Resolving drug selection and migration in an inbred South American *Plasmodium falciparum* population with identity-by-descent analysis

Manuela Carrasquilla[1,2,*,#], Angela M Early[1,2*,#], Aimee R Taylor[2,3], Angélica Knudson[4], Diego F Echeverry[5,6], Timothy JC Anderson[7], Elvira Mancilla[8], Samanda Aponte[9], Pablo Cárdenas[10], Caroline O Buckee[3], Julian C Rayner[11,12], Fabián E Sáenz[13], Daniel E Neafsey[1,2,#], Vladimir Corredor[9,#]

1. Department of Immunology and Infectious Disease, Harvard T.H.Chan School of Public Health, Boston, MA, USA
2. Infectious Disease and Microbiome Program, Broad Institute of MIT and Harvard, Cambridge, MA, USA
3. Center for Communicable Disease Dynamics, Harvard T.H.Chan School of Public Health, Boston, MA, USA
4. Departamento de Microbiología, Facultad de Medicina, Universidad Nacional de Colombia, Bogotá, Colombia
5. Departamento de Microbiología, Facultad de Salud, Universidad del Valle, Cali, Colombia
6. Centro Internacional de Entrenamiento e Investigaciones Médicas (CIDEIM), Cali, Colombia
7. Program in Disease Intervention and Prevention, Texas Biomedical Research Institution, San Antonio, Texas
8. Secretaría Departamental de Salud del Cauca, Popayán, Colombia
9. Departamento de Salud Pública, Facultad de Medicina, Universidad Nacional de Colombia, Bogotá, Colombia
10. Department of Biological Engineering, Massachusetts Institute of Technology
11. Wellcome Sanger Institute, Hinxton, Cambridge, UK
12. Cambridge Institute for Medical Research, University of Cambridge, Cambridge, UK
13. Centro de Investigación para la Salud en América Latina, Facultad de Ciencias Exactas y Naturales, Pontificia Universidad Católica del Ecuador, Quito, Ecuador

*Authors contributed equally
#Corresponding authors:
carrasquilla@mpiib-berlin.mpg.de
early@broadinstitute.org
neafsey@hsph.harvard.edu
vcorredore@unal.edu.co

Abstract

The human malaria parasite *Plasmodium falciparum* is globally widespread, but its prevalence varies significantly between and even within countries. Most population genetic studies in *P. falciparum* focus on regions of high transmission where parasite populations are large and genetically diverse, such as sub-Saharan Africa. Understanding population dynamics in low transmission settings, however, is of particular importance as these are often where drug resistance first evolves. Here, we use the Pacific Coast of Colombia and Ecuador as a model for understanding the population structure and evolution of *Plasmodium* parasites in small populations harboring low genetic diversity. The combination of low transmission and a high proportion of monoclonal infections means there are few outcrossing events and clonal lineages persist for long periods of time. Yet despite this, the population is evolutionarily labile and has successfully adapted to multiple drug regimes. Using 166 newly sequenced whole genomes, we measure relatedness between parasites, calculated as identity by descent (IBD), and find 17 distinct but highly related clonal lineages, six of which have persisted in the region for at least a decade. This inbred population structure is captured in more detail with IBD than with other common population structure analyses like PCA, ADMIXTURE, and distance-based trees. We additionally use patterns of intra-chromosomal IBD and an analysis of haplotypic variation to explore the role of recombination in spreading drug resistance mutations throughout the region. Two genes associated with chloroquine resistance, *crt* and *aat1*, show evidence of hard selective sweeps, while selection appears soft and/or incomplete at three other key resistance loci (*dhps*, *mdr1*, and *dhfr*). Overall, this work highlights the strength of IBD analyses for studying parasite population structure and resistance evolution in regions of low transmission, and emphasizes that drug resistance can evolve and spread in extremely small populations, as will occur in any region nearing malaria elimination.

Introduction

Plasmodium falciparum accounts for the vast majority of malaria mortality globally[1]. High-transmission regions like sub-Saharan Africa bear the greatest mortality, morbidity and economic burdens, but malaria caused by *P. falciparum* also imposes significant health and economic burdens in regions of low transmission, including South America[2]. Insights gained in regions of high transmission are often applied to regions of low transmission (and vice versa), however, multiple key biological conditions vary with transmission level including effective population size, outcrossing rates, intra-host competition, and host immunity. We do not fully understand the extent to which variation in these conditions impacts epidemiological, ecological, and evolutionary

dynamics under different transmission regimes, or whether these differences necessitate a shift in analysis approach. This is of growing relevance as more countries experience transmission declines and near malaria elimination.

The Americas are the global region with the lowest levels of endemic *P. falciparum* transmission and so present a rich opportunity for studying the evolutionary dynamics of small parasite populations. In the Americas, Venezuela accounts for 40% of all *P. falciparum* cases with important pockets of transmission also present in Colombia and Ecuador, with the former accounting for 20% of the regional cases[1]. More than 80% of cases in Colombia occur in the Pacific Coast Region[3]. After a period of sustained decline in malaria incidence between 2010-2017, there has been a recent increase in malaria in the Americas, reaching more than 900,000 reported cases in 2017[1]. The rise in cases is consistent with an increase in gold mining activities, an important driver of malaria transmission in the region, particularly for *P. falciparum*[4–7]. Another contributing factor is driven at least in part by instability in Venezuela impacting malaria control; with no imminent solution to the ongoing political and humanitarian crisis there, incidence is likely to remain on the rise, putting at risk elimination efforts not only in Venezuela but also in neighboring countries[8]. The evolution and spread of drug resistance have the potential to cause further setbacks. Drug resistance has arisen independently in South American *P. falciparum* populations multiple times in the past decades[9–11], and appears to be doing so again, with the recent novel emergence in Guyana of a C580Y mutation in Kelch13 that confers resistance to artemisinin, the current frontline antimalarial[12]. This highlights the importance of continued efforts in genomic surveillance in South America, particularly given the high human mobility currently taking place throughout the region.

The parasite population along the Pacific Coast of Colombia and Ecuador contains a low level of genetic diversity that reflects historical founding events as well as effective malaria control campaigns[13–15]. A large proportion of infections are monoclonal, which results in a low population-level (as opposed to meiotic) recombination rate and permits the long-term persistence of clonal genomes[16–18]. Standard population genetic theory would suggest that these are adverse conditions for adaptive evolution[19], but natural selection still operates effectively, as evidenced by the rapid spread of drug resistance alleles throughout the region in recent decades[20].

Prior genetic studies of parasites in the region have used neutral microsatellites[21] and SNP panels[16,22], which may lack the resolution required to reveal fine-scale differences among samples in this inbred parasite population. We therefore generated new whole-genome sequence data from 166 monoclonal *P. falciparum* samples collected between 2013 and 2017 from multiple sites along the Pacific Coast of Ecuador and Colombia. Our sampling focused on the

municipalities of Santa Bárbara de Iscuandé, Guapi, and Timbiquí in Colombia and on the municipalities of San Lorenzo and Esmeraldas in Ecuador. Relative to the region as a whole, these sites experienced high malaria caseloads —and in some instances local epidemics— during this time period[4,17,23,24]. To better understand population structure and selection dynamics in the region, we use a relatedness framework based on identity by descent (IBD) and compare these results to other common analytic approaches. Whole-genome IBD estimates confirm the presence of long-persisting clonal lineages across the region and more fully describe the high level of inbreeding that characterizes this population. In addition, we use IBD signatures to characterize instances of strong selection in both known and novel regions of the genome. Overall, we show that, despite the long-term persistence of clonal lineages, sufficient recombination occurs in this population to enable hard and soft selective sweeps that have responded to both the initiation and removal of drug pressure. This provides strong genomic evidence that *Plasmodium* populations, even when small and isolated, will not remain static as they approach elimination but will continue to adapt successfully to human interventions.

Results

Whole-genome sequencing finds predominantly monoclonal infections along the Ecuador-Colombia Pacific Coast Region

We performed whole genome sequencing on 207 samples collected from symptomatic *P. falciparum* malaria cases along the Ecuador-Colombia Pacific Coast between 2013 and 2017 (Figure 1; Supplementary Table S1). The Colombian data set includes 151 samples obtained from venous blood (leucocyte-depleted) from a 2014-2017 study that drew from the Guapi diagnostic microscopy post, which serves a range of interconnected rural and urban communities along the river networks of Guapi and neighboring municipalities[18]. When possible, travel history was documented, and from these data three sampled infections were assumed to have originated in Venezuela, with infected individuals traveling to the Guapi region for gold mining. In total, Colombia registered 104,074 *P. falciparum* cases between 2015-2017, from which approximately 92,000 were diagnosed in the four departments of the Pacific Coast region[25–28], and 3,920 in the municipalities of Guapi and Timbiquí (Cauca Department) and Santa Bárbara de Iscuandé (Nariño).

The Ecuador data set contains 56 samples collected at two study sites between 2013-2015: Esmeraldas and San Lorenzo. During these three years, Ecuador registered 396 cases of *P. falciparum* malaria, and 88% of the infections originated from these two Pacific Coast sites [29].

DNA from these samples was extracted from whole venous blood or filter paper blood spots and underwent selective whole-genome amplification prior to sequencing[30].

We restricted our downstream analysis to samples that had variant calls with greater than or equal to 5X coverage for at least 30% of the genome. Due to the different extraction and sequencing methods, the success rate for the two data sets varied, leaving 139, 38, and 3 samples from Colombia, Ecuador, and Venezuela, respectively. From within these high-coverage samples, we identified a high fraction of putatively monoclonal samples (126, 38, and 2 samples originating from Colombia, Ecuador, and Venezuela, respectively). This final set of 166 high quality, monoclonal samples was retained for further analysis.

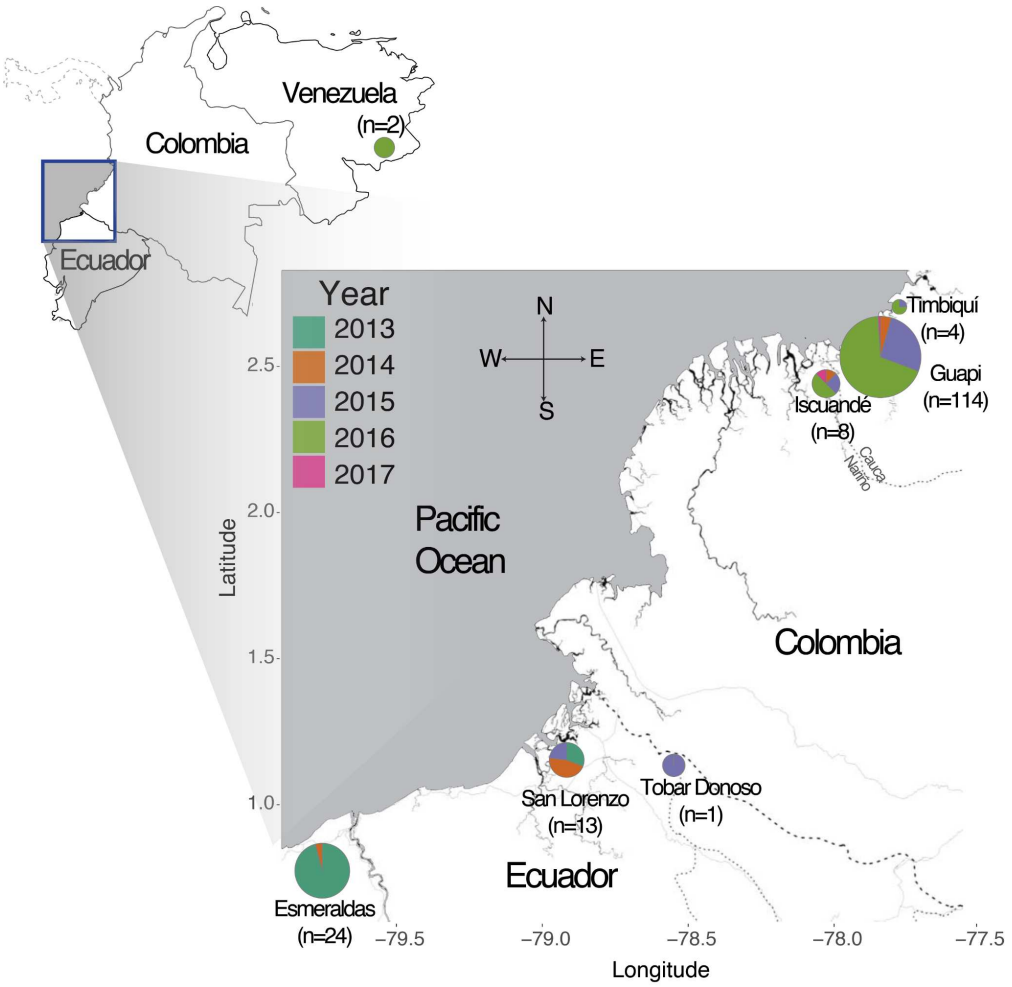


Figure 1. Geographic and temporal distribution of 166 monoclonal *P. falciparum* samples collected along the Pacific Coast of Colombia and Ecuador. Samples in Colombia originated in three municipalities (Santa Bárbara de Iscuandé in Nariño and Guapi and Timbiquí in Cauca) and were diagnosed at the Guapi diagnostic microscopy post, with the exception of the two monoclonal samples that originated in Venezuela. Pie chart divisions are colored by collection year and the area of the pie chart is proportional to the per-location sample count. Samples in Ecuador originated in three sites: Esmeraldas, San Lorenzo and Tobar Donoso. Esmeraldas and San Lorenzo are located 120 kms apart by road and Tobar Donoso is located 40 km east of San Lorenzo but there is no road to reach that locality. The number of samples indicates the number of high-coverage, monoclonal samples, which were used in all subsequent analyses. Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under ODbL.

The Pacific Coast parasite population shows high relatedness and a large proportion of clonal relationships

A previous analysis of 12 polymorphic microsatellites and 272 SNPs found two genetically distinct populations of *P. falciparum* in South America, separated by the Andes mountain range[15]. To explore this pattern at a whole-genome scale, we conducted a principal component analysis (PCA) that combined our Pacific Coast samples with whole-genome sequenced samples from Guyana in the eastern portion of the continent[12]. The results support the previously observed separation of eastern (Guyana) and western (Colombia-Ecuador) parasite populations (Supplementary Figure 1). This reinforces the evidence for strong structure at the continental scale in the South American *P. falciparum* population, underlining the potential for sub-regional elimination planning.

We next explored structure within the Western population, where prior studies using microsatellites and SNP panels have documented connectivity between Colombia, Ecuador, and potentially Peru, including the maintenance of clonal lineages along the Colombian coast for at least eight years[16,17,31]. In concordance with these studies, our PCA found no meaningful separation between Colombia and Ecuador samples (Supplementary Figure 1). To further dissect population structure across the region, we analyzed relatedness between parasite pairs by calculating identity-by-descent (IBD). Among all pairs of parasites originating from this region, IBD is high (median IBD = 0.29; Fig 2A). Median IBD within Colombia alone was similar to this region-wide estimate (median IBD = 0.27), whereas median IBD in Ecuador was extremely high (0.76). It is worth noting that the Ecuador data set was enriched for samples collected during an outbreak in Esmeraldas that was dominated by a single clone[32] (E3; Fig 2B). This likely inflates the IBD estimate and highlights how robust genomic inference needs to account for biased sampling. IBD between Colombia and Ecuador was also high (median IBD = 0.36), supporting the hypothesis of a high connectivity between countries. High levels of IBD were also previously observed in a genomic study of Guyana parasites[12]. However, outside of South America, such high levels of relatedness are unusual, even in low transmission settings like the Greater Mekong Subregion[33] (Fig 2A).

In addition to this overall high level of relatedness, we observed 1,737 parasite pairs that had a clonal relationship (IBD >= 0.99). We grouped clonal samples into clusters, members of which can be separated by *de novo* mutations but not by recombination events (Fig 2B). In total, we identified 19 distinct genomic lineages. Of these, 14 were clusters containing two or more samples, while five—including the two Venezuelan samples—are singletons observed only once in the data set. The 14 clonal clusters range in size from two to 43 samples (Supplementary Table

S2). Eight contain only samples from Colombia, three contain only samples from Ecuador, and three contain samples from both countries (clusters D, E2 and G). Relatedness between clonal clusters is variable (Figure 2C), which explains the spikiness apparent in the pairwise IBD distribution (Fig 2A). For example, members of cluster B and cluster D are IBD across 14% of the genome. These clusters are both large, leading to 688 pairwise comparisons with IBD of approximately 0.14.

To further understand relatedness at a sub-clonal scale, we iterated the clustering algorithm across a range of IBD thresholds (0.1-0.9; Supplementary Figure 2). This identified two highly-related “superclusters” containing clonal clusters that connect at IBD ≥ 0.8. For ease of identification, we named multi-member clonal clusters with letters and singleton genomes with numbers. We denote the clusters forming the two superclusters as E1-E2-E3 and F1-F2-F3 (Fig 2B).

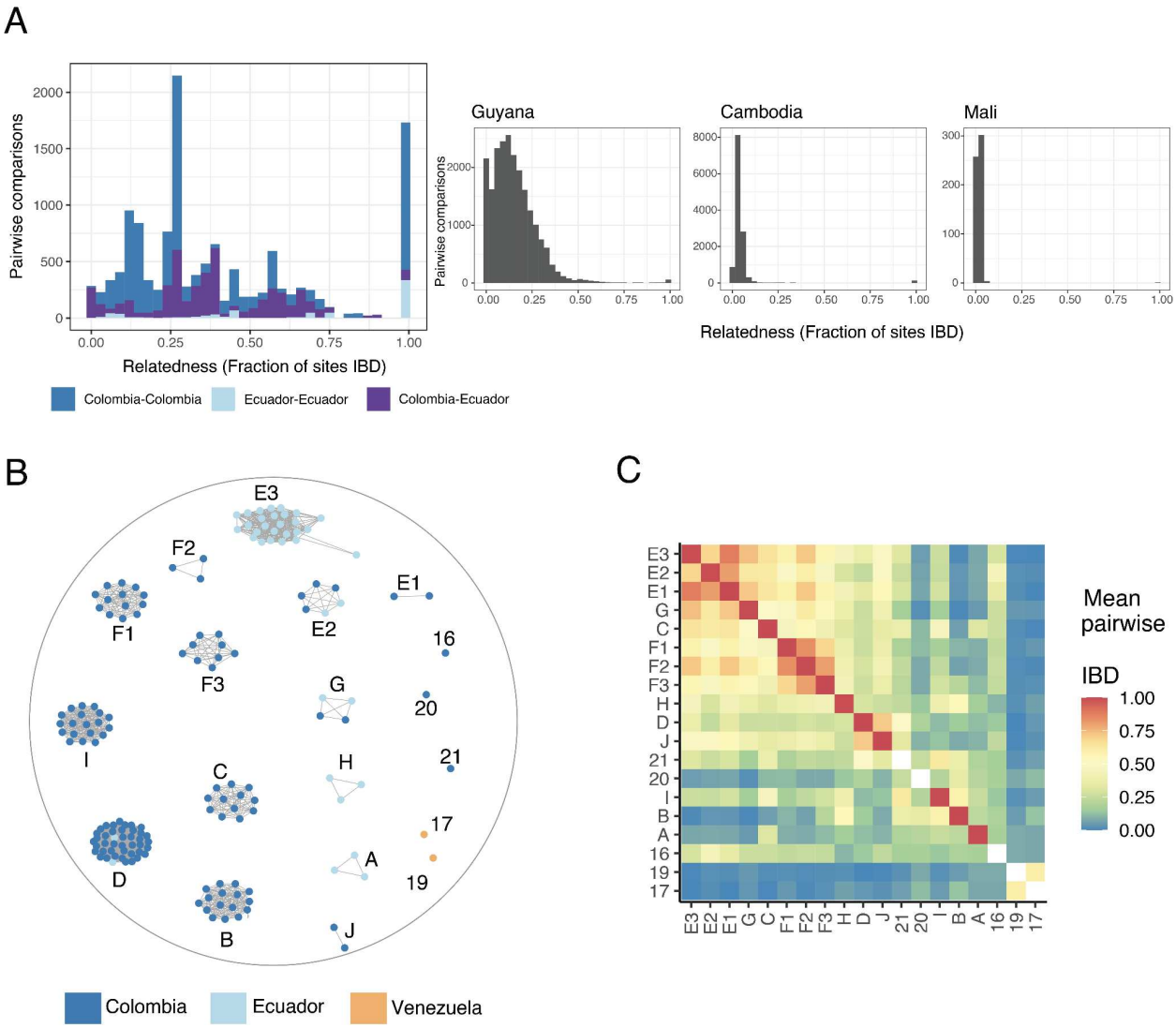


Figure 2. Relatedness between parasites along the Pacific Coast of Colombia and Ecuador. (A) Distribution of pairwise IBD in different transmission intensities: Pacific Coast of Colombia and Ecuador, Guyana[12], Cambodia[34] and Mali[34]. Samples from Western Cambodia belonging to the clonal expansion of K13 C580Y-harboring samples were excluded. (B) Network of 163 monoclonal samples collected from Colombia or Ecuador. Edges connecting nodes correspond to IBD≥0.99. (C) The mean IBD between clusters ranges from 0 to 0.89 and shows the presence of two larger "superclusters" at IBD>0.8. Two Venezuelan samples (17 and 19), show low IBD with all Colombian and Ecuadorian samples but intermediate IBD with each other. Single samples (numeric labels) are missing diagonal values as no intra-cluster comparisons could be made.

IBD analysis resolves the population structure of the Pacific Coast Region

With the exception of Taylor et al. (2020), previous analyses of *P. falciparum* population structure along the Pacific Coast have relied not on IBD but on alternative approaches. We further explored our data using common alternative methods to contextualize our work with prior studies and evaluate the value added by whole-genome IBD analysis.

First, we used the tool ADMIXTURE[35] to delineate related groups within the data set. Knudson et al (2020) recently took an analogous approach using the related tool STRUCTURE with a smaller set of genotyped SNPs generated from the same 2014-2017 Colombia samples[22]. The ADMIXTURE model assigned the Colombia samples to five qualitatively distinct groups (Fig. 3A) that recapitulate the previously identified groups[18] (Supplementary Figure 3) and resolve additional sub-structure, possibly as a result of using whole-genome data. The division of samples into groups, however, is driven by cluster frequency, which might not reflect true evolutionary or demographic patterns. We repeated the analysis after slightly altering cluster frequency on the order of what we expect to see with small localized outbreaks or uneven sampling. One cluster (C) was decreased from a frequency of 0.09 to 0.016 and a second (E1) was increased from a frequency of 0.015 to 0.06. This amount of simulated drift was sufficient to change which clonal clusters are described as “pure” versus “admixed” (Supplementary Figure 4).

We next analyzed the data using genetic distance (or the complement of identity by state; 1-IBS) methods and visualized these data with the common approach of creating a neighbor-joining (NJ) tree (Fig 3B). The NJ tree captures the identity of the clonal clusters, but it cannot accurately represent the relationships between clusters because they arose through recombination of standing variation, not divergence from a common ancestor. For instance, cluster E3 is more highly related to cluster F2 (mean IBD = 0.72) than to cluster F1 (mean IBD = 0.59) or cluster F3 (mean IBD = 0.56). While these relationships are apparent on the heatmap, the branch lengths on the NJ tree appear comparable for all three comparisons. The discrepancy is caused by the visualization approach, not differences in IBD vs IBS calculations, as these two measurements are highly correlated at the cluster level (Pearson’s r = -0.96; Supplementary Figure 5). It is worth noting, however, that only the use of IBD enables direct comparisons across studies and populations[36]. A heatmap for displaying IBD relationships may not be tractable with larger sample sizes and more outbred populations. In these instances, dimension reduction approaches like PCoA would be practical. Watson et al. (2020) have recently laid out considerations for rigorously applying these methods to *Plasmodium* data sets[37].

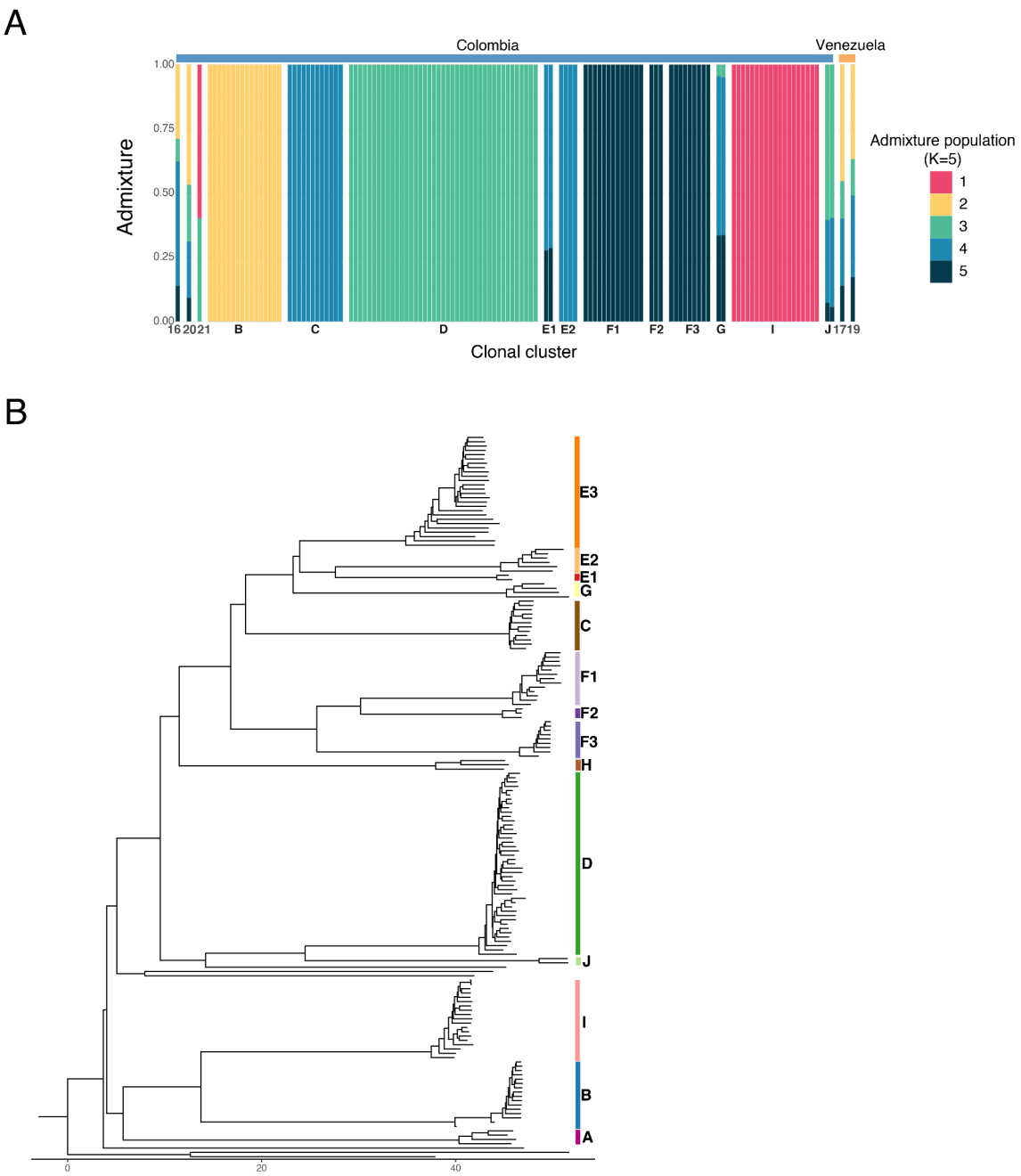


Figure 3. Comparison of an IBD approach to genetic distance analysis and visualization methods. (A) ADMIXTURE analysis of the Colombia-Venezuela samples identified five major groups, which correspond with the four largest clonal clusters (B, C, D and I) and one highly related supercluster (F1-F2-F3) in the data set. These results are dependent on cluster frequency (Supplementary Figure 4). (B) Depicting cluster relationships with a neighbor-joining tree based on genetic distance shows inter-cluster relationships in a qualitative way whereas the IBD heat map in Figure 2C provides quantitative relatedness estimates.

Six clonal lineages have persisted for a decade or longer within the Pacific Coast Region

We next assessed whether an IBD approach could be robustly applied to distinct data sets in order to explore further the spatial and temporal genetic structure of the region. To do this, we combined our data with 325 monoclonal Colombian parasites sampled between 1993 and 2007 that were previously genotyped at 250 SNPs[16,31]. We repeated our IBD and clustering analysis with a modified protocol that incorporated confidence intervals to accommodate the high proportion of missing calls at the 250 SNPs in some whole-genome sequenced samples (20% of samples were missing calls at ≥50% of the 250 sites). The initial analysis identified 16 distinct genetic backgrounds, which included 12 with multiple members and four singletons. Eleven of the 12 multi-member clonal clusters corresponded uniquely to a single WGS clonal cluster, and one encompassed three highly related WGS clusters (E1, E3, and G). The mean IBD point estimate between samples within this latter cluster was lower than that for the other 15 clusters (0.91 versus >0.99), and upon inspection, there was strong support for breaking this final group into the same three smaller clusters identified with WGS data (Supplementary Figure 6A). With the sparser SNP data, the approach had lower power than a genome-wide analysis. Nevertheless, the two data sets provided IBD point estimates between clonal clusters that were highly correlated (Pearson's r = 0.93; Supplementary Figure 6B).

We found that six of the multi-member clonal clusters from 2013-2017 held clonal relationships with parasites sampled prior to 2008 (Figure 4, Supplementary Figure 6C). The longest identified lineage (cluster E3) persisted back to at least 1999 and was the major driver of an outbreak in Ecuador in 2013[17]. Four of these persistent lineages incorporated samples from both Colombia and Ecuador.

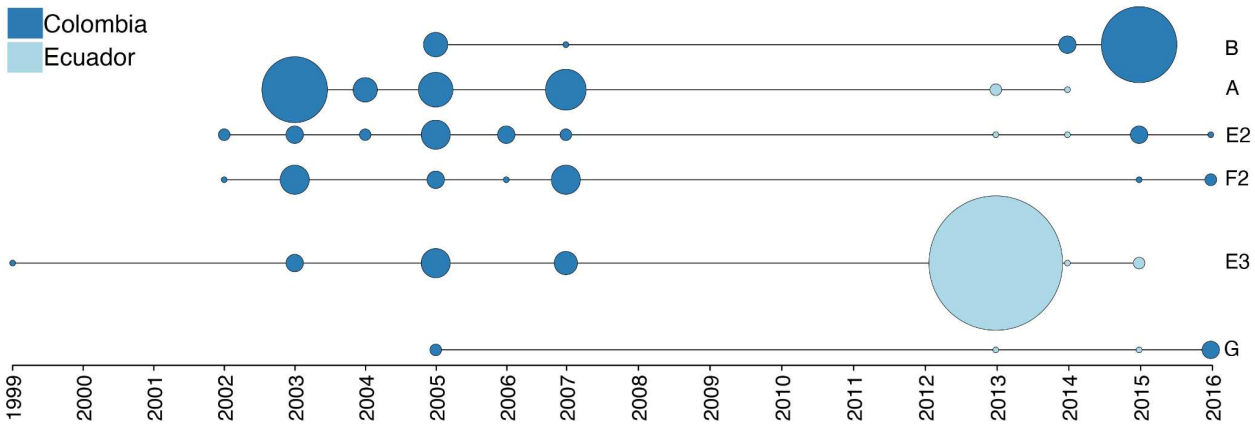


Figure 4. Spatiotemporal distribution of long-persisting clonal clusters across Ecuador and Colombia. A combined analysis of recent whole-genome sequencing data (2013-2017) and older parasites genotyped at 250 SNPs (1993-2007; N=325; Taylor et al, 2020) identified six clonal clusters that have persisted in the region for a decade or longer. Four of these clusters have been sampled in both Colombia and Ecuador, albeit not within the same year. Clonal clusters detected in a given year are depicted by the vertices with the size of the vertex corresponding to the number of samples. Supplementary Figure 6B breaks down the clonal components by within-country location.

Known mutations conferring drug resistance are common throughout the region

Our analysis provides further evidence that the Pacific Coast Region harbors a *P. falciparum* population with a small effective population size and a low population-level recombination rate. Evolutionary theory predicts that selection will not be highly efficacious under these conditions, so we were therefore interested in exploring how recent selection has progressed in the region. Multi-drug resistance is documented throughout the area, but antimalarial use—and therefore specific selection pressures—historically differed in Colombia and Ecuador (Fig 5A). Prior to the mid-1950s, both countries relied primarily on chloroquine. After emergence of clinical resistance in the 1950s, Colombia began replacing chloroquine with pyrimethamine, first as a monotherapy and, later in the 1980s, in combination with either sulfadoxine or sulfadoxine and chloroquine (SP, SP+chloroquine)[38,39]. Contrary to this, Ecuador continuously used chloroquine as the frontline malaria treatment until its replacement with artemisinin-combination therapies in the mid-2000s[3].

Given these specific drug regimes, we first assessed the prevalence of known resistance haplotypes in four genes that have strong prior evidence of contributing to drug resistance: Chloroquine-resistance transporter (*crt*, Pf3D7_0709000), dihydrofolate reductase (*dhfr*, PF3D7_0417200), dihydropteroate synthetase (*dhps*, PF3D7_0810800), and multidrug resistance protein 1 (*mdr1*, PF3D7_0523000). Mutations in a fifth gene, *kelch13* (PF3D7_1343700), have been associated with resistance to the antimalarial artemisinin,

however, in our data set no known resistance mutations were present in this gene, so it was excluded from further analysis.

Using all samples, we calculated the frequencies of these known resistance haplotypes at the country level. Overall, there was a high proportion of resistance-associated mutations in both Colombia and Ecuador, although wildtype (drug sensitive) haplotypes were present at appreciable frequencies for both *dhfr* and *dhps*. No sample contained wildtype haplotypes at *crt* or *mdr1*, and no sample contained the CRT C350R mutation, which has been documented as restoring chloroquine sensitivity in French Guiana[40]. Both samples originating from Venezuela contained highly resistant haplotypes at all successfully genotyped loci, and were the only samples containing the DHFR triple mutation, highlighting important differences between the two regions (Fig 5B). The different patterns in Colombia and Ecuador may reflect the different historical drug usage, but they are also affected by the highly clonal structure and small population size, which lead to repeated stochastic sampling of identical genomic backgrounds at the population level (drift) and at the study level. It is therefore unlikely that these raw allele counts accurately reflect selection. For instance, a single clonal cluster (E3) dominates the Ecuador sample set because of its high prevalence during the 2013 Esmeraldas outbreak (Fig 2 and 4). However, it is unknown whether this clone reached high frequency due to high intrinsic fitness or simply because it was present when ecological conditions became conducive for an outbreak.

In addition to calculating population-level allele frequencies, we mapped the resistance mutations to our clonal clusters (Supplementary Figure 7). This mapping suggests that frequent *de novo* mutation is not a major force structuring drug resistance at these loci as all samples within a clonal cluster contained matching haplotypes. Further, a subset of 17 pre-2007 samples from four of the six persistent clonal lineages (Fig 4) were previously genotyped at *crt* and *dhps*, and all genotypes were concordant through time.

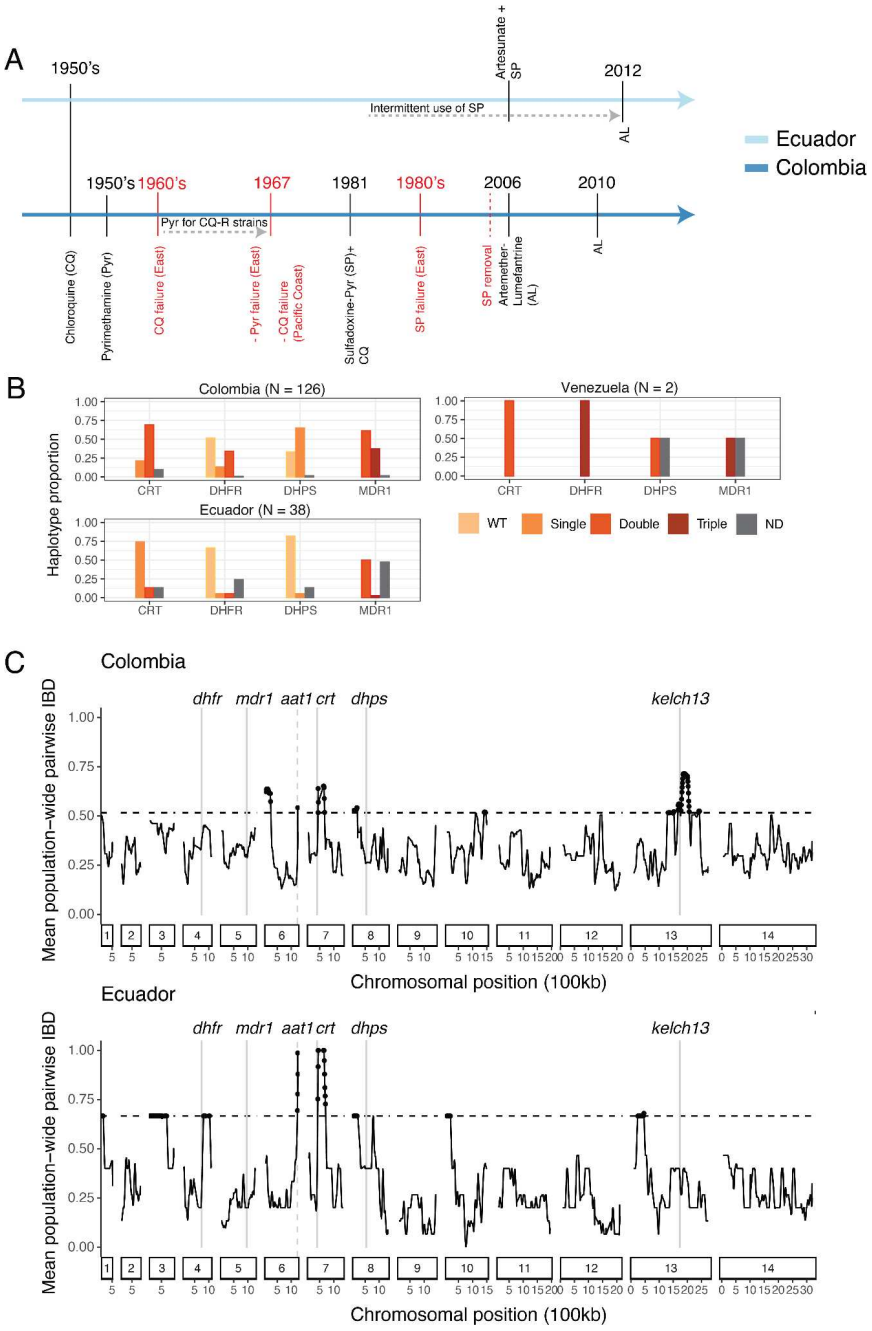


Figure 5. Drug selection across the Pacific Coast Region (A) Timeline of official antimalarial usage for treatment of *P. falciparum* malaria (black) and documented emergence of antimalarial resistance (red) in Ecuador (light blue) and Colombia (dark blue). Data are based on epidemiological information from the respective country-level Institutes of Health. (B) Drug-resistance haplotypes in the 166 monoclonal samples, colored by wild type (WT) allele, by the number of known functional mutations, or in gray for haplotypes not-determined due to low read coverage (ND). Drug-resistance haplotypes at the clonal cluster level are available in Supplementary Table S2. (C) Mean pairwise IBD within 50-kb overlapping windows across the genome (y-axis) for Colombia and Ecuador. IBD estimates were made using one representative sample per clonal cluster. Horizontal lines mark the 95th percentile for the respective country. Vertical gray lines mark the chromosomal positions of genes with known involvement in antimalarial resistance.

Shared IBD segments identify putative selective sweeps at known and novel targets

We next examined shared ancestry both genome-wide and around these specific resistance loci to understand how mutation and recombination have shaped patterns of selection. To investigate intrachromosomal patterns of shared ancestry, we calculated the mean pairwise IBD within overlapping 50-kb windows. Unlike the allele frequency estimates, distinct genomic backgrounds were included only once in the calculations to minimize the impact of stochastic re-sampling. While migration between countries is observed, we initially analyzed the two data sets separately to look for potential regional differences. As a means of focusing on genomic regions that likely experienced the strongest selective sweeps, we examined windows with mean relatedness above the 95th percentile. Six and nine peaks crossed this threshold in Colombia and Ecuador, respectively, with three of these regions shared by both countries (Fig 5C). Two of these shared peaks contain genes implicated in resistance to chloroquine, a drug that is known to have imposed strong selection on *P. falciparum* in the 20th century. While sample sizes are small and the effect of stochastic sampling is likely strong, it is interesting to note that the mean IBD within both windows is higher for Ecuador, where chloroquine selection took place for a longer period of time (Fig 5A).

One chloroquine-associated IBD peak begins at *crt* on chromosome 7, where a strong selective sweep is known to have occurred in South America leading to the fixation of the large-effect variant K76T along the Pacific Coast[10,11,18,34,41]. As we discuss further below, we observe an extreme depletion of segregating polymorphisms around *crt,* making the locus of selection difficult to pinpoint. The observation that *crt* is on the edge rather than the center of the IBD peak could be an artifact stemming from this low resolution or it could reflect an additional sweep that occurred after the K76T fixation.

A second chloroquine-associated peak found on chromosome 6 is a novel result in this geographic region. The peak contains (in the case of Ecuador) or is adjacent to (for Colombia) amino acid transporter 1 (*aat1,* Pf3D7_0629500), which has been identified as mediating resistance to chloroquine and other drugs[42,43] and as being under selection in natural populations[44,45]. In this data set, we observe a high-frequency derived allele that causes a nonsynonymous serine to leucine change at amino acid 258 (S258L), which falls within the protein's predicted transmembrane domain. Only one Pacific Coast sample (SPT26239) carries the ancestral allele. Interestingly, this sample is an outlier in our data and shows the lowest overall relatedness to the other Pacific Coast samples in our collection. Both Venezuelan samples also exhibit the ancestral allele, evidence that this putative selection signal at *aat1* may be population-specific rather than continent-wide like *crt* (Supplementary Figure 8). To further understand the

timing of selection, we examined the frequency of this mutation in 16 Colombian samples collected prior to 2005 that were whole-genome sequenced together with the sample set used for the longitudinal analysis of clonal persistence. The older samples were fixed for the known resistance-conferring K76T CRT mutation, but the ancestral AAT1 S258 allele was present in 25% of these early samples. In sum, while the genomic IBD scan cannot definitively pinpoint the causal allele driving selection, these results provide the first evidence of potential *aat1*-mediated chloroquine resistance in South America.

The third IBD peak within the 95th percentile for both countries is located on chromosome 8. This peak does not contain any known selection or drug targets, but it does include high-frequency coding variants in a gene associated with the proteasome/ubiquitination pathway (SENP2; PF3D7_0801700), which was previously identified in a study of artemisinin-driven selection in Southeast Asia[46].

Joint analysis of the full data set yielded similar results to the country-specific analyses. There are eight peaks above the 95th percentile (Supplementary Figure 9), and all but one overlap outlier regions in at least one country-level analysis. In addition to the peaks discussed above, two other large peaks stand out in both the joint and Colombia-specific analyses: one is on chromosome 13 and the second covers an additional region of chromosome 6. The chromosome 13 peak contains *kelch13*, a gene for which several amino acid variants are associated with resistance to the antimalarial drug artemisinin. In our data set, however, the only identified Kelch13 variant is a common, wide-spread polymorphism (K189T) that has shown no evidence of carrying a phenotypic effect[47]. We therefore hypothesize that the driver of the sweep may be an alternative locus. Candidates include genes with prior evidence of drug-associated phenotypes such as tyrosine kinase-like protein 3 (PF3D7_1349300)[48] and lysine--tRNA ligase *(krs1;* PF3D7_1350100)[49–51]. In this data set, both of these genes contain high-frequency variants that alter the coded protein sequence. In addition, a key gene involved in *Plasmodium* adaptation to novel vectors (*pfs47*; PF3D7_1346700) falls in the center of this window[52]. While we did not identify any high-frequency coding variants in this gene, expression-level variation could be mediated via upstream or downstream polymorphisms, which we did not analyze with this short-read data.

In contrast to *crt*, the genomic regions around the three other focal drug resistance genes (*dhfr*, *mdr1*, and *dhps*) do not show IBD patterns indicative of universal hard selective sweeps (Fig 5C, Supplementary Figure 9). A genomic segment bordering *dhfr* shows elevated IBD in Ecuador, suggesting strong selection in this country, but the same signal is not present in Colombia or the combined data set. The high frequencies of known resistance-conferring alleles

within these genes therefore suggest that selection may have mainly occurred through soft or incomplete sweeps. To evaluate this hypothesis, we examined the diversity of haplotype sequences in the 30-kb regions surrounding these loci (Fig 6). We selected one representative sample from each clonal cluster and additionally included the 16 pre-2005 Colombian samples mentioned above. We did not attempt to statistically evaluate these regions with haplotype-based selection tests due to the limited number of distinct genomic backgrounds and the presence of both geographic and temporal structure in the dataset.

Haplotypic diversity around *crt* is very low, consistent with there having been a hard selective sweep at this locus (Fig 6). We observe only three distinct 30-kb haplotypes in the expanded data set of 33 genomes. One haplotype contains two *crt* mutations (CVMET). This double-mutant haplotype was the only one found in the 16 pre-2005 Colombian samples showing that selection had already occurred before this time. A large proportion of the post-2013 samples from Colombia and Ecuador instead carry one of two 30-kb haplotypes that carry a single *crt* mutation (CVMNT; Fig 5B). These single-mutant haplotypes differ by only one nucleotide change, making it probable that they represent the same ancestral haplotype separated by a single, recent mutation event (Fig 6). The double mutant did not occur on the single-mutant background, showing different origins and separate selection events for these two haplotypes[10,11,41,53].

Contrasting with the haplotypic pattern around *crt*, combinations of resistance mutations in *mdr1*, *dhfr*, and *dhps* appear on multiple haplotypic backgrounds in both the older and more recent data sets (Fig 6). This could arise from either multiple *de novo* mutation events, or weaker selection (relative to that at *crt*), which allowed time for mutations to recombine onto different backgrounds. Of the four loci, only *dhps* maintained numerous distinct wildtype haplotypes, suggesting that selection has been weakest at this resistance locus, consistent with the history of drug therapy in both countries.

Interestingly, at *dhfr*, wildtype alleles are common and found on 11 of the 17 Pacific Coast genomic lineages. All of these genomes, however, are identical across the surrounding 30-kb haplotype. In contrast, variation is much higher between haplotypes carrying the single S108N mutation. The loss of haplotypic diversity around the ancestral allele is consistent with there having been strong selection against the wildtype allele when the drug pyrimethamine was in use, followed by a re-expansion of the wildtype allele as pyrimethamine dosage was reduced in 2006 in Colombia and in 2011 in Ecuador, where SP was used intermittently as a second-line treatment[54]. This is in agreement with laboratory observations that the DHFR S108N mutation carries a fitness cost in the absence of pyrimethamine[55], and parallels field observations of an

increase in wild-type alleles[56,57] or the rise of compensatory mutations[40] in *crt* following chloroquine cessation in other global populations.

In most instances, the two samples of Venezuelan origin carry haplotypes with distinct polymorphisms and even distinct combinations of resistance alleles relative to Pacific Coast parasites. Although the sample size is too small to be conclusive, this pattern suggests that the Venezuelan *P. falciparum* population has experienced selection on independent *de novo* mutations. The one exception is *dhps.* At this locus, one Venezuelan sample carries a haplotype that matches a subset of Pacific Coast samples.

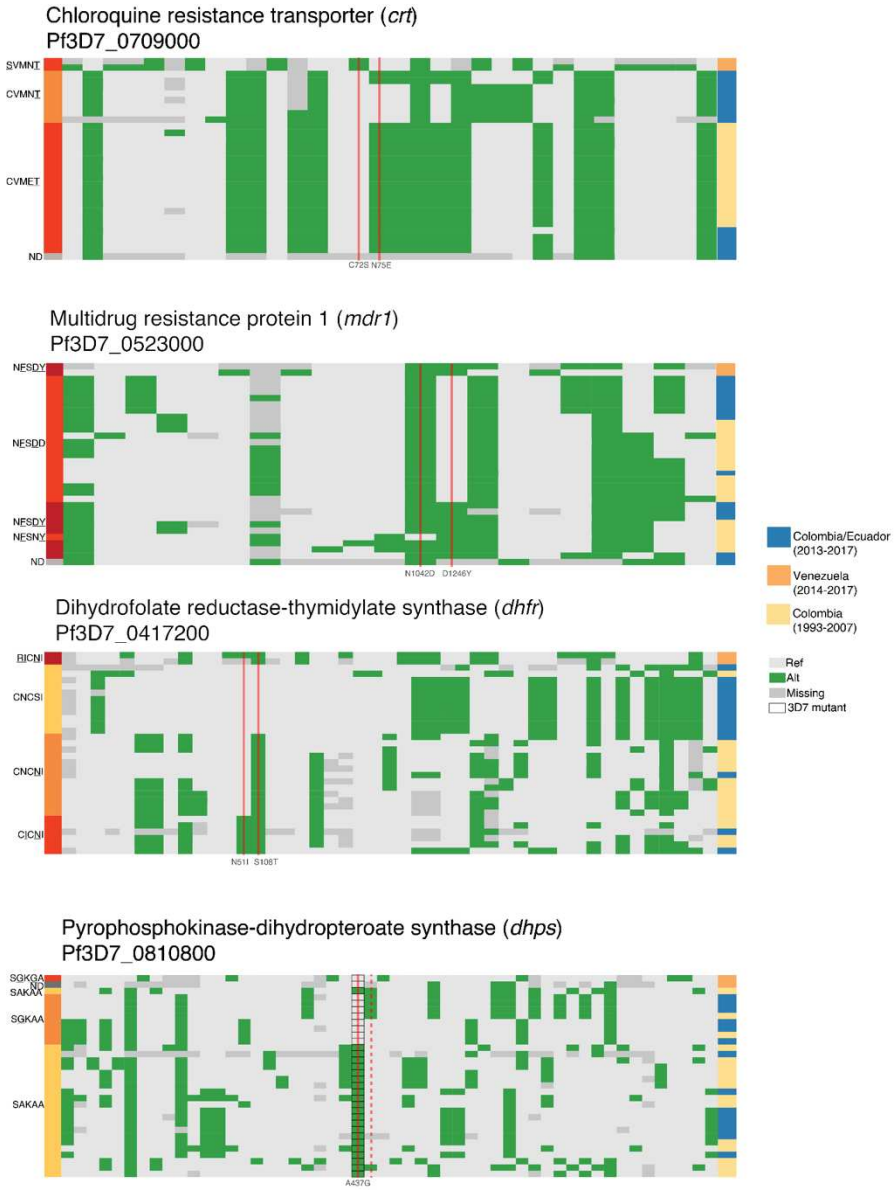


Figure 6. Haplotypes surrounding known drug resistance mutations in *crt*, *mdr1*, *dhfr*, and *dhps* show varying degrees of diversity. Plots display segregating SNPs within the 30-kb flanking regions around mutations of interest within four genes involved in antimalarial drug resistance. One high-coverage sample from each clonal cluster is displayed (rows). Variant positions were removed if they were within five nucleotides of a called indel or if they displayed a high rate of heterozygous calls within monoclonal samples (Materials and Methods). Calls are colored based on matching (light grey) or mismatching (green) the 3D7 reference. Calls are marked as missing (dark grey) if they were heterozygous or had fewer than five reads. Black outlines at alleles coding for *dhps* position 437 indicate that 3D7 contains the A437G mutation; wildtype calls are therefore green at this position. Nonsynonymous mutations of interest are depicted with a solid vertical red line. One synonymous mutation at codon 540 in *dhps* is depicted with a dashed vertical red line. Represented genes and downstream codon changes are (top to bottom): *crt* (72 and 75); *mdr1* (1042 and 1246); *dhfr* (50, 51 and 108); and *dhps* (437 and a synonymous mutation coding for 540). Supplementary Table S2 lists the mutations detected in the clonal clusters.

Discussion

The epidemiological, ecological, and evolutionary dynamics of malaria are expected to change as disease transmission declines in response to control efforts, and these transitions may necessitate different analytical tools. Studying these dynamics in current areas of low transmission can therefore inform best practices for future genomic epidemiological studies and prepare us to track malaria decline more globally. Here, we used IBD metrics to gain insight into both the population structure and recent evolution of *P. falciparum* parasites along the Pacific Coast Region of Ecuador and Colombia, a region of low transmission. The results demonstrate that IBD accurately characterizes the highly clonal structure of this population, and may be useful for studying other parasite populations as control efforts advance. Other common analysis approaches such as PCA, ADMIXTURE, and Neighbor-Joining trees only partially recapitulate these results, and do not lend themselves to a straightforward biological interpretation in the same manner as measurements of relatedness between parasites.

Because relatedness estimates can be compared between populations and time points[36], an IBD approach further enables a more detailed understanding of global and temporal variation in inbreeding and clonality—even between areas like the Pacific Coast Region and Greater Mekong Subregion, which are often both included under the same “low transmission” umbrella. These comparisons have practical importance as several transmission-related phenomena can impact relatedness in a population. These include: local sub-structure, which can inform the spatial implementation of control strategies; the proportion of imported versus endogenous parasites; and general transmission declines, for which genomic surveillance may prove a cost-effective means of assessment.

Our study incorporated new whole-genome sequence data from 166 monoclonal infections. This level of genomic resolution deepens our understanding of how selection has progressed in the region by enabling whole-genome scans and genetic analysis of large haplotype blocks. These two approaches provide strong support for there being at least two loci that experienced hard selective sweeps as a consequence of chloroquine pressure: *crt* and *aat1*. The phenotypic effects of *crt* mutations have been previously studied in South America, but *aat1* was only recently implicated in chloroquine resistance. This is the first evidence of this gene’s potential role in the Americas and raises concerns regarding cross-resistance to other quinoline-based combination therapies. In contrast to *crt* and *aat1*, three other genes with known drug-resistance phenotypes show evidence of soft, rather than hard, sweeps. Given the small number of distinct genomic backgrounds and the long-term persistence of these haplotypes—most of which trace back a decade or longer—we cannot definitively differentiate between recurrent *de*

novo mutations and recombination. Regardless of source, however, the persistence of multiple mutation-bearing haplotypes at *mdr1*, *dhfr*, and *dhps* versus the hard sweeps at *crt* and *aat1* highlights differences in the strength of selection connecting these drug-gene pairs. Interestingly, we also observe that the removal of drug pressure may have aided the expansion of one wildtype *dhfr* haplotype.

These results contribute to a growing body of literature that demonstrates how *P. falciparum*'s evolutionary potential is maintained even when effective population size is small and within-host competition is low. *Plasmodium* experiences extreme fluctuations in cell count over the course of its life cycle. These expansions and contractions cause the site frequency spectrum and strength of selection to differ from Wright-Fisher expectations[58] and create higher levels of standing variation than some measures of effective population size would suggest[59]. Here, we find that *de novo* mutation alone does not govern adaptation in small populations. Recombination also plays a role, and when selection is strong, even low population-level recombination is sufficient to passage new beneficial alleles onto multiple genomic backgrounds and through the population. Outcrossing can only occur when multiple genotypes are present in a mosquito's blood meal (COI > 1), and so transmission dynamics likely play a large role in governing evolutionary potential across this region. In both Colombia and Ecuador, localized outbreaks periodically drive infection counts well above baseline levels (API>10) and may increase the likelihood of outcrossing. Analyzing the temporal and spatial distribution of these events, determining their contribution to outcrossing events[32], and assessing the role of epistasis in driving clonal dynamics are the next steps for describing evolutionary trajectories in populations dominated by monoclonal infections. This will increase our capacity to anticipate the course of adaptation to drugs—and other novel interventions—as other populations near elimination.

The whole-genome data generated here also serve as validation for results obtained with other genotyping approaches. We found that a mid-sized panel of 250 SNPs[16] recapitulates the clonal clusters identified with whole genome sequencing and provides reasonable IBD point estimates for partially related parasites (Supplementary Figure 6B). However, uncertainty can overwhelm estimates of IBD-based relatedness based on sparse marker data. Moreover, samples from different data sets are liable to miss data at many markers, as they do here. Consequently, confidence intervals are critical for both maximal sample retention and for quality control: with confidence intervals, we can estimate relatedness as tolerantly as possible (e.g. estimate relatedness for samples that share any data) and then use the confidence intervals to filter highly uncertain estimates; without confidence intervals, we must resort to an arbitrary SNP cut-off (e.g.

only estimate relatedness for samples that share data on at least 100 SNPs) and then hope that the estimates are reliable.

Recent efforts in the malaria field have borne several amplicon panels with wide geographic breadth that can quickly and affordably genotype hundreds or thousands of samples[22,60,61]. While whole-genome sequencing will remain the bedrock of selection analyses, these methodological advances in targeted sequencing will facilitate the use of IBD analysis with confidence intervals for describing local population structure (as we do here)[62,63], measuring connectivity between populations[64], identifying likely importation event [65], and tracking changes in transmission[66]. These advances in genomic epidemiology are enhancing established malaria surveillance toolkits and enabling responses tailored to individual country's needs[22,60,67,68].

Methods

Sample collection

In Colombia, sample collection at the Guapi Health Post took place between 2014-2017 from individuals reporting malaria symptoms, and in diagnostic posts in the Guapi municipality (El Carmelo), as well as in Santa Bárbara de Iscuandé (Chanzará) and Timbiquí (El Cuerval) as previously reported[18]. Upon arrival, participants were diagnosed with malaria via microscopy, and after obtaining informed consent, 2-5mL of venous blood were collected into ethylenediaminetetracetic acid (EDTA) vacutainer tubes (BD Vacutainer). Sample origin was determined through a travel history survey. Cases were coded as local if a patient had remained at the site throughout the previous two weeks and imported if they had spent the majority of the previous two weeks at an alternate location. National and international health research standards were considered, and the project was presented and discussed with the local health authorities. The present work corresponded to minimal risk. It was approved by the ethics committee (Evaluation Report 127-14 and 003-021-16) of the Medical School of the Universidad Nacional de Colombia[69–71].

Colombian samples from between 1993 and 2007 were collected in municipalities from four departments on the Pacific Coast: Tadó and Quibdó in Chocó, Buenaventura in Valle, Guapi in Cauca, and Tumaco in Nariño. Informative samples (N=325) reported in this study were genotyped at 250 SNPs from blood spots collected on filter papers as reported in Echeverry et al[16]. In addition, 16 samples were used for whole-genome sequencing. Briefly, 5 mL of venous blood was collected, followed by adaptation to *in vitro* culture[72]. Genomic DNA was extracted using the Purelink extraction kit (ThermoFisher Scientific Waltham, MA, USA).

In Ecuador, samples were collected between 2013 and 2015 by Ministry of Health personnel from individuals reporting malaria symptoms, and in diagnostic posts in the San Lorenzo and Esmeraldas municipalities as well as in the locality of Tobar Donoso (on the border with Colombia). Upon arrival, participants were diagnosed with malaria via microscopy. Written informed consent was provided by study participants and/or their legal guardians and 2-5mL of venous blood were collected into CPD vacutainer tubes (BD Vacutainer). Alternatively, 2-4 drops of blood were spotted on 3M filter papers. Sample origin was determined through a travel history survey. Cases were coded as local if a patient had remained at the site throughout the previous two weeks and imported if they had spent the majority of the previous two weeks at an alternate location. The protocol was approved by the Ethical Review Committee of Pontificia Universidad Católica del Ecuador (approvals #: CBE-016-2013 and 20-11-14-01).

Whole-genome sequencing

Samples collected in Colombia, belonging to the studies undertaken in Knudson et al., and in Echeverry et al., were sequenced at the Wellcome Sanger Institute, as part of the MalariaGEN *Plasmodium falciparum* community project[34]. An Illumina HiSeqX platform was used to generate 200 bp paired-end reads. Samples collected in Ecuador underwent selective whole genome amplification at the Harvard School of Public Health [30] before library construction with Nextera XT library kit and sequencing at the Broad Institute on an Illumina HiSeqX.

We aligned reads to the *P. falciparum* 3D7 v.3 reference assembly and called variants following the best practices established by the Pf3K consortium (https://www.malariagen.net/projects/pf3k). In brief, raw reads were aligned with BWA-MEM[73] and duplicate reads were removed with Picard tools. SNPs and indels were called with GATK v3.5 HaplotypeCaller[74]. Base and variant recalibration (BQSR and VQSR) steps were performed using a set of Mendellian-validated SNPs. Downstream analysis was limited to variants found in the core region of the genome, as defined by Miles et al[75]. We masked any site that was called as heterozygous in >10% of samples and masked any individual call supported by fewer than five reads. Unless otherwise noted, we also excluded from analysis any variant within 5 nucleotides of a GATK-identified indel.

In multiple samples from the 2015-2017 Colombia data set, we found evidence of exogenous PCR amplicon contamination around the *kelch13* gene, and so this gene was masked from downstream analysis. We visually inspected samples in this region to determine if there was any remaining evidence for *kelch13* variants and results were cross-checked with previous

genotyping that had been performed on these samples. We detected only one valid polymorphism in the gene, which codes for the known, common variant K189T.

Population structure in Guapi

To determine the population structure of samples obtained in Colombia (2014-2017) and ancestral composition of the sympatric populations circulating in Guapi previously reported in Knudson et al., we combined Principal Component Analysis (PCA), as well as a Bayesian model-based model of admixture (ADMIXTURE)[35]. For both analyses, VCF filtering was performed beforehand to remove non-biallelic sites and to exclude heterozygous calls[76]. A pedigree format file was generated with Plink[77] and the output was used as input for ADMIXTURE, which we then performed for K=3 to 10 populations (Supplementary Figure 3). PCA analysis was performed at the local and regional levels with the same filtering parameters used with Admixture (Supplementary Figure 1).

IBD analysis

IBD analysis of the post-2012 parasites, included all monoclonal samples with high quality whole genome sequence data (>= 5x coverage for >30% of the genome). We estimated IBD on these whole-genome samples with a set of 16,460 SNPs using hmmIBD[78]. The SNPs included in the analysis satisfied the following requirements: (1) found within the core genome as defined by Miles et al, 2016[75]; (2) >5nt from any GATK-called indel; (3) called in at least 80% of samples; (4) minor allele frequency >= 0.05; (5) called as heterozygous in <10% of declared monoclonal samples. Individual calls with <5x read support were marked as missing. We estimated population-level allele frequencies in three ways: (1) using the full dataset (default parameters), (2) using a sample set with only a single representative per clonal cluster; and (3) using only samples from the pre-2008 data set. For genome-wide IBD estimates, the differences among these methods were minimal, and the default estimates were used in downstream analyses. We performed clustering based on the fraction of sites called IBD as estimated with the Viterbi algorithm (fract_vit_sites_IBD) by constructing an adjacency matrix in the R package igraph[79].

We performed an additional extended IBD analysis that included 325 pre-2008 samples from Echeverry et al (2013)[16]. This analysis included 250 GoldenGate SNP calls made for the Echeverry samples and GATK-calls made at the same positions for the whole-genome sequenced samples. GoldenGate calls were decoded using genotyping results from a 3D7 lab strain that were then compared to the PlasmoDB 3D7 reference sequence, as well as Dd2, Santa Lucía, HB3 and 7G8 strains[80,81]. Sixteen samples from Colombia had undergone both

GoldenGate genotyping and whole genome sequencing, and we confirmed that the two platforms made identical calls in all cases with the exception of one site that was then masked from analysis. In addition to decoding, we reordered some SNPs whose names we discovered were previously missordered.

The extended analysis relied upon confidence intervals around IBD-based relatedness estimates, which were computed using the statistical framework described in Taylor et al. Genetics 2019[36]. After removing one SNP for which there was no data among the WGS samples and seven WGS samples that had no data among the remaining 248 SNPs, we were left with 248 SNPs and 519 samples. Missing data was high, so to maximize sample retention, we estimated relatedness as tolerantly as possible (i.e. using all samples that share any data). To ensure quality control, we then used confidence intervals to filter uncertain estimates. As in Taylor et al. 2020[31], confidence intervals were also used to circumvent an arbitrary SNP cut-off for clonal classification and the igraph package in R [79] was used to construct clonal components (referred to as clusters in the results). Since the igraph package does not support the construction of clonal components using samples with missing relatedness estimates, those samples were removed, leaving 496 samples with 122760 relatedness estimates based on data from 12 to 248 SNPs. A sample pair was considered clonal if its 95% confidence interval included one and exceeded 0.75. This clonal definition, which is more stringent than that used before (95% confidence interval includes one, was necessary to minimize the number of cliques within clonal components (ideally all clonal components should be cliques - fully connected subgraphs - but they are not due to greater uncertainty among relatedness estimates). The same principle (minimizing cliques within components) was used to break down a clonal component containing six cliques into three clusters: the largest containing two cliques, the other two continuing a single clique each (Supplementary Figure 6A]. For further details of the extended analysis see https://github.com/aimeertaylor/ColombianBarcode/blob/master/Code/Extended_analysis/Analysis_summary.Rmd.

As a final aside, we did a marker-reordered analysis of Taylor 2020 for continuity[31]. The results were qualitatively consistent. In the marker-reordered analysis we count 45 clonal components whereas previously there were 46. Among the original 46, 44 are identical, while two differ: two samples constituting one clonal component were no-longer considered clonal; another clonal component had one additional sample; see https://github.com/aimeertaylor/ColombianBarcode. Due to these minor differences, some original clonal component labels are offset by one.

Acknowledgements

The work performed in Colombia was supported by the Newton Caldas Fund Institutional Links G1854 Award to JC and VC. Additionally, the Medical Faculty at Universidad Nacional de Colombia provided support with awards HERMES 35988 and 32309 to VC. Whole genome sequencing of all samples obtained in Guapi was financially supported by MalariaGEN and the Wellcome Trust (206194, 090770). From Colombia we thank the Guapi communities, the Secretaría Municipal de Salud del Cauca and Secretaría Departamental de Salud del Cauca. As well as the VEuPathDB outreach team for assistance in the GoldenGate decoding SNP process, with the Colombian samples. We thank Marco Galardini for helpful input and discussion. Support for Ecuador was provided by Pontificia Universidad Catolica del Ecuador, grants M13416, N13416 and O13087 to FES and Ministerio de Salud Pública del Ecuador. In particular we thank the communities in Esmeraldas and San Lorenzo and the Health districts (especially Drs. Javier Obando, César Diaz and Julio Valencia). This project has been funded in whole or in part with Federal funds from the National Institute of Allergy and Infectious Diseases (NIAID), National Institutes of Health, Department of Health and Human Services, under Grant Number U19AI110818 to the Broad Institute. TA is supported by 5R37 AI048071 from NIAID.

Author contributions:

- Conceptualization MC, AME, AK, ART, JCR, FES, DEN, VC
- Data Curation MC, AME, DFE, VC
- Formal Analysis MC, AME, ART
- Funding Acquisition TA, COB, JCR, FES, DEN, VC
- Investigation MC, AME, AK, ART, SA, PC
- Methodology MC, AME, ART
- Project Administration SA, JER, DEN, VC
- Resources EM, SA, TA, DFE, JCR, FES, DEN, VC
- Supervision TA, COB, DEN, VC
- Visualization MC, AME, ART
- Writing - Original Draft Preparation MC, AME
- Writing - Review & Editing [All authors]

Competing interests

The authors declare no competing interests.

Data availability

Whole genome sequence data for Ecuador samples are available on the Sequence Read Archive as BioProject PRJNA759192. Whole Genome sequence data from Colombia samples are available on the European Nucleotide Archive (Supplementary Table S3). Recoded GoldenGate calls from Echeverry, *et al* 2013 are available in Supplementary Table S4.

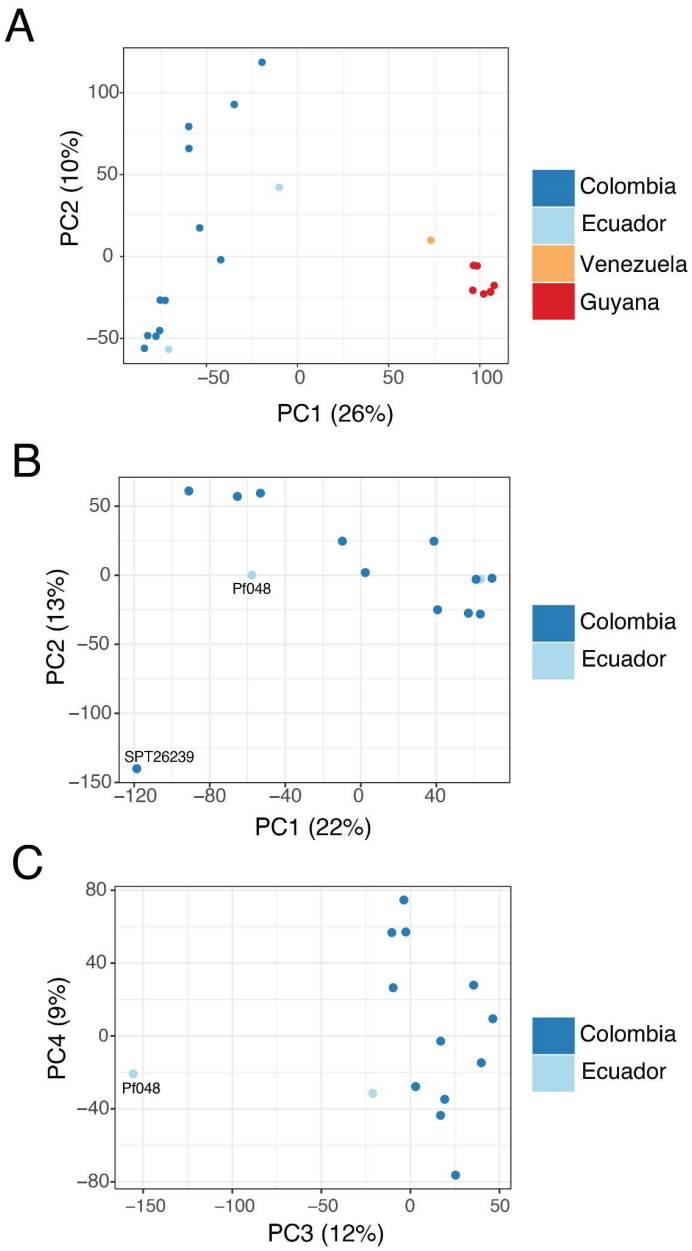
Supplementary Tables:

Supplementary Table S1. IDs and metadata for all WGS samples used in study

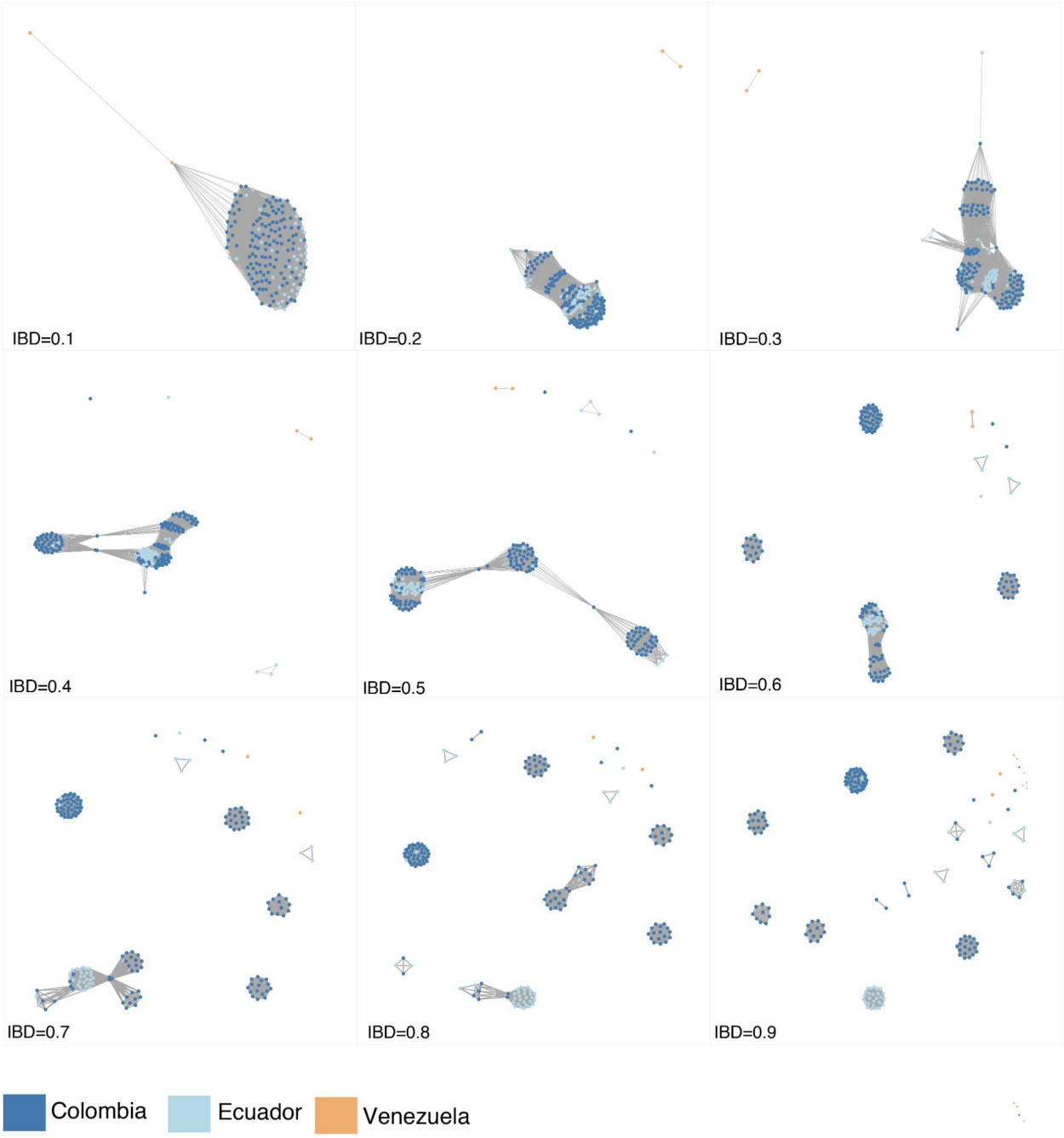
Supplementary Table S2. Information on clonal clusters defined in study

Supplementary Table S3. WGS sample accession numbers for Colombia

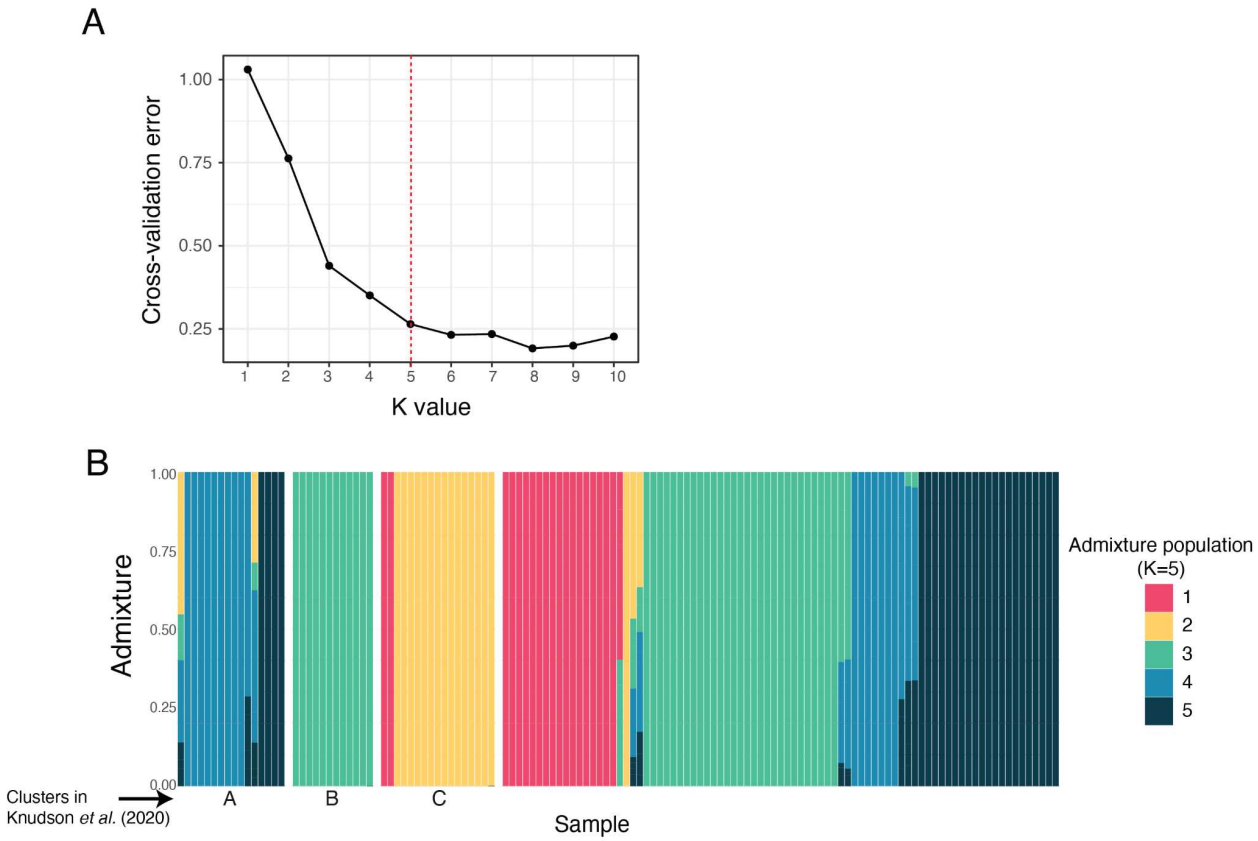
Supplementary Table S4. Recoded genotype information from Echeverry, *et al* 2013.



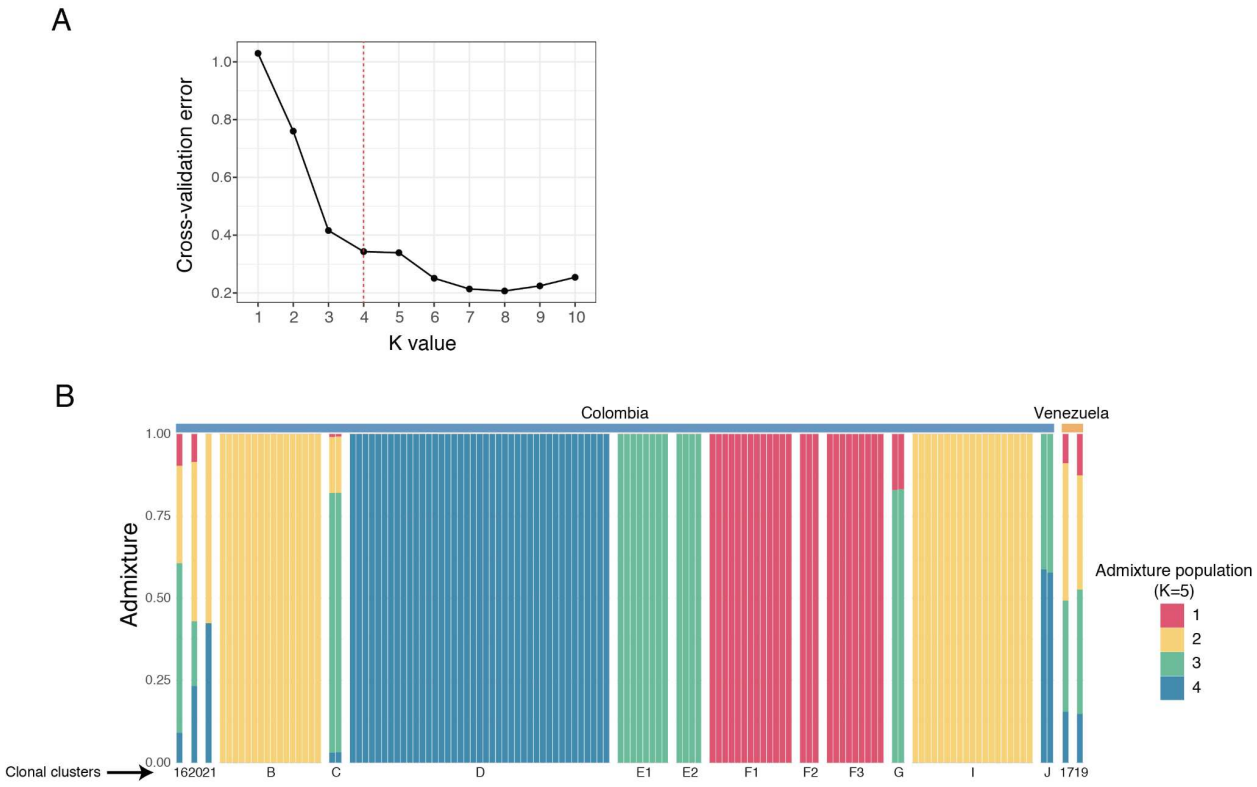
Supplementary figure 1. Principal component analysis of *P. falciparum* genomes obtained in different geographical regions of South America. (A) Parasite samples from the East of the Andes mountain range (Guyana and Venezuela) are separated by principal component 1 from parasites from the West (Colombia and Ecuador). (B) Parasite sample SPT26239 is separated from other Ecuador and Colombia samples by principal component 2. (C) Parasite samples from Colombia and Ecuador are separated from Pf048 by principal component 3.



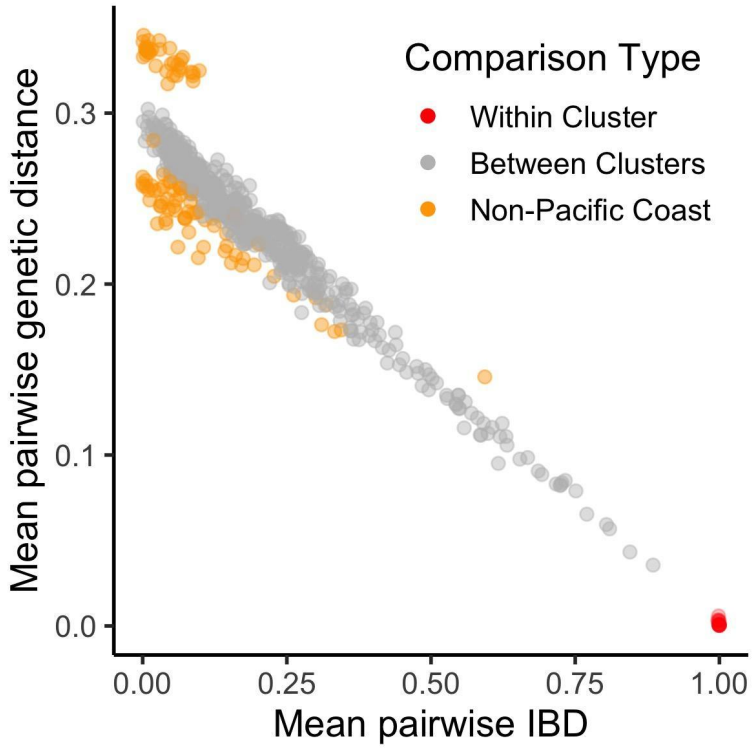
Supplementary Figure 2. Network analysis of pairwise relatedness visualized using incremental thresholds of IBD, ranging from 0.1 to 0.9, and colored by country included in the analysis from the sampling period 2013-2017.



Supplementary Figure 3. ADMIXTURE groups shown in Figure 3 and sample equivalency to previously reported groups using STRUCTURE. (A) Elbow method to identify the optimal number of ancestral populations from ADMIXTURE (K) based on cross-validation error. Red line indicates K=5 ancestral populations selected following best practices. (B) ADMIXTURE results highlighting the three ancestral populations identified in Knudson *et al.* 2020 with STRUCTURE[18].

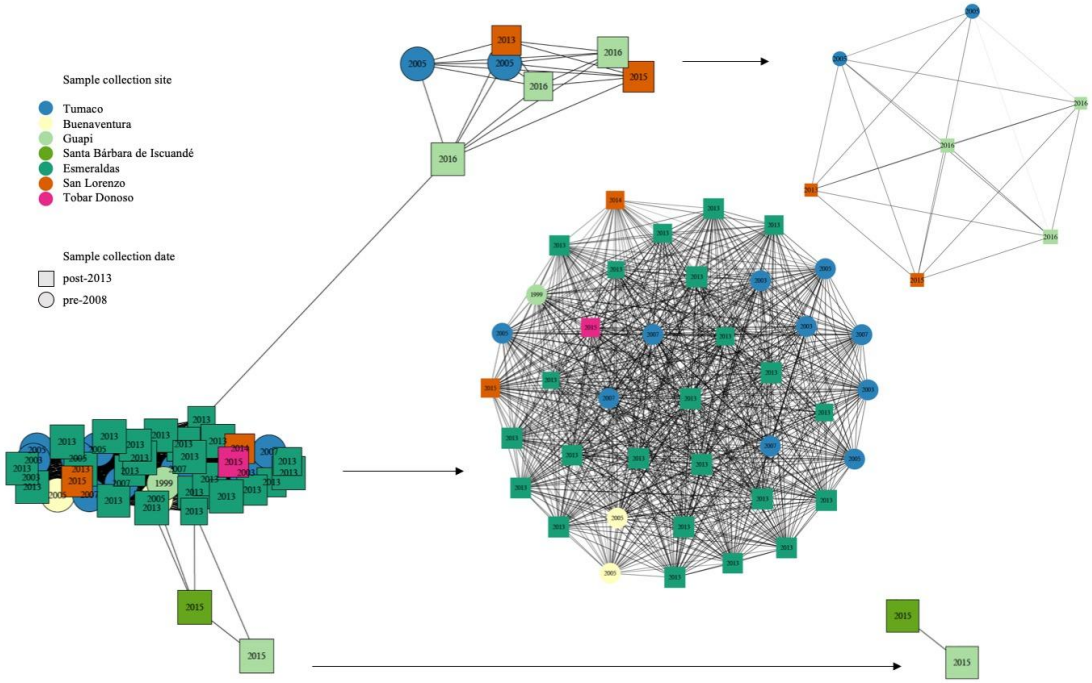


Supplementary Figure 4. ADMIXTURE groups using altered clonal clusters frequencies (A) Elbow method to identify the optimal number of ancestral populations from ADMIXTURE (K) based on cross-validation error. (B) ADMIXTURE analysis of the Colombia-Venezuela samples with altered cluster frequencies. Cluster C was decreased from a frequency of 0.09 to 0.016 and cluster E1 was increased from a frequency of 0.015 to 0.06. This analysis best supports a division into four groups (*K*=4), of which only two fully match with the original analysis. As expected, the cluster that increased in frequency (E1) is fully ascribed to a single ADMIXTURE group (along with E2) whereas the cluster that decreased in frequency (C) is described as “admixed”. In addition, clusters B and I are now included together in a single group. Increasing the group number to five (as was used in the original analysis), perpetuates these differences with the additional fifth cluster comprising the two Venezuelan samples.

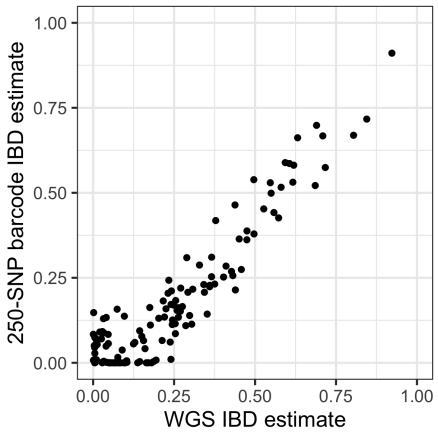


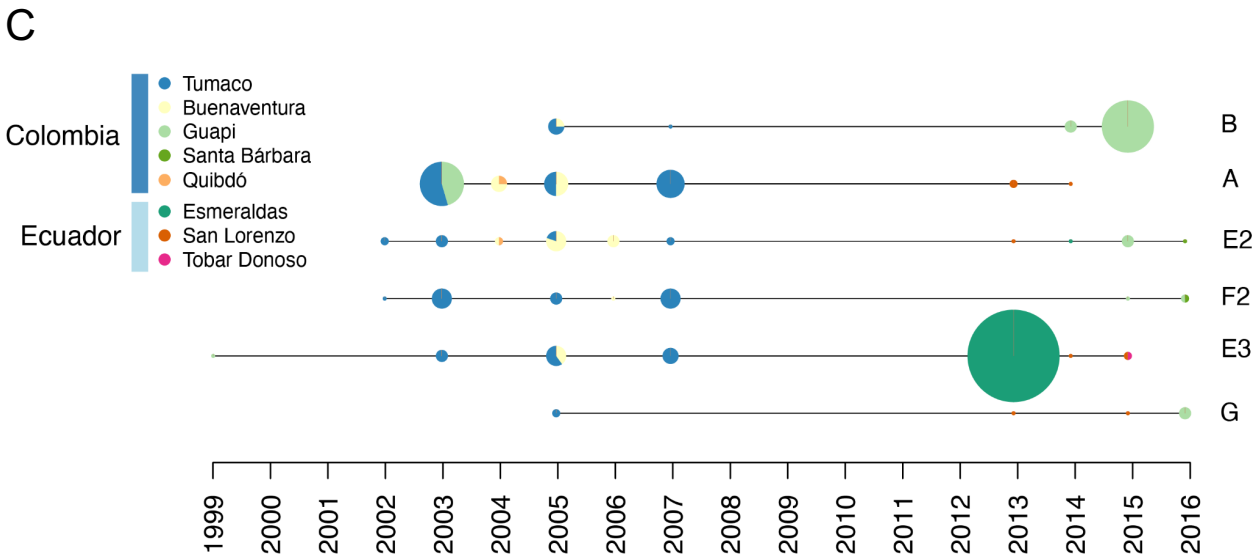
Supplementary Figure 5. IBD and genetic distance calculations are highly correlated within the Pacific Coast sample set at the cluster level (Pearson's r = -0.96). Mean genetic distance was calculated as the proportion of dissimilar calls within a set of 28,278 high-confidence SNPs. Both mean pairwise genetic distance and mean pairwise IBD were averaged over all inter-sample comparisons between each pair of clusters. Estimates for pairwise comparisons incorporating either of the two Venezuelan samples diverge to a greater extent, perhaps reflecting the inaccuracy of IBD estimation for samples originating outside the focal population. This is expected since population-level allele frequency estimates are required for IBD calculations, and these likely differ for the samples' true population of origin.

A

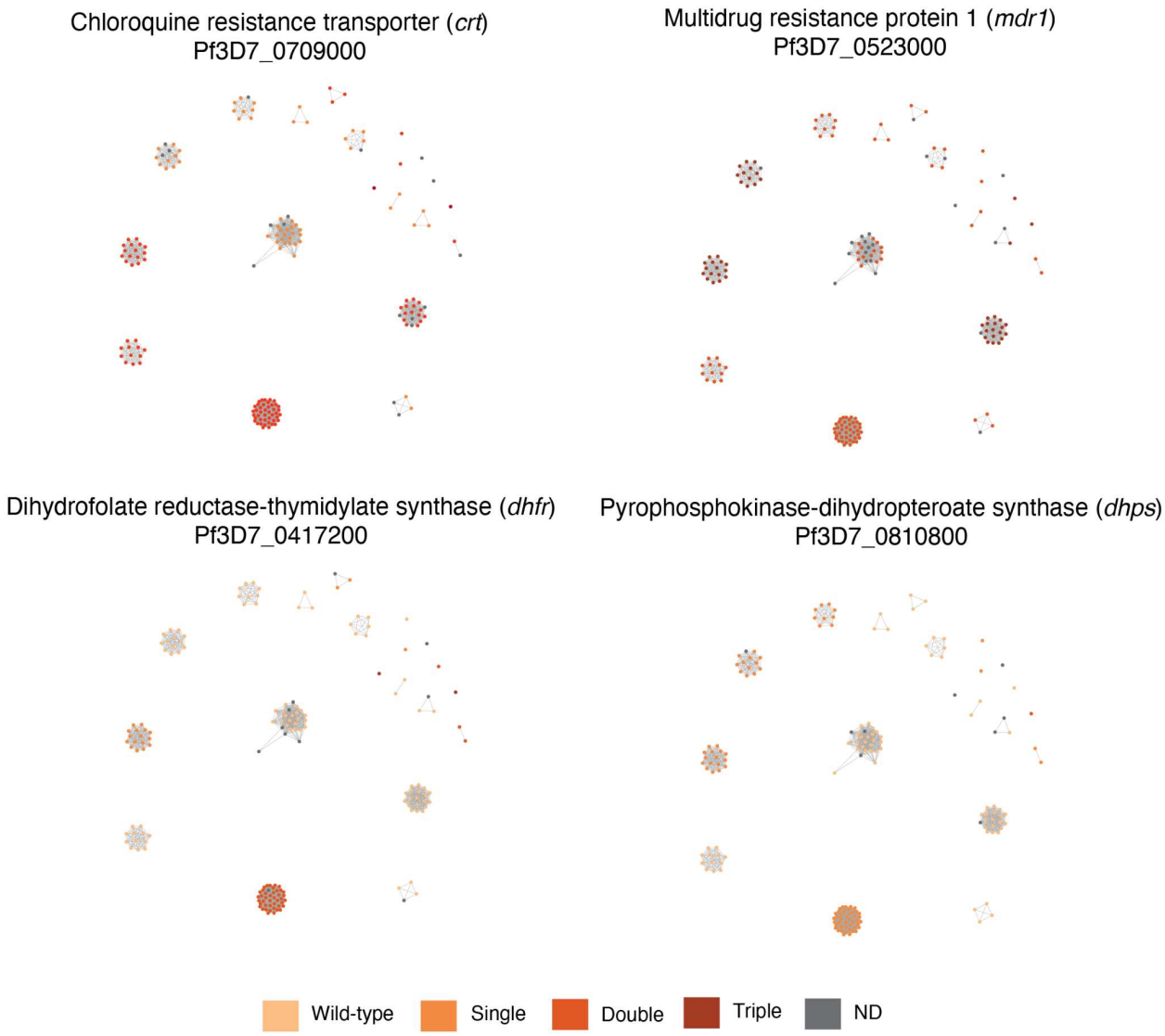


B

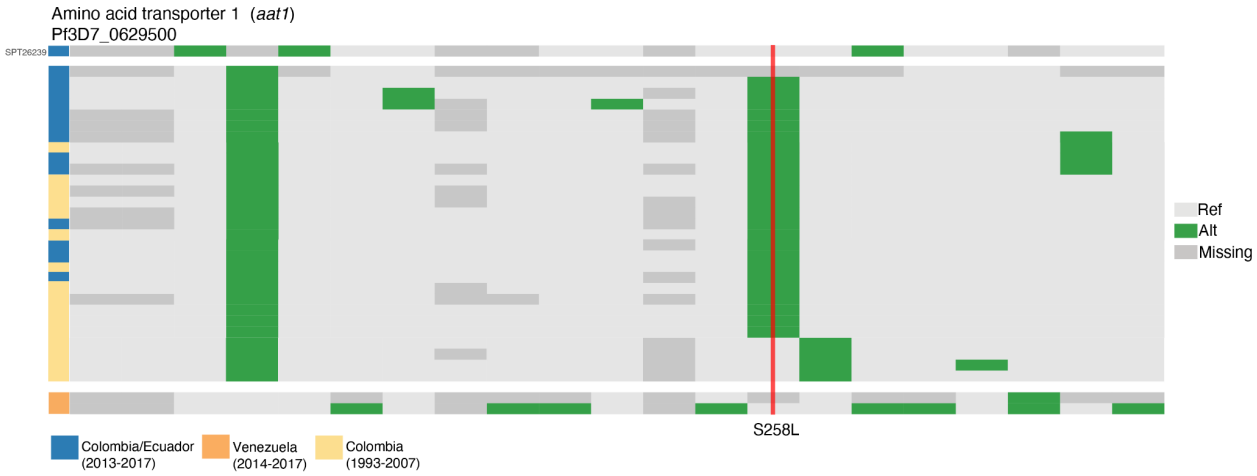




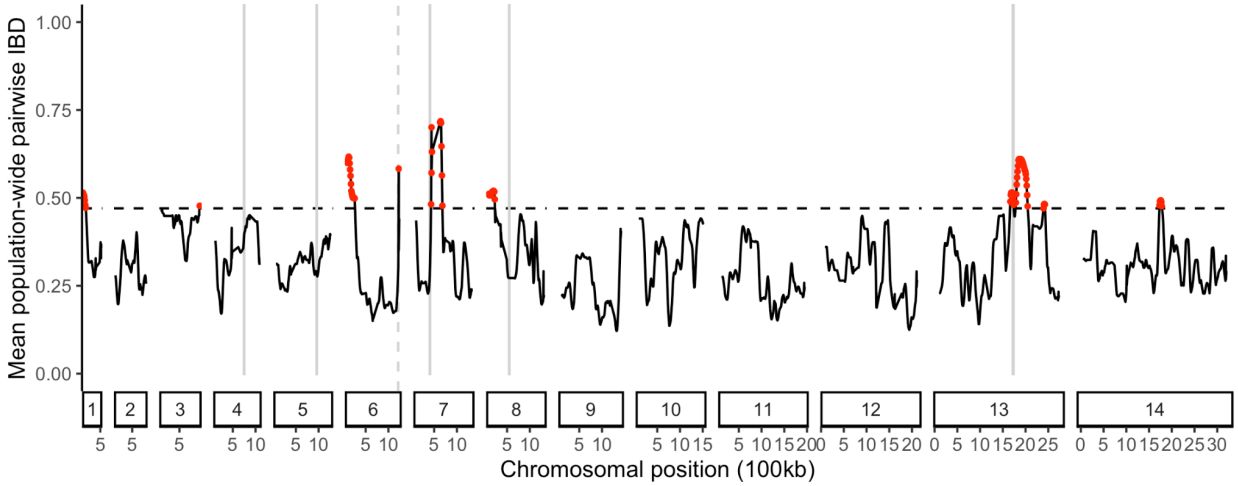
Supplementary Figure 6. IBD analysis performs robustly across samples with different genotyping approaches. (A) In order to minimize cliques within components in the extended analysis, one clonal component that contained six cliques was broken down into three clusters: the largest containing two cliques, the other two continuing a single clique each. (B) IBD point estimates between clonal clusters were obtained with both full WGS data (x-axis) and using information from up to 249 sites (y-axis). These estimates are highly correlated (Pearson's r = 0.93). Values are calculated as the mean IBD of all between-cluster pairwise comparisons. (C) Exact sampling locations for clonal clusters depicted in Figure 4.



Supplementary Figure 7. Network plot of clonal clusters (IBD≥0.99) harboring distinct haplotypes for four main genes involved in antimalarial drug resistance. Top panels correspond to *crt* (left) and *mdr1* (right), involved in resistance to chloroquine. Bottom panels are for *dhfr* (left) and *dhps* (right), involved in resistance to pyrimethamine and sulfadoxine, respectively. No clusters show signs of *de novo* mutations at known resistance alleles.



Supplementary Figure 8. Plot displaying haplotype surrounding the gene amino acid transporter 1 (aat1). Samples carrying ancestral alleles isolated in Guapi (SPT26239) and two samples originating in Venezuela are shown at the top and bottom, respectively.



Supplementary Figure 9. Mean population-wide pairwise IBD between clusters with origin in Ecuador and Colombia (2013-2017) calculated within 50-kb windows. Each cluster was represented by the sample member with the most complete genome coverage. The dashed horizontal line marks the genome-wide 95th percentile. Windows falling above this threshold are marked with red points. Vertical lines mark known resistance loci: *dhfr, mdr1*, *aat1*, *crt*, *dhps* and *kelch13*.

References

1. World Health Organization. World Malaria Report 2021 (2021).
2. Recht, J. *et al.* Malaria in Brazil, Colombia, Peru and Venezuela: current challenges in malaria control and elimination. *Malar. J.* **16**, 273 (2017).
3. PAHO. Situation of Malaria in the Region of the Americas. (2017).
4. Feged-Rivadeneira, A. *et al*. Malaria intensity in Colombia by regions and populations. *PLoS One* **13**, e0203673 (2018).
5. Douine, M. *et al.* Prevalence of *Plasmodium* spp. in illegal gold miners in French Guiana in 2015: a hidden but critical malaria reservoir. *Malar. J.* **15**, 315 (2016).
6. Salazar, P. M. D. *et al.* The association between gold mining and malaria in Guyana: a statistical inference and time-series analysis. *Lancet Plan. Health*. **5**, e731–e738 (2021).
7. Castellanos, A. *et al.* Malaria in gold-mining areas in Colombia. *Mem. Inst. Oswaldo Cruz* **111**, 59–66 (2016).
8. Grillet, M. E. *et al.* Venezuela's humanitarian crisis, resurgence of vector-borne diseases, and implications for spillover in the region. *Lancet Infect. Dis.* **19**, e149–e161 (2019).
9. Fidock, D. A. *et al.* Mutations in the *P. falciparum* Digestive Vacuole Transmembrane Protein PfCRT and Evidence for Their Role in Chloroquine Resistance. *Mol. Cell* **6**, 861–871 (2000).
10. Cortese, J. F. *et al*. Origin and Dissemination of *Plasmodium falciparum* Drug-Resistance Mutations in South America. *J. Infect. Dis.* **186**, 999–1006 (2002).
11. Wootton, J. C. *et al.* Genetic diversity and chloroquine selective sweeps in *Plasmodium falciparum. Nature* **418**, 320–323 (2002).
12. Mathieu, L. C. *et al.* Local emergence in Amazonia of *Plasmodium falciparum* k13 C580Y mutants associated with in vitro artemisinin resistance. *Elife* **9**, (2020).
13. Ramirez, A. P. *et al*. Frequency and tendency of malaria in Colombia, 1990 to 2011: a descriptive study. *Malar. J.* **13**, 202 (2014).
14. Rodríguez, J. C. P. *et al*. Epidemiology and control of malaria in Colombia. *Mem. Inst. Oswaldo Cruz* **106**, 114–122 (2011).
15. Yalcindag, E. *et al.* Multiple independent introductions of *Plasmodium falciparum* in South America. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 511–516 (2012).
16. Echeverry, D. F. *et al.* Long term persistence of clonal malaria parasite *Plasmodium falciparum* lineages in the Colombian Pacific region. *BMC Genet.* **14**, 2 (2013).
17. Sáenz, F. E. *et al.* Clonal population expansion in an outbreak of *Plasmodium falciparum* on the northwest coast of Ecuador. *Malaria Journal* vol. 14 (2015).
18. Knudson, A. *et al.* Spatio-temporal dynamics of *Plasmodium falciparum* transmission within a spatial unit on the Colombian Pacific Coast. *Sci. Rep.* **10**, 3756 (2020).
19. Charlesworth, B. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat. Rev. Genet.* **10**, 195–205 (2009).
20. Mita, T., Tanabe, K. & Kita, K. Spread and evolution of *Plasmodium falciparum* drug resistance. *Parasitol. Int.* **58**, 201–209 (2009).
21. Anderson, T. J. *et al.* Microsatellite markers reveal a spectrum of population structures in the malaria parasite *Plasmodium falciparum. Mol. Biol. Evol.* **17**, 1467–1482 (2000).
22. Jacob, C. G. *et al.* Genetic surveillance in the Greater Mekong subregion and South Asia to support malaria control and elimination. *Elife* **10**, (2021).

23. Padilla-Rodríguez, J. C. *et al.* Malaria risk stratification in Colombia 2010 to 2019. *PLoS One* **16**, e0247811 (2021).
24. Vera-Arias, C. A., Enrique Castro, L., Gómez-Obando, J.*et al*. Diverse origin of *Plasmodium falciparum* in northwest Ecuador. *Malar. J.* **18** (2019).
25. Ministerio de Salud y Protección Social. Boletín epidemiológico semana 52 de 2016. *Boletín epidemiológico semanal* (2016).
26. Ministerio de Salud y Protección Social. Boletín epidemiológico semana 52 de 2017. *Boletín epidemiológico semanal* (2017).
27. Ministerio de Salud y Protección Social. Boletín epidemiológico semana 52 de 2015. *Boletín epidemiológico semanal* (2020).
28. Instituto Nacional de Salud. *Sistema Nacional de Vigilancia en Salud Publica* (Portal SIVIGILA: http://portalsivigila.ins.gov.co).
29. Ministerio de Salud Pública del Ecuador. Reporte de Datos de Malaria del periodo 2008 al inicios del 2018. Quito. *Ministerio de Salud Pública del Ecuador* (2018).
30. Oyola, S. O. *et al.* Whole genome sequencing of *Plasmodium falciparum* from dried blood spots using selective whole genome amplification. *Malar. J.* **15**, 597 (2016).
31. Taylor, A. R. *et al*. Identity-by-descent with uncertainty characterises connectivity of *Plasmodium falciparum* populations on the Colombian-Pacific coast. *PLoS Genet.* **16**, e1009101 (2020).
32. Ruybal-Pesántez, S. *et al.* Clinical malaria incidence following an outbreak in Ecuador was predominantly associated with *Plasmodium falciparum* with recombinant variant antigen gene repertoires. *bioRxiv* (2021) doi:10.1101/2021.04.12.21255093.
33. Henden, L. *et al*. Identity-by-descent analyses for measuring population dynamics and selection in recombining pathogens. *PLoS Genet.* **14**, e1007279 (2018).
34. MalariaGEN *et al.* An open dataset of *Plasmodium falciparum* genome variation in 7,000 worldwide samples. *Wellcome Open Research* **6**, (2021).
35. Alexander, D. H. *et al*. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
36. Taylor, A. R. *et al*. Estimating Relatedness Between Malaria Parasites. *Genetics* **212**, 1337–1351 (2019).
37. Watson, J. A. *et al.* A cautionary note on the use of unsupervised machine learning algorithms to characterise malaria parasite population structure from genetic distance matrices. *PLoS Genet.* **16**, e1009037 (2020).
38. Corredor, V. *et al*. Origin and dissemination across Colombian Andes mountain range of *Plasmodium falciparum* Sulfadoxine-Pyrimethamine resistance. *Antimicrob. Agents Chemother.* **54,** 3121-3125 (2010).
39. Ministerio de Salud y Protección Social. Informe sobre el estado actual del programa de erradicación de la malaria en Colombia. *Servicio Nacional de Erradicación de la Malaria* (1969).
40. Pelleau, S. *et al.* Adaptive evolution of malaria parasites in French Guiana: Reversal of chloroquine resistance by acquisition of a mutation in pfcrt. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 11672–11677 (2015).
41. Echeverry, D. F. *et al.* Short report: polymorphisms in the pfcrt and pfmdr1 genes of *Plasmodium falciparum* and in vitro susceptibility to amodiaquine and desethylamodiaquine.

Am. J. Trop. Med. Hyg. **77**, 1034–1038 (2007).

42. Cowell, A. N. *et al.* Mapping the malaria parasite druggable genome by using *in vitro* evolution and chemogenomics. *Science* **359**, 191–199 (2018).
43. Modrzynska, K. *et al.* Quantitative genome re-sequencing defines multiple mutations conferring chloroquine resistance in rodent malaria. *BMC Genomics* **13**, 106 (2012).
44. Wang, Z. *et al.* Genome-wide association analysis identifies genetic loci associated with resistance to multiple antimalarials in *Plasmodium falciparum* from China-Myanmar border. *Sci. Rep.* **6**, 33891 (2016).
45. Amambua-Ngwa, A. *et al.* Major subpopulations of *Plasmodium falciparum* in sub-Saharan Africa. *Science*. **365,** 813–816 (2019).
46. Cerqueira, G. C. *et al.* Longitudinal genomic surveillance of *Plasmodium falciparum* malaria parasites reveals complex genomic architecture of emerging artemisinin resistance. *Genome Biol.* **18**, 78 (2017).
47. WWARN K13 Genotype-Phenotype Study Group. Association of mutations in the *Plasmodium falciparum* Kelch13 gene (Pf3D7_1343700) with parasite clearance rates after artemisinin-based treatments—a WWARN individual patient data meta-analysis. *BMC Medicine.* **17,** 1 (2019).
48. Vanaerschot, M. *et al.* Inhibition of Resistance-Refractory *P. falciparum* Kinase PKG Delivers Prophylactic, Blood Stage, and Transmission-Blocking Antiplasmodial Activity. *Cell Chem Biol* **27**, 806–816.e8 (2020).
49. Ross, L. S. & Fidock, D. A. Elucidating Mechanisms of Drug-Resistant *Plasmodium falciparum. Cell Host Microbe* **26**, 35–47 (2019).
50. Hoepfner, D. *et al.* Selective and specific inhibition of the *Plasmodium falciparum* lysyl-tRNA synthetase by the fungal secondary metabolite cladosporin. *Cell Host Microbe* **11**, 654–663 (2012).
51. Manary, M. J. *et al.* Identification of pathogen genomic variants through an integrated pipeline. *BMC Bioinformatics* **15**, 63 (2014).
52. Molina-Cruz, A. *et al. Plasmodium* evasion of mosquito immunity and global malaria transmission: The lock-and-key theory. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 15178–15183 (2015).
53. Comer, R. D., Young, M. D., Porter, J. A., Jr, Gauld, J. R. & Merritt, W. Chloroquine resistance in *Plasmodium falciparum* malaria on the Pacific coast of Colombia. *Am. J. Trop. Med. Hyg.* **17**, 795–799 (1968).
54. Aguilar-Velasco, H. M. Malaria y espacio en el Ecuador del verde de París a la eliminación de la enfermedad. (2021).
55. Brown, K. M. *et al.* Compensatory mutations restore fitness during the evolution of dihydrofolate reductase. *Mol. Biol. Evol.* **27**, 2682–2690 (2010).
56. Okombo, J. et al. Temporal trends in prevalence of *Plasmodium falciparum* drug resistance alleles over two decades of changing antimalarial policy in coastal Kenya. *Int. J. Parasitol. Drugs Drug. Resist.* **4**, 152–163 (2014).
57. Ocan, M. *et al.* Persistence of chloroquine resistance alleles in malaria endemic countries: a systematic review of burden and risk factors. *Malar. J.* **18**, 76 (2019).
58. Chang, H.-H. *et al.* Malaria life cycle intensifies both natural selection and random genetic drift. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 20129–20134 (2013).

59. Anderson, T. J. C. *et al.* Population Parameters Underlying an Ongoing Soft Sweep in Southeast Asian Malaria Parasites. *Mol. Biol. Evol.* **34**, 131–144 (2017).
60. Tessema, S. K. *et al.* Sensitive, highly multiplexed sequencing of microhaplotypes from the *Plasmodium falciparum* heterozygome. *J. Infect. Dis.* (2020) doi:10.1093/infdis/jiaa527.
61. LaVerriere, E. *et al.* Design and implementation of multiplexed amplicon sequencing panels to serve genomic epidemiology of infectious disease: a malaria case study. *bioRxiv* (2021) doi:10.1101/2021.09.15.21263521.
62. Shetty, A. C. *et al.* Genomic structure and diversity of *Plasmodium falciparum* in Southeast Asia reveal recent parasite migration patterns. *Nat. Commun.* **10**, 2665 (2019).
63. Verity, R. *et al.* The impact of antimalarial resistance on the genetic structure of *Plasmodium falciparum in the DRC. Nat. Commun.* **11**, 2107 (2020).
64. Taylor, A. R. *et al.* Quantifying connectivity between local *Plasmodium falciparum* malaria parasite populations using identity by descent. *PLoS Genet.* **13**, e1007065 (2017).
65. Buyon, L. E. *et al.* Population genomics of *Plasmodium vivax* in Panama to assess the risk of case importation on malaria elimination. *PLoS Negl. Trop. Dis.* **14**, e0008962 (2020).
66. Daniels, R. *et al.* Modeling malaria genomics reveals transmission decline and rebound in Senegal. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 7067–7072 (2015).
67. Dalmat, R., Naughton, B., Kwan-Gett, T. S., Slyker, J. & Stuckey, E. M. Use cases for genetic epidemiology in malaria elimination. *Malar. J.* **18**, 163 (2019).
68. Sy, M. *et al.* Genomic investigation of atypical malaria cases in Kanel, northern Senegal. *Malar. J.* **20**, 103 (2021).
69. General Assembly of the World Medical Association and others. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *General Assembly of the World Medical Association and others.*
70. Presidencia de la República de Colombia. Código del menor. Decreto 2737 de 1989. *Presidencia, República de Colombia* **9**, (1989).
71. Ministerio de Salud y Protección Social. Scientific, technical and administrative guidelines for health research. *Resolution 8430 from 1993* **19**, (1993).
72. Trager, W. & Jensen, J. B. Human malaria parasites in continuous culture. *Science* **193**, 673–675 (1976).
73. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv q-bio.GN* (2013).
74. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
75. Miles, A. *et al.* Indels, structural variation, and recombination drive genomic diversity in Plasmodium falciparum. *Genome Res.* **26**, 1288–1299 (2016).
76. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
77. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
78. Schaffner, S. F. *et al*. hmmIBD: software to infer pairwise identity by descent between haploid genotypes. *Malar. J.* **17**, 196 (2018).
79. Csardi, G. *et al*. The igraph software package for complex network research. *Inter. J. Com. Syst.* **1695**, 1–9 (2006).

80. Aurrecoechea, C. *et al.* EuPathDB: the eukaryotic pathogen genomics database resource. *Nucleic Acids Res.* **45**, D581–D591 (2017).
81. Amos, B. *et al.* VEuPathDB: the eukaryotic pathogen, vector and host bioinformatics resource center. *Nucleic Acids Res.* **50**, D898–D911 (2022).