by pairwise total IBD > 0.5 of genome size until no close relatives are present in the remaining subgroup [10].

For population structure inference, we first built the pairwise total IBD matrix, each element being the total IBD for a pair of genomes. The matrix was then squared and used as a weighted adjacency matrix to construct an IBD-sharing network. We then ran the InfoMap algorithm to infer community membership. Genomes assigned the same membership were inferred to be of the same subpopulation. For empirical data, we excluded IBD segments shorter than 4 cM when calculating the total IBD matrix to help reduce noise due to false positives and set each element with a value < 5 to zero in the unsquared IBD matrix to decrease the density of the IBD matrix.

Before all high-level benchmarking analyses, we pruned highly related samples as mentioned above. As IBD-based estimates can be biased by strong positive selection in empirical data, we removed high IBD sharing peaks with peak impact index > 0.01 as previously described [10] before

any downstream analyses.

Processing empirical data sets

We constructed empirical data sets for validation using genotype data from whole genome sequencing samples from the MalariaGEN *Pf*7 database [34]. We used *malariagen_data* 7.13 to download high-quality monoclonal samples that pass quality control. Monoclonal samples were determined by $F_{\text{ws}} > 0.95$ (F_{ws} table available from the malariaGEN website). Quality control labels were extracted directly from the metadata (provided in the *malariagen_data* package).

We then generated haploid genomes (phased genotype data) using the dominant allele from genotype calls. The dominant allele of each genotype call was determined by the per-sample allele depth (AD) fields. For each sample and site, the allele supported by 90% of total reads (total AD values) in a genotype call with at least 5 total reads was used as the dominant allele. Genotype calls without dominant alleles were marked as missing; those with dominant alleles were replaced with a phased genotype homozygous for the dominant allele. The genotype data were further filtered by sample missingness and SNP minor allele frequency and missingness. The resulting genotype data had per-SNP and per-sample missingness < 0.1 and minor allele frequency ≥ 0.01. Genotypes based on dominant alleles were further imputed without a panel using Beagle 5.1 [68]. These processing steps generated phased, imputed, pseudo-homozygous diploid genotype data, ready for IBD detection.

We constructed different data sets, including, two "single" population data sets and a "structured, multiple" population data set, by subsampling the above haploid genomes according to sampling time and location. For each data set, we set the time window to 2-3 years to reduce the sample time heterogeneity, and then shifted the window within all possible sampling years and chose one that maximized the sample size. For the "single" population data sets, we further restricted the sample locations to a relatively small geographic region, such as eastern Southeast Asia, as the data set was used for N_e estimation which assumes a homogeneous population. For the "multiple" population data set, we included samples from different continental or subcontinental regions using "Population" labels from the meta-information table provided with the MalariaGEN *Pf*7 database. To make the sample size of each "population" more balanced, we set a maximum number of samples of 300. Populations with samples larger than 300 are subsampled to a size of 300; populations with

a size smaller than 100 were not included in the multiple population data set. The details of the sampling location and time information were summarized in Supplementary Tables 3-5.

Details about the preparation and analysis of the empirical data sets can be found at (`https://github.com/bguo068/bmibdcaller_empirical`).

Measuring computational runtime and memory usage

The genomes were simulated with $N_0 = 1{,}000$ and s = 0.0 under the single population model. The runtime and maximum memory usage were measured using the GNU time 1.7 utility. To allow a more appropriate comparison, we ensured: (1) the time used for data pre- and post-processing was excluded; (2) memory resource allocation was capped at 30 gigabytes per IBD call; (3) input genotype data included only common SNPs with minor allele frequency ≥ 0.01; (4) the minimum reported IBD segment length was set to 2.0 cM.

Code availability

Custom tools or scripts were provided in the following GitHub repositories: (1) bmibdcaller_simulations: a Nextflow pipeline to benchmark different IBD detection methods and optimize IBD caller-specific parameters by simulating *Plasmodium falciparum*-like genomes and using true IBD (`https://github.com/bguo068/bmibdcaller_simulations`). (2) bmibdcaller_empirical: a Nextflow pipeline to benchmark IBD callers with empirical data by comparing IBD-based estimates with expected patterns (`https://github.com/bguo068/bmibdcaller_empirical`). (3) ishare/ibdutils: a Rust crate and command line tools designed to facilitate the analysis of rare-variant sharing and identity-by-descent (IBD) sharing (used here mainly for fast IBD segment overlapping analysis) (`https://github.com/bguo068/ishare`, command line tool `ibdutils`).

Data availability

All empirical data used were publicly available from MalariaGEN *Pf*7 [34].

Acknowledgement

This publication uses MalariaGEN data as described in "Pf7: an open dataset of Plasmodium falciparum genome variation in 20,000 worldwide samples" MalariaGEN et al, Wellcome Open

Research 2023, 8:22 https://doi.org/10.12688/wellcomeopenres.18681.1. This work was supported by NIH 1R01AI145852 granted to ST-H and TDO by the U.S. National Institutes of Health.

References

1. World Health Organization. *World Malaria Report 2023* (World Health Organization, 2023).
2. Neafsey, D. E., Taylor, A. R. & MacInnis, B. L. Advances and Opportunities in Malaria Population Genomics. *Nature Reviews Genetics,* 1–16. ISSN: 1471-0064. (2021) (Apr. 2021).
3. Wesolowski, A. *et al.* Mapping Malaria by Combining Parasite Genomic and Epidemiologic Data. *BMC medicine* **16,** 190–190 (2018).
4. Shetty, A. C. *et al.* Genomic Structure and Diversity of Plasmodium Falciparum in Southeast Asia Reveal Recent Parasite Migration Patterns. *Nature Communications* **10,** 1–11. ISSN: 4146701910 (2019).
5. Taylor, A. R., Jacob, P. E., Neafsey, D. E. & Buckee, C. O. Estimating Relatedness between Malaria Parasites. *Genetics* **212,** 1337–1351 (2019).
6. Henden, L., Lee, S., Mueller, I., Barry, A. & Bahlo, M. Identity-by-Descent Analyses for Measuring Population Dynamics and Selection in Recombining Pathogens. *PLoS genetics* **14,** e1007279–e1007279. ISSN: 1111111111 (2018).
7. Gerlovina, I., Gerlovin, B., Rodríguez-Barraquer, I. & Greenhouse, B. Dcifer: An IBD-based Method to Calculate Genetic Distance between Polyclonal Infections. *Genetics* **222** (ed Browning, S.) iyac126. ISSN: 1943-2631. (2024) (Sept. 2022).
8. Schaffner, S. F., Taylor, A. R., Wong, W., Wirth, D. F. & Neafsey, D. E. HmmIBD: Software to Infer Pairwise Identity by Descent between Haploid Genotypes. *Malaria Journal* **17,** 10–13 (2018).
9. Schaffner, S. F. *et al.* Malaria Surveillance Reveals Parasite Relatedness, Signatures of Selection, and Correlates of Transmission across Senegal. *Nature Communications* **14,** 7268. ISSN: 2041-1723. (2024) (Nov. 2023).

10. Guo, B. *et al.* Strong Positive Selection Biases Identity-by-Descent-Based Inferences of Recent Demography and Population Structure in Plasmodium Falciparum. *Nature Communications* **15,** 2499. ISSN: 2041-1723. (2024) (Mar. 2024).

11. Browning, S. R. & Browning, B. L. Probabilistic Estimation of Identity by Descent Segment Endpoints and Detection of Recent Selection. *The American Journal of Human Genetics* **107,** 895–910. ISSN: 0002-9297, 1537-6605 (Nov. 2020).

12. Morgan, A. P. *et al.* Falciparum Malaria from Coastal Tanzania and Zanzibar Remains Highly Connected despite Effective Control Efforts on the Archipelago. *Malaria Journal* **19,** 1–14 (2020).

13. Nait Saada, J. *et al.* Identity-by-Descent Detection across 487,409 British Samples Reveals Fine Scale Population Structure and Ultra-Rare Variant Associations. *Nature Communications* **11,** 6130. ISSN: 2041-1723 (Nov. 2020).

14. Al-Asadi, H., Petkova, D., Stephens, M. & Novembre, J. Estimating Recent Migration and Population-Size Surfaces. *PLOS Genetics* **15** (ed DeGiorgio, M.) e1007908–e1007908. ISSN: 1111111111 (Jan. 2019).

15. Zhou, Y., Browning, S. R. & Browning, B. L. A Fast and Simple Method for Detecting Identity-by-Descent Segments in Large-Scale Data. *American Journal of Human Genetics* **106,** 426–437 (2020).

16. Freyman, W. A. *et al.* Fast and Robust Identity-by-Descent Inference with the Templated Positional Burrows–Wheeler Transform. *Molecular Biology and Evolution* **38,** 2131–2151. ISSN: 1537-1719 (May 2021).

17. Browning, S. R. & Browning, B. L. Accurate Non-parametric Estimation of Recent Effective Population Size from Segments of Identity by Descent. *American Journal of Human Genetics* **97,** 404–418 (2015).

18. Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLOS Genetics* **5,** e1000695. ISSN: 1553-7404 (Oct. 2009).

19. Joy, D. A. *et al.* Early Origin and Recent Expansion of Plasmodium Falciparum. *Science* **300,** 318–321. (2023) (Apr. 2003).

20. Gardner, M. J. *et al.* Genome Sequence of the Human Malaria Parasite Plasmodium Falciparum. *Nature* (2002).

21. Su, X.-z. *et al.* A Genetic Map and Recombination Parameters of the Human Malaria Parasite Plasmodium Falciparum. *Science (New York, N.Y.)* **286,** 1351–1353 (Nov. 1999).

22. Jiang, H. *et al.* High Recombination Rates and Hotspots in a Plasmodium Falciparum Genetic Cross. *Genome Biology* (2011).

23. Amambua-Ngwa, A. *et al.* Major Subpopulations of Plasmodium Falciparum in Sub-Saharan Africa. *Science (New York, N.Y.)* **365,** 813–816 (2019).

24. Miles, A. *et al.* Indels, Structural Variation, and Recombination Drive Genomic Diversity in Plasmodium Falciparum. *Genome Research* **26,** 1288–1299. ISSN: 1088-9051, 1549-5469 (Sept. 2016).

25. Kong, A. *et al.* A High-Resolution Recombination Map of the Human Genome. *Nature genetics* **31,** 241–247. ISSN: 1546-1718 (July 2002).

26. Camponovo, F., Buckee, C. O. & Taylor, A. R. Measurably Recombining Malaria Parasites. *Trends in Parasitology* **39,** 17–25. ISSN: 14714922. (2024) (Jan. 2023).

27. Bopp, S. E. R. *et al.* Mitotic Evolution of Plasmodium Falciparum Shows a Stable Core Genome but Recombination in Antigen Families. *PLOS Genetics* **9,** e1003293. ISSN: 1553-7404. (2023) (Feb. 2013).

28. Hamilton, W. L. *et al.* Extreme Mutation Bias and High AT Content in Plasmodium Falciparum. *Nucleic Acids Research* **45,** 1889–1901. ISSN: 0305-1048 (Feb. 2017).

29. McDew-White, M. *et al.* Mode and Tempo of Microsatellite Length Change in a Malaria Parasite Mutation Accumulation Experiment. *Genome Biology and Evolution* **11** (ed Baer, C.) 1971–1985. ISSN: 1759-6653. (2023) (July 2019).

30. Huber, J. H., Johnston, G. L., Greenhouse, B., Smith, D. L. & Perkins, T. A. Quantitative, Model-Based Estimates of Variability in the Generation and Serial Intervals of Plasmodium Falciparum Malaria. *Malaria Journal* **15,** 490. ISSN: 1475-2875 (Sept. 2016).

31. Churcher, T. S. *et al.* Public Health. Measuring the Path toward Malaria Elimination. *Science (New York, N.Y.)* **344,** 1230–1232. ISSN: 1095-9203 (June 2014).

32. Campbell, C. D. *et al.* Estimating the Human Mutation Rate Using Autozygosity in a Founder Population. *Nature Genetics* **44,** 1277–1281. ISSN: 1061-4036, 1546-1718. (2024) (Nov. 2012).

33. Taliun, D. *et al.* Sequencing of 53,831 Diverse Genomes from the NHLBI TOPMed Program. *Nature* **590,** 290–299. ISSN: 0028-0836, 1476-4687. (2024) (Feb. 2021).

34. MalariaGEN *et al.* Pf7: An Open Dataset of Plasmodium Falciparum Genome Variation in 20,000 Worldwide Samples. *Wellcome Open Research* **8,** 22. ISSN: 2398-502X. (2023) (Jan. 2023).

35. Tang, K., Naseri, A., Wei, Y., Zhang, S. & Zhi, D. Open-Source Benchmarking of IBD Segment Detection Methods for Biobank-Scale Cohorts. *GigaScience* **11,** giac111. ISSN: 2047-217X. (2022) (Jan. 2022).

36. Shemirani, R. *et al.* Rapid Detection of Identity-by-Descent Tracts for Mega-Scale Datasets. *Nature Communications* **12,** 3546. ISSN: 2041-1723. (2021) (June 2021).

37. Browning, B. L. & Browning, S. R. A Fast, Powerful Method for Detecting Identity by Descent. *The American Journal of Human Genetics* **88,** 173–182. ISSN: 0002-9297, 1537-6605 (Feb. 2011).

38. Thompson, E. A. Identity by Descent: Variation in Meiosis, Across Genomes, and in Populations. *Genetics* **194,** 301–326. ISSN: 1943-2631 (June 2013).

39. Browning, S. R. & Browning, B. L. Identity by Descent Between Distant Relatives: Detection and Applications. *Annual Review of Genetics* **46,** 617–633 (2012).

40. Kelleher, J., Thornton, K. R., Ashander, J. & Ralph, P. L. Efficient Pedigree Recording for Fast Population Genetics Simulation. *PLoS Computational Biology* **14,** 1–21. ISSN: 1111111111 (2018).

41. Kelleher, J. *et al.* Inferring Whole-Genome Histories in Large Population Datasets. *Nature Genetics* **51,** 1330–1338 (2019).

42. Speidel, L., Forest, M., Shi, S. & Myers, S. R. A Method for Genome-Wide Genealogy Estimation for Thousands of Samples. *Nature Genetics* **51,** 1321–1329 (Sept. 2019).

43. Rosvall, M., Axelsson, D. & Bergstrom, C. T. The Map Equation. *European Physical Journal Special Topics* **178,** 13–23. ISSN: 1951-6401 (Nov. 2009).

44. Csardi, G. & Nepusz, T. The Igraph Software Package for Complex Network Research. *InterJournal, complex systems* **1695,** 1–9 (2006).

45. Harris, D. N. *et al.* Evolutionary History of Modern Samoans. *Proceedings of the National Academy of Sciences* **117,** 9458–9465. (2023) (Apr. 2020).

46. Miotto, O. *et al.* Multiple Populations of Artemisinin-Resistant Plasmodium Falciparum in Cambodia. *Nature Genetics* **45,** 648–655 (2013).

47. Koenderink, J. B., Kavishe, R. A., Rijpma, S. R. & Russel, F. G. M. The ABCs of Multidrug Resistance in Malaria. *Trends in Parasitology* **26,** 440–446. ISSN: 1471-4922. (2023) (Sept. 2010).

48. Martin, R. E. & Kirk, K. The Malaria Parasite's Chloroquine Resistance Transporter Is a Member of the Drug/Metabolite Transporter Superfamily. *Molecular Biology and Evolution* **21,** 1938–1949. ISSN: 0737-4038 (Oct. 2004).

49. Brooks, D. R. *et al.* Sequence Variation of the Hydroxymethyldihydropterin Pyrophosphokinase: Dihydropteroate Synthase Gene in Lines of the Human Malaria Parasite, Plasmodium Falciparum, with Differing Resistance to Sulfadoxine. *European journal of biochemistry* **224,** 397–405. ISSN: 0014-2956 (Sept. 1994).

50. Early, A. M. *et al.* Declines in Prevalence Alter the Optimal Level of Sexual Investment for the Malaria Parasite Plasmodium Falciparum. *Proceedings of the National Academy of Sciences* **119,** e2122165119 (July 2022).

51. Miotto, O. *et al.* Genetic Architecture of Artemisinin-Resistant Plasmodium Falciparum. *Nature Genetics* **47,** 226–234 (Mar. 2015).

52. Browning, B. L. & Browning, S. R. Improving the Accuracy and Efficiency of Identity-by-Descent Detection in Population Data. *Genetics* **194,** 459–471. ISSN: 1943-2631 (June 2013).

53. Ishigohoka, J. & Liedvogel, M. *High-Recombining Genomic Regions Affect Demography Inference* Feb. 2024. (2024).

54. Mehra, S., Neafsey, D. E., White, M. & Taylor, A. R. *Systematic Bias in Malaria Parasite Relatedness Estimation* Apr. 2024. (2024).

55. MalariaGEN *et al.* An Open Dataset of Plasmodium Falciparum Genome Variation in 7,000 Worldwide Samples. *Wellcome open research* **6,** 42. ISSN: 2398-502X (2021).

56. Amir, A., Cheong, F. W., de Silva, J. R., Liew, J. W. K. & Lau, Y. L. Plasmodium Knowlesi Malaria: Current Research Perspectives. *Infection and Drug Resistance* **11,** 1145–1155. ISSN: 1178-6973. (2024) (Aug. 2018).

57. Bright, A. T. *et al.* A High Resolution Case Study of a Patient with Recurrent Plasmodium Vivax Infections Shows That Relapses Were Caused by Meiotic Siblings. *PLoS neglected tropical diseases* **8,** e2882. ISSN: 1935-2735 (June 2014).

58. Ibrahim, A. *et al.* Population-Based Genomic Study of Plasmodium Vivax Malaria in Seven Brazilian States and across South America. *The Lancet Regional Health - Americas* **18,** 100420. ISSN: 2667193X. (2023) (Feb. 2023).

59. Escalante, A. A., Cornejo, O. E., Rojas, A., Udhayakumar, V. & Lal, A. A. Assessing the Effect of Natural Selection in Malaria Parasites. *Trends in Parasitology* **20,** 388–395. ISSN: 1471-4922 (Aug. 2004).

60. Loy, D. E. *et al.* Out of Africa: Origins and Evolution of the Human Malaria Parasites Plasmodium Falciparum and Plasmodium Vivax. *International journal for parasitology* **47,** 87–97. ISSN: 0020-7519. (2024) (Feb. 2017).

61. Lee, K.-S. *et al.* Plasmodium Knowlesi: Reservoir Hosts and Tracking the Emergence in Humans and Macaques. *PLoS Pathogens* **7,** e1002015. ISSN: 1553-7366. (2024) (Apr. 2011).

62. Anderson, T. J. C., Paul, R. E. L., Donnelly, C. A. & Day, K. P. Do Malaria Parasites Mate Non-Randomly in the Mosquito Midgut? *Genetics Research* **75,** 285–296. ISSN: 1469-5073, 0016-6723. (2023) (June 2000).

63. Smukowski, C. S. & Noor, M. A. F. Recombination Rate Variation in Closely Related Species. *Heredity* **107,** 496–508. ISSN: 0018-067X. (2023) (Dec. 2011).

64. Haller, B. C. & Messer, P. W. SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model. *Molecular Biology and Evolution* **36,** 632–637. ISSN: 0737-4038 (Mar. 2019).

65. Haller, B. C., Galloway, J., Kelleher, J., Messer, P. W. & Ralph, P. L. Tree-Sequence Recording in SLiM Opens New Horizons for Forward-Time Simulation of Whole Genomes. *Molecular Ecology Resources* **19,** 552–566. ISSN: 1755-0998 (2019).

66. Baumdicker, F. *et al.* Efficient Ancestry and Mutation Simulation with Msprime 1.0. *Genetics* **220,** iyab229. ISSN: 1943-2631 (Mar. 2022).

67. Conway, D. J. *et al.* High Recombination Rate in Natural Populations of Plasmodium Falciparum. *Proceedings of the National Academy of Sciences of the United States of America* **96,** 4506–4511. ISSN: 0027-8424 (Apr. 1999).

68. Browning, B. L., Zhou, Y. & Browning, S. R. A One-Penny Imputed Genome from Next-Generation Reference Panels. *The American Journal of Human Genetics* **103,** 338–348. ISSN: 00029297 (Sept. 2018).

Supplementary Information

S1 Fig. Overview of methods used in benchmarking IBD detection methods. Benchmarking and optimization of IBD callers for *Pf* includes simulation analyses (top, green shading) and empirical data-based validation analysis (bottom, gray shading). (1) For the simulation study, the genealogical trees and phased genotype data are generated via the combination of forward (`SLiM`) and coalescent (`msprime`) simulations (indicated by the superscript a in the diagram). True IBD is obtained from a simulated genealogy tree via `tskibd` (b) and inferred IBD from the phased genotype using different IBD callers, including `hap-IBD`, `hmmIBD`, `isoRelate`, `Refined IBD` and `phased IBD`. IBD benchmarking is performed at two levels. The first is at the IBD segment levels. The metrics include false positive and false negative rates (c), population-level total IBD per length bin (d), and total IBD per isolate pair. The second is at the level of IBD-based downstream analyses, including the effective population size (N_e) by `IBDNe` (e), community membership through the InfoMap algorithm (f) and selection signals by statistics X_{iHS} (g). As the default parameters of different IBD callers, particularly those developed for human data, may not be ideal for *Pf* genomes, we performed grid searches for key parameters for each IBD caller so that the comparison was based on the best performance of each caller. (2) For validation in empirical data where true IBD is not available, we obtained IBD-based estimates using IBD from different callers and assessed which version of IBD can generate the expected patterns. The empirical data sets are subsamples of the MalariaGEN *Pf*7 data set (h). Sources of tools/methods/data: (a) `SLiM` [64, 65], `pyslim` [65], and `msprime` [66] for simulation; (b) `tskibd` [10] for obtaining true IBD from simulated genealogy trees; (c) False negative/positive rates are defined in

Zhou *et al.* 2020 [15]; (d) Population-level total IBD per length bin is defined consistent with Browning *et al.* 2015 [17]; (e) `IBDNe` [17] for the inferring trajectory of effective population size in recent history; (f) InfoMap algorithm [43, 44] for IBD-network based communities detection (population structure inference); (g) X_{iHS} statistics [10] for positive selection test; (h) Data used for validation analysis are from the MalariaGEN *Pf*7 [34].

S2 Fig. High recombination rates affect the quality of detected IBD segments (extending Fig 1). a-b, Accuracy of detected IBD segments from genomes simulated with different recombination rates. The accuracy of IBD segments is measured by the false negative rates (a) and the false positive rates (b). The plotted error rates reflect the genome-wide rates (as defined in Methods) of IBD segments called via default IBD caller parameters. Note, for both (a) and (b), the genomes were simulated under the single-population model (Methods). Data for low recombination rates is missing for `isoRelate` as genome sizes in bair-pair are too large for `isoRelate` to finish IBD calling in a timely manner.

S3 Fig. Pairwise total IBD of simulated *Pf*- like genomes from the different IBD callers compared with those based on true IBD (before parameter optimization). Quality of detected IBD as measured by genome-wide pairwise total IBD (total length of IBD segments $\geq$ 2cM for each genome pair), an estimate of genetic relatedness. Each dot represents a genome pair with the x, and y coordinates being true and the inferred total IBD. Data for each IBD caller were plotted in separate subplots as indicated in the titles. Note: both the x and y axes use a log10 scale. The blue dots are pairs with nonzero true and inferred total IBD. Whereas, red dots are pairs with the true total IBD or the inferred total IBD being 0; "0" total IBD was replaced with 1.0 cM for visualization. The red dotted line of $y = x$ indicates the expected pattern, i.e. true total IBD equal to inferred total IBD if the inferred IBD was 100% accurate.

S4 Fig. IBD caller-specific parameter optimization can improve the quality of IBD segments inferred from simulated *Pf* genomes (extending Fig 3). a, Quality of detected IBD measured by false positive and false negative rates before (left column) and after (right column) IBD caller-specific parameter optimization. As indicated in the legend of the axes, the error rates were calculated for different IBD segment length ranges (in centimorgans), including [3-4), [4-6), [6-10), [10-18), [18, inf), and at the genome-wide level. b, Quality of detected IBD measured by genome-wide pairwise total IBD, an estimate of genetic relatedness, before (left column) and after (right column) parameter optimization. Each dot represents a genome pair with the x, and y coordinates being true and inferred total IBD. Note: both the x and y axes in (b) use a log10 scale. In (b), the blue dots are pairs with non-zero true and inferred total IBD. Whereas, red dots are pairs with either true total IBD or inferred total IBD being 0; "0" total IBD was replaced with 1.0 cM for visualization. The red dotted line of $y = x$ indicates the expected pattern, that is,

the true total IBD equal to the inferred total IBD if the inferred IBD was 100% accurate. Note: This figure is similar to Fig 3 except that it includes the results for the five IBD callers analyzed. For both (a) and (b), each row represents results for a caller, as indicated by the vertical text label with a gray background on the left of the figure.

S5 Fig. Quality of IBD segments called via human-oriented callers from simulated genomes with UK human population demographic history but different recombination rates (human versus *Pf* rates). a, Quality of detected IBD as measured by false positive and false negative rates from genomes simulated with human (left axes) and *Pf* (right axes) recombination rates. As indicated in the legend of the axes (at the bottom of the panel), the error rates were calculated for different length ranges (centimorgan), including [3-4), [4-6), [6-10), [10-18), [18, inf) and at the genome-wide level. b, Quality of detected IBD measured by genome-wide pairwise total IBD, an estimate of genetic relatedness, from genomes simulated with human (left column) and *Pf* (right column) recombination rate. Each dot represents a genome pair with the x, and y coordinates being true and inferred total IBD. Note: (1) both the x and y axes in (b) use a log10 scale. In (b), the blue dots are pairs with non-zero true and inferred total IBD. Whereas, red dots are pairs with either true total IBD or inferred total IBD being 0; "0" total IBD was replaced with 1.0 cM for visualization. The red dotted line of $y = x$ indicates the expected pattern, i.e. true total IBD equal to inferred total IBD if the inferred IBD was 100% accurate. (2) Genomes were simulated under the demographic model mimicking the history of the UK human population (described in Zhou *et al.* 2020 [15]).

S6 Fig. Post-optimization benchmarking of different IBD callers by comparing downstream estimates of selected loci and population structure(extending Figure 4). a, IBD coverage and estimates of selected loci based on IBD segments called after parameter optimization. Each subplot corresponds to one of the five benchmarked IBD callers. In each subplot, lines of different colors represent the IBD coverage over one of 14 chromosomes; text in the upper right corner indicates the observed (on the left the "/") and expected (on the right of "/") numbers of true peaks and false peaks. A peak is defined as true if the span of the peak contains a selected site determined from the simulation. b, Post-optimization inference of population structure via the InfoMap algorithm. In each subplot, the row labels indicate the true population labels set in the simulation, and the column labels indicate the detected community labels. C1 - C5 represent the largest five detected communities. The color within each block represents the number of samples assigned to the given true population label and the detected community label. Columns are rearranged for better visualization. See S9 Fig for pre-optimization results.

S7 Fig. Population-level total IBD per length bin. The population-level total length of IBD segments with a given length range is an intermediate quantity used by `IBDNe` internally to estimate the trajectory of N_e at different times in recent history. The population-level total IBD inferred by an IBD caller (as shown in the title of each subplot) is compared with that of true IBD. The log ratio of inferred to true population level total IBD was plotted against length bins (represented by bin centers in centimorgans). The x-axis uses the log10 scale to better visualize the bias related to shorter IBD segments. Blue dots represent a length bin with non-zero population-level total IBD for both true and inferred sets. Red dots indicate either the true or inferred population-level total IBD is zero for the corresponding length bin (x coordinate). A line of $y = 0$ is added as the expected line, meaning that the log ratios for all length bins would be 0's if there were no biases.

S8 Fig. The effect of including and excluding short IBD segments ($<$ 4cM) on N_e estimation. a. N_e estimates for genomes simulated under the single population model without selective sweep. In the upper row, IBD segments with TMRCA $<$ 1.5 generations were excluded from the IBD input for N_e estimation; in the bottom row, they are kept. For true IBD segments (from `tskibd`), the TRMCA for each IBD segment was obtained directly from genealogical trees; For inferred IBD segments, any of them that overlapped with a segment of the truth set with TMRCA $<$ 1.5 generations was determined to have TMRCA $<$ 1.5 generations. b. N_e estimates for genomes for different empirical data sets. Each row corresponds to a data set labeled on the right. The columns are estimates for different IBD callers. Solid red line: N_e estimates when short IBD segments are included; red dashed line: N_e estimates when short IBD segments are excluded; black dashed line: parameter population sizes used in simulations.

S9 Fig. Pre-optimization benchmarking of different IBD callers by comparing downstream estimates of selected loci and population structure (for comparison with S6 Fig). a, IBD coverage and estimates of selected loci based on IBD segments called before parameter optimization. Each subplot corresponds to one of five benchmarked IBD callers. In each subplot, lines of different colors represent the IBD coverage over one of 14 chromosomes; text in the upper right corner indicates the observed (on the left the "/") and expected (on the right of "/") numbers of true peaks and false peaks. A peak is defined as true if the span of the peak contains selected sites determined from simulation. b, Post-optimization inference of population structure via InfoMap algorithm. In each subplot, the row labels indicate the true population labels set in the simulation and the column labels indicate the detected community labels. C1 - C5 represent the largest five detected communities. The color within each block represents the number of samples assigned to the given true population label and the detected community label. Columns are rearranged for visualization purposes.

S10 Fig. Pre-optimization benchmarking of different IBD callers by comparing downstream estimates of N_e (for comparison with Fig 4). With mainly default values for IBD caller parameters, the performance of IBD callers was evaluated by comparing the N_e trajectory for the recent 100 generations estimated via `IBDNe` based on true (black dashed line) IBD versus inferred IBD (red solid line). True IBD was calculated from simulated genealogical trees via `tskibd` inferred IBD includes those inferred from `tskibd`, `hap-IBD`, `hmmIBD`, `isoRelate`, `Refined IBD`, and, `phased IBD`, with their N_e estimates shown from left to right. Red shading area in each plot indicate the 95% confidence interval for N_e estimates as generated by `IBDNe`.

S11 Fig. Performance of IBD callers with empirical data sets before parameter optimization (in comparison with post-optimization performance shown in Fig 5). a, IBD coverage and detected selection signals in the SEA data set using different IBD callers (rows 1 to 5). Annotations and corresponding vertical dotted lines at the top indicate the center of known and putative drug resistance genes and genes related to sexual commitment; red shading indicates regions that are inferred to be under positive selection (see Methods for definitions). Note the difference in the peak height of IBD and the counts of `hap-IBD` and `phased IBD` compared to Fig 5a). b, N_e estimates of the SEA data set based on IBD inferred from different callers. The line plots are the point estimates; the shading areas around the line plots indicate confidence intervals based on bootstrapping (generated by `IBDNe`). c, Inference of the population structure of the structured data set by the InfoMap community detection algorithm using IBD network inferred from different IBD callers. The rows of the heatmap are geographic regions of isolates, the columns of the heatmap are the largest, inferred communities, labeled as c1 to c6. The color of the heatmap represents the number of isolates in each block with the given row and column labels. The columns are rearranged such as diagonal blocks tend to have the largest values per row for better visualization. Note the difference in size of the main detected main communities for `hap-IBD` and `phased IBD` for parasites from Africa when compared with Fig 5c.

S12 Fig. Comparison of maximum computer memory used for IBD calling process for different callers. a, Memory usage for different IBD callers to detect IBD from genomes of different sizes in single-thread mode. The comparison is based on *Pf* genomes simulated under the single population model. The tick labels on the x axis include the number of pairs of genomes analyzed (below the plot, in linear scale) and the sample size (number of genomes, in nonlinear scale) analyzed. The analyzed IBD callers included the five tools benchmarked previously. The line style and marker for different callers/tools are provided in the legend box on the far right of the figure, which is shared across the three subplots. b, Memory usage under the multi-threading mode. (b) is organized similarly to (a), except that the IBD calling processes were

run under the multithreading mode with 10 threads.

S1 Table. The default and optimized values for parameters used in inferring IBD segments via different callers. A link to the source code and a citation of the corresponding article are provided for each IBD caller. For `hap-IBD`, `Refined IBD`, and `hmmIBD`, the parameters are used on the command line except that the parameter rec_rate of `hmmIBD` needs to be specified in the source code file `hmmIBD`.c. For `phased IBD` and `isoRelate`, the parameters are specified within a Python or R script. The details of how the parameters are specified can be found in the scripts on GitHub (`https://github.com/bguo068/bmibdcaller_simulations/tree/main/bin`). Note that `mincm` and `minmaf` are values shared across IBD callers to allow fair comparisons.

S2 Table. Grid/Line search to optimize IBD caller-specific parameters. Column 2 lists the optimized parameters (other parameters are not explored). Column 3 shows the tested values for each parameter. Some parameters are optimized via two steps, such as coarse search (coarse) and fine-tuning (finetune). Column 4 provides comments on the impact of the values on IBD accuracy (measured as false positive rates (FP) and false negative rates(FN)).

S3 Table. Isolates in the "multi-population" data set used in empirical validation. Rows are counts of isolates from different locations ("Population" labels from MalariaGEN *Pf*7 meta information table).

S4 Table. Isolates in the "single-population" data set for "AS-SE-E" population used in empirical validation. Rows are counts of isolates from different locations ("Country" labels from MalariaGEN *Pf*7 meta information table); columns are counts of isolates collected at a given year.

S5 Table. Isolates in the "single-population" data set for "AF-W-Ghana" population used in empirical validation. Rows are counts of isolates from different locations ("admin level 1" labels from MalariaGEN *Pf*7 meta information table); columns are counts of isolates collected at a given year.

S1 Data. Detailed IBD-level benchmarking results shown in heatmaps. Each panel (a-r) represents the FN/FP rates for a specific combination of IBD callers (`hap-IBD`, `hmmIBD`, `isoRelate`, `phased IBD` and `Refined IBD`), demographic models (single-population model, multiple-population model, and UK human demographic model) and recombination rates (human *versus Pf*) as indicated in the text above the panel. Note that benchmarking for genomes with human recombination rates was not performed for `hmmIBD`

and `isoRelate` as they did not scale well for large genome sizes (in base pair). For each heatmap, the searched parameters and their values are indicated as the x and y labels and tick labels, respectively; the labels in grey background on the top indicate the IBD length bin that were used to calcuate FN/FP rates (labels in grey on the left); the bold labels in grey at the bottom, show either different groups (e.g., coarse search or finetune) or the third parameter searched (such as min-maf=0.01 and min-maf=0.00).