Title:

Malaria molecular surveillance in the Peruvian Amazon with novel highly multiplexed *Plasmodium falciparum* Ampliseq assay

Authors

J.H. Kattenberg, 1
C. Fernandez-Miñope, 2,3
N.J. van Dijk, 1[†]
L. Llacsahuanga Allcca, 2
P. Guetens, 1
H.O. Valdivia, 4
J.P. Van geertruyden, 3
E. Rovira-Vallbona, 1*
P. Monsieurs, 1
C. Delgado-Ratto, 2,3
D. Gamboa, 2,5
A. Rosanas-Urgell 1

Affiliations:

1) Institute of Tropical Medicine Antwerp, Biomedical Sciences Department, Antwerp, Belgium
2) Instituto de Medicina Tropical "Alexander von Humboldt", Universidad Peruana Cayetano Heredia, Lima, Peru
3) Global Health Institute, University of Antwerp, Antwerp, Belgium
4) Department of Parasitology, U.S. Naval Medical Research Unit No 6 (NAMRU-6), Lima, Peru
5) Departamento de Ciencias Celulares y Moleculares, Facultad de Ciencias y Filosofía, Universidad Peruana Cayetano Heredia, Lima, Peru

† Current affiliation: Department of Medical Microbiology, Amsterdam University Medical Centers, Amsterdam, the Netherlands

* Current affiliation: Barcelona Institute for Global Health, ISGlobal, Barcelona, Spain

Abstract

Background

Malaria molecular surveillance has great potential to support local national malaria control programs (NMCPs) to inform policy for malaria control and elimination. Molecular markers associated with drug resistance are good predictors of treatment responses. In addition, molecular detection of deletions in *hrp2* and *hrp3* genes are indicative of potential failure of HRP2-based rapid diagnostic tests. However, there is an urgent need for feasible, cost-effective and fast molecular surveillance tools that NMCPs can implement.

Methods

Here we present a new 3-day workflow for targeted resequencing of markers in 13 resistance-associated genes, *hrp2&3*, a country-specific 28 SNP-barcode for population genetic analysis, and *ama1.* The assay was applied to control isolates and retrospective samples collected between 2003-2018 in the Loreto region (n = 254) in Peru. Pf AmpliSeq libraries were prepared using a multiplex PCR

simultaneously amplifying a high number of targets from dried blood spots and sequenced at high coverage (median 1336, range 20-43795).

Results

There was no evidence suggesting the emergence of artemisinin resistance in Peru. However, alleles in *ubp1* and *coronin* contributed to recent genetic differentiation of the parasite population. After 2008, predominant parasite lineages in Peru are resistant to sulfadoxine-pyrimethamine (sextuple *dhfr/dhps* mutant) and chloroquine (SVMNT in *crt* and NDFCDY in *mdr1*) and can escape HRP2 based RDTs.

Conclusions

These findings indicate a parasite population under drug pressure, and demonstrates the added value of molecular surveillance systems and offers a highly multiplexed surveillance tool. The targets in the assay can be easily adjusted to suit the needs of other settings.

Funding

This work was funded by the Belgium Development Cooperation (DGD) under the Framework Agreement Program between DGD and ITM (FA4 Peru, 2017-2021) and the sample collections in 2018 were supported by VLIR-UOS (project PE2018TEA470A102; University of Antwerp). Funding for the sample collections lead by the U.S. Naval Medical Research Unit 6 (NAMRU-6) in 2011 and 2012 was provided by the Armed Forces Health Surveillance Division (AFHSD) and its Global Emerging Infections Surveillance and Response (GEIS) Section (P0144_20_N6_01, 2020-2021).

Introduction

Surveillance of *Plasmodium* parasites has been identified by the World Health Organization (WHO) as one of the essential pillars to move towards malaria elimination in malaria endemic areas (1). Although the WHO has developed and updated protocols to standardize study designs to assess antimalarial therapeutic efficacy *in vivo*, this requires considerable logistical efforts to reach a minimum sample size (2). These requirements are especially challenging and costly in areas of moderate to low malaria transmission and with hard-to-reach populations at risk of malaria (3-5). Molecular parasite genotyping tools can strengthen malaria surveillance systems by monitoring the emergence and spread of drug resistance, *hrp2* and *hrp3* deletions, quantification of malaria importation risk and characterization of changing transmission intensity. As molecular surveillance platforms transition from proof-of-concept research studies to operational incorporation into NMCPs, simple standardized laboratory protocols feasible on benchtop sequencers and automated analysis pipelines are necessary to generate reproducible results and decrease the time from sample collection to results, thereby ensuring rapid turnover of up-to-date reports for decision-making and policy (6, 7).

Molecular markers associated with drug resistance are good predictors of treatment responses (3, 4, 7), and provide early warning signals of drug resistance emergence that can inform and guide treatment policy (6). A key example is the *P. falciparum* resistance against the first-line treatment component artemisinin (ART), associated with non-synonymous mutations in the gene encoding the *P. falciparum* Kelch domain protein on chromosome 13 (*K13*) in Southeast Asia (8-11). Various methods have been developed to investigate drug resistance markers, usually based on the amplification of loci of interest using PCR-based techniques, with amplicons detected in real time

assays, by amplicon sequencing (9, 12, 13), or analyzed using restriction fragment length polymorphism (RFLP) (14, 15).

Monitoring population genetic markers in addition to resistance markers, can guide malaria control by identifying the emergence and spread of new foci of drug resistance, and the distinction between imported and indigenous cases and sources of reintroduced cases in malaria-free areas (7, 16-18). However, it is still rare for these molecular tools to be used on a countrywide scale beyond research (19-21). In the past, population surveillance widely utilized genotyping tools targeting surface antigens by PCR to distinguish parasite clones (22, 23), followed by panels of microsatellites that are not under evolutionary selection pressure and, therefore, more suitable to inform population genetic changes (24-26). Despite its popularity, MS-typing is difficult to standardize, complicating comparisons between research studies (27) and is not very suitable to genotype using short-read sequencing technologies due to the repetitive nature of the alleles. More recently, genome-wide single nucleotide polymorphism (SNP) panels capable of defining a "molecular barcode" to capture the diversity of parasite populations have been developed and can be investigated with methods such as microarrays, real-time PCR, and deep sequencing (28-35). While most SNP panels have been designed to be informative on a global scale, they lack resolution to distinguish closely related clones and study subtle patterns of population dynamics on a smaller geographical scale, such as within-country, thereby limiting their informative power for national malaria control programs (NMCPs).

Deletions in the genes histidine rich proteins 2 and 3 (*hrp2* and *hrp3*), the products of which are detected by rapid diagnostic tests (RDTs), are increasingly reported throughout the world, complicating point-of-care diagnosis, and threatening the gains that have been made in malaria control (36). Many health facilities in malaria-endemic countries depend on RDTs to diagnose malaria. Failure to detect and subsequently treat the infection can result in increased disease and death. These gene deletions were first described in Peru (2) and neighbouring countries (37-40), and as a result HRP2-based RDTs are not used in several countries in South America. More recently, an increasing number of African and Asian countries have reported *hrp2* and *hrp3* deletions, which raises the threat to a new level, as these countries case management strategies rely heavily on HRP2-based RDTs for diagnosis. Recent surveys in the Horn of Africa region found that more than half of all *P. falciparum* cases are missed by HRP2-based RDTs (41). Surveys to detect these gene deletions rely on molecular methods based on PCR assays that classify deletions based on failure to amplify targets in the *hrp2* and *hrp3* exon region (and sometimes flanking genes). While this should be combined with confirmation of parasite DNA quality by amplification of single copy control genes (*e.g. msp1* and *msp2)*, this is not always adequately performed (36, 42). In addition, the sensitivity and specificity of PCR assays varies between protocols, therefore there is a need for a standardized approach.

In contrast to the continued high malaria burden in Africa, South America has made much progress with 40% reduction of malaria cases and 39% of deaths in the last decades (43), putting malaria elimination on the regional agenda. The majority of malaria cases in South America occur in the Amazon rain forest areas and are caused by *Plasmodium vivax* (76%) and *Plasmodium falciparum* (24%) (43). Several countries, including Peru, are first targeting *P. falciparum* elimination due to its lower case load, but also due to fear of emerging antimalarial resistance. Historically, South America has been a hotspot for the evolution of resistance against chloroquine (CQ) and antifolates (44) that resulted in the switch to artemisinin and derivatives (collectively referred to as ART) used in combination therapies. Mutations in K13 associated with ART resistance in South America have

independently emerged in Guyana in 2010 (45, 46), while, Suriname has reported delayed parasite clearance after ART treatment (47, 48) without the presence of validated *K13* mutations (45, 48), suggesting that an alternative *K13* independent resistance mechanism might be involved (49-53). As emerging ART resistance is a threat to the whole region, there is an increasing sense of the importance of monitoring the efficacy of ART and partner drugs in artemisinin combination therapy (ACT).

Peru has been on track to reach the goal of reducing case incidence by ≥40% by 2020. Despite this significant progress, the proportion of *P. falciparum* cases had increased from 7.8% in 2010 (lowest in the past decade) to 23.9% in 2017 (highest in the past decade). Peru has a history of drug resistance, with high levels of *P. falciparum* CQ and sulfadoxine-pyrimethamine (SP) resistance (54-57), which led to the introduction of ACT in 2001 (Figure 1) as one of the first countries in the Americas (57-59). Artesunate + mefloquine (AS+MQ) was used as first-line treatment in the Amazon region (54) and AS+SP on the North Coast. In 2015, AS+MQ+PQ became the recommended first-line treatment in the whole country (60). Efficacy of AS+MQ in 2006 was 99% (61); however, ACT efficacy against *P. falciparum* and drug-resistant marker prevalence have not been evaluated in Peru since 2009. Monitoring efficacy of the antimalarial treatment policy and molecular surveillance of resistance markers are not currently part of the NMCP activities.

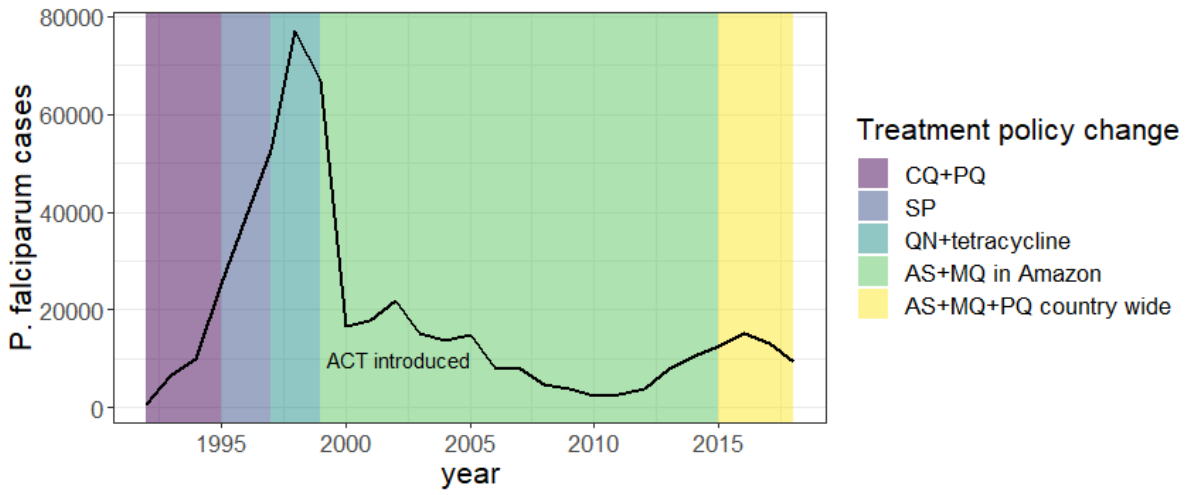


Figure 1 – Annual *P. falciparum* cases in Peru (black line) from 1992-2018, with first-line treatment policy changes (source of data: Centro Nacional de Epidemiología, Prevención y Control de Enfermedades (2020) Sistema de Atención de Solicitudes de Acceso a la Información Pública Vía internet del Ministerio de Salud. Available at: http://www.minsa.gob.pe/portada/transparencia/solicitud/). Artemisinin combination therapies were first introduced in the Amazon region in Peru in 2001, contributing to a dramatic decline in cases. Artesunate mefloquine with primaquine (AS + MQ +PQ) has been the recommended first-line treatment in the whole country since 2015. CQ + PQ = chloroquine + primaquine; SP = sulfadoxine-pyrimethamine; QN = quinine.

In order to fill this gap, we have developed a targeted amplicon Next Generation Sequencing (NGS) assay for molecular surveillance of *P. falciparum* parasites. The assay uses AmpliSeq deep sequencing technology (62), and combines (i) resistance markers to evaluate drug resistance, (ii) a 28-SNP barcode with in-country resolution to investigate parasite gene flow, and (iii) *hrp2*/3 deletion detection. Noteworthy, the seq-based tool can be performed on routinely collected filter paper blood spots, and can be easily adapted to different regions to investigate either regional trends or in-country epidemiological changes. The technology applies a multiplex PCR approach to amplify all genomic regions of interest in a rapid and easily standardizable procedure, and allows for simultaneous

amplification of a high number of targets at once, and therefore has great potential for implementation into routine surveillance practice by national malaria control programs (NMCPs). Because amplicons are allowed to overlap, coverage of long target regions (large genes for example) is possible. In this study, we have applied the Pf AmpliSeq assay to a case study of samples from Peru for assay validation and to investigate resistant markers and changes in the *P. falciparum* parasite population between 2003 and 2018. We have observed a marked change in the parasite population between 2005 and 2008, with the predominance of multiple drug resistant haplotype and *hrp2/3* deletions.

Results

Pf AmpliSeq design

A panel of *P. falciparum* drug resistant makers was selected including validated mutations for drug resistance (3, 63) and markers from literature relevant to Peru, resulting in the inclusion of 13 drug resistance associated genes (or regions of genes) (Table 1) and corresponding variants of interest (Supplementary file 1) of mutations that were (putatively) reported to have an association with resistance. Markers and genes were selected based on the following criteria: i) previously reported in the country, and ii) associated to resistance to antimalarial drugs relevant for Peru, i.e. historically used drugs, currently used drugs, or potentially alternative ACT treatments.

Table 1. Genes of interest for drug resistance molecular surveillance of malaria in Peru.

Gene ID	Chromosome	Gene with resistance-associated SNP(s)	Drug associated with resistance*
PF3D7_1218300	12	*ap2-mu*	ART (51, 64-66); QN (64)
PF3D7_1251200	12	*coronin*	DHA (49, 50)
PF3D7_0709000	7	*crt*	CQ (67-71); PPQ (72-76); AMQ (77)
mal_mito_3	mitochondrial	*cytochrome B*	ATQ (78-80)
PF3D7_0417200	4	*dhfr*	PYR (81-86); PG (87)
PF3D7_0810800	8	*dhps*	SULF (81-84, 86)
PF3D7_1362500	13	*exonuclease*	PPQ (88)
PF3D7_1343700	13	*k13*	ART (8, 45)
PF3D7_0523000	5	*mdr1*	CQ (89, 90); PPQ (91); MQ (92), QN, HF, ART (81, 84, 91-94); AMQ (77, 91); LF (94, 95)
PF3D7_0112200	1	*mrp1*	ART, MQ, LF (91, 96)
PF3D7_1408000	14	*plasmepsin II*	PPQ (88, 97, 98)
PF3D7_API04900	apicoplast	*23s rRNA*	CM (99)
PF3D7_0104300	1	*ubp-1*	ART (51, 64, 65, 100)

*AMQ: amodiaquine; ART: artesunate; ATQ: atovaquone; CM: clindamycin; CQ: chloroquine; DHA: dihydroartemisinin; HF: halofantrine; LF: lumefantrine; MQ: mefloquine; PG: proguanil; PPQ: piperaquine; PYR: pyrimethamine; SULF: sulfadoxine; QN: quinine. *The list of drugs to which resistance has been observed is non-exhaustive. Validated mutations are listed in the WHO Report on antimalarial drug efficacy, resistance and response: 10 years of surveillance (2010–2019) (63).*

Pfhrp2 and *pfhrp3* genes were included in the panel, and we targeted *pfama1* and *cpmp* (101) and MS markers *poly-alpha, ARAII, TA81* and *PfPK2* (24) as these have been frequently used markers for population surveillance in past studies. A barcode of 28 biallelic SNPs was designed for surveillance of gene flow with in-country resolution in Peru. Target regions (whole genes for drug resistance genes, and *hrp2/3;* SNP positions for barcode; variable regions for *ama1*, *cpmp* and MS markers) were submitted for assay design to Illumina resulting in 233 amplicons (Figure 2) in 2 primer pools (of 117 and 116 amplicons respectively), with amplicon length varying from 55 – 323 bp (Supplementary file

2). The final design included all targeted regions, except the *cpmp* variable region which could not be covered by the amplicon design. Overall, only 12.4% of the desired target region could not be targeted, missing one variant of interest in the *crt* gene (I218F).

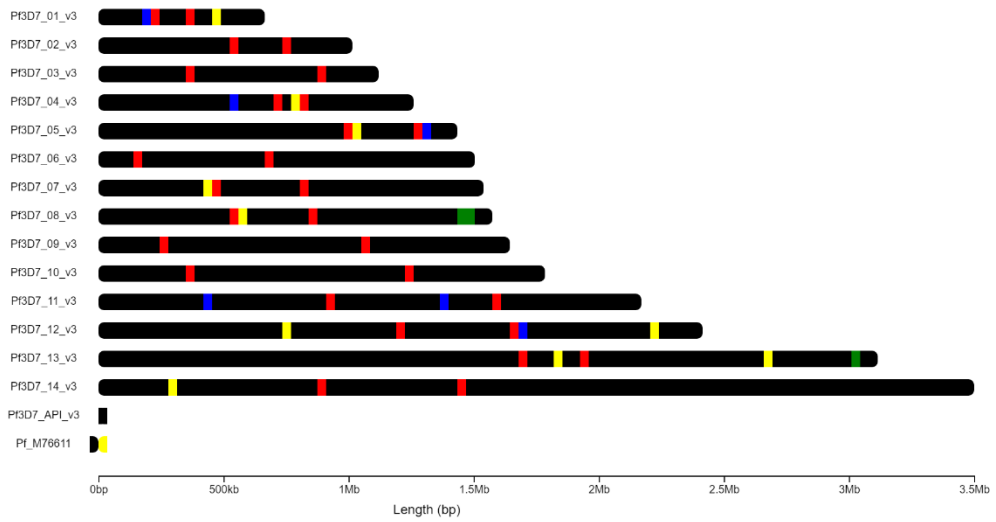


Figure 2 – Chromosomal position of amplicons in the Pf AmpliSeq design for Peru. Amplicons are depicted on the 14 nuclear chromosomes and apicoplast and mitochondrial DNA of the Pf 3D7 reference genome. Amplicons targeting drug resistance associated genes are colored in yellow, amplicons targeting SNP Barcode position in red, *hrp2* and *hrp3* in green and microsatellites, *ama1* and *cpmp* in blue. In the interactive figure (online version), the position and number of amplicons per target region can be visualized.

Pf AmpliSeq assay sequencing performance

The assay was applied to laboratory strains (n = 5), previously genotyped isolates from Vietnam (n = 6), and *P. falciparum* positive samples (n = 312) from multiple previous studies in Peru for assay validation. The mean number of reads per sample (including samples and controls with more than 50% of loci with genotype calls) after trimming was 217037 ± 490478, with a median of 99.6% [range 3.9-99.9%] of trimmed reads aligning to the reference genome. Mean coverage was 1336 ± 3627 [range 20-43795] and was not associated with parasite density. Due to the high AT-richness of the *P. falciparum* genome, nucleotide diversity in the library is relatively low (median 30% GC [range 24-54]). This can have an impact on sequencing quality because the MiSeq instrument can have trouble identifying the location of the clusters and make quality bases calls when all clusters provide a signal primarily in one or two channels. Nucleotide diversity in the run can be improved by adding PhiX spike-in during the sequencing run, at the expense of reducing coverage (geom. mean 110 ± 6.9 *vs.* 549 ± 4.2, p= 2e-16, t-test of log-transformed coverage). However, 20% spike-in did not significantly reduce the proportion of reads that are trimmed (9.7% ± 10.6 *vs.* 10.3% ± 7.3).

Depth of coverage

The Pf AmpliSeq assay is a deep sequencing assay targeting a high number of target regions with high coverage for accurate genotyping. During data processing, low quality or low likelihood genotype calls are filtered out, resulting in only high quality genotypes past filter for final analysis. Depth of coverage (DP) past filter per amplicon was relatively consistent at median 82.9 [IQR 42.9 – 116.1]

(Supplementary figure S1). Fifteen (6.4%) amplicons showed poor depth (<10) in most study samples and controls (>70% of included samples & controls; n = 267; Supplementary table S1). Six (2.6%) amplicons had a higher median depth (>150) than average (Supplementary table S2).

As the depth of coverage was calculated based on variant genotypes only, for amplicons where we detected no variants (*i.e.* only reference alleles) in all samples and controls (n = 391), depth of coverage was not determined. This was the case for 23/233 amplicons (Supplementary table S3), despite observed good amplification and sequencing for these amplicons (as evidenced from the alignment bam files in IGV browser). All 23 amplicons were amplicons in relatively conserved regions of drug resistance associated genes (Supplementary table S3), and the only variant of interest in these regions is the A220S mutation in the *crt* gene associated with PPQ resistance in South East Asia.

Negative controls

In the 5 negative controls (DNA from uninfected human blood) included in the MiSeq runs, the majority (n=4) contained only reads that did not align to the 3D7 genome in the target region. To investigate the source of the reads FastQ screen analysis revealed that most reads were human sequences and unspecific plasmodium sequences (multiple hits with different Plasmodium species). Less than 4% of reads aligned specifically to the *P. falciparum* genome (<0.5% of loci), all with reference genotypes, so no alternate alleles were detected in these controls. This indicated that cross contamination between wells is limited with little to no effect on identified mutations. One negative control sample was contaminated during the laboratory procedures or mislabeled, with 20% of reads aligning to *P. falciparum* with genotypes similar to the Vietnamese control samples; other negative controls in the same run were not contaminated.

Parasite density limit and selective whole genome amplification

To determine the lower parasite density limit of the assay, a serial dilution of 3D7 extracted DNA was prepared, mimicking concentrations that are generally eluted from a DBS (parasite density ranging from 6000 to 6 p/µl). While in all dilutions amplification was observed and sequences retrieved, at the lowest density (6 p/µl), the depth of coverage was only 10, and 60% of variant loci were not detected. The depth and proportion of loci detected increased with increasing density. However, at high parasite densities it is important not to exceed the maximum recommended DNA starting concentration (100ng). Selective whole genome amplification (sWGA) of *P. falciparum* DNA applied prior to library preparation improved the number of high quality reads at parasite densities below 60 p/µl by reducing the amount of reads trimmed (Supplementary figure S2). At 60 p/µl, 10-fold more reads were generated in samples with sWGA, while at 6 p/µl this increased to 90-fold. Selective WGA did not improve the amount of reads at densities of >60 p/µl and resulted in a higher proportion of low quality reads being trimmed than in the same samples with equal input density but without sWGA (Supplementary figure S2).

Error rate and reproducibility

To determine the sequencing accuracy of the assay, error rates were calculated using variant positions and comparing replicates of 3D7 controls (all parasite densities) in the different runs. Overall, error rates in 3D7 control samples were lower in bi-allelic SNPs (mean 0.008% ± 0.004) compared to indels (mean 0.02% ± 0.005) (Supplementary table S4). In the 3D7 serial dilution (60000-6 p/µl) without and with sWGA, the error-rate increased from 0.05% ± 0.01 in samples without sWGA to 0.13% ± 0.06 in samples with sWGA (Supplementary table S4). Hence, at parasite densities below approximately 6 p/µl sWGA improves the number of reads that passed filter per sample, but it is not recommended at

higher parasite densities, as sufficient coverage can be achieved with fewer sequencing errors without sWGA. In the run with PhiX spike-in, the error rate as calculated by the Miseq reporter was 3.6%.

Sample replicates (from Peru and Vietnam) were included in different runs to determine the reproducibility of the assay. A median difference of four bi-allelic SNPs (0.6% of all called biallelic SNP loci or 0.007% of all 57445 bases targeted in the assay) between replicates was observed. When indels and multiallelic variants were also included, a median difference of 50 alleles (2.8% of variant positions, or 0.09% over targeted bases) was observed. Genotyping known variants in previously genotyped controls (genotypes from (72, 91, 102, 103)) and samples from Vietnam (104) (n = 19, incl. replicates) resulted in accurate genotyping in 97.6% of genotypes, and detection of previously undetected additional variants or heterozygous calls in 7.9% (Supplementary table S5). For the Vietnamese control samples, it is possible that the samples contained multiple clones with different genotypes at these positions but were previously not detected due to lower sequencing depth.

Complexity of Infection

The limit of detecting minority clones in the Pf AmpliSeq assay was tested by preparing mixtures of 3D7 and Dd2 strains at various ratios at DNA concentrations mimicking a DBS sample. In the equal mixture of both strains, 94.0% (63/67) of variant loci of the Dd2 strain were detected, and when comprising 20% of the mixture, the minority clone (Dd2) was detected at 43.8% (28/64) of variant loci. However, below 5% of the mixture, very few Dd2-genotypes were detected (Supplementary table S6). Depth of coverage of these mock samples was lower (median 69.2 [IQR 18-83]) than with most study samples as input concentration was relatively low, and detection limit of minority clones is expected to be dependent on the sequencing depth.

The complexity of infection (COI) is a molecular measure estimating the number of genetically distinct clones in an infection, which can be used as a proxy for transmission intensity (18, 33). Traditionally microsatellite (MS) markers with length polymorphic alleles estimate the number of clones in a single infection. In most NGS pipelines, variants are called using diploid variant files allowing no more than two variants at a given position, therefore algorithms have been developed to estimate the complexity of infection (COI) from biallelic SNP variants. We applied different analysis methods to determine COI (*i.e.* number of clones) based on the sequencing reads, but obtained varying results with different numbers of SNPs. The McCOIL algorithms (categorical and proportional methods) were applied to different subsets of biallelic SNPs detected in the Pf AmpliSeq assay: 1) all biallelic variants; 2) core variants (*i.e.* all variants excluding *hrp2*, MS regions and mitochondrial and apicoplast variants); and 3) only the 28 barcode SNPs. The highest number of clones estimated with either the McCOIL categorical or proportional algorithm (categorial uses diploid genotype calls, proportional uses allele depths) was two clones (COI =2). However, a considerably larger proportion of single clone infections was predicted with the categorical method, especially when using more than 28 SNPs (Supplementary table S7).

Because of the large differences observed between the McCOIL methods, we estimated the proportions of single and multiple clone infections with an additional method based on the number of heterozygous genotypes in 1) the 28-SNP barcode, 2) *ama1* region, 3) core variants, and 4) MS markers. Estimates of single clone infections using the heterozygous loci in the 28-SNP barcode and *ama1* were similar to the 28 SNP proportional McCOIL method (83.9%, 85.0% and 83.1% single clone infections, respectively), but Cohen's Kappa agreement was weak (*i.e.* <0.4; poor agreement beyond chance) (Supplementary table S7). With the MS alleles determined from the short-read sequences, a

much larger proportion of multiple clone infections was estimated (61.1%), but there was no agreement beyond chance with the McCOIL proportional methods that also estimated high proportion multiple clone infections. Using the mode (most frequent value) of the different approaches, the proportion of multiple clone infections was higher in 2008-2012 and 2014-2018 than in 2003-2005 ($p = 0.0005$, X^2) (Supplementary figure S3).

Barcode performance

All 28 SNPs in the barcode were successfully amplified in the assay. Median depth (DP) was above 30 for 26/28 amplicons and >50 in 22/28 of them. Both reference and alternate alleles in the barcode were genotyped in controls and study samples (Supplementary table S8), except for SNP Pf3D7_12_1127001 that was genotyped as an indel (ref T *vs.* alt TA) on chromosome 12, position 1127000. The 28-SNPs in the barcode were correctly genotyped in all 3D7 control replicates without sWGA, while in 2/5 3D7 control replicates with sWGA (at densities of 6000 and 600 p/µl) incorrect genotypes were called.

In the study samples (n = 254), minor allele frequencies (MAF) of the 28-SNP barcode positions were between 0.35-0.55 at 7 loci and <0.1 at 6 loci (Supplementary table S8). The minor allele at position 921893 on chromosome 5 (Pf3D7_05_921893) was not observed in the study samples, but detected in controls, *e.g.* Dd2 strain.

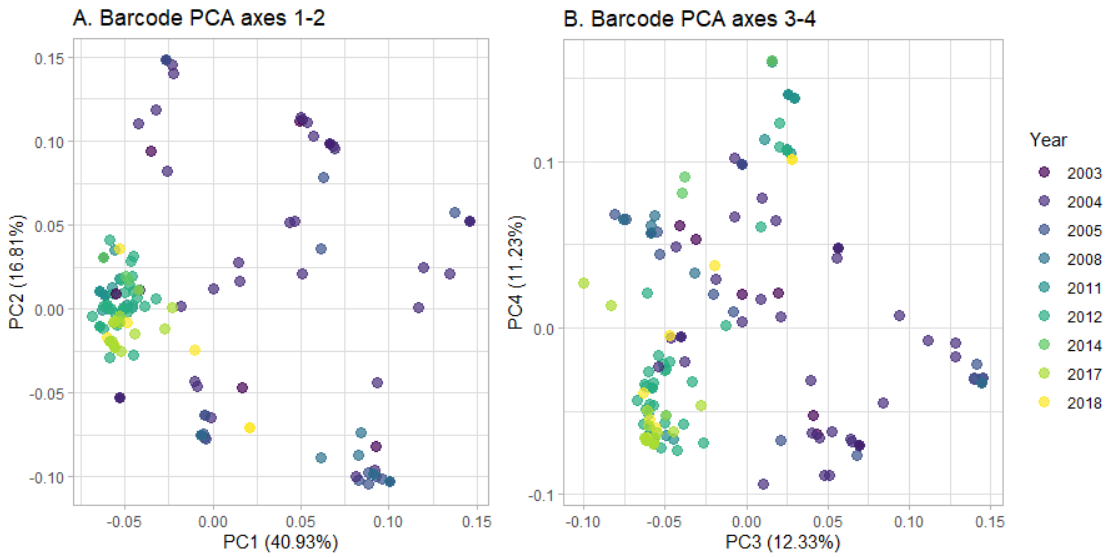


Figure 3 – **Principal component analysis (PCA) of samples collected in Peru between 2003 and 2018**. PCA is shown for 28-SNP Barcode loci along the first 2 principal components (A) and 3rd and 4th PCs (B). Isolates from earlier years (blue and purple colors) are more diverse than later isolates (greens & yellows), which cluster closer together.

Genetic diversity & Population structure

We explored population structure with principal component analysis (PCA) on 1) the 28-SNP barcode positions (Figure 3) and 2) all biallelic variants in the core region (Supplementary figure S4). Clustering was observed by year, rather than collection site. Older study samples (purple-bluish colors, ≤2005) displayed higher variability than more recent samples (yellow-greenish colors ≥2008) in the first four axes of the PCA (Figure 3), even though the majority of early samples was collected in a single district (San Juan Bautista) and later samples originated from multiple districts. Therefore, for subsequent analyses the samples were divided in three time periods: 1) early time period from 2003-2005, 2)

intermediate time period from 2008-2012 where the population is changing, and 3) most recent period 2014-2018 where little variability is observed.

Genetic diversity measured as expected heterozygosity (*He*) at the barcode positions was lower in the period of 2008-2012 (p=0.003) and 2014-2018 (p=0.00004) than in 2003-2005 (Figure 4A). Likewise, genetic diversity in the San Juan Bautista district, the only district sampled at all three time-periods, decreased slightly from 2003-2005 to 2008-2012, and was more pronounced in 2014-2018 (p >0.001, Supplementary figure S5 and table S9). These results indicate that the overall decrease in diversity observed over time is not due to unequal sampling of districts through time. Genetic diversity was very low (median *He* =0) in the far North (Pastaza) that made up 83% of samples from the 2008-2012 period and in Mazan that made up 34% of samples from 2014-2018 (Supplementary figure S5).

Observed heterozygosity was lower than the expected heterozygosity at all time-periods (Wilcoxon signed rank test, p<0.005) indicating that genetic diversity is lower than expected from a population in Hardy-Weinberg equilibrium, possibly due to a small total population size, *e.g.* due to a bottleneck or population sub-structure.

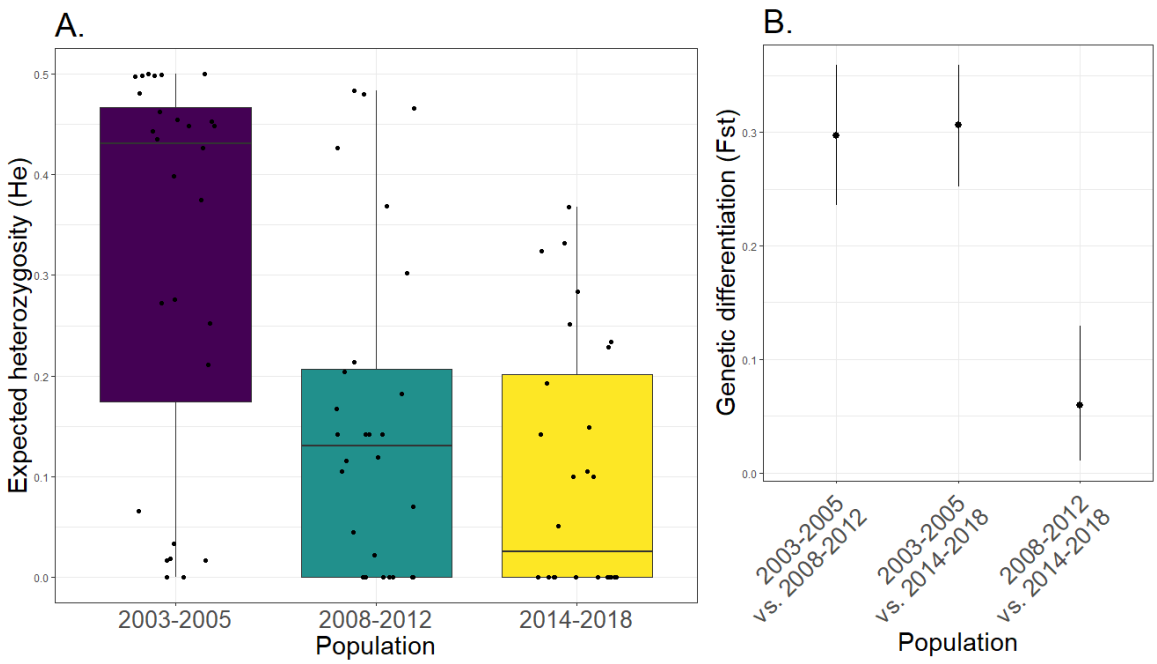


Figure 4 - Population genetic statistics. A) Genetic diversity (expected heterozygosity, *He*) for each barcode locus (black dots) grouped by time-period. Boxplots show a decrease in the median [interquartile range] of *He* after the first time period (2003-2005). B) Genetic differentiation between the three populations measured as Fst (Weir & Cockerham, 1984) using the R package diveRsity. Number of individuals for each population: $n_{2003-2005}$ = 118; $n_{2008-2012}$ = 65; $n_{2014-2018}$ = 38. In Supplementary figure S6 the differentiation measures $G`_{ST}$ (Hedrick, 2005) and Jost's D (Jost 2008) are also plotted.

Genetic differentiation was very great (F_{ST}>0.25) between 2003-2005 *vs*. 2008-2012 and 2014-2018, and moderate (0.05<F_{ST}<0.15) between 2008-2012 *vs*. 2014-2018 (Figure 4B). Within San Juan Bautista district very great genetic differentiation (F_{ST} = 0.31 and F_{ST} = 0.48, respectively) was observed between 2003-2005 *vs.* 2014-2018, and 2008-2012 *vs.* 2014-2018. Together these results indicate that the *P. falciparum* parasite population in Peru has changed considerably in the past decade, and especially in most recent years. Between 2014-2018, a very great genetic differentiation (F_{ST}>0.33) was observed

between San Juan Bautista and Mazan in Central/East Loreto surrounding the capital city of Iquitos, and Pastaza in North Loreto, close to the border with Ecuador, indicating population substructure.

Significant linkage disequilibrium (LD) was observed in the three time periods (p=0.001, Table 2), while an increasing number of barcode alleles became fixed in the populations through time, *i.e.* two fixed alleles in 2003-2005, eight in 2008-2012 and fourteen in 2014-2018 (Supplementary table S10). LD was not caused by population substructure as significant LD remained at the district level (Supplementary table S11).

Table 2. Linkage disequilibrium expressed as $\bar{r}D$ per population, measured with 999 resamplings using the poppr package in R.

population	n	$\bar{r}D$	p-value
all samples	221	0.155	0.001
2003-2005	118	0.149	0.001
2008-2012	65	0.250	0.001
2014-2018	38	0.299	0.001

Discriminant analysis of principal components (DAPC), which is similar to PCA but emphasizes differences between predefined populations, was used to further explore population structure in spatial-temporal populations (defined based on the three time periods and district) using all biallelic variants in the core region. We observed a clear distinction between samples collected before 2008 and those afterwards (x-axis, PC1 Figure 5), and a subsequent drift from 2014 to more recent years (along the y-axis, PC2 Figure 5). Using DAPC, the alleles that contribute most to the observed differentiation can be determined along with the principal components (Supplementary table S12). The highest contributors to the differentiation between 2003-2005 and later populations (2008-2012 and 2014-2018) were mutations C50R and N51I in *dhfr*, D1246Y in *mdr1*, and K540E in *dhps,* implicating selection on drug resistance-associated markers. Differentiation along the second axis, (more recent years (2014-2018) and Pastaza), are characterized by variants in seven barcode SNPs and in *dhfr*, *mdr1*, *coronin* and *ubp1* (Supplementary table S12).

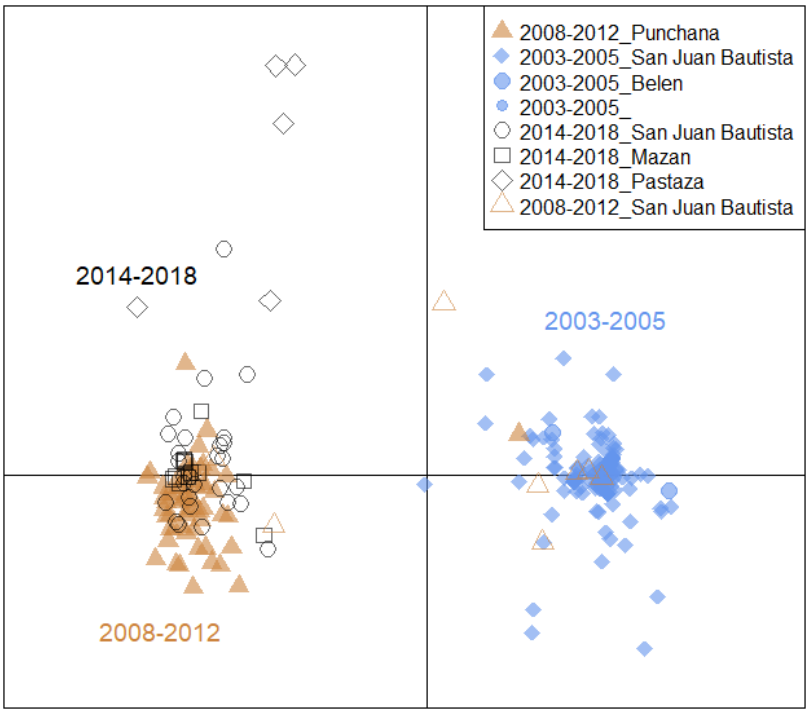


Figure 5 - Discriminant analysis of principal components (DAPC). Scatter plot of DA eigenvalues 1 and 2 using all biallelic SNPs in the core targeted region (excl. amplicons targeting repetitive regions and non-nuclear targets), showing a differentiation between pre-2005 and post-2008 isolates along the 1st PC (x-axis) and a subsequent drift to later years along the 2nd PC (y-axis). SNPs contributing most to the DAPC are listed in supplementary table S12. Samples from 2003-2005 are displayed in blue, samples from 2008-2012 in brown and from 2014-2018 in black.

hrp2 and *hrp3* gene deletions

HRP2 is an antigen targeted by RDTs, and deletions in the *hrp2* and *hrp3* genes causing these RDTs to fail have been reported at high levels in Peru (2). The presence or absence of the *hrp2* and *hrp3* genes was determined from Pf AmpliSeq data using the read depth for the amplicons for the respective genes compared to the mean depth of all amplicons in each sample. Compared to PCR classification (n = 10), all samples were correctly classified for *hrp3,* and six of these for both *hrp2* and *hrp3*, while four samples remained undefined for *hrp2* (Supplementary table S13).

Correctly classifying *hrp2* gene deletions in the study samples (n = 254) was more difficult when the *hrp3* gene was present. Therefore, while *hrp3* presence *vs.* absence could be determined for most samples (96.5%), 26.8% of *hrp3*+ samples could not be classified for *hrp2*. The proportion of individuals with both genes deleted increased over the years, from 24.8% in 2003-2005 to 42.6% in 2008-2012 up to 68.9% in 2014-2018 (Table 3). However, the proportion of samples with missing classification for one or both genes was highest (45.4%) in 2003-2005, which may be underestimating the proportion of deletions. Only few *hrp3* deletions were observed in Pastaza (1/5,20%) in 2014-2018, but high frequency of *hrp2* deletions were observed in Punchana (59/60, 98.3%) in 2008-2012 and Mazan (12/12, 100%) in 2014-2018.

Table 3. Frequency of *hrp2* and *hrp3* deletions.

	hrp2+/ hrp3-	*hrp2- /hrp3-*	*hrp2- /hrp3+*	*hrp2* NA/ *hrp3+*	*hrp2* NA/ *hrp3-*	*hrp2-/ hrp3* NA	*hrp2* NA/ *hrp3* NA
2003-2005	14 (9.9%)	35 (24.8%)	28 (19.9%)	15 (10.6%)	44 (31.2%)	2 (1.4%)	3 (2.1%)
2008-2012	1 (1.5%)	29 (42.6%)	34 (50.0%)	1 (1.5%)	1 (1.5%)	1 (1.5%)	1 (1.5%)
2014-2018	0	31 (68.9%)	9 (20.0%)	3 (6.7%)	0	2 (4.4%)	0
All	15 (5.9%)	95 (37.4%)	71 (28.0%)	19 (7.5%)	45 (17.7%)	5 (2.0%)	4 (1.6%)

NA = not applicable; deletion could not be reliably determined.

Artemisinin resistance-associated genes

The full length *K13* gene was amplified in the Pf AmpliSeq assay with 12 overlapping amplicons. While we did not observe validated SNP mutations in *K13* (F446I, N458Y, M476I, Y493H, R539T, I543T, P553L, R561H, P574L and C580Y) in the study samples (n = 254), we observed the mutation G449C in two samples (both mixed infections; Table 4). Other non-validated variants in the *K13* propellor region (V581I, I448K, V445A, premature stop codon at 613) were observed at low frequency (1-3 isolates), and only in mixed infections (*i.e.* as heterozygous genotypes). Outside of the propellor region, the K189T mutation was observed at a relatively high frequency in the three time periods at relatively high frequency (range 63% - 84%; Table 4).

Mutations in *ubp1* have been associated with *in vivo* ACT treatment failure, and *in vitro* ART resistance (100), and a change in *ubp1* haplotypes was reported paralleling the emergence of C580Y in *K13* in Thailand (105). The *ubp1* gene was covered for 91.8% with 44 amplicons, but missing regions did not contain variants of interest. Although we did not detect any previously reported variants (incl. R3138H), we observed a change in the predominant *ubp1* haplotype over time. In 2003-2005, a haplotype with Q107L, commonly combined with a mutation in K1193T, was most frequently observed, followed by wild type at all positions and R1133S single haplotype (Table 4). In later years (2008-2012 and 2014-2018), these variants were replaced by a predominant R1133S variant, most often in combination with E1011K. These two variants were frequently observed in an haplotype with four other mutations: K764N, K774N, D777G and a large insert, EQKY between amino acid positions 2826/2827. The Q107L/K1193T and R1133S/E1011K types were only observed together in mixed infections, suggesting these are mutually exclusive haplotypes.

Mutations in *coronin*, an actin-binding protein with structural similarity to the *K13* beta-propellor domain, have been associated with *in vitro* resistance to ART in selection and gene manipulation experiments (49, 50, 53, 106); but these variants have not yet been reported in field isolates (107), and were not observed in study samples in Peru. The full length *coronin* gene was amplified with 11 amplicons. We identified two new mutations V62M and V424I with opposite AF trends over time. The AF of the V62M mutation was 20.6% in 2003-2005, decreased over time and was absent in 2014-2018 (Table 4). The AF of mutation V424I increased from 55.4% in 2003-2005 to 80.9% in 2014-2018 (Table 4). While the V62M mutation is located in the WD-40 beta-propeller domain region (containing the validated mutations G50E, R100K and E107V) and has structural similarity to the *K13* beta-propellor

domain, the V424I mutation is located just outside this region and therefore not expected to contribute to ACT resistance.

CQ and ACT partner drug resistance associated genes

SP resistance associated genes, *dhfr* and *dhps* were each targeted with 9 amplicons, which covered all variants of interest, with 100% of the *dfhr* gene and 87% of the *dhps* gene covered. Following the discontinued use of SP in the Amazon region in Peru in 1999 due to high-level resistance, in 2003-2005 the single *dhfr* mutant (S108N) is seen at high frequency (68.8%), as well as the double mutant (S108N + N51I) at 22.7% (Table 5). The double mutant also carried secondary mutations in I164L and the "Bolivian Repeat" (BR), which is a silent 5 amino acid insertion before codon 30 only observed in South America. In this same period, predominantly wild type *dhps* was seen (≥56.7%), followed by a *dhps* silent mutant (K540K at 11.3%) and the double mutant (A437G, A581G) at 9.9% (Table 5).

Parasite resistance to SP increased during the period of 2008-2012 with the appearance of a triple mutant in *dhfr* (C50R, N51I and S108N) at 88.2% and triple mutant in *dhps* (A437G, K540E and A581G) at 70.6%. In 2014-2018, both triple-mutant parasites further increase in frequency (97.8% and 66.7%, respectively).

1

2 **Table 4. Haplotype frequencies in ART resistance associated genes.**

Years	*K13*					*ubp1*						*coronin*				
	G449C *	K189T	wt	wt incomplete	missing	Q107L and/or K1193T type	R1133S type	R1133S + E1011K type	mixed Q107L/K1193T+ R1133S	wt or wt incomplete	missing	V424I	V62M	wt	mixed	incomplete
2003-2005	1 (0.7%)	94 (66.7%)	15 (10.6%)	29 (20.6%)	2 (1.4%)	60 (42.6%)	21 (14.9%)	7 (5.0%)	3 (2.1%)	24 (17.0%)	26 (18.4%)	78 (55.3%)	29 (20.6%)	30 (21.3%)	0 (0%)	4 (2.8%)
2008-2012	0	42 (61.8%)	16 (23.5%)	10 (14.7%)	0 (0%)	6 (8.8%)	7 (10.3%)	31 (45.6%)	1 (1.5%)	1 (1.5%)	22 (32.4%)	41 (60.3%)	5 (7.4%)	20 (29.4%)	1 (1.5%)	1 (1.5%)
2014-2018	1 (2.2%)	25 (55.6%)	3 (6.7%)	14 (31.1%)	2 (4.4%)	3 (6.7%)	1 (2.2%)	23 (51.1%)	0 (0%)	0 (0%)	18 (40%)	37 (82.2%)	0 (0%)	8 (17.8%)	0 (0%)	0 (0%)

*1/2 G449C also accompanied with K189T; both cases mixed inf.

wt = wild type

3

4

1 **Table 5. Haplotype frequencies in ACT partner drug resistance associated genes, *dhfr and dhps*.**

Years	*dhfr*				*dhps*						
	single mutant (S108N)	double mutant + BR + I164L	triple mutant	incomplete	Single mutant K540K	Single mutant A437G	Double mutant	triple mutant	incomplete mutant	wt	wt incomplete
2003-2005	97 (68.8%)	32 (22.7%)	0 (0%)	12 (8.5%)	16 (11.3%)	3 (2.1%)	14 (9.9%)	2 (1.4%)	7 (5.0%)	80 (56.7%)	19 (13.5%)
2008-2012	6 (8.8%)	0 (0%)	60 (88.2%)	2 (2.9%)	0 (0%)	1 (1.5%)	0 (0%)	48 (70.6%)	13 (19.1%)	4 (5.9%)	2 (2.9%)
2014-2018	0 (0%)	0 (0%)	44 (97.8%)	1 (2.2%)	0 (0%)	0 (0%)	0 (0%)	30 (66.7%)	12 (26.7%)	2 (4.4%)	1 (2.2%)

dhfr double mutant = (N51I + S108N) *dhps* double mutant (A437G +A581G)

dhfr triple mutant = C50R + N51I + S108N *dhps* triple mutant (A437G +A581G + K540E)

2

3 **Table 6. Haplotype frequencies in CQ resistance associated genes, *crt* and *mdr1*.**

Years	*crt*				*mdr1*							
	CVMNT	SVMNT	CVMNT/SVMNT mixed	incomplete	NDFCDD	ND_CDD	NDFCDY	NGFSDD	NG_SDD	NDFSDD	mixed	incomplete
2003-2005	65 (46.1%)	49 (34.8%)	0 (0%)	27 (19.1%)	61 (43.3%)	2 (1.4%)	17 (12.1%)	43 (30.5%)	4 (2.8%)	1 (0.7%)	5 (3.5%)	8 (5.7%)
2008-2012	2 (2.9%)	60 (88.2%)	1 (1.5%)	5 (7.4%)	1 (1.5%)	0 (0%)	60 (88.2%)	5 (7.4%)	0 (0%)	0 (0%)	0 (0%)	2 (2.9%)
2014-2018	0 (0%)	32 (71.1%)	0 (0%)	12 (28.9%)	0 (0%)	0 (0%)	42 (93.3%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	3 (6.7%)

crt haplotypes positions 72-76 *mdr1* haplotypes : positions 86, 144, 184, 1034,1042, 1246

4

Mutations in the *crt* gene, associated with *P. falciparum* CQ resistance (CQR) at amino acids 72-76 (63) increased over time in the study samples (Table 6). The *crt* gene was targeted with 16 amplicons, covering 91.1% of the gene, missing only one variant of interest (I218F) associated with PPQ resistance. The proportion of CVMNT (CQR) and SVMNT (highly CQR) haplotypes in 2003-2005 was 46.1% and 34.8%, respectively. SVMNT increased to ≥88.2% in 2008-2012 and was observed in all (100%) genotyped samples in 2014-2018, indicating a fully CQR population. Amino acid positions 74 and 75 in *crt* were wildtype (wt) in all samples, as mutations at these positions are rare in South America. The mutation I356V, whose contribution to CQR remains unclear, was observed at high frequency (>93%) at all time periods.

Mutations in *mdr1* have been associated with CQ and MQ resistance (63), and this gene was targeted with 18 amplicons covering 87% of the gene and all variants of interest. The *mdr1* haplotypes NDFCDD (positions 86, 144, 184, 1034,1042, 1246) and NGFSDD were predominantly observed in 2003-2005 (34.7% and 48.6% respectively), with a low proportion of NDFCDY haplotypes (13.5%; Table 6). In 2008-2012 the NDFCDD haplotype had decreased to 7.3% while the NDFCDY haplotype had increased to 89% and further increased to 93.3% in 2014-2018.

The reference allele at position 86 (N86) observed in this study has been associated with decreased sensitivity to MQ, LF and DHA (91). However, there are conflicting reports of associations with MQ resistance and the haplotype (NDFCDY): while NDFCDY haplotype has been associated with decreased *in vitro* MQ sensitivity in Peru (81), in another study it has been associated with increased *in vitro* sensitivity to MQ and AS, and further increase of resistance in CQR genetic backgrounds (93). The *mdr1* D144G mutation has so far only been reported in Peru (annotated as D142G in (81)) and its role in antimalarial resistance is unknown, however, at this locus we see mostly the reference allele (D144) in more recent years, which, in conjunction with S1034C mutation was reported as less sensitive to MQ (81).

<u>Drug resistance evolution/parasite lineages</u>

In the current study we observe similar drug and diagnostic resistant haplotypes and their hybrids as previously reported in Peru (37, 81, 82), with predominantly single and double *dhfr* mutants, wt *dhps*, CVMNT and SVMNT *crt* haplotypes and multiple *mdr1* haplotypes in 2003 to 2005 (Figure 6). Gene deletions of *hrp2* and *hrp3* were observed alone and in combination. However, in 2008 we see the introduction of a new *dhfr* triple mutant (C50R/N51I /S108N) combined with triple *dhps* (sextuple *dhfr/dhps* mutant), *crt* SVMNT, and *mdr1* NDFCDY and both *hrp2* and *hrp3* deleted at ≥26.5% (Figure 6). Both the I164L and BR mutations are no longer observed (0%; Table 5) in all haplotypes with the triple *dhfr* mutant, suggesting the C50R mutation did not emerge from the circulating double *dhfr* mutant observed at high frequency before 2008. Consistent with previous reports from Peru and South America (44, 81), the *dhfr* mutant C59R was not observed here, and therefore the 'sextuple' mutant reported here is different from the sextuple mutant observed in parts of Africa, which have the C59R mutation instead of the C50R mutation (108).

This haplotype with sextuple *dhfr/dhps, crt* SVMNT, *mdr1* NDFCDY and both *hrp2* and *hrp3* deleted (called B_{V1} lineage) was previously reported in a few cases (3/62) from Peru in 2006 (81) and later associated with outbreaks in Tumbes (109) and Cusco (110). However, here we observe hybrids of the B_{V1} lineage with the same *dhps/dhfr/crt/mdr1* haplotype, but with the *hrp3* gene present (49 of 105 isolates, 47%), and varying in *coronin* V424I and barcode genotypes (Supplementary Figure S7).

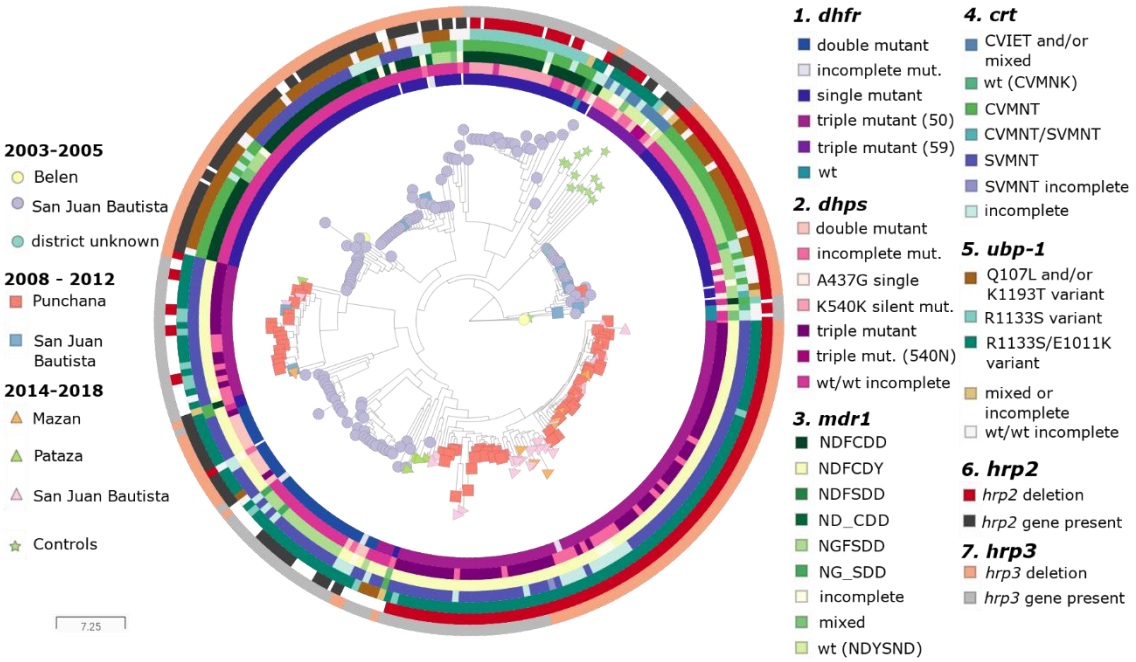


Figure 6. Neighbor-joining tree of genetic distances and colored by haplotypes of partner-drug resistance associated markers. Neighbor-joining tree using Euclidean distances between all biallelic SNPs in the core region and was rooted on a 3D7 control isolate made using ape package in R. Nodes are colored and shaped by year and district. A Metadata ring of drug and diagnostic resistance marker haplotypes. Inner ring: *dhfr* haplotypes with predominantly single and double mutants in 2003-2005 to triple mutants from 2008 onwards. 2nd ring: *dhps* haplotypes with predominantly triple mutants after 2008. 3rd ring: *mdr1* haplotypes with predominantly NDFCDY haplotype after 2008. 4th ring: *crt* haplotypes, with predominantly SVMNT haplotype after 2008. 5th ring: *ubp1* haplotypes with a high proportion of Q107L/K1193T haplotypes until 2005, and mainly R1133S/E1011K haplotypes after 2008. 6th ring: *hrp2* deletions. Outer ring: *hrp3* deletions. Interactive image and data: https://microreact.org/project/aV7RHNCBmG3sJ2rwzchE4k/279b0013

Copy number variations (CNVs) in *mdr1* and *plasmepsin II* have been associated with MQ and PPQ resistance in South East Asia (92), and have not been previously reported in Peru (91). The absence of CNVs in a random subset of 78 samples (spanning all time periods and districts) were assessed by qPCR, confirming single copies of both genes in all samples (Supplementary figure S8).

Other drug resistance associated markers were included and genotyped in the Pf AmpliSeq assay. None of the variants of interest were observed in the samples in the genes *pfmrp1* (22 amplicons covering 95.1% of the gene), *pfexonuclease* (10 amplicons covering 98.6% of the gene), *cytochrome B* (5 amplicons covering 100% of the gene) and *23S rRNA* gene (9 amplicons covering 100% of the desired region). The *ap2-mu* gene was targeted by 9 amplicons covering 99.1% of the gene, and the S160N mutant was observed at 10.6% in 2003-2005, but decreases in later periods (1.5% and 2.2%) (supplementary data for all variants of interest available at https://microreact.org/project/aV7RHNCBmG3sJ2rwzchE4k/279b0013).

Discussion

We successfully designed a new highly-multiplexed deep sequencing assay (Pf AmpliSeq Peru) targeting markers in 13 genes associated with *P. falciparum* antimalarial resistance, combined with a country-specific 28-SNP barcode for population genetic analysis in Peru and *hrp2&3* genes to detect deletions that can cause false-negative results in RDTs targeting these proteins. The assay described

here has a low error rate, high accuracy, high depth of coverage for an optimal performance to analyze blood samples collected on filter papers with *P. falciparum* parasite densities ≥60 p/µl. In *P. falciparum* samples with parasite densities < 60 p/µl, sWGA prior to the Pf AmpliSeq assay increases the number of reads and genotype calls, but also the error rate. The assay also allows for the estimation of multiplicity of infection using detected variants in the assay.

The assay was validated with a collection of retrospective samples (n = 254) collected in the Peruvian Amazon region between 2003 and 2018. While the Pf AmpliSeq assay was capable to detect ART-resistance validated SNP mutations in the propellor region of the *K13* gene (63) in control samples from South East Asia, these mutations were not detected in the Peruvian samples. However, we identified other SNP variants in the *K13* propeller domain (G449C, V581I, I448K, V445A, and a premature stop codon at 613), although at a low frequency (1-3 samples) and as minor allele in mixed clone infections. In Guyana, the C580Y mutation has independently emerged in *P. falciparum* strains, on a genetic background with the K189T mutation (46), although this mutation was common and could be a coincidence. Overall, without the detection of validated molecular markers for ART resistance, there is no evidence suggesting the emergence of ART resistance in Peru in recent years (up to 2018), since the introduction of ACTs in Peru in 1999.

Other genes in the endocytosis pathway such as *ubp1, ap-2-mu* and *coronin*, can cause decreased sensitivity to ART in *in vitro* studies and a potential importance of these genes in *in vivo* emergence of ART resistance has been suggested (51-53, 66, 111), and were amplified in the Pf AmpliSeq assay. Evidence of an African ancestry of *P. falciparum* lineages in South America (112, 113) might predict a genetic background prone to accumulate mutations in these genes (111, 114). In the samples analyzed in this study, we did not detect any of the confirmed *ubp1, coronin* and *ap-2-mu* ART resistance associated mutations (51, 64, 65, 111); however, other observed mutations in the *ubp1* and *coronin* genes contributed to the genetic differentiation of the more recent *P. falciparum* populations in Peru resulting in changes in the predominant haplotypes in these genes after 2008. We do not know the phenotypic effect of the newly discovered *ubp1* and *coronin* variants, which will have to be evaluated with *in vitro* and/or *in vivo* studies. Finally, we did not investigate the presence of genetic variants in KIC7, Eps15, Formin2 (52, 53, 66, 100, 115), associated with *in vitro* ART resistance while we were conducting this study; these genes can be included in future versions of the assay.

To our knowledge, this is the first report of a *P. falciparum* NGS amplicon assay targeting *hrp2* and *hrp3* deletions, and we present a first proof-of-principle of the ability to detect these deletions using amplicon sequencing from samples in Peru, where a high proportion of deletions in both genes that would cause HRP2-based RDTs to fail, has previously been reported (2, 37). The molecular structure of *hrp2* and *hrp3* deletions is not well characterized, and potentially differs between countries and lineages. Better characterized start and end points for the large deletions would improve amplicon design and analysis of deletions, as now some amplicons seem better at classifying deletions than others, especially in the case of *hrp2*.

By combining the phenotypic markers with the 28-SNP-barcode in the Pf AmpliSeq assay, we were able to detect not only changing allele frequencies in known resistance associated genes, but also to identify temporal changes in genetic differentiation in the parasite population under study, alongside a decrease in genetic diversity. In addition, the SNP-barcode helped to distinguish different lineages of the multiple resistant (*dhfr/dhps/crt/mdr1*) haplotypes circulating in recent years, and enables the investigation of drug resistance evolution. Multidrug resistance in South America spread from a single

origin in the lower Amazon (44). In Peru an admixed population of sensitive and resistant parasites was reported before 2006, with five distinct clonal lineages linked with specific *dhfr, crt, dhps* and *mdr1* haplotypes (81, 82) and *hrp2* haplotypes (37). Since 2006, frequent outcrossing and evolution of the parasite populations under the changing drug policy was reported in studies combining MS markers with drug resistance genes (82). With the Pf AmpliSeq, we observe the same resistant haplotypes and the evolution of the hybrids in time, with predominantly single and double *dhfr* mutants, wildtype *dhps*, CVMNT and SVMNT *crt* haplotypes and multiple *mdr1* haplotypes in 2003 to 2005 and introduction of the BV_1-lineage in 2008.

The BV_1 lineage and hybrids seem very successful, as they increasingly dominate the parasite population in Peru after 2008, which is resistant to SP (sextuple *dhfr/dhps* mutant) and CQ (SVMNT in *crt*) and can escape HRP2 based RDTs. The selective pressure for these parasite lineages is unclear, as there is no clear evidence of ART-MQ resistance (absence of *K13* variants and *mdr1* CNV) and HRP2-based RDTs are not used in common practice in Peru. The high level of CQ resistance markers (in *crt* and *mdr1)* could be the result of CQ-treatment of co-endemic *P. vivax* cases. Co-infections of both species are frequent (116), and can occur with low density *P. falciparum*, which could be misdiagnosed and treated as *P. vivax* mono-infection. Potentially, the continued CQ pressure is applying a contrasting selective pressure than the current ACT treatment (AS+MQ+PQ), which might contribute to selection of variants in *mdr1* that are more sensitive to MQ.

The strong population differentiation observed between 2005-2008, fits with the suggestion in Baldeviano et al. (109) that the BV_1 lineage was introduced in Peru. The lineage was previously reported in the Amazonas Department in Colombia in 2005 (38, 39). Similar *dhfr/dhps/crt/mdr1* haplotypes have also been reported in Suriname, with a high level of gene flow between neighbouring countries in the Guiana shield (Venezuela, Guyana, Brazil) (48), thought to be driven by mobile populations such as gold miners that have access to non-recommended treatments (116, 117) and RDTs for self-diagnosis. Predicting the origin of imported infections based on SNP barcodes is technically possible (118-120), however, for this study samples from neighboring countries were not available to fully investigate the spatial dynamics within the continent.

With the Pf AmpliSeq data, we compared several methods to determine the COI with poor agreement among them. The McCOIL categorical method (the recommended tool for a dataset with many biallelic SNPs like ours) estimated all samples and controls as single clone infections, even though we identified good quality heterozygous calls in drug resistance or barcode alleles. Therefore, we also estimated the COI using heterozygous calls in varying subsets of the genotype data (including the highly diverse *ama1* regions and MS genotypes), however, these do not allow an exact estimate of the number of clones, due to the nature of the genotype calling process from NGS data applied here. In the case of MS markers, this way we lose the great capability of these polymorphic alleles at calling multiple clones in mixtures of parasites, as can be achieved with PCR and capillary electrophoresis. Alternative haplotype-based amplicon sequencing analysis approaches have been developed (*e.g.* SeekDeep (121) and HaplotypR (101)) that might have a greater capability at unraveling complex infections (*i.e.* more than two clones) and relatedness analysis, however, applying these methods to our Pf AmpliSeq data is complicated by the overlapping amplicons in the design and requires further validation.

Despite the good performance, there are some limitations to the Pf AmpliSeq design. First of all, due to the high AT-richness of the *P. falciparum* genome, primer design is not always possible at the desired regions, and small sections of some genes are missing (total 12.4% missing of desired target

regions). Nevertheless, most of the variants of interest except one were included, resulting in a highly multiplexed assay with many important markers for the purpose of malaria molecular surveillance. Second, the assay was designed specifically for molecular surveillance of malaria in Peru, and the application of the assay in its current form is limited to South America. Third, several SNPs in the barcode (14/28) became fixed in the most recent parasite population analyzed, limiting the usefulness of these SNPs in future years. Of note, the sample size in the 2014-2018 period was smaller than earlier periods (n =45 *vs.* n = 141 & n = 68), what may have lowered the likelihood of detecting minor alleles, although with 45 samples a theoretical frequency of 0.01 with 1.5% precision and a frequency of 0.5 with 7.5% precision can be detected, therefore the sample size should be sufficient to detect a true decrease in diversity. The sample size included in this study, is comparable or greater than earlier reports from Peru (*e.g.* 220 samples across multiple sites in 1999 (82), 104 samples from 1999 and 62 samples from 2006-07 (81), 97 samples from a therapeutic efficacy trial (122). The 28-SNP barcode in the design can be replaced with other SNP-barcodes, making it suitable for implementation in other regions. The *in silico* design has good predictability for successful wet-lab & sequencing performance. Ideally, the Pf AmpliSeq assay could be combined with periodic WGS of a subset of samples to investigate the emergence of mutations in other genes that could be included in a future design. We are in process of designing a similar assay for South East Asia and Africa, which also includes the latest ART resistance markers. In addition, we are developing assays for molecular surveillance of *P. vivax,* which is the most predominant species in South America.

Compared to WGS, where the entire genome is targeted, the Pf AmpliSeq Peru assay offers a much greater sequencing depth of the targets of interest (100-1000 fold *vs*. 50-100 fold in WGS). In addition, WGS from DBS samples often requires prior whole genome amplification to reduce the amount of human DNA contamination, which reduces balanced coverage across the genome. While WGS approaches allow the identification of newly emerging alleles in resistance genes, it is less feasible for routine surveillance because of the higher cost, lower throughput on a midrange benchtop sequencer and higher computational needs. Compared to other amplicon sequencing assays (12, 35, 123) and genetic surveillance platforms (124), the Pf AmpliSeq includes by far the highest number of drug resistance associated genes, a SNP barcode, as well as diagnostics resistance markers all in one assay. Library preparation and pooling takes 2-3 days, sequencing ≥96 samples can be performed at once on a smaller benchtop sequencer, and data processing can be automated and standardized and does not require an advanced computational infrastructure. As an alternative to our pipeline, Illumina offers a program on their cloud environment where the user generates variant files from the sequencing assay (minor bioinformatics skills are required), which can be analyzed using a regular bench- or laptop computer.

Concluding remarks

The findings describe a *P. falciparum* population in Peru with increasing accumulation of drug and diagnostic resistance. These lineages seem more successful under the current treatment regimen than lineages observed prior to 2008 and have likely been introduced. Spread of resistant lineages throughout the Amazon region and risk of (re)introduction of parasites is a serious threat to *P. falciparum* elimination in Peru. Changes in haplotypes in genes in the endocytosis pathway further warrant caution for emerging resistance to artemisinins. This study highlights the importance of systematic resistance monitoring in the country, which is currently lacking, and demonstrates the added value of molecular surveillance systems for early detection of emerging resistance and offers a

surveillance tool. While the data presented here is based on a limited number of samples, and does not represent a full overview of the Peruvian *P. falciparum* population, a more systematic sampling approach of dried blood spots at many more sites within the country would be ideal for continuous surveillance. Molecular markers can be an alternative for mapping and monitoring trends of drug resistance, but implementation will require a structured approach that will most importantly not unnecessarily burden existing health systems and is sustainable for the long term. Genetic surveillance should be incorporated in national malaria strategies and within Peru a molecular surveillance network (GENMAL) has been established to stimulate wide-scale implementation and standardization of the markers used for surveillance. This network will be a platform for coordination of research in Peru, data sharing and efficient translation of results into policy advice and practical guidelines for malaria elimination in Peru. The type of multiplex assay as presented here could be more broadly applicable for molecular surveillance in the wider Amazon area and South America. Offering a standardized platform enabling easier data exchange between neighboring countries to supplement the malaria control strategies is especially important as countries near the pre-elimination phase.

Materials and methods

Study sites and samples

P. falciparum positive samples (n = 312) from multiple previous studies were selected for the Pf AmpliSeq assay analysis based on geographical representation (Figure 7) and parasite density (≥100 parasites/µl by PCR; geometric mean density 4300 ± 3.1 p/µl). With a sample size of 96 samples (~1 plate) the assay can detect a minor allele frequency of 0.01 with 0.02 (2%) precision, and with 381 samples with 0.01 (1%) precision (https://epitools.ausvet.com.au/oneproportion).

P. falciparum selected samples from 2018 were collected in Loreto region by the NMCP in Peru (Zero Malaria Programme) in collaboration with the Universidad Peruana Cayetano Heredia (UPCH) (n = 5). *P. falciparum* samples between 2003 and 2018 were collected in studies led by UPCH (N=269) ((2, 125) and ongoing studies), and the U.S. Naval Medical Research Unit 6 (NAMRU-6) (n = 63, collected between 2011 and 2012).

These studies have been reviewed and approved by local ethical review boards and/or authorities prior to sample collection. Protocols were registered in the Decentralized System of Information and Follow-up to Research (SIDISI) of the University Directorate of Research, Science and Technology for approval by UPCH Institutional Research Ethics Committee (CIEI) prior to its execution (SIDISI numbers: 52707, 61703, 101645, 66235, 102725) in Lima, Peru, and in one case a clinical trial registered at clinicaltrials.gov (NCT00373607). The NAMRU-6 study was approved by its Institutional Review Board in compliance with all applicable federal regulations governing the protection of human subjects (protocol NMRCD.2007.0004). Individuals were included in this study only if signed informed consent forms included a future-use clause. Secondary use of these samples for assay validation and molecular surveillance was approved through the Institutional Review Board of the Institute of Tropical Medicine Antwerp (reference 1417/20).

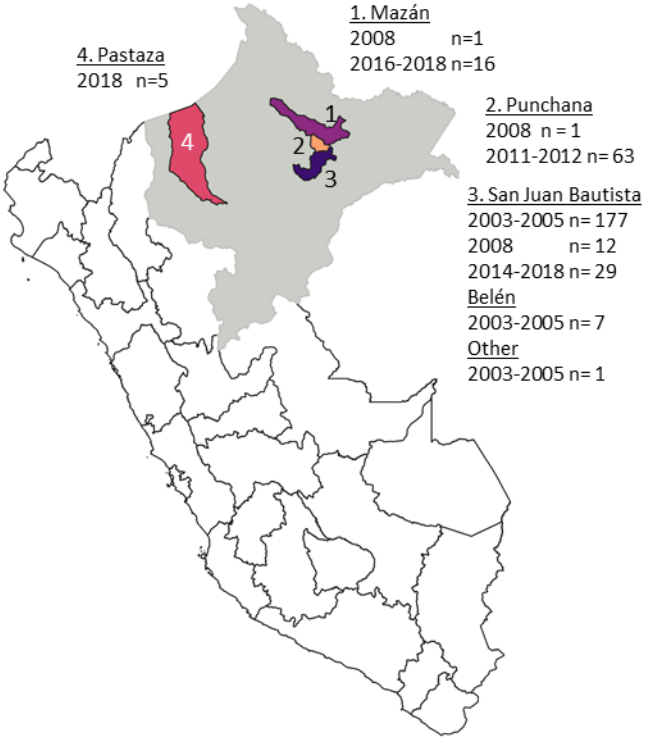


Figure 7 – Map of study sites of retrospective sample collection from the Peruvian Amazon region Loreto (grey). The majority of samples were collected in Maynas Province (areas 1, 2, and 3), which could be divided into 2 groups: samples collected in/near the San Juan Bautista district (area 3), which covers part of the urban area of Loreto's capital Iquitos and peri-urban communities south of Iquitos (n=226); samples from the Mazán district (area 1; n =17) and Punchana district (area 2; n =64) with remote communities living at forest rivers such as the Amazon River and Napo River, North of Iquitos. A final collection of samples from 2018 was available from the Northern Amazon region in Datem del Marañón Province, Pastaza District (area 4, n=5).

For the final analysis we selected 254/312 (81%) samples with good quality data (<50% of genotype calls missing and mean coverage >15) and retaining only one library in case of replicates, with the lowest proportion of missingness and highest aligned coverage.

DNA extraction and qPCR

DNA was extracted from stored dried blood spots using the E.Z.N.A Blood DNA Mini Kit (Omega Bio-tek, Georgia, USA) following manufacturer's instructions at the UPCH laboratory. Per sample, two pieces of approximately 0.5 cm^2 each were cut from the dried blood spot and DNA was eluted in 100 µL final volume. Subsequently, quantitative PCR targeting the 18s rRNA gene for the diagnosis of *P. falciparum* and *P. vivax* infection was performed as previously described (126). DNA of control samples was extracted using QIAamp DNA Mini Kit (Qiagen, Hilden, Germany) following manufacturer's instructions at the ITM Antwerp laboratory.

Control samples for Pf AmpliSeq validation

Controls in the library preparation and sequencing reactions included DNA extracted from laboratory strains (3D7, Dd2 (MRA-150), CamWT_C580Y (MRA-1251), Dd2_R539T (MRA-1255), IPC 4912 (MRA-1241). Sensitivity of the assay was tested in a ten-fold serial dilution of 3D7 parasite DNA (60.000 p/µlto 6 p/µl) mixed with human DNA from whole blood. Selective whole genome amplification (sWGA) was performed on aliquots of the 3D7 dilution series as previously described (127) to determine whether this improved sensitivity at low densities. For minority clone detection, DNA of

3D7 and Dd2 strains at similar parasite density (20,000 p/µL) were mixed at 50-50% ratio, 80-20%, 95-5%, 99-1% and 99.5-0.5% to mimic field isolates with multiple clone infections at different ratio's. A selection of samples (n = 6) with known variants including *K13*, *crt* and *mdr1* from a previous study in Vietnam (104) were also included as controls. Uninfected human DNA (extracted from whole blood) was used as negative control (at least one negative control in each run).

Pf AmpliSeq Design: SNP-barcode

A barcode of 28 biallelic SNPs was designed for molecular surveillance in Peru as described in Supplementary file 3. Briefly, MalariaGEN *Plasmodium falciparum* Community Project data (2015 and 2016 releases (128)) were used to select SNPs with a minor allele frequency between 0.35-0.5 and contribution to differentiation between countries (within the dataset) using discriminant analysis of principal components (DAPC) (129). Two SNPs per chromosome were selected with priority given to synonymous SNPs with low pairwise linkage disequilibrium.

Pf AmpliSeq assay and sequencing

Library preparation for the targeted resequencing was performed using AmpliSeq Library PLUS for Illumina kit (Illumina), AmpliSeq Custom Panels (*i.e.* designed panel primer pools) (Illumina) and AmpliSeq CD Indexes (Illumina) as per the manufacturer's instructions for two primer pools. Standard recommended input for the library preparation is 10 ng of DNA (with 1 ng as recommended minimal input). DNA concentration for controls was determined using Qubit v3 High sensitivity DNA kit (Invitrogen) and diluted with nuclease-free water, so that the maximum input volume (7.5 µl) would contain 1 ng of total DNA (both human and parasite).

Qubit results indicated a mean DNA concentration of 6.1 ± 0.3 ng/µl of eluted DNA from 25 study samples, therefore study samples were not further diluted and 7.5 µl of eluted DNA was added to the reaction mix Ampliseq HiFi Mix (from Ampliseq library preparation kit; Illumina), and split in two reactions with one custom DNA panel primer pool (Illumina) each as per manufacturer's instructions. Amplification was performed with cycling conditions adjusted for low input DNA and long amplicons (1 cycle 99°C for 2 min; 21 cycles of 99°C for 15s and 60°C for 8 min). Samples from the same time period and geographical locations were distributed across different library preparation and sequencing batches. After target amplification, the two reactions of each sample were combined on a new plate and partially digested (FuPa reagent, Illumina), followed by ligation with AmpliSeq CD Indexes, DNA ligase and Switch solution (Illumina). Next, the libraries were cleaned with AMPure XP beads (Agencourt) to remove remaining reagents and unbound indexes according to the library preparation kit procedures. Subsequently, libraries were amplified with the library Amp primers and mix provided by the kit (Illumina), followed by two additional clean-up steps using AMPure XP beads. Libraries were quantified using KAPA Library Quantification Kit for Illumina Platforms (KAPA Biosystems) according to the manufacturer's instructions. Then, libraries were diluted to 2 nM with low Tris-EDTA buffer supplied by the kit (Illumina), pooled, denatured with 0.2 N NaOH, and 18 pM of the library pool in Hybridization Buffer was loaded on a Miseq system (Illumina) for paired end sequencing using the Miseq Reagent Kit v3 (600 cycle; Illumina). 20% PhiX spike-in (Illumina) was used in one of the sequencing runs to determine the effect on sequencing quality and trade-off in coverage. A detailed protocol is available at bio-protocol.org/###.

Detection of *mdr1* and *pm2* copy number variations by PCR

CNVs in *pm2* and *mdr1*, associated respectively, with PPQ (98) and MQ (92) resistance, were determined in a subset of samples (n=78) from all time periods and districts using a previously described qPCR method using the ubiquitin-conjugated enzyme gene as the reference gene and the 3D7 strain as calibrator (104). Previously genotyped samples from Vietnam (104) were used as controls in addition to 3D7 and Dd2/MRA-150. Samples were tested in triplicate; 20% of samples was re-tested in an independent experiment to test reproducibility with 100% of results in agreement (*i.e.* single or multiple copy gene).

Read processing, alignment and variant calling

FASTQ files generated with the Miseq sequencer were processed with an in-house analysis pipeline, running on a Unix operating system desktop computer. As a first step, the FASTQ files were analyzed with FastQC for quality control of sequencing data (130), and a multi-sample QC report was generated using the R-package fastqcr. Reads were trimmed using Trimmomatic (settings: ILLUMINACLIP: 2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36) to remove adapter sequences and poor-quality reads. Trimmed reads were aligned to the 3D7 reference genome (version plasmoDB-44) using Burrows-Wheeler aligner (v0.7.17) (131). Alignment statistics were generated using Picard's CollectAlignmentSummaryMetrics. Variants were called using HaplotypeCaller of the Genome Analysis Toolkit (GATK, v4.1.2) (132) and individual sample gVCF files were used in GenomicsDBimport and GenotypeGVCFs to jointly call the genotypes of all samples and controls together. Called variants were hard filtered (QUAL score >30, overall DP >100, MQ > 50, QD>1.0, ReadPosRakSum >-10, SOR<4, GT depth >5) and variants that passed the filtering were annotated with SnpEff (v4.3T) (133). After filtering, 2146 loci were included in the vcf file. Per sample and locus filtered depth of coverage (format field DP) from the multi-sample vcf file was used for depth calculations. Median depth was calculated over each position within one amplicon or gene target (*i.e.* combination of overlapping amplicons spanning one gene), and per sample mean depth calculated for all genotyped loci for that sample.

Analysis of performance

Sequencing performance and alignment statistics were analyzed using R and RStudio. Aligned coverage was calculated as the number of bases passed filter divided by the target region size (57445 bases targeted in the Pf AmpliSeq assay design). The number of reads in the fastq before and after trimming files (from the multisample FastQC report prepared using the R-package fastqcr) were used to calculate the proportion of reads trimmed. To determine sources of contamination in negative controls, fastQ screen analysis (134) was performed. As the AmpliSeq manufacturers protocol does not recommend including a Phi-X spike in, error rates were determined using n = 10 replicates of 3D7 control DNA. One of the 3D7 replicates (with highest mean depth - normalized DP from vcf file - of 161 and only 0.1% missing loci and lowest amount (n = 4/847) of non-reference alleles) was defined as the reference sample. Allelic difference among 3D7 replicates (calculated as a pairwise matrix using the R package poppr v2.8.6. (135)) was determined compared to this reference sample. The error rate for each sample was calculated as the number of allelic differences with the reference sample over the target region (57445 bases), and the mean error rate was calculated over all 3D7 replicates (n = 10) and all 3D7 replicates with prior sWGA (n = 5). In addition, error rates were determined separately for only allelic differences in 1) biallelic variants and 2) indels, as "errors" in indels are often alignment errors rather than sequencing errors.

Analysis of Complexity of infection

COI was first estimated using the Real McCOIL categorical and proportional methods (136) in R. As the Real McCOIL was designed for use with SNP panels of around 100 SNPs, we explored the effect of using different subsets of our data to estimate COI: 1) all biallelic variants detected in the Pf AmpliSeq target regions, 2) 25 biallelic SNPs from the 28 SNP barcode (one excluded because typed as indel; two excluded because 3rd allele detected, see results section), 3) "core variants": all biallelic variants excluding repetitive regions (*hrp2*, *hrp3* and MS regions) and non-nuclear targets (mitochondrial and apicoplast). For the analysis with Real McCOIL categorical method (using genotype calls) a fixed error was used with default settings, an upper bound for COI of 10, and initial COI of 5. For the proportional method (using allele depths for each of the 2 alleles per locus), a fixed error was used with default settings, an upper bound for COI of 15, and initial COI of 7. For both methods we allowed a proportion of missing genotypes of 50% for all biallelic variants, 20% for the barcode and 40% for the "core variants".

Since different results were obtained with the two McCOIL methods, next we estimated the proportions of single and multiple clone infections with an additional method based on the number of heterozygous genotypes in 1) the 28-SNP barcode, 2) *ama1*, 3) core variants, and 4) MS markers. We included only biallelic variants as the error rate was lowest for these variants (see results). Samples with one or more heterozygous SNP were considered to contain multiple clones (COI≥2).

Then, we estimated COI from the MS amplified regions. The raw fastq files were aligned to reference sequences containing only the four MS amplicon regions (*poly-alpha, TA81, ARAII and PfPK2*) using Burrows-Wheeler aligner (v0.7.17) (131). Subsequently, reads in resulting bam files were realigned on repeats using Genotan v0.1.5 (137) and short tandem repeat (STR) length was determined using HipSTR (138). HipSTR performs de novo stutter estimation and STR calling and de novo allele generation for specified regions of short tandem repeats in the reference sequence using the sample alignments. As HipSTR (and other short tandem repeat calling algorithms) are made for diploid genomes, only the 2 most predominant MS genotypes present in the sequencing reads of a sample are determined in this way. While this does not allow us to give exact estimates of COI by MS, we can distinguish between single clone (COI =1) *vs.* multiple clone infections (COI≥2), if 2 MS alleles are found for one or multiple MS markers.

Finally, we compared all methods of determining single clone *vs.* multiple clone infection (COI by McCOIL categorical, COI by McCOIL proportional, heterozygous barcode SNPs, heterozygous AMAI SNPS, COI by MS). Agreement between methods was determined using Cohen's Kappa in R. As all methods are only an approximation of the truth (and the true COI is unknown), we sought to determine the optimal method from those tested based on the area under receiver operating characteristic (ROC) curves (using pROC package in R (139)), assuming the optimal method would have the highest area-under-the-curve compared to all other methods. As methods generated diverging results, the mode (most frequent value - single *vs.* multiple clone - observed with the 4 methods) was determined as final measure of COI for each sample.

Population genetic analysis

Population genetic analysis was performed using the 28-SNP barcode. Samples with missing data at more than 7 SNP loci (25%) were excluded from the analysis. Samples with heterozygous SNP genotype at one or multiple loci were included in the analysis, justified by the observation that little genetic differentiation (F_{ST} =0.05) was observed between samples with single clone barcodes *vs.*

samples with heterozygous barcodes. Genetic differentiation was measured as F_{ST} (140), $G`_{ST}$ (141) and Jost's D (142) using 1000 bootstraps with the R package DiveRsity (143). F_{ST} measures genetic distance with biallelic markers, but was generalized as G_{ST} for multiple alleles. However, for loci with high mutation rates, such as microsatellites, the amount of mutations can influence the values of F_{ST}, and as a result, new indices such as G'_{ST} and Jost's D have been proposed, and are also reported here (144).

Similarity between the samples was explored over space (district) and time (year of collection) using a PCA, which is a visual tool that clusters together samples with a similar genotype, while contrasting genotypes will be further away from each other. PCA was performed on the genotype matrix using 'prcomp' function in the stats package v4.0.5 in R base, where prior to PCA missing genotypes were replaced by the mean of the allelic frequency observed in all samples. PCA was performed using 1) 28-SNP barcode positions only and 2) all biallelic SNPs in the core region. Populations for subsequent analyses were defined based on observed clustering in the PCA as followed: temporal groups with samples from 2003-2005, 2008-2012 and 2014-2018 and district. Genetic diversity was calculated as expected heterozygosity (*He*) from 28-SNP barcode loci using the adegenet package (145, 146) in R using diploid genotype calls. Wilcoxon signed rank test with continuity correction was used to test whether the mean observed *He* was lower than the expected *He*. Linkage disequilibrium (LD) was measured as the standardized index of association ($\bar{r}D$) as pairwise LD and multilocus LD using 999 resamplings (method: permutation of alleles) using the poppr package v2.8.6. (135) in R. Discriminant Analysis of Principal Components (DAPC) (129) calculates the discriminant components using predefined populations described above in such a way that samples from the same population will group together, while it simultaneously maximizes the distance with samples from other populations. DAPC was performed using the R package adegenet using 40 principal components (PCs) and 7 discriminants (determined through cross-validation). To get insight into the underlying causes of the differentiation between the populations, we inspected the associated allele loadings (the contribution of an allele to a principal component) of the DAPC analysis for the first and second principal components (129).

Analysis of genetic variants

Allele frequencies (AF) for barcode loci were calculated from allele depths from the selected samples (N=254) in order to better reflect the true allele frequencies in complex infections using an in-house R script. First, AF was calculated per locus and sample using the allele depths (AD format field in gvcf). Next, we summed the AF at each locus (SUM-AF) from all samples and then divided the SUM-AFs by the sum of within-sample AFs for all alleles at that locus within the defined populations to obtain an overall AF for each allele for that population. Haplotypes for each drug resistance gene were created using the genotype calls for variants of interest (Supplementary file 1) with an in-house R script, allowing for mixed haplotypes if more than one allele was detected at one or more locus of interest.

A neighbor-joining tree was made using core variants build in a matrix of Euclidean distances between all selected samples (n = 254) and controls (n = 13) using the R packages stats and ape (147, 148). The tree was rooted on the 3D7 control strain and visualized in Microreact (149) with sample meta data, including drug resistant haplotypes, date and geographical location (https://microreact.org/project/aV7RHNCBmG3sJ2rwzchE4k/279b0013).

Validation of detection of *hrp2*/*hrp3* deletions

We explored the potential of using short read sequencing data to determine the presence/absence of deletion in the *hrp2* and *hrp3* genes in 10 samples for which we had previously performed *hrp2* and

hrp3 PCRs (2). Depth ratio for each locus in each sample was calculated as the depth of coverage (format field DP) divided by the sample mean DP. Mean depth ratio per *hrp2* and *hrp3* amplicon was calculated for each sample. Log transformed mean depth ratios of samples with *hrp2*/*hrp3* deletions determined by PCR were compared, to arbitrarily define thresholds for each amplicon to classify the sample as deletion or presence (or undetermined) for *hrp2* and *hrp3* based on the read depth of that amplicon (Supplementary figure S9 and table S14). As both *hrp2* and *hrp3* are covered by multiple amplicons, a final classification of presence/absence of *hrp2*/*hrp3* was based on the proportion of amplicons with a deletion in all 5 amplicons. One amplicon for *hrp2* (AMPL3593062) was not used for the classification, as it offered no discriminatory power.

To classify a sample as *hrp2* and *hrp3* deleted or non-deleted, the number of amplicons per sample and gene with deletions was summed and then divided by the total number of amplicons (with or without the deletion). If the resulting ratio was > 0.8 a sample was classified as having a deletion in *hrp3* or *hrp2*; if the ratio was < 0.2 for *hrp2* or 0.4 for *hrp3*, the samples was classified as without deletion in that gene. Due to the repetitive nature and homologies of the *hrp2* and *hrp3* genes, misalignment between reads from these two genes can occur especially in *hrp2*. Therefore, the thresholds were more difficult to determine for *hrp2*, when the *hrp3* gene was also present, resulting in a "grey-zone" of *hrp2* mean depth ratio where we did not distinguish between presence or absence of this gene (Supplementary figure S9).

Data availability: Sample meta data and drug resistant and *hrp2* and *hrp3* haplotypes, barcodes, MS and COI is accessible at https://microreact.org/project/aV7RHNCBmG3sJ2rwzchE4k. Raw data (fastq), variant files (vcf), and scripts are available on request. All other data is included in the manuscript and supporting files.

Acknowledgements: We wish to thank all clinical, microscopy and field staff who supported the sample collections and all participants for making their material available for malaria studies. We acknowledge the support of the Dirección Regional de Salud de Loreto for the study authorization and sample collection activities in Loreto. The following reagents were obtained through BEI Resources, NIAID, NIH: *P. falciparum*, Strain Dd2, MRA-150, contributed by David Walliker; *P. falciparum*, Strain Dd2_R539T, MRA-1255, and *P. falciparum*, Strain CamWT_C580Y, MRA-1251, contributed by David A. Fidock; and *P. falciparum*, Strain IPC 4912, MRA-1241, contributed by Didier Ménard.

Funding: This work was funded by the Belgium Development Cooperation (DGD) under the Framework Agreement Program between DGD and ITM (FA4 Peru, 2017-2021) and the sample collections in 2018 were supported by VLIR-UOS (project PE2018TEA470A102; University of Antwerp). Funding for the sample collections lead by the U.S. Naval Medical Research Unit 6 (NAMRU-6) in 2011 and 2012 was provided by the Armed Forces Health Surveillance Division (AFHSD) and its Global Emerging Infections Surveillance and Response (GEIS) Section (P0144_20_N6_01, 2020-2021).

Competing interests: The authors declare that they have no competing interests.

Copyright statement: Some authors of this manuscript are military service members and employees of the U.S. Government. This work was prepared as part of their official duties. Title 17 U.S.C. §105 provides that "Copyright protection under this Title is not available for any work of the United States Government". Title 17 U.S.C. §101 defines a U.S. Government work as a work prepared by a military service member or employee of the U.S. Government as part of that person's official duties. The views

expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, nor the U.S. Government.

Author contributions (in alphabetical order per item): Conception of ideas for the study: ARU, DG, JHK; Selection of targets and design of the assay: ERV, JHK, NvD; Sample collections and coordination of field work: CDG, CFM, DG, HOV, JPvG; Laboratory experiments: CFM, JHK, NvD, PG; Bioinformatics & Data analysis: JHK, LLA, NvD, PM; Writing first draft manuscript: ARU, JHK, CFM, NvD. All authors reviewed and contributed to the final version of the manuscript.

References

1. WHO. World Malaria Report 2019. Geneva: World Health Organization; 2019.
2. Gamboa D, Ho MF, Bendezu J, Torres K, Chiodini PL, Barnwell JW, et al. A large proportion of P. falciparum isolates in the Amazon region of Peru lack pfhrp2 and pfhrp3: implications for malaria rapid diagnostic tests. PloS one. 2010;5(1):e8091.
3. WHO. Malaria Surveillance, Monitoring & Evaluation: A Reference Manual. Geneva: World Health Organization; 2018.
4. WHO. Methods for surveillance of antimalarial drug efficacy. Geneva: World Health Organization; 2009.
5. Wangdi K, Pasaribu AP, Clements ACA. Addressing hard-to-reach populations for achieving malaria elimination in the Asia Pacific Malaria Elimination Network countries. Asia & the Pacific Policy Studies.n/a(n/a).
6. WHO. Minutes of the Evidence Review Group meeting on the emergence and spread of multidrug-resistant Plasmodium falciparum lineages in the Greater Mekong subregion. . Geneva: World Health Organization; 2016.
7. Committee MPA. Technical consultation on the role of parasite and anopheline genetics in malaria surveillance. Geneva: World Health Organization; 2019.
8. Ariey F, Witkowski B, Amaratunga C, Beghain J, Langlois AC, Khim N, et al. A molecular marker of artemisinin-resistant Plasmodium falciparum malaria. Nature. 2014;505(7481):50-5.
9. Menard D, Khim N, Beghain J, Adegnika AA, Shafiul-Alam M, Amodu O, et al. A Worldwide Map of Plasmodium falciparum K13-Propeller Polymorphisms. The New England journal of medicine. 2016;374(25):2453-64.
10. Noedl H, Se Y, Schaecher K, Smith BL, Socheat D, Fukuda MM, et al. Evidence of artemisinin-resistant malaria in western Cambodia. The New England journal of medicine. 2008;359(24):2619-20.
11. Woodrow CJ, White NJ. The clinical impact of artemisinin resistance in Southeast Asia and the potential for future spread. FEMS Microbiol Rev. 2017;41(1):34-48.
12. Taylor SM, Parobek CM, Aragam N, Ngasala BE, Martensson A, Meshnick SR, et al. Pooled deep sequencing of Plasmodium falciparum isolates: an efficient and scalable tool to quantify prevailing malaria drug-resistance genotypes. The Journal of infectious diseases. 2013;208(12):1998-2006.
13. Nsanzabana C, Djalle D, Guérin PJ, Ménard D, González IJ. Tools for surveillance of anti-malarial drug resistance: an assessment of the current landscape. Malaria journal. 2018;17(1):75.
14. Duraisingh MT, Curtis J, Warhurst DC. Plasmodium falciparum: detection of polymorphisms in the dihydrofolate reductase and dihydropteroate synthetase genes by PCR and restriction digestion. Experimental parasitology. 1998;89(1):1-8.
15. Eldin de Pecoulas P, Basco LK, Abdallah B, Dje MK, Le Bras J, Mazabraud A. Plasmodium falciparum: detection of antifolate resistance by mutation-specific restriction enzyme digestion. Experimental parasitology. 1995;80(3):483-7.
16. Programme WGM. WHO Malaria Policy Advisory Committee (MPAC) meeting - meeting report. Geneva; 2019.
17. Mita T, Tanabe K, Kita K. Spread and evolution of Plasmodium falciparum drug resistance. Parasitology International. 2009;58(3):201-9.
18. Auburn S, Barry AE. Dissecting malaria biology and epidemiology using population genetics and genomics. International journal for parasitology. 2017;47(2-3):77-85.
19. Fola AA, Harrison GLA, Hazairin MH, Barnadas C, Hetzel MW, Iga J, et al. Higher Complexity of Infection and Genetic Diversity of Plasmodium vivax Than Plasmodium falciparum Across All Malaria Transmission Zones of Papua New Guinea. Am J Trop Med Hyg. 2017;96(3):630-41.
20. Fola AA, Nate E, Abby Harrison GL, Barnadas C, Hetzel MW, Iga J, et al. Nationwide genetic surveillance of Plasmodium vivax in Papua New Guinea reveals heterogeneous transmission

dynamics and routes of migration amongst subdivided populations. Infect Genet Evol. 2018;58:83-95.
21. Waltmann A, Koepfli C, Tessier N, Karl S, Fola A, Darcy AW, et al. Increasingly inbred and fragmented populations of Plasmodium vivax associated with the eastward decline in malaria transmission across the Southwest Pacific. PLoS Negl Trop Dis. 2018;12(1):e0006146.
22. Falk N, Maire N, Sama W, Owusu-Agyei S, Smith T, Beck HP, et al. Comparison of PCR-RFLP and Genescan-based genotyping for analyzing infection dynamics of Plasmodium falciparum. Am J Trop Med Hyg. 2006;74(6):944-50.
23. Koepfli C, Mueller I, Marfurt J, Goroti M, Sie A, Oa O, et al. Evaluation of Plasmodium vivax genotyping markers for molecular monitoring in clinical trials. The Journal of infectious diseases. 2009;199(7):1074-80.
24. Anderson TJ, Haubold B, Williams JT, Estrada-Franco JG, Richardson L, Mollinedo R, et al. Microsatellite markers reveal a spectrum of population structures in the malaria parasite Plasmodium falciparum. Mol Biol Evol. 2000;17(10):1467-82.
25. Imwong M, Nair S, Pukrittayakamee S, Sudimack D, Williams JT, Mayxay M, et al. Contrasting genetic structure in Plasmodium vivax populations from Asia and South America. International journal for parasitology. 2007;37(8-9):1013-22.
26. Karunaweera ND, Ferreira MU, Munasinghe A, Barnwell JW, Collins WE, King CL, et al. Extensive microsatellite diversity in the human malaria parasite Plasmodium vivax. Gene. 2008;410(1):105-12.
27. Kattenberg JH, Razook Z, Keo R, Koepfli C, Jennison C, Lautu-Gumal D, et al. Monitoring Plasmodium falciparum and Plasmodium vivax using microsatellite markers indicates limited changes in population structure after substantial transmission decline in Papua New Guinea. Mol Ecol. 2020;29(23):4525-41.
28. Baniecki ML, Faust AL, Schaffner SF, Park DJ, Galinsky K, Daniels RF, et al. Development of a single nucleotide polymorphism barcode to genotype Plasmodium vivax infections. PLoS Negl Trop Dis. 2015;9(3):e0003539.
29. Daniels R, Volkman SK, Milner DA, Mahesh N, Neafsey DE, Park DJ, et al. A general SNP-based molecular barcode for Plasmodium falciparum identification and tracking. Malaria journal. 2008;7:223.
30. Neafsey DE, Schaffner SF, Volkman SK, Park D, Montgomery P, Milner DA, Jr., et al. Genome-wide SNP genotyping highlights the role of natural selection in Plasmodium falciparum population divergence. Genome biology. 2008;9(12):R171.
31. Daniels R, Chang HH, Sene PD, Park DC, Neafsey DE, Schaffner SF, et al. Genetic surveillance detects both clonal and epidemic transmission of malaria following enhanced intervention in Senegal. PloS one. 2013;8(4):e60780.
32. Daniels RF, Schaffner SF, Wenger EA, Proctor JL, Chang H-H, Wong W, et al. Modeling malaria genomics reveals transmission decline and rebound in Senegal. Proceedings of the National Academy of Sciences. 2015;112(22):7067-72.
33. Koepfli C, Mueller I. Malaria Epidemiology at the Clone Level. Trends Parasitol. 2017;33(12):974-85.
34. Fola AA, Kattenberg E, Razook Z, Lautu-Gumal D, Lee S, Mehra S, et al. SNP barcodes provide higher resolution than microsatellite markers to measure Plasmodium vivax population genetics. Malaria journal. 2020;19(1):375.
35. Tessema SK, Hathaway NJ, Teyssier NB, Murphy M, Chen A, Aydemir O, et al. Sensitive, Highly Multiplexed Sequencing of Microhaplotypes From the Plasmodium falciparum Heterozygome. The Journal of infectious diseases. 2020.
36. Thomson R, Parr JB, Cheng Q, Chenet S, Perkins M, Cunningham J. Prevalence of Plasmodium falciparum lacking histidine-rich proteins 2 and 3: a systematic review. Bull World Health Organ. 2020;98(8):558-68F.

37. Akinyi S, Hayden T, Gamboa D, Torres K, Bendezu J, Abdallah JF, et al. Multiple genetic origins of histidine-rich protein 2 gene deletion in Plasmodium falciparum parasites from Peru. Sci Rep. 2013;3:2797.
38. Murillo Solano C, Akinyi Okoth S, Abdallah JF, Pava Z, Dorado E, Incardona S, et al. Deletion of Plasmodium falciparum Histidine-Rich Protein 2 (pfhrp2) and Histidine-Rich Protein 3 (pfhrp3) Genes in Colombian Parasites. PloS one. 2015;10(7):e0131576.
39. Dorado EJ, Okoth SA, Montenegro LM, Diaz G, Barnwell JW, Udhayakumar V, et al. Genetic Characterisation of Plasmodium falciparum Isolates with Deletion of the pfhrp2 and/or pfhrp3 Genes in Colombia: The Amazon Region, a Challenge for Malaria Diagnosis and Control. PloS one. 2016;11(9):e0163137.
40. Rachid Viana GM, Akinyi Okoth S, Silva-Flannery L, Lima Barbosa DR, Macedo de Oliveira A, Goldman IF, et al. Histidine-rich protein 2 (pfhrp2) and pfhrp3 gene deletions in Plasmodium falciparum isolates from select sites in Brazil and Bolivia. PloS one. 2017;12(3):e0171150.
41. WHO. Statement by the Malaria Policy Advisory Group on the urgent need to address the high prevalence of pfhrp2/3 gene deletions in the Horn of Africa and beyond 2021 [Available from: https://www.who.int/news/item/28-05-2021-statement-by-the-malaria-policy-advisory-group-on-the-urgent-need-to-address-the-high-prevalence-of-pfhrp2-3-gene-deletions-in-the-horn-of-africa-and-beyond.
42. Agaba BB, Yeka A, Nsobya S, Arinaitwe E, Nankabirwa J, Opigo J, et al. Systematic review of the status of pfhrp2 and pfhrp3 gene deletion, approaches and methods used for its estimation and reporting in Plasmodium falciparum populations in Africa: review of published studies 2010–2019. Malaria journal. 2019;18(1):355.
43. WHO. World Malaria Report 2020. Geneva: World Health Organization; 2020.
44. Cortese JF, Caraballo A, Contreras CE, Plowe CV. Origin and dissemination of Plasmodium falciparum drug-resistance mutations in South America. The Journal of infectious diseases. 2002;186(7):999-1006.
45. Chenet SM, Akinyi Okoth S, Huber CS, Chandrabose J, Lucchi NW, Talundzic E, et al. Independent Emergence of the Plasmodium falciparum Kelch Propeller Domain Mutant Allele C580Y in Guyana. The Journal of infectious diseases. 2016;213(9):1472-5.
46. Mathieu LC, Cox H, Early AM, Mok S, Lazrek Y, Paquet JC, et al. Local emergence in Amazonia of Plasmodium falciparum k13 C580Y mutants associated with in vitro artemisinin resistance. Elife. 2020;9.
47. Vreden SG, Jitan JK, Bansie RD, Adhin MR. Evidence of an increased incidence of day 3 parasitaemia in Suriname: an indicator of the emerging resistance of Plasmodium falciparum to artemether. Mem Inst Oswaldo Cruz. 2013;108(8):968-73.
48. Chenet SM, Okoth SA, Kelley J, Lucchi N, Huber CS, Vreden S, et al. Molecular Profile of Malaria Drug Resistance Markers of Plasmodium falciparum in Suriname. Antimicrob Agents Chemother. 2017;61(7).
49. Demas AR, Sharma AI, Wong W, Early AM, Redmond S, Bopp S, et al. Mutations in Plasmodium falciparum actin-binding protein coronin confer reduced artemisinin susceptibility. Proc Natl Acad Sci U S A. 2018.
50. Sharma AI, Shin SH, Bopp S, Volkman SK, Hartl DL, Wirth DF. Genetic background and PfKelch13 affect artemisinin susceptibility of PfCoronin mutants in Plasmodium falciparum. PLoS Genet. 2020;16(12):e1009266.
51. Henriques G, Hallett RL, Beshir KB, Gadalla NB, Johnson RE, Burrow R, et al. Directional selection at the pfmdr1, pfcrt, pfubp1, and pfap2mu loci of Plasmodium falciparum in Kenyan children treated with ACT. The Journal of infectious diseases. 2014;210(12):2001-8.
52. Sutherland CJ, Henrici RC, Artavanis-Tsakonas K. Artemisinin susceptibility in the malaria parasite Plasmodium falciparum: propellers, adaptor proteins and the need for cellular healing. FEMS Microbiol Rev. 2021;45(3).

53. Xie SC, Ralph SA, Tilley L. K13, the Cytostome, and Artemisinin Resistance. Trends in Parasitology. 2020;36(6):533-44.
54. Política Nacional de Medicamentos para el Control de la Malaria en el Perú. Lima, Perú: Ministerio de Salud del Perú; 1999.
55. Marquino W, MacArthur JR, Barat LM, Oblitas FE, Arrunategui M, Garavito G, et al. Efficacy of chloroquine, sulfadoxine-pyrimethamine, and mefloquine for the treatment of uncomplicated Plasmodium falciparum malaria on the north coast of Peru. Am J Trop Med Hyg. 2003;68(1):120-3.
56. Aramburú Guarda J, Ramal Asayag C, Witzig R. Malaria reemergence in the Peruvian Amazon region. Emerg Infect Dis. 1999;5(2):209-15.
57. Ruebush TK, 2nd, Neyra D, Cabezas C. Modifying national malaria treatment policies in Peru. J Public Health Policy. 2004;25(3-4):328-45.
58. Williams HA, Vincent-Mark A, Herrera Y, Chang OJ. A retrospective analysis of the change in anti-malarial treatment policy: Peru. Malaria journal. 2009;8:85.
59. Marquino W, Huilca M, Calampa C, Falconi E, Cabezas C, Naupay R, et al. Efficacy of mefloquine and a mefloquine-artesunate combination therapy for the treatment of uncomplicated Plasmodium falciparum malaria in the Amazon Basin of Peru. Am J Trop Med Hyg. 2003;68(5):608-12.
60. Norma téchnica de salud para la atencion de la malaria y malaria grave en el Peru. Iquitos, Peru: Ministerio de Salud Perú, Direccion regional de Salud Loreto; 2015.
61. de Oliveira AM, Chavez J, de Leon GP, Durand S, Arrospide N, Roberts J, et al. Efficacy and effectiveness of mefloquine and artesunate combination therapy for uncomplicated Plasmodium falciparum malaria in the Peruvian Amazon. Am J Trop Med Hyg. 2011;85(3):573-8.
62. Illumina. AmpliSeq for Illumina Custom and Community Panels: Reference Guide. San Diego, California, USA2018.
63. WHO. Report on antimalarial drug efficacy, resistance and response: 10 years of surveillance (2010–2019). Geneva: World Health Organization; 2020.
64. Henriques G, van Schalkwyk DA, Burrow R, Warhurst DC, Thompson E, Baker DA, et al. The Mu subunit of Plasmodium falciparum clathrin-associated adaptor protein 2 modulates in vitro parasite response to artemisinin and quinine. Antimicrob Agents Chemother. 2015;59(5):2540-7.
65. Adams T, Ennuson NAA, Quashie NB, Futagbi G, Matrevi S, Hagan OCK, et al. Prevalence of Plasmodium falciparum delayed clearance associated polymorphisms in adaptor protein complex 2 mu subunit (pfap2mu) and ubiquitin specific protease 1 (pfubp1) genes in Ghanaian isolates. Parasit Vectors. 2018;11(1):175.
66. Henrici RC, Edwards RL, Zoltner M, van Schalkwyk DA, Hart MN, Mohring F, et al. The Plasmodium falciparum Artemisinin Susceptibility-Associated AP-2 Adaptin mu Subunit is Clathrin Independent and Essential for Schizont Maturation. Mbio. 2020;11(1).
67. Martin RE, Marchetti RV, Cowan AI, Howitt SM, Broer S, Kirk K. Chloroquine transport via the malaria parasite's chloroquine resistance transporter. Science. 2009;325(5948):1680-2.
68. Sidhu AB, Verdier-Pinard D, Fidock DA. Chloroquine resistance in Plasmodium falciparum malaria parasites conferred by pfcrt mutations. Science. 2002;298(5591):210-3.
69. Fidock DA, Nomura T, Talley AK, Cooper RA, Dzekunov SM, Ferdig MT, et al. Mutations in the P. falciparum digestive vacuole transmembrane protein PfCRT and evidence for their role in chloroquine resistance. Mol Cell. 2000;6(4):861-71.
70. Nagesha HS, Casey GJ, Rieckmann KH, Fryauff DJ, Laksana BS, Reeder JC, et al. NEW HAPLOTYPES OF THE PLASMODIUM FALCIPARUM CHLOROQUINE RESISTANCE TRANSPORTER (PFCRT) GENE AMONG CHLOROQUINE-RESISTANT PARASITE ISOLATES. The American Journal of Tropical Medicine and Hygiene Am J Trop Med Hyg. 2003;68(4):398-402.
71. Roepe PD. Molecular and physiologic basis of quinoline drug resistance in Plasmodium falciparum malaria. Future Microbiology. 2009;4(4):441-55.

72. Ross LS, Dhingra SK, Mok S, Yeo T, Wicht KJ, Kumpornsin K, et al. Emerging Southeast Asian PfCRT mutations confer Plasmodium falciparum resistance to the first-line antimalarial piperaquine. Nat Commun. 2018;9(1):3314.
73. Dhingra SK, Redhi D, Combrinck JM, Yeo T, Okombo J, Henrich PP, et al. A Variant PfCRT Isoform Can Contribute to Plasmodium falciparum Resistance to the First-Line Partner Drug Piperaquine. MBio. 2017;8(3).
74. Hamilton WL, Amato R, van der Pluijm RW, Jacob CG, Quang HH, Thuy-Nhien NT, et al. Evolution and expansion of multidrug-resistant malaria in southeast Asia: a genomic epidemiology study. The Lancet Infectious Diseases. 2019;19(9):943-51.
75. Pelleau S, Moss EL, Dhingra SK, Volney B, Casteras J, Gabryszewski SJ, et al. Adaptive evolution of malaria parasites in French Guiana: Reversal of chloroquine resistance by acquisition of a mutation in pfcrt. Proceedings of the National Academy of Sciences. 2015;112:11672-7.
76. van der Pluijm RW, Imwong M, Chau NH, Hoa NT, Thuy-Nhien NT, Thanh NV, et al. Determinants of dihydroartemisinin-piperaquine treatment failure in Plasmodium falciparum malaria in Cambodia, Thailand, and Vietnam: a prospective clinical, pharmacological, and genetic study. Lancet Infect Dis. 2019;19(9):952-61.
77. Folarin OA, Bustamante C, Gbotosho GO, Sowunmi A, Zalis MG, Oduola AM, et al. In vitro amodiaquine resistance and its association with mutations in pfcrt and pfmdr1 genes of Plasmodium falciparum isolates from Nigeria. Acta Trop. 2011;120(3):224-30.
78. Korsinczky M, Chen N, Kotecka B, Saul A, Rieckmann K, Cheng Q. Mutations in Plasmodium falciparum cytochrome b that are associated with atovaquone resistance are located at a putative drug-binding site. Antimicrob Agents Chemother. 2000;44(8):2100-8.
79. Akhoon BA, Singh KP, Varshney M, Gupta SK, Shukla Y, Gupta SK. Understanding the mechanism of atovaquone drug resistance in Plasmodium falciparum cytochrome b mutation Y268S using computational methods. PloS one. 2014;9(10):e110041.
80. Peters JM, Chen N, Gatton M, Korsinczky M, Fowler EV, Manzetti S, et al. Mutations in cytochrome b resulting in atovaquone resistance are associated with loss of fitness in Plasmodium falciparum. Antimicrob Agents Chemother. 2002;46(8):2435-41.
81. Bacon DJ, McCollum AM, Griffing SM, Salas C, Soberon V, Santolalla M, et al. Dynamics of malaria drug resistance patterns in the Amazon basin region following changes in Peruvian national treatment policy for uncomplicated malaria. Antimicrob Agents Chemother. 2009;53(5):2042-51.
82. Griffing SM, Mixson-Hayden T, Sridaran S, Alam MT, McCollum AM, Cabezas C, et al. South American Plasmodium falciparum after the malaria eradication era: clonal population expansion and survival of the fittest hybrids. PloS one. 2011;6(9):e23486.
83. Griffing SM, Viana GM, Mixson-Hayden T, Sridaran S, Alam MT, de Oliveira AM, et al. Historical shifts in Brazilian P. falciparum population structure and drug resistance alleles. PloS one. 2013;8(3):e58984.
84. Costa GL, Amaral LC, Fontes CJF, Carvalho LH, de Brito CFA, de Sousa TN. Assessment of copy number variation in genes related to drug resistance in Plasmodium vivax and Plasmodium falciparum isolates from the Brazilian Amazon and a systematic review of the literature. Malaria journal. 2017;16(1):152.
85. Basco LK, Ringwald P. Molecular epidemiology of malaria in Yaounde, Cameroon. VI. Sequence variations in the Plasmodium falciparum dihydrofolate reductase-thymidylate synthase gene and in vitro resistance to pyrimethamine and cycloguanil. The American journal of tropical medicine and hygiene Am J Trop Med Hyg Am J Trop Med Hyg. 2000;62(2):271-6.
86. Hyde JE. Drug-resistant malaria − an insight. The FEBS Journal. 2007;274(18):4688-98.
87. Staines HM, Burrow R, Teo BH-Y, Chis Ster I, Kremsner PG, Krishna S. Clinical implications of Plasmodium resistance to atovaquone/proguanil: a systematic review and meta-analysis. Journal of Antimicrobial Chemotherapy. 2017;73(3):581-95.

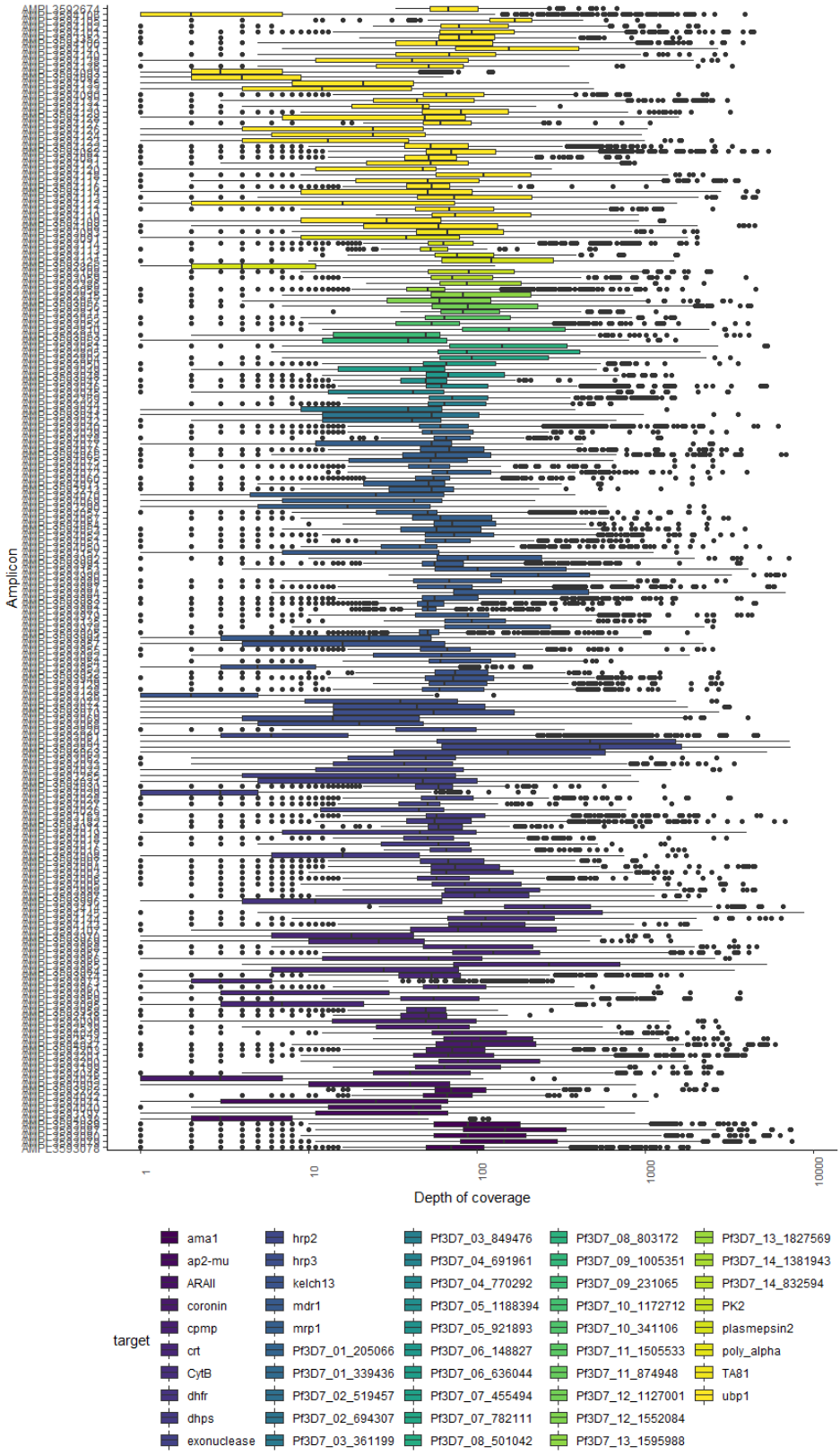
88. Amato R, Lim P, Miotto O, Amaratunga C, Dek D, Pearson RD, et al. Genetic markers associated with dihydroartemisinin-piperaquine failure in Plasmodium falciparum malaria in Cambodia: a genotype-phenotype association study. Lancet Infect Dis. 2017;17(2):164-73.
89. Reed MB, Saliba KJ, Caruana SR, Kirk K, Cowman AF. Pgh1 modulates sensitivity and resistance to multiple antimalarials in Plasmodium falciparum. Nature. 2000;403(6772):906-9.
90. Pickard Amy L, Wongsrichanalai C, Purfield A, Kamwendo D, Emery K, Zalewski C, et al. Resistance to Antimalarials in Southeast Asia and Genetic Polymorphisms in pfmdr1. Antimicrobial Agents and Chemotherapy. 2003;47(8):2418-23.
91. Veiga MI, Dhingra SK, Henrich PP, Straimer J, Gnadig N, Uhlemann AC, et al. Globally prevalent PfMDR1 mutations modulate Plasmodium falciparum susceptibility to artemisinin-based combination therapies. Nat Commun. 2016;7:11553.
92. Price RN, Uhlemann AC, Brockman A, McGready R, Ashley E, Phaipun L, et al. Mefloquine resistance in Plasmodium falciparum and increased pfmdr1 gene copy number. Lancet (London, England). 2004;364(9432):438-47.
93. Sidhu AB, Valderramos SG, Fidock DA. pfmdr1 mutations contribute to quinine resistance and enhance mefloquine and artemisinin sensitivity in Plasmodium falciparum. Mol Microbiol. 2005;57(4):913-26.
94. Mbaye A, Dieye B, Ndiaye YD, Bei AK, Muna A, Deme AB, et al. Selection of N86F184D1246 haplotype of Pfmrd1 gene by artemether-lumefantrine drug pressure on Plasmodium falciparum populations in Senegal. Malaria journal. 2016;15(1):433.
95. Venkatesan M, Gadalla NB, Stepniewska K, Dahal P, Nsanzabana C, Moriera C, et al. Polymorphisms in Plasmodium falciparum chloroquine resistance transporter and multidrug resistance 1 genes: parasite risk factors that affect treatment outcomes for P. falciparum malaria after artemether-lumefantrine and artesunate-amodiaquine. Am J Trop Med Hyg. 2014;91(4):833-43.
96. Dahlstrom S, Ferreira PE, Veiga MI, Sedighi N, Wiklund L, Martensson A, et al. Plasmodium falciparum multidrug resistance protein 1 and artemisinin-based combination therapy in Africa. The Journal of infectious diseases. 2009;200(9):1456-64.
97. Bopp S, Magistrado P, Wong W, Schaffner SF, Mukherjee A, Lim P, et al. Plasmepsin II-III copy number accounts for bimodal piperaquine resistance among Cambodian Plasmodium falciparum. Nat Commun. 2018;9(1):1769.
98. Witkowski B, Duru V, Khim N, Ross LS, Saintpierre B, Beghain J, et al. A surrogate marker of piperaquine-resistant Plasmodium falciparum malaria: a phenotype-genotype association study. Lancet Infect Dis. 2017;17(2):174-83.
99. Dharia NV, Plouffe D, Bopp SE, Gonzalez-Paez GE, Lucas C, Salas C, et al. Genome scanning of Amazonian Plasmodium falciparum shows subtelomeric instability and clindamycin-resistant parasites. Genome Res. 2010;20(11):1534-44.
100. Birnbaum J, Scharf S, Schmidt S, Jonscher E, Hoeijmakers WAM, Flemming S, et al. A Kelch13-defined endocytosis pathway mediates artemisinin resistance in malaria parasites. Science. 2020;367(6473):51-9.
101. Lerch A, Koepfli C, Hofmann NE, Messerli C, Wilcox S, Kattenberg JH, et al. Development of amplicon deep sequencing markers and data analysis pipeline for genotyping multi-clonal malaria infections. BMC Genomics. 2017;18(1):864.
102. Straimer J, Gnädig NF, Witkowski B, Amaratunga C, Duru V, Ramadani AP, et al. Drug resistance. K13-propeller mutations confer artemisinin resistance in Plasmodium falciparum clinical isolates. Science. 2015;347(6220):428-31.
103. Miotto O, Sekihara M, Tachibana S-I, Yamauchi M, Pearson RD, Amato R, et al. Emergence of artemisinin-resistant Plasmodium falciparum with kelch13 C580Y mutations on the island of New Guinea. PLOS Pathogens. 2020;16(12):e1009133.

104. Rovira-Vallbona E, Van Hong N, Kattenberg JH, Huan RM, Hien NTT, Ngoc NTH, et al. Efficacy of dihydroartemisinin/piperaquine and artesunate monotherapy for the treatment of uncomplicated Plasmodium falciparum malaria in Central Vietnam. J Antimicrob Chemother. 2020;75(8):2272-81.
105. Cerqueira GC, Cheeseman IH, Schaffner SF, Nair S, McDew-White M, Phyo AP, et al. Longitudinal genomic surveillance of Plasmodium falciparum malaria parasites reveals complex genomic architecture of emerging artemisinin resistance. Genome biology. 2017;18(1):78.
106. Olshina MA, Angrisano F, Marapana DS, Riglar DT, Bane K, Wong W, et al. Plasmodium falciparum coronin organizes arrays of parallel actin filaments potentially guiding directional motility in invasive malaria parasites. Malaria journal. 2015;14:280.
107. Velavan TP, Nderu D, Agbenyega T, Ntoumi F, Kremsner PG. An alternative dogma on reduced artemisinin susceptibility: A new shadow from east to west. Proceedings of the National Academy of Sciences. 2019;116:12611-2.
108. Sridaran S, McClintock SK, Syphard LM, Herman KM, Barnwell JW, Udhayakumar V. Anti-folate drug resistance in Africa: meta-analysis of reported dihydrofolate reductase (dhfr) and dihydropteroate synthase (dhps) mutant genotype frequencies in African Plasmodium falciparum parasite populations. Malaria journal. 2010;9:247.
109. Baldeviano GC, Okoth SA, Arrospide N, Gonzalez RV, Sanchez JF, Macedo S, et al. Molecular Epidemiology of Plasmodium falciparum Malaria Outbreak, Tumbes, Peru, 2010-2012. Emerg Infect Dis. 2015;21(5):797-803.
110. Okoth SA, Chenet SM, Arrospide N, Gutierrez S, Cabezas C, Matta JA, et al. Molecular Investigation into a Malaria Outbreak in Cusco, Peru: Plasmodium falciparum BV1 Lineage is Linked to a Second Outbreak in Recent Times. Am J Trop Med Hyg. 2016;94(1):128-31.
111. Behrens HM, Schmidt S, Spielmann T. The newly discovered role of endocytosis in artemisinin resistance. Medicinal Research Reviews.n/a(n/a).
112. Yalcindag E, Elguero E, Arnathau C, Durand P, Akiana J, Anderson TJ, et al. Multiple independent introductions of Plasmodium falciparum in South America. Proc Natl Acad Sci U S A. 2012;109(2):511-6.
113. Rodrigues PT, Valdivia HO, de Oliveira TC, Alves JMP, Duarte AMRC, Cerutti-Junior C, et al. Human migration and the spread of malaria parasites to the New World. Scientific Reports. 2018;8(1):1993.
114. Rosenthal MR, Ng CL. Plasmodium falciparum Artemisinin Resistance: The Effect of Heme, Protein Damage, and Parasite Cell Stress Response. ACS Infectious Diseases. 2020;6(7):1599-614.
115. Oberstaller J, Zoungrana L, Bannerman CD, Jahangiri S, Dwivedi A, Silva JC, et al. Integration of population and functional genomics to understand mechanisms of artemisinin resistance in Plasmodium falciparum. International Journal for Parasitology: Drugs and Drug Resistance. 2021;16:119-28.
116. Recht J, Siqueira AM, Monteiro WM, Herrera SM, Herrera S, Lacerda MVG. Malaria in Brazil, Colombia, Peru and Venezuela: current challenges in malaria control and elimination. Malaria journal. 2017;16(1):273.
117. Evans L, 3rd, Coignez V, Barojas A, Bempong D, Bradby S, Dijiba Y, et al. Quality of anti-malarials collected in the private and informal sectors in Guyana and Suriname. Malaria journal. 2012;11:203.
118. Schmedes SE, Patel D, Kelley J, Udhayakumar V, Talundzic E. Using the Plasmodium mitochondrial genome for classifying mixed-species infections and inferring the geographical origin of P. falciparum parasites imported to the U.S. PloS one. 2019;14(4):e0215754.
119. Preston MD, Campino S, Assefa SA, Echeverry DF, Ocholla H, Amambua-Ngwa A, et al. A barcode of organellar genome polymorphisms identifies the geographic origin of Plasmodium falciparum strains. Nature Communications. 2014;5(1):4052.
120. Trimarsanto H, Amato R, Pearson RD, Sutanto E, Noviyanti R, Trianty L, et al. A molecular barcode and online tool to identify and map imported infection with Plasmodium vivax. bioRxiv. 2019:776781.

121. Hathaway NJ, Parobek CM, Juliano JJ, Bailey JA. SeekDeep: single-base resolution de novo clustering for amplicon deep sequencing. Nucleic Acids Research. 2017;46(4):e21-e.
122. Pillai DR, Hijar G, Montoya Y, Marouino W, Ruebush TK, 2nd, Wongsrichanalai C, et al. Lack of prediction of mefloquine and mefloquine-artesunate treatment outcome by mutations in the Plasmodium falciparum multidrug resistance 1 (pfmdr1) gene for P. falciparum malaria in Peru. Am J Trop Med Hyg. 2003;68(1):107-10.
123. Talundzic E, Ravishankar S, Kelley J, Patel D, Plucinski M, Schmedes S, et al. Next-Generation Sequencing and Bioinformatics Protocol for Malaria Drug Resistance Marker Surveillance. Antimicrobial Agents and Chemotherapy. 2018;62(4):e02474-17.
124. Jacob CG, Thuy-Nhien N, Mayxay M, Maude RJ, Quang HH, Hongvanthong B, et al. Genetic surveillance in the Greater Mekong subregion and South Asia to support malaria control and elimination. eLife. 2021;10:e62997.
125. Grande T, Bernasconi A, Erhart A, Gamboa D, Casapia M, Delgado C, et al. A Randomised Controlled Trial to Assess the Efficacy of Dihydroartemisinin-Piperaquine for the Treatment of Uncomplicated Falciparum Malaria in Peru. PloS one. 2007;2(10):e1101.
126. Mangold KA, Manson RU, Koay ES, Stephens L, Regner M, Thomson RB, Jr., et al. Real-time PCR for detection and identification of Plasmodium spp. J Clin Microbiol. 2005;43(5):2435-40.
127. Oyola SO, Ariani CV, Hamilton WL, Kekre M, Amenga-Etego LN, Ghansah A, et al. Whole genome sequencing of Plasmodium falciparum from dried blood spots using selective whole genome amplification. Malaria journal. 2016;15(1):597.
128. MalariaGEN. MalariaGEN Plasmodium falciparum Community Project: MRC Centre for Genomics and Global Health; 2015 [Available from: https://www.malariagen.net/apps/pf/4.0/.
129. Jombart T, Devillard S, Balloux F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet. 2010;11:94.
130. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010.
131. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60.
132. Genome Analysis Toolkit (GATK). 4.0.12.0 ed: Broad Institute; 2018.
133. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012;6(2):80-92.
134. Wingett S, Andrews S. FastQ Screen: A tool for multi-genome mapping and quality control [version 2; peer review: 4 approved]. F1000Research. 2018;7(1338).
135. Kamvar ZN, Tabima JF, Grunwald NJ. Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. PeerJ. 2014;2:e281.
136. Chang HH, Worby CJ, Yeka A, Nankabirwa J, Kamya MR, Staedke SG, et al. THE REAL McCOIL: A method for the concurrent estimation of the complexity of infection and SNP allele frequency for malaria parasites. PLoS Comput Biol. 2017;13(1):e1005348.
137. Tae H, Kim Dy Fau - McCormick J, McCormick J Fau - Settlage RE, Settlage Re Fau - Garner HR, Garner HR. Discretized Gaussian mixture for genotyping of microsatellite loci containing homopolymer runs. (1367-4811 (Electronic)).
138. 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.
139. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;12(1):77.
140. Weir BS, Cockerham CC. Estimating F-Statistics for the Analysis of Population Structure. Evolution. 1984;38(6):1358-70.
141. Hedrick PW. A standardized genetic differentiation measure. Evolution. 2005;59(8):1633-8.
142. Jost L. G(ST) and its relatives do not measure differentiation. Mol Ecol. 2008;17(18):4015-26.

143. Keenan K, McGinnity P, Cross TF, Crozier WW, Prodöhl PA. diveRsity: An R package for the estimation and exploration of population genetics parameters and their associated errors. Methods in Ecology and Evolution. 2013;4(8):782-8.
144. Whitlock MC. G'ST and D do not replace FST. Mol Ecol. 2011;20(6):1083-91.
145. Jombart T. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics. 2008;24(11):1403-5.
146. Jombart T, Ahmed I. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics. 2011;27(21):3070-1.
147. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics. 2004;20(2):289-90.
148. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 2019;35(3):526-8.
149. Argimón S, Abudahab K, Goater RJE, Fedosejev A, Bhai J, Glasner C, et al. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. Microbial Genomics. 2016;2(11).

Supplementary tables and figures



Supplementary Figure S1. Distribution of depth of coverage past filter per amplicon region.

Supplementary table S1. Amplicons with low genotype depth. The *hrp2* and *hrp3* amplicons have lower mean depth in the samples than in the controls most likely due to gene deletions. There is one *ubp1* amplicon that also has poorer performance in samples than controls, possibly due to variations in primer regions in the study sample.

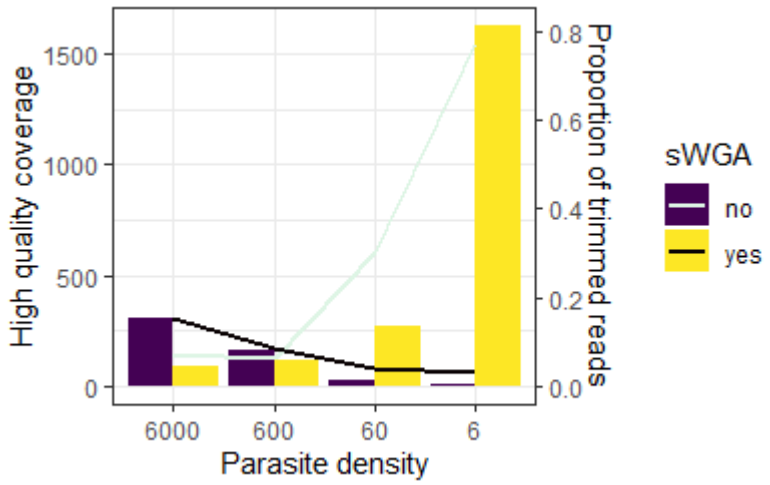
Amplicon_ID	CHR	START	END	POOL	Target	Amplicon name	% Samples with mean depth <10	% controls with mean depth <10
AMPL3594045	Pf3D7_12_v3	2091952	2092317	1	coronin	coronin_1	93%	100%
AMPL3593126	Pf3D7_13_v3	1724575	1724896	1	K13	K13_1	94%	99%
AMPL3594029	Pf3D7_13_v3	2503988	2504341	2	exonuclease	exonuclease_4	94%	99%
AMPL3594038	Pf3D7_12_v3	717766	718035	1	ap2-mu	ap2-mu_1	91%	93%
AMPL3594106	Pf3D7_01_v3	200597	200953	1	ubp1	ubp1_43	95%	93%
AMPL3593973	Pf3D7_07_v3	404305	404638	1	crt	crt_7	93%	92%
AMPL3594093	Pf3D7_01_v3	197290	197637	1	ubp1	ubp1_29	91%	89%
AMPL3592866	Pf3D7_14_v3	293335	293566	2	plasmepsin 2	plasmepsin2_1	83%	85%
AMPL3594092	Pf3D7_01_v3	197119	197440	2	ubp1	ubp1_28	86%	85%
AMPL3593953	Pf3D7_13_v3	1725868	1726167	1	K13	K13_7	78%	77%
AMPL3593061	Pf3D7_08_v3	1374965	1375281	1	hrp2	hrp2_5	83%	72%
AMPL3594112	Pf3D7_01_v3	190943	191243	1	ubp1	ubp1_5	76%	69%
AMPL3593072	Pf3D7_13_v3	2841623	2841818	1	hrp3	hrp3_5	84%	55%
AMPL3593068	Pf3D7_13_v3	2840615	2840970	1	hrp3	hrp3_1	82%	45%
AMPL3593069	Pf3D7_13_v3	2840910	2841262	2	hrp3	hrp3_2	72%	35%

Supplementary table S2. Amplicons with high genotype depth of coverage (>150).

Amplicon_ID	CHR	START	END	POOL	Target	Amplicon name	median DP	25% percentile	75% percentile
AMPL3592823	Pf3D7_08_v3	1374462	1374730	1	hrp2	hrp2_3	268	15	1045
AMPL3593064	Pf3D7_08_v3	1374687	1375025	2	hrp2	hrp2_4	207	9	905
AMPL3593414	Pf_M76611	4421	4659	1	CytB	CytB_5	183	63	329.5
AMPL3593965	Pf3D7_07_v3	404974	405187	2	crt	crt_10	176	59	626
AMPL3593990	Pf3D7_05_v3	960959	961225	2	mdr1	mdr1_14	155	79	344
AMPL3594105	Pf3D7_01_v3	200424	200657	2	ubp1	ubp1_42	161	89.5	208

Supplementary table S3. Amplicons in conserved regions. These amplicons had no variants in the vcf, and in most cases have few variants detected in South America or even global (source: Pf4 - P. falciparum Community Project Data - Variant catalogue, https://www.malariagen.net/apps/pf/4.0/#variation). Only chromosomal variants are included in the Pf4 data-app.

Amplicon_ID	CHR	START	END	Target	Amplicon name	nr. variants in Pf4	nr. variants with NRAF>0 in SAM Pf4
AMPL3594061	Pf3D7_01_v3	467940	468211	mrp1	mrp1_14	9	1
AMPL3594073	Pf3D7_01_v3	468287	468636	mrp1	mrp1_16	5	0
AMPL3594094	Pf3D7_01_v3	197577	197841	ubp1	ubp1_30	22	1
AMPL3594103	Pf3D7_01_v3	199990	200337	ubp1	ubp1_41	22	1
AMPL3594125	Pf3D7_01_v3	194496	194777	ubp1	ubp1_18	39	0
AMPL3594137	Pf3D7_01_v3	198077	198428	ubp1	ubp1_32	38	1
AMPL3594139	Pf3D7_01_v3	198618	198838	ubp1	ubp1_34	11	0
AMPL3593191	Pf3D7_04_v3	749304	749551	dhfr	dhfr_7	0	0
AMPL3593194	Pf3D7_04_v3	749721	749960	dhfr	dhfr_9	0	0
AMPL3593986	Pf3D7_05_v3	960142	960478	mdr1	mdr1_10	24	0
AMPL3593988	Pf3D7_05_v3	960549	960738	mdr1	mdr1_12	20	0
AMPL3593972	Pf3D7_07_v3	404272	404376	crt	crt_6	3	0
AMPL3594014	Pf3D7_08_v3	550217	550357	dhps	dhps_9	3	0
AMPL3593202	Pf3D7_12_v3	2093070	2093321	coronin	coronin_6	10	0
AMPL3594039	Pf3D7_12_v3	717975	718318	ap2-mu	ap2-mu_2	10	0
AMPL3594043	Pf3D7_12_v3	719101	719460	ap2-mu	ap2-mu_7	14	0
AMPL3594044	Pf3D7_12_v3	719401	719727	ap2-mu	ap2-mu_8	12	1
AMPL3592420	Pf3D7_13_v3	2503916	2504048	exonuclease	exonuclease_3	13	0
AMPL3592435	Pf3D7_13_v3	2504281	2504427	exonuclease	exonuclease_5	16	0
AMPL3593127	Pf3D7_13_v3	1724836	1725019	K13	K13_2	13	0
AMPL3593956	Pf3D7_13_v3	1726758	1726897	K13	K13_11	7	0
AMPL3593110	Pf3D7_14_v3	293605	293961	plasmepsin2	plasmepsin2_3	23	0
AMPL3593186	Pf_M76611	4210	4483	CytB	CytB_4	NA	NA



Supplementary figure S2. Effect of selective whole genome amplification (sWGA) on high quality coverage and amount of trimming of a 3D7 serial dilution at different parasite densities (6000 - 6 p/µl) at DNA concentrations mimicking DBS samples. At parasite densities below 60 p/µl sWGA increases the number of high-quality reads and reduces the number of low-quality reads that are trimmed away.

Supplementary table S4. Error rates in 3D7 replicates without and with sWGA in different subsets of loci.

Type of variants	3D7 without sWGA	3D7 with sWGA
All variants in entire target region	0.05% ± 0.01	0.13% ± 0.06
Bi-allelic SNPs only	0.008% ± 0.004	0.03% ± 0.01
Indels only	0.02% ± 0.005	0.04% ± 0.02
“core” region only	0.03% ± 0.01	0.11% ± 0.06
Bi-allelic SNPs in core region only	0.006% ± 0.004	0.02% ± 0.01

Supplementary table S5. Genotyping known variants in previously genotyped controls: MRA 1241, MRA 1251, MRA 1255, MRA 150 (genotypes from literature (72, 92, 102, 103) and samples from Vietnam (104). Several replicates of each samples were tested. NA = no genotype was obtained at this position; wt = wildtype.

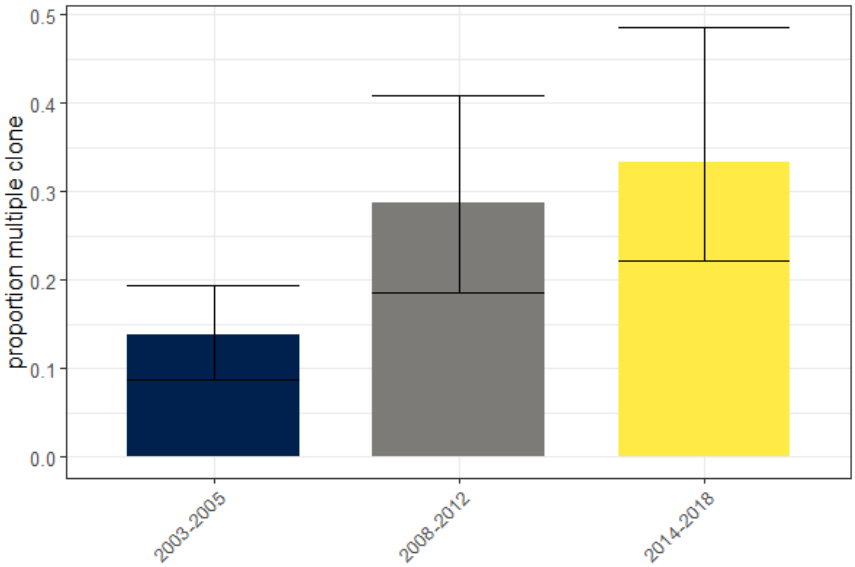
Sample name	PF Ampliseq results					Previous data					error	additional mixed	total calls ampliseq	% error
	Crt: 72-76	*K13*	*mdr1*: 86, 184, 1246	*dhfr:* 51, 59, 108, 164	*dhps:* 436, 437, 540, 581, 613	*crt:* 72-76	*K13*	*mdr1:* 86, 184, 1246	*dhfr*: 51, 59, 108, 164	*Dhps:* 436, 437, 540, 581, 613				
MRA 150	CVIET	wt	YYD	IRNI	F,G,K,A,S	CVIET	wt	YYD	IRNI	FGKAS	0	0	18	0.0%
	NA	wt	YYD	IRNI	F,_,K,A,S						0	0	12	0.0%
	CVIET	wt	NA	NA	NA						0	0	6	0.0%
MRA 1241	NA	I543T	NYD	IRNL	S/F,G/A,E,A,S	CVIET	I543T	NYD	IRNL	FAEAS	0	2	13	15.4%
	NA	I543T	N_D	__NI/__NL	_,_,E,A,S						0	1	8	12.5%
	CVIET	I543T	N_	IR__	S/F,G/A,E,A,A/S						0	3	15	20.0%
MRA 1251	CVIET	C580Y	NYD	IRNI	A/S,G,E,A,A	CVIET	C580Y	NYD	IRNI	AGEAA	0	1	18	5.6%
	NA	NA at pos 580	N_D	__NI	_,_,E,A,A						0	0	7	0.0%
	CVIET	NA at pos 580	N_	IR__	NA						0	0	8	0.0%
MRA 1255	CVIET	NA at pos 539	YYD	IRNI	F,G,K,A,S	CVIET	R539T	YYD	IRNI	FGKAS	0	0	17	0.0%
	NA	wt	YYD	IRNI	_,_,K,A,S						1	0	11	9.1%
	CVIET	NA at pos 539	NA	IR__/NC__	NA						0	2	7	28.6%
VTN 1	CVIET	T511H	NFD	IRNI	A,G,K/E,A,A	CVIET	Y511H	NFD	IRNI	-GKAA	0	1	18	0.0%
	NA	T511+K189T	NFD	NCNI	S,G/A,K,_,A						3	1	13	23.1%
VTN 2	CVIDT/CVIET	C580Y/I543T	NYD	IRNI/IRNL	A/S/F,G,E,A,A/S	CVIDT	C580Y	NYD	IRNI	AGEAA	0	4	19	0.0%
	CVIDT/CVIET	het I543T	NYD	IRNI/IRNL	F,G,E,A,A/S						1	2	18	5.6%
VTN 3	CVIET	C580Y/I543T	NYD	IRNI/IRNL	A/S/F,G,K/N,A/G,A/S	CVIET	C580Y	_F_	IRNL	S_NG_	1	3	19	5.3%
VTN 4	CVIET	C580Y	NFD	IRNL	S,G,N,G,A	CVIET	C580Y	_F_	IRNL	S_NG_	0	0	17	0.0%
	NA	C580Y	_FD	IR_L	S,G,-,-,-						0	0	8	0.0%
											6	**20**	**216**	**2.8%**

Supplementary table S6. Detection limit of variant loci in 3D7-Dd2 mock samples

proportion 3D7 DNA	proportion Dd2 DNA	nr of heterozygote SNP genotypes	nr of homozygote reference SNP genotypes	Proportion variant alleles detected
50%	50%	63	4	94.0%
80%	20%	28	36	43.8%
95%	5%	3	65	4.4%
99%	1%	0	68	0%
99.5%	0.5%	0	69	0%

Supplementary table S7. Complexity of infection. COI determined with McCOIL algorithms (categorical and proportional) with different subsets of biallelic variants: 1) all biallelic variants; 2) all variants (core variants) excluding *hrp2*, MS regions and mitochondrial and apicoplast variants; and 3) the 28-SNP barcode variants. Because of the large differences observed between the two McCOIL methods, we estimated the proportions of single and multiple clone infections with an additional methods based on the number of heterozygous variants in 1) the 28-SNP barcode, 2) *ama1*, 3) core variants, and 4) MS targeted regions. The mode (most frequent value) from four measurements for single *vs.* multiple clone (heterozygotes in 1) MS, 2) *ama1* and 3) barcode regions and 4) McCOIL proportional barcode) was determined. The top diagonal of the pairwise comparisons shows the Cohen's Kappa values for agreement, values >0.4 are in bold and represent fair to good agreement beyond chance. The bottom diagonal shows the area-under-the-curve for the ROC, as a measure of pairwise classification performance using one of the methods as reference (column variable as reference/row variable as reference), with values >0.6 in bold.

		COI=1	COI>1	% single clone	95% CI	McCOIL categorical			McCOIL proportional			Heterozygous variants				**Mode**
						all biallelic	core	28-SNP barcode	all biallelic	core	28-SNP barcode	MS from NGS	28-SNP barcode	SNPs in AMA1	SNPs in core (>5 SNPs)	
McCOIL categorical	all biallelic variants	208	0	100%		X	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	core variants	166	1	99.4%	99-100%	NA	X	**0.664**	0	0.002	0.065	0.009	-0.012	-0.012	-0.012	0.036
	28-SNP barcode variants.	192	2	99.0%	98-100%	NA	**0.997 / 0.750**	X	0	0.002	0.065	0.008	0.037	-0.020	-0.020	0.033
McCOIL proportional	all biallelic variants	3	64	4.5%	1.5-9.6%	NA	NA / 0.5	NA/ 0.5	X	**0.646**	0.014	-0.028	0.014	0.020	-0.049	0.024
	core variants	12	112	9.7%	6-15%	NA	0.549 / 0.504	0.549/ 0.504	**0.977 / 0.750**	X	0.033	-0.102	-0.003	-0.013	-0.053	-0.017
	28-SNP barcode variants.	123	25	83.1%	78-89%	NA	**0.918** / 0.520	**0.918** / 0.520	0.570 / 0.526	0.576 / 0.556	X	0.037	0.254	-0.079	0.095	0.331
Heterozygous variants	MS from NGS	77	121	38.9%	32-46%	NA	**0.705** / 0.506	**0.701** / 0.505	0.425 / 0.487	0. 366 / 0.454	0.538 / 0.521	X	0.050	0.007	0.092	0.302
	28-SNP barcode	213	41	83.9%	80-88%	NA	0.419 / 0.496	**0.664** / 0.512	0.570 / 0.526	0.493 / 0.495	**0.623 / 0.632**	0.531 / 0.560	X	0.116	0.116	**0.515**
	SNPs in AMA1	216	38	85.0%	81-89%	NA	0.410 / 0.496	0.411 / 0.494	0.594 / 0.527	0.473 / 0.485	0.458 / 0.462	0.504 / 0.506	0.559 / 0.557	X	**0.496**	**0.511**
	SNPs in core (>5 SNPs)	85	169	71.3%	66-77%	NA	0.352 / 0.496	0.352 / 0.493	**0.701** / 0.458	0.598 / 0.461	0.562 / 0.542	0.553 / 0.560	**0.692 / 0.627**	**0.817 / 0.718**	X	**0.458**
Mode		191	51	78.9%	74-84%	NA	**0.877** / 0.512	**0.881** / 0.511	**0.609** / 0.528	0.467 / 0.486	**0. 702 / 0.646**	**0.677 / 0.729**	**0.817** / **0.724**	**0.787 / 0.733**	**0.711 / 0.757**	X

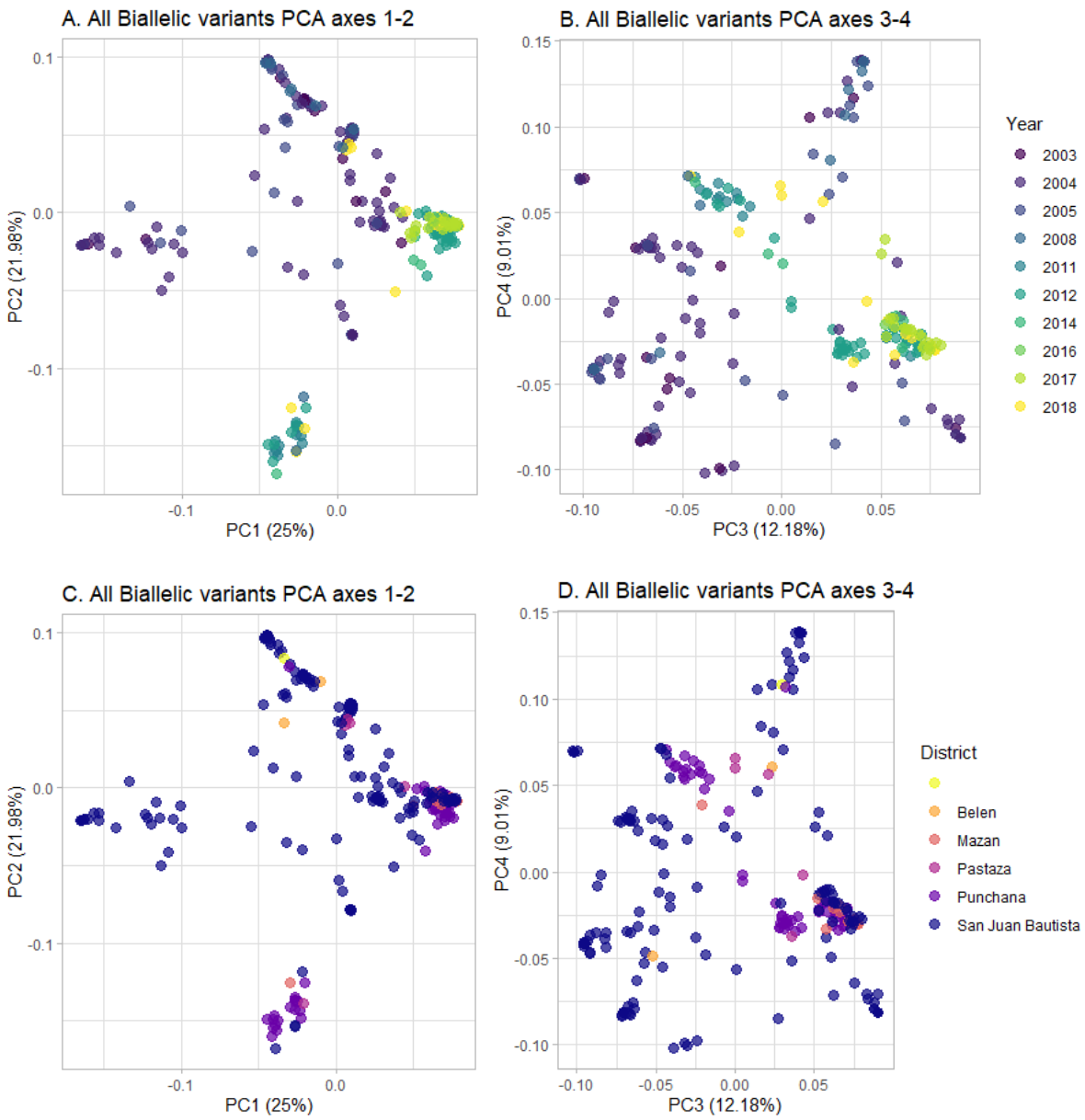


Supplementary Figure S3. Proportion multiple clone infections in Peru. The proportion of multiple clone infections (with 95% confidence interval) was plotted for three time periods, and was higher in 2008-2018 than in 2003-2005 (p = 0.0005, X^2). Multiple clone infections determined as mode of the different approaches.

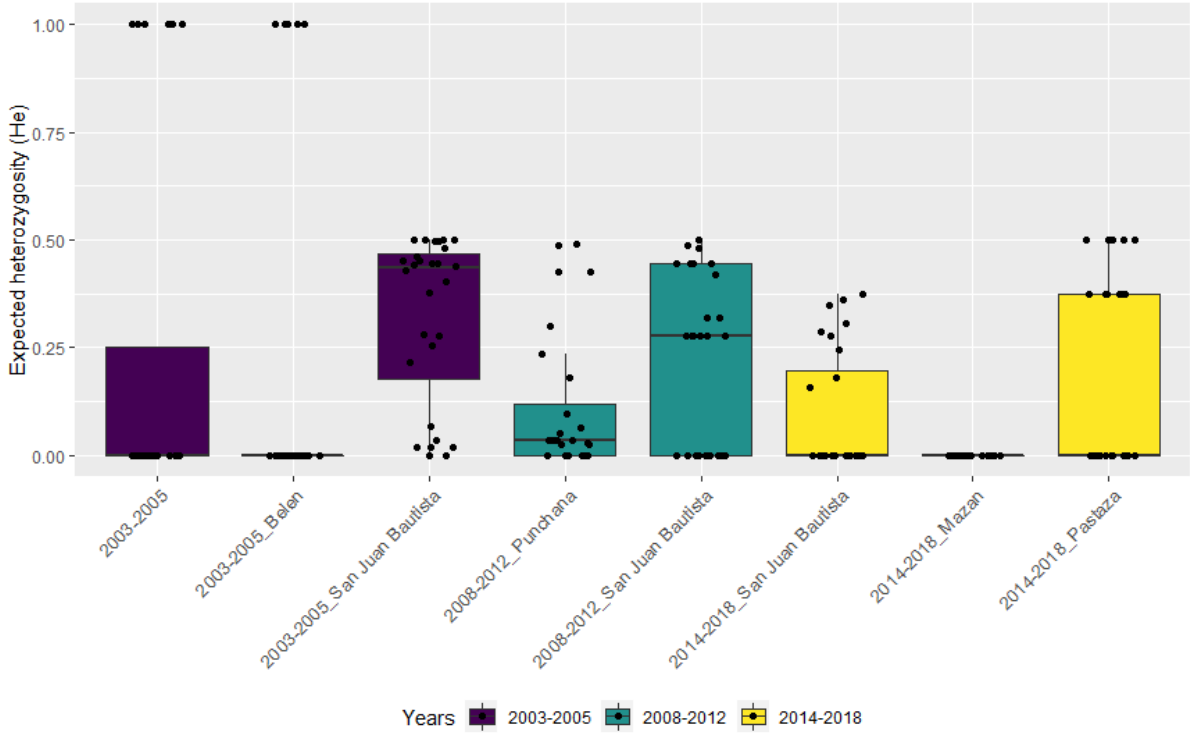
Supplementary Table S8 Barcode alleles (position and type), alleles annotated in plasmoDB (https://plasmodb.org/) with major allele frequency. In addition allele frequenchies (AF) for the reference allele (REF) and alternate allele (ALT) in study samples (n = 254) from Peru.

			plasmoDB				Samples (n=254)		
CHROM	POS	TYPE	Reference allele	major allele	Major AF	Minor allele (if not ref)	REF AF	ALT AF	major allele
Pf3D7_01_v3	205066	SNP	G	G	0.66	A	0.81	0.19	G
Pf3D7_01_v3	339436	SNP	A	G	0.59		0.15	0.85	G
Pf3D7_02_v3	519457	SNP	C	T	0.68		0.61	0.39	C
Pf3D7_02_v3	694307	SNP	A	A	0.54	G	0.03	0.97	G
Pf3D7_03_v3	361199	SNP	C	C	0.65	T	0.68	0.32	C
Pf3D7_03_v3	849476	SNP	C	C	0.64		0.99	0.01	C
Pf3D7_04_v3	691961	SNP	C	T	0.53	T	0.57	0.43	C
Pf3D7_04_v3	770292	SNP	A	A	0.74	G	0.77	0.23	A
Pf3D7_05_v3	921893	SNP	T	T	0.58	A	1.00	0.00	T
Pf3D7_05_v3	1188394	SNP	C	C	0.84	T	0.79	0.21	C
Pf3D7_06_v3	148827	SNP	T	T	0.8	C	0.64	0.36	T
Pf3D7_06_v3	636044	SNP	A	C	0.58		0.67	0.33	A
Pf3D7_07_v3	455494	SNP	A	A	0.57		0.79	0.21	A
Pf3D7_07_v3	782111	SNP	T	T	0.92	G	0.91	0.09	T
Pf3D7_08_v3	501042	SNP	T	T	0.56	C	0.27	0.73	C
Pf3D7_08_v3	803172	SNP	T	T	0.82	G	0.10	0.90	G
Pf3D7_09_v3	231065	SNP	C	A	0.5		0.80	0.20	C
Pf3D7_09_v3	1005351	SNP	G	G	0.65	C	0.49	0.51	C
Pf3D7_10_v3	341106	SNP	A	A	0.54	G	0.01	0.99	G
Pf3D7_10_v3	1172712	SNP	A	G	0.65	G	0.25	0.75	G
Pf3D7_11_v3	874948	SNP	G	A	0.7	A	0.43	0.57	A
Pf3D7_11_v3	1505533	SNP	T	T	0.92	C	0.98	0.02	T
Pf3D7_12_v3	1127000	INDEL	T				0.30	0.70	TA
Pf3D7_12_v3	1552084	SNP	T	C	0.86	C	0.29	0.71	C
Pf3D7_13_v3	1595988	SNP	T	C	0.54	C	0.68	0.32	T
Pf3D7_13_v3	1827569	SNP	T	T	0.62	A	0.50	0.50	T
Pf3D7_14_v3	832594	SNP	T	T	0.87	C	0.61	0.39	T
Pf3D7_14_v3	1381943	SNP	T	T	0.94	C	0.01	0.99	C

Supplementary Figure S4. Principal component analysis of samples collected in Peru between 2003 and 2018. PCA is shown all biallelic loci in the core region along the first 2 principal components (A) and 3rd and 4th PCs (B). Isolates are colored by year (A& B) and by district (C &D), and from earlier years (blue and purple colors) are more diverse than later isolates (greens & yellows), which form two clusters. All samples with unknown district were collected in the rural communities south of Iquitos.



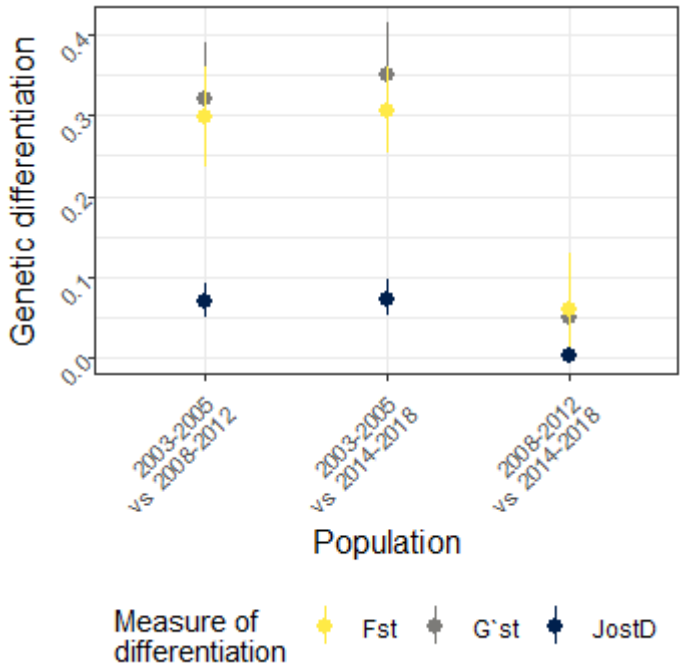
Supplementary Figure S5. Expected Heterozygosity by year group and district. Number of individuals for each population: $n_{2003\text{-}2005} = 1$, $n_{2003\text{-}2005_Belen} = 1$, $n_{2003\text{-}2005_San\ Juan\ Bautista} = 116$, $n_{2008\text{-}2012_Punchana} = 59$, $n_{2008\text{-}2012_San\ Juan\ Bautista} = 6$, $n_{2014\text{-}2018_San\ Juan\ Bautista} = 24$, $n_{2014\text{-}2018_Mazan} = 10$, $n_{2014\text{-}2018_Pastaza} = 4$.



Supplementary table S9. p-values for Pairwise comparisons of *He* using Wilcoxon rank sum test with Benjamini-Hochberg correction for multiple testing. Significant p-values (<0.05) are indicated in bold.

	2003-2005 unspecified	2003-2005 Belen	2003-2005 San Juan Bautista	2008-2012 Punchana	2008-2012 San Juan Bautista	2014-2018 San Juan Bautista	2014-2018 Mazan
2003-2005 Belen	0.61	-	-	-	-	-	-
2003-2005 San Juan Bautista	**0.009**	**>0.001**	-	-	-	-	-
2008-2012 Punchana	0.117	**0.018**	**0.0010**	-	-	-	-
2008-2012 San Juan Bautista	0.20	**0.039**	**0.045**	0.24	-	-	-
2014-2018 San Juan Bautista	0.95	0.58	**>0.001**	0.12	**0.018**	-	-
2014-2018 Mazan	**0.015**	**0.039**	**>0.001**	**>0.001**	**>0.001**	**0.0046**	-
2014-2018 Pastaza	0.66	0.23	**0.034**	0.93	0.72	0.12	**>0.001**

Supplementary Figure S6. Genetic differentiation between the three populations measured as Fst (Weir & Cockerham, 1984), $G`_{ST}$ (Hedrick, 2005) and Jost's D (Jost 2008) using the R package diveRsity. Number of individuals for each population: $n_{2003\text{-}2005}$ = 118; $n_{2008\text{-}2012}$ = 65; $n_{2014\text{-}2018}$ = 38.



Supplementary Table S10. 28-SNP barcode loci that become fixed over time

Chromosome	Position	Fixed in
Pf3D7_02_v3	519457	2014-2018
Pf3D7_02_v3	694307	2014-2018
Pf3D7_03_v3	849476	2008-2018
Pf3D7_04_v3	691961	2014-2018
Pf3D7_05_v3	921893	2003-2018
Pf3D7_06_v3	636044	2014-2018
Pf3D7_07_v3	455494	2014-2018
Pf3D7_07_v3	782111	2008-2018
Pf3D7_08_v3	803172	2008-2018
Pf3D7_10_v3	341106	2003-2018
Pf3D7_11_v3	1505533	2008-2018
Pf3D7_12_v3	1127000	2014-2018
Pf3D7_12_v3	1552084	2008-2018
Pf3D7_14_v3	1381943	2008-2018

Supplementary Table S11 LD at district level.

years	district	n	rbarD	p-value
2003-2005	San Juan Bautista	116	0.15	0.001
2008-2012	San Juan Bautista	6	0.38	0.001
2008_2012	Punchana	59	0.222	0.001
2014-2018	San Juan Bautista	24	0.408	0.001

other populations too small or containing too much missing loci

Supplementary Table S12 Contributions of SNPs to DAPC.

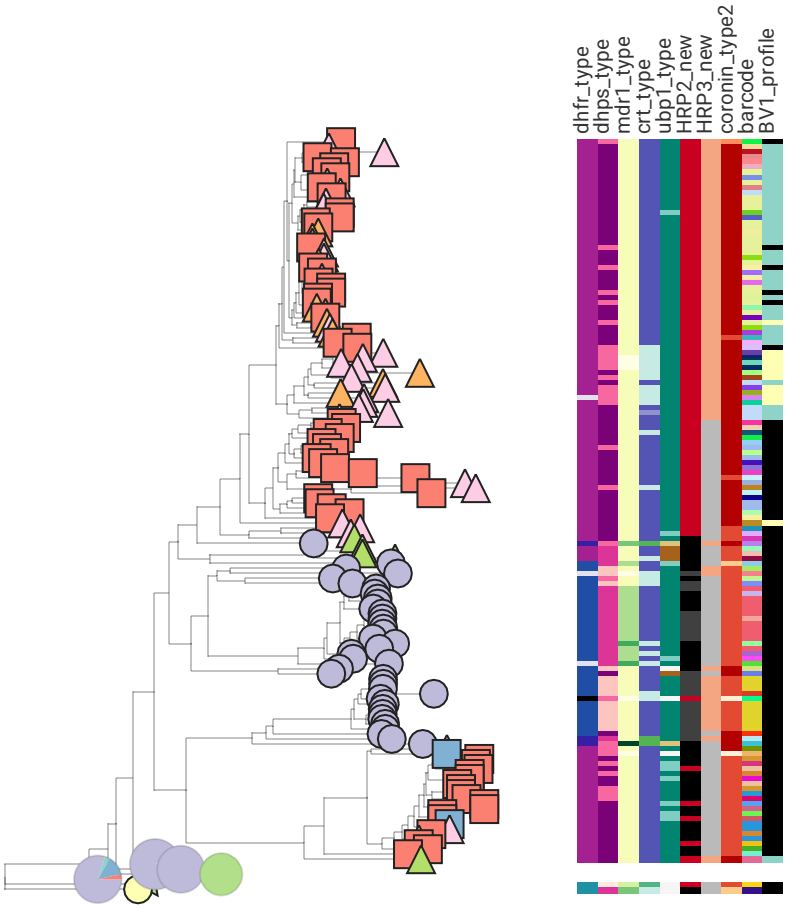
variant position	contributing factor	PC in DAPC	annotation of alt allele	gene ID	gene name	28-SNP barcode
Pf3D7_01_v3_180379	0.009194983					
Pf3D7_01_v3_192590	0.013224389	2	Lys774Asn	PF3D7_0104300	ubp1	
Pf3D7_01_v3_196974	0.011249826	2	Leu2236Leu	PF3D7_0104300	ubp1	
Pf3D7_01_v3_199237	0.008442663	2	Lys2912Asn	PF3D7_0104300	ubp1	
Pf3D7_01_v3_205066	0.007943108	2	Phe130Phe	PF3D7_0104500	unknown protein	Yes
Pf3D7_01_v3_339432	0.035879398	1	Glu755Glu	PF3D7_0108300	unknown protein	in barcode amplicon
Pf3D7_02_v3_519457	0.025872597	2	Ile3228Ile	PF3D7_0212500	unknown protein	in barcode amplicon
Pf3D7_03_v3_361195	0.007454305	2	Arg1626Lys	PF3D7_0308100	zinc finger protein, putative	in barcode amplicon
Pf3D7_03_v3_361199	0.008475745	2	Thr1627Thr	PF3D7_0308100	zinc finger protein, putative	Yes
Pf3D7_04_v3_748235	0.0860365	1	Cys50Arg	PF3D7_0417200	pfdhfr	
Pf3D7_04_v3_748235	0.009881925	2	Cys50Arg	PF3D7_0417200	pfdhfr	
Pf3D7_04_v3_748239	0.01967936	1	Asn51Ile	PF3D7_0417200	pfdhfr	
Pf3D7_04_v3_748239	0.009615098	2	Asn51Ile	PF3D7_0417200	pfdhfr	
Pf3D7_04_v3_748577	0.02383201	1	Ile164Leu	PF3D7_0417200	pfdhfr	
Pf3D7_04_v3_770292	0.032732908	2	Lys4667Glu	PF3D7_0417400	unknown protein	Yes
Pf3D7_05_v3_1188491	0.032529119	1	Phe152Phe	PF3D7_0529000	unknown protein	in barcode amplicon
Pf3D7_05_v3_960989	0.01175819	1	Ser1034Cys	PF3D7_0523000	pfmdr1	
Pf3D7_05_v3_960989	0.009963482	2	Ser1034Cys	PF3D7_0523000	pfmdr1	
Pf3D7_05_v3_961625	0.031696855	1	Asp1246Tyr	PF3D7_0523000	pfmdr1	
Pf3D7_06_v3_148827	0.010701516	1	Val39Val	PF3D7_0603600	AT-rich interactive domain-containing protein, putative	Yes
Pf3D7_06_v3_148827	0.017232716	2	Val39Val	PF3D7_0603600	AT-rich interactive domain-	Yes

					containing protein, putative	
Pf3D7_06_v3_636044	0.046900879	2	Ile1310Met	PF3D7_0615400	ribonuclease, putative	Yes
Pf3D7_07_v3_455494	0.007594591	2	Asn927Asn	PF3D7_0710100	unknown protein	Yes
Pf3D7_07_v3_455550	0.007107321	2	Ile909Leu	PF3D7_0710100	unknown protein	in barcode amplicon
Pf3D7_08_v3_501042	0.018331353	2	Glu176Glu	PF3D7_0809700	RuvB-like helicase 1	Yes
Pf3D7_08_v3_501054	0.017953778	2	Val172Val	PF3D7_0809700	RuvB-like helicase 1	in barcode amplicon
Pf3D7_08_v3_549993	0.024643639	1	Lys540Glu	PF3D7_0810800	pfdhps	
Pf3D7_08_v3_549993	0.007753441	2	Lys540Glu	PF3D7_0810800	pfdhps	
Pf3D7_09_v3_1005351	0.036558277	2	Glu1919Gln	PF3D7_0924600	unknown protein	Yes
Pf3D7_11_v3_874948	0.012525102	2	Gly973Asp	PF3D7_1122800	calcium-dependent protein kinase 6	Yes
Pf3D7_12_v3_2092606	0.013989583	2	Val62Met	PF3D7_1251200	pfcoronin	
Pf3D7_12_v3_2093692	0.014452113	2	Val424Ile	PF3D7_1251200	pfcoronin	
Pf3D7_12_v3_2094242	0.00941445	1	downstream variant	PF3D7_1251200	pfcoronin	
Pf3D7_13_v3_1827569	0.016409733	1	Tyr345Asn	PF3D7_1345600	inner membrane complex protein	Yes
Pf3D7_13_v3_1827569	0.010051953	2	Tyr345Asn	PF3D7_1345600	inner membrane complex protein	Yes
Pf3D7_14_v3_294796	0.022437109	2	Gln442His	PF3D7_1408000	plasmepsin II	

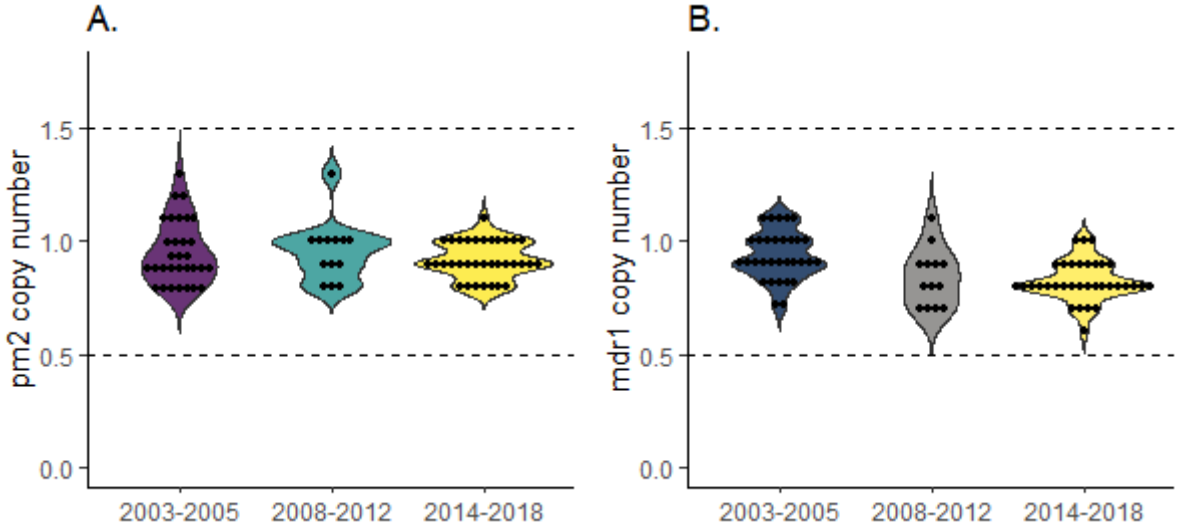
Supplementary table S13. Pairwise comparison of *hrp2/hrp3* classification by PCR and Pf AmpliSeq in study samples tested with both methods (n = 10). PCR genotypes from Gamboa *et al.* 2010 (2).

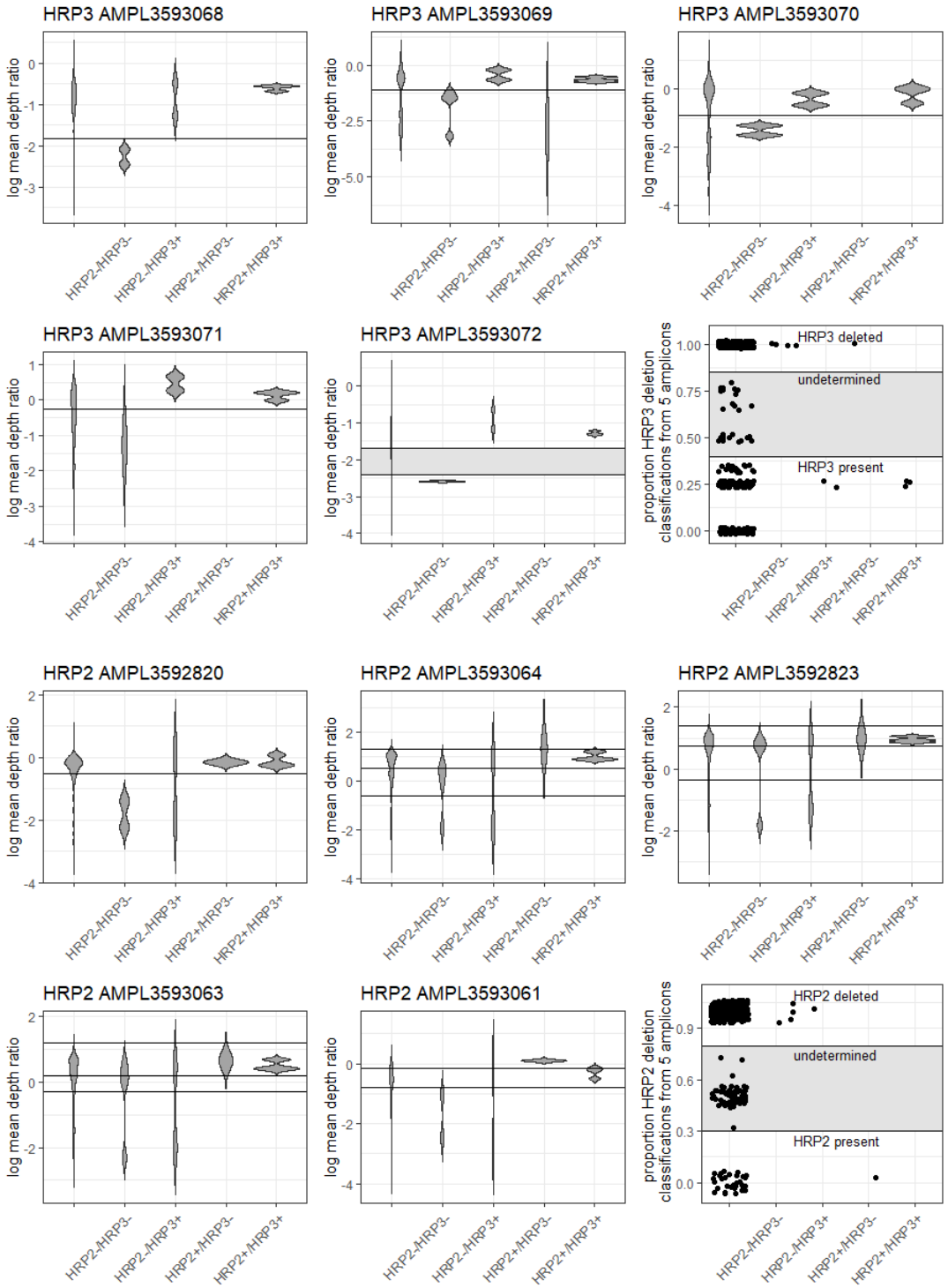
PCR	Pf AmpliSeq			
	hrp2+/ hrp3-	*hrp2- /hrp3-*	*hrp2- /hrp3+*	*hrp2* NA/ *hrp3+*
hrp2+/hrp3-	1			
hrp2-/hrp3-		4		
hrp2+/hrp3+				3
hrp2-/hrp3+			1	1

Supplementary figure S7. Neighbor-joining tree of genetic distances, depicted sub-tree of BV_1-like lineages colored by haplotypes of drug resistance associated markers, *hrp2* & *hrp3* deletions, and 28-SNP barcode haplotypes.



Supplementary figure S8. Copy number variations in A) *plasmepsin II* gene (*pm2*) and B) *multidrug resistance gene 1* (*mdr1*) in a subset of samples from Peru collected between 2003-2018. Samples with copy numbers between 0.5 - 1.5 (dotted lines) relative to 3D7 are considered to have single copies of the respective genes. Sample sizes: $n_{2003\text{-}2005} = 31$, $n_{2008\text{-}2012} = 13$, $n_{2014\text{-}2018} = 34$.





Supplementary figure S9. Distributions of log mean depth ratio's for all samples for each *hrp3* amplicon (AMPL3593072, AMPL3593071, AMPL3593070, AMPL3593069, AMPL3593068) and *hrp2* amplicon (AMPL3592820, AMPL3593064, AMPL3592823, AMPL3593063, AMPL3593061), plotted by *hrp2*/*hrp3* PCR results (not tested, *hrp2-*/*hrp3-*, *hrp2-*/*hrp3+*, *hrp2+*/*hrp3-*, *hrp2+*/*hrp3+*), with thresholds used to define deletions or presence of the genes (Supp. Table 2). The proportion of deletions classified for the amplicons was determined for each samples and below a certain threshold (0.4 for *hrp3* and 0.3 for HPR2)

Supplementary table S14. Cutoff thresholds for *hrp2* and *hrp3* determination of deletions for each amplicon.

Gene	Amplicon	Threshold log mean depth ratio	Result	
hrp3	AMPL3593072	<= -2.4	*hrp3-*	
	AMPL3593072	>= -1.7	*hrp3+*	
	AMPL3593071	<= -0.25	*hrp3-*	
	AMPL3593070	< -0.9	*hrp3-*	
	AMPL3593070	>= -0.9	*hrp3+*	
	AMPL3593069	< -1.10	*hrp3-*	
	AMPL3593068	< -1.82	*hrp3-*	
	AMPL3593068	>= -1.82	*hrp3+*	
hrp2	AMPL3592820	< -0.50	*hrp2-*	
	AMPL3592820	>= -0.50	& *hrp3-*	*hrp2+*
			& *hrp3+*	undetermined
	AMPL3593064	< -0.60	*hrp2-*	
	AMPL3593064	< 0.5	& *hrp3+*	*hrp2-*
	AMPL3593064	>1.30	& *hrp3-*	*hrp2+*
	AMPL3592823	<-0.35	*hrp2-*	
	AMPL3592823	<0.75	& *hrp3+*	*hrp2-*
	AMPL3592823	>1.40	& *hrp3-*	*hrp2+*
	AMPL3593063	< -0.3	*hrp2-*	
	AMPL3593063	< 0.2	& *hrp3+*	*hrp2-*
	AMPL3593063	>1.2	& *hrp3-*	*hrp2+*
	AMPL3593061	< -0.15	& *hrp3-*	*hrp2-*
	AMPL3593061	>= -0.15	& *hrp3-*	*hrp2+*
	AMPL3593061	< -0.8	& *hrp3+*	*hrp2-*
	AMPL3593062	Not used; no discriminatory power		

Supplementary methods - SNP Barcode selection

Introduction

This material accompanies the article "Novel highly multiplexed targeted NGS assay for molecular surveillance of *P. falciparum* reveals selection on drug and diagnostic resistance associated genes in the Peruvian Amazon from 2003 to 2018." by J. H. Kattenberg et al. submitted for publication. The paper presents the design and validation of a *P. falciparum* Ampliseq assay for the purpose of molecular surveillance of *P. falciparum* in Peru. Included in this PF ampliseq assay is a SNP barcode which was designed to monitor parasite strains circulating in Peru over space and time. This document presents the details of how the SNPs were selected for this barcode.

Online whole genome data from the MalariaGEN *Plasmodium falciparum* Community Project (Catalogue of genetic variation v4.0 (2015) (1) and 2016 data release (2)) were used for population genetic analyses, resulting in a final selection of 28 SNP for a genetic barcode for *P. falciparum* parasites in South America, and Peru in particular. These SNPs were common within the Peruvian *P. falciparum* samples in the WGS dataset (*i.e.* showed a high minor allele frequency (>0.35)) and differentiated these samples from other populations in the dataset (using discriminant analysis of principle components). Moreover, they were broadly distributed across the *P. falciparum* genome and were not under selective pressure from parasite environmental factors, like drug exposure or host immunity.

Detailed methods and results

As a first selection, the MalariaGEN Plasmodium falciparum Community Project Catalogue of genetic variation v4.0 (2015) was used with the online data app (900.000+ high quality SNPs) to select SNPs with minor allele frequency (MAF) in South America ranging between 0.35-0.5, resulting in 1880 selected SNPs. Subsequently, these 1880 loci in the MalariaGEN *Plasmodium falciparum* Community Project 2016 data release (2), were investigated for heterozygous genotypes at these loci in all 3394 samples. Loci that were heterozygous in one or more of the 7 Peruvian samples in the database were removed, resulting in a selection of 1778 SNPs. With all samples (from all countries) that had homozygous genotype calls at these loci (N=338) we proceeded with examining country level population structure using discriminant analysis of principal components (DAPC with the adegenet package in R (3). On a per chromosome basis, the contribution of each SNP to the first component of the DAPC (i.e. allele loadings) were scored and sorted. DAPC was performed using countries as populations for all countries (150 principal components retained and 5 discriminants), as well as a subset analysis with South American countries only (Peru and Colombia; 5 principal components retained and 1 discriminant). For the top contributing alleles (Supplementary file 4), pairwise linkage disequilibrium (LD) between selected SNPs was calculated using the R package poppr in R and a selection of 4-13 SNPs/chromosome was made with the lowest LD. Next pairwise LD was examined between the selected SNPs at all chromosomes, and any known antigens or genes that could potentially be under selection (*e.g.* exposed on outer membrane) were removed from the list. Finally, 2 SNPs per chromosome were selected, with priority given for synonymous SNPs with low pairwise LD, resulting in a barcode of 28 SNPs (Table 1).

Table 1. Selected SNP positions and annotated gene location in the final barcode. *Syn = synonymous mutation, Non-syn = non-synonymous mutation.*

SNP ID	chr	position	gene	type
Pf3D7_01_205066	1	205066	"conserved Plasmodium protein, unknown function"	Syn
Pf3D7_01_339436	1	339436	"conserved Plasmodium protein, unknown function"	Non-syn
Pf3D7_02_694307	2	694307	"conserved Plasmodium protein, unknown function"	Syn
Pf3D7_02_519457	2	519457	conserved Plasmodium protein, unknown function"	Syn
Pf3D7_03_361199	3	361199	"conserved Plasmodium protein, unknown function"	Syn
Pf3D7_03_849476	3	849476	oocyst capsule protein (Cap380)	Non-syn
Pf3D7_04_691961	4	691961	"conserved Plasmodium protein, unknown function"	Non-syn
Pf3D7_04_770292	4	770292	"conserved Plasmodium protein, unknown function"	Non-syn
Pf3D7_05_1188394	5	1188394	"conserved Plasmodium protein, unknown function"	Non-syn
Pf3D7_05_921893	5	921893	"conserved Plasmodium protein, unknown function"	Non-syn
Pf3D7_06_148827	6	148827	AT-rich interactive domain-containing protein, putative	Syn
Pf3D7_06_636044	6	636044	"ribonuclease, putative"	Non-syn
Pf3D7_07_455494	7	455494	conserved Plasmodium protein, unknown function"	Syn
Pf3D7_07_782111	7	782111	"dynein heavy chain, putative"	Non-syn
Pf3D7_08_501042	8	501042	"RuvB DNA helicase, putative"	Syn
Pf3D7_08_803172	8	803172	"conserved Plasmodium protein, unknown function"	Non-syn
Pf3D7_09_1005351	9	1005351	"conserved Plasmodium protein, unknown function"	Non-syn
Pf3D7_09_231065	9	231065	"Cu2++-transporting ATPase, putative (CUP)"	Non-syn
Pf3D7_10_341106	10	341106	"conserved Plasmodium protein, unknown function"	Syn
Pf3D7_10_1172712	10	1172712	partial CSTF domain-containing protein, putative	non-syn
Pf3D7_11_874948	11	874948	calcium dependent protein kinase 6+(CDPK6)	non-syn
Pf3D7_11_1505533	11	1505533	guanylyl cyclase (GCalpha)	non-syn
Pf3D7_12_1127001	12	1127001	"cyclin related protein, putative"	non-syn
Pf3D7_12_1552084	12	1552084	"conserved Plasmodium protein, unknown function"	Syn
Pf3D7_13_1595988	13	1595988	"conserved Plasmodium protein, unknown function"	Syn
Pf3D7_13_1827569	13	1827569	inner membrane complex protein	non-syn
Pf3D7_14_832594	14	832594	"conserved Plasmodium protein, unknown function"	non-syn
Pf3D7_14_1381943	14	1381943	"conserved Plasmodium protein, unknown function"	non-syn

References supplementary file

1. MalariaGEN. MalariaGEN Plasmodium falciparum Community Project: MRC Centre for Genomics and Global Health; 2015 [Available from: https://www.malariagen.net/apps/pf/4.0/.
2. Malaria GENPfCP. Genomic epidemiology of artemisinin resistant malaria. Elife. 2016;5.
3. Jombart T, Devillard S, Balloux F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet. 2010;11:94.