Bayes-optimal estimation of overlap between populations of fixed size

Daniel B. Larremore[1,2,*]

[1]*Department of Computer Science, University of Colorado Boulder, Boulder, CO, USA*
[2]*BioFrontiers Institute, University Colorado at Boulder, Boulder, CO, USA*

Measuring the overlap between two populations is, in principle, straightforward. Upon fully sampling both populations, the number of shared objects—species, taxonomical units, or gene variants, depending on the context—can be directly counted. In practice, however, only a fraction of each population's objects are likely to be sampled due to stochastic data collection or sequencing techniques. Although methods exists for quantifying population overlap under subsampled conditions, their bias is well documented and the uncertainty of their estimates cannot be quantified. Here we derive and validate a method to rigorously estimate the population overlap from incomplete samples when the total number of objects, species, or genes in each population is known, a special case of the more general β-diversity problem that is particularly relevant in the ecology and genomic epidemiology of malaria. By solving a Bayesian inference problem, this method takes into account the rates of subsampling and produces unbiased and Bayes-optimal estimates of overlap. In addition, it provides a natural framework for computing the uncertainty of its estimates, and can be used prospectively in study planning by quantifying the tradeoff between sampling effort and uncertainty.

I. INTRODUCTION

Quantifying the similarity between two populations, environments, or ecosystems, based on their constituent members or species, is a fundamental problem in ecology. Some methods quantify this pairwise similarity, often called β-diversity [1], based on only the presence or absence of species [2], while other methods take into account species abundance as well [3]. Still other methods, more common in microbial ecology, make use of genetic sequence data, measuring similarities through phylogenetic relationships [4–6]. In practical applications of all three types of methods, the populations being compared are almost always undersampled, meaning that estimators which are principled in the context of perfect sampling show substantial bias in practice [7].

Consider, as an example of estimator bias, the oldest of pairwise similarity measures, which have roots in botany with Jaccard, Dice, and Sørenson. Their 1901 [8] and 1940s [9, 10] publications introduced the eponymous Jaccard index and Sørenson-Dice coefficient. Both are simple ratios involving the number of distinct species observed in each population, n_a and n_b, and the number of species shared by both populations, n_{ab}, so that each measure quantifies overlap as a fraction,

$$\mathring{J} = \frac{n_{ab}}{n_a + n_b - n_{ab}}\,, \quad \mathring{S} = \frac{n_{ab}}{\frac{1}{2}(n_a + n_b)}\,. \tag{1}$$

Intuitively, when the two populations are identical, both $\mathring{J}$ and $\mathring{S}$ are one, and when two populations are entirely distinct, both are zero. However, imagine two populations of 10 species each in which 5 species are found in both populations. With perfect sampling, $\mathring{J} = \frac{1}{3}$ and $\mathring{S} = \frac{1}{2}$, but when only 9 of 10 species are drawn from each population, these indices, computed with empirically observed values, average $\mathrm{E}[\mathring{J}] = 0.29$ and $\mathrm{E}[\mathring{S}] = 0.45$, representing relative biases of -12% and -10% respectively. These biases, which are well documented [7], become worse as sampling rates fall.

Bias is not unique to the Jaccard and Sørenson-Dice coefficients, but instead affects all algebraic combinations of n_a, n_b, and n_{ab}, of which over 20 have been proposed [2]. This is due to the fact that observed values of n_a, n_b, and n_{ab} are realizations of random variables, and in particular, n_{ab} necessarily shows a nonlinear dependence on n_a and n_b. As a consequence, guides to navigating the multiple measures of β diversity—including both presence/absence and abundance measures—emphasize matching of estimators' principles and the scientific questions they are meant to answer [11], but do not address underlying bias, variation, or uncertainty itself.

Progress toward unbiased estimators has been made, however. If estimators account for not just the presence or absence of species, but their abundance as well, they can be adjusted based on the effects of both observed and *unobserved* species. Chao et al. took a probabilistic view of such adjustments based on the underlying sampling process, resulting in modified Jaccard and Sørenson-Dice coefficients with substantially reduced bias [7]. Other successful approaches directly model the sampling process itself. For instance, Kery and Royle introduced a hierarchical Bayes approach to species-richness estimation by posing a spatial sampling process and using it to improve richness estimates [12]. Importantly, this Bayesian approach allowed them to quantify uncertainty in their estimates via credible intervals. Outside of ecology entirely, the estimation of overlap between sets arises in large-data scenarios, e.g. when comparing two individuals' sets of interests or friends on Facebook using streaming algorithms in distributed settings [13]. These approaches show that better estimates are possible when the presence of uncertainty due to stochastic sampling is addressed directly, even when much about the underlying populations is unknown.

Here, we solve a special case of the more general pairwise similarity problem described above—one which is particularly relevant to the genomic epidemiology and disease

* daniel.larremore@colorado.edu

Plain-English Summary. Understanding when two populations are composed of similar species is important for ecologists, epidemiologists, and population geneticists, and in principle it is easy: just sample the two populations, compare the sets of species identified in each, and count how many appear in both populations. In practice, however, this is difficult because sampling methods typically produce only a random subset of the total population, leaving current population overlap estimates biased. Knowing only the number of shared members between two of these partial population samples, this paper shows how we can nevertheless estimate the true overlap between the full populations, when those full populations' sizes are known. Using Bayesian statistics, we can also compute credible intervals to produce error bars. We show that using this unbiased approach has a dramatic impact on the conclusions one might draw from previously published studies in the malaria literature, which used simple but biased methods. Because the method in this paper quantifies the tradeoff between sampling effort and uncertainty, we also show how to compute the number of samples required to ensure high-confidence results, which may be useful for planning future studies or budgeting lab reagents and time.

ecology of *Plasmodium falciparum*, the most virulent of the human malaria parasites. Rather than comparing two ecosystems based on their shared species, we consider the problem of comparing two genomic repertoires based on their shared gene variants. Mathematically, the problems are similar but with one important difference: when estimating genomic repertoire overlap, the total number of variants per genome is known. This additional specification opens the door to unbiased and Bayes-optimal estimation of true repertoire overlap, given a single noisy measurement of n_a, n_b, and n_{ab}, while also quantifying the increased uncertainty inherent in decreased sampling.

The *P. falciparum* repertoire overlap problem. Of the diverse multigene families of *P. falciparum*, the *var* family is the most heavily studied because of its direct links to both malaria's duration of infection and its virulence [14–17]. Each parasite genome contains a repertoire of $\sim$ 60 hypervariable and mutually distinct *var* genes, but repertoires differ between parasites, evolving rapidly through recombination and reassortment. Recent studies of *P. falciparum* epidemiology and evolution have generated insights by comparing of the sets of genomic *var* repertoires between parasites [18–23]. Indeed, since *var* repertoires are, themselves, under selection, theory suggests that if a human population has been exposed to particular *var* genes, then repertoires containing those *var* genes will have a lower fitness than repertoires that are entirely unrecognized by local hosts, shaping the *var* population structure [21, 22, 24, 25]. Methods by which we estimate the extent to which *var* repertoires overlap are therefore important, particularly as studies of the population genetics and genetic epidemiology of malaria's antigens become more sophisticated and data rich. However, as with estimates of β-diversity in ecology, traditional estimates of overlap between *var* repertoires also suffer bias due to subsampling.

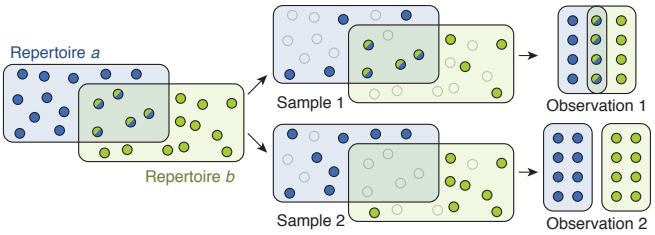

FIG. 1. **Stochastic sampling leads to variation in observed overlap**. The members of two hypothetical populations are represented by blue and green circles, respectively. Each population has 16 members, and $s = 5$ are shared members of both populations. In two independent sampling experiments, shown in top and bottom rows, $n_a = n_b = 8$ members are sampled at random from each population (dark circles) while the other 8 members are not sampled (transparent circles). Observation of the first experiment finds an overlap of $n_{ab} = 4$, while observation of the second finds $n_{ab} = 0$.

Due to the massive diversity and recombinant structure of *var* genes, researchers are restricted to using degenerate PCR primers targeting a small "tag" sequence within a particular *var* domain called DBLα [26]. Due to their experimental accessibility, DBLα tags have been widely used to study the structure and function of *var* genes [17, 18, 21, 26–30]. Still, these PCR techniques generate a random sample of 60 or fewer unique tag sequences from each parasite. This means that experimental measurements of repertoire overlap are performed using stochastic subsamples whose *empirical* overlap may fluctuate from experiment to experiment (Figure 1), motivating the three questions answered by this paper: First, how can we estimate the true overlap between repertoires when we can only measure the overlap between samples from repertoires? Second, how can we quantify the uncertainty around our repertoire overlap estimates? Third, what are the implications of uncertainty for the design and budgeting of *var* repertoire studies?

In the malaria literature, repertoire overlap is most commonly computed using the Sørenson-Dice coefficient where it is often called *pairwise type sharing* [18]. When PCR methods have produced n_a and n_b tags from parasites a and b, respectively, and when a sequence-level comparison has found n_{ab} tags are shared by both repertoires then repertoire similarity is computed using the coefficient $\mathring{S}$ in Eq. (1). When n_a and n_b are nearly 60, the performance of $\mathring{S}$ is excellent. For instance, when two parasites are completely different, $n_{ab} = 0$, so $\mathring{S} = 0$; when two parasites are identical, and both repertoires have been fully sampled, $n_{ab} = n_a = n_b$, so $\mathring{S} = 1$. However, when n_a or n_b is smaller (as is overwhelmingly the case in existing studies [18–23]) $\mathring{S}$ is conservative and systematically underestimates the true overlap between repertoires [7].

Organization. In this manuscript, I introduce a method that estimates repertoire overlap using Bayesian inference. By modeling the stochastic process by which repertoires are sampled, I show that this method produces unbiased *a posteriori* estimates of true repertoire overlap. I then show how the Bayesian framework can be used to estimate uncertainty and produce error bars which represent credible intervals, a Bayesian analog of confidence intervals. These methods are then used to reevaluate past results which used the Sørenson-Dice coefficient $\mathring{S}$. Finally, in the case of *P. falciparum*, since each successful PCR amplification randomly samples just one of 60 available tags, I extend the Bayesian approach to compute the tradeoff between increasing sampling and decreasing the uncertainty of overlap estimates. These calculations allow the cost of sampling to be weighed against scientific confidence, illustrating the use of this statistical framework for planning and budgeting experiments. Open-source code and a web tool are freely available (see Acknowledgements).

II. METHODS

Suppose that there are two *P. falciparum* parasites, each with a repertoire of 60 *var* types. Our goal is to estimate the true repertoire overlap s (were we to fully sample each parasite) from the knowledge that n_a samples from parasite a and n_b samples from parasite b share n_{ab} types. Due to the fact that the underlying sampling process is stochastic (Figure 1), our secondary goal is to quantify the uncertainty in the method's estimates. Both goals can be met by writing down the process that creates the data in the first place. Therefore, in what follows, we will at first assume that the true overlap s is fixed, model the process of generating data via stochastic sampling, and use that model to compute a likelihood. We will then use Bayes' Rule to compute the posterior probability for each value of s, given the evidence in the data and the likelihood computed in the first step.

Consider the following sampling process, written in the slightly more rigid and generic language of a probability textbook. Suppose that there are s special objects among a total of N objects. We draw n objects uniformly at random without replacement. The number of special objects chosen during this sampling procedure will be distributed according to a hypergeometric distribution, which we write as $\mathcal{H}(s, N, n)$.

First, with this definition in mind, consider drawing n_a *var* genes from parasite a's 60 total. Of the 60 total, suppose that exactly s are considered special because they are also shared by parasite b. The number of shared sequences that are captured by sequencing parasite a will be a random variable $S_a = \mathcal{H}(s, 60, n_a)$. Depending on the luck of the draw, this number could be as small as zero, or as high as s or n_a (whichever is smaller).

Now consider drawing n_b *var* genes from parasite b's 60 total, in which exactly s_a are special because they are shared by both parasites *and* were actually drawn during the sequencing of parasite a. This process is identical in construction to the process for sampling parasite a, but with s_a special sequences instead of s, and so the number of shared sequences that are captured after sequencing both parasites will be $\mathcal{H}(s_a, 60, n_b)$. Substituting the random variable S_a for a fixed value s_a, which we derived in the paragraph above, yields a hypergeometric inside a hypergeometric, which means that the probability of a particular number of shared sequences in the samples n_{ab} is given by these sequential (or nested) hypergeometric distributions,

$$P(n_{ab} \mid n_a, n_b, s) \sim \mathcal{H}\big(\mathcal{H}(s, 60, n_a), 60, n_b\big) \; . \qquad (2)$$

Reassuringly, one can switch the order in which the imagined sampling took place, first sequencing parasite b and then sequencing parasite a, or sequencing them both at once, and show that these are mathematically equivalent.

In practice, we want to go the other direction, and estimate s from our empirical measurements of n_a, n_b, and n_{ab}. Since the distributions above allow us to compute the likelihood of empirical observations, given s, we use Bayes' rule to formulate the posterior distribution for s,

$$P(s \mid n_a, n_b, n_{ab}) = \frac{P(n_{ab} \mid n_a, n_b, s) P(s)}{P(n_{ab})} \; , \qquad (3)$$

where $P(s)$ is the prior distribution for overlap. In practice, we generally wish to remain agnostic about the level of overlap s and therefore we consider an uninformative prior $P(s) \sim \text{unif}[0, 60]$, i.e. $P(s) = \frac{1}{61}$. Using the law of total probability to rewrite the denominator, and canceling the factors of $\frac{1}{61}$, we get

$$P(s \mid n_a, n_b, n_{ab}) = \frac{P(n_{ab} \mid n_a, n_b, s)}{\sum_{s'=0}^{60} P(n_{ab} \mid n_a, n_b, s')} \; . \qquad (4)$$

Each term on the right hand side of Eq. (4) can now be computed directly from the nested hypergeometric distributions in Eq. (2) as follows. To generate a specific empirical overlap n_{ab}, two things must have happened in succession and independently of each other: first, s_a of the original s shared sequences must have been sampled; and second, n_{ab} of the intermediate s_a shared sequences must then have been sampled. We therefore multiply these two hypergeometric probabilities. However, because this sequential process may occur for any value of the intermediate variable s_a, we sum over all possible values of s_a,

$$P(n_{ab} \mid n_a, n_b, s) = \sum_{s_a=0}^{60} P(n_{ab} \mid n_b, s_a) P(s_a \mid n_a, s) \; . \qquad (5)$$

Thus, computing the posterior probability that the true overlap was s, given the empirical overlap between samples, is

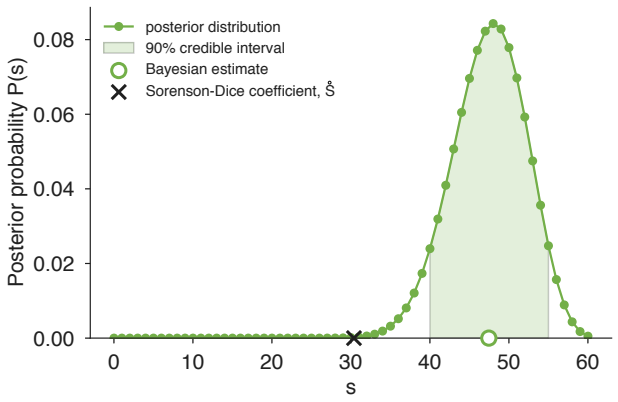

FIG. 2. **Inference and uncertainty using the posterior.** The posterior distribution over s is plotted for the realistic scenario of $n_a = 47$, $n_b = 32$, and $n_{ab} = 20$ [line; Eq. (6)]. The posterior mean provides our estimate of the true overlap $\hat{s}$ [open circle; Eq. (7)], and the interval accounting for at least 90% of the area under the posterior curve provides an equal-tailed 90% credible interval [shading; Eq. (8)]. The $\mathring{S}$ estimate is shown for comparison [black cross; Eq. (1)], and is typically less than or equal to $\hat{s}$.

given by substituting Eq. (5) into Eq. (4), yielding

$$P(s \mid n_a, n_b, n_{ab}) = \frac{\sum_{s_a=0}^{60} P(n_{ab} \mid n_b, s_a) P(s_a \mid n_a, s)}{\sum_{s'=0}^{60} \sum_{s_a=0}^{60} P(n_{ab} \mid n_b, s_a) P(s_a \mid n_a, s')} . \tag{6}$$

The term $P(s \mid n_a, n_b, n_{ab})$ is a posterior distribution over s, meaning that it tells us the probability for each value of s, given the evidence provided by the actual data. While this equation appears notation-heavy, its inference requires only calls to the hypergeometric probability distribution. To illustrate this graphically, the posterior distribution is plotted for $n_a = 47$, $n_b = 32$, and $n_{ab} = 20$ in Figure 2.

The posterior distribution can now be used (i) to estimate the true value of s, and (ii) to quantify the uncertainty of that estimate. First, our estimate for the true value of s, which we call $\hat{s}$, is the expected value of the posterior,

$$\hat{s} = \sum_{s=0}^{60} s \, P(s \mid n_a, n_b, n_{ab}) . \tag{7}$$

This value is typically (in 99.85% of all possible cases) larger than the estimate provided by $\mathring{S}$ (Fig. 2).

The framework here is easily extended to repertoire sizes other than 60, generalizing to applications beyond *var* genes. Suppose that populations a and b have total sizes of N_a and N_b, and without loss of generality, assume that $N_a \leq N_b$. The values N_a and N_b need only be substituted into Eqs. (6) and (7), with $P(s) = (N_a + 1)^{-1}$. This is shown in Eqs. (S1) and (S2), but not shown here for conciseness.

The posterior distribution provides a convenient way to quantify the uncertainty associated with an estimate $\hat{s}$. Intuitively, if the posterior is sharply peaked around $\hat{s}$, then our confidence in $\hat{s}$ is high; if the posterior is broadly distributed then our confidence in $\hat{s}$ is low. Making use of the Bayesian construction once more, we compute a credible interval by finding the range of s values that account for 90% of the posterior probability (Fig. 2). Due to the fact that the posterior distribution is a discrete distribution over only 61 values, it is possible (indeed, highly probable) that no interval will contain exactly 90% of the probability. Nevertheless, we define a conservative equal-tailed 90% credible interval $[s_{\text{min}}, s_{\text{max}}]$ as the smallest index s_{min} and the largest index s_{max} for which

$$\sum_{s=s_{\text{max}}}^{60} P(s \mid n_a, n_b, n_{ab}) \geq 0.05$$
$$\sum_{s=0}^{s_{\text{min}}} P(s \mid n_a, n_b, n_{ab}) \geq 0.05 . \tag{8}$$

III. RESULTS

A. Estimator performance

We first demonstrate that the $\hat{s}$ computed in Eq. (7) produces accurate estimates by simulating the sampling process with known s and evaluating our ability to accurately recover it. Specifically, for each simulation, we consider two *var* repertoires a and b, of 60 genes each, and specify a priori that they share exactly s sequences. We then choose the number of samples taken from each, n_a and n_b respectively, and draw from each repertoire uniformly at random, without replacement. These draws are compared to compute the number of empirically shared sequences n_{ab}. Equation (7) is used to compute the Bayesian repertoire overlap (BRO) estimate $\hat{s}$, while Eq. (1) is used to compute $\mathring{S}$ using the same data. These estimates are then compared to the true value of s to evaluate accuracy. Varying the values of s, n_a, and n_b allows us to quantify the performance of BRO and $\mathring{S}$ in a variety of realistic sampling scenarios.

Figure 3 shows the results of this simulation for sampling rates of 30, 40, and 50 genes, with two independent simulations at each value of s. Intuitively, both BRO and $\mathring{S}$ are more accurate when n_a and n_b are larger. However, the two methods' behaviors are fundamentally different. When n_a and n_b are below 60, BRO provides estimates that are distributed around the true overlap, with variance decreasing as sampling rates increase. In contrast, $\mathring{S}$ systematically underestimates the true overlap, while also showing decreasing variance as sampling rates increase [7]. For realistic sampling rates, BRO provides estimates centered at the true value, while $\mathring{S}$ provides estimates centered below the true value. These general patterns hold even when $n_a \neq n_b$ or when total repertoire sizes are unequal (Figure S1).

Credible intervals, which visually show uncertainty in each

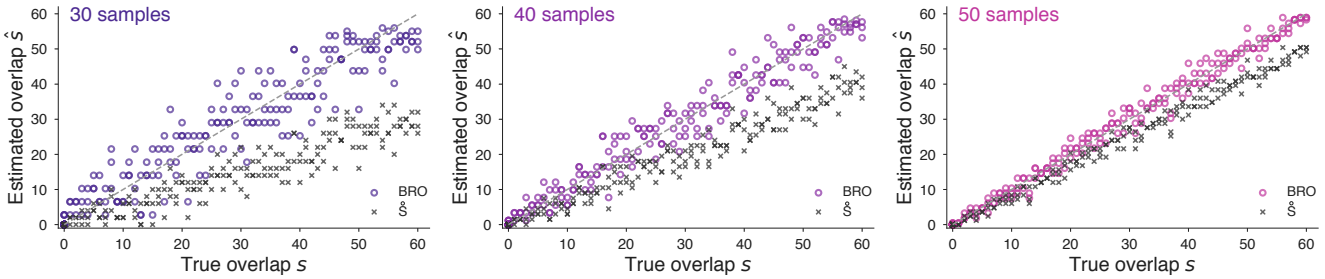

FIG. 3. **Bayesian repertoire overlap consistently estimates true overlap**. Repertoires with true overlaps ranging from 0 to 60 were subsampled in simulations. As sampling rates increase from $n_a = n_b = 30$ (left) to 40 (middle) and to 50 (right), the estimates of BRO (colored circles) approach the true values (dotted lines) symmetrically. Estimates from $\mathring{S}$ (crosses) approach the true values from below, systematically underestimating the true overlap. This bias is worse with lower sampling rates [7]. Similar results are found when $n_a \neq n_b$, and when the total repertoire sizes are different from each other (Figure S1).

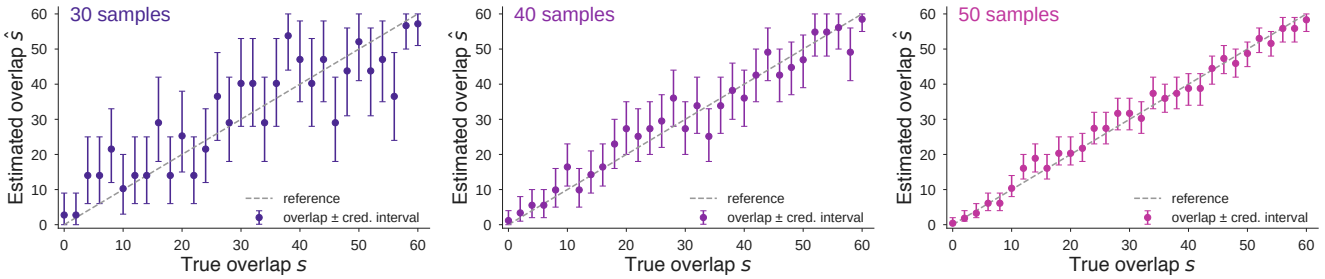

FIG. 4. **Credible intervals quantify uncertainty in overlap estimates**. By using Eq. (8), 90% credible intervals are show above as error bars around the point estimates $\hat{s}$ for varying true overlap s. As sampling rate increases from $n_a = n_b = 30$ (left) to 40 (middle) and to 50 (right), credible intervals shrink, indicating a reduction in uncertainty. In expectation, 90% of intervals cover the true overlap (dotted line).

estimate, can also be easily computed from the simulations described above. For each simulation, Eq. (8) uses the posterior distribution over s to produce error bars around the point estimate $\hat{s}$, shown for sampling rates of 30, 40, and 50 in Figure 4. This illustrates the substantial reduction in uncertainty that comes with increased sampling rates. While all simulations shown here use $n_a = n_b$, this is by no means required (see Figure S1), and in real data scenarios, is rare.

B. Revisiting past results

We now show how the methods of this paper can be used in practical contexts by applying them to data from three published studies. In particular, this reanalysis highlights the impact of variation in sampling rates across studies, which creates variable bias in $\mathring{S}$ calculations and produces misleading results. However, we also show that while using BRO in place of $\mathring{S}$ sidesteps bias problems, the ability to quantify uncertainty with error bars highlights new problems. In short, the conclusions of previous studies may be worth reevaluating.

In 2007, Barry et al. introduced $\mathring{S}$, which they referred to as *pairwise type sharing*, in an analysis of *var* data from Amele, Papua New Guinea [18]. In 2010, Albrecht et al. included Barry's data in a broader analysis of *var* data from Ariquemes, Brazil [19] which also included sequences from a study by Bull et al. from Kilifi, Kenya for comparison [27]. Each one of these studies, individually, sequenced parasite isolates to a particular target depth, yet the studies varied in their coverage of repertoires. Since the bias of $\mathring{S}$ depends on the number of samples (Fig. 3; see also [7]), the variation of sampling rates across study populations means that different populations are biased downward by different amounts.

Albrecht et al.conveniently provide *var* type data from all three studies, from which we can rebuild their first figure which shows a $\mathring{S}$ comparison of five populations (Fig. 5; left). Overlaps between pairs of parasites can then be recomputed using BRO (Fig. 5; middle). The conclusions drawn from these two figures differ substantially.

First, according to $\mathring{S}$, identical clones from Ariquemes share only around 30 sequences with themselves, illustrating the downward bias produced by subsampling—clones ought to share all of their genes with their genetically identical siblings. Indeed, the reanalysis using BRO finds over 75% of overlap estimates to be greater than 50 (and over 50% over 55), far closer to what is expected.

Second, the inter-clone overlap and inter-isolate overlap distributions in Ariquemes appear to be similar and overlapping through the lens of $\mathring{S}$. However, the recalculation using

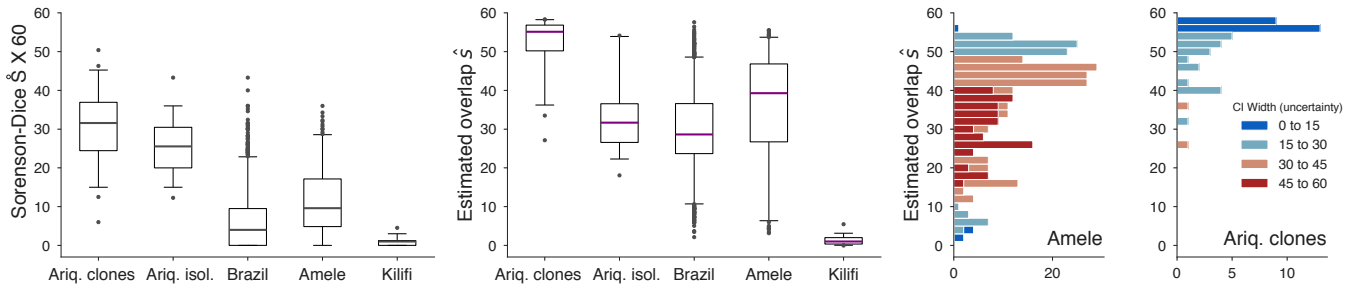

FIG. 5. **Reevaluation of published results.** In 2010, Albrecht et al. compared *var* repertoires from 5 populations using pairwise type sharing (see Refs. [18, 19, 27] for original data details). (left) Reproduction of $\mathring{S}$ analysis of [19], rescaled from $[0,1] \to [0,60]$. (middle) Reanalysis using Bayesian repertoire overlap [Eq. (7)]. For all boxplots, boxes span inner quartiles; center lines show medians; whiskers extend to 2.5 and 97.5 percentiles. (right) Histograms of Bayesian repertoire overlap distributions from Amele and Ariquemes clones (data identical to those in middle boxplots) colored by width of credible interval [Eq. (8)], a measure of uncertainty. Differences in uncertainties are driven primarily by sampling rates: Amele samples average $\bar{n} = 15.6$ sequences per parasite while Ariquemes clones average $\bar{n} = 26.5$.

BRO shifts the clones' distribution dramatically upward but leaves the isolates' distribution more or less untouched. This is due to the dramatic difference in *var* coverage: the average number of sequences per clone is $\bar{n} = 26.5$ while for isolates it is $\bar{n} = 45.8$, meaning that relatively different amounts of bias are inherited from $\mathring{S}$ (illustrated in simulations in Fig. 3).

Finally, the distributions from Brazil ($\bar{n} = 17.3$) and Amele ($\bar{n} = 15.6$) also shift dramatically upward when the bias of $\mathring{S}$ is removed (Fig. 5; left, middle). However, this does not necessarily mean that they should be reinterpreted. For each pairwise comparison, Eq. (8) allows us to compute the width of the credible interval, $s_{\max} - s_{\min} + 1$, quantifying our uncertainty in each estimate. Due to low average coverage, the uncertainty of estimates in the Amele dataset tends to be extremely large (Fig. 5; right), with the majority of estimates showing an uncertainty greater than 30 sequences (50% overlap). For comparison, estimates from Thiès, Senegal [21] ($\bar{n} = 36.0$) are also shown, whose dramatically lower uncertainty enables more confident conclusions to be drawn.

There are two main methodological findings that result from using rigorous and unbiased methods. First, the boxplots of Fig. 5 clearly illustrate that sampling rates can have a dramatic impact on findings, reinforcing the simulation results of Fig. 3. Second, uncertainty is an issue when $\bar{n}$ is too small, and datasets with low sampling rates may have such wide error bars that their estimates should not be trusted, as shown in the histograms of Fig. 5, reinforcing the simulation results of Fig. 4. Additional sequencing efforts come at a cost, however, and so in the next subsection we use the methods of this paper to quantify the tradeoff between increased sequencing efforts and decreased uncertainty.

C. The cost of reduced uncertainty

In the previous section, the reanalysis of published results shows clearly that the number of samples per parasite has a dramatic impact on the uncertainty (and therefore the interpretability) of painstakingly collected parasite sequence data. Naturally, increasing the sampling rates, n_a and n_b, decreases the uncertainty in $\hat{s}$, our estimate of s (Figure 4). However, additional samples cost time, effort, and money. Complicating matters, generating additional *var* sequences may or may not increase n_a, since the previously sequenced *var* tags may be redundantly sequenced. Thus, there is a stochastic tradeoff between increased laboratory effort and decreased uncertainty about repertoire overlap, which we now calculate.

To obtain *var* tags, the DNA is PCR amplified using degenerate primers that are designed to universally capture all *var* genes with DBLα domains. This product is then cloned into a vector that allows single products to integrate, and these vectors are then transformed into bacteria and plated such that each colony contains one vector and one insert (see e.g. [21] for detailed methods, but see also [25] which uses a different pipeline based on next-generation sequencing). Therefore, among a large number of colonies, there are likely to be multiple colonies with the same *var* gene while some genes may not be covered by any colony. How many colonies should be separated and sequenced in order to get an accurate estimate of the repertoire overlap between two parasites? Put more formally, if we repeatedly perform an experiment in which we sequence c colonies each from two parasites and estimate their overlap $\hat{s}$, how much more accurate will $\hat{s}$ become if we increase c?

To answer this question, we split it into two parts. First, if we sequence c colonies, how many unique *var* genes n are we likely to have sampled? Second, what implications will this have for our repertoire overlap estimates, discussed in the previous section?

The first question can be answered by considering a process in which there are $k = 60$ distinct sequences in total and we draw c of them, one at a time, independently and with replacement. For a fixed c, we can compute the probability mass function for the number of distinct sequences by a straightforward recursion: At any point during the process of drawing sequences, if n distinct sequences have already been drawn, then the probability of drawing an already-discovered sequence is n/k, making the probability of drawing a new

sequence $1 - n/k$. Each draw is independent of the previous draws, so the incremental accumulation of distinct sequences can be written as a Markov chain with transition matrix π whose non-zero entries are

$$\pi_{n\to n} = \frac{n}{k} \quad \text{and} \quad \pi_{n\to n+1} = 1 - \frac{n}{k}\,. \tag{9}$$

Initially, zero sequences have been drawn ($c=0$), making $n=0$ with probability 1. For each additional sequence drawn, the probability distribution over the number of distinct sequences evolves according to the transition matrix π, so that after c draws the distribution over distinct sequences is given by the entries of the vector $\mathbf{x}$,

$$\mathbf{x} = \mathbf{x}_0^T \pi^c\,, \tag{10}$$

where $\mathbf{x}_0$ is initial condition vector of zeros, except for the entry corresponding to the state $n = 0$, which equals one. This allows us to analytically compute the distribution of the number of unique *var* genes sampled by a PCR process with c colonies. In other words, we now have a map between laboratory efforts c and the distribution of actual unique *var* genes sampled, and we write this as $P(n \mid c)$. A variant of this problem was previously considered with the goal of computing the value of c that would cover at least 60% of each repertoire [31]. Although those calculations can be shown to produce incorrect estimates, Eq. (10) can be used to solve that problem variant as well. More widely, this general problem has been charmingly named *the coupon collector's problem* by statisticians.

The second question focuses on the implications of Eq. (10), and specifically requires that we quantify how an increase in sequencing efforts c affects the noisy distribution of estimates $\hat{s}$. Intuitively, for low c, both n_a and n_b will tend to be small, leading to broad distributions of $\hat{s}$ around the correct value of s. Similarly, as c grows very large, we expect the distribution of $\hat{s}$ to concentrate on exactly s. This distribution, $P(\hat{s} \mid s, c)$, can be computed by integrating the distribution of estimates, conditioned on particular data, over the probability distribution of having produced those data, conditioned on c and s. Symbolically, the distribution of estimators $\hat{s}$, given true overlap s and colonies c is given by

$$P(\hat{s} \mid s, c) = \sum_{n_a, n_b, n_{ab}} \Big\{ P(\hat{s} \mid n_a, n_b, n_{ab}) \times P(n_{ab} \mid n_a, n_b, s) P(n_b \mid c) P(n_a \mid c) \Big\}\,. \tag{11}$$

$P(\hat{s} \mid n_a, n_b, n_{ab})$ is the probability of getting a particular estimate $\hat{s}$, given information about coverage and overlap. In fact, this is a distribution concentrated at a single point, i.e., a Dirac δ function, since each triple (n_a, n_b, n_{ab}) maps to exactly one point estimate $\hat{s}$. As a result, this term tells us the locations at which there will be probability mass, while the remaining terms in Eq. (11) tell us how much mass there will be at those locations. In other words, this distribution is a discrete probability distribution, and we have written down a fancy form of it above. By aggregating into bins, this distribution can be conveniently visualized as a histogram, which shows how the uncertainty of estimators depends on the true overlap s and the sequencing effort c. Figure 6 shows the effect of increasing sequencing efforts from a half plate ($c=48$) to a full 96-well plate ($c = 96$) and beyond. These calculations succinctly quantify intuition: additional laboratory efforts lead to higher accuracy guarantees.

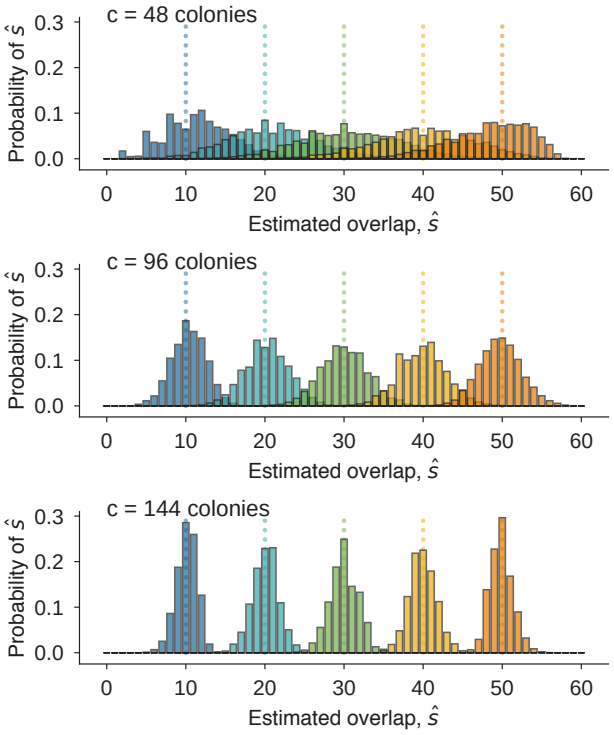

FIG. 6. **Quantifying the decrease in uncertainty from increased sequencing**. Histograms show distributions of overlap estimates $\hat{s}$, computed using Eq. (11), for various values of s which are indicated by color-matched dotted lines. While all estimates are distributed around the true values of s, increasing the number of colonies c from 48 (top) to 96 (middle) and to 144 (bottom) substantially decreases the error of estimates. For example the bottom plot shows that successfully sequencing $c = 144$ colonies from each parasite is guaranteed to produce estimates $\hat{s}$ that are off by at most 5 (8.3%) in either direction of the true s.

The calculations and distributions in this section show how the Bayesian framework in this manuscript can also be used to plan sequencing studies and estimate study costs. If a desired downstream analysis of repertoire overlap requires results that are accurate to within a particular number of shared sequences, BRO methods can easily specify the sequencing efforts needed.

IV. DISCUSSION

This manuscript places the estimation of overlap between fixed-size repertoires or populations from incomplete samples on firm statistical ground. While myriad indices of β-diversity for presence-absence data exist [2], they implicitly treat species counts as complete, leading to bias. Notable exceptions which embrace imperfect sampling exist [7], but require species abundance data to compute. Here, we clearly define a stochastic process for fixed-size repertoires that generates sample presence/absence data, opening the door to rigorous Bayesian inference. In particular, Eq. (7) provides point estimates of true repertoire overlap, while Eq. (8) provides error bars and uncertainty estimates via credible intervals. If desired, improved estimates of $\hat{s}$ can be plugged directly into any of the dozens of presence/absence measures of similarity reviewed in Ref. [2]. Figures 3 and 4 show the consistency and accuracy of these calculations across simulated sampling regimes in which the correct answer is known.

Bayesian repertoire overlap (BRO) is also useful in real-data scenarios, when the correct answer is unknown. By revisiting previously published studies of the *var* genes of *P. falciparum* [18, 19, 21, 27], we showed that switching from the Sørenson-Dice coefficient $\mathring{S}$ (called pairwise type sharing in the malaria literature) to BRO leads to different conclusions (Fig. 5 left, middle) or high uncertainty (Fig. 5 right). In particular, these reanalyses point to a clear recommendation for the design of future malaria studies: the number of unique *var* sequences per isolate should be at least 30. Since each additional PCR product may not contribute an additional unique sequence, we again used the Bayesian framework to translate increased PCR efforts to decreased uncertainty (Fig. 6). Accuracy requirements can now be weighed against laboratory costs during the planning of studies.

While BRO clearly outperforms $\mathring{S}$ in practical contexts, it is also more cumbersome to compute. Indeed, $\mathring{S}$ can be calculated on the back of an envelope while Eq. (7) requires a computer, or at least a lot more envelopes. However, as it turns out, there are only around $77,500$ possible combinations of n_a, n_b, and n_{ab}, which means that a lookup table of every conceivable $\hat{s}$ value can be computed on a laptop in minutes and attached to an email. Links to open-source code and a convenient web tool can be found in the Acknowledgements.

The models introduced in this paper are as correct as their assumptions, which we now revisit. During the construction of the Bayesian repertoire overlap, we assumed that our prior distribution $P(s)$ was uniform, meaning that we treated each possible level of overlap as equally likely. This is easily defensible in practice, as any other choice would introduce unacceptable bias.

We also assumed, when computing the tradeoff between sequencing effort and uncertainty, that each sequence in each repertoire was just as likely to have been sampled, which may or may not be true, for two distinct reasons. First, due to the fact that sequences are obtained using degenerate primers, the effects of primer bias may cause some sequences to be amplified more often than others. Second, a single parasite genome might have multiple copies of the exact same *var* gene, or might have distinct *var* genes whose DBLα tags are nevertheless identical. This scenario is arguably more likely among South American genomes whose overall *var* diversity is lower. Experimentally, the probability that a sequence is observed will be scaled upward by its genomic multiplicity, but the scaling may be non-linear since PCR protocols include many rounds of amplification, magnifying the deviations from uniformity. Fully addressing either of these possibilities would require that we modify the probabilities in both the coupon collector's problem and the repertoire subsampling processes, and then use Monte Carlo methods to numerically compute posterior distributions.

New sampling protocols for *var* genes, based on next-generation sequencing methods [25], may or may or may not meet the assumptions of the estimator presented here. To use the hypergeometric distribution, we require that, if an entire sampling protocol were to be technically replicated many times, that eventually each member of the repertoire would be observed with equal probability. In other words, while fluctuations in any particular set of observations are expected, with technical replication those fluctuations must eventually even out, approaching uniformity. Thus, the issues of primer bias and gene multiplicity violate the assumption of uniformity, but the magnification of random initial fluctuations, e.g., by the repeated amplification rounds of PCR, do not.

Could deviations from the modeling assumption of uniformity could be inferred from the data themselves? If so, this idea could in principle be applied to cloning-based methods and next-generation methods alike. This is an interesting direction for future work, and could draw from advances in abundance-based estimators for β-diversity [7, 12], or could incorporate explicit knowledge of the effects of protocols and pipelines, in order to mathematically undo their effects.

More practically, the assumption that the *var* repertoire size is 60 makes the methods of this paper less useful in the context of complex infections with multiple parasite genomes [23, 27]. In cases where the multiplicity of infection is known, overlap estimates could be computed using generalizations of the statistics in this paper, computing overlap between infections (instead of between parasites). This would be complicated by possible overlap of parasite repertoires within each infection, but repertoires tend to be quite different in areas of high transmission so the methods herein may be approximately correct. Nevertheless, development of more sophisticated methods would be especially useful in the context of *var*-based epidemiological studies.

This paper focuses on malaria's *var* genes, and assumes a total repertoire size of 60, but mathematically relaxing this assumption (Supplementary Materials) broadens the set of possible applications. First, within studies of malaria's *var* genes, the total repertoire size fluctuates slightly from parasite to parasite. As larger whole-genome datasets become available, this information can be incorporated directly as a prior over the distribution of repertoire sizes, improving estimates further. This opens the door to the analysis of *Plasmodium spp.*

multigene antigen families such as *rif* and *stevor* [32, 33], or more general studies of β-diversity in multigene families in which population sizes are fixed or their size distributions have been sampled [34]. Second, outside of malaria, improved estimators may also be useful in comparing, for instance, the genomic archives of antigen-encoding *vsg* genes used by *Trypanosoma brucei* for immune evasion [35]. Finally, the mathematics of this paper need not be applied to genetics or even within ecology; large-data applications like Facebook and other online social networks compute the sizes of intersections of sets—how many interests or friends do two individuals have in common?—but use only subsamples of data to decrease computation time in distributed computing settings [13].

Finally, this work presents a Bayesian approach to inference of β-diversity under particular assumptions, which contrasts the vast majority of indices, coefficients, and metrics to date which remain non-probabilistic [2]. Changing the underlying assumptions of the Bayesian repertoire overlap method, or the statistics of the sampling process, would lead to additional estimators for other common cases. Across applications, unbiased estimation combined with the quantification of uncertainty will allow for more reliable results and better prospective study design.

V. ACKNOWLEDGEMENTS

DBL was supported by the Ruth and Sidney Weiss Fund, and would like to thank Amy K. Bei, Samuel F. Way, Caroline O. Buckee, and Allison C. Morgan for helpful conversations and suggestions. DBL also warmly thanks the authors of Refs. [18, 19, 27] whose commitment to open science and methods means that their data were freely available for reanalysis.

Open-source code is freely available in Python at github.com/dblarremore/BayesianRepertoireOverlap. A web tool version produces estimates, credible intervals, and figures like Fig. 2 and is available at bro.colorado.edu.

[1] Whittaker RH. Vegetation of the Siskiyou mountains, Oregon and California. Ecological Monographs. 1960;30(3):279–338.
[2] Koleff P, Gaston KJ, Lennon JJ. Measuring beta diversity for presence–absence data. Journal of Animal Ecology. 2003;72(3):367–382.
[3] Barwell LJ, Isaac NJ, Kunin WE. Measuring β-diversity with species abundance data. Journal of Animal Ecology. 2015;84(4):1112–1122.
[4] Rao CR. Diversity and dissimilarity coefficients: a unified approach. Theoretical Population Biology. 1982;21(1):24–43.
[5] Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. Applied and Environmental Microbiology. 2005;71(12):8228–8235.
[6] Stirling A. A general framework for analysing diversity in science, technology and society. Journal of the Royal Society Interface. 2007;4(15):707–719.
[7] Chao A, Chazdon RL, Colwell RK, Shen TJ. A new statistical approach for assessing similarity of species composition with incidence and abundance data. Ecology Letters. 2005;8(2):148–159.
[8] Jaccard P. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bull Soc Vaudoise Sci Nat. 1901;37:547–579.
[9] Dice LR. Measures of the amount of ecologic association between species. Ecology. 1945;26(3):297–302.
[10] Sørensen T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Biol Skr. 1948;5:1–34.
[11] Anderson MJ, Crist TO, Chase JM, Vellend M, Inouye BD, Freestone AL, et al. Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist. Ecology Letters. 2011;14(1):19–28.
[12] Kéry M, Royle JA. Hierarchical Bayes estimation of species richness and occupancy in spatially replicated surveys. Journal of Applied Ecology. 2008;45(2):589–598.
[13] Ting D. Towards optimal cardinality estimation of unions and intersections with sketches. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM; 2016. p. 1195–1204.
[14] Avril M, Tripathi AK, Brazier AJ, Andisi C, Janes JH, Soma VL, et al. A restricted subset of var genes mediates adherence of Plasmodium falciparum-infected erythrocytes to brain endothelial cells. Proceedings of the National Academy of Sciences. 2012;109(26):E1782–E1790.
[15] Claessens A, Adams Y, Ghumra A, Lindergard G, Buchan CC, Andisi C, et al. A subset of group A-like var genes encodes the malaria parasite ligands for binding to human brain endothelial cells. Proceedings of the National Academy of Sciences. 2012;109(26):E1772–E1781.
[16] Ochola LB, Siddondo BR, Ocholla H, Nkya S, Kimani EN, Williams TN, et al. Specific receptor usage in Plasmodium falciparum cytoadherence is associated with disease outcome. PLOS One. 2011;6(3):e14741.
[17] Warimwe GM, Fegan G, Musyoki JN, Newton CR, Opiyo M, Githinji G, et al. Prognostic indicators of life-threatening malaria are associated with distinct parasite variant antigen profiles. Science Translational Medicine. 2012;4(129):129ra45–129ra45.
[18] Barry AE, Leliwa-Sytek A, Tavul L, Imrie H, Migot-Nabias F, Brown SM, et al. Population genomics of the immune evasion (var) genes of Plasmodium falciparum. PLOS Pathogens. 2007;3(3):e34.
[19] Albrecht L, Castiñeiras C, Carvalho BO, Ladeia-Andrade S, da Silva NS, Hoffmann EH, et al. The South American Plasmodium falciparum var gene repertoire is limited, highly shared and possibly lacks several antigenic types. Gene. 2010;453(1):37–44.
[20] Chen DS, Barry AE, Leliwa-Sytek A, Smith TA, Peterson I, Brown SM, et al. A molecular epidemiological study of var gene diversity to characterize the reservoir of Plasmodium falciparum in humans in Africa. PLOS One. 2011;6(2):e16629.

[21] Bei AK, Diouf A, Miura K, Larremore DB, Ribacke U, Tullo G, et al. Immune characterization of Plasmodium falciparum parasites with a shared genetic signature in a region of decreasing transmission. Infection and Immunity. 2015;83(1):276–285.

[22] Tessema SK, Monk SL, Schultz MB, Tavul L, Reeder JC, Siba PM, et al. Phylogeography of var gene repertoires reveals fine-scale geospatial clustering of Plasmodium falciparum populations in a highly endemic area. Molecular Ecology. 2015;24(2):484–497.

[23] Day KP, Artzy-Randrup Y, Tiedje KE, Rougeron V, Chen DS, Rask TS, et al. Evidence of strain structure in Plasmodium falciparum var gene repertoires in children from Gabon, West Africa. Proceedings of the National Academy of Sciences. 2017; p. 201613018.

[24] Buckee CO, Bull PC, Gupta S. Inferring malaria parasite population structure from serological networks. Proceedings of the Royal Society of London B: Biological Sciences. 2009;276(1656):477–485.

[25] He Q, Pilosof S, Tiedje KE, Ruybal-Pesantez S, Artzy-Randrup Y, Baskerville EB, et al. Networks of genetic similarity reveal non-neutral processes shape strain structure in Plasmodium falciparum. Nature Communications. 2018;9(1):1817.

[26] Taylor HM, Kyes SA, Newbold CI. Var gene diversity in Plasmodium falciparum is generated by frequent recombination events. Molecular and Biochemical Parasitology. 2000;110(2):391–397.

[27] Bull PC, Berriman M, Kyes S, Quail MA, Hall N, Kortok MM, et al. Plasmodium falciparum variant surface antigen expression patterns during malaria. PLOS Pathogens. 2005;1(3):e26.

[28] Bull PC, Kyes S, Buckee CO, Montgomery J, Kortok MM, Newbold CI, et al. An approach to classifying sequence tags sampled from Plasmodium falciparum var genes. Molecular and Biochemical Parasitology. 2007;154(1):98.

[29] Normark J, Nilsson D, Ribacke U, Winter G, Moll K, Wheelock CE, et al. PfEMP1-DBL1α amino acid motifs in severe disease states of Plasmodium falciparum malaria. Proceedings of the National Academy of Sciences. 2007;104(40):15835–15840.

[30] Warimwe GM, Keane TM, Fegan G, Musyoki JN, Newton CR, Pain A, et al. Plasmodium falciparum var gene expression is modified by host immunity. Proceedings of the National Academy of Sciences. 2009;106(51):21801–21806.

[31] Fowler EV, Peters JM, Gatton ML, Chen N, Cheng Q. Genetic diversity of the DBLα region in Plasmodium falciparum var genes among Asia-Pacific isolates. Molecular and Biochemical Parasitology. 2002;120(1):117–126.

[32] Cheng Q, Cloonan N, Fischer K, Thompson J, Waine G, Lanzer M, et al. stevor and rif are Plasmodium falciparum multicopy gene families which potentially encode variant antigens. Molecular and Biochemical Parasitology. 1998;97(1-2):161–176.

[33] Niang M, Yam XY, Preiser PR. The Plasmodium falciparum STEVOR multigene family mediates antigenic variation of the infected erythrocyte. PLOS Pathogens. 2009;5(2):e1000307.

[34] Otto TD, Gilabert A, Crellen T, Böhme U, Arnathau C, Sanders M, et al. Genomes of all known members of a Plasmodium subgenus reveal paths to virulent human malaria. Nature Microbiology. 2018;3(6):687.

[35] Marcello L, Barry JD. Analysis of the VSG gene silent archive in Trypanosoma brucei reveals that mosaic gene expression is prominent in antigenic variation and is favored by archive substructure. Genome Research. 2007;17(9):000–000.

S1 TEXT:
BAYES-OPTIMAL ESTIMATION OF OVERLAP BETWEEN POPULATIONS OF FIXED SIZE

How should one estimate the size of the intersection of two sets of arbitrary size from subsamples? As in the main text, assume that set a has total size N_a, and n_a objects are drawn from it uniformly at random. Similarly, assume that set b has total size N_b, and that n_b objects are drawn from it uniformly at random. Suppose that the number of objects found among the samples of sizes n_a and n_b is, as in the main text, n_{ab}.

Without loss of generality, assume that $N_a \leq N_b$. We make this assumption because the maximum value of the true overlap s is then N_a, since the two sets cannot have an intersection larger than the smaller of the two sets. The estimator $\hat{s}$ is given by

$$P(s \mid n_a, n_b, n_{ab}, N_a, N_b) = \frac{\sum_{s_a=0}^{N_a} P(n_{ab} \mid n_b, s_a, N_b) P(s_a \mid n_a, s, N_a)}{\sum_{s'=0}^{N_a} \sum_{s_a=0}^{N_a} P(n_{ab} \mid n_b, s_a, N_b) P(s_a \mid n_a, s', N_a)}, \tag{S1}$$

and

$$\hat{s} = \sum_{s=0}^{N_a} s \, P(s \mid n_a, n_b, n_{ab}, N_a, N_b), \tag{S2}$$

where we are now explicit about the total number of objects in the hypergeometric distributions—in the main text, these were implicitly 60. In other words, $P(x \mid t, u, v)$ is the hypergeometric probability of drawing exactly x special objects out of t draws, from a population of size v, in which there are u special objects total.

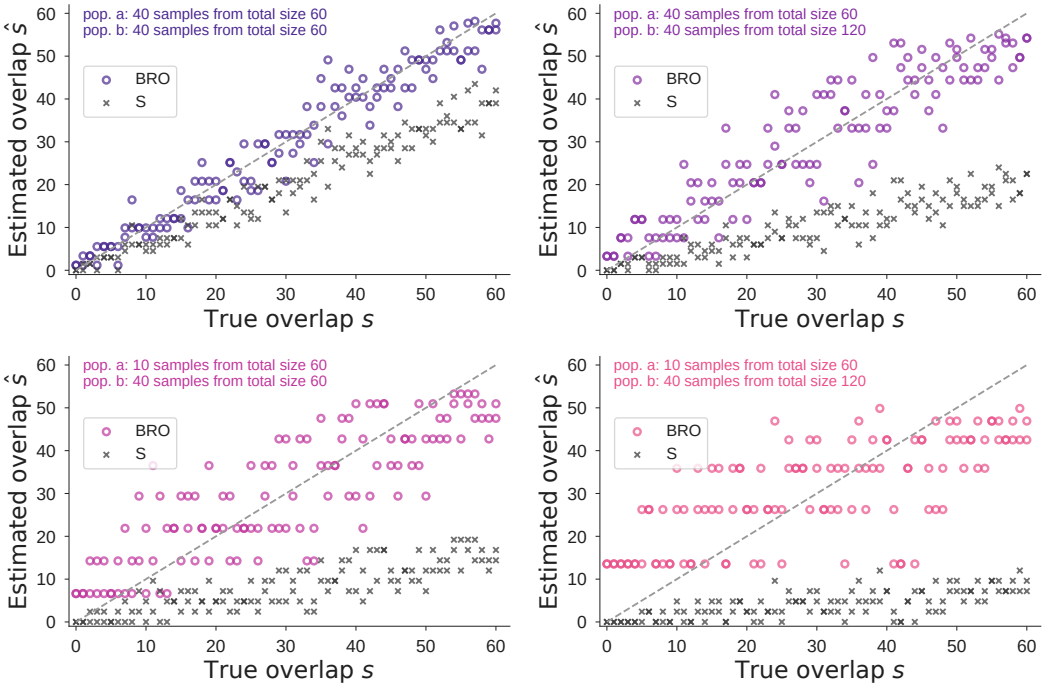

FIG. S1. **Bayesian repertoire overlap consistently estimates true overlap for varying population size and sampling rates.** Repertoires with true overlaps ranging from 0 to 60 were subsampled in simulations. While the main text shows results when $n_a = n_b$ and when $N_a = N_b = 60$, these assumptions can also be relaxed. Increasing N_b from 60 (left column) to 120 (right column) does not affect the consistency of BRO estimates, nor does decreasing the number of samples from population a from $n_a = 40$ (top row) to $n_a = 10$ (bottom row). As in the main text, the underestimating bias of $\hat{S}$ is worse with lower sampling rates [7].