A cautionary note on the use of machine learning algorithms to characterise malaria parasite population structure from genetic distance matrices

James A Watson[1,2], Aimee R Taylor[3,4], Elizabeth A Ashley[1,2], Arjen Dondorp[1,2], Caroline O Buckee[3,4], Nicholas J White[1,2], Chris C Holmes[5,6]

1 Mahidol-Oxford Tropical Medicine Research Unit, Faculty of Tropical Medicine, Mahidol University, Thailand
2 Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford, United Kingdom
3 Center for Communicable Disease Dynamics, Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, MA 02115, USA
4 Broad Institute of MIT and Harvard, Cambridge, MA, 02142, USA
5 Department of Statistics, University of Oxford, United Kingdom
6 Nuffield Department of Medicine, University of Oxford, United Kingdom

Abstract

Genetic surveillance of malaria parasites supports malaria control programmes, treatment guidelines and elimination strategies. Surveillance studies often pose questions about malaria parasite ancestry (e.g. about the spread of antimalarial resistance) and employ methods that characterise parasite population structure. Many of the methods used to characterise structure are algorithms developed in machine learning (ML) and depend on a genetic distance matrix, e.g. principal coordinates analysis (PCoA) and hierarchical agglomerative clustering (HAC). However, PCoA and HAC are sensitive to both the definition of genetic distance and algorithmic specification. Importantly, neither algorithm generates an inferred malaria parasite ancestry. As such, PCoA and HAC can support (e.g. via exploratory visualisation and hypothesis generation), but not answer comprehensively, key questions about malaria parasite ancestry. We illustrate the sensitivity of PCoA and HAC using 393 *P. falciparum* whole genome sequences collected from Cambodia and neighbouring regions (where antimalarial resistance has emerged and spread recently) and we provide tentative guidance for the use of PCoA and HAC in malaria parasite genetic epidemiology. This guidance includes a call for fully transparent and reproducible analysis pipelines that feature (i) a clearly outlined scientific question; (ii) clear justification of methods used along with discussion of any inferential limitations; (iii) publicly available genetic distance matrices; and (iv) sensitivity analyses. To bridge the inferential disconnect between the output of the ML algorithms and the scientific questions of interest, tailor-made statistical models are needed to infer malaria parasite ancestry. In the absence of such models speculative reasoning should feature only as discussion but not as results.

Introduction

As part of a global push to eliminate malaria, there are ongoing efforts to roll out routine molecular surveillance to understand how the parasites causing malaria are spreading, to track drug resistance and identify its emergence, and to understand where to target interventions. Malaria parasite genetic data have been accrued increasingly rapidly in recent years, but the analytical methods for making sense of them lag behind. In particular, the methods used currently to infer malaria parasite ancestry, which is often central to the questions posed by studies of malaria parasite genetic epidemiology, have significant limitations. An important goal in malaria parasite genetic epidemiology is inference of the full ancestral recombination graph. However, there are no scalable methods directly applicable to malaria parasites. Instead, the first step towards inferring

ancestry involves characterisation of contemporary population genetic structure using computationally tractable methods. Many questions of clinical and public health relevance, for example, interpreting reduced haplotype diversity as a selective sweep, rely on methods that first characterise the underlying population structure within the *Plasmodium* species.

Seminal methods for characterising and quantifying population structure in general include EIGENSTRAT, STRUCTURE, fineSTRUCTURE and ADMIXTURE. EIGENSTRAT provides a continuous estimate of population structure (a projection onto a linear combination of genetic variants) and is typically used in the context of genome wide association studies to adjust for population stratification [1]. Methods such as STRUCTURE [2], fineSTRUCTURE [3] and ADMIXTURE [4] perform population assignment (i.e. assign individuals to one or more inferred ancestral populations). These methods have been developed primarily in the context of human statistical genetics. They are only loosely applicable to analysing genetic data from human malaria parasites: eukaryotic pathogens of the genus *Plasmodium* that are transmitted by anopheline mosquitoes. Malaria parasites are haploid during the human host stage of their life-cycle, but undergo an obligate stage of sexual recombination in the mosquito. Although both humans and malaria parasites recombine sexually, the added complexity of the malaria parasite life cycle complicates and confounds *Plasmodium* population genetics. For example, unlike humans, malaria parasites can self (recombination between genetically identical male and female gametes), and the rate of selfing varies with transmission intensity.

In addition to methods borrowed from human genetics, context agnostic unsupervised machine learning (ML) algorithms have been applied to *Plasmodium* genetic data to characterise population structure (e.g. [5–7]). Unsupervised methods (in statistics and ML) can detect patterns and structure in data without needing classifications or categories as inputs. Principal coordinates analysis (PCoA) and hierarchical agglomerative clustering (HAC) are the two most popular classes of unsupervised ML algorithms used in genetic epidemiological studies of malaria. PCoA is a generalisation of principal components analysis (PCA). PCA has deep connections with statistical model-based clustering analysis (e.g. STRUCTURE and related methods) [3] and, under certain assumptions, has a genealogical interpretation [8]. In contrast, unsupervised ML clustering algorithms such as HAC do not, in general, have a genealogical interpretation and are non-inferential (they do not have parameters whose values are inferred from the data). This limits the scope of the scientific questions they can answer. There are however important stopgap methods and remain important conceptual aids.

A common feature of many computational analysis pipelines used in both human and *Plasmodium* population genetics is the genetic distance matrix. Typically, the genetic data for all (or a subset of) polymorphic loci for a set of n isolates are summarised into an n-by-n matrix of scalar distances or similarities (distance is inversely related to similarity). The distance or similarity used is context specific and can either be computed directly as a function of the genetic data or inferred under a statistical model fit to the data [9]. For example, fineSTRUCTURE is a program that performs statistical model-based clustering of a co-ancestry matrix (a similarity matrix) generated by a process called chromosome painting (using the program ChromoPainter) which is based on a statistical model [3, 10]. PCoA and HAC both take as inputs arbitrary genetic distance matrices. The assumption in fineSTRUCTURE is that the co-ancestry matrix is a (nearly) sufficient statistic: i.e. all of the information in the data that is needed to infer the parameters of the clustering model is contained in the co-ancestry matrix. In a similar way, PCoA and HAC only use as input the summarised information in the distance matrix. Summarising the data in this succinct way has many advantages. For example, when dealing with whole genome sequence (WGS) data, the number of isolates is usually many orders of magnitude smaller than the number of loci. Thus the distance matrix is computationally easier to analyse (proportional to the square of the number of isolates). In addition it is not possible to reverse engineer the distance matrix to obtain the original genetic data, making it a privacy preserving summary of the genetic data. Careful consideration of which genetic distance should be used to construct this matrix is needed. Making the distance matrix fully available along with the code renders all analyses dependent on this data summary fully reproducible.

ML algorithms applied to genetic distance matrices summarising WGS data from the Eastern Greater Mekong subregion (GMS) have been used to characterise and quantify the current epidemic of multi-drug resistant *P. falciparum* parasites [5–7, 11]. For example, two recent studies used PCoA and HAC to determine *P. falciparum* population structure in the Eastern GMS [5, 6]. Along with other studies, they made two important observations. The first observation, made before the

discovery of the causative mutations in the *Pfkelch13* gene which confer artemisinin resistance, was the identification of multiple apparently sympatric sub-populations in the Eastern GMS [5]. These sub-populations were identified primarily using PCoA of a genetic distance matrix based on identity-by-state (IBS; see Methods) [5]. The second observation was the spread of a single multi-drug resistant parasite lineage across the eastern GMS. This observation was first made using a small panel of microsatellite markers flanking the *Pfkelch13* gene which showed that a single long haplotype (bearing the *Pfkelch13* mutation C580Y and named *Pf*Pailin, as it was observed first in the Pailin area of Western Cambodia) had spread across Cambodia into North East Thailand, Southern Lao PDR and Southern Vietnam [12, 13]. This parasite lineage also acquired resistance to piperaquine, manifest by amplification in the *Pfplasmepsin* gene and mutations in *Pfcrt*. A subsequent study used WGS data and presented qualitatively similar results; *Pf*Pailin was renamed the KEL1/PLA1 lineage [6]. This latter study was based on HAC of a genetic distance matrix derived from chromosome painting [6]; it is the second of our examples that features PCoA and HAC [5, 6]. These two observations and associated studies [5–7, 11–13] are of considerable importance as it is in this same region of Southeast Asia that antimalarial drug resistance arose previously, first to chloroquine and then to sulphadoxine-pyrimethamine, before spreading to India and Africa at a cost of millions of lives. Now *P. falciparum* has evolved resistance to both artemisinin [14] and to piperaquine [15], the two active components of the artemisinin combination therapy (ACT) dihydroartemisinin-piperaquine [16]. Dihydroartemisinin-piperaquine was the first line treatment for falciparum malaria in Eastern Thailand, Cambodia and Vietnam until worsening resistance forced a change in policy. There has been controversy over the role of artemisinin resistance in facilitating the emergence of partner drug resistance. Thus understanding whether artemisinin resistance promoted the selection of piperaquine resistance in the Eastern GMS (or mefloquine resistance on the Thailand-Myanmar border), and if so by how much, is of contemporary relevance for antimalarial drug policies and practices. Genetic epidemiological analyses have helped to answer these questions but there is a disconnect between the scientific questions of interest and those that methods such as PCoA and HAC can address. In addition, there is a lack of statistical models articulated around specific hypotheses which could help bridge this disconnect (an area of open research).

In this work we undertake an empirical sensitivity analysis of PCoA and HAC with respect to the genetic distance measure used and to the algorithm specification. Our analysis uses sequence data from *P. falciparum* parasites isolated from a subset of patients enrolled in the Eastern GMS sites (Cambodia and neighbouring areas of Thailand, Laos and Vietnam) of a large clinical trial that took place between 2011 and 2013 [14]. We provide some tentative guidance as to the use and interpretation of PCoA and HAC in the context of *Plasmodium* population genetics, including ways to increase the transparency and reproducibility of analytic pipelines (e.g. by sharing key intermediate data summaries such as genetic distance matrices). In the discussion we make a call for the development of more statistical models of malaria parasite ancestry.

Methods

Data

The *P. falciparum* data used in this study are from isolates obtained from patients presenting to rural clinics and health centres in Eastern Thailand (Sri Saket and Ubon Ratchatani provinces), Western and North Eastern Cambodia, Southern Lao PDR and Southern Vietnam between 2011 and 2013 during the Tracking Resistance to Artemisinin (TRAC) study [14]. This is a subset of all isolates collected during the TRAC study. It was chosen for simplicity: the parasite isolates come from a subset of locations that represent a small (and thought to be interconnected in terms of transmission) geographic area of low seasonal malaria transmission. The subset contains WGS data on 52308 single nucleotide polymorphisms (SNPs) that vary across 393 *P. falciparum* isolates that were deemed mononclonal based on a measure of within-host diversity [17] (cutoff at 0.95 of the within-host diversity measure).

The data were used to construct matrices of genetic distances between pairs of monoclonal isolates (see next section). Strictly speaking, each "monoclonal" isolate represents a collection of malaria parasites that can be represented by a single parasite genotype.

Matrices of genetic distances between pairs of monoclonal isolates

WGS data are high dimensional. Many computational analysis pipelines first proceed by applying a data processing step to reduce dimensionality. This often takes the form of the construction of a genetic distance or similarity matrix: a n-by-n matrix that is much smaller than the original n-by-p data matrix (where n is the number of isolates and $p >> n$ is the number of loci sequenced). Hereafter we refer explicitly to distances only, noting that similarities can be derived from distances and vice versa. A distance or similarity matrix is not necessarily symmetric, e.g. the co-ancestry matrix produced by the chromosome painting is in general asymmetric [3, 10].

The genetic distances that feature within the distance matrix can be either computed directly as a function of the genetic data (e.g. the fraction of loci where observed alleles differ) or inferred under a statistical model (e.g. the model underpinning chromosome painting [10]). The distance tends be more interpretable when inferred under a statistical model that references a particular evolutionary process, however the inferential process is more involved, and thus possibly less transparent.

A genetic distance can reflect one or more evolutionary processes: mutation, recombination, or both. For example, between a pair of non-recombining organisms (e.g. viruses), the fraction of loci where observed alleles differ is measured in units of mutation since the most recent common ancestor (MRCA), and can be used to infer the time since the MRCA, e.g. under a model that assumes constant rate mutation. Between a pair of recombining organisms (e.g. malaria parasites), the fraction of loci where observed alleles differ can be measured in units of mutation using data on regions of the genome that do not recombine, e.g. mitochondrial DNA. Otherwise, both mutation and recombination contribute to observed differences. A genetic distance that explicitly references recombination can be obtained via inference under a statistical model.

The best choice of genetic distance depends on which process generates variation over the appropriate time-scale. For example, among malaria parasites that outcross frequently, distances (either inferred or not) to which recombination contributes tend to reflect the recent past, as recombination introduces variation at each outcrossing event. Mutation-based distances reflect the more distant past because the region of the genome that does not recombine is small and the mutation rate is relatively low (e.g. compared to that of rapidly evolving viruses). As an aside, population genetic distance measures based on allele frequencies (e.g. F_{ST}) can also reflect the more distant past for similar reasons; however, our current focus is on distances that are defined between pairs of isolates, not populations, and measures such as F_{ST} are available only on a population level. The aspect of time-scale is sometimes ignored in genetic analyses of *Plasmodia* but it is essential.

In this study, we use data on 52308 biallelic SNPs from the core *P. falciparum* genome and consider three distinct genetic distances. In the following, i, j index parasite isolates, k indexes the p sequenced loci, and y_{ik} denotes the observed allele for i-th isolate at the k-th locus.

1. The fraction of loci where observed alleles differ (denoted 1-IBS hereafter):

$$1 - \widehat{\mathrm{IBS}}_{ij} = 1 - \frac{1}{p}\sum_{k=1}^{p} \mathrm{IBS}_{ijk}, \tag{1}$$

   where $\mathrm{IBS}_{ijk} = 1$ if $y_{ik} = y_{jk}$ and 0 otherwise. In other words, $\mathrm{IBS}_{ijk} = 1$ if observed alleles y_{ik} and y_{jk} are identical-by-state (IBS). Each IBS_{ijk} is observed. As such, $1 - \widehat{\mathrm{IBS}}_{ij}$ can be computed directly as a function of genetic data. Note that equation (1) does not feature a linkage disequilbrium correction (used elsewhere e.g. [18]). It does not feature allele frequencies and it does not reference any particular process. As stated above, both mutation and recombination contribute to its value. Although it is not inferred, 1-IBS is thus a distance to which recombination contributes. Its expectation is a linear function of relatedness [9].

2. Unrelatedness estimate (denoted 1-IBD):

$$1 - \widehat{r}_{ij} = 1 - \mathbb{P}\left(\mathrm{IBD}_{ijk} = 1\right) \text{ for all } k = 1, \ldots, p, \tag{2}$$

   where $\mathrm{IBD}_{ijk} = 1$ if $y_{ik} = y_{jk}$ and are both descended from a recent common ancestor, 0 otherwise. In other words, $\mathrm{IBD}_{ijk} = 1$ if observed alleles y_{ik} and y_{jk} are identical-by-descent (IBD). Each IBD_{ijk} is unobserved. As such, each relatedness estimate, $\widehat{r}_{ij}$, is inferred under a statistical model. The statistical model is typically a hidden Markov model (HMM) that

features allele frequencies and explicitly references sexual recombination to some extent, e.g. [19]. We used R code available via [9] to estimate relatedness. Unrelatedness is thus an inferred genetic distance to which recombination contributes.

3. An approximation of the number of outcrossed generations since the MRCA (denoted $-\log_2 \text{IBD}$):

$$-\log_2 \widehat{r}_{ij}. \tag{3}$$

This approximation is based on the expectation that relatedness is halved each time fully outbred parasites recombine sexually. The number of outcrossed generations is less than or equal to the number of life cycle generations: although the malaria parasite life cycle involves an obligate stage of sexual recombination, malaria parasites can self.

Unsupervised learning algorithms in statistical genetics

Principal coordinates analysis

Principal components analysis (PCA), also referred to as eigenanalysis, is an essential method for the assessment of population structure [1]. PCA usually refers to an eigenanalysis of the sequence covariance matrix. However, it is possible to do an eigenanalysis on any genetic distance matrix, for any user-defined genetic distance. This is referred to as principal coordinates analysis (PCoA).

PCA on the covariance matrix has a solid theoretical footing in statistical genetics [8, 20]. Different patterns of divergent evolution and admixture lead to specific patterns in the top principal components. A genealogical interpretation can be given to the output of a PCA under the coalescent model, but with limitations [8]. Multiple histories may lead to the same PCA structure and uneven sampling can skew the output severely [8, 21]. There does not exist currently the same theoretical footing for PCoA on arbitrary genetic distances matrices. Therefore PCoA is primarily a technique for visualising structure in the genetic distance matrix and for reducing dimensionality to some number of large eigenvalues that approximates the number of dimensions on which the data vary. The detection of clusters from PCoA or PCA is sometimes estimated directly from the first two (or more) principal components of the distance matrix, as reported recently in a large analysis of *P. falciparum* parasites from the African continent [22].

Hierarchical agglomerative clustering

Unsupervised clustering methods are important tools in ML and statistics more broadly. The goal of an unsupervised ML clustering algorithm is to partition a set of observations (agnostic of the data type, whether genetic or other) into a discrete number of subsets (clusters). This partition should in theory minimise the distances between members of the same clusters (intra-cluster), and maximise the distances between members of different clusters (inter-cluster). The main class of unsupervised ML clustering algorithms used in malaria genetic epidemiology is hierarchical agglomerative clustering (HAC). This is a class of bottom-up clustering algorithms which operate directly on distance matrices. HAC includes as special cases UPGMA (UPGMA stands for unweighted pair group method with arithmetic mean), and neighbour-joining. By bottom-up, we mean that at the first iteration of the algorithm, each observation is in its own cluster. At subsequent iterations, the two 'nearest' distinct clusters are merged. At the first iteration of the algorithm, 'nearest' is defined by the distance matrix, subsequently it is defined by the linkage function applied to the distance matrix (see below). Therefore, a HAC algorithm necessitates specification of the following:

- A definition of distance between pairs of observations,
- A linkage function that determines the distance between sets of observations (the inter-cluster distance), which is used to assess a proposed joining of clusters.

HAC algorithms are 'greedy' (greedy in computer science refers to the process of local optimisation, i.e. only looking one step ahead at each iteration of the algorithm). The linkage function determines the update at each iteration and is key to the specification of the algorithm. The most common choices reported in the literature are to use average linkage (average inter-cluster distance), single linkage (the minimum inter-cluster distance, known as closest neighbours), complete linkage (the maximum inter-cluster distance, known as furthest neighbours), or Ward's criterion (the total within cluster variance, applicable to cases where the distance matrix is given by squared

Euclidean distances). HAC algorithms do not output a set of discrete clusters explicitly but instead they provide a set of nested clustering arrangements. This nested clustering arrangement can be visualised as a dendrogram which depicts a possible underlying data structure based on the distance matrix. The y-axis scale on the dendrogram (corresponding to the branch lengths) is proportional to the inter-cluster distance. It is possible to 'flatten' this dendrogram by choosing a cut-off point on the y-axis. This cut is either done by specifying a number of desired clusters and thus finding an appropriate value on the y-axis to obtain this number, or by just specifying a particular value on the y-axis (in units of inter-cluster distance). Importantly, both are arbitrary choices.

Neighbour-joining is a special case of a HAC algorithm [23]. This was designed specifically for the construction of non-rooted phylogenetic trees based on an evolutionary distance measured in units of mutation. The motivation for the neighbour-joining algorithm is finding the most parsimonious phylogenetic tree with respect to the data, whereby parsimonious means that the fewest number of evolutionary events (e.g. mutation events) are needed to explain the differences between observations. The algorithm starts from a star-like tree structure and then iteratively joins 'neighbours' where neighbours are determined by a joining event that minimises the total length of all the tree branches in the subsequent tree (the minimum total branch length here implies the most parsimonious tree representation). When attempting to determine ancestry the objective is to build a phylogenetic structure. But for observations where differences are mainly driven by recombination and not mutation or drift (e.g. sexually recombining malaria parasites), genetic distance based on IBD (or its correlate, IBS) is not linearly proportional to evolutionary genetic distance driven by mutation.

Statistical analysis

All statistical analyses were performed in R version 3.6.0 [24]. PCoA and HAC were applied to three distance matrices each featuring different distances outlined above: $1 - \text{IBS}$ (equation (1)), $1 - \text{IBD}$ (equation (2)), and $-\log_2 \text{IBD}$ (equation (3)). For the $-\log_2 \text{IBD}$ genetic distance, if relatedness was estimated to be exactly zero (i.e. 1-IBD=1), then we replaced it by the approximated lower limit of quantification of IBD (the lowest observed non-zero IBD value). The PCoA was performed using the R function *cmdscale*. HAC was performed using the R package *fastcluster* [25]. Four separate specifications of the HAC algorithm were considered, each using a different linkage function: average linkage, complete linkage, single linkage, and Ward's criterion. The function *cutree* (R package *dendextend* [26]) was used to collapse the resulting dendrograms into nine discrete clusters. The choice of nine was arbitrary: there is no principled approach for choosing the number of clusters into which to cut the dendrograms.

Code and data availability

The three genetic distance matrices (data summaries) and code used to generate the results reported here are openly available at the following github repository: `https://github.com/jwatowatson/sensitive-plasmodium-structure`. The raw sequencing data files are available from the Pf3K project (`https://www.malariagen.net/projects/pf3k`). Identification codes for the isolates used in this study are in the github repository.

Results

Computational analysis pipelines in malaria genetic epidemiology studies

The computational analysis pipelines of many genetic epidemiology studies of *Plasmodium spp.* follow the steps shown in the schematic outline in Figure 1. Genetic data (e.g. SNP calls, which are themselves the output of a complex computational pipeline) are summarised into a genetic distance matrix. The distance matrix is then input into methods which can help visualise structure such as PCoA, or can help assign a population structure such as fineSTRUCTURE or HAC algorithms. Under this schema, the genetic distance matrix is used as a (nearly) sufficient statistic for all downstream analyses. The two examples that we use as illustration, the discovery of multiple sympatric populations [5], and the determination of the dominant multi-drug resistant lineage [6], both follow the schematic analysis plan shown in Figure 1. In the following sections we show that

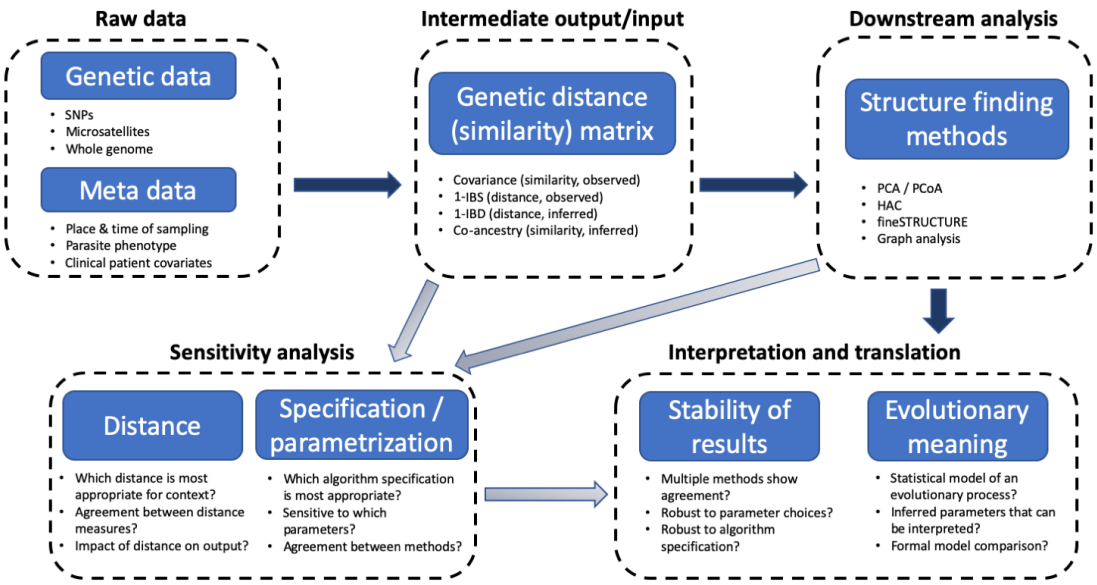

Figure 1: Typical computational analysis pipeline in a malaria genetic epidemiology study dependent on a genetic distance (similarity) matrix. The main analysis pipeline is shown with the dark arrows. Additional sensitivity analyses feed into the final interpretation and translation of results and are shown by the light arrows.

when using similar analytic procedures as those that underpinned these important results, outputs can be sensitive to both the choice of distance and to the algorithm specification in downstream analyses.

Genetic distance between *P. falciparum* 'monoclonal' isolates from South-east Asia

We computed three distance matrices for the 393 isolates, using as genetic distance $1 - \text{IBS}$, $1 - \text{IBD}$, and $-\log_2 \text{IBD}$, respectively. Figure 2 shows the distribution of the distances between pairs of isolates for the three choices of distance measure, and the agreement between the 1-IBS and 1-IBD distances. For closely related pairs (IBD close to 1), $1 - \text{IBS}$ and $1 - \text{IBD}$ strongly agree. However, when comparing unrelated pairs of parasites (e.g. $1 - \text{IBD}$ greater than 0.95), there is considerable variation in the $1 - \text{IBS}$ distance (variation across approximately 20% of its observed range).

Sensitivity of PCoA to the genetic distance

PCoA applied to the three distance matrices gave qualitatively similar results for $1 - \text{IBS}$ and $1 - \text{IBD}$, but differed considerably for $-\log_2 \text{IBD}$ (Figure 3). For both $1 - \text{IBS}$ and $1 - \text{IBD}$, the first principal component distinguishes clearly between *Pfkelch13* wild type parasites (green) and *Pfkelch13* mutated parasites (not green and predominantly the artemisinin resistant C580Y mutation shown in blue). This first principal component also distinguishes between *Pfplasmepsin* wild type (circles) and *Pfplasmepsin* amplified (triangles) parasites (associated with piperaquine resistance). This pattern also emerges from the the first principal component of the PCoA of the $-\log_2 \text{IBD}$ distance matrix, while the second shows the same level of variation between the wild type parasites (green circles) as amongst multi-drug resistant parasites with the C580Y mutation (blue triangles). It may be tempting to give an evolutionary interpretation to the PCoA pattern shown in panel C. For example, this PCoA pattern could demonstrate how what was initially a soft sweep became a hard sweep over time (leftmost contains almost exclusively WT parasites, centre contains the many resistant alleles seen initially in the Eastern GMS, rightmost contains almost exclusively C580Y isolates with amplification in *Pfplasmepsin*).This was the interpretation applied earlier to the Pfkelch13 flanking sequence haplotype data [12] However, there is no available

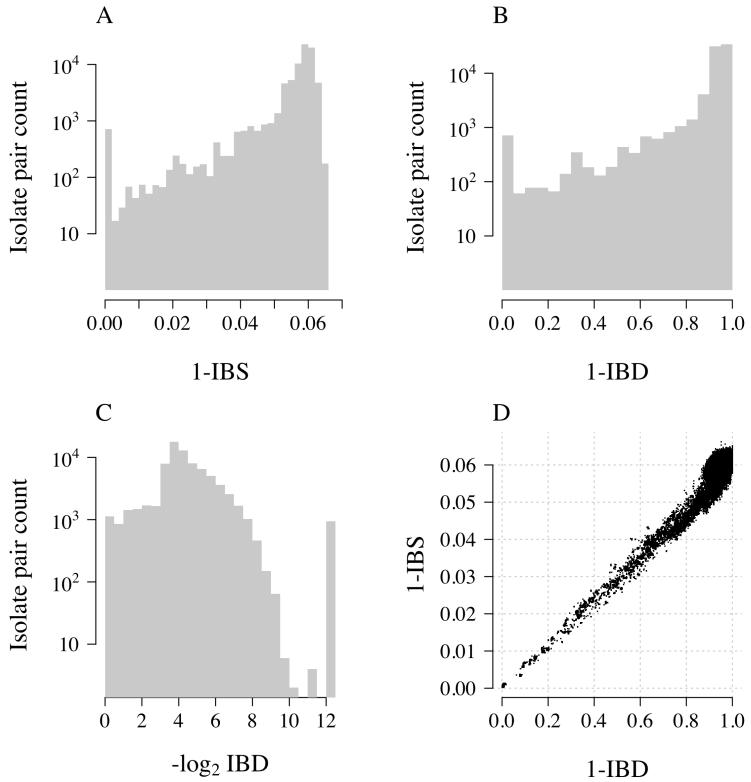

Figure 2: Comparison of three genetic distances. Each panel represents distances between 77028 pairs of 393 *P. falciparum* isolates from the Eastern GMS. A: $1-\text{IBS}$ distance; B: $1-\text{IBD}$ distance; C: $-\log_2 \text{IBD}$ distance; D: agreement between $1-\text{IBS}$ and $1-\text{IBD}$. Note that the y-axes in panels A-C are on a $\log_{10}$ scale. The set of distances such that $-\log_2 \text{IBD}$ is approximately 12.5 are those with estimated IBD equal to zero, replaced by a lower limit of quantification equal to the smallest IBD value greater than zero. Note that $1-\text{IBD}$ spans the zero to one range, whereas $1-\text{IBS}$ does not.

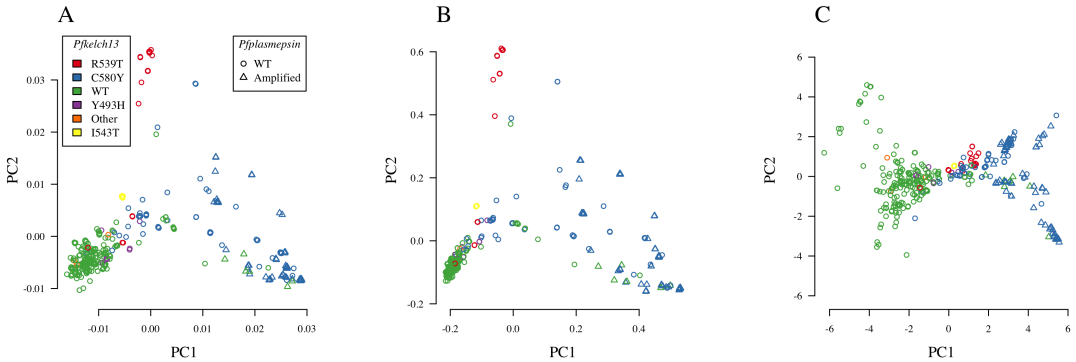

Figure 3: PCoA applied to the $1 - \text{IBS}$ (A), $1 - \text{IBD}$ (B), and $-\log_2 \text{IBD}$ (C) n-by-n distance matrices for n =393 isolates. Isolates are plotted along the first two principal components. Colours correspond to the different known causative mutations in the *Pfkelch13* gene, where green is wild type (WT) and blue is C580Y. Triangles correspond to *Pfplasmepsin* amplified parasites, and circles correspond to parasites that are WT in *Pfplasmepsin*.

mathematical theory supporting possible interpretations of the patterns observed in panels A-C of Figure 3. As an aside, there would be for 1-IBS between isolates from disconnected populations (e.g. different continents) or 1-IBS based on mitochondrial DNA, since mitochondrial DNA does not recombine and thus the coalescent model applies [8].

PCoA applied to arbitrary genetic distance matrices is a powerful visual tool, but the validity of the interpretation of the output is sensitive to the choice of genetic distance.

Sensitivity of HAC-based discrete clustering

Dendrograms constructed using HAC algorithms do not, in general, approximate the unknown phylogeny of the isolates. Therefore the topology of the dendrogram is not of primary interest. In addition, the clustering arrangement which underlies all possible topologies is, in general, highly dependent on the genetic distance and on the HAC algorithm specification. Comparing dendrogram topologies constructed using different genetic distances or algorithm specifications for a large set of parasite isolates is difficult. For example, a common approach is to use tanglegrams (also known as co-phylo plots) which show a side-by-side comparison of leaf placements with joining lines for two dendrograms [27]. The implicit assumption is that the level of entanglement of the lines is proportional to the level of incongruence between the dendrograms. However, the correlation between entanglement and dendrogram incongruence is weak and so tanglegrams can be misleading [28]. A key reason for this is that the ordering of the leaves within the dendrogram is arbitrary. Many leaf orderings correspond to the same clustering arrangement, see for example Figure 4. The visual topology of the tree is therefore unstable. This is especially important when using meta-data to check coherency of clustering structure found by HAC. For example, when colouring the leaves or adding a colour bar at the bottom of the tree, see Figure 4. For these reasons, we focus on the sensitivity of discrete cluster assignments based on HAC.

A dendrogram (which is a visualisation of a set of nested clustering arrangements) can be collapsed into discrete clusters by choosing a cut-point on the y-axis. For example, the KEL1 lineage reported in [6] was determined on the basis of cutting an HAC dendrogram into a set of discrete clusters. The largest cluster (named KEL1) was shown to be dispersed across Cambodia and Vietnam. This process of 'cutting' or 'flattening' a dendrogram into a set of discrete clusters necessitates specifying either the level at which to cut the dendrogram (in terms of the distribution of distances) or the number of clusters desired. Both of these choices are subjective. In this section, we show how the resulting cluster assignments are sensitive to both the genetic distance and to the algorithm specification (linkage function).

To illustrate the sensitivity of the discrete cluster assignment to the genetic distance, we applied HAC with average linkage (the algorithmic choice used in [6]) to the 1-IBD, 1-IBS and $-\log_2$ IBD distance matrices. Figure 5 tracks how cluster assignment for each isolate varies according to the genetic distance used. The mix of colours present in the stacked barplots of panels B & C of Figure 5 illustrates the sensitivity of cluster assignment to the genetic distance. This applies to constructing

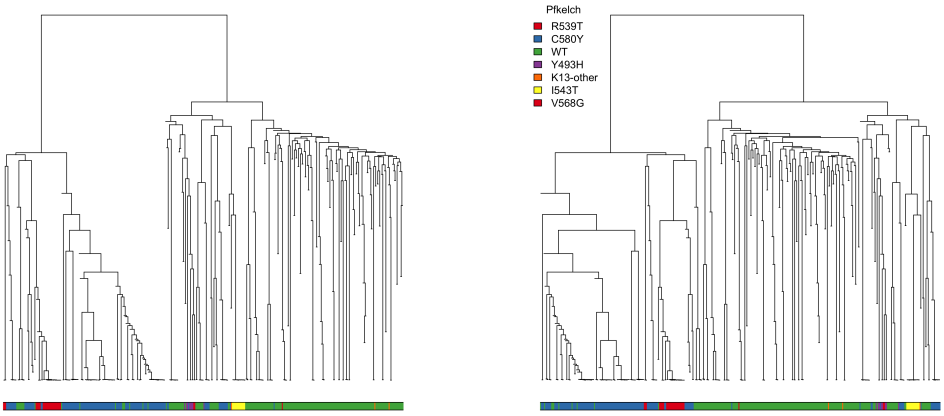

Figure 4: Two distinct dendrograms depict the same underlying clustering arrangement. The ordering of the leaves was changed by randomly rotating the internal nodes. The clustering arrangement was produced by applying HAC with average linkage to the -$\log_2$ IBD distance matrix. The coloured bars below the dendrograms visualise the corresponding *Pfkelch13* mutation of each isolate (green: wild type; blue: C580Y). This illustrates how the ordering of the meta-data is sensitive to arbitrary choices for the dendrogram topology.

discrete clusters using HAC with average linkage. The cluster assignment differs even when the two distances compared are exactly proportional to each other (1-IBD and $-\log_2$ IBD, panels A and C of Figure 5, respectively). We note that cluster assignment using HAC with single linkage is invariant to monotonic transformations of the distance (i.e. HAC with single linkage of a matrix of 1-IBD distances would give an identical clustering arrangement as HAC with single linkage of the matrix of $-\log_2$ IBD distances). However, the single linkage specification is not often used as it produces highly unbalanced tree structures (left or right branching, resulting in unbalanced clusters as shown in panel C of Figure 6). It is important to note that none of the linkage functions, when applied to arbitrary genetic distance matrices have a theoretical grounding. The resulting dendrograms have no theoretical guarantees to approximate the unknown phylogenies. They only provide a graphical representation of the genetic distance matrix.

To illustrate the sensitivity to the linkage function when constructing discrete clusters using HAC algorithms, we computed HAC-based dendrograms using average, complete, and single linkage, and Ward's criterion, applied to the $-\log_2$ IBD distance matrix, which is similar but not identical to the distance matrix used in [6]. We cut each of these dendrograms into nine discrete clusters. Figure 6 tracks how the cluster assignment varies according to the linkage function specified. This Figure shows how the clusters discovered under the four algorithmic choices have very few similarities. This highlights the strong sensitivity to the algorithm specification. If there were perfect agreement between cluster assignments, the stacks should contain only one colour each.

Visualising distance matrices using heatmaps ordered with HAC algorithms

HAC algorithms applied to genetic distance matrices are useful for certain tasks. Structure in the genetic distance matrix can be visualised directly via a heatmap. However, the clarity of a heatmap visualisation of the distance matrix is dependent on the ordering of the isolates in the matrix. Finding optimal orderings (known as seriation) whereby all the small distances are concentrated around the diagonal provides improved visualisation. Finding the optimal ordering (e.g. a perfect anti-Robinson matrix [29]) is an NP hard problem and so heuristic seriation approaches such as HAC are typically used [30]. HAC algorithms are fast and empirically determine a good ordering of isolates. As an illustration, Figure 7 shows heatmaps for the 1-IBD, 1-IBS, and $-\log_2$ IBD distance matrices, whereby the ordering was determined by HAC with average linkage.

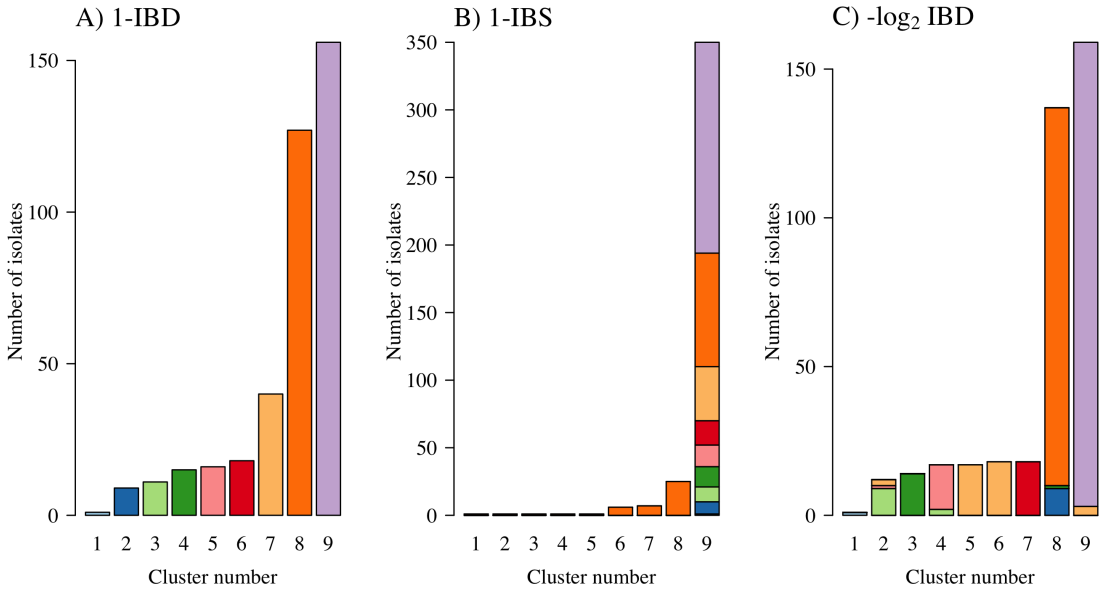

Figure 5: Tracking the discrete cluster assignment derived from HAC (specification is average linkage) according to the genetic distance matrix used. Each *P. falciparum* isolate is assigned a colour based on their cluster assignment according to a flattened dendrogram (a dendrogram cut at a given y-axis point) of a clustering arrangement generated by HAC of the 1-IBD distance matrix. In this case the y-axis cut-point was chosen to produce nine distinct clusters (panel A). These colours are then used to track cluster membership when the same HAC algorithm is applied to the 1-IBS and $-\log_2$ IBD distance matrices (panels B and C, respectively).

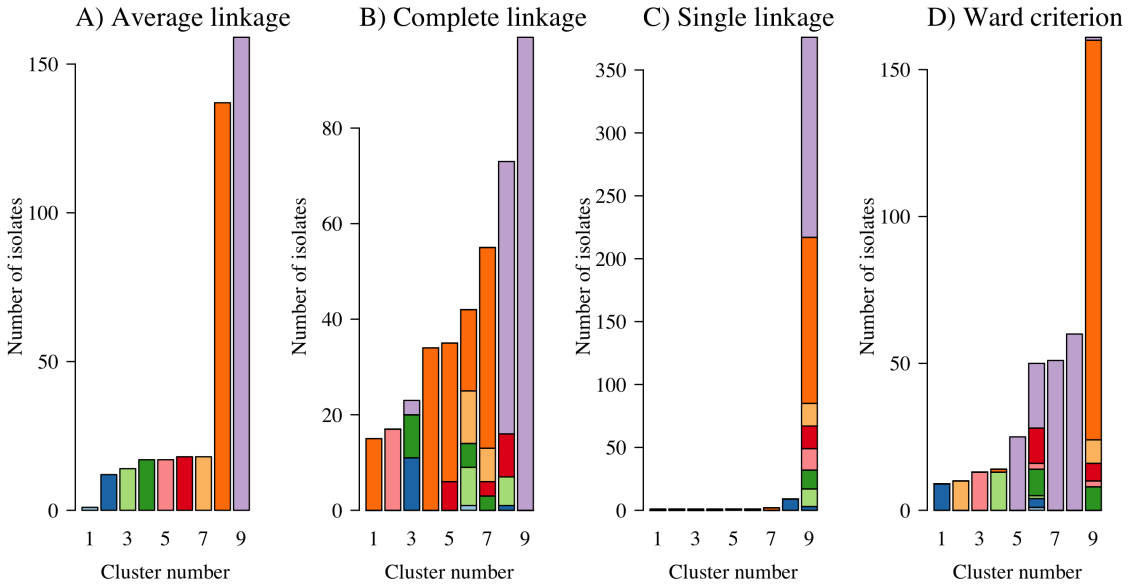

Figure 6: Tracking the discrete cluster assignment derived from HAC according to the linkage function used applied to the $-\log_2$ IBD distance matrix. Each *P. falciparum* isolate is assigned a colour based on their cluster assignment from the average linkage algorithm whereby the y-axis cut-point was chosen to produce nine distinct clusters (panel A). These colours are then used to produce stacked barplots for cluster membership derived from three other algorithm specifications (complete, single and Ward's criterion, panels B-D, respectively).

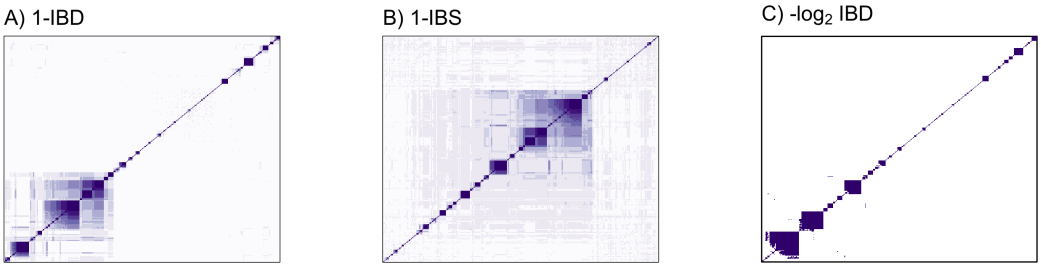

Figure 7: Heatmaps of genetic distance matrices, whereby isolates have been ordered with the output of the HAC algorithm (average linkage specified here). The colour shading was chosen by applying nine shades of purple to a uniform grid over the range of observed values (see panels A-C of Figure 2) in the distance matrix. The visual effect of the clustering in the heatmap is sensitive to this specification, for example, a grid over observed quantiles would produce a different visual effect.

Reporting recommendations

Panel 1 summarises our proposed recommendations to improve the transparency and reproducibility of genetic epidemiology studies in malaria. A key recommendation is that HAC algorithms should not be used to construct discrete clusters as they are sensitive to both the genetic distance and to the algorithm specification. Analyses that rely on genetic distance matrices can be made fully reproducible by making the distance matrices available along with the analysis code. In particular we recommend the use of statistical notebooks, such as RMarkdown or Jupyter, which allow for entire analyses to be re-run with little burden to the analyst. Visualisation of the analytic output of clustering methods applied to distance matrices is best done via heatmaps. Dendrograms should not be used as the main analytic output unless they are labelled properly, as their structure is unstable.

We also propose that genetic epidemiology studies that use ML algorithms justify explicitly their algorithmic choice(s). ML algorithms such as PCoA and HAC are non inferential. The interpretation of their output is restricted. They can be used only to support answers to a given clinical or biological question and they cannot compute the evidence (answer) for a particular model. There is therefore a disconnect between the restricted scientific questions that these ML algorithms can answer and the difficult scientific questions that genetic studies usually pose. It should be made very clear when speculative reasoning is appended to quantitative results from these algorithms to bridge this disconnect. This is analogous to the use of causal reasoning in observational studies: most analyses of observational data have as their ultimate goal providing an answer to a causal question but they are inherently limited to characterising correlation not causation. For this reason, medical journals usually impose associative language in the results (what the study can answer) and restrict causal reasoning to the discussion of the results (speculation: what the study desires to answer). We propose that genetic epidemiological studies of malaria that feature use of ML algorithms should be subject to analogous restrictions.

Discussion

The last decade has seen a huge increase in available malaria parasite genetic data, offering profound insights into the genetic epidemiology of both *P. falciparum* and *P. vivax* malaria. However, the analytical tools with which to analyse and interpret these data are still in their infancy. Summarizing dense genetic marker data such as WGS data into a genetic distance matrix is a key step in many analytical pipelines. The utility of this dimensionality reduction step has been demonstrated clearly in human genetics. The method fineSTRUCTURE (operating on the co-ancestry matrix) has provided key insights into recent human evolution, e.g. [31]. In malaria parasite population genetics the choice of genetic distance has been given less consideration than in human population genetics. Recent work has demonstrated the utility of assessing genetic relatedness based on IBD

(and its correlate IBS) and using this to measure genetic distance between proximal parasites populations [32]. Output of analyses using an IBD-based genetic distance are interpretable insofar as IBD-based genetic distance explicitly references recent evolutionary events driven by recombination. As for which methods should be applied to parasite genetic distance matrices, our work shows that PCoA is a useful exploratory tool. We suggest researchers apply PCoA to multiple distance matrices (e.g. 1-IBS, 1-IBD and $-\log_2$ IBD) to characterise sensitivity to the genetic distance. This study shows that HAC, although useful for ordering samples and visualising distance matrices, should not be used to construct discrete clusters of samples. PCoA and HAC provide a basis from which to explore the key biological questions of interest in the malaria parasite populations under study (e.g. "what is the most likely sequence of events that led to the spread of a single multi-drug resistant *P. falciparum* lineage across Southeast Asia?"), but they do not provide the complete answer.

The emergence and spread of a single multidrug resistant *P. falciparum* lineage across four countries in the eastern Greater Mekong sub-region (GMS) [12, 13] is of serious public health concern. It leaves few options for effective antimalarial therapy in the region. Understanding how drug resistance originates and evolves assists in developing strategies to prevent or curb its spread. The GMS is an area of particular interest for global malaria control and elimination because resistance of *P. falciparum* to all the major antimalarial drugs (chloroquine, sulfadoxine, pyrimethamine, mefloquine, atovaquone, piperaquine and the artemisinins) has originated there and, in the case of chloroquine and sulfadoxine-pyrimethamine, spread across India and Africa with devastating consequences. We believe there are opportunities to improve and extend the use of the available genetic data to answer important questions such as "did artemisinin resistance promote the spread of piperaquine resistance (and if so, by how much)?". Tailor-made statistical methods need to be developed that best utilise the available data to answer these important questions.

ML algorithms are powerful tools that can extract structure from vast quantities of complex data and thus generate hypotheses regarding ancestral events. On their own, these methods can neither prove nor disprove an evolutionary model. It is important to note that model-based clustering developed specifically for population genetics - e.g. STRUCTURE - are also easily mis-used. As for PCA, multiple evolutionary models can give rise to identical patterns in the admixture bar plots which are the main analytic output of these models. The output patterns are also affected by uneven sampling [33]. These bar plots are often mis-interpreted and the strong underlying model assumptions ignored [34]. For the majority of applications, there is no true K (the number of clusters inferred by STRUCTURE). It is highly plausible that *Plasmodium spp.* populations strongly diverge from the restrictive model assumptions of the program STRUCTURE. This calls into question its applicability to malaria phylogenetics beyond exploratory data analysis.

A large class of ML algorithms take as input a distance matrix (genetic in our context) and output a summary of the data structure. These algorithms are powerful as they are fast and versatile. However, they have important analytical limitations which may have been under appreciated. In this work we show that two popular methods, PCoA and HAC, are sensitive to the definition of genetic distance and to other algorithmic specification. It is also important to understand the interpretative limitations of the ML algorithm outputs. By interpretative limitations, we mean that the interpretation of the algorithmic output is restricted. This has sometimes led to a disconnect between the questions posed by a study and the questions addressed by the methods used in the study. It is important that genetic epidemiology studies convey clearly the interpretative limitations of methods applied to the data. Continuing the analogy with the the use of causal reasoning when the underlying data are observational. Speculative causality based on associations is an essential scientific exercise, but medical journals usually restrict this to the discussion of results. At present, in genetic epidemiology studies, speculation often bridges the disconnect between valid interpretation of the analytic output and desired answers to the clinical or epidemiological questions of interest. When this speculation does not make the disconnect clear, it has been described colourfully as 'cargo-cult' science [35, 36], or 'trans-science' [37]; in other words, the use of quantitative methods to give a scientific veneer to speculative results without highlighting the analytical and inferential limitations. Ideally, inferential interpretation would replace speculative interpretation. Statistical models articulated around specific hypotheses allow inferential interpretation but, in many cases, fitting these models is not feasible due to computational limitations (e.g. intractable likelihoods). By using ML algorithms first (to convert a possibly obscure signal in the data into a summary from which a manageable set of hypotheses can be generated), and statistical models second (to do statistical inference under a manageable set of hypotheses), we can offset

the advantages and disadvantages of both ML algorithms and statistical models (fast and versatile but non-inferential versus slow and constrained but inferential) and thus address appropriately the questions of interest. The development of statistical models to characterise malaria parasite population ancestry is an open area of research with many exciting methodological challenges (resulting from the complexity of the malaria parasite life cycle). Importantly, if the signal in the data is strong (e.g. as it is in the malaria studies that we use here as running examples), our calls for more sensitivity checks (to address algorithmic sensitivity) and more statistical models (to bridge the inferential disconnect) are not likely to change the overall narrative, but they will make the narrative more robust and, critically, falsifiable (i.e. scientific).

Panel 1: Overview of the proposed guidelines for the use of ML algorithms in population genetic epidemiology studies of *Plasmodium*

Introduction

- *State the overarching scientific question(s) of interest*

Methods

- *Justification of ML algorithms* Clearly state why the ML algorithms used were chosen (indeed, the use of any method should be justified) and state precisely the scientific questions they can answer. If the questions that can be answered differ to the overarching question above, we suggest explaining why an inferential method was not used (e.g. because of computational complexity).
- *Hierarchical agglomerative clustering algorithms* These should not be used to construct discrete population clustering assignments.
- *Sensitivity analysis* If clustering analyses have been performed, and the presence of biologically meaningful sub-populations are reported, there needs to be a robust analysis examining sensitivity to the genetic distance definition and sensitivity to the underlying methods (both the method used and specification of the method).
- *Publicly available code.* Computational analysis pipelines processing high-dimensional genetic data are usually complex with multiple filtering and data processing steps. The best way to ensure that the analysis is fully reproducible and transparent is to provide readable code. For readability we recommend the use of computational notebooks (e.g. RMarkdown, Jupyter). Although it may not be possible to provide all the raw data used to generate the results, a minimum working example (for example using simulated data or a small subset of isolates) is essential.
- *Publicly available intermediate output/input and meta-data* Many computational analysis pipelines follow the schematic steps shown in Figure 1. All downstream and sensitivity analyses are thus fully reproducible if the intermediate output/input are provided. Examples of intermediate data summaries include matrices of genetic distances (1-IBS, 1-IBD, -$\log_2$IBD); IBD segments for all pairs of isolates as output by e.g. *hmmIBD* [19], *isoRelate* [38], *DEploidIBD* [39]; co-ancestry matrix as output by chromosome painting [3, 10]. These intermediate data summaries are privacy-preserving: it is not possible to reverse engineer them and retrieve the genetic data.

Results

- *Dendrograms* HAC dendrograms should not be shown as a main result. We recommend using them as a data visualisation and exploration tool only. It should be made clear (e.g. in Figure captions) that dendrograms do not approximate the underlying phylogeny (not family trees).
- *Heatmaps.* Genetic distance (similarity) matrices are best represented using heatmaps whereby isolates are ordered using seriation algorithms (this includes HAC algorithms which empirically work well for this purpose).
- *Clustering agreement* Agreement between clustering algorithms should be represented graphically. We recommend against using tanglegrams for comparing clustering arrangements. Colour coded stacked barplots can help to understand agreement and dis-agreement between discrete cluster membership (e.g. Figures 5 & 6).

Discussion

- *Speculative reasoning* If the overarching scientific question(s) of interest do not align with the scientific questions answered by the applied methods, the speculative nature of any reasoning should be made clear. This is analogous to causal reasoning based on observational data.

Acknowledgements

This study was part of the Mahidol-University Oxford Tropical Medicine Research Programme funded by the Wellcome Trust of the United Kingdom (core grant 106698/B/14/Z). The clinical studies were supported by the United Kingdom Department for International Development (programme code 201900). Genotyping for this study was funded by the Wellcome Trust (098051, 206194), Bill & Melinda Gates Foundation (OPP11188166) and the Medical Research Council (G0600718). Genome sequencing was done by the Wellcome Sanger Institute (WSI). A.R.T. and C.O.B. are supported by a NIGMS Maximizing Investigators' Research Award (MIRA) for Early Stage Investigators (R35GM-124715-02). R35GM124715-02

References

[1] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904, 2006.

[2] Jonathan K Pritchard and Marcus W Feldman. Statistics for microsatellite variation based on coalescence. *Theoretical Population Biology*, 50(3):325–344, 1996.

[3] Daniel John Lawson, Garrett Hellenthal, Simon Myers, and Daniel Falush. Inference of Population Structure using Dense Haplotype Data. *PLoS Genetics*, 8(1):e1002453, 2012.

[4] David H Alexander, John Novembre, and Kenneth Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, 19(9):1655–1664, 2009.

[5] Olivo Miotto, Jacob Almagro-Garcia, Magnus Manske, Bronwyn MacInnis, Susana Campino, Kirk A Rockett, Chanaki Amaratunga, Pharath Lim, Seila Suon, Sokunthea Sreng, et al. Multiple populations of artemisinin-resistant plasmodium falciparum in cambodia. *Nature Genetics*, 45(6):648, 2013.

[6] Roberto Amato, Richard D Pearson, Jacob Almagro-Garcia, Chanaki Amaratunga, Pharath Lim, Seila Suon, Sokunthea Sreng, Eleanor Drury, Jim Stalker, Olivo Miotto, et al. Origins of the current outbreak of multidrug-resistant malaria in southeast asia: a retrospective genetic study. *The Lancet Infectious Diseases*, 18(3):337–345, 2018.

[7] William L Hamilton, Roberto Amato, Rob W van der Pluijm, Christopher G Jacob, Huynh Hong Quang, Nguyen Thanh Thuy-Nhien, Tran Tinh Hien, Bouasy Hongvanthong, Keobouphaphone Chindavongsa, Mayfong Mayxay, Rekol Huy, Rithea Leang, Cheah Huch, Lek Dysoley, Chanaki Amaratunga, Seila Suon, Rick M Fairhurst, Rupam Tripura, Thomas J Peto, Yok Sovann, Podjanee Jittamala, Borimas Hanboonkunupakarn, Sasithon Pukrittayakamee, Nguyen Hoang Chau, Mallika Imwong, Mehul Dhorda, Ranitha Vongpromek, Xin Hui S Chan, Richard J Maude, Richard D Pearson, T Nguyen, Kirk Rockett, Eleanor Drury, Sónia Gonçalves, Nicholas J White, Nicholas P Day, Dominic P Kwiatkowski, Arjen M Dondorp, and Olivo Miotto. Evolution and expansion of multidrug-resistant malaria in southeast Asia: a genomic epidemiology study. *The Lancet Infectious diseases*, 0(0), jul 2019.

[8] Gil McVean. A genealogical interpretation of principal components analysis. *PLoS Genetics*, 5(10):e1000686, 2009.

[9] Aimee R Taylor, Pierre E Jacob, Daniel E Neafsey, and Caroline O Buckee. Estimating relatedness between malaria parasites. *Genetics*, pages genetics–302120, 2019.

[10] Na Li and Matthew Stephens. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4):2213–2233, 2003.

[11] Olivo Miotto, Roberto Amato, Elizabeth A Ashley, Bronwyn MacInnis, Jacob Almagro-Garcia, Chanaki Amaratunga, Pharath Lim, Daniel Mead, Samuel O Oyola, Mehul Dhorda, et al. Genetic architecture of artemisinin-resistant plasmodium falciparum. *Nature Genetics*, 47(3):226, 2015.

[12] Mallika Imwong, Kanokon Suwannasin, Chanon Kunasol, Kreepol Sutawong, Mayfong Mayxay, Huy Rekol, Frank M Smithuis, Tin Maung Hlaing, Kyaw M Tun, Rob W van der Pluijm, et al. The spread of artemisinin-resistant plasmodium falciparum in the greater mekong subregion: a molecular epidemiology observational study. *The Lancet Infectious Diseases*, 17(5):491–497, 2017.

[13] Mallika Imwong, Tran T Hien, Nguyen T Thuy-Nhien, Arjen M Dondorp, and Nicholas J White. Spread of a single multidrug resistant malaria parasite lineage (pfpailin) to vietnam. *The Lancet Infectious Diseases*, 17(10):1022–1023, 2017.

[14] Elizabeth A Ashley, Mehul Dhorda, Rick M Fairhurst, Chanaki Amaratunga, Parath Lim, Seila Suon, Sokunthea Sreng, Jennifer M Anderson, Sivanna Mao, Baramey Sam, et al. Spread of artemisinin resistance in plasmodium falciparum malaria. *New England Journal of Medicine*, 371(5):411–423, 2014.

[15] Rob W van der Pluijm, Mallika Imwong, Nguyen Hoang Chau, Nhu Thi Hoa, Nguyen Thanh Thuy-Nhien, Ngo Viet Thanh, Podjanee Jittamala, Borimas Hanboonkunupakarn, Kitipumi Chutasmit, Chalermpon Saelow, et al. Determinants of dihydroartemisinin-piperaquine treatment failure in Plasmodium falciparum malaria in Cambodia, Thailand, and Vietnam: a prospective clinical, pharmacological, and genetic study. *The Lancet Infectious Diseases*, 19(9):952–961, 2019.

[16] World Health Organization. *Guidelines for the treatment of malaria.* World Health Organization, 2015.

[17] Magnus Manske, Olivo Miotto, Susana Campino, Sarah Auburn, Jacob Almagro-Garcia, Gareth Maslen, Jack O'Brien, Abdoulaye Djimde, Ogobara Doumbo, Issaka Zongo, et al. Analysis of plasmodium falciparum diversity in natural infections by deep sequencing. *Nature*, 487(7407):375–379, 2012.

[18] MalariaGEN Plasmodium falciparum Community Project et al. Genomic epidemiology of artemisinin resistant malaria. *eLife*, 5:e08714, 2016.

[19] Stephen F Schaffner, Aimee R Taylor, Wesley Wong, Dyann F Wirth, and Daniel E Neafsey. hmmibd: software to infer pairwise identity by descent between haploid genotypes. *Malaria Journal*, 17(1):196, 2018.

[20] Nick Patterson, Alkes L Price, and David Reich. Population structure and eigenanalysis. *PLoS Genetics*, 2(12):e190, 2006.

[21] Lan Huong Nguyen and Susan Holmes. Ten quick tips for effective dimensionality reduction. *PLOS Computational Biology*, 15(6):e1006907, 2019.

[22] Alfred Amambua-Ngwa, Lucas Amenga-Etego, Edwin Kamau, Roberto Amato, Anita Ghansah, Lemu Golassa, Milijaona Randrianarivelojosia, Deus Ishengoma, Tobias Apinjoh, Oumou Maïga-Ascofaré, Ben Andagalu, William Yavo, Marielle Bouyou-Akotet, Oyebola Kolapo, Karim Mane, Archibald Worwui, David Jeffries, Vikki Simpson, Umberto D'Alessandro, Dominic Kwiatkowski, and Abdoulaye A. Djimde. Major subpopulations of plasmodium falciparum in sub-saharan africa. *Science*, 365(6455):813–816, 2019.

[23] Naruya Saitou and Masatoshi Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, 1987.

[24] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2019.

[25] Daniel Müllner. fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python. *Journal of Statistical Software*, 53(9):1–18, 2013.

[26] Tal Galili. dendextend: an r package for visualizing, adjusting, and comparing trees of hierarchical clustering. *Bioinformatics*, 2015.

[27] Celine Scornavacca, Franziska Zickmann, and Daniel H. Huson. Tanglegrams for rooted phylogenetic trees and networks. *Bioinformatics*, 27(13):i248–i256, 06 2011.

[28] Damien M De Vienne. Tanglegrams are misleading for visual evaluation of tree congruence. *Molecular Biology and Evolution*, 36(1):174–176, 2019.

[29] W. S. Robinson. A method for chronologically ordering archaeological deposits. *American Antiquity*, 16(4):293–301, 1951.

[30] Michael Hahsler, Kurt Hornik, and Christian Buchta. Getting things in order: an introduction to the r package seriation. *Journal of Statistical Software*, 25(3):1–34, 2008.

[31] Stephen Leslie, Bruce Winney, Garrett Hellenthal, Dan Davison, Abdelhamid Boumertit, Tammy Day, Katarzyna Hutnik, Ellen C Royrvik, Barry Cunliffe, Daniel J Lawson, et al. The fine-scale genetic structure of the british population. *Nature*, 519(7543):309–314, 2015.

[32] Aimee R Taylor, Stephen F Schaffner, Gustavo C Cerqueira, Standwell C Nkhoma, Timothy JC Anderson, Kanlaya Sriprawat, Aung Pyae Phyo, François Nosten, Daniel E Neafsey, and Caroline O Buckee. Quantifying connectivity between local plasmodium falciparum malaria parasite populations using identity by descent. *PLoS Genetics*, 13(10):e1007065, 2017.

[33] EC Anderson and KK Dunham. The influence of family groups on inferences made with the program structure. *Molecular Ecology Resources*, 8(6):1219–1229, 2008.

[34] Daniel J Lawson, Lucy Van Dorp, and Daniel Falush. A tutorial on how not to over-interpret structure and admixture bar plots. *Nature Communications*, 9(1):3258, 2018.

[35] Richard Phillips Feynman and Ralph Leighton. *"Surely you're joking, Mr. Feynman!": adventures of a curious character*. Random House, 1992.

[36] Philip B Stark and Andrea Saltelli. Cargo-cult statistics and scientific crisis. *Significance*, 15(4):40–43, 2018.

[37] Andrea Saltelli. A short comment on statistical versus mathematical modelling. *Nature Communications*, 10(1):1–3, 2019.

[38] Lyndal Henden, Stuart Lee, Ivo Mueller, Alyssa Barry, and Melanie Bahlo. Identity-by-descent analyses for measuring population dynamics and selection in recombining pathogens. *PLoS genetics*, 14(5):e1007279, 2018.

[39] Sha Joe Zhu, Jason A Hendry, Jacob Almagro-Garcia, Richard D Pearson, Roberto Amato, Alistair Miles, Daniel J Weiss, Tim CD Lucas, Michele Nguyen, Peter W Gething, et al. The origins and relatedness structure of mixed infections vary with local prevalence of *P. falciparum* malaria. *Elife*, 8:e40845, 2019.