Approaches to estimating inbreeding coefficients in clinical isolates of Plasmodium falciparum from genomic sequence data

Lucas Amenga-Etego[1,2], Ruiqi Li[3], and John D. O'Brien[3,*]

[1]Department of Mathematics, Bowdoin College, Brunswick, Maine, USA

[2]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK

[3]Navrongo Health Research Centre, Navrongo, Upper East Region, Ghana

[*]To whom correspondence should be addressed

ABSTRACT

A recent genomic characterization of more than 200 *Plasmodium falciparum* samples isolated from the bloodstreams of clinical patients across three continents further supports the presence of significant strain mixture within infections. Consistent with previous studies, these data suggest that the degree of genetic strain admixture within infections varies signficantly both within and across populations. The life cycle of the parasite implies that the mixture of multiple genotypes within an infected individual controls the outcrossing rate across populations, making methods for measuring this process *in situ* central to understanding the genetic epidemiology of the disease. Peculiar features of the *P. falciparum* genome mean that standard methods for assessing structure within a population – inbreeding coefficients and related F-statistics – cannot be used directly. Here we review an initial effort to estimate the degree of mixture within clinical isolates of *P. falciparum* using these statistics, and provide several generalizations using both frequentist and Bayesian approaches. Using the Bayesian approach, based on the Balding-Nichols model, we provide estimates of inbreeding coefficients for 168 samples from northern Ghana and find significant admixture in more than 70% of samples, and characterize the model fit using posterior predictive checks. We also compare this approach to a recently introduced mixture model and find that for a significant

minority of samples the F-statistic-based approach provides a significantly better explanation for the data. We show how to extend this model to a multi-level testing framework that can integrate other data types and use it to demonstrate that transmission intensity signficantly associates with degree of structure of within-sample mixture in northern Ghana.

INTRODUCTION

The protozoan parasite *Plasmodium falciparum* causes most cases of severe malaria and presents one of humanity's most significant public health burdens, killing at least half a million people a year [16, 39]. The parasite's ability to develop resistance to drugs and the rapid spread of that resistance across geographically-separated populations presents a constant threat to international control efforts [23, 27]. While genetic factors play a crucial role in resistance, many aspects of the genetic epidemiology of the parasite remain obscure [33, 35]. The beginnings of a global perspective on the genetic structure of this problem emerged from the analysis of whole-genome sequencing (WGS) data derived from $\sim$ 200 parasite genomes collected directly from clinicial patients in six countries on three continents [20]. This study provides strong evidence for signficant continental-scale structure in the amount of genetic variation present within populations, as well as indicating frequent but variable amounts of within-isolate strain mixture, often referred to as multiplicity of infection (MOI) [8].

Within-isolate *P. falciparum* strain mixture may result from a host individual being infected by a mosquito carrying several distinct strains of the parasite, being infected sucessively by several mosquitos carrying single strains, or some combination of the two. The parasites's sexual mating process occurs only in the period immediately after being taken up into the mosquito hindgut as a bloodmeal implying that mixed infections – either within the host or vector – provide the essential grist for the maintainance of population-level genetic diversity by creating an opportunity for outcrossing [15, 7]. The degree of apparent mixture within an infected individual's bloodstream then would largely depend on the number of infective events and the effective population size of the surrounding *P. falciparum* population [3]. Other effects, such as genetic interactions with the vector, density-dependent selection, non-random mating and host immune response, may also play a role. Consequently, reliable techniques for measuring the structure of mixture within clinical infections will give researchers a quantitative measurement of potential outcrossing – a key determinant of

genetic epidemiological control – and may provide insight into other important biological effects, such as host immune response. This paper presents a collection of statistical methods for capturing the amount of mixture within clinical isolates using WGS data.

Historically, the *P. falciparum* research focuses on MOI, the minimal number of distinct genetic strains identifiable in a clinical sample. Researchers have largely used microsatellite, RFLP, or SNP data to infer the number of strains [8, 1, 3, 18] and have associated these values with morbidity [21, 26], patient age [9, 32], the course of infection in pregnancy [31], the presence of bed nets [10], and a number of other conditions [6]. The focus on the number of strains naturally follows from the limits of earlier genotyping techniques since more subtle features, such as the mixture proportions of strains or recombination events among them, could not be readily identified.

The introduction of WGS data naturally enforces a generalization from MOI into more complex measures of mixture, since, in a strict sense, we find find evidence for some level of MOI in nearly all of our samples. For instance, we present four samples in Figure 1 with within-sample observed allele frequency plotted against the population-level allele frequency for all relevant SNPs. The left plot shows little apparent mixture. Previous technologies limited measurement of the number of genetic types to small regions of the genome, effectively bounding the maximum number of strains that could be identified. This has led to considerations of complexity of infection (COI), where a strain mixture model is used to explain these within sample mixture levels [11], expanding on traditional methods for inferring MOI [17]. Below, we show below that this model, while powerful in modeling patterns within some isolates, is not sufficient to explain observed mixture within all samples.

F-statistics, measurements of the departure of allelic heterozygosity observed within a population from those expected at Hardy-Weinberg equilibrium (HWE), make for a natural approach to quantifying the structure within mixed samples [37, 24]. HWE specifies the distribution of alleles assuming panmixia, a population exhibiting perfectly random mating with an absence of mutation, migration, drift, selection or other effects [40]. F-statistics calibrate the empirical allele distribution within a subpopulation against those expected under HWE, ranging from a value of one (no mixture) to zero (perfect mixture). In the context of comparing the parasites' genetic diversity within a single infected individual relative to the local geographic population (and absent any geographic structuring of the population, i.e. the Wahlund effect), these statistics effectively become inbreeding

coefficients. For a sample, i, we refer to this value as f_i.

These methods have proven to be an effective and extremely popular means for investigating species' population structure from both allelic and genomic data [37, 30, 38]. However, standard methods assume specific ploidy structures incommensurate with WGS data from *P. falciparum* and so cannot be used directly. The critical difference is that, within a human host, *P. falciparum* exists only in the haploid stage of its life-cycle [15]. Since short read WGS data cannot yet capture full haplotypes, individual reads cannot be uniquely identified with their strain of origin. Without being able to associate reads to individual *P. falciparum* strains, we cannot see any 'out-of-the-box' use of standard F-statistics approaches with this new data.

[20] provides an initial estimator for inbreeding coefificents using WGS based on the slope of a modified regression line between the expected heterozygosity assuming population-level HWE and the observed heterozygosity within a sample. [3] explores the connection between this estimator and standard MOI approaches by comparing estimates from WGS with MOI values inferrred by genotyping the *msp-1* and *msp-2* genes, showing strong correlation between these values in their sample sets. They note that the correlation is strongest at high f_i and low MOI values, where samples are close to being unmixed, with weaker correlation among more mixed samples, suggesting increasing divergence between these models. This estimator, while providing an effective first effort, does not clearly connect to the larger statistical tradition around F-statistics. This paper seeks to clarify this estimator by placing it more firmly within this larger discussion.

This paper proceeds as follows. First, we provide an overview of data collection and layout our notation. We present the initial estimator employed by [20] for estimating f_i and provide two additional frequentist estimators and detail their connection to classical F statistics. We then proceed to describe a Bayesian approach for these statistics that builds on the Baldings-Nichols model together with an inference scheme and framework for hierarchical modeling. We use this construction to show that observed transmission intensity, a measure of the amount of infective mosquito activity in the surrounding environment, signficantly associates with changes in mixture among 168 norhtern Ghanian samples. We then show that, in comparison with a COIL-like approach [11], the Bayesian F-statistic is a more powerful explanatory model for a substantial fraction of samples. We conclude with a discussion of the strengths and limitations of our approaches, and possible future directions for modeling within-sample mixture using WGS.

1 Model and Data

1.1 Data and preparation

The WGS data come from Illumina HiSeq sequencing applied to *P. falciparum* extracted from 235 clinical blood samples collected from infected patients from the Kassena-Nankana district (KND) region of northern Ghana. Collection occurred over approximately 2 years, from June 2009 to June 2011. The full sequencing protocol and collection regime are described in [20, 2]. After quality control measures, sequencing was performed on 235 samples, and, following a documented protocol using comparison against world-wide variation, 198,181 single-nucleotide polymorphisms (SNPs) were called within each sample [20]. Each call provides the number of reference and non-reference read counts observed at each variant position within the genome, mapped to the 3D7 reference [12]. Positions that exhibited no variation within the KND samples, any level of missingness (no read counts observed), or minor allele frequency less than 0.05 were excluded. Samples that possesed more than 4000 SNPs called with fewer than 10 read counts were also excluded, following an observed inflection point. These cleaning measures left 1470 SNPs in 168 samples. We observe little apparent population structure among the samples, evidenced either principal components analysis or a neighbor-joining tree of the pairwise difference among samples, as in Supplementary Figure S1.

1.2 Notation

We label the samples $i = 1, \cdots, N$ and the SNPs by $j = 1, \cdots, M$, with $N = 194$ and $M = 1,470$ if all samples and all SNPs are considered. In some contexts below M may be the number of SNPs within a chromosome, which should be clear by context. At SNP j within sample i, we observe r_{ij} reads that agree with the reference, and n_{ij} reads that are different from the reference. We write p_{ij} for the allele frequency for reference allele for SNP j in sample i and estimate it via the maximum-likelihood estimator (MLE) for proportions: $\hat{p}_{ij} = \frac{r_{ij}}{r_{ij}+n_{ij}}$. Similarly, we write p_j as population-level reference allele frequency for each SNP and estimate according to the across-sample MLE:

$$\widehat{p}_j = \sum_{i=1}^{N} n_{ij} \Big/ \sum_{i=1}^{N} (r_{ij} + n_{ij}).$$

To slightly streamline our notation, we relabel the inbreeding coefficient, F_{is}, for each sample i as f_i. We provide Table 1 as a reference to the reader for notation.

1.3 A previous estimator for f_i, and two alternatives

In [20], the authors provide an initial approach to estimating f_i. We refer to this estimator as $f_i^{(m)}$ to contrast it with subsequent estimators. This method relies on minor allele frequencies rather than reference allele frequency, which we mark with a tilde so that p_j becomes $\widetilde{p}_j$. The two quantities are naturally related so that $\widetilde{p}_j = p_j$ if $p_j < 0.5$ and $\widetilde{p}_j = 1 - p_j$ otherwise. $\widetilde{p}_{ij}$ and p_{ij} are related in the same fashion and we continue to use hats to denote estimates. The estimator $f_i^{(m)}$ proceeds sample by sample, so we will consider a generic sample i. The estimator first partitions alleles into 11 equally-spaced bins based on their minor allele frequency: $(0, 0.05), \cdots, (0.45, 0.50)$. Within each bin, b, the averaged expected heterozygosity assuming population-level HWE is calculated by

$$H_e(b) = \frac{1}{M_b} \sum_{k \in b}^{M_b} 2 \cdot \hat{\widetilde{p}}_k \cdot (1 - \hat{\widetilde{p}}_k),$$

where M_b is the number of SNPs within bin b. The averaged observed heterozygosity within each bin and each sample is calculated by

$$H_o(b, i) = \frac{1}{M_b} \sum_{k \in b}^{M_b} 2 \cdot \hat{\widetilde{p}}_{ik} \cdot (1 - \hat{\widetilde{p}}_{ik}).$$

Finally, $\hat{f}_i^{(m)}$ is calculated as $1 - \beta$ where β is the slope found by regressing the $H_o(b, i)$ values against H_b^e values centered within their respective allele frequency bins and constrained to pass through the origin.

The binning procedure stablizes the estimator against influence by the low frequency alleles that dominate the samples. Consequently, this has the result of biasing the estimates towards high frequency alleles. We can remove this effect by discarding the binning procedure in favor of directly regressing observed heterozygoity for each SNP against the expected value, still constrained to pass through the origin. This provides a closed for expression for a new estimator, $f_i^{(r)}$, as

$$\hat{f}_i^{(r)} = 1 - \frac{\sum_{j=1}^{M} \hat{\widetilde{p}}_j \cdot (1 - \hat{\widetilde{p}}_j) \cdot \hat{\widetilde{p}}_{ij} \cdot (1 - \hat{\widetilde{p}}_{ij})}{\sum_{j=1}^{M} \hat{\widetilde{p}}_j^2 \cdot (1 - \hat{\widetilde{p}}_j)^2}.$$

We can also create a similar estimator but one more transparently derived from the ideas underpinning traditional F-statistics in the following way. For a single SNP j, suppose f_i to be the

fraction of the population-level heterozygosity equal to the difference between the population-level heterozygosity, H_j^p and the sample-level heterozygosity, H_j^i that is,

$$f_i \cdot H_j^p = H_j^p - H_j^i.$$

Dividing through by H_j^p gives an estimate for f_i for the SNP j. Averaging across all SNPs, and taking the ratio of expections to be the expectation of the ratios, gives the estimator

$$\hat{f}_i^{(d)} = 1 - \frac{\sum_{j=1}^{M} \tilde{p}_{ij}(1-\tilde{p}_{ij})}{\sum_{j=1}^{M} \tilde{p}_j(1-\tilde{p}_j)}.$$

For each of these estimators, a corresponding variance calculation is possible. For $f_i^{(i)}$ and $f_i^{(r)}$ these can be made by recoursing to known properties of regression lines. For $f_i^{(d)}$, a delta approximation can be used. However, we instead employ a more convenient bootstrap approach to capture the variance in the estimates for confidence intervals. For the Bayesian estimates presented below, we can establish credible intervals based on the inferred posterior distribution.

Figures 1 and Table 2 compare the f_i estimates made by the initial, regressed and direct estimators. In Figure 1, we present the estimates for four samples, together with the SNP data and the binned values from the initial estimator. For the direct estimator, we construct the line shown by connecting the origin with the (x, y) point of the denominator and numerator of Equation 1.3. The other two estimators' lines come naturally from their regression procedure. The slope of each line is $1 - f_i$ for that estimator. As shown in Table 2, the correlation of the three estimators is greater than 0.98. In particular, the direct and regressed estimates differ by at most 1% across all samples. The initial estimator produces values that are almost invariably slightly lower than the two other estimates, by as much as 15% of the higher value for highly mixed samples.

Despite these differences, these estimators provide strongly consistent portraits of the f_i values for the samples in our data. However, they all possess two less-than-desirable properties: they rely on a separate estimate of the allele frequency; and cannot be easily incorporated into a more involved analysis for use in hypothesis testing. Researchers will presumably seek to use estimates of f_i as a means of testing clinical or epidemiological differences between subpopulations. A preferable approach would simultaneously allele frequency across all SNPs and the inbreeding coefficient for

each sample, as well as permitting extension to more complex modeling contexts. We submit that our Bayesian models below satisify these requirements.

1.4 Bayesian model framework

We present two models, with the second as a multi-level extension of the first. In the first, we estimate inbreeding coeffcients comparable to the above estimators but employing the Baldings-Nichols model [4]. In the second model, we show how we can exploit the more flexible Bayesian approach to estimate these values inside of a nested structure that allows us to test how different transmission regimes affect inbreeding coefficients. In both cases, we make several simplifying assumptions. We treat SNPs as being unlinked (i.e. no linkage disequilibrium) and assume that individual parasites within a sample respresent a random sample of the surrounding population. We also assume that read counts are sampled identically, independently, and represent an unbiased sample of variation at each position. We will discuss the evidence for and against these assumptions and possibilities for modeling extensions in the discussion.

1.4.1 Likelihood and priors

Our approach adapts the Balding-Nichols model of allele frequency within inbred subpopulations to the specific context of *P. falciparum* WGS data [5]. In *P. falciparum* the relevant subpopulation is the collection of parasites within a clinical sample. For sample i and SNP j, we assume that each read count arises as an identical and independent Bernoulli process with the probability of a reference read given by the unobserved reference allele frequency p_{ij}. Conditional upon an inbreeding coefficient f_i and a population-level allele frequency p_j, the Balding-Nichols model gives the allele frequency p_{ij} as a Beta distribution:

$$p_{ij} \sim \mathcal{B}\left(\frac{1-f_i}{f_i}p_j, \frac{1-f_i}{f_i}(1-p_j)\right).$$

Since the read counts are assumed to be i.i.d, p_{ij} is drawn from a Beta, and the probability of the data is binomial, we use the conjugancy of these distributions to eliminate the dependence on the unknown p_{ij} and give a Beta-binomial distribution for the likelihood:

$$\mathbb{P}(r_{ij}, d_{ij}|p_j, f_i) = \binom{r_{ij}+n_{ij}}{n_{ij}} \frac{\mathrm{B}(n_{ij}+\frac{1-f_i}{f_i}p_j, r_{ij}+\frac{1-f_i}{f_i}(1-p_j))}{\mathrm{B}(\frac{1-f_i}{f_i}p_j, \frac{1-f_i}{f_i}(1-p_j))}, \quad (1)$$

where $\mathrm{B}(\cdot\,,\,\cdot)$ is the beta function. Since we assume independence by site and by sample, the complete likelihood of the data, $\mathcal{D}$, conditional upon the inbreeding coefficients for all samples, $\mathbf{f} = (f_1, \cdots, f_N)$ and the allele frequency for all SNPs $\mathbf{p} = (p_1, \cdots, p_M)$ becomes

$$\mathbb{P}(\mathcal{D}|\mathbf{f}, \mathbf{p}) = \prod_{i=1}^{N}\prod_{j=1}^{M} \mathbb{P}(r_{ij}, r_{ij}|f_i, p_j).$$

In this first model, where we seek to estimate only the inbreeding coefficients for a set of samples, prior specification is straight-forward. The only prior information we have about the f_i values suggests that high levels of inbreeding are common but not obligatory in west African populations, and we quantitatively interpret this as a uniform prior on each f_i. We place a uniform prior on each allele frequency, although we have eliminated rare variants as part of data cleaning described in Section 2.1.

1.4.2 A hierarchical extension

For nearly all samples we posssess additional metadata on the assessed transmission intensity (TI) in the KND area at the time of *P. falciparum* sample collection. Field researchers categorize TI as low, medium or high based on the percevied probability of infection from observed mosquito counts, temperature, precipitation, and number of malaria cases entering area clinics. We write c_i for the TI of sample i, with $c_i \in \{0, 1, 2, 3\}$, with 0 representing no record and $1, 2,$ and 3 denoting low, medium and high transmission, respectively. The collection of all c_i's we write as $\mathbf{c}$.

We extend the previous model to also model the relationship between the distrbution of inbreeding coefficients and TI by constructing a model of the inbreeding coefficients in terms of the $\mathbf{c}$. Conditional upon c_i, we assume that each f_i is drawn independently from a Beta distribution with parameters α_{c_i} and β_{c_i}. There are consequently four α values and four β values and we label the vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. We then decompose the posterior distribution of the unobserved parameters conditional upon the read count data and TI values by noting that

$$\begin{aligned}\mathbb{P}(\mathbf{f}, \mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\beta}|\mathcal{D}, \mathbf{c}) \quad \propto \quad & \mathbb{P}(\mathcal{D}, \mathbf{c}|\mathbf{f}, \mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \cdot \mathbb{P}(\mathbf{f}, \mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \\ & \mathbb{P}(\mathcal{D}|\mathbf{c}, \mathbf{f}, \mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \cdot \mathbb{P}(\mathbf{c}, \mathbf{f}, \mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\beta}). \qquad (2)\end{aligned}$$

As the introdcution of the additional structure does not affect the probability of the read count data, we retain the same likelihood as in Equation 3.

The dependency of $\mathbf{f}$ on $\mathbf{c}, \boldsymbol{\alpha}$, and $\boldsymbol{\beta}$ we specified above. Together with the assumption that the remaining parameters are otherwise independent of each other these facts decompose Equation 2 to

$$\mathbb{P}(\mathbf{f}, \mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\beta} | \mathcal{D}, \mathbf{c}) \propto \mathbb{P}(\mathcal{D}|\mathbf{f}, \mathbf{p}) \cdot \mathbb{P}(\mathbf{f}|\mathbf{c}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \cdot \mathbb{P}(\mathbf{p}) \cdot \mathbb{P}(\mathbf{c}) \cdot \mathbb{P}(\boldsymbol{\alpha}) \cdot \mathbb{P}(\boldsymbol{\beta}).$$

It remains to specify the four prior terms on the right-hand side of the equation. We assume that the prior distribution should be the same as in the previous model. We take the observations of $\mathbf{c}$ to have probability one since they are the researchers' own assessment technique. For $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, we assume that they are drawn independently from an exponential distribution with mean one, except α_0 and β_0, corresponding to the unrecorded category. For those parameters we fix the values to one, ensuring a uniform prior on the corresponding f_i's.

1.4.3 Inference

We use a Metropolis-Hastings Markov chain (MCMC) approach to inference. The Metropolis-Hastings ratio gives the probability that a proposed parameter update x' will be accepted from a current state x with probability α such that

$$\alpha = \min\left(\frac{\mathbb{P}(x')}{\mathbb{P}(x)} \cdot \frac{\mathbb{P}(x' \to x)}{\mathbb{P}(x \to x')}, 1\right) = \min\Big(\alpha_1 \cdot \alpha_2,\ 1\Big).$$

The first ratio is that the posterior probabilities of x and x', and we write this as α_1. The second ratio gives probability of choosing the current state from the proposed state over the reverse move and we write this α_2. Since α_1 constitutes assessment of the likelihood and the prior functions that can be calculated directly from the likelihood and priors above, we subsequently only consider α_2. We denote proposed parameters with an apostrophe.

f - We randomly select i and propose f_i from $\mathcal{B}(\alpha_{c_i}, \beta_{c_i})$, leading to $\alpha_2 = \frac{\mathcal{B}(f_i|\alpha_{c_i},\beta_{c_i})}{\mathcal{B}(f_i'|\alpha_{c_i},\beta_{c_i})}$.
p - We randomly select j and then draw the proposed parameter p_j from the uniform prior, leading to $\alpha_2 = 1$.
$\boldsymbol{\alpha}, \boldsymbol{\beta}$ - For both of these parameters, we randomly select individual components and propose new values directly from the prior distribution, leading to $\alpha_2 = \frac{\exp(-x)}{\exp(-x')}$ where x and x' are the current and proposed state of the relevant component.

We examined the autocorrelation of the log-posterior and find that the lag was minimal (Supplementary Figure S2). As a secondary check, we ran chains both for all of the chromosomes

individually, as well as all together. Since we treat SNPs as independent, the performance of the model should be unaffected if indeed the model performs similarly across chromosomes. We find that across all chromosomes performance is nearly identical, with greater than 95% correlation among estimates.

1.5 Implementation

All code was implemented in the R computational environment [28]. The set of scripts implementing each of the estimators, the MCMC algorithm, and visualizations, together with subsets of the data sets are available at github.com/jacobian1980/pfmix. This repository includes a manual and workflow for completing analyses using these approaches. All materials are released under a Creative Commons License.

2 Results

2.1 Comparison with frequentist estimates

In Table 2 we compare between the Bayesian model estimates and each of the frequentist estimators for each of the 168 samples, observing strongly consistent estimates for all samples. We note that the least consistency among estimates on highly mixed samples, although still strongly similar. The Bayesian estimates are noticeably more distinct from the frequentist estimates than they are from each other. In general, the Bayesian estimates are less extreme than their frequentist comparators, likely due to mitigated influence from low-frequency variants. We also observe consistent estimates of variation among the two estimation procedures. For the Bayesian estimate, variation was determined using the maximum *a posteriori* (MAP) parameters and the properties of the Beta-binomial distribution. For the frequentist estimators, we employed 250 bootstrap samples on the set of SNPs.

2.2 Temporal changes in Northern Ghana

We describe above additional metadata collected at the time of the samples, the perceived tranmission intensity (PTI). This categorical data gives a measure of the frequency of malaria transmission and incorporates the rate of malaria cases presenting at the clinic, amount of standing water, irrigation status, and other factors. Given the role that transmission intensity is believed to play

in the process of outcrossing, researchers may naturally hypothesize that the f_i value of samples is partially determined by PTI. Similar effects have been reported in a variety of investigations [14, 29].

We plot the 95% credible interval of $1 - f_i$ values for each sample across the two year period of collection with the PTI coded by the color (Figure 2). We find evidence for significant mixture ($f_i > 0.95$) in more than 70% of samples (119/168). We plot $1 - f_i$ to show highly mixed sample as having high $1 - f_i$ values. The plot suggests that the PTI at samples collection corresponds to f_i, with high PTI yielding low values, low PTI giving high values, and moderate PTI somewhere in between. In a frequentist framework, this hypothesis could be tested either by pairwise comparison of means or by ANOVA. Grouping samples by PTI, we plot the distribution of MAP f_i values in Figure 3(upper right), noting that there appears some difference in distribution across the groups. However, pairwise comparison of the mean or variance between groups indicates no difference among the groups, even at a liberal signficance level of 0.1.

In Figure 2, we see the distribution of α and β values for each PTI suggests the posterior distribution differs between categories, although in a more complex way than a simple shift in mean. Our hierarchical model allows us to test this hypothesis in a different fashion, using Bayes factors, a measure of the support provided in the data for comparable models [19]. We can consider the hypothesis as a comparison between two models that we label M_0 and M_1. Under M_0, the α and β values are equal for all categories of PTI, i.e. $\alpha_1 = \alpha_2 = \alpha_3$ and $\beta_1 = \beta_2 = \beta_3$. This describes the sitatuation where the distribution for f_i is constant across PTI categories. Under M_1, the α and β values are unconstrained, and the f_i distribution may vary by PTI class. Notice that M_0 is nested within M_1. Because of this, we may use the Savage-Dickey ratio to calculate the Bayes factor. Using the methods set forth in [36], we calculate the Bayes factor using a standard kernel density estimator, as in [34]. The Bayes factor is 11.52, indicating M_1 provides a a moderately preferable explanation for the data relative to M_0.

2.3 Comparison with COI model

A recently introduced mixture model attempts to model within-sample allele frequency variation, in keeping with the MOI tradition within Pf genetics [11, 17]. To contrast this model with the one presented here, we use the BIC to compare model fit between the two for each of the 168 clinical

samples. Unfortunately, the implementation of this model was not designed for WGS data, so we rely on a reduced version of the model in [25] that amounts to an equivalent model, detailed in the Appendix. For each sample,

2.4 Posterior predictive assessment

While F statistics provide a convenient way to summarize the degree of heterogeneity in a clinical sample, researchers may also be interested in the degree that the model captures the complexities of the biological mixture process. We examine this discrepancy using posterior predictive checks (PPC) [22, 13]. PPCs measure the discrepancy between predictive data and the observed data by some discrepancy measure, for which here we take as a χ^2 statistic. For each point in the posterior, we generate a realization of data from that model. By sampling from the posterior and generating data for each sample, we create a predictive data distribution, y^{pred}. We then use the χ^2 statistic to generate a p-value comparing the observed data. For each SNP, we also plot the distribution of predicted SNP data versus and the observed value, across the allele frequency (see Figure 4 for examples).

The PPCs indicate that the model performs best for nearly unmixed samples and a subset of highly mixed samples, where fit appears strong. For a majority of samples, the fit is reasonable for a large section of SNPs but poor for a noticeable subset of variants. The PPCs also indicate that a zero-inflation in the data is not fully accounted for within the model. Taken across all samples, this suggest that the F-statistic model is insufficient to fully capture the within-sample heterogeneity. However, the model provides strong fit, better even than the COI model above, for a certain subset of samples and SNPs, indicating that a similar admixture process likely contributes to observed data patterns.

3 Discussion

This work presents a number of related approaches to inferring inbreeding coefficients, and connects them to an extensive body of research on multiplicity of infection (MOI) in *P. falciparum* suggesting the importance of MOI in characterizing the epidemiology of malaria. We provide the attendant code and workflows in an open-source platform for other researchers to implement these methods.

In developing the model, we make a number of assumptions about the underlying structure of

the read count data and the biological mixing process that may affect our inference. For the read count data, we assume that read counts are unbiased and the SNPs are unlinked. While short read data can be biased in several ways, previous research indicates that mixture proportions calculated by ratios of read counts is largely unbiased [20]. However, *P. falciparum* exhibits significant linkage disequilibrium on scales significantly larger than the average distance between neigboring SNPs in our data. While we do not expect this violation to bias our estimates as this absence of independence likely occurs roughly evenly across the genome.

A more troublesome assumption is embedded in the underlying structure of the F-statistic. An F-statistic measures the departure of the observed number of heterozygotes relative to those expected under Hardy-Weinberg equilibrium. In the context of mixed *P. falciparum* infections, the equilibrium assumptions – random mating, no selection, large population size, genetic isolation – are likely each violated at some level. For example, the mixture within a sample may be the result of a small number of founding individuals or be strongly selected by the human immune system. Without a more general approach to understanding the mixing process, we cannot anticipate the robustness of our estimates to this sort of misspecification. While looking at the SNP plots (e.g. Figure **??**) indicate that the f_i values inferred do correspond qualitatively to their apparent degree of mixture, the PPC analysis suggests that the model does not always capture the data's full complexity.

We suspect that as the genomic data enables more elaborate statistical models for mixed infections to develop that these considerations will become increasingly important to the biological community. In our presentation of PPCs, we only discuss model fit in a statistical fashion although there may be biological implications as well. Genes or regions of the genome that are either more or less mixed relative to the levels observed in the remainder of the genome could indicate either positive or negative selective pressure from the host immune system, intraspecific competition, or other processes. Examining the PPCs produced we find no strong indications of these effects. This may be because there is no signal to be discovered, or because the underlying model is too simple to allow these distinctions to emerge or the signature is not apparent without more involved statistical approaches. The COIL approach and other recent work [11, 25] indicate that strain-based mixtures may be a complementary line of inquiry. However, the substantial minority of samples for which the inbreeding model did provide the most powerful explanation, strongly suggests that considerations of inbreeding or similar processes will have to be included in the next generation of

statistical models.

Appendix: COI inference

Following the approach in [11], we implemented a finite mixture model for the data. As the method described there does not easily accomodate the amount of data in our samples ($\sim 1,500$ SNPs), we relied on a reduced version of the model presented in [25]. As in [11], the model presumes that a finite number of strains K give rise to 2^K 'bands' of the within-sample allele frequency owing different combinations of the present strains. Following the presentation in [25], for SNP j within sample i showing read counts (r_{ij}, n_{ij}), the within-sample allele frequency within band r is given by

$$q_{ijr} \quad = \quad \sum_{k=1}^{K} w_k \cdot \mathbf{1}_{\{s_k \in r\}},$$

where w_k is the sample proportion for strain s_k and $\mathbf{1}_{\cdot}$ is an indicator function. Supposing that read counts are *i.i.d.* conditional upon their band of origin, this leads to Beta-binomial likelihood given r,

$$\mathbb{P}(n_{ij}, r_{ij} | r, q_{ijr}, \nu) \quad = \quad \binom{n_{ij} + r_{ij}}{n_{ij}} \cdot \frac{\mathrm{B}(n_{ij} + q_{ijr} \cdot \nu, r_{ij} + (1 - q_{ijr}) \cdot \nu)}{\mathrm{B}(q_{ijr} \cdot \nu, (1 - q_{ijr}) \cdot \nu)}, \tag{3}$$

where B is the beta function and ν is an inverse variance parameter. Assuming no population structure within the local population, we can then write the probability of a SNP being in band r as binomial random variable with C_r being the number of non-reference allele states present in band r, that is, $\mathbb{P}(\text{SNP } j \in \text{ band } r | p_j) = p_j^{C_r} \cdot (1 - p_j)^{2^K - C_r}$. By summing over all bans, we get a likelihood independent of r,

$$\mathbb{P}(r_{ij}, n_{ij} | q_{ij.}, p_j, \nu, K) \quad = \quad \sum_{r=1}^{2^K} p_j^{C_r} \cdot (1 - p_j)^{2^K - C_r} \cdot \mathbb{P}(n_{ij}, r_{ij} | r, q_{ijr}, \nu).$$

Assuming independence across SNPs yields a product over j as the full data likelihood. Inference is performed in a Bayesian fashion using standard MCMC approaches, detailed in [25].

Competing interests

The authors declare that they have no competing interests.

Author's contributions

JO'B designed and implemented the study and wrote the manuscript. RL provided visualization of the data and model results. LA-E collected the data, contributed to the study design, and edited the manuscript.

Acknowledgements

We are grateful for many helpful discussions with Jason Wendler.

References

[1] Timothy JC Anderson, Bernhard Haubold, Jeff T Williams, Jose G Estrada-Franco, Lynne Richardson, Rene Mollinedo, Moses Bockarie, John Mokili, Sungano Mharakurwa, Neil French, et al. Microsatellite markers reveal a spectrum of population structures in the malaria parasite *Plasmodium falciparum*. *Molecular Biology and Evolution*, 17(10):1467–1482, 2000.

[2] Sarah Auburn, Susana Campino, Taane G Clark, Abdoulaye A Djimde, Issaka Zongo, Robert Pinches, Magnus Manske, Valentina Mangano, Daniel Alcock, Elisa Anastasi, et al. An effective method to purify *Plasmodium falciparum* dna directly from clinical blood samples for whole genome high-throughput sequencing. *PLoS ONE*, 6(7):e22213, 2011.

[3] Sarah Auburn, Susana Campino, Olivo Miotto, Abdoulaye A Djimde, Issaka Zongo, Magnus Manske, Gareth Maslen, Valentina Mangano, Daniel Alcock, Bronwyn MacInnis, et al. Characterization of within-host *Plasmodium falciparum* diversity using next-generation sequence data. *PloS one*, 7(2):e32891, 2012.

[4] David J Balding. Likelihood-based inference for genetic correlation coefficients. *Theoretical population biology*, 63(3):221–230, 2003.

[5] David J Balding and Richard A Nichols. Dna profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. *Forensic Science International*, 64(2):125–140, 1994.

[6] H-P Beck, I Felger, P Vounatsou, R Hirt, M Tanner, P Alonso, and C Menendez. 8. effect of iron supplementation and malaria prophylaxis in infants on *Plasmodium falciparum* genotypes and multiplicity of infection. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 93(Supplement 1):41–45, 1999.

[7] RS Bray and PCC Garnham. The life-cycle of primate malaria parasites. *British medical bulletin*, 38(2):117–122, 1982.

[8] DJ Conway, BM Greenwood, and JS McBride. The epidemiology of multiple-clone plasmodium falciparum infections in gambian patients. *Parasitology*, 103(Pt 1):1–6, 1991.

[9] I Felger, T Smith, D Edoh, A Kitua, P Alonso, M Tanner, and H-P Beck. 6. multiple *Plasmodium falciparum* infections in Tanzanian infants. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 93(Supplement 1):29–34, 1999.

[10] N Fraser-Hurt, I Felger, D Edoh, S Steiger, M Mashaka, H Masanja, T Smith, F Mbena, and H-P Beck. 9. effect of insecticide-treated bed nets on haemoglobin values, prevalence and multiplicity of infection with *Plasmodium falciparum* in a randomized controlled trial in tanzania. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 93(Supplement 1):47–51, 1999.

[11] Kevin Galinsky, Clarissa Valim, Arielle Salmier, Benoit de Thoisy, Lise Musset, Eric Legrand, Aubrey Faust, Mary L Baniecki, Daouda Ndiaye, Rachel F Daniels, et al. Coil: a methodology for evaluating malarial complexity of infection using likelihood from single nucleotide polymorphism data. *Malaria journal*, 14(1):4, 2015.

[12] Malcolm J Gardner, Neil Hall, Eula Fung, Owen White, Matthew Berriman, Richard W Hyman, Jane M Carlton, Arnab Pain, Karen E Nelson, Sharen Bowman, et al. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*, 419(6906):498–511, 2002.

[13] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

[14] Carlos A Guerra, Priscilla W Gikandi, Andrew J Tatem, Abdisalan M Noor, Dave L Smith, Simon I Hay, and Robert W Snow. The limits and intensity of *Plasmodium falciparum* trans-

mission: implications for malaria control and elimination worldwide. *PLoS Medicine*, 5(2):e38, 2008.

[15] Neil Hall, Marianna Karras, J Dale Raine, Jane M Carlton, Taco WA Kooij, Matthew Berriman, Laurence Florens, Christoph S Janssen, Arnab Pain, Georges K Christophides, et al. A comprehensive survey of the plasmodium life cycle by genomic, transcriptomic, and proteomic analyses. *Science*, 307(5706):82–86, 2005.

[16] Simon I Hay, Carlos A Guerra, Peter W Gething, Anand P Patil, Andrew J Tatem, Abdisalan M Noor, Caroline W Kabaria, Bui H Manh, Iqbal RF Elyazar, Simon Brooker, et al. A world malaria map: *Plasmodium falciparum* endemicity in 2007. *PLoS Medicine*, 6(3):e1000048, 2009.

[17] William G Hill and Hamza A Babiker. Estimation of numbers of malaria clones in blood samples. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 262(1365):249–257, 1995.

[18] Jonathan J Juliano, Kimberly Porter, Victor Mwapasa, Rithy Sem, William O Rogers, Frédéric Ariey, Chansuda Wongsrichanalai, Andrew Read, and Steven R Meshnick. Exposing malaria in-host diversity and estimating population diversity by capture-recapture using massively parallel pyrosequencing. *Proceedings of the National Academy of Sciences*, 107(46):20138–20143, 2010.

[19] Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

[20] Magnus Manske, Olivo Miotto, Susanna Campino, Sarah Auburn, Jacob Almagro-Garcia, Gareth Maslen, Jack O'Brien, and Dominic Kwiatkowski. Analysis of *Plasmodium falciparum* diversity in natural infections by deep sequencing. *Nature*, AOP, 2012.

[21] Alfredo Mayor, Francisco Saute, John J Aponte, Jesús Almeda, F Xavier Gómez-Olivé, Martinho Dgedge, and Pedro L Alonso. *Plasmodium falciparum* multiple infections in mozambique, its relation to other malariological indices and to prospective risk of malaria morbidity. *Tropical Medicine & International Health*, 8(1):3–11, 2003.

[22] Xiao-Li Meng. Posterior predictive p-values. *The Annals of Statistics*, pages 1142–1160, 1994.

[23] Toshihiro Mita, Kazuyuki Tanabe, and Kiyoshi Kita. Spread and evolution of *Plasmodium falciparum* drug resistance. *Parasitology international*, 58(3):201–209, 2009.

[24] Masatoshi Nei. F-statistics and analysis of gene diversity in subdivided populations. *Annals of human genetics*, 41(2):225–233, 1977.

[25] John D O'Brien, Zamin Iqbal, and Lucas Amenga-Etego. An integrative statistical model for inferring strain admixture within clinical plasmodium falciparum isolates. *arXiv preprint arXiv:1505.08171*, 2015.

[26] A Ofosu-Okyere, MJ Mackinnon, MPK Sowa, KA Koram, F Nkrumah, YD Osei, WG Hill, MD Wilson, and DE Arnot. Novel *Plasmodium falciparum* clones and rising clone multiplicities are associated with the increase in malaria morbidity in ghanaian children during the transition into the high transmission season. *Parasitology*, 123(02):113–123, 2001.

[27] D Payne. Spread of chloroquine resistance in *Plasmodium falciparum*. *Parasitology today*, 3(8):241–246, 1987.

[28] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[29] Hugh Reyburn, Redempta Mbatia, Chris Drakeley, Jane Bruce, Ilona Carneiro, Raimos Olomi, Jonathan Cox, WMMM Nkya, Martha Lemnge, Brian M Greenwood, et al. Association of transmission intensity and age with clinical manifestations and case fatality of severe plasmodium falciparum malaria. *JAMA*, 293(12):1461–1470, 2005.

[30] François Rousset. Genetic differentiation and estimation of gene flow from f-statistics under isolation by distance. *Genetics*, 145(4):1219–1228, 1997.

[31] Dietlind Schleiermacher, Christophe Rogier, Andre Spiegel, Adama Tall, Jean-Francois Trape, and Odile Mercereau-Puijalon. Increased multiplicity of *Plasmodium falciparum* infections and skewed distribution of individual msp1 and msp2 alleles during pregnancy in ndiop, a senegalese village with seasonal, mesoendemic malaria. *The American journal of tropical medicine and hygiene*, 64(5):303–309, 2001.

[32] T Smith, H-P Beck, A Kitua, S Mwankusye, I Felger, N Fraser-Hurt, A Irion, P Alonso, T Teuscher, and M Tanner. 4. age dependence of the multiplicity of *Plasmodium falciparum* infections and of other malariological indices in an area of high endemicity. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 93(Supplement 1):15–20, 1999.

[33] Robert W Snow, Carlos A Guerra, Abdisalan M Noor, Hla Y Myint, and Simon I Hay. The global distribution of clinical episodes of *Plasmodium falciparum* malaria. *Nature*, 434(7030):214–217, 2005.

[34] Marc A Suchard, Robert E Weiss, and Janet S Sinsheimer. Bayesian selection of continuous-time markov chain evolutionary models. *Molecular Biology and Evolution*, 18(6):1001–1013, 2001.

[35] Michel Tibayrenc. Genetic epidemiology of parasitic protozoa and other infectious agents: the need for an integrated approach. *International journal for parasitology*, 28(1):85–104, 1998.

[36] Isabella Verdinelli and Larry Wasserman. Computing bayes factors using a generalization of the savage-dickey density ratio. *Journal of the American Statistical Association*, 90(430):614–618, 1995.

[37] Bruce S Weir and C Clark Cockerham. Estimating F-statistics for the analysis of population structure. *evolution*, pages 1358–1370, 1984.

[38] Bruce S Weir and WG Hill. Estimating f-statistics. *Annual Review of Genetics*, 36(1):721–750, 2002.

[39] World Health Organization. *World malaria report 2008*. World Health Organization, 2008.

[40] Sewall Wright. The interpretation of population structure by f-statistics with special regard to systems of mating. *Evolution*, pages 395–420, 1965.

Tables

Parameter	Description
$j = 1, \cdots, M$	index over number of SNPs, M
$i = 1, \cdots, N$	index over number of samples, N
$d_{ij} = (r_{ij}, n_{ij})$	Read count data in sample i at variant j for reference and non-reference counts.
p_j	population-level non-reference allele frequency for SNP j
$\widehat{p}_j$	estimate of non-reference allele frequency for SNP j
$\tilde{p}_j$	minor-allele frequency for SNP j
$\widehat{\tilde{p}}_j$	estimate of minor allele frequency for SNP j
p_{ij}	within-sample non-reference allele frequency for SNP j in sample i
f_i	Inbreeding coefficient
$H_o(b, i)$	Observed heterozygosity for sample i in bin b.
$H_e(b)$	Expected heterozygosity for bin b.
$\widehat{f}_i^*$	Estimator of f_i by method $*$.
$\mathbf{f}, \mathbf{p}$	Vector of f_i and p_j's.
α_i, β_i	Parameters of beta distribution by tranmission intensity group i.
$\mathbf{c}$	Vector of parameters for PTI.

Table 1: Notation for parameters used throughout the manuscript. Note that additional parameters in the Appendix are not included.

Initial	Regressed	Direct	Bayesian
1.000	0.999	0.996	0.930
-	1.000	0.998	0.930
-	-	1.000	0.929
-	-	-	1.000

Table 2: Correlation coefficient among the four inbreeding estimators across 168 samples.

Figures

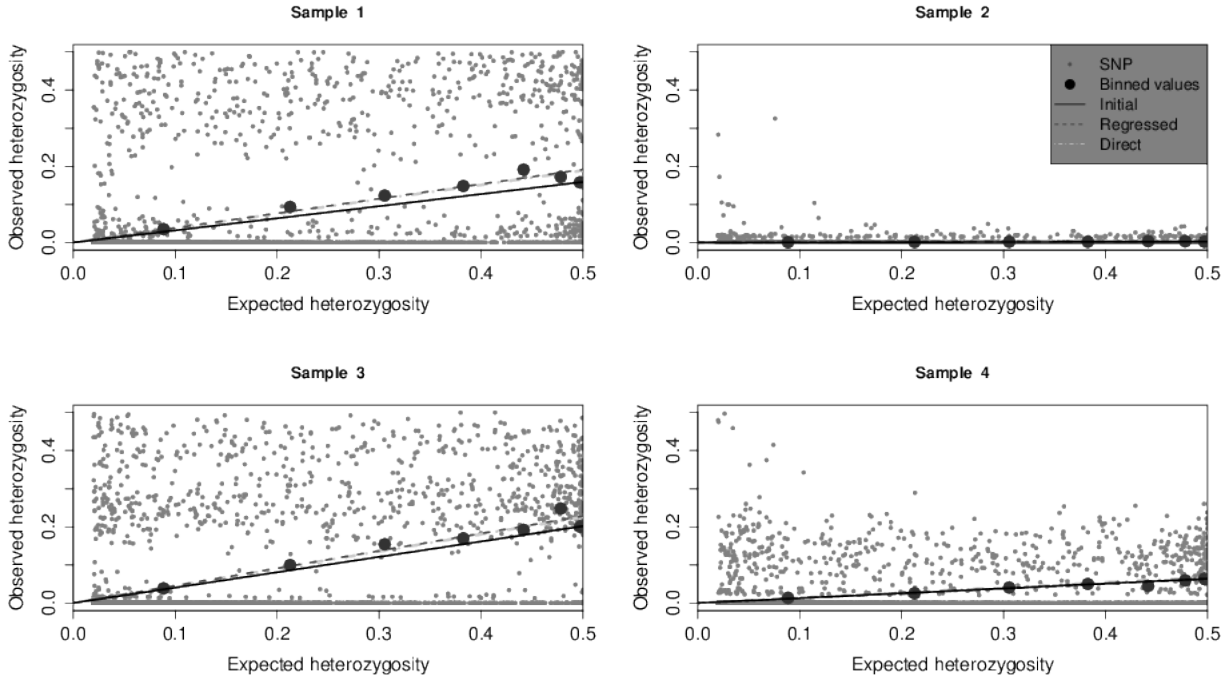

Figure 1: Raw SNP data for four representative samples with initial, regressed, and direct estimates of f_i overlaid. Grey dots represent individual SNPs with x-axis showing expected heterozygosity under HWE and y-axis showing observed heterozygosity.

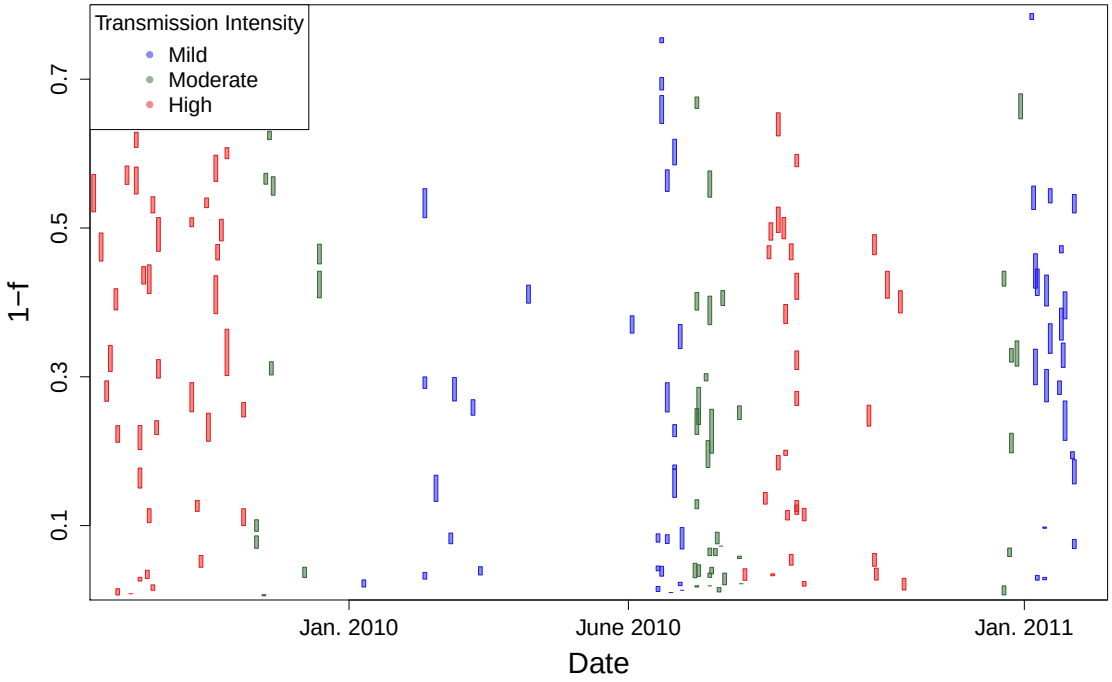

Figure 2: 95% credible intervals for $1 - f$ over the study interval, colored by tranmission intensity. Unmixed samples correspond tot $1 - f < 0.05$ (grey dashed line).

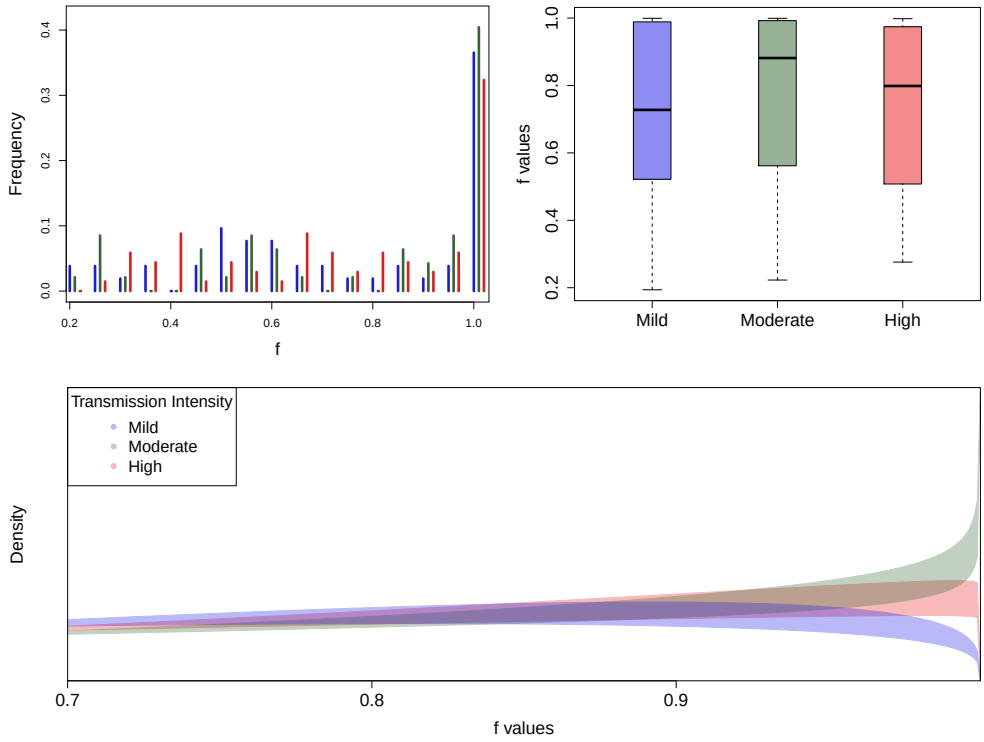

Figure 3: Comparisons of the inferred f values under mild, moderate and high transmission intensity. (Upper left) Frequency of binned f values by tranmission intensity. (Upper right) Boxplot of f values by transmission intensity. (Bottom) 90% credible interval of posterior density by transmission intensity.

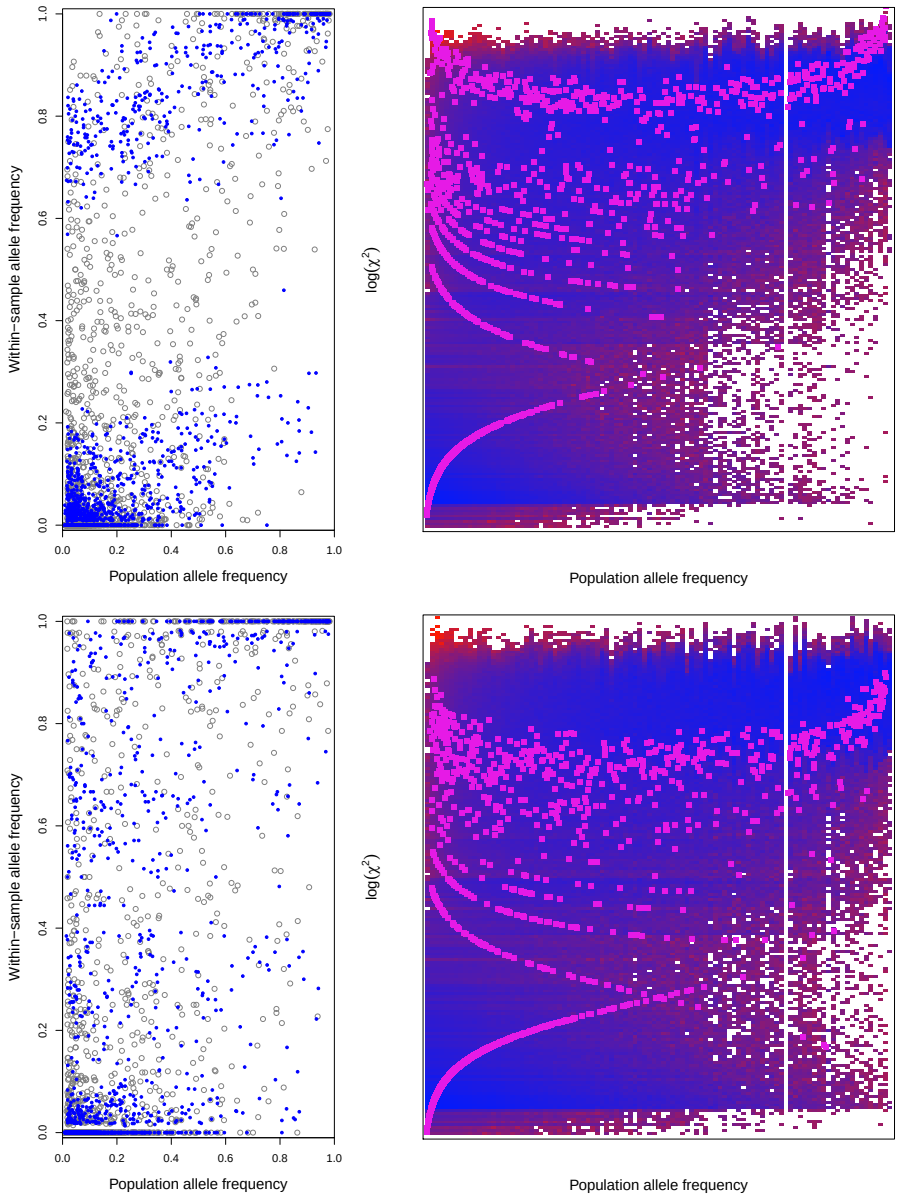

Figure 4: Posterior predictive distributions for two representative samples: $f = 0.45$ above; $f = 0.54$ below.

Supplementary Figures

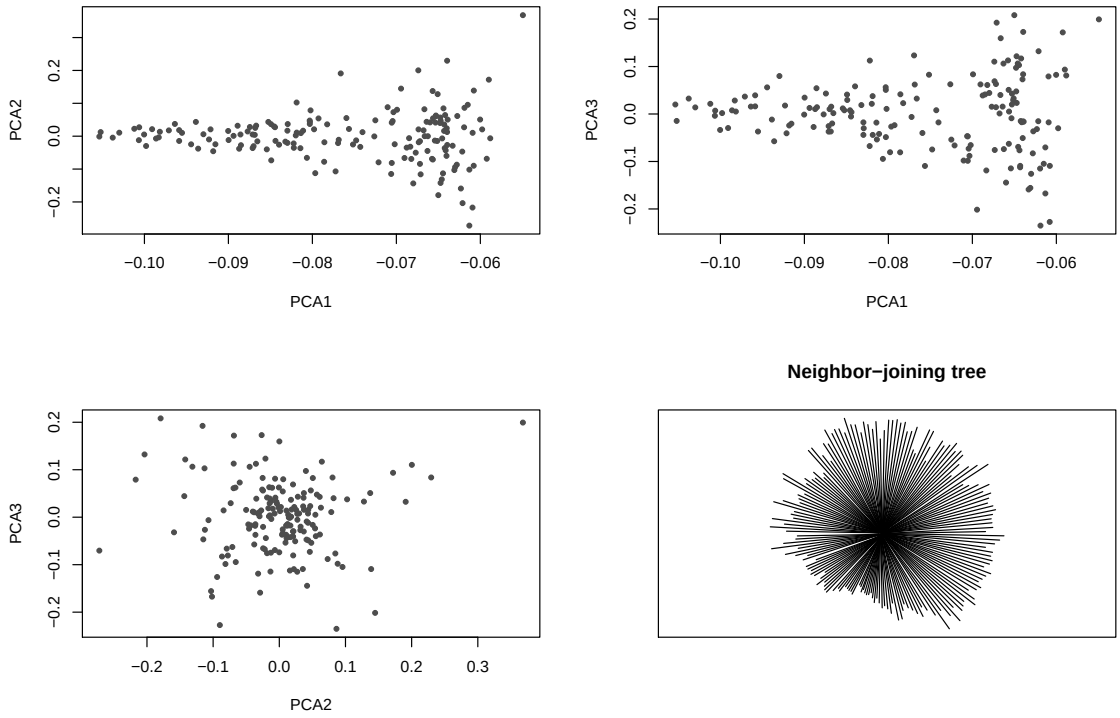

Figure S1: Observed population structure by principal components (upper left, upper right, lower left panels) and neighbor-joining tree.

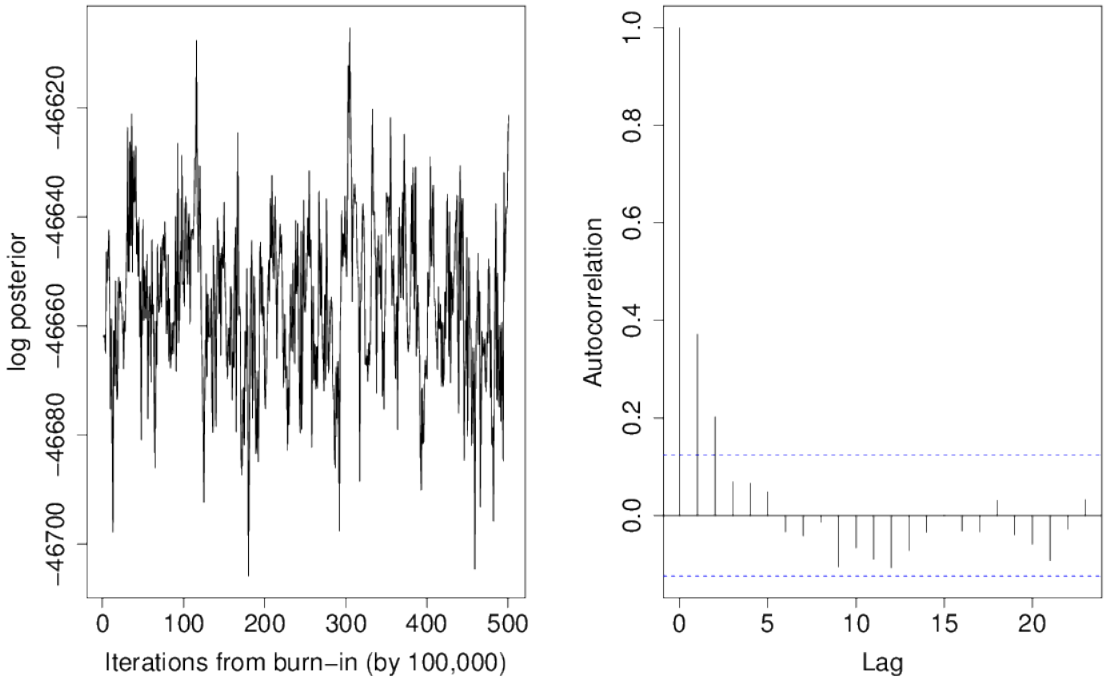

Figure S2: Log-likelihood for thinned MCMC chain (left) and autocorrelation for same chain (right).