PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Inferring Strain Mixture within Clinical *Plasmodium falciparum* Isolates from Genomic Sequence Data

John D. O'Brien[1]*, Zamin Iqbal[2], Jason Wendler[3], Lucas Amenga-Etego[2,4]

1 Mathematics Department, Bowdoin College, Brunswick, Maine, United States of America, **2** Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, Oxfordshire, United Kingdom, **3** Pacific Northwest National Laboratory, Richland, Washington, United States of America, **4** Navrongo Health Research Centre, Navrongo, Upper East Region, Ghana

* jobrien@bowdoin.edu

OPEN ACCESS

Citation: O'Brien JD, Iqbal Z, Wendler J, Amenga-Etego L (2016) Inferring Strain Mixture within Clinical *Plasmodium falciparum* Isolates from Genomic Sequence Data. PLoS Comput Biol 12(6): e1004824. doi:10.1371/journal.pcbi.1004824

Editor: Sergei L. Kosakovsky Pond, Temple University, UNITED STATES

Received: August 13, 2015

Accepted: February 17, 2016

Published: June 30, 2016

Data Availability Statement: These data are available via the PF3K data release 3: www.malariagen.net/data/pf3k-3 and the European Nucleotide Archive: www.ebi.ac.uk/ena. The sample numbers are included in the manuscript.

Funding: ZI was funded by a Sir Henry Dale Fellowship jointly awarded by the Wellcome Trust and the Royal Society (102541/Z/13/Z). (http://www.wellcome.ac.uk/Funding/Biomedical-science/Funding-schemes/Fellowships/Basic-biomedical-fellowships/wtdv031823.htm) LAE was funded by a MRC Centre for Genomics and Global Health grant from the Medical Research Council UK (G0600718)

Abstract

We present a rigorous statistical model that infers the structure of *P. falciparum* mixtures—including the number of strains present, their proportion within the samples, and the amount of unexplained mixture—using whole genome sequence (WGS) data. Applied to simulation data, artificial laboratory mixtures, and field samples, the model provides reasonable inference with as few as 10 reads or 50 SNPs and works efficiently even with much larger data sets. Source code and example data for the model are provided in an open source fashion. We discuss the possible uses of this model as a window into within-host selection for clinical and epidemiological studies.

Author Summary

Since the 1960's researchers have observed that *Plasmodium falciparum* infections, the cause of most severe malaria, are frequently composed of several different strains of the parasite. In this work, the authors use Bayesian methods on whole genome sequence data to model the structure of these mixtures. Results from this method are consistent with previous approaches but provide finer resolution of these mixtures. As whole genome data in malaria research becomes increasingly common, this work will allow researchers to delve further into the within-host dynamics of the parasite, a much-discussed but previously difficult-to-access aspect of this disease.

This is a *PLOS Computational Biology* Methods paper.

(http://www.mrc.ac.uk/funding/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

The protozoan parasite *Plasmodium falciparum* (Pf) is the cause of the vast majority of fatal malaria cases, killing at least half a million people a year [1–3]. 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 [4–6]. While research has elucidated many genetic factors this process, much of the genetic epidemiology of the parasite —including the effective recombination rate and the rate of gene flow across populations—is still unclear [5, 7, 8].

Understanding the implications of multiplicity of infection (MOI), where multiple strains appear to be present within a single patient's bloodstream, is a long-standing challenge [9–13]. While MOI-focused studies implicate MOI levels with a range of conditions, including clinical severity [14], age-specific severity [15–18], parasitemia levels during pregnancy [19], and other effects [20–23], there is no broad consensus about its role in controlling the course of an infection. Still, a wide variety of studies and genetic assays—most commonly through typing the *MSP* genes—show MOI as a regular feature of clinical Pf isolates [24–26].

WGS technologies applied to Pf extracted directly from infected patients' bloodstreams provide an unprecedented window into the structure of genetic mixture within samples [27, 28]. Initial work on this data shifted focus from estimating MOI to analysis based on inbreeding coefficients [13, 29–31]. These metrics, a form of *F*-statistic, give an estimate of the departure of within-sample allele frequencies from those expected under a Hardy-Weinberg-type equilibrium with the ambient population. From this perspective, each patient's bloodstream is seen as a subpopulation comprised of an admixture of all strains in the local environment, ranging from a perfectly random sampling of all nearby strains (panmixia) to the repeated sampling of just a single strain (unmixed).

The initial study applying WGS to clinical Pf isolates from eight countries on three continents showed the parasite to exhibit significant population structure at continental scales, with the amount of subpopulation structure varying significantly among regions [27]. Employing an F-statistic approach to measure the inbreeding coefficient from thousands of single nucleotide polymorphisms (SNPs), this work also argued that the degree of mixture varies significantly across populations, with highly mixed samples occurring relatively frequently in west Africa but only occasionally in Papua New Guinea. The authors suggest an association between increased levels of observed mixture and increased transmission intensity in the local environment. Transmission intensity, the rate at which individuals are infected with Pf, likely determines some part of the frequency of out-crossing within parasite populations and so may be critical to understanding gene flow and strategies for resistance control [32].

In this paper, we present a statistically rigorous model that synthesizes these two distinct and previously disparate approaches to analyzing Pf clinical mixtures: assessing the number of distinct genetic types within a sample (the MOI approach [31]) and measuring the degree of panmixia with respect to the local population (the panmixia approach [33]). The model makes two significant innovations: first, a reversible jump Markov Chain Monte Carlo (MCMC) implementation to capture uncertainty in the number of strains, and second the inclusion of a panmixia term to deal with unexplained variation in the mixture. This work possesses similarities in character to the COIL algorithm [34], but can capture more complex mixture structure and is geared toward analyzing WGS data ($>$1000 SNPs) rather than a small number of SNPs ($\sim$50 SNPs).

This model centers around how the two sub-models—MOI and panmixia—contribute to the observed *within-sample* non-reference allele frequencies (WSAF) as they relate to the *population-level* non-reference allele frequencies (PLAF). For clarity, we will deprecate the use of

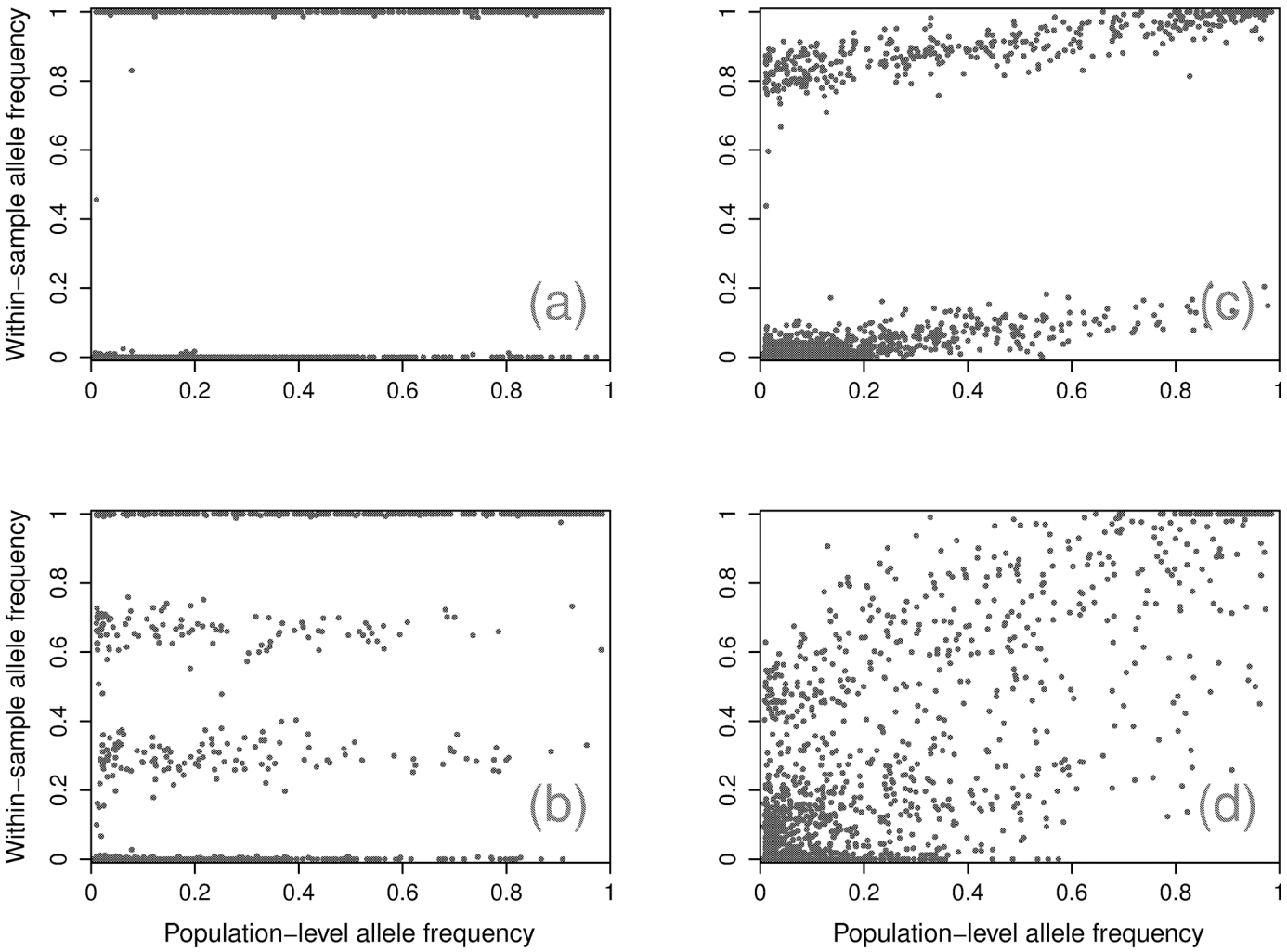

Fig 1. Example samples. Four representative samples with WSAF for each SNP plotted against the PLAF, showing an absence of mixture (a), a partially panmixed sample (b), a simple mixture (c), and a complex mixture (d).

doi:10.1371/journal.pcbi.1004824.g001

non-reference in front of the term allele frequency, since they are all calibrated in this fashion. We will use the acronyms WSAF to denote the within-sample allele frequency and PLAF to denote population-level allele frequency to avoid confusion about the particularly allele frequency being indicated. The goal of the model is to explain observed 'bands' that emerge when examining SNPs WSAF as a function of their PLAF (Fig 1).

The model assumes (1) that the number of bands is a consequence of the number of distinct strains present within a sample, (2) that SNPs are unlinked, and (3) that unexplained variation is assumed to be due to a small fraction of genomes sampled under panmixia. To distinguish from an inbreeding coefficient—a similar but distinct concept—we refer to this fraction as a panmixia coefficient. The collection of WSAF bands then appears as a function of the finite mixture of the strains, with the slope in WSAF bands with respect to the PLAF explained by both the SNP distribution and the panmixia coefficient.

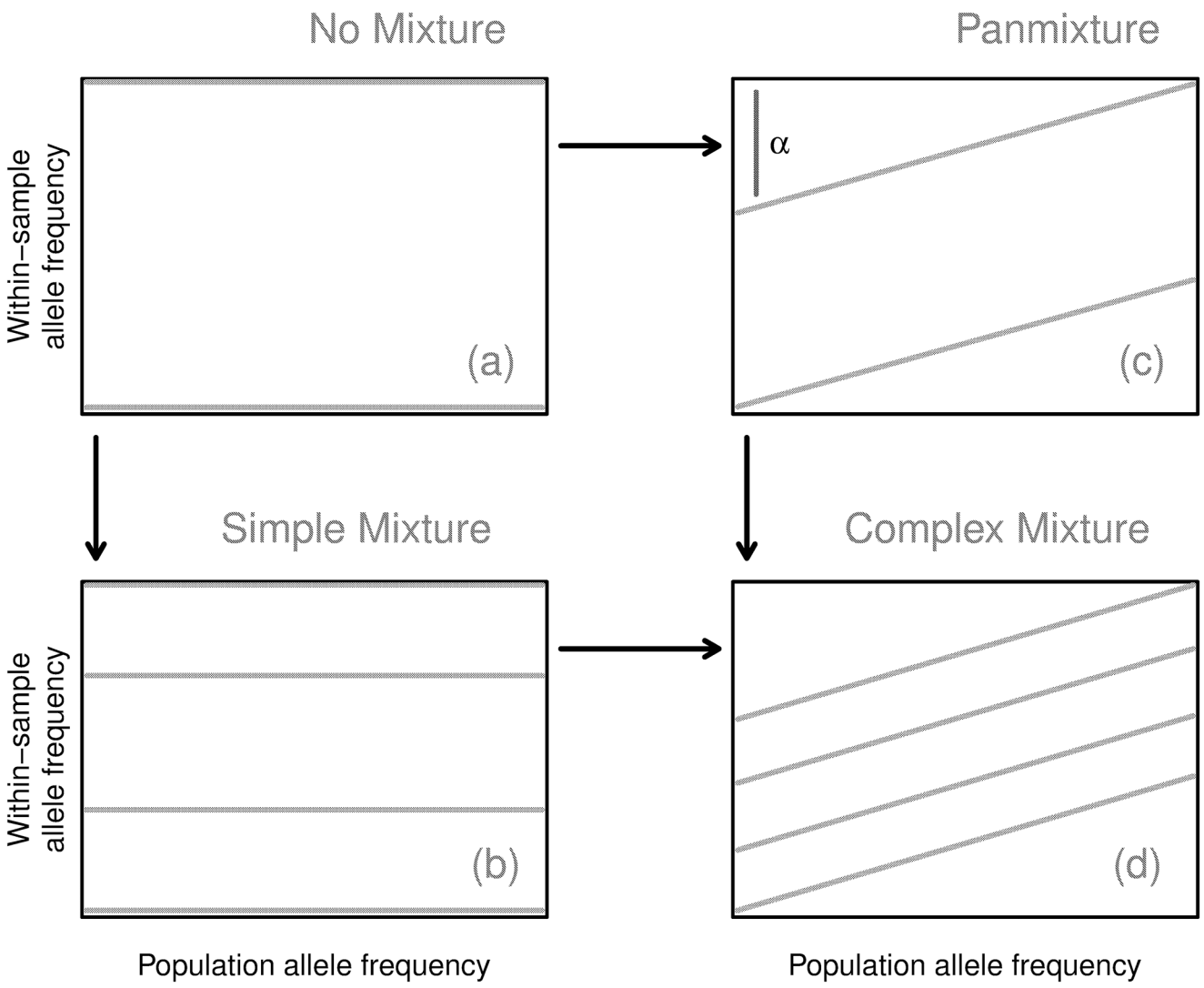

Fig 2. Model diagram. The structure of the model can be understood in terms of four related states connecting the WSAF to the PLAF: no mixture (upper left); simple mixture (lower left); panmixture (upper right); and complex mixture (lower right).α is exaggerated for explanation; realistic values are less than 0.05.

doi:10.1371/journal.pcbi.1004824.g002

Fig 2 lays out how the consequent banding patterns can arise. In the simplest case, a sample is composed of a single, unmixed strain, and all SNPs exhibit a WSAF of zero or one (see Fig 2 (a)), based on whether they agree with the reference. Consequently, the WSAF is independent of PLAF, leading to two flat bands at these values. We call these samples unmixed, since this is how a single strain with some divergence from the reference will appear. In the case where there are a finite number of strains mixed within a sample, then at a given SNP position some number of the strains will exhibit a reference allele and some a non-reference allele. The WSAF for that SNP is determined by the proportions of non-reference strains in the sample mixture. Observing many SNPs displays 'bands' of constant WSAF across the PLAF. Thus, for K component strains there are 2^K possible combinations of biallelic states, leading to that number of bands.

A fraction of the Pf organisms present within the blood may not be from any of the dominant strains. We model these as randomly sampled from the local population according to simple panmixia. Observationally, this leads to a consistent change in the slope of each of the bands. To see this, consider an admixture of two distinct Pf populations: a single strain, representing $1 - \alpha$ of the within-sample genomes, and the remaining α that we assume follow panmixia. The α tilt in the WSAF arises from the fact that for this proportion of organisms the probability of sampling non-reference allele is proportional to the PLAF (Figs 1(c) and 2(c)). Samples with high K appear to have additional tilt due to the higher probability of non-reference alleles occurring at high PLAF (Figs 1(d) and 2(d)).

The paper proceeds as follows. We first detail the structure of the WGS data, introduce some notation, and the essential mathematical structure of the model. We then present an extensive simulation study on the performance of the model, an application of the model to artificial laboratory mixtures, and an examination of its application to field isolates collected from northern Ghana. We conclude by discussing the strengths and weaknesses of the model, a means of experimental validation, and potential consequences for the etiology of clinical malaria.

Materials and Methods

Data

The field WGS data come from Illumina HiSeq sequencing applied to Pf extracted from 419 clinical blood samples collected from infected patients in the Kassena-Nankana district (KND) region of Upper East Region of northern Ghana. Collection occurred over approximately 2 years, from June 2009-June 2011. The raw sequence reads for these data are accessible through the PF3K project https://www.malariagen.net/projects/parasite/pf3k. This includes data from the MalariaGEN Plasmodium falciparum Community Project on www.malariagen.net/projects/parasite/pf. On the website for this method, we provide read count data subsampled from the full data set. The artificial laboratory samples were sequenced and called per protocols given in [35]. The raw sequence data is available through the European Nucleotide Archive with the accessions available in the S1 Text.

The full sequencing protocol and collection regime are described in [27]. After quality control measures, all samples were examined, and following a documented protocol comparing against world-wide variation, 198,181 single-nucleotide polymorphisms (SNPs) were called [27]. These are exclusively coding SNPs found outside of the telomeric and subtelomeric regions that exhibit unusual structural properties. Each SNP xcall provides the number of reference and non-reference read counts observed at each variant position within the genome, ascertained against the the 3D7 reference [36]. These data were exhaustively examined for spurious heterozygosity and evidence of DNA contamination, with mixed calls verified using time-of-flight mass spectrometry at greater than 99% accuracy [27].

For this project, we further filtered the data. First, multiallelic positions were reclassed as biallelic. We then excluded positions that exhibited no variation within the KND samples, any level of missingness (no read counts observed), or minor allele frequency less than 0.01. To remove low quality samples, we removed those with more than 4,000 SNPs missing and fewer than 20 read counts, following an inflection point observed in S1 Fig. These cleaning measures left 2,429 SNPs in 168 samples. These SNPs exhibit desirable properties for model inference—high and consistent coverage across all samples—that could likely be expanded to non-coding or less stringent cleaning standards without issue. More than 95% of the remaining samples' sequencing was completed without PCR amplification. We observe little apparent population structure among the samples, evidenced either by principal components analysis or a

Table 1. Parameters and definitions for the model and its description.

Parameter	Definition
N	Number of samples
M	Number of SNPs
K	Number of strains
$i = 1, \cdots, N$	Index for samples
$j = 1, \cdots, M$	Index for SNPs
$r = 1, \cdots 2^K$	Index for bands / strain mixtures
p_j	(Non-reference) allele frequency for SNP j
$\mathcal{P} = [p_j]$	The PLAF for all SNPs
$\mathcal{Q} = [q_{ij}]$	Within-sample allele frequency for SNP j in sample i
α	Degree of panmixia within a sample, panmixia coefficient
$\mathcal{S} = [s_1, \cdots, s_K]$	Strains in a sample
$\mathcal{W} = [w_1, \cdots, w_K]$	Strain proportions in a sample
λ_r	Band proportions within sample
v	Variation parameter for Beta-binomial
WSAF	Within-sample allele frequency
PLAF	Population-level allele frequency

doi:10.1371/journal.pcbi.1004824.t001

neighbor-joining tree of the pairwise difference among samples (S2 Fig). The data preparation scripts are available with the source code for the model, https://github.com/jacobian1980/pfmix/.

Notation

Following the data preparation and cleaning, our analysis begins with a set of $N = 168$ clinical samples, each composed of $M = 2{,}429$ SNPs. At each SNP j within each clinical sample i, we observe r_{ij} reads that agree with the reference genome and n_{ij} reads that do not agree. The total number of read counts in sample i at SNP j is then $n_{ij} + r_{ij}$. For a sample i, we write the complete data across all SNPs as $\mathcal{D}_i = [(r_{i1}, n_{i1}), \cdots, (r_{iM}, n_{iM})]$. For each SNP j, we associate a PLAF p_j. The collection of all p_j we refer to as $\mathcal{P}$.

Conditional upon the number of strains K, there are 2^K bands, indexed by $r = 1, \cdots, 2^K$. The full collection of bands we call $\mathcal{Q}$, with q_{ijr} indicating the WSAF for sample i at SNP j in band r. The probability of a SNP lying within the distinct bands across the PLAF is specified by a mixture component λ_r, which is a function of the PLAF detailed below. The degree of panmixia in a sample is given by α, a value between zero and one. A complete list of the model parameters is given in Table 1.

Model

Statistically, the model takes the form of a finite mixture model with the mixture components associated with individual bands [37, 38]. We take a Bayesian approach to inference and construct the model by giving an overall rationale for the decomposition of the posterior distribution, and then justify the appropriate choice of probability distributions for each of the terms [39].

Decomposition. We assume that samples are independent of each other and that the SNP data for each sample depends solely on the number of bands (K), the WSAF ($\mathcal{Q}$), the PLAF ($\mathcal{P}$), and a shape parameter v. As samples are treated independently, we deprecate sample-specific subscripts for the model parameters. Considering the data for a single sample, $\mathcal{D}_i$, the posterior distribution can then be written as:

$$\begin{aligned}\mathbb{P}(\mathcal{Q},\mathcal{P},\mathcal{W},\alpha,v,K|\mathcal{D}_i) &\propto \mathbb{P}(\mathcal{D}_i|\mathcal{Q},\mathcal{P},\mathcal{W},\alpha,v,K)\cdot\mathbb{P}(\mathcal{Q},\mathcal{P},\mathcal{W},\alpha,v,K)\\ &= \mathbb{P}(\mathcal{D}_i|\mathcal{Q},\mathcal{P},v,K)\cdot\mathbb{P}(\mathcal{Q},\mathcal{P},v,K,\mathcal{W},\alpha).\end{aligned} \tag{1}$$

We also assume that the WSAF depends only on the PLAF, the panmixia coefficient, the number of strains, and their proportions within the sample, allowing the right-hand side of Eq (1) to be further decomposed, by noting that:

$$\mathbb{P}(\mathcal{Q},\mathcal{P},v,K,\mathcal{W},\alpha) = \mathbb{P}(\mathcal{Q}|\mathcal{P},v,K,\mathcal{W},\alpha)\cdot\mathbb{P}(\mathcal{P},v,K,\mathcal{W},\alpha). \tag{2}$$

While the strain proportions clearly depend on the number of strains, the remaining parameters we take to be independent of this value and of each other. This means that the last right-hand side term in Eq (2) becomes:

$$\mathbb{P}(\mathcal{P},v,K,\mathcal{W},\alpha) = \mathbb{P}(\mathcal{P})\cdot\mathbb{P}(v)\cdot\mathbb{P}(\mathcal{W}|K)\cdot\mathbb{P}(K)\cdot\mathbb{P}(\alpha). \tag{3}$$

Substituting Eqs (2) and (3) into Eq (1), yields the final decomposition:

$$\begin{aligned}\mathbb{P}(\mathcal{Q},\mathcal{P},\mathcal{W},\alpha,v,K|\mathcal{D}_i) \propto\ & \mathbb{P}(\mathcal{D}_i|\mathcal{Q},\mathcal{P},v,K)\cdot\mathbb{P}(\mathcal{Q}|\mathcal{P},v,K,\mathcal{W},\alpha)\cdot\\ & \mathbb{P}(\mathcal{P})\cdot\mathbb{P}(v)\cdot\mathbb{P}(\mathcal{W}|K)\cdot\mathbb{P}(K)\cdot\mathbb{P}(\alpha).\end{aligned} \tag{4}$$

We now specify each of the terms on the right-hand side above as probability distributions.

Likelihood: $\mathbb{P}(\mathcal{D}_i|\mathcal{Q},\mathcal{P},v,K)$. Within band r, the WSAF at SNP j in sample i is q_{ijr}. Supposing that read counts at j are identically and independently distributed with probability q_{ijr}, we model the probability of the data (r_{ij}, n_{ij}) as a Beta-binomial distribution, allowing us to fit greater dispersion than expected under a pure binomial. We parameterize this distribution in terms of q_{ijr} and v rather than the more commonly used shape and scale parameters, α and β, with the relationship $q_{ijr}\cdot v=\alpha$ and $(1-q_{ijr})\cdot v=\beta$. This parameterization allows us to write the model in terms of the allele frequency that defines each band. The additional v is a shape parameter that serves as an over-dispersion parameter. These give a likelihood expression for a SNP within a band as:

$$\mathbb{P}(n_{ij},r_{ij}|r,q_{ijr},v) = \binom{n_{ij}+r_{ij}}{n_{ij}}\cdot\frac{\mathrm{B}(n_{ij}+q_{ijr}\cdot v,r_{ij}+(1-q_{ijr})\cdot v)}{\mathrm{B}(q_{ijr}\cdot v,(1-q_{ijr})\cdot v)}, \tag{5}$$

where B is the beta function.

As any SNP could lie within any band, we employ a novel version of the finite mixture model to capture this segregation. Given K strains, there are then 2^K ways that the strains can be assorted into non-reference and reference allele states at any given position j. A given band r arises from C_r strains exhibiting the non-reference allele and $2^K - C_r$ strains exhibiting the reference allele. Supposing no population structure among the strains and neglecting linkage among SNPs, the probability that a given SNP will be in that band is simply the probability of drawing C_r non-reference alleles and $2^K - C_r$ reference alleles, conditional upon p_j:

$$\begin{aligned}\mathbb{P}(\text{SNP } j\in\text{band } r|p_j) &= p_j^{C_r}\cdot(1-p_j)^{2^K-C_r}\\ &= \lambda_r(p_j).\end{aligned}$$

Consequently, the density of the mixture coefficients for each band varies across the PLAF but such that they always sum to 1 across all bands at any SNP position j. This gives a likelihood across all bands as:

$$\begin{aligned}\mathbb{P}(\mathcal{D}_{ij}|\mathcal{Q},\mathcal{P},v,K) &= \sum_{r=1}^{2^K}\mathbb{P}(\text{SNP } j \in \text{band } r|p_j)\cdot\mathbb{P}(n_{ij},r_{ij}|r,q_{ijr},v)\\ &= \sum_{r=1}^{2^K}\lambda_r(p_j)\cdot\mathbb{P}(n_{ij},r_{ij}|r,q_{ijr},v).\end{aligned}$$

Following from the assumption of no linkage, SNPs will independently assort into bands. This leads to a product-sum form for the likelihood for $\mathcal{D}_i$:

$$\mathbb{P}(\mathcal{D}_i|\mathcal{Q},\mathcal{P},v,K) = \prod_{j=1}^{M}\left[\sum_{r=1}^{2^K}\lambda_r(p_j)\cdot\mathbb{P}(n_{ij},r_{ij}|r,q_{ijr},v)\right]. \quad (6)$$

Band structure: $\mathbb{P}(\mathcal{Q}|\mathcal{P},v,K,\mathcal{W},\alpha)$**.** The complex mixture model contains two distinct subcomponents that we call the simple mixture model and the panmixture model, respectively. Both models generalize the unmixed case, though in different ways. We first characterize the unmixed model and the two extensions before showing how these can be combined to create the complex model. In practice, we only fit data using the full model and allow it to indicate the number of strains, their proportions, and the degree of panmixia. We do not know the number of strains *a priori* so we employ a reversible jump approach to infer the posterior distribution on K. However, for the purpose of detailing the model, we assume that K is known.

Unmixed model. In an unmixed sample only one strain is present and the panmixture coefficient is zero (i.e. $K = 1$ and $\alpha = 0$). Consequently, all SNPs exhibit a WSAF of either zero or one (Fig 2(a)). There are then two bands, $r = 1, 2$ and $q_{ij1} = 0$ and $q_{ij2} = 1$.

Simple mixture model. Conditional upon K, the distinct strains, $s_1, \cdots, s_K$, are combined together in the sample with proportions, $\mathcal{W} = (w_1, \cdots, w_K)$, but that $\alpha = 0$. Necessarily, $\Sigma_k\, w_k = 1$. For each SNP j, the probability of being within band r is given by $\lambda_r(p_j)$, as above. Band r is defined by a vector $v_r = (\mathbf{1}_{\{s_1 \in r\}}, \cdots, \mathbf{1}_{\{s_K \in r\}})$, where $\mathbf{1}_{\{s_k \in r\}}$ is a function indicator function on whether strain k exhibits a non-reference allele within the sample. The WSAF of band r for SNP j (q_{ijr}) is then given by the sum of proportions for strains that exhibit a non-reference allele:

$$q_{ijr} = \sum_{k=1}^{K} w_k \cdot \mathbf{1}_{\{s_k \in r\}}. \quad (7)$$

Taken across all r bands, this leads to 2^K bands with zero slope and corresponding proportions $(0, w_1, \cdots, w_K, w_1 + w_2, w_1 + w_3, \cdots, 1)$.

Panmixture model. In the simplest case, the panmixture model represents the admixture of two distinct Pf populations when $K = 1$: a single strain, representing $1 - \alpha$ of the within-sample genomes, and a random sample of alleles from the local population for the remaining α genomes. α can be considered the fraction of unexplained variation in the sample. When $\alpha = 0$ the model reduces to the unmixed case (see Figs 1(b) and 2(b)). For each position j, there are still only two bands: the higher one corresponding to the non-reference allele being present in the dominant strain, and the lower one corresponding to its absence. However, the WSAF for these bands varies according to p_j with slope α. To resolve q_{ijr}, first consider the upper band, $r = 2$. At any position j, $1 - \alpha$ of the reads come from the dominant strain. The remaining

reads, each sampled randomly from the local population, each have probability p_j of being non-reference. This leads to $q_{ij2} = (1 - \alpha) + \alpha \cdot p_j$. For the lower band, the dominant strain contributes no non-reference reads so $q_{ij1} = \alpha \cdot p_j$.

Complex mixture model. The complex model synthesizes the simple mixture and panmixture models so that both K and α may vary. In this case, at position j, α of the reads are sampled randomly from the across the local population, contributing a fraction of $\alpha \cdot p_j$ non-reference alleles. The state of the remaining reads are determined by $\mathcal{W}$ as in Eq (7). For band r at position j, the WSAF is then given by:

$$q_{ijr} = (1-\alpha) \cdot \left(\sum_{k=1}^{K} w_k \cdot \mathbf{1}_{\{s_k \in r\}} \right) + \alpha \cdot p_j. \quad (8)$$

There are then 2^K bands with proportions $(0, w_1, \cdots, w_K, w_1 + w_2, w_1 + w_3, \cdots, 1)$ and slope α.

Priors. For the remaining four probability distributions we place the following vague prior distributions:

$$\begin{aligned} \mathcal{W}|K &\sim \text{DIRICHLET}(\mathbf{1}_K) \\ \alpha &\sim \text{UNIFORM}(0,1) \\ v &\sim \text{EXPONENTIAL}(5) \\ K &\sim \text{zero-truncated} \quad \text{POISSON}(2), \end{aligned}$$

where $\mathbf{1}_K$ is a vector of K ones.

Inference

We infer the model parameters using a standard reversible jump MCMC approach [40, 41] with one exception: we first calculate maximum-likelihood estimates (MLE) for $\mathcal{P}$ across all samples and then treat these as fixed when inferring the remaining parameters [42]. This choice is motivated by statistical expedience and computational speed: except for $\mathcal{P}$, the parameters of the model are independent across samples and so this approximation enables the algorithm to infer parameters in parallel rather than jointly. This avoids the difficulties of performing inference on the number of strains within all samples simultaneously. Running in parallel also increases the computational speed of the implementation by at least an order of magnitude. Since the sample collection is large enough that the PLAF is nearly independent of any given sample, we do not expect this approximation to significantly bias inference.

For each SNP j, the MLE derives from treating the non- and reference reads within a sample as coming from a binomial distribution with parameter p_j. This leads to:

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

To infer the number of strains, K, for each sample, we employ a pair of complementary split/merge reversible jump MCMC moves. To specify the split move first not that in moving from $K \to K + 1$ that the transformation only affects the parameter $\mathcal{W}$. If we randomly select w_k, $1 \leq k \leq K$, then we can split this into two components, $u \cdot w_k$ and $(1 - u) \cdot w_k$, where u is drawn from a uniform distribution. This establishes a diffeomorphism between parameters at K and $K + 1$ with Jacobian w_k. The proposal ratio is $(K^2 - K)/K = K - 1$. The acceptance ratio then is the product of the proposal ratio, Jacobian, the likelihood ratio, and the prior likelihood. The merge move randomly selects two states, k_1 and k_2, and merges them to k' by setting $w' = w_{k_1} + w_{k_2}$. The Jacobian and proposals are the reciprocal of those for the split move.

Table 2. Table of simulated parameter values. *C* is the number of read counts while *M*, *K* and *α* are as in Table 1.

Parameter	Values:			
M	50	150	500	2500
C	10	25	100	250
α	0.001	0.01	0.1	
K	1	3		

doi:10.1371/journal.pcbi.1004824.t002

Conditional on $\mathcal{P}$ and K, for each of the three parameters, α, $\mathcal{W}$, and v, we propose new values directly from the prior distribution. This leads to Metropolis-Hastings ratios almost solely dependent on the ratio between the likelihood and priors for the proposed state to those for the current. The inference scheme is implemented in set of scripts for the R computing language, and can be found under the Academic Free License at https://github.com/jacobian1980/pfmix/s. For a single sample with $K = 5$, a sufficiently long MCMC run takes approximately 10 minutes on a single high-performance computing core.

Results

Simulations under the model

To demonstrate the efficacy of the model and our implementation, we present a simulation study examining the algorithm's performance under a range of simulated data. We consider two distinct aspects of the inference: how well the model infers the number of strains, and, conditional upon that number, how well it infers the model's other parameters. We simulate data from the model in the following way. Conditional upon the number of SNPs (M), panmixture coefficient (α), number of strains (K) and the sum of the read counts (C) we draw a vector of probabilities ($\mathcal{W}$) from a uniform Dirichlet distribution. We combine the values of $\mathcal{W}$ in all possible permutations to create the 2^K bands and assign the PLAF for the SNPs evenly from $1/M$ to 1, so that the j^{th} SNP has PLAF $\frac{j}{M}$. For each SNP, we first probabilistically select the band it occupies according to Eq (6). We then simulate read counts from the likelihood (Eq 5) with q_{ijr} per Eq (8). For all simulations, we set $v = 10$. We run the simulation across the range of values for M, α, K and C given in Table 2. For each parameter set, we create 10 independent realizations.

Number of components. Fig 3 shows the algorithm's performance for inferring the number of components becomes more accurate with the number of SNPs and the number of reads, with 50 SNPs and 25 read counts sufficient to reliably recover the simulated values. With more SNPs, the requirement on read counts can be reduced to 10 with similar performance. Conditional upon α, the simulations indicate that the number of SNPs is the largest determinant of performance, and the sum of the read counts playing an important supporting role. Inference of the number of strains is generally strong at low panmixture levels (small α values), but is noticeably more conservative for $\alpha = 0.1$.

Parameters. Fig 4 shows similar performance for inference of the strain proportions, $\mathcal{W}$, and panmixture coefficient, α. For $\mathcal{W}$, we report the mean squared deviation. For α, we report the absolute normalized deviation to account for relative difference from the true value. For both parameters, we observe that the number of SNPs is the strongest determinant of accuracy, with $M = 150$ ensuring moderately strong performance. Again, high α moderately decreases the quality of inference for the strain proportions.

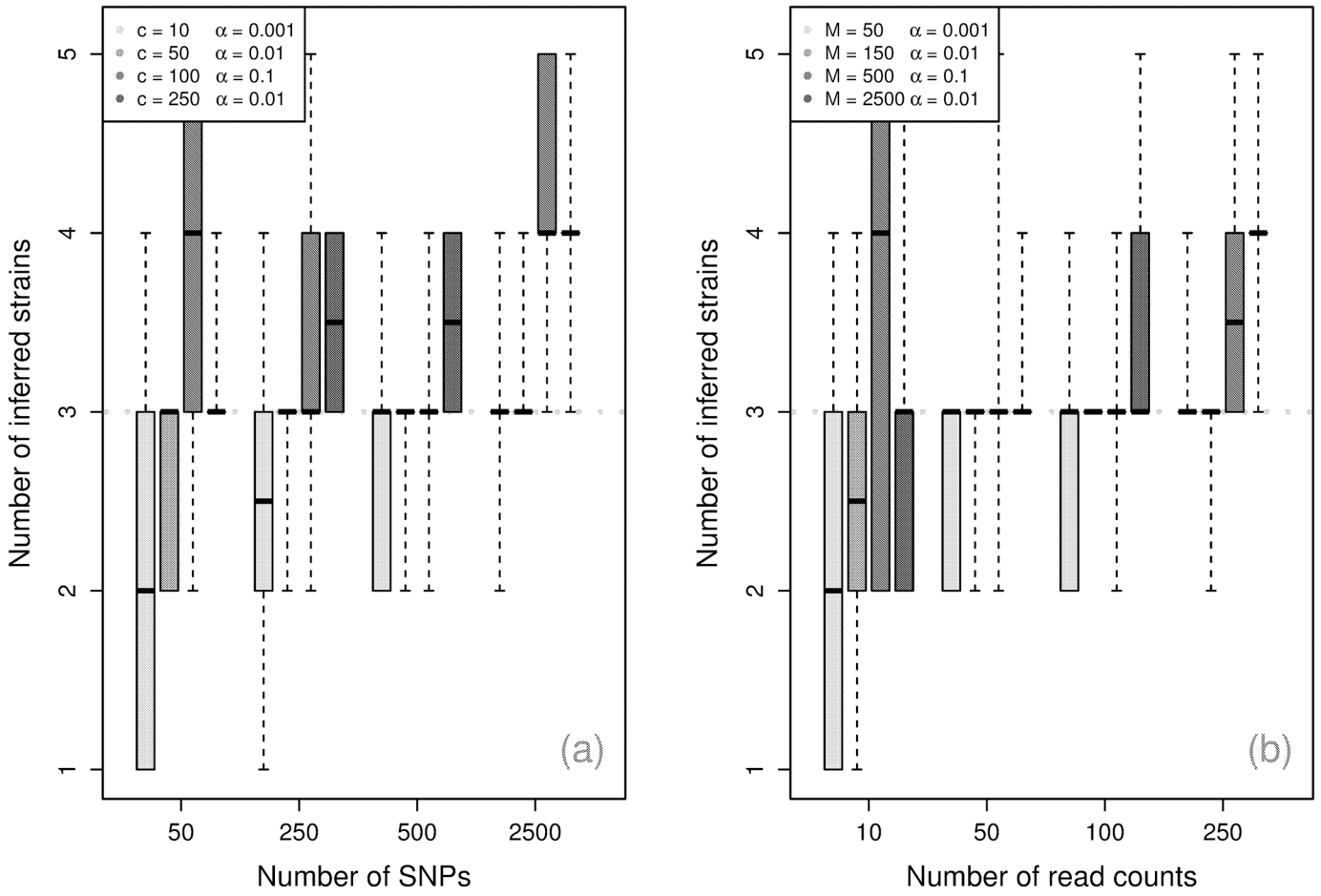

Fig 3. Component inference. *Maximum a posteriori* (MAP) inferred number of components by number of read counts across 10 simulations, with dotted line at the true number of components.

doi:10.1371/journal.pcbi.1004824.g003

Laboratory artificial mixtures

We apply the algorithm to 18 artificial laboratory mixtures. These artificial samples were generated by taking stock of two standard Pf lines, DD2 and 7G8, and adding them together in the fixed proportions given in S1 Table, and completing then Illumina sequencing and variant-calling with using the same protocols as [27]. Samples had a median of 65 reads for the variants considered here. Complete sequencing protocols and laboratory methods detailed in [35] (data available at European Nucleotide Archive). Both strains have high-confidence reference sequences. We subsample 2000 SNPs from the 23,109 SNPs available for comparison based on non-reference WSAF. The results in S1 Table show very strong agreement between the laboratory and inferred mixtures. The inferred α for all samples was less than 0.001 and had Bayes factor for non-zero α as less than 1, indicating that the samples have little unexplained mixture observed relative to the field samples.

Clinical samples from northern Ghana

Applying the algorithm to the 168 high-quality samples from KND, we observe numbers of strains ranging from 1 to 7, with α falling between 0 and 0.14, and a moderate correlation

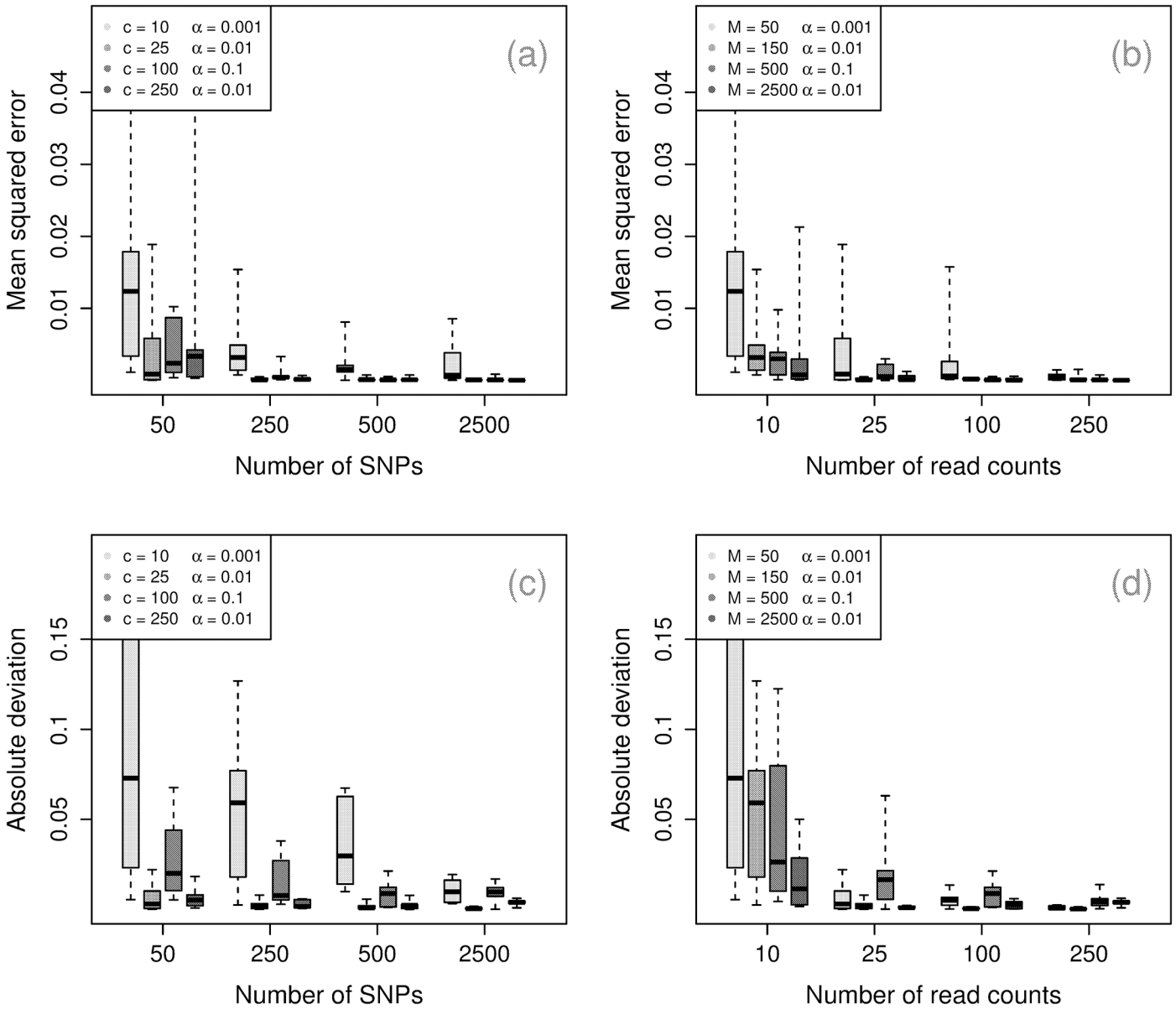

Fig 4. Performance for parameter inference. Upper row: mean squared deviation for strain frequencies by number of read counts (left) and by number of SNPs (right). Lower row: absolute normalized deviation for panmixia coefficient by number of read counts (left) and by number of SNPs.

doi:10.1371/journal.pcbi.1004824.g004

between K and α (Fig 5). The largest subset of samples were unmixed, with $K = 1$ and $\alpha < 0.01$, though the majority of samples exhibit low but noticeable levels of admixture, with $K = 2, 3, 4$ and $0.01 \leq \alpha \leq 0.03$. A small number of samples exhibit complex mixtures, with $K > 4$ and α typically greater than 0.02. These samples also exhibit the most variance in the posterior estimate of K, frequently ranging from 3 to 8. To examine the necessity of the panmixia model to capture unexplained variation in the field samples, we calculate a Bayes factor for each sample under the two models, M_0: $\alpha = 0$ and M_1: $\alpha \neq 0$. Since this is a single parameter, we employ the Savage-Dickey ratio calculation as in [43]. We find that 78 samples give factors larger than 10, indicating strong evidence for M_1, and 9 samples give factors larger than 100, indicating overwhelming evidence for M_1.

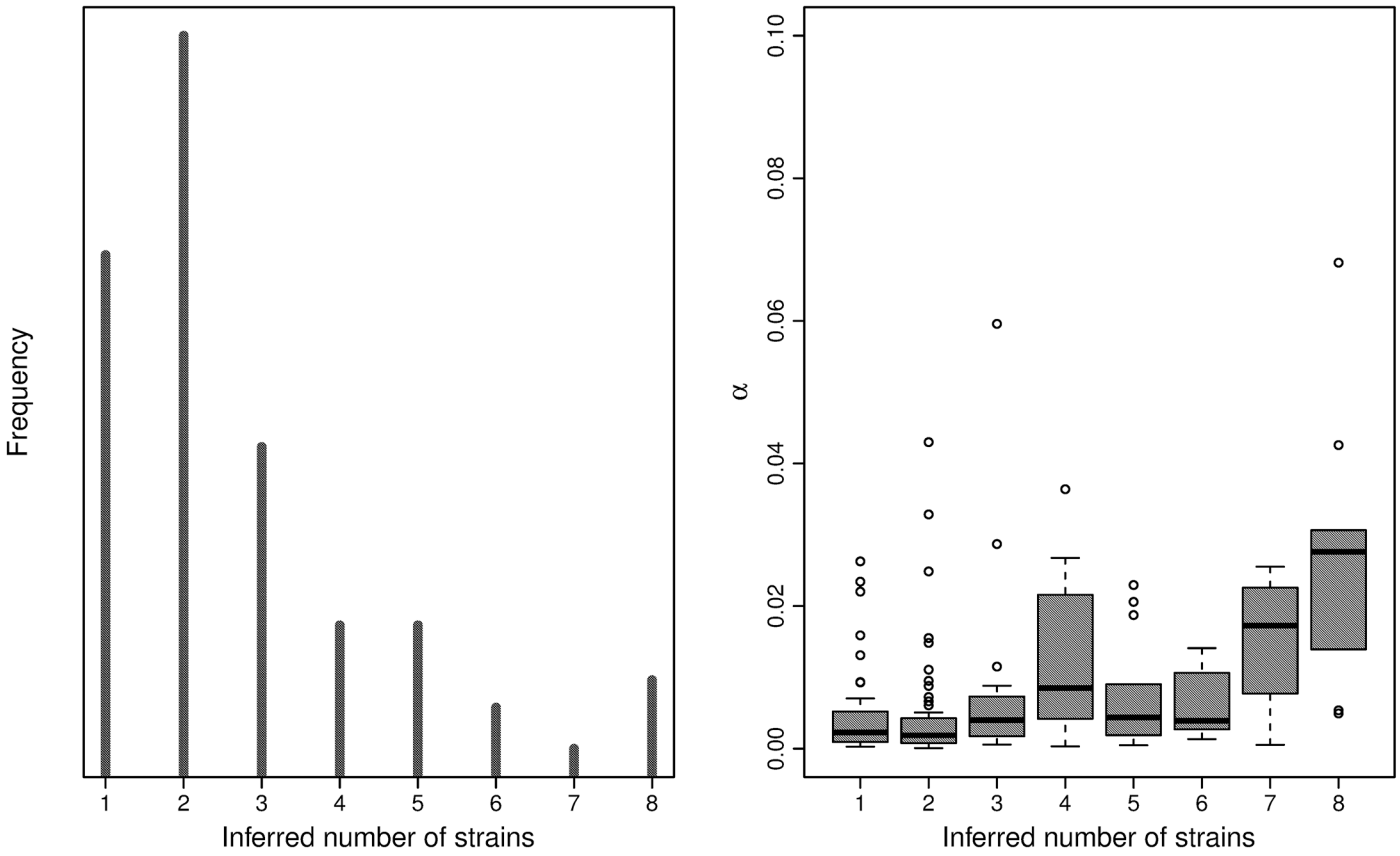

Fig 5. Ghanian sample summary. The frequency of inferred number of strains per sample (left) and and the panmixia coefficient by number of strains (right). MAP estimates used.

doi:10.1371/journal.pcbi.1004824.g005

To visually inspect the quality of the results, we generate figures for each of the samples showing the observed WSAF and PLAF data, the inferred model structure, and data simulated under the inferred model following the observed PLAF. We show examples of these plots for three typical samples in Fig 6. Nearly all samples (158/168), across all different mixture patterns, show strong visual correspondence between the observed and model-simulated data. Samples where PCR amplification was used (9 samples) exhibit no unusual features other than low values for α relative to the remaining samples. We also observe a strong correlation between the inferred number of components and an estimate for the inbreeding coefficient for each sample (Fig 7) [29]. These results are consistent with the high rate of MOI previously observed in Ghanaian clinicial samples [24, 44, 45].

Discussion

In this work we show how to infer strain mixture within Pf isolates using WGS with two improvements over previous efforts: an additional model for unexplained variation based on a panmixia and a reversible jump implementation that accounts for uncertainty in the underlying number of strains. Simulations show that the model can perform accurate inference (MSE < 0.05 for strain proportions) with as few as 50 SNPs and 10 read counts per SNP. Simulations with more than 100 SNPs or at least 25 read counts give highly accurate results (MSE < 0.02). In artificial laboratory mixtures the model provides excellent agreement with baseline mixture.

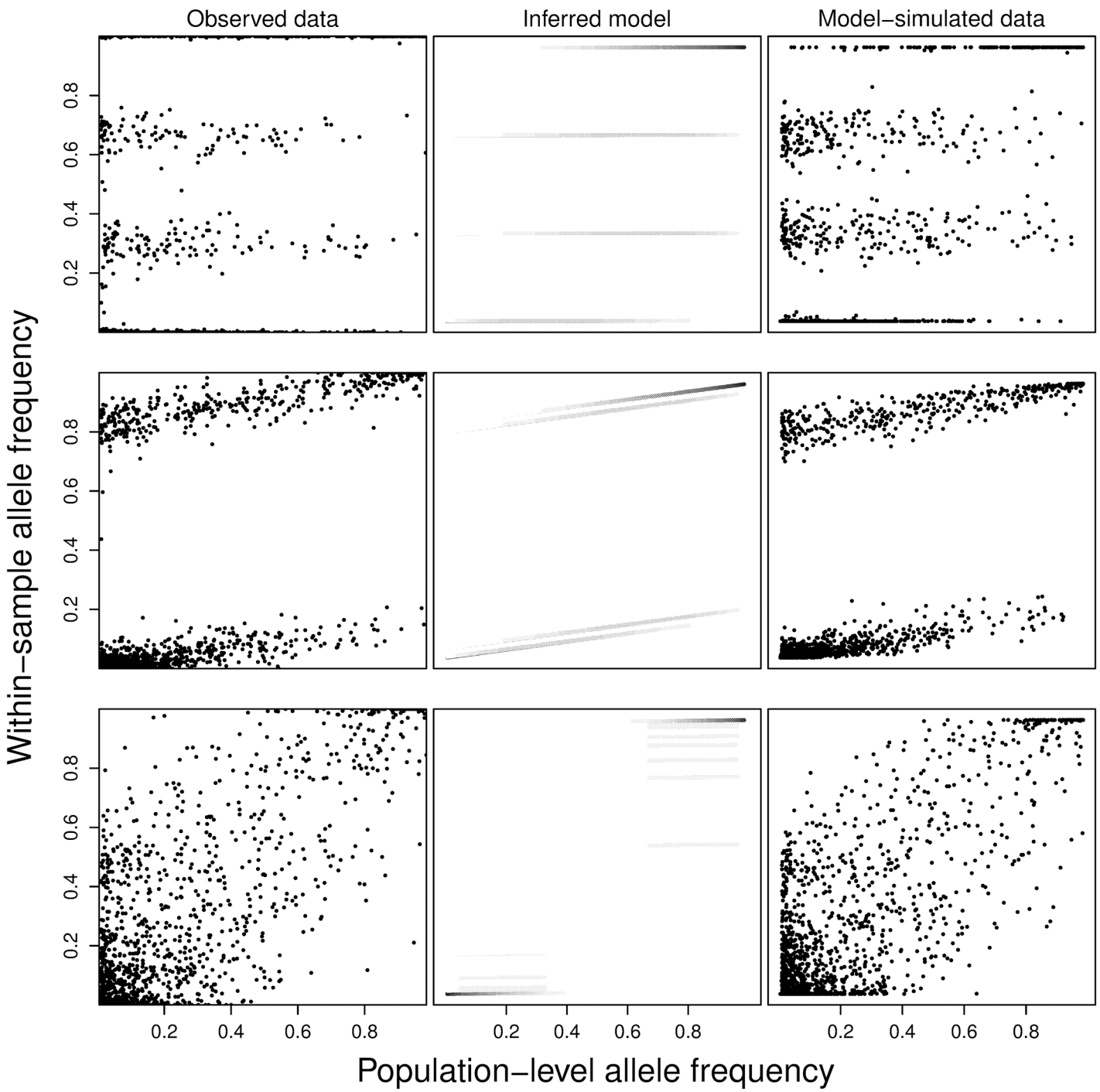

Fig 6. Examples of real and model-simulated data. For three samples (rows), we present the observed data WSAF plotted against the PLAF (first column), a diagram of the inferred model indicating the bands, proportions, and panmixia coefficient (second column), and data simulated under the inferred model. Panmixia coefficient and strain proportions are the MAP values. In the second column, the model's PLAF-varying mixture densities are shown in grey scale, with black equal to one.

doi:10.1371/journal.pcbi.1004824.g006

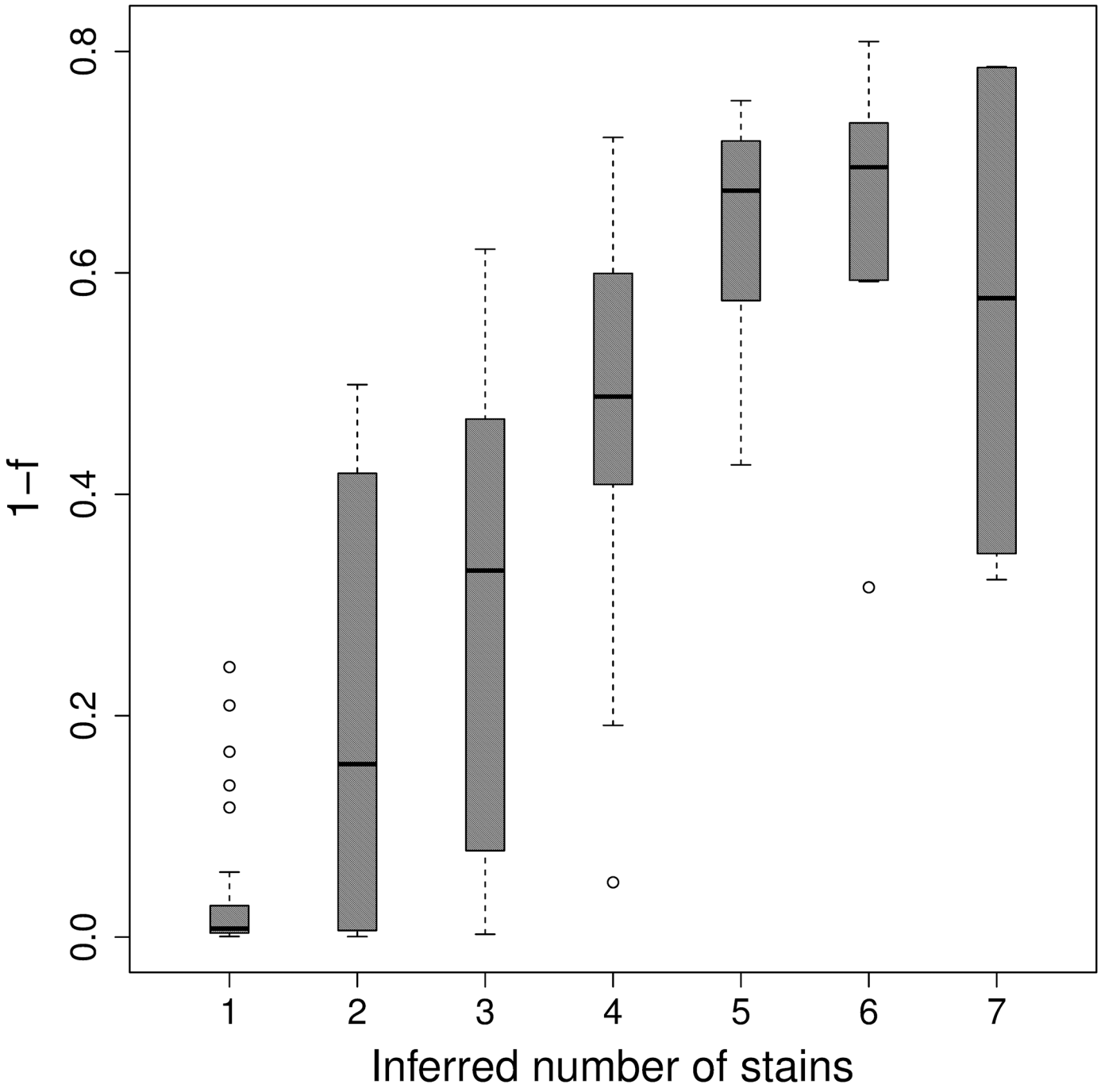

Fig 7. Number of strains by F-statistic. Boxplot of the inbreeding coefficient ($1 - F_{is}$) for each sample grouped by the MAP number of inferred strains.

doi:10.1371/journal.pcbi.1004824.g007

In field samples the model provides strong agreement with observed data and evidence based on Bayes factors indicates that some unexplained variation is present in a significant fraction of samples.

While the method works efficiently in practice, a number of possible improvements could strengthen its statistical performance. Most immediately, creating a full Bayesian approach rather than the parallelizing implementation here—while likely not improving the parametric inference for individual samples—would provide the full posterior distribution across all samples. The panmixia model is one of several possible approaches to dealing with additional within-sample variation that rigorous model comparison could reveal. The model also does not perform haplotype phasing to resolve the sequence of the underlying strains [46–48]. The analysis here suggests that a method for estimating haplotypes would be straight-forward for some samples but difficult for others (say, when α is greater than 0.05). Researchers may be particularly interested in whether, in these phased samples, particular stretches of the genome appear more or less frequently in the dominant strains than others, indicating structures of immunological or environmental selection. This is a natural avenue for statistical methods development.

The model makes a number of simplifying assumptions that may be violated in practice. The model presumes that SNPs are unlinked and consequently independent for the purpose of calculating the likelihood. Given the high recombination rate of Pf this assumption may hold for the majority of pairs of SNPs, but neglects correlations that appear locally ($\sim$ 10 kB). However, we expect that this independence assumption serves to moderately weaken the inferential power of the model rather than cause any type of bias since it effectively fails to include possibly informative data. More problematic is the model's implicit assumption of limited population structure. In the case of the KND samples, and perhaps in much of west Africa, this assumption appears supported [27, 49]. In other contexts, specifically southeast Asia, recent population bottlenecks and selection suggest that this assumption will be violated [50]. The consequences on this model inference are unknown but may be partially resolved with appropriate simulation studies.

The model will work with any technology capable of typing multiple variants and where the measurement of the fraction of non-reference variants is unbiased. It was developed for WGS data but is not specific to the sequencing employed and should work similarly for Illumina, 454 and Pacific Bioscience read technologies. As noted in the results, we observe that the small number of field samples where PCR amplification was used did not appear unusual other than exhibiting relatively low α values. This is could be due to preferential amplification of the dominant strains, suggesting that PCR-based approaches may obscure some aspects of natural infections. This model is not appropriate for data from RFLP assays or DNA microarrays without substantial modification.

In principle, the model can be explicitly tested against experimental mixtures more complex than those presented above. Laboratory facilities with the capacity to store many field strains ($>$100) could generate artificial samples in an experimental analog of our simulation procedure. Starting with N unmixed strains at equal dilution, they could create mixtures by first fixing the required sequencing volume as η, and the desired parameters for panmixia (α), number of component strains (K), and their mixture parameters, $\mathcal{W}$. For the finite mixture component, they would then combine volumes of $\eta \cdot \mathcal{W}$ from the K strains. For the panmixture component, they would then fix some large number but experimentally feasible number of strains (say 50) and randomly sample from all of them a volume of $\eta/50$. Combining these into a final sample and applying WGS sequencing, will yield data that we hypothesize will closely follow the integrated model outlined above, with ν capturing the experimental variation. Naturally, consistent results would indicate the sufficiency of the model but not its necessity, holding out the

possibility of a more minimal description. These results could be further compared against other next-generation technologies—such as single-cell sequencing—that have been deployed to understand Pf clinical mixtures [51].

The model presents an important new tool for interrogating the biology of clinical Pf infections. The ability to measure the structure of strain mixture connects to many aspects of Pf epidemiology including seasonality, transmission intensity, outcrossing, and rates of gene flow. It also presents a means for clarifying the poorly detailed structure of intra-host infection dynamics, such as strain selection or density-dependent selection [52], by resolving how the model parameters change within the course of an infection or in response to drug intervention. This approach can serve as a means for researchers to empirically resolve these hypotheses.

Supporting Information

S1 Table. Table of output values from algorithm applied to artificial laboratory mixture data.
(PDF)

S1 Text. Accession numbers for raw data.
(PDF)

S1 Fig. Cut-off for low-quality samples. Number of missing SNPs for each sample in ascending order (black dots) with the threshold used for cleaning (dotted blue line).
(PDF)

S2 Fig. Population structure of samples. Principal components (1–2, 1–3, 2–3) for samples and neighbor-joining tree of pairwise distance among samples both indicate limited population structure.
(PDF)

Acknowledgments

We thank Ana Lagunez for careful editing of the manuscript.

Author Contributions

Wrote the paper: JDO ZI LAE JW. Designed and implemented the study: JDO. Commented on the study design: ZI LAE. Generated artificial mixture data: JW. Performed all computational experiments: JDO. Performed data analysis: JDO. Commented on the analysis: ZI. Contributed all of the clinical data: LAE. Wrote and implemented the analysis tools: JDO. Commented on the development of the analysis tools: ZI.

References

1. Hay SI, Guerra CA, Gething PW, Patil AP, Tatem AJ, Noor AM, et al. A world malaria map: *Plasmodium falciparum* endemicity in 2007. PLoS Medicine. 2009; 6(3):e1000048. doi: 10.1371/journal.pmed.1000048 PMID: 19323591
2. Snow RW, Guerra CA, Noor AM, Myint HY, Hay SI. The global distribution of clinical episodes of *Plasmodium falciparum* malaria. Nature. 2005; 434(7030):214–217. doi: 10.1038/nature03342 PMID: 15759000
3. World Health Organization. World malaria report 2008. World Health Organization; 2008.
4. Wootton JC, Feng X, Ferdig MT, Cooper RA, Mu J, Baruch DI, et al. Genetic diversity and chloroquine selective sweeps in *Plasmodium falciparum*. Nature. 2002; 418(6895):320–323. doi: 10.1038/nature00813 PMID: 12124623
5. Mita T, Tanabe K, Kita K. Spread and evolution of *Plasmodium falciparum* drug resistance. Parasitology International. 2009; 58(3):201–209. doi: 10.1016/j.parint.2009.04.004 PMID: 19393762

6. Payne D. Spread of chloroquine resistance in *Plasmodium falciparum*. Parasitology Today. 1987; 3 (8):241–246. doi: 10.1016/0169-4758(87)90147-5 PMID: 15462966
7. Sidhu ABS, Verdier-Pinard D, Fidock DA. Chloroquine resistance in *Plasmodium falciparum* malaria parasites conferred by pfcrt mutations. Science. 2002; 298(5591):210–213. doi: 10.1126/science.1074045 PMID: 12364805
8. Roper C, Pearce R, Nair S, Sharp B, Nosten F, Anderson T. Intercontinental spread of pyrimethamine-resistant malaria. Science. 2004; 305(5687):1124–1124. doi: 10.1126/science.1098876 PMID: 15326348
9. Wilson R, McGregor I, Williams K, Hall P, Bartholomew R. Antigens associated with *Plasmodium falciparum* infections in man. The Lancet. 1969; 294(7613):201–205. doi: 10.1016/S0140-6736(69)91437-8
10. McGregor I. Immunology of malarial infection and its possible consequences. British Medical Bulletin. 1972; 28(1):22–27. PMID: 4118010
11. Jamjoom GA. Dark-field microscopy for detection of malaria in unstained blood films. Journal of Clinical Microbiology. 1983; 17(5):717–721. PMID: 6863496
12. Conway D, Greenwood B, McBride J. The epidemiology of multiple-clone *Plasmodium falciparum* infections in Gambian patients. Parasitology. 1991; 103(Pt 1):1–6. doi: 10.1017/S0031182000059217 PMID: 1682870
13. Auburn S, Campino S, Miotto O, Djimde AA, Zongo I, Manske M, et al. Characterization of within-host *Plasmodium falciparum* diversity using next-generation sequence data. PloS one. 2012; 7(2):e32891. doi: 10.1371/journal.pone.0032891 PMID: 22393456
14. Müller D, Charlwood J, Felger I, Ferreira C, Do Rosario V, Smith T. Prospective risk of morbidity in relation to multiplicity of infection with *Plasmodium falciparum* in São Tomé. Acta tropica. 2001; 78(2):155–162. doi: 10.1016/S0001-706X(01)00067-5 PMID: 11230825
15. Henning L, Schellenberg D, Smith T, Henning D, Alonso P, Tanner M, et al. A prospective study of *Plasmodium falciparum* multiplicity of infection and morbidity in Tanzanian children. Transactions of the Royal Society of Tropical Medicine and Hygiene. 2004; 98(12):687–694. doi: 10.1016/j.trstmh.2004.03.010 PMID: 15485698
16. Smith T, Beck HP, Kitua A, Mwankusye S, Felger I, Fraser-Hurt N, et al. 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. 1999; 93(Supplement 1):15–20. doi: 10.1016/S0035-9203(99)90322-X PMID: 10450421
17. Färnert A, Rooth I, Svensson Å, Snounou G, Björkman A. Complexity of *Plasmodium falciparum* infections is consistent over time and protects against clinical disease in Tanzanian children. Journal of infectious diseases. 1999; 179(4):989–995. doi: 10.1086/314652 PMID: 10068596
18. Stirnadel HA, Felger I, Smith T, Tanner M, Beck HP, et al. Malaria infection and morbidity in infants in relation to genetic polymorphisms in Tanzania. Tropical Medicine & International Health. 1999; 4 (3):187–193. doi: 10.1046/j.1365-3156.1999.43381.x
19. Beck S, Mockenhaupt FP, Bienzle U, Eggelte TA, Thompson W, Stark K. Multiplicity of *Plasmodium falciparum* infection in pregnancy. The American Journal of Tropical Medicine and Hygiene. 2001; 65 (5):631–636. PMID: 11716126
20. Beck HP, Felger I, Vounatsou P, Hirt R, Tanner M, Alonso P, et al. 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. 1999; 93(Supplement 1):41–45. doi: 10.1016/S0035-9203(99)90326-7 PMID: 10450425
21. Smith T, Felger I, Fraser-Hurt N, Beck HP. 10. Effect of insecticide-treated bed nets on the dynamics of multiple *Plasmodium falciparum* infections. Transactions of the Royal Society of Tropical Medicine and Hygiene. 1999; 93(Supplement 1):53–57. doi: 10.1016/S0035-9203(99)90328-0 PMID: 10450427
22. Paganotti GM, Babiker HA, Modiano D, Sirima BS, Verra F, Konate A, et al. Genetic complexity of *Plasmodium falciparum* in two ethnic groups of Burkina Faso with marked differences in susceptibility to malaria. The American Journal of Tropical Medicine and Hygiene. 2004; 71(2):173–178. PMID: 15306706
23. Mayengue PI, Luty AJ, Rogier C, Baragatti M, Kremsner PG, Ntoumi F. The multiplicity of *Plasmodium falciparum* infections is associated with acquired immunity to asexual blood stage antigens. Microbes and Infection. 2009; 11(1):108–114. doi: 10.1016/j.micinf.2008.10.012 PMID: 19028595
24. Kobbe R, Neuhoff R, Marks F, Adjei S, Langefeld I, Von Reden C, et al. Seasonal variation and high multiplicity of first Plasmodium falciparum infections in children from a holoendemic area in Ghana, West Africa. Tropical Medicine & International Health. 2006; 11(5):613–619. doi: 10.1111/j.1365-3156.2006.01618.x

25. Atroosh WM, Al-Mekhlafi HM, Mahdy MA, Saif-Ali R, Al-Mekhlafi AM, Surin J. Genetic diversity of Plasmodium falciparum isolates from Pahang, Malaysia based on MSP-1 and MSP-2 genes. Parasit Vectors. 2011; 4(4):233. doi: 10.1186/1756-3305-4-233 PMID: 22166488

26. Joshi H, Valecha N, Verma A, Kaul A, Mallick PK, Shalini S, et al. Genetic structure of Plasmodium falciparum field isolates in eastern and north-eastern India. Malar J. 2007; 6(60):10–1186.

27. Manske M, Miotto O, Campino S, Auburn S, Almagro-Garcia J, Maslen G, et al. Analysis of Plasmodium falciparum diversity in natural infections by deep sequencing. Nature. 2012; 487(7407):375–379. doi: 10.1038/nature11174 PMID: 22722859

28. Auburn S, Campino S, Clark TG, Djimde AA, Zongo I, Pinches R, et al. An effective method to purify *Plasmodium falciparum* DNA directly from clinical blood samples for whole genome high-throughput sequencing. PLoS ONE. 2011; 6(7):e22213. doi: 10.1371/journal.pone.0022213 PMID: 21789235

29. O'Brien J, Li R, Amenga-Etego L. Approaches to estimating inbreeding coefficients in clinical isolates of *Plasmodium falciparum* from genomic sequence data. bioRxiv. 2015;p. e021519.

30. Weir BS, Cockerham CC. Estimating F-statistics for the analysis of population structure. Evolution. 1984;p. 1358–1370. doi: 10.2307/2408641

31. Hill WG, Babiker HA. Estimation of numbers of malaria clones in blood samples. Proceedings of the Royal Society of London Series B: Biological Sciences. 1995; 262(1365):249–257. doi: 10.1098/rspb.1995.0203 PMID: 8587883

32. Guerra CA, Gikandi PW, Tatem AJ, Noor AM, Smith DL, Hay SI, et al. The limits and intensity of *Plasmodium falciparum* transmission: implications for malaria control and elimination worldwide. PLoS medicine. 2008; 5(2):e38. doi: 10.1371/journal.pmed.0050038 PMID: 18303939

33. Balding DJ. Likelihood-based inference for genetic correlation coefficients. Theoretical population biology. 2003; 63(3):221–230. doi: 10.1016/S0040-5809(03)00007-8 PMID: 12689793

34. Galinsky K, Valim C, Salmier A, de Thoisy B, Musset L, Legrand E, et al. COIL: a methodology for evaluating malarial complexity of infection using likelihood from single nucleotide polymorphism data. Malaria Journal. 2015; 14(1):4. doi: 10.1186/1475-2875-14-4 PMID: 25599890

35. Wendler J. Accessing complex genomic variation in Plasmodium falciparum natural infection. Doctoral dissertation, University of Oxford. 2015;.

36. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, et al. Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature. 2002; 419(6906):498–511. doi: 10.1038/nature01097 PMID: 12368864

37. Redner RA, Walker HF. Mixture densities, maximum likelihood and the EM algorithm. SIAM review. 1984; 26(2):195–239. doi: 10.1137/1026034

38. McLachlan G, Peel D. Finite mixture models. John Wiley & Sons; 2004.

39. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian data analysis. CRC press; 2013.

40. Gilks WR. Markov chain Monte Carlo. Wiley Online Library; 2005.

41. Geyer CJ. Practical Markov chain Monte Carlo. Statistical Science. 1992;p. 473–483. doi: 10.1214/ss/1177011137

42. Scholz F. Maximum likelihood estimation. Encyclopedia of statistical sciences. 1985;.

43. Suchard MA, Weiss RE, Sinsheimer JS. Bayesian selection of continuous-time Markov chain evolutionary models. Molecular Biology and Evolution. 2001; 18(6):1001–1013. doi: 10.1093/oxfordjournals.molbev.a003872 PMID: 11371589

44. Owusu-Agyei S, Asante KP, Adjuik M, Adjei G, Awini E, Adams M, et al. Epidemiology of malaria in the forest-savanna transitional zone of Ghana. Malar J. 2009; 8(1):220. doi: 10.1186/1475-2875-8-220 PMID: 19785766

45. Felger I, Maire M, Bretscher MT, Falk N, Tiaden A, Sama W, et al. The dynamics of natural Plasmodium falciparum infections. PLoS One. 2012; 7(9):e45542. doi: 10.1371/journal.pone.0045542 PMID: 23029082

46. Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. The American Journal of Human Genetics. 2001; 68(4):978–989. doi: 10.1086/319501 PMID: 11254454

47. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nature Genetics. 2012; 44(8):955–959. doi: 10.1038/ng.2354 PMID: 22820512

48. O'Brien JD, Didelot X, Iqbal Z, Amenga-Etego L, Ahiska B, Falush D. A Bayesian approach to inferring the phylogenetic structure of communities from metagenomic data. Genetics. 2014; 197(3):925–937. doi: 10.1534/genetics.114.161299 PMID: 24793089

49. Anderson TJ, Haubold B, Williams JT, Estrada-Franco JG, Richardson L, Mollinedo R, et al. Microsatellite markers reveal a spectrum of population structures in the malaria parasite *Plasmodium falciparum*. Molecular Biology and Evolution. 2000; 17(10):1467–1482. doi: 10.1093/oxfordjournals.molbev.a026247 PMID: 11018154

50. Miotto O, Almagro-Garcia J, Manske M, MacInnis B, Campino S, Rockett KA, et al. Multiple populations of artemisinin-resistant *Plasmodium falciparum* in Cambodia. Nature Genetics. 2013; 45(6):648–655. doi: 10.1038/ng.2624 PMID: 23624527

51. Nair S, Nkhoma SC, Serre D, Zimmerman PA, Gorena K, Daniel BJ, et al. Single-cell genomics for dissection of complex malaria infections. Genome research. 2014; 24(6):1028–1038. doi: 10.1101/gr.168286.113 PMID: 24812326

52. Kwiatkowski D, Nowak M. Periodic and chaotic host-parasite interactions in human malaria. Proceedings of the National Academy of Sciences. 1991; 88(12):5111–5113. doi: 10.1073/pnas.88.12.5111