# Population-level genome-wide STR typing in *Plasmodium* species reveals higher resolution population structure and genetic diversity relative to SNP typing

Jiru Han[1,2], Jacob E. Munro[1], Anthony Kocoski[1,3], Alyssa E. Barry[1,2,4,5], Melanie Bahlo[1,2]*

**1** Population Health and Immunity Division, The Walter and Eliza Hall Institute of Medical Research, Melbourne, VIC, Australia, **2** Department of Medical Biology, The University of Melbourne, VIC, Australia, **3** Department of Mathematics and Statistics, The University of Melbourne, VIC, Australia, **4** Disease Elimination Program, Burnet Institute, Melbourne, VIC, Australia, **5** Present Address: IMPACT Institute for Innovation in Mental and Physical Health and Clinical Translation, Deakin University, 75 Pigdons Road, Waurn Ponds, Geelong, VIC 3216, Australia.

* bahlo@wehi.edu.au

# Abstract

Short tandem repeats (STRs) are highly informative genetic markers that have been used extensively in population genetics analysis. They are an important source of genetic diversity and can also have functional impact. Despite the availability of bioinformatic methods that permit large-scale genome-wide genotyping of STRs from whole genome sequencing data, they have not previously been applied to sequencing data from large collections of malaria parasite field samples. Here, we have genotyped STRs using HipSTR in more than 3,000 *Plasmodium falciparum* and 174 *Plasmodium vivax* published whole-genome sequence data from samples collected across the globe. High levels of noise and variability in the resultant callset necessitated the development of a novel method for quality control of STR genotype calls. A set of high-quality STR loci (6,768 from *P. falciparum* and 3,496 from *P. vivax*) were used to study *Plasmodium* genetic diversity, population structures and genomic signatures of selection and these were compared to genome-

wide single nucleotide polymorphism (SNP) genotyping data. In addition, the genome-wide information about genetic variation and other characteristics of STRs in *P. falciparum* and *P. vivax* have been made available in an interactive web-based R Shiny application PlasmoSTR (https://github.com/bahlolab/PlasmoSTR).

# Author summary

Malaria is a severe disease caused by a genus of parasites called *Plasmodium* and is transmitted to humans through infected *Anopheles* mosquitoes. *P. falciparum* and *P. vivax* are the predominant species responsible for more than 95% of all human malaria infections which continue to pose a significant challenge to human health. Antimalarial drug resistance is a serious threat hindering the elimination of malaria. As such, it is important to understand the role of genomic variation in the development of antimalarial drug resistance. STRs are an important source of genomic variation that, from a population genetics perspective, have several advantages over SNPs, including being highly polymorphic, having a higher mutation rate, and having been widely used to study the population structure and genetic diversity. However, STRs are not routinely genotyped with bioinformatic tools across the whole genome with short read sequencing data because they are difficult to identify and genotype accurately, as they vary in size and may align poorly to the reference genome, therefore requiring rigorous quality control (QC). In this study, we genotype STRs using HipSTR[1] in more than 3,000 *P. falciparum* and 174 *P. vivax* whole-genome sequence samples collected world-wide. We develop a multivariable logistic regression model for the measurement and prediction of the quality of STRs. In addition, we use a set of genome-wide high-quality STRs to study parasite population genetics and compare them to genome-wide SNP

genotyping data, revealing both high consistency with SNP based signals, as well as identifying some signals unique to the STR marker data. These results demonstrate that the identification of highly informative STR markers from large numbers of population samples is a powerful approach to study the genetic diversity, population structures and genomic signatures of selection in *P. falciparum* and *P. vivax*. Furthermore, we built an interactive web-based R Shiny application PlasmoSTR (https://github.com/bahlolab/PlasmoSTR) that includes genome-wide information about genetic variation and other characteristics of the high quality STRs identified in *P. falciparum* and *P. vivax*, allowing researchers to explore and visualize the specific STRs.

# Introduction

Short tandem repeats (STRs), also known as microsatellites, are tandem nucleotide repeats (1-9 base pairs) that are both abundant throughout the genome and highly polymorphic. Unlike many other types of genetic markers, STRs have a high mutation rate that is highly variable across different loci. *P. falciparum* has the most AT-rich eukaryotic genome known, with 80.6% A + T content overall and approaching 90% in introns and intergenic regions[2]. As a consequence, many regions in *P. falciparum* genome are highly repetitive, and STRs are found in abundance in both coding and noncoding regions throughout the *P. falciparum* parasite genome[2, 3] leading to about 10.74% of the *P. falciparum* genome being composed of STRs[2, 4]. In contrast, the total A+T content in *P. vivax* is 57.7%[5, 6]. In organisms with AT content close to 50%, such as *Drosophila* or humans, STRs only account for 1–3% of the genome[4, 7, 8]. These repetitive sequences can arise, expand or contract rapidly. In many cases, the simple homopolymer repeats tend to evolve neutrally and may not have a function, representing non-functional 'junk DNA', however more

64 complex sequences seem to be under selective pressure indicating a functional role[3, 9, 10]. The
65 repetitive protein sequences of *Plasmodium* have been previously shown to alter protein activity,
66 protein folding efficiency, stability, or aggregation and play an important role in the formation of
67 key structural elements of protein function[10]. STRs in coding regions with a motif size that is a
68 multiple of three (e.g. trinucleotide or hexanucleotide repeats) will not result in a frame-shift
69 mutation when repeats are deleted or added, but can change protein sequences[11]. For example,
70 the *Pfnhe-1* protein contains a polymorphic amino acid motif DNNND (GATAACAATAATGAT)
71 and DDNHNDNHNND (GATGATAACCATAATGATAATCATAATAATGAT) which affects
72 the *P. falciparum* Na+/H+ exchanger capabilities, and influences quinine resistance by combining
73 *Pfcrt* and *Pfmdr1*[12, 13].

74 STRs have also been widely used to study the population structure and genetic diversity of *P.*
75 *falciparum* and *P. vivax* populations in many countries[14-17]. However, most studies used
76 relatively few (< 20) polymorphic STR markers. These STRs were typed using a variety of low-
77 throughput lab-based methods, most recently with capillary electrophoresis[18, 19]. Extending
78 these low-throughput methods to hundreds of STR markers or genome-wide is prohibitive in both
79 time and cost. A few previous studies of genome-wide STRs analyses used *Plasmodium* reference
80 genome or limited *in vitro Plasmodium* samples and mainly focused on examining compositions
81 and function of STRs[10, 11, 20] and mutation rates[4, 21]. The overall contributory effect of STR
82 variation in *Plasmodium* field samples has not been evaluated at a genome-wide level.

83 Recently developed bioinformatic methods that infer the length of STR alleles using short-read
84 sequencing data permit STR genotyping from large collections of samples. There are many tools
85 for genotyping STRs, such as GATK HaplotypeCaller[22], LobSTR[23], RepeatSeq[24],

HipSTR[1] and GangSTR[25]. We used HipSTR[1], which is a haplotype-based method specifically designed for STR analysis. While other STR tools were mainly developed for calling the STR length per individual sample, HipSTR considers the entire sequence across all samples in the dataset for each STR site, and has been shown to outperform other tools when considering genotyping error rate, even with low coverage[26].

Here, we report the first large-scale STR typing study in more than 3,000 *P. falciparum*[27] and 174 *P. vivax*[28, 29] short-read whole genome sequencing samples sourced from global malaria hot-spots. A central aim of this work was to develop a filtering strategy to discover a set of high-quality STR variants and build a publically available and easy to use resource available for researchers who are interested in looking at the role of specific STRs throughout the *Plasmodium* genome. We then aimed to compare the performance of genome-wide SNPs data and STRs data in the following aspects: delineate population structure, genetic diversity, and genetic differentiation metrics. We also explored the biological importance of STR variation in different populations, and identified STR loci that may be linked to antimalarial drug resistance.

# Results

## SNP genotyping

### MalariaGEN global *P. falciparum* dataset.

Variations at more than three million positions were discovered in the *P. falciparum* dataset in the first stage of variant analysis. These included 1,542,905 SNPs and 1,545,263 indels. After performing all the filtering procedures (see Methods for more details), a total of 213,757 biallelic

106 SNPs were retained. We removed 194 samples with higher than 10% missing genotypes or other
107 quality control issues leaving a total of 3,047 high-quality samples from the 26 countries (remove
108 Burkina Faso) of 8 populations. Sample size varied by population with South America (SAM) =
109 31, West Africa (WAF) = 959, Central Africa (CAF) = 100, East Africa (EAF) = 327, South Asia
110 (SAS) = 32, the western part of Southeast Asia (WSEA) = 690, the eastern part of Southeast Asia
111 (ESEA) = 827 and Oceania (OCE) = 81[27]. The downstream analysis of the *P. falciparum* dataset
112 in this paper is based on the filtered dataset of high-quality SNP genotypes in 3,047 samples.

113 **Global *P. vivax* dataset.**

114 Variations included 1,345,364 SNPs and 715,369 indels discovered in the *P. vivax* dataset in the
115 first stage of variant analysis. After performing all the filtering procedures, a total of 188,571
116 biallelic SNPs were retained. We removed samples with multiple infections as determined by the
117 within-host infection fixation index ($F_{ws}$) metric[30, 31], or higher than 10% missing genotypes or
118 other quality control issues, leaving 174 high-quality samples from 11 countries. Sample size
119 varied by country with Brazil = 2, Cambodia = 16, Colombia = 27, Indonesia = 2, Malaysia = 3,
120 Mexico = 17, Myanmar = 7, Peru = 30, Papua New Guinea (PNG) = 7, Thailand = 59, Vietnam =
121 4 samples respectively. The downstream analysis of the *P. vivax* dataset in this paper is based on
122 the filtered dataset of high-quality SNP genotypes in 174 samples.

123 **STR genotyping**

124 We identified 104,649 high quality STRs from the *P. falciparum* 3D7 reference genome
125 (accounting for 9.29% of the genome) and 40,224 STRs from the *P. vivax* PvP01 reference genome
126 (accounting for 3.16% of the genome) by using Tandem Repeats Finder (TRF)[32]. The number

of STRs in *P. falciparum* is almost three times that of *P. vivax*. STRs with a 1-6 bp repeat unit account for 97.32% of *P. falciparum* 3D7 reference genome and 95.27% of *P. vivax* PvP01 reference genome STRs. Of these, homopolymeric tracts account for 40.09% of the *P. falciparum*, 64.16% of the *P. vivax*. The dinucleotide repeats account for 24.66% of the *P. falciparum*, while only 6.38% of the *P. vivax*. The higher proportion of dinucleotide repeats in *P. falciparum* can be attributed to the overall high AT content of the *P. falciparum* genome, with 24% of dinucleotide repeats in *P. falciparum* having the 'AT' motif, and 3% of dinucleotide repeats in *P. vivax* having the 'AT' motif.

A total of 20,196 STRs remained for downstream analysis across 3,047 *P. falciparum* samples after the initial QC filtering steps (see Methods for more details). The majority of STRs were located in promoter region (53.02%), coding region (25%), followed by the intergenic region (12.91%), intron region (8.61%) and other regions. *P. falciparum* is an extremely AT-rich genome but with higher GC content in coding and promoter regions, probably leading to more confident calling and higher quality STRs in those regions. STRs with a 1–6 bp repeat unit accounted for 96.24% of all the STRs. Of these, 9,382 (46.45%) are homopolymeric tracts (mononucleotide STRs), 3,563 (17.64%) are dinucleotide repeats, and 3,767 (18.65%) are trinucleotide repeats. Almost all the STR motifs (99.99%) have a repeat unit containing 'A' or 'T'.

A total of 23,146 STRs were retained for downstream analysis across 174 *P. vivax* samples after performing all the filtering procedures. The number of STRs varies in different genomic regions: promoter region (48.88%), coding region (27.35%), intergenic region (16.12%), followed by the intron region (7.47%) and other regions. STRs with a 1–6 bp repeat unit accounted for 94.59% of

all the STRs. Of these, 14,386 (62.15%) are homopolymeric tracts, 1,324 (5.72%) are dinucleotide repeats, and 4,251 (18.37%) are trinucleotide repeats.

To measure HipSTR's quality of prediction, we inspected HipSTR's genotype calls for the markers against the gel electrophoresis (GE) calls (see Methods for more details) and found HipSTR is calling length polymorphisms accurately with respect to the GE calls (S1 Fig).

**Multivariable logistic regression modeling for measurement and prediction of the quality of STRs**

A set of metrics including QC metrics from HipSTR and other metrics that were deemed useful were derived and used for the prediction of the STR quality. These are summarised in S1 Table. A multivariable logistic regression analysis was then performed to identify potential predictors of STR quality (see Methods for more details). Because STRs with mononucleotide repeats are more abundant and have higher error rates, we built separate regression models for the mononucleotide (1 bp motif) STRs and the polynucleotide (2-9 bp motif) STRs. Examination of Spearman's correlation coefficients ($R^2$) suggested that selecting the first five SNP principal components (PCs) were sufficient to capture the signal STRs. For the *P. falciparum* dataset, features that were significantly associated with the STR quality and the results obtained for the estimated coefficients for both the mononucleotide STR and polynucleotide STR models are presented in Table 1. For both the mononucleotide STR and polynucleotide STR models, the STR quality tends to be more associated with the STR features that captured population specific aspects of the dataset. Compared to the polynucleotide STR model, 'Mean_Posterior' which is the mean posterior probability of the STR genotype across all samples derived from HipSTR, exhibited a large effect, but only in the mononucleotide STR model. The closer this quantity is to 1, the higher the

confidence of the called genotype. We also built the model for the *P. falciparum* dataset that did not use the reported population origins of each sample as some samples were clearly distinct to the majority of samples from a particular country. The model of the *P. falciparum* sample without the population origin label was found to produce very similar results (S2 Table). For all further analysis only the model with the given population label was used.

**Table 1**. Multivariable logistic regression model's estimated coefficients and respective 95% confidence intervals. The model was fitted on the *P. falciparum* dataset.

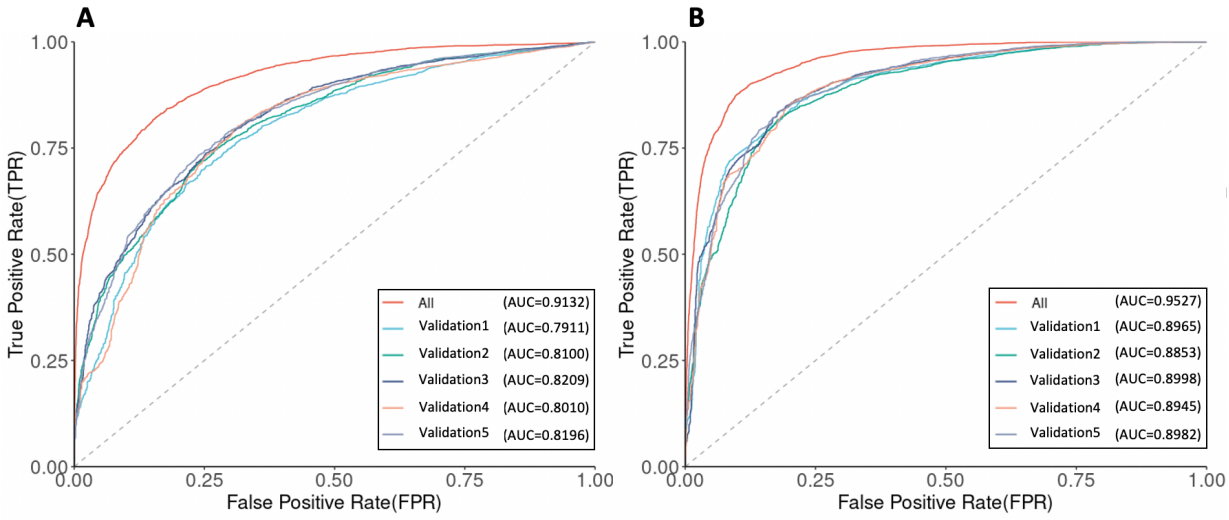
| | **Variable** | **Coefficient estimate** | **Lower 95% CI** | **Upper 95% CI** | ***P* values** |
|---|---|---|---|---|---|
| **Modeling Mononucleotide STR** | (Intercept) | 0.48 | 0.38 | 0.59 | <0.001 |
| | Repeat units | 0.12 | 0.05 | 0.19 | <0.001 |
| | GC_Diff | 0.09 | 0.02 | 0.16 | 0.01 |
| | Missingness | 0.06 | -0.01 | 0.12 | 0.1 |
| | Mean_Posterior | 0.81 | 0.70 | 0.92 | <0.001 |
| | Mean_stutter | -0.15 | -0.24 | -0.07 | <0.001 |
| | He | 7.00 | 6.38 | 7.64 | <0.001 |
| | MeanHe | -3.87 | -4.57 | -3.17 | <0.001 |
| | MinimumHe | -0.81 | -1.03 | -0.57 | <0.001 |
| | MaximumHe | 2.33 | 2.08 | 2.59 | <0.001 |
| **Modeling Polynucleotide STR** | (Intercept) | 1.55 | 1.40 | 1.71 | <0.001 |
| | Length | -0.11 | -0.18 | -0.04 | 0.003 |
| | Repeat units | 0.16 | 0.06 | 0.26 | 0.001 |

| | | | | | |
|---|---|---|---|---|---|
| | GC_Flank | 0.11 | 0.05 | 0.17 | <0.001 |
| | Mean_Posterior | 0.10 | 0.00 | 0.20 | 0.05 |
| | Mean_Stutter | -0.11 | -0.20 | 0.03 | 0.04 |
| | He | 10.20 | 9.02 | 11.41 | <0.001 |
| | MeanHe | -12.41 | -13.73 | -11.10 | <0.001 |
| | JostD | -1.10 | -1.42 | -0.75 | <0.001 |
| | MinimumHe | -0.48 | -0.84 | -0.12 | 0.009 |
| | MaximumHe | 7.02 | 6.52 | 7.52 | <0.001 |

We also compared the prediction performance of the *P. falciparum* complete dataset with five-fold cross-validation (see section Methods). The performance of each dataset is compared in terms of their receiver–operator–characteristic (ROC) curves and the area-under-the-curve (AUC). The AUC values of the *P. falciparum* complete dataset is 0.9132 in the mononucleotide STR model and 0.9527 in the polynucleotide STR model, and these five validation-datasets range from 0.7911 to 0.8209 in the mononucleotide STR model, and 0.8853 to 0.8998 in the polynucleotide STR model (Fig 1). The performance fluctuation depends on the size of the datasets. The predictive performance in the whole dataset was generally superior to that in the smaller training datasets (S2 and S3 Figs). This was observed in both the *P. falciparum* mononucleotide STR and polynucleotide STR models, suggesting that larger datasets may improve prediction power. In this work, we select the whole *P. falciparum* dataset with a higher AUC value to perform all subsequent analyses. The model of the *P. falciparum* sample without the population origin label also showed high performance for both the mononucleotide STR (AUC=0.9198) and polynucleotide STR (AUC=0.9589). The *P. falciparum* model was observed to be very stable, with regards to which

192 measures of quality were used in the model, and reproducible, with the five-fold cross validation
193 sets giving very similar results in both the mononucleotide and polynucleotide STR models.

194




195

196 **Fig 1.** ROC curves and AUC values of the *P. falciparum* complete dataset and five validation-datasets. (A) The mononucleotide
197 STR model. (B) The polynucleotide STR model.

198 For the *P. vivax* dataset, features that were most significantly associated with the STR quality are
199 summarised in S3 Table. The *P. vivax* model also showed high performance for both the
200 mononucleotide STR (AUC=0.9186) and polynucleotide STR (AUC=0.9548). For the *P. vivax*
201 mononucleotide STR and polynucleotide STR models, the STR quality is also more associated
202 with the STR features that capture population specific aspects of the dataset, showing similar
203 results as *P. falciparum*. However one of the STR variables, MinimumHe, displayed an opposite
204 relationship compared to the *P. falciparum* STR models. For *P. vivax*, four countries have few
205 samples (< 5), which may affect the robustness of the MinimumHe estimate and thus may have
206 led to this difference. To investigate this, we also built the model of the *P. vivax* sample without
207 the population origin labels, wherein the smallest group was now 23 samples. The results obtained

for the estimated coefficients for both the mononucleotide STR and polynucleotide STR models are in S4 Table. The model of the *P. vivax* sample without the population origin label also showed high performance for both the mononucleotide STR (AUC=0.9295) and polynucleotide STR (AUC=0.9596). It was found to produce very similar results for the statistically significantly ($P < 0.001$) associated STR variables, and additionally, the MinimumHe variable showed the same negative relationship as in the *P. falciparum* STR models.

Based on the predicted probability of the logistic regression model, for the *P. falciparum* mononucleotide STR model, we select the predicted probability greater than 0.6 as the cut-off value to retain high-quality STRs, while for the *P. falciparum* polynucleotide STR model we chose 0.8 (see Methods for more details). A total of 6,768 high-quality STR (2,563 mononucleotide STR and 4,205 polynucleotide STR) loci were thus selected. The high-quality STRs have been made available through an interactive web-based application for instant data exploration and visualization, and can be accessed at https://github.com/bahlolab/PlasmoSTR. The STRs with a 1-3 bp repeat unit in the 3,047 *P. falciparum* 3D7 samples account for 83.05% of all retained STRs. Of these, homopolymeric tracts account for 37.87%. The dinucleotide repeats account for a further 26.06% (Fig 2A). The motif size of STRs showed differential distribution among various genomic features (Fig 2B). The frequency of the trinucleotide repeats is higher in coding regions than in intronic, intergenic, and promoter regions, which has been previously observed in other species, including humans and is an example of survivorship bias with non 3-mer motifs likely to disrupt the transcript and be deleterious causing strong selection against such STRs in coding regions[33-35]. Introns and intergenic regions mostly show a similar distribution except for the mononucleotide STRs which are enriched in intergenic regions. Promoters also show a difference

in the proportions of different motif sizes compared to the non-coding regions with an abundance of highly polymorphic dinucleotide (2 bp) STRs.

For the *P. vivax* mononucleotide STR model, we selected the predicted probability greater than 0.6 as the cut-off value to retain high-quality STRs, while for the *P. vivax* polynucleotide STR model we chose 0.2. A total of 3,496 high-quality STR (1,648 mononucleotide STR and 1,848 polynucleotide STR) loci were therefore selected (Fig 2C and 2D). Compared with *P. falciparum*, *P. vivax* has fewer 2 bp repeats and more 1, 8 and 9 bp repeats. The distribution of high quality STRs motif sizes (1-9 bp) of the 3,047 *P. falciparum* samples and 174 *P. vivax* samples were significantly associated with the motif size composition of the reference genome (Chi-square test, *P. falciparum*: $P < 2.2 \times 10^{-16}$; *P. vivax*: $P < 2.2 \times 10^{-16}$), as well as the distribution of STRs among various genomic features (Chi-square test, *P. falciparum*: $P = 0.0059$; *P. vivax*: $P < 1.32 \times 10^{-6}$).

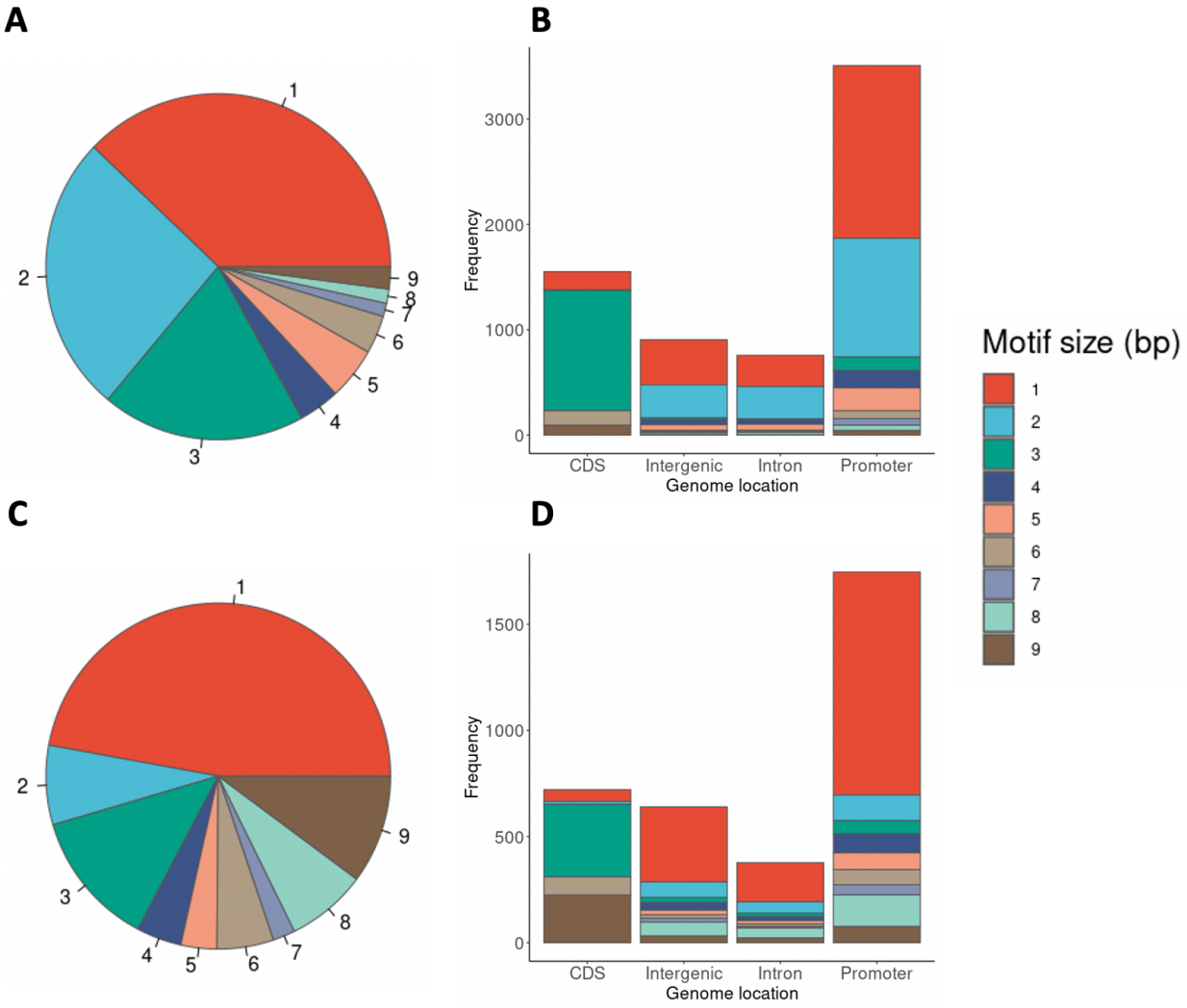



**Fig 2.** (A) Distribution of motif sizes (1-9 bp) of the 3,047 *P. falciparum* samples, colored by the motif size. (B) Motif size dependent distribution of STRs among various genomic features of the 3,047 *P. falciparum* samples. (C) Distribution of motif sizes (1-9 bp) of the 174 *P. vivax* samples. (D) Motif size dependent distribution of STRs among various genomic features of the 174 *P. vivax* samples. The genomic features are labeled along the X-axis for (B) and (D). The frequencies of each motif size are calculated as the total bases covered by STRs of a given motif size divided by the total bases covered by all STRs, labeled along the Y-axis.

## Population structure analysis

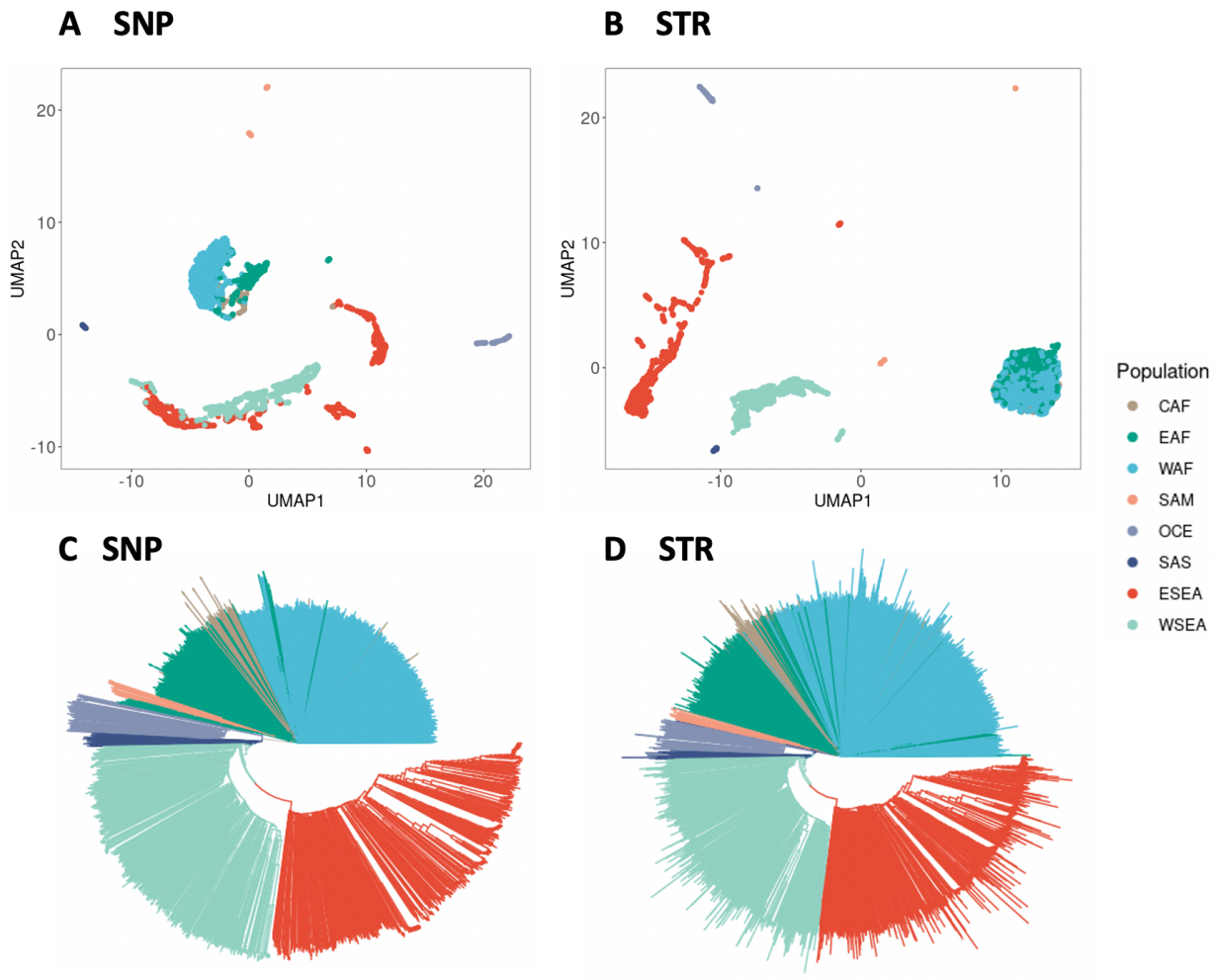
We investigated the population genetic structure of the global *P. falciparum* and *P. vivax* parasite population by performing dimensionality reduction analyses, applying both uniform manifold approximation and projection (UMAP) and principal component analysis (PCA), and generating

neighbour-joining trees (NJTs) for all *P. falciparum* and *P. vivax* samples based on the SNP and STR genotypes. For the *P. falciparum* dataset, UMAP on the top five PCs of both SNP and STR genotypes can distinguish the SAM, OCE, SAS, Asia (WSEA, ESEA), and Africa (WAF, CAF, EAF) parasite populations from different geographic regions, with each of these populations being more strongly differentiated from all other populations, but the SNP data from the WSEA and ESEA populations suggest greater genetic similarity between these populations than the matching STR data. Conversely, the STR data of African sub-regions appear to be genetically more similar than the corresponding SNP data suggests (Fig 3A and 3B). All population subdivisions supported by the UMAP analyses were also present in the PCA analysis (S4 Fig). In general, the global *P. falciparum* parasite population formed four distinct clusters: SAM, Africa (WAF, CAF, EAF), OCE, and the Asia (WSEA, ESEA) region. This clustering may be affected by a variety of factors, including vector species, varying malaria transmission intensity, and the historical usage of antimalarial drugs, all of which are confounded by the time of collection of the samples. To further explore clustering patterns and investigate the average genetic dissimilarity between pairs of individuals, phylogenetic analysis was performed to produce a neighbor-joining tree. The neighbor-joining trees also recapitulate the population structure from the clustering analyses (Fig 3C and 3D). Overall the STR data recapitulates the broad geographical structure of the SNP data, but provides greater resolution of distinct samples at the local scale.

For the *P. vivax* dataset, the PCA analysis (S5 Fig) of both SNP and STR genotypes revealed several distinct clusters that were similar to previous studies[28, 29]: South America (Brazil, Colombia, Peru), Mexico, Southeast Asia (Thailand, Vietnam, Myanmar, Cambodia), PNG, Indonesia and Malaysia. The UMAP analyses and neighbor-joining trees (S6 Fig) also recapitulates the population structure from the clustering analyses.

277 For the sub-population structure analysis from the *P. falciparum* dataset, we found that Ethiopia
278 was genetically distinct from other EAF countries, and that the two countries of Colombia and
279 Peru in the SAM population were also genetically distinct (S7 Fig). This was observed in both the
280 STR and SNP data.

281




282

283 **Fig 3.** Population structure analysis of the 3,047 *P. falciparum* samples of SNP and STR data. (A) UMAP clustering of the top five
284 principal components of the SNP data, colours representing the eight different populations. (B) UMAP clustering of the top five
285 principal components of the STR data with different. (C) NJTs based on the SNP data. (D) NJTs based on the STR data. Branches
286 are colored according to the population.

## Genome-wide genetic differentiation

Estimates of population and country differentiation in the *P. falciparum* and *P. vivax* dataset calculated using the STR data were highly correlated with those calculated using the SNP data for both *Jost's D* and $F_{ST}$ (S8 and S9 Figs). Similar to some previous studies[36-38], we found *Jost's D* and $F_{ST}$ tended to produce values higher in magnitude with STRs than with SNPs. The bi-allelic SNP loci limit the information content per locus compared to the more polymorphic STR markers, which have higher allelic diversity per locus and therefore result in higher estimates of *Jost's D* and $F_{ST}$.

There were significant associations between geographic and genetic distances at the population and country level for both SNP and STR data (*P. falciparum*: Mantel test based on pairwise $F_{ST}$, Fig 4; Mantel test based on pairwise *Jost's D*, S10 Fig; *P. vivax*: *Jost's D* and $F_{ST}$, S11 Fig), indicating that genetic differentiation in populations might be the result of isolation by geographic distance. For the *P. falciparum* dataset, we observed that the genetic differentiation between SAM and ESEA is the largest for both SNP and STR data (genome-wide average SNP $F_{ST}$ 0.11, STR $F_{ST}$ 0.30), and that this geographic distance is also the largest among the populations for both SNP and STR $F_{ST}$. It is worth noting that the genetic differentiation was much larger within the Asia region (SAS, ESEA, WSEA) than within the Africa region (CAF, WAF, EAF), despite the geographic distances being much larger in Africa. This may be due to the higher transmission intensity within Africa[39]. For the country level of the *P. falciparum* dataset, within the Africa region, we identified some country pairs that had higher genetic differentiation both in SNP and STR data, this being driven by the Ethiopian genetic differences, which is consistent with the previous studies demonstrating that Ethiopia is a distinct sub-population[27, 40].

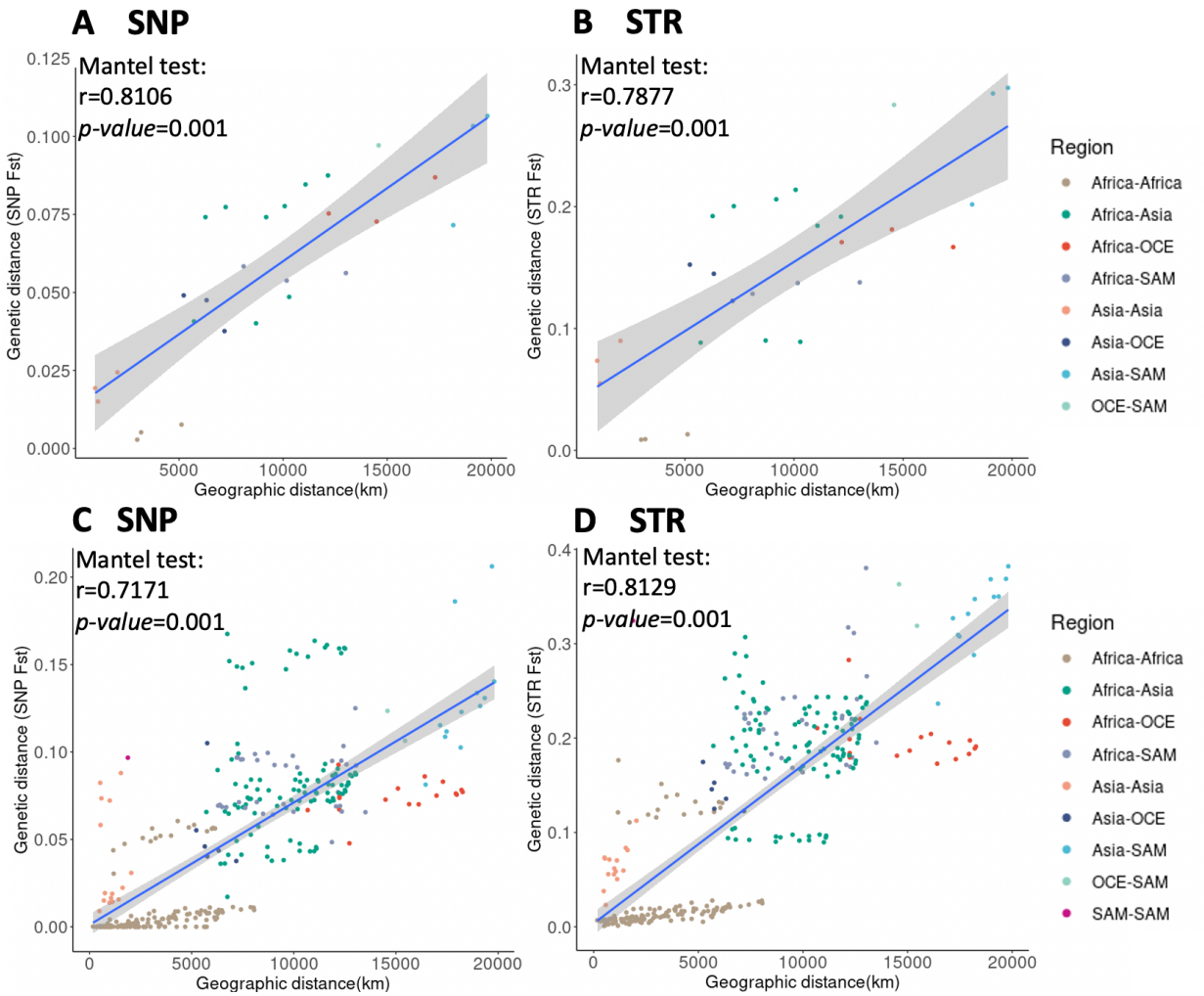



**Fig 4.** Pairwise genetic distance ($F_{ST}$) and geographical distances (km) between populations and countries of *P. falciparum*. (A) SNP data of population pairs. (B) STR data of population pairs. (C) SNP data of country pairs. (D) STR data of country pairs. A Mantel test was used to measure the association.

## Selection signatures related to geographic differentiation

We performed genome-wide scans of the pairwise *Jost's D* values in an attempt to identify regions that were differentiated between the populations or the countries in the *P. falciparum* and *P. vivax* dataset. Several genomics regions with high *Jost's D* values were detected at both the global level

(different regions) and the local level (different countries). A summary of these comparisons between the *P. falciparum* populations is shown in S5 Table, and the comparison between the *P. vivax* countries is shown in S6 Table.

At the global level in the *P. falciparum* dataset, we found three STRs located in coding regions (within PF3D7_0810600, PF3D7_0810900, PF3D7_0811200) which were highly differentiated between the CAF, EAF, and WAF populations. All three STRs were located on chromosome 8, 0.59–12.3 kb from the drug resistance gene *Pfdhps* (PF3D7_0810800, dihydropteroate synthase) (Fig 5A). The first of these STRs (*Jost's D*=0.52), located within PF3D7_0810600 (chromosome 8 544,455-544,481 kb) is composed of an 'AAT' motif (Fig 5A). 78% samples in CAF and 77.99% samples in WAF have the same genotype as the Pf3D7 reference genome of nine 'AAT' motifs, whereas 86.23% samples in EAF have two 'AAT' insertions. The nonsynonymous mutation (Pf3D7_08_v3:g.543210G>T) in PF3D7_0810600 were detected in artemisinin-resistant cell lines by Frances et al. that might play a role in gene expression regulation and subsequently contribute to the artemisinin resistance phenotypes[41].

The second of the three coding STRs is located between the genome reference coordinates 2,260,430-2,260,449 kb, within PF3D7_1455300. It consists of an 'AAT' motif, which was highly differentiated between the Africa region (CAF, EAF, WAF) and the Southeast Asia region (WSEA, ESEA) (Fig 5B). 81.02% of samples in the Africa region have the Pf3D7 reference genotype of seven 'AAT' motifs, whereas 97.17% of samples in the Southeast Asia region have two 'AAT' deletions. PF3D7_1455300 is a conserved *Plasmodium* protein that plays a role in DNA mismatch repair. According to previous work[42], it is a candidate molecular marker for altered DNA repair capability. SNP mutations previously found in this gene may be associated with the phenotype of

accelerated resistance to multiple drugs (ARMD). Also, the SNP mutations identified in PF3D7_1455300 (Pf3D7_14_v3:g.2260945T>G) by Xiong et al. have a high frequency in the Southeast Asia population, but cannot be found in African populations, which may be due to selection and thus be a signature for the Southeast Asia population selection signal[42]. In our study, we also found that the STR mutations in PF3D7_1455300 are significantly different in Southeast Asia and Africa. Aside from this locus, another STR within PF3D7_1431400 (surface-related antigen) is located between the genome reference coordinates 1,234,853-1,234,865 kb and consists of a monomer 'T' motif, which also showed high differentiation (Fig 5B). 97.76% samples in Africa have the Pf3D7 reference genotype of 13 'T' motif repeats, whereas 89.52% samples in Southeast Asia have an insertion of 24 'T' repeats.

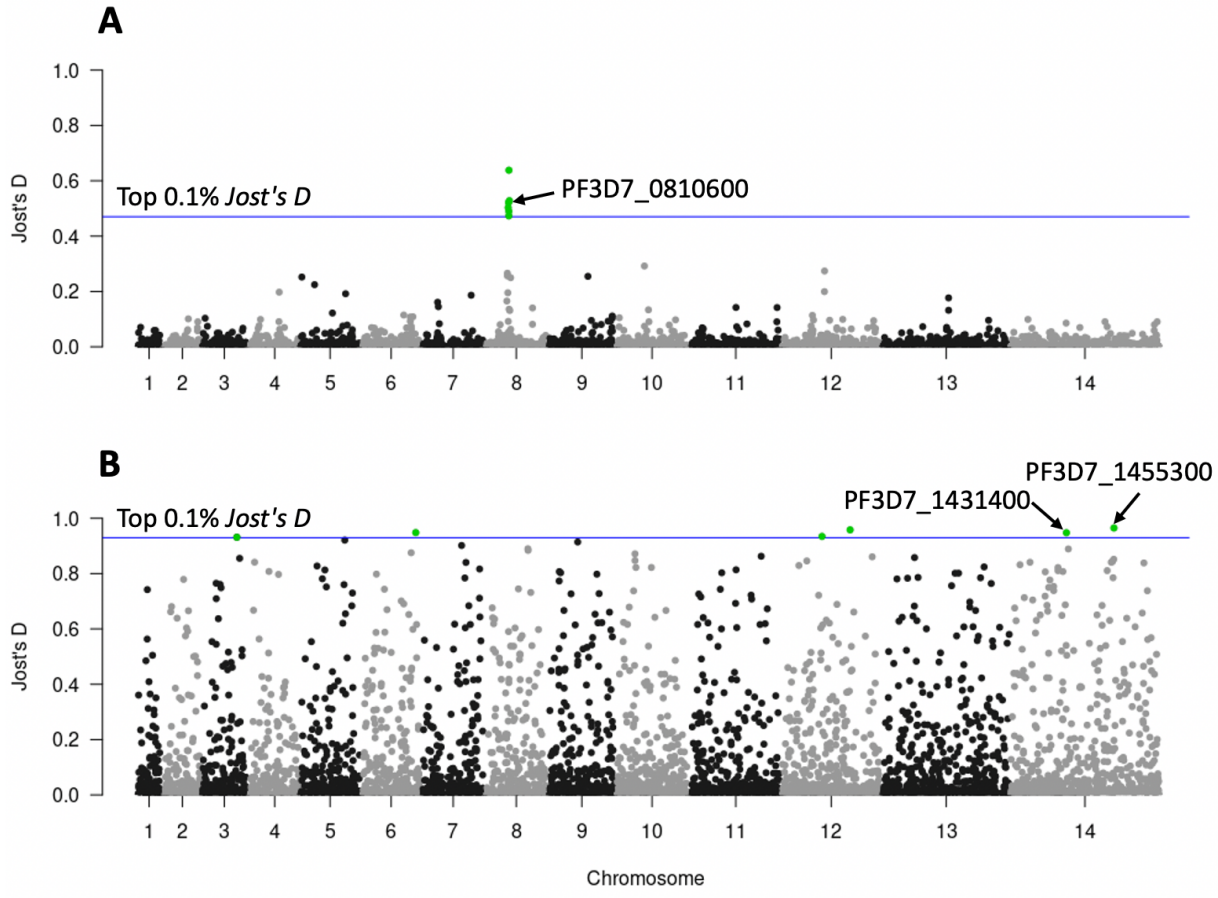


**Fig 5.** Genome scans for differentiation, as measured by *Jost's D* values. (A) CAF, EAF, and WAF samples. (B) Africa region (CAF, EAF, WAF) with Southeast Asia (WSEA, ESEA) region samples. The x-axis represents the chromosomes and the y-axis the *Jost's D* values. Each point represents an STR locus with a total 6,768 STRs represented. The blue horizontal line represents the threshold based on the top 0.1% *Jost's D* values.

At the local level in the *P. falciparum* dataset, we also found several STRs which showed differentiation between countries. Within WAF countries, we identified a set of STRs located in coding regions of the genome, with potentially direct functional impact. These STRs appeared to be under positive directional selection: PF3D7_0627800 (acetyl-CoA synthetase), which was predicted as being under balancing selection[43]; and PF3D7_0826100 (HECT-like E3 ubiquitin ligase), found to be possibly involved in a mechanism of drug resistance to pyrimethamine[44, 45] and which may also be involved in reduced susceptibility to quinine and quinidine[46]. Additionally we also identified STRs in the coding regions of PF3D7_0416000 (RNA-binding protein); PF3D7_0811000 (cullin-1); PF3D7_0811200 (ER membrane protein complex subunit 1); PF3D7_1210400 (general transcription factor 3C polypeptide 5), PF3D7_1409100 (aldo-keto reductase) and two conserved proteins with unknown function: PF3D7_0107100 and PF3D7_0604000. Within EAF countries, selection signals included STRs in PF3D7_0628100 (HECT-domain (ubiquitin-transferase), which was previously observed to have a strong signature of deviation from neutrality in Gambia based on STR analysis[47]; PF3D7_0527900 (ATP-dependent RNA helicase DDX41); PF3D7_1212900 (bromodomain protein 2); PF3D7_1331100 (DNA polymerase theta); and three conserved proteins with unknown function: PF3D7_0526600, PF3D7_0810900 (0.59kb away from the drug resistance gene *Pfdhps*) and PF3D7_1448500. Two STRs within PF3D7_1433400 (PHD finger protein PHD2) and PF3D7_0926600 (conserved *Plasmodium* membrane protein, unknown function) were highly differentiated between WSEA countries. Several STRs located in the coding region were also found to be highly differentiated

375 between ESEA countries. PF3D7_1225100 (isoleucine--tRNA ligase); PF3D7_0826100 (HECT-
376 like E3 ubiquitin ligase); PF3D7_1317900 (nucleolar complex protein 4); and two conserved
377 protein with unknown function: PF3D7_1233200 and PF3D7_1303800.

378 We extracted the top ten most highly differentiated STRs from each pairwise population and
379 country comparison to determine if a small number of highly differentiated STRs could represent
380 population structure. The minimum spanning network can distinguish between two distinct groups
381 of samples (*P. falciparum*: Fig 6; *P. vivax*: S12 Fig). The top ten most highly differentiated STRs
382 from each pairwise comparison of the *P. falciparum* and *P. vivax* dataset have been made available
383 through the R shiny PlasmoSTR, accessible at https://github.com/bahlolab/PlasmoSTR. From the
384 analyses, we can identify STR mutations that are fixed in one population but that are distinct from
385 other populations and which may become a signature for the specific population.

386

387

388

389

390

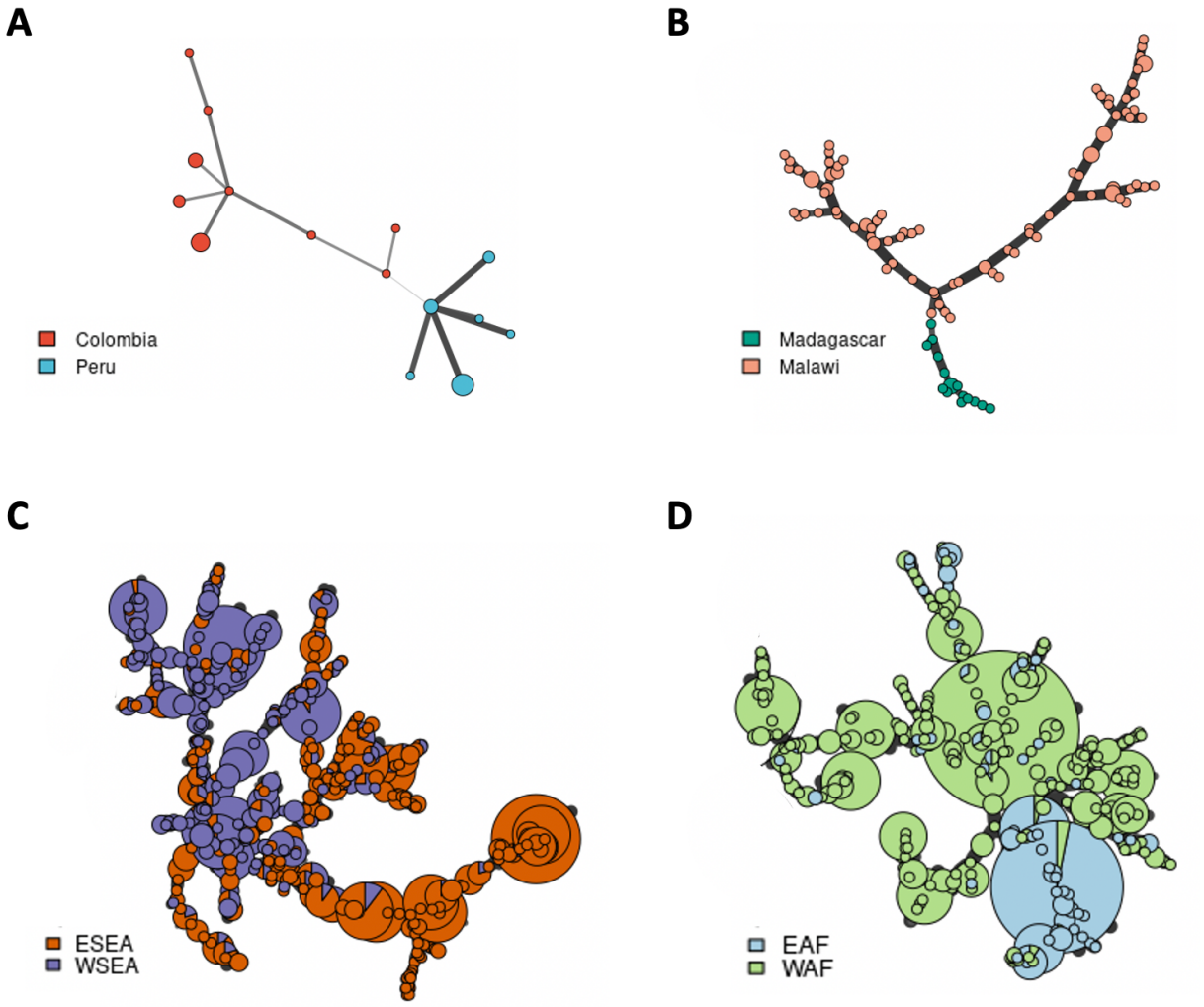



**Fig 6.** Minimum spanning network using Bruvo's distances based on the ten most informative STR markers showing the relationship among two groups of *P. falciparum* isolates. (A) Colombia and Peru from the SAM population. (B) Madagascar and Malawi from the EAF population. (C) ESEA and WSEA populations. (D) EAF and WAF populations. Colors correspond to the country or population. Node sizes correspond to the number of samples. Edge lengths are arbitrary.

## Selection signatures related to drug resistance

Resistance of malaria parasites to chloroquine is known to be associated with the parasite protein *Pfcrt*. Samples were classified as chloroquine-resistant if they carried the *Pfcrt* 76T allele[27]. Chloroquine-resistance was found in almost all samples from SAM, OCE, SAS, WSEA, and ESEA.

It was also found across the Africa region (WAF, CAF, EAF), but the frequency is low, especially in EAF. However, it is noteworthy that all samples from Ethiopia were classified as chloroquine-resistant as they all carried the *Pfcrt* 76T allele, and also displayed a higher genetic differentiation with both SNP and STR data with other EAF countries. To identify regions with signatures of selection that may be associated with chloroquine resistance, we calculated the *Jost's D* per STR genome-wide among the 57 EAF drug-resistant and 269 EAF drug-sensitive samples. Average genome-wide *Jost's D* estimates were 0.011, considering the top 0.1% *Jost's D* threshold, signatures of selection were detected at six STR loci that had *Jost's D* values > 0.36, located on chromosomes 7, 11, and 14 (Fig 7A). Three STRs were located on chromosome 7, 1.7–22.2 kb from the drug resistance associated gene *Pfcrt* (PF3D7_0709000, chloroquine resistance transporter), demonstrating that drug selection produces chromosomal segments of selective sweeps as we have previously demonstrated with SNP data[48]. One of these STRs (*Jost's D*=0.82) was located in the genomic promoter region between the genome reference coordinates 392,230-392,268 kb which consists of an 'AT' motif. Interestingly, all of the drug-resistant samples in Ethiopia have one 'AT' deletion compared to the reference genome (S13A Fig), where the EAF drug-sensitive samples range from the six 'AT' deletion to 11 'AT' insertion (S13B Fig). The length of the promoter region is gene-specific, and the levels of gene expression can be increased or decreased by expanding and contracting in length[49]. This is known as an STR expression quantitative trait locus (STReQTL). To demonstrate this is a true STReQTL would require a dual WGS, RNA sequencing (RNA-seq) dataset for the parasite or alternatively a lab-based investigation of expression levels for the different STR genotypes.

Artemisinin resistance of malaria parasites is known to be associated with the Pfk13 (kelch 13) gene, and samples were classified as artemisinin-resistant if homozygous non-synonymous

mutations occurred in the kelch13 BTB/POZ and propeller domain1. Artemisinin-resistant samples were only found in samples from WSEA and ESEA. We observed that in the ESEA population almost all samples from Laos were classified as artemisinin sensitive, while almost all samples from Thailand were classified as artemisinin resistant. To identify genomic regions under selection due to artemisinin resistance, we calculated the *Jost's D* per STR genome-wide among the 77 drug-sensitive samples in Laos and 16 drug-resistant samples in Thailand. The average genome-wide *Jost's D* estimate was 0.086. Signatures of selection were detected at six STR loci that had *Jost's D* values > 0.95, located on chromosome 5, 12, 13, and 14 (Fig 7B). One particular STR (*Jost's D*=0.99) on chromosome 12, within the api-IRS (isoleucine--tRNA ligase) gene (PF3D7_1225100; STR position 1,023,293-1,023,313), has a variable length 'AAT' motif. Fifteen out of the 16 of the drug-resistant samples in Thailand have three 'AAT' insertions, whereas the drug-sensitive samples in Laos vary from the two 'AAT' deletions to two 'AAT' insertions. Strong signatures of differentiation (*Jost's D*=0.98) were also observed within the NOC4 (nucleolar complex protein 4) gene (PF3D7_1317900; STR position 744,342-744,366) which consists of an 'ATT' motif. Fifteen out of 16 of the drug-resistant samples in Thailand have three 'ATT' insertions, where the drug-sensitive samples in Laos range from the one 'ATT' deletion to three 'AAT' insertions, 70.13% have the same genotype as the reference genome.

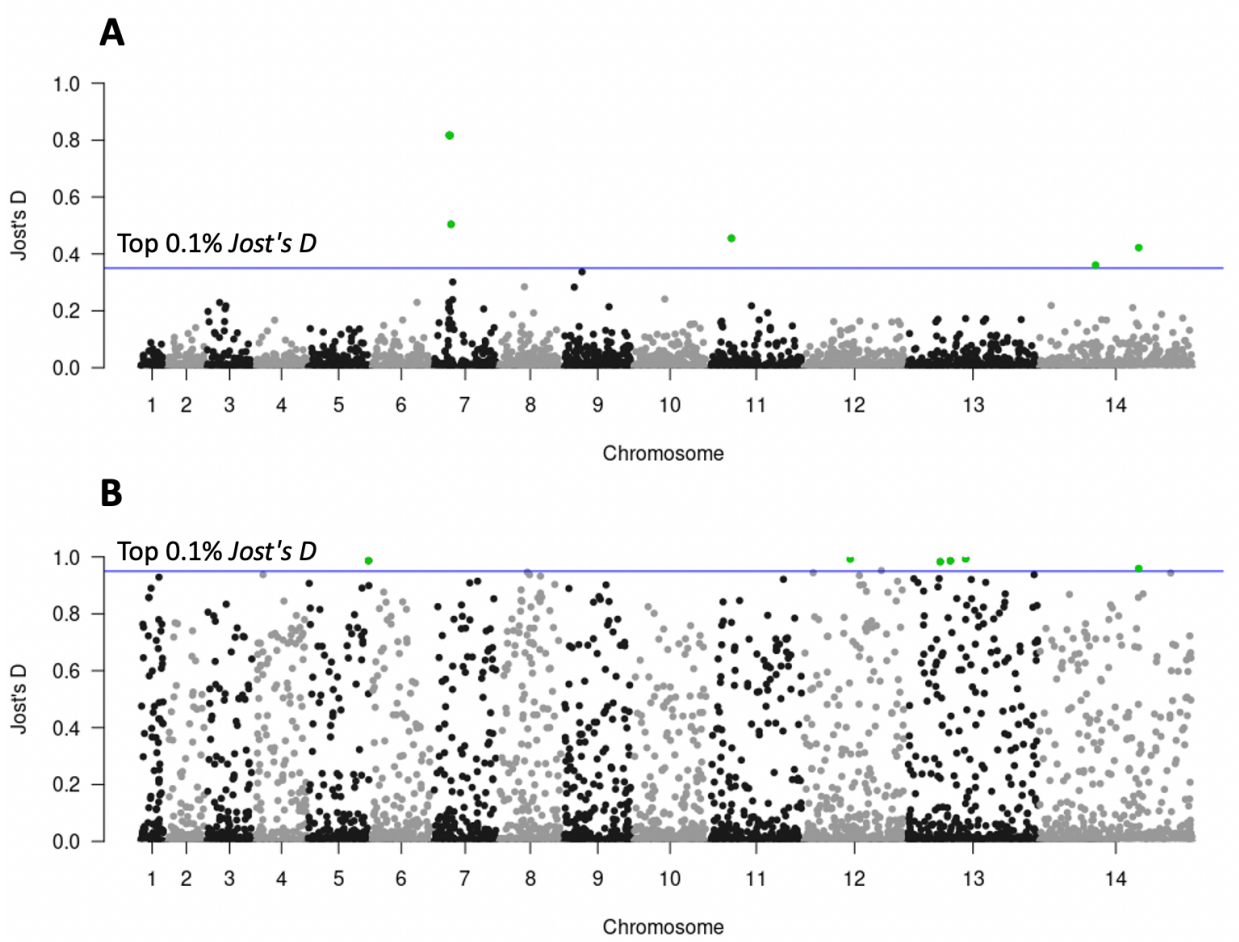



**Fig 7.** Genome scans for differentiation for kelch13 drug resistance. (A) 57 EAF drug-resistant and 269 EAF drug-sensitive samples. (B) 16 drug-resistant samples in Thailand and 77 drug-sensitive samples in Laos. The x-axis represents the chromosomes and the y-axis the *Jost's D* values. Each point represents a STR loci. The blue horizontal line represents the threshold based on the top 0.1% *Jost's D* values.

# Discussion

In this study we genotyped thousands of STRs applying an *in-silico* or bioinformatic approach (HipSTR). We performed STR genotyping for the first time on the *P. falciparum* and *P. vivax* genomes comprising more than 3,000 *P. falciparum* and 174 *P. vivax* WGS samples, obtained from across the world. To our knowledge this is the first time this has been attempted for this

population dataset and a dataset of this size. To measure HipSTR's quality of prediction, we use a set of *P. falciparum* STR markers[50], which have been genotyped with gel electrophoresis (GE) within a subset of the in-house *P. falciparum* samples. A strong linear relationship between GE allele calls against HipSTR's calls can indicate that HipSTR is predicting genotypes adequately.

STRs are an important source of genetic diversity, and are generally more informative than SNP markers due to the higher number of variants per locus. STRs also have a much higher mutation rate ($4.43 \pm 0.37 \times 10^{-7}$ per locus per asexual cycle for *P. falciparum*), ~1,000 times higher than SNPs ($3.18 \pm 0.74 \times 10^{-10}$ base substitutions per site per asexual cycle for *P. falciparum*)[4]. This may capture very recent evolution more robustly and could be used to distinguish closely related samples in clonal outbreaks and could potentially be used to distinguish whether the recurrent infection represents reinfection or recrudescence. The ability to do so is an important tool for countries aiming for elimination of *P. falciparum* or *P. vivax* malaria and is particularly important for *P. vivax* with its ability to reactivate malaria from its dormant liver stage.

STRs are not routinely analyzed across the whole genome with short read sequencing data because they are difficult to identify and genotype accurately, requiring rigorous QC. In order to attain reliable sets of STRs, some studies have sequenced each sample twice as a technical replicate and used multiple STR calling algorithms to test variant calling accuracy and keep the high-quality STRs[4, 51]. However, it is not possible or even practical to use these filtering approaches in large-scale field samples where there are unlikely to be technical replicates. Furthermore, technical replicates are only able to identify a limited set of problematic STRs. To overcome these limitations in the current study, we developed a novel method for quality control of STR genotyping data based on gold standard SNP genotyping data from the same cohort. We

demonstrated that this was a successful approach and replicated it in a second *Plasmodium* species, *P. vivax*, demonstrating that this is a method which can be broadly applied to many other species. Our results provided new insights for further exploration of STRs across the whole genome.

We built separate multivariable logistic regression modelling for measurement and prediction of the quality of STRs for the mononucleotide STR and polynucleotide STRs. This is because genotyping homopolymers is particularly challenging for many STR tools, including HipSTR, and has a high error rate[26] which has led them to be discarded altogether for other studies[23]. For both the mononucleotide STR and polynucleotide STR model, the STR quality is highly influenced by the STR features that capture population specific aspects of the cohort, while the mononucleotide STR quality is also highly influenced by HipSTR's Mean_Posterior metric, a parameter that indicates the quality of the called STR genotype. Higher allelic diversity STRs indicate greater genetic variability among the samples, hence we also considered the effect of extremely polymorphic STRs using some additional parameters in our models: population differentiation (*Jost's D*), and mean, maximum and minimum heterozygosity across the different populations, to tease apart the drivers of high-quality STRs based on population-dependent measures. This population-dependence also led us to fit models using naive clustering based labels, however these produced very similar results but are useful QC steps in cohort studies, especially if reported population membership is uncertain. Our models showed high AUC values both in the mononucleotide STR and polynucleotide STR model, which can effectively combine a range of different STR features to predict the STR quality.

Genome-wide SNP genotyping[27, 52, 53] and a set of STR markers (<20) genotyped using lab-based approaches such as capillary genotyping [14-16] have been employed in several studies to

494 reveal the *P. falciparum* and *P. vivax* genetic diversity and population structures. However, the
495 overall contributory effect of STR variation in *P. falciparum* and *P. vivax* has not been evaluated
496 using genome-wide sequencing data from a larger collection of samples representing the global
497 distribution due to the limitations of scaling up of the lab-based STR genotyping approaches.

498 Our clustering results, investigating parasite population genetics, demonstrated general agreement
499 of clustering by the population of origin between SNP and STR markers. For the *P. falciparum*
500 dataset, the overall population genetic structure of parasites represents four distinct groups: SAM,
501 OCE, Asia (SAS, WSEA, ESEA), and Africa (WAF, CAF, EAF) populations. PCA shows the
502 SNP data resulted in tighter groups of individuals compared to the somewhat loose clusters of
503 individuals with STRs data. The neighbor-joining trees based on IBS analysis showed that STR
504 data have a higher power to identify groups in SAM, EAF, and CAF (213,757 SNP loci versus
505 6,768 STR loci), this is likely due to STRs that are highly polymorphic and have multiple alleles
506 thus providing higher information content as compared to the biallelic SNPs. There is some
507 circularity in these results since our QC measure was predicated on capturing geographic
508 information. Nonetheless this was a general approach and the clustering analysis is a finer-scale
509 method which further supported the validity of the approach.

510 Significant Mantel correlations between geographical and genetic distances (based on $F_{ST}$ and
511 *Jost's D*) at the population and country level for both SNP and STR data in the *P. falciparum* and
512 *P. vivax* dataset were detected, suggesting that genetic differentiation in populations are likely the
513 result of geographic isolation. The *P. falciparum* parasites from the Africa region (CAF, EAF,
514 WAF) have lower levels of population structure, and the genetic differentiation within the African
515 region is lower than within the Asian region (SAS, ESEA, WSEA), although noting that the

516 geographic distances are greater in Africa. This is likely due to the high transmission intensity in
517 Africa, where individuals are more likely to be infected by more than one *P. falciparum* parasite,
518 which increases the frequency of recombination, leading to a highly diverse population with low
519 linkage disequilibrium[39]. At the country level in the *P. falciparum* dataset, it is worth noting
520 Ethiopia in the EAF population which displayed a higher genetic differentiation with other EAF
521 countries both in SNP and STR data, which was consistent with the previous study that stated
522 Ethiopia is a distinct sub-population[27, 40]. In Ethiopia, over 75% of the land surface is at risk,
523 with varying intensities of malaria, unlike other many African countries, Ethiopia is also unique
524 in that *P. vivax* is co-transmitted with *P. falciparum*, further evidence of a high malaria burden.
525 Higher rainfall, temperature, humidity and seasonal transmission in Ethiopia could also be the
526 driving factors of the higher genetic differentiation[54-56]. Additionally, all of the samples in
527 Ethiopia were classified as chloroquine-resistant as they carried the *Pfcrt* 76T allele. In contrast,
528 the same allele was observed at much lower frequencies in other EAF countries[27]. This could
529 also explain the higher genetic differentiation between Ethiopia and other EAF isolates.

530 To scan the STR genomic loci under divergent selection that might occur due to varying
531 antimalarial drug use or local differences, *P. falciparum* and *P. vivax* samples were analysed from
532 the population pairs and the country pairs. For the *P. falciparum* dataset, we identified several
533 STRs with outlier *Jost's D* values, including strong signatures of genetic differentiation, likely due
534 to selection, around the chloroquine resistance transporter, *Pfcrt* (PF3D7_0709000), and
535 dihydropteroate synthase, *Pfdhps* (PF3D7_0810800, dihydropteroate synthase). We also identified
536 several selection signals including the ATP-dependent RNA helicase *Pfdbp1* (PF3D7_0810600),
537 which may be associated with artemisinin resistance[41]; the HECT-like E3 ubiquitin ligase *Pfheul*
538 (PF3D7_0826100), found to be possibly involved in a mechanism of drug resistance to

539 pyrimethamine[44, 45] and may also be involved in reduced susceptibility to quinine and
540 quinidine[46]; and a conserved *Plasmodium* protein PF3D7_1455300 that may be associated with
541 the phenotype of accelerated resistance to multiple drugs (ARMD)[42]. The STR variations
542 observed in *P. falciparum* drug-resistant samples may reflect the differences in the historical use
543 of antimalarial drugs of different countries and may contribute to the development of local malaria
544 treatment guidelines. Multiple STR loci that had strong signatures of deviation from neutrality
545 were also detected, which was consistent with previous studies[43, 47]. However, most of the
546 previous studies detected the selection signatures through association with SNP-based signals,
547 whereas in our study we found that the STR mutations within different genes also play an important
548 role. The key question is whether some of these STRs may actually be the driver mechanism
549 underpinning the selection signals rather than merely showing association due to linkage
550 disequilibrium. The small changes in the length of STR mutations may alter protein activity,
551 protein folding efficiency, stability, or aggregation[10], and the levels of gene expression can be
552 increased or decreased by expanding and contracting in length that allows the parasite to adapt
553 under selective pressure[49]. These signals could be actively pursued in laboratories to investigate
554 whether the STR signals directly affect relevant expression signals as STReQTL. Unlike the
555 human Genotype–Tissue Expression Project (GTEx)[57], a comprehensive public resource,
556 including both WGS and RNA-seq datasets, which can identify STRs associated with expression
557 of nearby genes is not available. One limitation in our study is that we are unable to determine if
558 these STRs are true STReQTL due to the absence of RNA-seq data in the *Plasmodium* datasets
559 analyzed. Many novel candidate genomic regions that were likely under recent positive directional
560 selection were also detected in our study, possibly revealing recent signals of selection not yet
561 observable with SNP markers.

# Conclusions

In this paper, we report the first large-scale *in-silico* STR study performed in more than 3,000 *P. falciparum* and 174 *P. vivax* WGS worldwide samples. We developed a novel method for quality control of STR genotyping data based on gold standard SNP genotyping data, which provides new insights for further exploration of STRs across the whole genome. Furthermore, a set of genome-wide high-quality STRs were then used to study parasite population genetics and compared to genome-wide SNP genotyping data, revealing both high consistency with SNP based signals, as well as identifying some signals unique to the STR marker data. These results demonstrate that the identification of highly informative STR markers from large population screening is a powerful approach to study the genetic diversity, population structures and genomic signatures of selection on *P. falciparum* and *P. vivax*. In addition, the genome-wide information about genetic variation and other characteristics of STRs in plasmodium have been made readily available in an interactive web-based R Shiny application PlasmoSTR (https://github.com/bahlolab/PlasmoSTR).

# Materials and Methods

## Data

### MalariaGEN global *P. falciparum* dataset.

All samples and metadata were obtained through the MalariaGEN *Plasmodium falciparum* Community Project (https://www.malariagen.net/resource/26)[27]. We retrieved the data in fastq file format from the Sequence Read Archive (SRA). The *P. falciparum* dataset consists of 3,241

monoclonal (within-host infection fixation index $F_{ws}$ > 0.95 downloaded from MalariaGEN) samples. Metadata was available in the form of population labels for all samples representing the country of origin of each sample at both a population (8 levels) and country level (27 levels). The populations were SAM, WAF, CAF, EAF, SAS, WSEA, ESEA and OCE, and the countries were: Ghana, Cambodia, Bangladesh, Thailand, Colombia, Malawi, Guinea, Uganda, Ethiopia, Mali, Senegal, Gambia, Mauritania, Peru, Nigeria, Myanmar, Laos, Viet Nam, Kenya, Tanzania, Papua New Guinea, Burkina Faso, Congo DR, Madagascar, Cameroon, Ivory Coast, and Benin[27]. Based on published genetic markers including SNPs and copy number variations (CNVs), all samples are classified into different types of drug resistance in the MalariaGEN *Plasmodium falciparum* Community Project[27]. As previously described[27], sequencing was performed using Illumina HiSeq 2000 paired-end sequencing platform. The *P. falciparum* 3D7 (v3 PlasmoDB-41) was used as the reference genome and was downloaded from PlasmoDB[58]. In this study, we only considered monoclonal samples because STR genotyping algorithms such as HipSTR are optimised for diploid and haploid chromosomes, without considering the possibility of multiplicity of infection (MOI)[1, 23].

**Global *P. vivax* dataset.**

The dataset for *P.vivax* comprised 353 previously published samples in *Plasmodium vivax* Genome Variation project (https://www.malariagen.net/projects/p-vivax-genome-variation) as described in Pearson et al.[29] and data from Hupalo et al.[28], which are sampled from multiple countries around the world. Fastq files were also downloaded from SRA. The whole genome sequencing was performed using Illumina-based sequencing platforms. The *P. vivax* genome PvP01 (PlasmoDB release 41) was used as the reference genome and was downloaded from PlasmoDB[58].

## Methods

### SNP genotyping

SNPs and deletions/insertions (Indels) were called using the standard best practice from Genome Analysis Toolkit (GATK) version 4.0.12.0 implemented in nextflow[22, 59] (https://github.com/gatk-workflows/gatk4-germline-snps-indels). The pipeline generates a final joint VCF file for all samples. Variants were further removed with the following filtering thresholds: Quality of Depth (QD) < 20, Mapping Quality (MQ) < 50, MQ Rank Sum (MQRankSum) < -2, Strand Odds Ratio (SOR) > 1, and Read Position Rank Sum (ReadPosRankSum) less than -4 or greater than 4. SnpEff was used to annotate variants based on the *P . falciparum* 3D7 and *P. vivax* P01 reference genome[60]. SNPs were excluded if they were: (i) indels, (ii) not biallelic, (iii) variants in genes from the surface antigen (VSA)[61] families, (iv) not in core genome region defined by Miles et al.[20] for *P . falciparum* and Pearson et al.[29] for *P. vivax*, (v) if their minor allele frequency (MAF) was less than 1% in all populations, or (vi) their missing genotype frequency was higher than 10%.

### STR genotyping

We initially identified the composition and distribution of STRs in the *P. falciparum* 3D7 and *P. vivax* P01 reference genome using Tandem Repeats Finder (TRF Version 4.09)[32]. The parameters used for TRF were: the alignment weights for matching (Match) equal to 2, mismatching (Mismatch) penalty is 7, indel (Delta) penalty is 7, the match probability (PM) is 80, the indel probability (PI) is 10, the minimum alignment score (Minscore) is 20, the maximum period size (MaxPeriod) (the pattern size of the tandem repeat) to report is 500bp. Additional post-processing steps of the TRF output files were performed by removing STRs: (i) with overlapping

626 STRs, (ii) with motif period size > 9bp, (iii) repeat number of the motif < 3, (iv) where the percent
627 of matches ≤ 85%, (v) where the percentage of indels ≥ 5%, (vi) the repeat length was larger than
628 70bp, as the genotype call rate declined for longer tandem repeats. Genome-wide STR genotyping
629 was performed with HipSTR (Version 0.6.2)[1] using the haploid version under the default
630 parameters. STRs were then excluded if they were from VSA[61] families, or if they were not in
631 the core genome[20, 29], or if their missing genotype rate was higher than 10%. The
632 VariantAnnotation[62] R package (Version 1.32.0) was used to annotate variants making use of
633 the Pf3D7 and PvP01 gene annotation in GFF format.

634 To measure HipSTR's quality of prediction, we used 10 *P. falciparum* STR markers proposed by
635 Anderson et al. (1999)[50]. These markers have been genotyped with GE on Applied Biosystems
636 3700 (ABI3700) within a subset of in-house *P. falciparum* samples. Whole-genome sequencing
637 was also performed on these samples. GE is typically taken to be the gold standard for STR
638 genotyping. GE data was only available for Milne Bay and East Sepik samples (90 samples). S7
639 Table represents the set of markers we have for *P. falciparum* in addition to how many samples
640 had genotypes called by HipSTR at these specific markers. The locations of the markers were
641 obtained by using a BLAST search with PlasmoDB[58] on the primer sequences for each STR
642 marker which were presented in Anderson et al. (1999)[50]; Figan et al. (2018)[16]; Greenhouse
643 et al. (2006)[63]. Given that the locations of the markers are known, we can compare HipSTR's
644 genotype calls to the length of the STR markers as determined by the GE procedure. It is important
645 to note that the reported length of STRs from GE are generally shifted by a fixed number of base
646 pairs due to the primers being used, which add to the product length[50]. Regardless there should
647 still be a linear relationship between the two classifications if HipSTR is predicting genotypes well.

## Characterization of within-host diversity

We applied the $F_{ws}$ metric to the *P. vivax* dataset to determine samples that had multiple infections[30]. Samples with $F_{ws} < 0.95$ were considered multiple infections[31]. $F_{ws}$ was calculated using the moimix (Version 0.0.2.9001)[64] R package. Samples with multiple infections were excluded from further analysis.

## Multivariable logistic regression modeling for measurement and prediction of the quality of STRs

Although *in-silico* STR genotyping methods have QC metrics that can be used to identify well performing STRs, these have been shown to retain many poorly performing STRs, which are not easy to identify. In order to collate a set of high-quality STRs, we developed a complex filtering strategy based on leveraging genetic distance between samples as determined by SNPs and STRs, aiming to identify further, more precise STR relevant metrics that could be applied to identify high quality STRs. The rationale here is that variants with high genotyping accuracy should more accurately represent the population structure of field samples. Based on this, we developed a SNP PCA based approach to capture the signal STRs. PCA was first performed to investigate potential population structure using the SNP genotype data (*P. falciparum*: 213,757 SNPs loci; *P. vivax*: 188,571 SNPs loci). The top ten principal components were chosen. The squared Spearman's correlation coefficient ($R^2$) was then used to assess correlations between the top ten SNP PCs values and each STR by using the repeat units (number of times the motif is repeated in tandem) for each sample, using only those STRs retained after the initial QC step described above (*P. falciparum*: 20,196 STRs; *P. vivax*: 23,146 STRs). STRs that correlated with an $R^2$ above a permutation derived threshold with any one of the ten significant SNP PCs were deemed to be

high quality STRs. The optimal cut-point of the correlation to distinguish high-quality or low-quality STRs was determined using the resampling permutation test where the population labels were permuted between samples to derive a null distribution to determine an appropriate correlation threshold that maximised the difference between high and low quality STRs. This was determined using the ROC AUC.

A set of metrics including QC metrics from HipSTR and metrics that were deemed useful were derived and used for the prediction of the STR quality. These are summarised in S1 Table. The Z-score standardization method was used to normalize these metrics. Considering the large difference of sample size between the *P. falciparum* and *P. vivax* dataset, when calculating the metrics that are associated with population structure, the *P. falciparum* dataset is based on 8 population-level labels and the *P. vivax* dataset is based on 11 country-level labels. A multivariable logistic regression analysis was then performed to identify potential predictors of STR quality. Empirical clustering and subsequent labelling was performed by using SNP genotype data to calculate the identity-by-state (IBS) pairwise distance between samples using the SNPRelate (Version 1.20.1)[65] R package. Clustering analysis was then performed to assign the samples clusters which were assumed to represent geographical regions.

Model selection in the multivariable regression models was employed for stepwise regression analysis based on the Akaike information criterion (AIC) using the R package MASS (Version 7.3-51.5). The effectiveness of the prediction was evaluated by calculating the AUC on the ROC curve. To assess the predictive performance of the logistic regression model, the large *P. falciparum* dataset was randomly separated into five combinations of training and test sets in an 80/20 split, and fivefold cross-validation was performed. Predictive performance was measured

with the AUC in the testing model. To assess the robustness of the performance of the model with respect to different size datasets, the large *P. falciparum* dataset was also randomly separated into five combinations of training and test sets of each 70/20, 60/20, 50/20, 40/20, 30/20, 20/20, and 10/20 splits, and fivefold cross-validation was performed. To select the optimal cut-off value to remove low-quality STRs, the predicted probabilities are sorted into five bins ([0, 0.2), [0.2, 0.4), [0.4, 0.6), [0.6, 0.8), [0.8, 1]). For each bin we randomly selected 500 STRs and calculated the correlation of sample pairwise distances of STR and SNP based on PCA analysis, and repeated this 100 times to calculate the mean value of correlation.

The R script used to perform the multivariable logistic regression modeling for measurement and prediction of the quality of STRs is available on https://github.com/bahlolab/PlasmoSTR.

**Population structure analysis**

To investigate the major geographical division of population structure that could be determined with the final set of STR markers, the SNP-based and STR-based PCA of all 3,047 *P. falciparum* and 174 *P. vivax* samples were performed separately. In the SNP-based PCA we used 213,757 SNPs for *P. falciparum* and 188,571 SNPs for *P. vivax*. In the STR-based PCA, we used 6,768 (2,563 mononucleotide STR and 4,205 polynucleotide STR) high-quality loci for *P. falciparum* and 3,496 (1,648 mononucleotide STR and 1,848 polynucleotide STR) for *P. vivax* selected by the logistic regression model based on predicted probabilities. PCA plots were constructed from the analysis. UMAP[66] was performed after selecting the significant PCs using the umap (Version 0.2.4.1) R package. SNP and high-quality STR loci across the whole genome were used to calculate the average IBS distances as the average genetic dissimilarity between pairs of individuals. The R package SNPRelate[65] was used to calculate the IBS values for SNP data and

a method based on Bruvo's distance[67] was used for STR data, providing a stepwise mutation model appropriate for microsatellite markers. Neighbour-joining trees were then produced using the R package ape (Version 5.4-1)[68] and ggtree (Version 2.0.4)[69].

**Genome-wide genetic differentiation**

Pairwise estimates of genetic differentiation ($F_{ST}$) between all pairs of populations and countries defined by geographic origin were calculated using the R package SNPRelate[65] for SNP data and the R package hierfstat (Version 0.5-7)[70] for microsatellites based on the method of Weir and Cockerham (1984)[71]. The degree of population differentiation was also measured by calculating *Jost's D* using the R package mmod (Version 1.3.3)[72], which is a superior diversity measure for highly polymorphic loci proposed by Jost[73]. The geographic distance between different populations and countries (km), were calculated using the R packages sf, maps, units, and rnaturalearth. The Mantel test was performed by the R packages vegan (Version 2.5-7)[74] to study the correlations between pairwise values of genetic distance and geographical distance between populations and countries, and also used to check the correlation between pairwise differentiation measures (*Jost's D* and $F_{ST}$) from SNPs and STRs data.

**Selection signatures related to geographic differentiation**

Global *Jost's D* were calculated per STR for each pairwise population combination (*P. falciparum*: 8 populations = 28 comparisons) and country combination (*P. falciparum*: 26 countries = 325 combinations) using the mmod[72] R package. For *P. vivax*, we only compared the countries with sample sizes larger than 10 (*P. vivax*: 5 countries = 10 combinations) as many countries had few samples (< 5). To identify regions with strong signatures of selection, the top 0.1% *Jost's D* values

were used to set the threshold to represent a selection signature. Genome-wide distribution of selection signatures was visualized by plotting the *Jost's D* against chromosome positions. To determine if a small number of highly differentiated STRs, such as those routinely used in capillary genotyping based STR analysis, could possibly show the population structure, we extracted the top ten most highly differentiated STRs from each pairwise comparison. The highly differentiated multilocus genotypes (MLGs) from these ten STRs were then used to construct a minimum spanning network (MSN) plot using Bruvo's distance using the R package poppr (Version 2.9.0)[75].

**Selection signatures related to drug resistance**

Based on published genetic markers, all *P. falciparum* samples are classified into different types of drug resistance in the MalariaGEN *Plasmodium falciparum* Community Project[27]. *Jost's D* was calculated per STR among the drug-resistant and drug-sensitive samples using the mmod[72] R package to explore the genetic differentiation. The top 0.1% *Jost's D* values were used to set the threshold to represent a selection signature. Considering that the malaria parasite population genetic structure varies substantially among the different populations due to different malaria control efforts, signatures of selection related to drug resistance were only performed for the comparison within subpopulations.

# Acknowledgments

This publication uses data generated by the MalariaGEN *Plasmodium falciparum* Community Project (https://www.malariagen.net/resource/26)[27] and *Plasmodium vivax* Genome Variation

project (https://www.malariagen.net/projects/p-vivax-genome-variation) as described in Pearson et al.[29] and *P. vivax* data from Hupalo et al.[28]. We thank the MalariaGEN Consortium and Hupalo et al.[28] for allowing the use of this data. This work was made possible through the Victorian State Government Operational Infrastructure Support and Australian Government National Health and Medical Research Council (NHMRC) Independent Research Institute Infrastructure Support Scheme (IRIISS). JH was supported by a Melbourne Research Scholarship (The University of Melbourne) and a WEHI PhD Scholarship. MB was supported by an NHMRC Senior Research Fellowship (1102971).

# References

1. Willems T, Zielinski D, Yuan J, Gordon A, Gymrek M, Erlich Y. Genome-wide profiling of heritable and de novo STR variations. Nature Methods. 2017;14(6):590.
2. 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.
3. Zilversmit MM, Volkman SK, DePristo MA, Wirth DF, Awadalla P, Hartl DL. Low-Complexity Regions in Plasmodium falciparum: Missing Links in the Evolution of an Extreme Genome. Molecular Biology and Evolution. 2010;27(9):2198-209.
4. McDew-White M, Li X, Nkhoma SC, Nair S, Cheeseman I, Anderson TJC. Mode and Tempo of Microsatellite Length Change in a Malaria Parasite Mutation Accumulation Experiment. Genome Biol Evol. 2019;11(7):1971-85.
5. Carlton J. The Plasmodium vivax genome sequencing project. Trends in Parasitology. 2003;19(5):227-31.
6. Carlton JM, Adams JH, Silva JC, Bidwell SL, Lorenzi H, Caler E, et al. Comparative genomics of the neglected human malaria parasite Plasmodium vivax. Nature. 2008;455(7214):757-63.
7. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860-921.
8. Subramanian S, Kumar S. Neutral substitutions occur at a faster rate in exons than in noncoding DNA in primate genomes. Genome Res. 2003;13(5):838-44.
9. Battistuzzi FU, Schneider KA, Spencer MK, Fisher D, Chaudhry S, Escalante AA. Profiles of low complexity regions in Apicomplexa. BMC Evol Biol. 2016;16:47-.
10. Davies HM, Nofal SD, McLaughlin EJ, Osborne AR. Repetitive sequences in malaria parasite proteins. FEMS Microbiol Rev. 2017;41(6):923-40.

788 11. Tan JC, Tan A, Checkley L, Honsa CM, Ferdig MT. Variable numbers of tandem repeats
789 in Plasmodium falciparum genes. J Mol Evol. 2010;71(4):268-78.
790 12. Ekland EH, Fidock DA. Advances in understanding the genetic basis of antimalarial drug
791 resistance. Curr Opin Microbiol. 2007;10(4):363-70.
792 13. Andriantsoanirina V, Khim N, Ratsimbasoa A, Witkowski B, Benedet C, Canier L, et al.
793 Plasmodium falciparum Na+/H+ exchanger (pfnhe-1) genetic polymorphism in Indian Ocean
794 malaria-endemic areas. The American journal of tropical medicine and hygiene. 2013;88(1):37-
795 42.
796 14. Anderson TJC, Haubold B, Williams JT, Estrada-Franco§ JG, Richardson L, Mollinedo
797 R, et al. Microsatellite Markers Reveal a Spectrum of Population Structures in the Malaria
798 Parasite Plasmodium falciparum. Molecular Biology and Evolution. 2000;17(10):1467-82.
799 15. Schultz L, Wapling J, Mueller I, Ntsuke PO, Senn N, Nale J, et al. Multilocus haplotypes
800 reveal variable levels of diversity and population structure of Plasmodium falciparum in Papua
801 New Guinea, a region of intense perennial transmission. Malaria Journal. 2010;9(1):336.
802 16. Figan CE, Sá JM, Mu J, Melendez-Muniz VA, Liu CH, Wellems TE. A set of
803 microsatellite markers to differentiate Plasmodium falciparum progeny of four genetic crosses.
804 Malaria Journal. 2018;17(1):60.
805 17. Manrique P, Miranda-Alban J, Alarcon-Baldeon J, Ramirez R, Carrasco-Escobar G,
806 Herrera H, et al. Microsatellite analysis reveals connectivity among geographically distant
807 transmission zones of Plasmodium vivax in the Peruvian Amazon: A critical barrier to regional
808 malaria elimination. PLOS Neglected Tropical Diseases. 2019;13(11):e0007876.
809 18. Karkar S, Alfonse LE, Grgicak CM, Lun DS. Statistical modeling of STR capillary
810 electrophoresis signal. BMC Bioinformatics. 2019;20(16):584.
811 19. Willems T, Gymrek M, Highnam G, Mittelman D, Erlich Y. The landscape of human
812 STR variation. Genome Res. 2014;24(11):1894-904.
813 20. Miles A, Iqbal Z, Vauterin P, Pearson R, Campino S, Theron M, et al. Indels, structural
814 variation, and recombination drive genomic diversity in Plasmodium falciparum. Genome Res.
815 2016;26(9):1288-99.
816 21. Hamilton WL, Claessens A, Otto TD, Kekre M, Fairhurst RM, Rayner JC, et al. Extreme
817 mutation bias and high AT content in Plasmodium falciparum. Nucleic Acids Res.
818 2017;45(4):1889-901.
819 22. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The
820 Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA
821 sequencing data. Genome Res. 2010;20(9):1297-303.
822 23. Gymrek M, Golan D, Rosset S, Erlich Y. lobSTR: A short tandem repeat profiler for
823 personal genomes. Genome Res. 2012;22(6):1154-62.
824 24. Highnam G, Franck C, Martin A, Stephens C, Puthige A, Mittelman D. Accurate human
825 microsatellite genotypes from high-throughput resequencing data using informed error profiles.
826 Nucleic Acids Res. 2013;41(1):e32.
827 25. Mousavi N, Shleizer-Burko S, Yanicky R, Gymrek M. Profiling the genome-wide
828 landscape of tandem repeat expansions. Nucleic Acids Research. 2019;47(15):e90-e.
829 26. Halman A, Oshlack A. Accuracy of short tandem repeats genotyping tools in whole
830 exome sequencing data. bioRxiv. 2020:2020.02.03.933002.
831 27. MalariaGen, Ahouidi A, Ali M, Almagro-Garcia J, Amambua-Ngwa A, Amaratunga C,
832 et al. An open dataset of Plasmodium falciparum genome variation in 7,000 worldwide samples.
833 Wellcome Open Res. 2021;6:42-.

834 28. Hupalo DN, Luo Z, Melnikov A, Sutton PL, Rogov P, Escalante A, et al. Population
835 genomics studies identify signatures of global dispersal and drug resistance in Plasmodium
836 vivax. Nature genetics. 2016;48(8):953-8.
837 29. Pearson RD, Amato R, Auburn S, Miotto O, Almagro-Garcia J, Amaratunga C, et al.
838 Genomic analysis of local variation and recent evolution in Plasmodium vivax. Nature Genetics.
839 2016;48(8):959-64.
840 30. Manske M, Miotto O, Campino S, Auburn S, Almagro-Garcia J, Maslen G, et al.
841 Analysis of Plasmodium falciparum diversity in natural infections by deep sequencing. Nature.
842 2012;487(7407):375-9.
843 31. Auburn S, Campino S, Miotto O, Djimde AA, Zongo I, Manske M, et al.
844 Characterization of within-host Plasmodium falciparum diversity using next-generation sequence
845 data. PLoS One. 2012;7(2):e32891.
846 32. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids
847 Research. 1999;27(2):573-80.
848 33. Bolton KA, Ross JP, Grice DM, Bowden NA, Holliday EG, Avery-Kiejda KA, et al.
849 STaRRRT: a table of short tandem repeats in regulatory regions of the human genome. BMC
850 Genomics. 2013;14(1):795.
851 34. Gemayel R, Vinces MD, Legendre M, Verstrepen KJ. Variable Tandem Repeats
852 Accelerate Evolution of Coding and Regulatory Sequences. Annual Review of Genetics.
853 2010;44(1):445-77.
854 35. Li Y-C, Korol AB, Fahima T, Nevo E. Microsatellites Within Genes: Structure, Function,
855 and Evolution. Molecular Biology and Evolution. 2004;21(6):991-1007.
856 36. Haasl RJ, Payseur BA. Multi-locus inference of population structure: a comparison
857 between single nucleotide polymorphisms and microsatellites. Heredity. 2011;106(1):158-71.
858 37. Fischer MC, Rellstab C, Leuzinger M, Roumet M, Gugerli F, Shimizu KK, et al.
859 Estimating genomic diversity and population differentiation – an empirical comparison of
860 microsatellite and SNP variation in Arabidopsis halleri. BMC Genomics. 2017;18(1):69.
861 38. Zimmerman SJ, Aldridge CL, Oyler-McCance SJ. An empirical comparison of
862 population genetic analyses using microsatellite and SNP data for a species of conservation
863 concern. BMC Genomics. 2020;21(1):382.
864 39. Volkman SK, Neafsey DE, Schaffner SF, Park DJ, Wirth DF. Harnessing genomics and
865 genome biology to understand malaria biology. Nature Reviews Genetics. 2012;13(5):315-28.
866 40. Amambua-Ngwa A, Amenga-Etego L, Kamau E, Amato R, Ghansah A, Golassa L, et al.
867 Major subpopulations of Plasmodium falciparum in sub-Saharan Africa. Science.
868 2019;365(6455):813-6.
869 41. Rocamora F, Zhu L, Liong KY, Dondorp A, Miotto O, Mok S, et al. Oxidative stress and
870 protein damage responses mediate artemisinin resistance in malaria parasites. PLoS Pathog.
871 2018;14(3):e1006930.
872 42. Xiong A, Prakash P, Gao X, Chew M, Tay IJJ, Woodrow CJ, et al. K13-Mediated
873 Reduced Susceptibility to Artemisinin in Plasmodium falciparum Is Overlaid on a Trait of
874 Enhanced DNA Damage Repair. Cell Rep. 2020;32(5):107996.
875 43. Amambua-Ngwa A, Tetteh KK, Manske M, Gomez-Escobar N, Stewart LB, Deerhake
876 ME, et al. Population genomic scan for candidate signatures of balancing selection to guide
877 antigen characterization in malaria parasites. PLoS Genet. 2012;8(11):e1002992.
878 44. Hamilton MJ, Lee M, Le Roch KG. The ubiquitin system: an essential component to
879 unlocking the secrets of malaria parasite biology. Mol Biosyst. 2014;10(4):715-23.

880 45. Park DJ, Lukens AK, Neafsey DE, Schaffner SF, Chang HH, Valim C, et al. Sequence-
881 based association and selection scans identify drug resistance loci in the Plasmodium falciparum
882 malaria parasite. Proc Natl Acad Sci U S A. 2012;109(32):13052-7.
883 46. Sanchez CP, Liu CH, Mayer S, Nurhasanah A, Cyrklaff M, Mu J, et al. A HECT
884 ubiquitin-protein ligase as a novel candidate gene for altered quinine and quinidine responses in
885 Plasmodium falciparum. PLoS Genet. 2014;10(5):e1004382.
886 47. Amambua-Ngwa A, Danso B, Worwui A, Ceesay S, Davies N, Jeffries D, et al.
887 Exceptionally long-range haplotypes in Plasmodium falciparum chromosome 6 maintained in an
888 endemic African population. Malar J. 2016;15(1):515.
889 48. Henden L, Lee S, Mueller I, Barry A, Bahlo M. Identity-by-descent analyses for
890 measuring population dynamics and selection in recombining pathogens. PLOS Genetics.
891 2018;14(5):e1007279.
892 49. Sawaya SM, Bagshaw AT, Buschiazzo E, Gemmell NJ. Promoter microsatellites as
893 modulators of human gene expression. Adv Exp Med Biol. 2012;769:41-54.
894 50. Anderson TJ, Su XZ, Bockarie M, Lagog M, Day KP. Twelve microsatellite markers for
895 characterization of Plasmodium falciparum from finger-prick blo{Anderson, 1999 #694}od
896 samples. Parasitology. 1999;119 ( Pt 2):113-25.
897 51. Jakubosky D, Smith EN, D'Antonio M, Jan Bonder M, Young Greenwald WW,
898 D'Antonio-Chronowska A, et al. Discovery and quality analysis of a comprehensive set of
899 structural variants and short tandem repeats. Nature Communications. 2020;11(1):2928.
900 52. Ye R, Tian Y, Huang Y, Zhang Y, Wang J, Sun X, et al. Genome-Wide Analysis of
901 Genetic Diversity in Plasmodium falciparum Isolates From China–Myanmar Border. Frontiers in
902 Genetics. 2019;10(1065).
903 53. Mobegi VA, Duffy CW, Amambua-Ngwa A, Loua KM, Laman E, Nwakanma DC, et al.
904 Genome-Wide Analysis of Selection on the Malaria Parasite Plasmodium falciparum in West
905 African Populations of Differing Infection Endemicity. Molecular Biology and Evolution.
906 2014;31(6):1490-9.
907 54. Animut A, Lindtjorn B. Use of epidemiological and entomological tools in the control
908 and elimination of malaria in Ethiopia. Malar J. 2018;17(1):26.
909 55. Lo E, Hemming-Schroeder E, Yewhalaw D, Nguyen J, Kebede E, Zemene E, et al.
910 Transmission dynamics of co-endemic Plasmodium vivax and P. falciparum in Ethiopia and
911 prevalence of antimalarial resistant genotypes. PLOS Neglected Tropical Diseases.
912 2017;11(7):e0005806.
913 56. Degefa T, Zeynudin A, Godesso A, Michael YH, Eba K, Zemene E, et al. Malaria
914 incidence and assessment of entomological indices among resettled communities in Ethiopia: a
915 longitudinal study. Malaria Journal. 2015;14(1):24.
916 57. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue
917 Expression (GTEx) project. Nature Genetics. 2013;45(6):580-5.
918 58. Bahl A, Brunk B, Crabtree J, Fraunholz MJ, Gajria B, Grant GR, et al. PlasmoDB: the
919 Plasmodium genome resource. A database integrating experimental and computational data.
920 Nucleic Acids Res. 2003;31(1):212-5.
921 59. Brouard J-S, Schenkel F, Marete A, Bissonnette N. The GATK joint genotyping
922 workflow is appropriate for calling variants in RNA-seq experiments. Journal of Animal Science
923 and Biotechnology. 2019;10(1):44.

60. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly. 2012;6(2):80-92.
61. Kyes SA, Kraemer SM, Smith JD. Antigenic variation in Plasmodium falciparum: gene organization and regulation of the var multigene family. Eukaryot Cell. 2007;6(9):1511-20.
62. Obenchain V, Lawrence M, Carey V, Gogarten S, Shannon P, Morgan M. VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants. Bioinformatics. 2014;30(14):2076-8.
63. Greenhouse B, Myrick A, Dokomajilar C, Woo JM, Carlson EJ, Rosenthal PJ, et al. Validation of microsatellite markers for use in genotyping polyclonal Plasmodium falciparum infections. The American journal of tropical medicine and hygiene. 2006;75(5):836-42.
64. Lee S BM. moimix: an R package for assessing clonality in high-throughput sequencing data. 2016.
65. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics. 2012;28(24):3326-8.
66. McInnes L, Healy J, Saul N, Großberger L. UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software. 2018;3(29).
67. Bruvo R, Michiels NK, D'Souza TG, Schulenburg H. A simple method for the calculation of microsatellite genotype distances irrespective of ploidy level. Mol Ecol. 2004;13(7):2101-6.
68. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics. 2004;20(2):289-90.
69. Yu G. Using ggtree to Visualize Data on Tree-Like Structures. Curr Protoc Bioinformatics. 2020;69(1):e96.
70. GOUDET J. hierfstat, a package for r to compute and test hierarchical F-statistics. Molecular Ecology Notes. 2005;5(1):184-6.
71. Weir BS, Cockerham CC. Estimating F-Statistics for the Analysis of Population Structure. Evolution. 1984;38(6):1358-70.
72. Jost L. G(ST) and its relatives do not measure differentiation. Mol Ecol. 2008;17(18):4015-26.
73. Gerlach G, Jueterbock A, Kraemer P, Deppermann J, Harmand P. Calculations of population differentiation based on GST and D: forget GST but not all of statistics! Mol Ecol. 2010;19(18):3845-52.
74. Dixon P. VEGAN, a package of R functions for community ecology. Journal of Vegetation Science. 2003;14(6):927-30.
75. Kamvar ZN, Tabima JF, Grünwald NJ. Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. PeerJ. 2014;2:e281.

# Supporting information

**S1 Fig. Comparison with *P. falciparum* gel electrophoresis data.** Bubble plots representing the GE allele calls are plotted against HipSTR's calls, where the left plot has been shifted such that the bottom leftmost point lies on the origin. In plot (b) points are coloured according to the region where samples originated from. The line represents y = x. It is important to note that in most of these plots, the Milne Bay samples are typically off the line y = x, indicating potential underlying issues with the GE calls for Milne Bay samples.

**S2 Fig. ROC curves and AUC values of the *P. falciparum* complete dataset and five train datasets.** (A) The mononucleotide STR model. (B) The polynucleotide STR model.

**S3 Fig. AUC values of the *P. falciparum* fivefold cross validation-datasets (80/20, 70/20, 60/20, 50/20, 40/20, 30/20, 20/20, and 10/20 splits).** (A) The mononucleotide STR model. (B) The polynucleotide STR model.

**S4 Fig. Principal component analysis of the 3,047 *P. falciparum* samples of SNP and STR data.** (A) SNP-based PCA based on 213,757 loci. (B) STR-based PCA based on 6,768 (2,563 mononucleotide STR and 4,205 polynucleotide STR) high-quality loci.

**S5 Fig. Principal component analysis of the 174 *P. vivax* samples of SNP and STR data.** (A) SNP-based PCA based on 188,571 loci. (B) STR-based PCA based on 3,496 (1,648 mononucleotide STR and 1,848 polynucleotide STR) high-quality loci.

**S6 Fig. Population structure analysis of the 174 *P. vivax* samples of SNP and STR data.** (A) UMAP clustering of the top five principal components of the SNP data. (B) UMAP clustering of the top five principal components of the STR data with different colours representing the eight different countries. (C) NJTs based on the SNP data. (D) NJTs based on the STR data. Branches are colored according to the country.

**S7 Fig. Sub-population structure analysis of the SAM and EAF population *P. falciparum* samples of SNP and STR data.** (A) UMAP on the top five principal components of the SNP data (SAM countries). (B) UMAP on the top five principal components of the STR data (SAM countries). Colouring the points by the SAM countries. (C) UMAP on the top five principal components of the SNP data (EAF countries). (D) UMAP on the top five principal components of the STR data (EAF countries). Colouring the points by the EAF countries.

**S8 Fig. A comparison of measures of genetic differentiation (*Jost's D* and $F_{ST}$) estimates using SNP and STR data of *P. falciparum*.** (A) *Jost's D* of population pairs (*Mantel r* = 0.996, *P* = 0.001). (B) *Jost's D* of country pairs (*Mantel r* = 0.996, *P* = 0.001). (C) $F_{ST}$ of population pairs (*Mantel r* = 0.97, *P* = 0.001). (D) $F_{ST}$ of country pairs (*Mantel r* = 0.90, *P* = 0.001). Mantel tests were used to measure the correlation.

**S9 Fig. A comparison of measures of genetic differentiation (*Jost's D* and $F_{ST}$) estimates using SNP and STR data of *P. vivax*.** (A) *Jost's D* of country pairs (*Mantel r* = 0.9678, *P* = 0.001). (B) $F_{ST}$ of country pairs (*Mantel r* = 0.9185, *P* = 0.001). Mantel tests were used to measure the correlation.

**S10 Fig. Pairwise genetic distance (*Jost's D*) and geographical distances (km) between populations and countries of *P. falciparum*.** (A) SNP data of populations. (B) STR data of populations. (C) SNP data of countries. (D) SNP data of countries. A Mantel test was used to measure the association.

**S11 Fig. Pairwise genetic distance (*Jost's D* and $F_{ST}$) and geographical distances (km) between countries of *P. vivax*.** (A) SNP data of country pairs (*Jost's D*). (B) STR data of country pairs (*Jost's D*). (C) SNP data of country pairs ($F_{ST}$). (D) SNP data of country pairs ($F_{ST}$). A Mantel test was used to measure the association.

**S12 Fig. Minimum spanning network using Bruvo's distances based on the ten most informative STR markers showing the relationship among two groups of *P.* vivax isolates.** (A) Cambodia and Thailand. (B) Colombia and Peru. (C) Mexico and Peru. Colors correspond to the country. Node sizes correspond to the number of samples. Edge lengths are arbitrary.

**S13 Fig. Examples of corresponding genotypes.** (A) One drug-resistant sample. (B) One drug-sensitive sample.

**S1 Table. Baseline variables used for prediction of the STR quality.**

**S2 Table. Multivariable logistic regression model's estimated coefficients and respective 95% confidence intervals.** The model was fitted on the *P. falciparum* dataset without the population origin label.

**S3 Table. Multivariable logistic regression model's estimated coefficients and respective 95% confidence intervals.** The model was fitted on the *P. vivax* dataset.

**S4 Table. Multivariable logistic regression model's estimated coefficients and respective 95% confidence intervals.** The model was fitted on the *P. vivax* dataset without the population origin label.

**S5 Table. Detected selection signatures (located in the coding region) between the *P. falciparum* populations containing the top 0.1% of STR.**

**S6 Table. Detected selection signatures (located in the coding region) between the *P. vivax* countries containing the top 0.1% of STR.**

**S7 Table. *P. falciparum* STR markers used in the analysis and the number of samples genotyped at each STR marker.** The symbol "*" denotes markers which were dropped in the TRF post-processing phase rather than the HipSTR phase. The table also highlights the location of the markers on the Pf3D7 reference genome as determined by the BLAST search.

# Author Contributions

Conceptualization: Melanie Bahlo.

Data curation: Jiru Han, Jacob E. Munro.

Formal analysis: Jiru Han.

Funding acquisition: Melanie Bahlo.

Investigation: Jiru Han, Jacob E. Munro, Anthony Kocoski, Alyssa Barry, Melanie Bahlo.

Methodology: Jiru Han, Jacob E. Munro, Melanie Bahlo.

Project administration: Melanie Bahlo.

Resources: Alyssa Barry, Melanie Bahlo.

Supervision: Melanie Bahlo.

Validation: Jiru Han.

Visualization: Jiru Han.

Writing – original draft: Jiru Han

Writing – review & editing: Jiru Han, Jacob E. Munro, Anthony Kocoski, Alyssa Barry, Melanie Bahlo.