Identity-by-descent relatedness estimates with uncertainty characterise departure from isolation-by-distance between *Plasmodium falciparum* populations on the Colombian-Pacific coast

Aimee R. Taylor[*,1,2], Diego F. Echeverry[3,4,5], Timothy J. C. Anderson[6], Daniel E. Neafsey[2,7], Caroline O. Buckee[1]

[*]Corresponding author: ataylor@hsph.harvard.edu

1 Center for Communicable Disease Dynamics, Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, Massachusetts, USA
2 Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA
3 Centro Internacional de Entrenamiento e Investigaciones Médicas (CIDEIM), Cali, Colombia
4 Universidad Icesi, Cali, Colombia
5 Departamento de Microbiologia, Facultad de Salud, Universidad del Valle, Cali, Colombia
6 Department of Genetics, Texas Biomedical Research Institute, San Antonio, Texas, USA
7 Department of Immunology and Infectious Diseases, Harvard T. H. Chan School of Public Health, Boston, Massachusetts, USA

Keywords: malaria; *Plasmodium falciparum*; relatedness; identity-by-descent; molecular epidemiology; spatial connectivity; isolation-by-distance

Abstract

Characterising connectivity between geographically separated biological populations is a common goal in many fields. Recent approaches to understanding connectivity between malaria parasite populations, with implications for disease control efforts, have used estimates of relatedness based on identity-by-descent (IBD). However, uncertainty around estimated relatedness has not been accounted for to date. IBD-based relatedness estimates with uncertainty were computed for pairs of monoclonal *Plasmodium falciparum* samples collected from five cities on the Colombian-Pacific coast where long-term clonal propagation of *P. falciparum* is frequent. The cities include two official ports, Buenaventura and Tumaco, that are separated geographically but connected by frequent marine traffic. The fraction of highly-related sample pairs (whose classification accounts for uncertainty) was greater within cities versus between. However, based on both the fraction of highly-related sample pairs and on a threshold-free approach (Wasserstein distances between parasite populations) connectivity between Buenaventura and Tumaco was disproportionally high. Buenaventura-Tumaco connectivity was consistent with three separate transmission events involving parasites from five different clonal components (groups of statistically indistinguishable parasites identified under a graph theoretic framework). To conclude, *P. falciparum* population connectivity on the Colombian-Pacific coast abides by accessibility not isolation-by-distance, potentially implicating marine traffic in malaria transmission with opportunities for targeted intervention. Further investigations are required to test this and alternative hypotheses. For the first time in malaria epidemiology, we account for uncertainty around estimated relatedness (an important consideration for future studies that plan to use genotype versus whole genome sequence data to estimate IBD-based relatedness); we also use a threshold-free approach to compare parasite populations, and identify clonal components in a statistically principled manner. The approaches we employ could be adapted to other recombining organisms with mixed mating systems, thus have broad relevance.

Introduction

In many research fields genetic data are used to help characterise connectivity between geographically disparate biological populations, with numerous applications in conservation, agriculture, and public health. Patterns of genetic similarity between pathogen populations help us understand how the disease spreads. Patterns of relatedness (a measure of genetic similarity) between malaria parasites in different human populations, for instance, help characterise the connectivity between them, thus guide the design of targeted public health interventions [1].

Several methods are employed to measure genetic similarity and thus characterise connectivity. Phylogenetic methods, in which genetic distances between individuals are measured in units of mutation [2], are most applicable to rapidly mutating organisms that do not recombine (e.g RNA viruses) [3]. Studies of relatedness, in which relatedness is a measure of probability of identity-by-descent (IBD) between individuals, are applicable to organisms that do recombine (e.g. malaria parasites). Population genetic parameters of allelic variation (e.g. F_{ST}) are applicable to all organisms (those that do and do not recombine), but do not generate measures of genetic distance or similarity on an inter-individual level, thus provide less granularity. Moreover, among recombining organisms, inter-population allelic variation tends to accumulate more slowly than inter-individual variation in IBD [4]. As such, analyses of relatedness can sometimes nearby and recent connectivity where analyses of F_{ST} cannot [5].

Malaria parasites are protozoan parasites that undergo an obligate stage of sexual recombination in the mosquito midgut. Like many organisms (e.g. many plants [6, 7]), malaria parasites have a mixed mating system that encompasses both inbreeding and outcrossing. The extent to which malaria parasites outcross depends on transmission intensity and is not fully understood [8]. In any event, for outcrossing to occur a mosquito must ingest genetically distinct gametocytes. Humans can be infected by multiple genetically distinct parasite clones that are either co-transmitted via inoculation from a single mosquito, in which case they are likely recombinants so inter-related, or transmitted independently by multiple mosquitoes, in which case the parasite clones are likely unrelated [9, 10]. The latter can occur in a setting where the entomological inoculation rate is high; recent work suggests co-transmission is important in both low and high transmission settings [10].

Malaria genomic epidemiology studies of connectivity are increasingly common, especially in the context of public health and using genotype (versus whole genome sequence) data [5, 11–14]. Using IBD-based relatedness but not F_{ST}, evidence of isolation-by-distance among *P. falciparum* populations along a 100 km stretch of the Thailand-Myanmar border was found [5]. This study was based, in part, on analyses of monoclonal *P. falciparum* samples genotyped at 93 single nucleotide polymorphisms (SNPs). Based on F_{ST} estimated using *P. falciparum* samples genotyped at 250 SNPs, a different study found evidence of departure from isolation-by-distance among *P. falciparum* populations along a 500 km stretch of the Colombian-Pacific coast where transmission is mixed (low and high in some regions) and outcrossing limited [11, 15]. Departure from isolation-by-distance on the Colombian-Pacific coast was based on F_{ST} alone [11]. In the current study, we re-explore departure from isolation-by-distance with more granularity using IBD-based relatedness. For the first time in malaria epidemiology, we account for uncertainty in relatedness estimates; we also use a threshold-free approach to compare parasite populations, and identify clonal components in a statistically principled manner. The original study [11] and our response to it are described in more detail below.

Malaria epidemiology in Colombia is associated with a multitude of ecological, evolutionary and social factors (Table A.1), including human migration due to deforestation, illegal crops, gold mining [16–20], and the mass emigration of people fleeing the humanitarian crisis in Venezuela [21–24]. Understanding the interplay between e.g. migration, parasite population connectivity and the spread of antimalarial resistance is critical [16, 18]. In preparation for studies of resistance, Echeverry et al. genotyped *P. falciparum* samples from four provinces on the Colombian-Pacific coast [11]. Clonality, population structure and linkage disequilibrium (LD) were characterised using a suite of population genetic analyses. The results were highly informative: the vast majority of successfully genotyped *P. falciparum* samples were deemed monoclonal (325 of 400) with a strong association between clonality and incidence. Among the 325 monoclonal samples, 136 unique haploid multilocus genotypes (MLGs) were identified using relatedness based on identity-by-state (IBS), which is a correlate of IBD [25] (and has been used elsewhere to characterise connectivity between nearby malaria parasite populations [12–14]). Of the 136 MLGs, 44 infected two or more patients (max. 28 patients), 45 persisted for two or more days (max. 8 years), and 7 of the 15 most common MLGs were

sampled in two or more provinces (max. all four provinces). Panmixia was rejected based on evidence of four sympatric but geographically structured subpopulations; and, overall, LD decayed at a rate that was faster than expected for South American *P. falciparum* populations (compare with e.g. [26]). Echeverry et al. concluded that evidence of low genetic diversity, persistent MLGs and population structure is consistent with low transmission and limited outcrossing, while evidence of a relatively fast rate of LD decay and of shared MLGs across provinces is consistent with extensive human movement connecting *P. falciparum* populations.

Although the study by Echeverry et al. features analyses of IBS-based relatedness (i.e. MLGs), evidence of departure from isolation-by-distance was based on F_{ST} alone. To explore in more granularity while accounting for uncertainty, we compute IBD-based relatedness estimates and confidence intervals for all pairs of 325 monoclonal parasite samples. Akin to previous studies (e.g. [5]), highly-related parasites were classified using a threshold; however, confidence intervals allow uncertainty to be accounted for. This is important because relatedness estimated using limited genotype data can be overwhelmed by uncertainty [25]. Our approach includes two additional contributions. First, we complement our analysis of highly-related parasites with a threshold-free approach based on optimal transport using Wasserstein distances between parasite populations. Second, we identify groups of statistically indistinguishable parasites, which we call clonal components, using the simple concept of components from graph theory and confidence intervals. Confidence intervals circumvent reliance on an arbitrary clonal threshold (i.e. some number of differences tolerated between parasites samples considered clonal). Graph components circumvent reliance on unsupervised clustering methods that are notoriously brittle [27]. Overall, our approach could be adapted to viruses and bacteria that show recombination or reshuffling of segments as well as clonal propagation [28–31], to other protozoans (e.g. *Toxoplasma*, *Cryprosporidium* [32–34]), and to the many fungal pathogens [35], plants [6, 7], and animals with mixed mating systems. Due to our treatment of uncertainty, it is especially relevant for a growing number of studies that plan estimate IBD-based relatedness using genotype (versus sequence) data.

Results

Relatedness estimates between *P. falciparum* sample pairs

Relatedness was estimated for all 52650 pairwise comparisons of 325 previously published monoclonal *P. falciparum* samples with data on 250 biallelic single nucleotide polymorphisms (SNPs) [11]. The parasite samples were collected between 1993 and 2007 from symptomatic patients participating in studies at five cities on the Colombian-Pacific coast (Table A.2). Despite considerable uncertainty, all estimates are informative (Figure 1). That is to say, there are no relatedness estimates whose 95% confidence intervals span entirely from zero to one. The vast majority of relatedness estimates were classified unrelated.

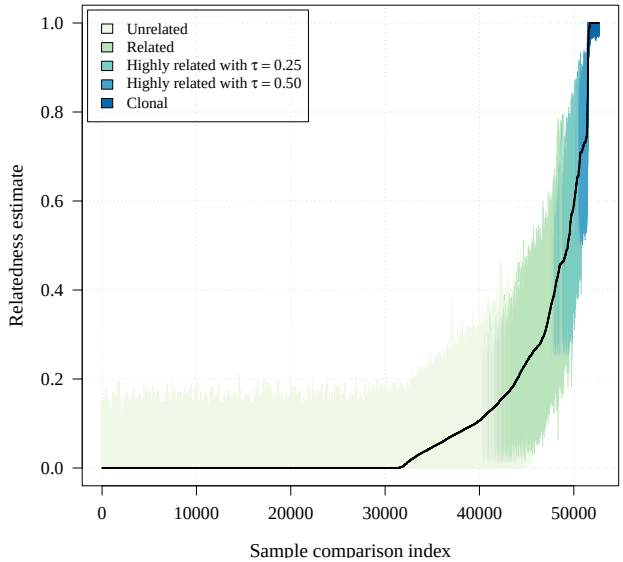


Figure 1: Estimates of relatedness with colour-coded 95% confidence intervals for all 325 choose two (52650) *P. falciparum* sample pairs in order of increasing relatedness estimate. Confidence intervals are coloured according to classifications based on lower and upper confidence interval bounds (Table 1).

Classification	Interpretation	Definition
Unrelated	$\hat{r}$ statistically indistinguishable from zero	LCI $< \epsilon$
Related	$\hat{r}$ statistically distinguishable from zero	LCI $> \epsilon$
highly-related	$\hat{r}$ statistically distinguishable from a specified threshold	LCI $> \tau$
Clonal	$\hat{r}$ statistically indistinguishable from one	UCI $> 1 - \epsilon$

Table 1: Classification of parasite sample pairs. Lower and upper confidence interval bounds (LCI and UCI, respectively) are used to classify pairs with $\epsilon = 0.01$, $\tau = 0.25$ (main analysis, Figure 2) and $\tau \in \{0.25, 0.50\}$ (sensitivity analysis, Figure A.1).

highly-related *P. falciparum* sample pair fractions partitioned in space and time

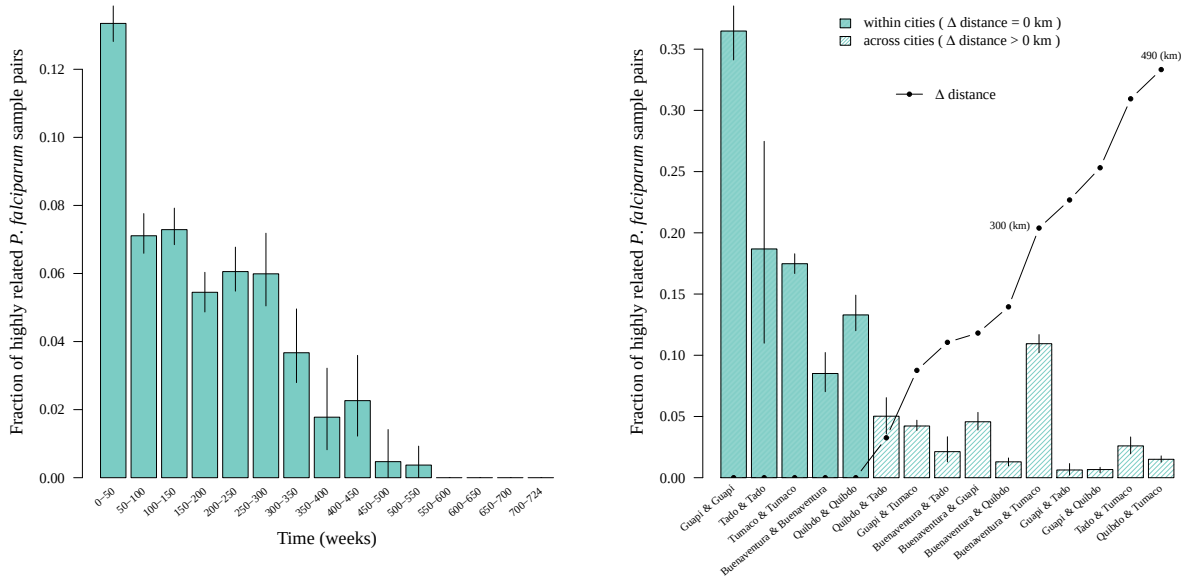
In our main analysis (Figure 2), highly-related parasite samples were classified using a high-relatedness threshold of 0.25 (Table 1). Despite few highly-related *P. falciparum* sample pairs overall, there are three notable observations regarding their fraction partitioned in space and time. First, there is a greater fraction of highly-related sample pairs among those collected close together in time versus far apart (Figure 2a). Second, the fraction of highly-related sample pairs is generally greater within cities versus between, with Guapi having the largest fraction of highly-related pairs and Buenaventura having the lowest (Figure 2b). However, third, the fraction shared between Buenaventura and Tumaco is exceptionally high considering inter-city distance (Figure 2b). These observations are largely robust to different high-relatedness thresholds (Figure A.1). Spatial trends evaluated using a threshold-free approach are also consistent: they show a general increase in 1-Wasserstein distance with geographic distance between cities besides Buenaventura and Tumaco (Figure 3). The 1-Wasserstein distance can be interpreted as the effort required to transform a distribution of parasite samples from one city into a distribution of parasite samples from another [16]. The small 1-Wasserstein distance between Buenaventura and Tumaco is thus consistent with elevated gene flow between *P. falciparum* populations sampled from these cities.

Figure 4a shows the inter-city *P. falciparum* population connectivity plotted in Figure 2b projected onto a map of the Colombian-Pacific coast. Buenaventura and Tumaco are the two largest official ports on the Colombian-Pacific coast (Buenaventura is the largest) and are connected by frequent marine traffic (Figure 4b). Although Tumaco is connected to Buenaventura via the pan-american highway, which connects all sites but Guapi, primary access to Tumaco is via the port due to difficult and unsafe country roads in Nariño. Guapi, which is effectively unreachable by road and not an official port, is connected by marine traffic but with less frequency (Figure 4b). Consistent with its isolation, the fraction of highly-related parasite pairs is relatively large within Guapi (Figure 2b), and very small between Guapi and the two inland cities, Quibdó and Tadó (Figures 2b and 4a). Moreover and importantly regarding the elevated fraction of highly-related samples pairs within both Guapi and Tadó (Figures 2b), all samples from Guapi and Tadó were collected within a single year (Table A.2). The low fraction of highly-related parasite sample pairs within Buenaventura (Figure 2b) is consistent with it having contributed samples over many years (Table A.2) and with it being the most important port on the Pacific coast (Figure 4b), i.e. a hub through which human traffic and thus potential parasite mixing is high [11].

The apparent association between *P. falciparum* population connectivity and the frequency of marine traffic raises questions about the latter's role in malaria transmission. However, other scenarios could lead to these relationships, for example the high connectivity could result from a single travel event between Buenaventura and Tumaco, followed by expansion of highly-related and clonal parasites. To further explore the genetic signal that supports this association we next consider clonal components.

Clonal components

We define clonal components as groups of statistically indistinguishable parasite samples identified under a graph theoretic framework; see Methods. In total, 46 distinct clonal components were detected, ranging in size from 2 to 28 statistically indistinguishable parasite samples (Figure 5). They are spatially clustered. Ten of the 46 contain parasite samples collected from two or more cities. Each clonal component besides one (clonal component four) is on average related to at least one other (Figure 5). The unrelated clonal



(a) Partitioned by time between collection dates.

(b) Partitioned by collection city.

Figure 2: Fractions of highly-related sample pairs partitioned in time and space.

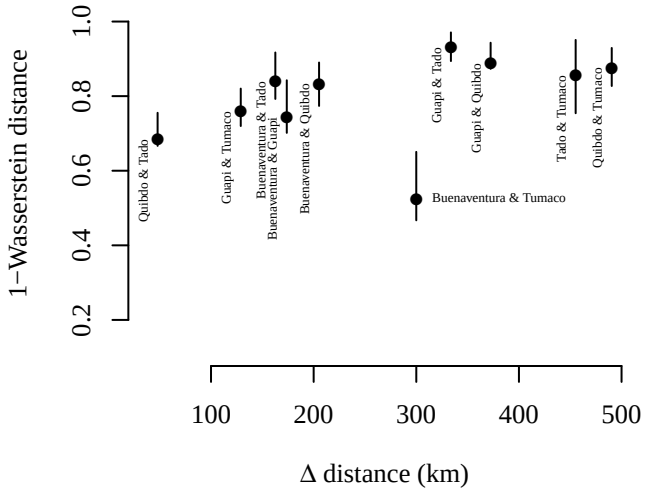
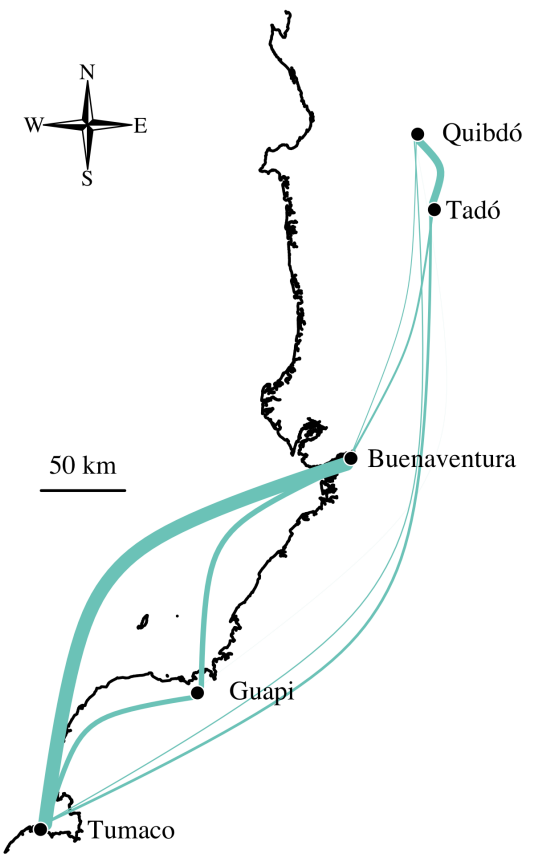


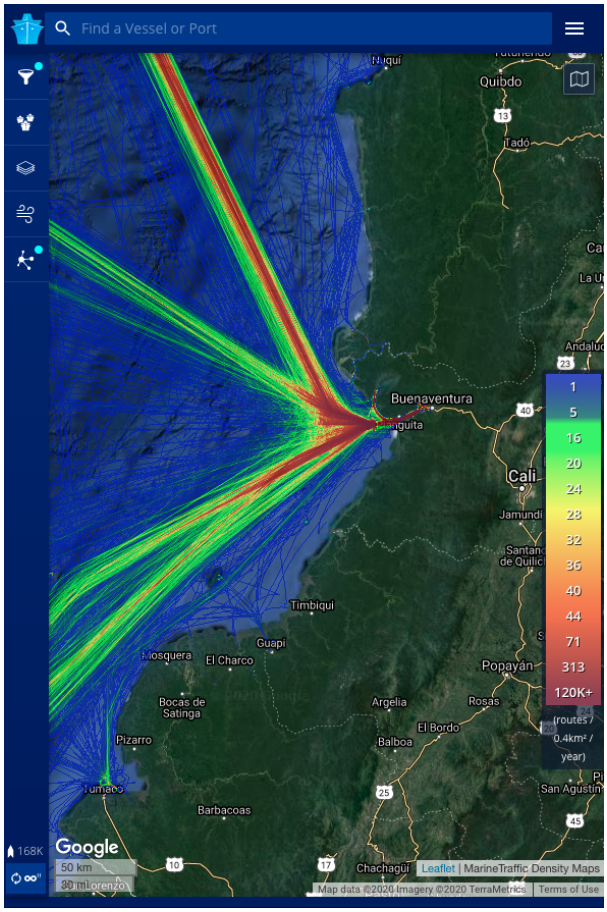
Figure 3: 1-Wasserstein distance between parasite populations across cities versus inter-city distance in kilometres (km).

component is likely a contaminant: it accords with MLG 036 reported in [11], where contamination during *in vitro* adaptation or DNA manipulation was suspected.

Clonal parasite samples detected in both Buenaventura and Tumaco belong to five distinct clonal components (one, 12, 14, 20 and 40, Figure 5). We thus dismiss a single travel event connecting Buenaventura and Tumaco involving a single parasite clone. We cannot dismiss a single travel event involving multiple parasite clones, however. Indeed, three of the five clonal components are inter-related on average (Table A.3). As such, they could derive from co-transmitted recombinant parasites transported in a single individual with a multiclonal infection. On the contrary, the remaining two clonal components have relatedness estimates that are not statistically distinguishable from zero. As such, they likely derive from different individuals with independent monoclonal infections. Given dates and cities of first detection (Table A.4), it is tempting to suggest some clonal components predate others and originate in specific locations. For example, it is possible that parasite samples from clonal components 1 and 20 in Buenaventura and Tumaco emanated from Guapi,



(a) *P. falciparum* population connectivity: the width of each inter-city edge is proportional to the fraction of highly-related sample pairs across cities plotted in Figure 2b. Note that the edges between Guapi and Quibdó and Guapi and Tadó are plotted but too thin to visually discern.



(b) Screen shot of official marine traffic frequency (routes per 0.4km^2 per year) taken from www.marinetraffic.com 2020-02-12. The road system, which includes the pan-american highway, is visible (faint orange line) on the land map (Google satellite), as are some of the province boundaries (dashed grey lines). Zoom in to see city names Tumaco, Guapi, Buenaventura, Quibdó and Tadó among others.

Figure 4: Comparison between *P. falciparum* population connectivity and the frequency of marine traffic.

creating a spurious link between Buenaventura and Tumaco. However, because these data are from sparsely sampled symptomatic cases in setting where clonal propagation is frequent, sample collection chronology is not necessarily representative of transmission chain events (Figure 6).

Regarding transmission chain events, we note that clonal component 20 relates to the three inter-related clonal components (1, 12 and 14) via an intermediate clonal component detected in Tumaco only (clonal component 15) as well as an intermediate parasite sample from Quibdó that does not belong to a clonal component (Figure A.2). These intermediates likely derive from recombination between parasites related to the clonal components they connect. Several connections consistent with recombinants can be found among the relatedness graphs (Figures 5 and A.2). As such, it seems it may be at least theoretically possible to construct approximate *P. falciparum* transmission chains given more dense sampling of malaria infections on the Colombian-Pacific coast.

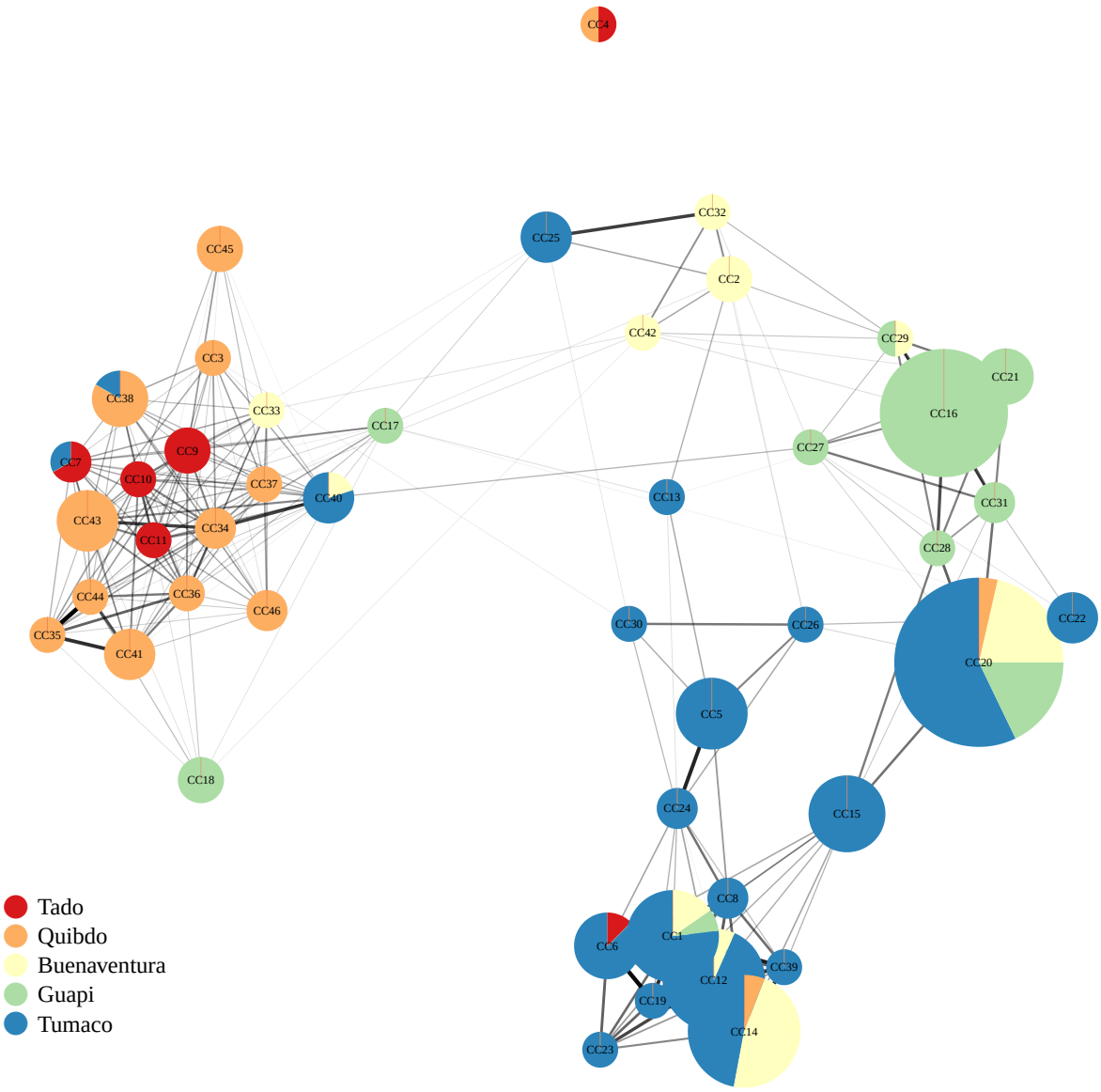


Figure 5: Clonal components (vertices) and the average relatedness between them (edges). Clonal components (CCs) are groups of two or more statistically indistinguishable parasite samples. CC vertices are plotted using the Fruchterman-Reingold layout algorithm [36], thereby clustering inter-related CCs. The size of each CC vertex is proportional to the number of parasite samples per CC, ranging from 2 to 28 statistically indistinguishable parasite samples. CCs are named in order of the collection date of the earliest parasite sample per CC (Table A.4. CCs with parasite samples collected from two or more cities are depicted as pie charts. Colour denotes the city of parasite sample collection. Edge transparency and weight is proportional to average relatedness, ranging from 0.003 to 0.840. Relatedness estimates that are indistinguishable from zero were set to zero. Edges whose average relatedness is zero are not plotted. Each CC besides CC4 is related to at least one other. CC4 contains two samples (one from Tadó, another from Quibdó). It is likely a contaminant; see main text. A plot of CCs that includes singletons (individual parasite samples that do not belong to a CC) can be found in the Appendix (Figure A.2).

Discussion

Here we show that estimates of IBD-based relatedness and their associated uncertainty can be used to uncover epidemiologically meaningful connectivity between *P. falciparum* populations on a relatively local

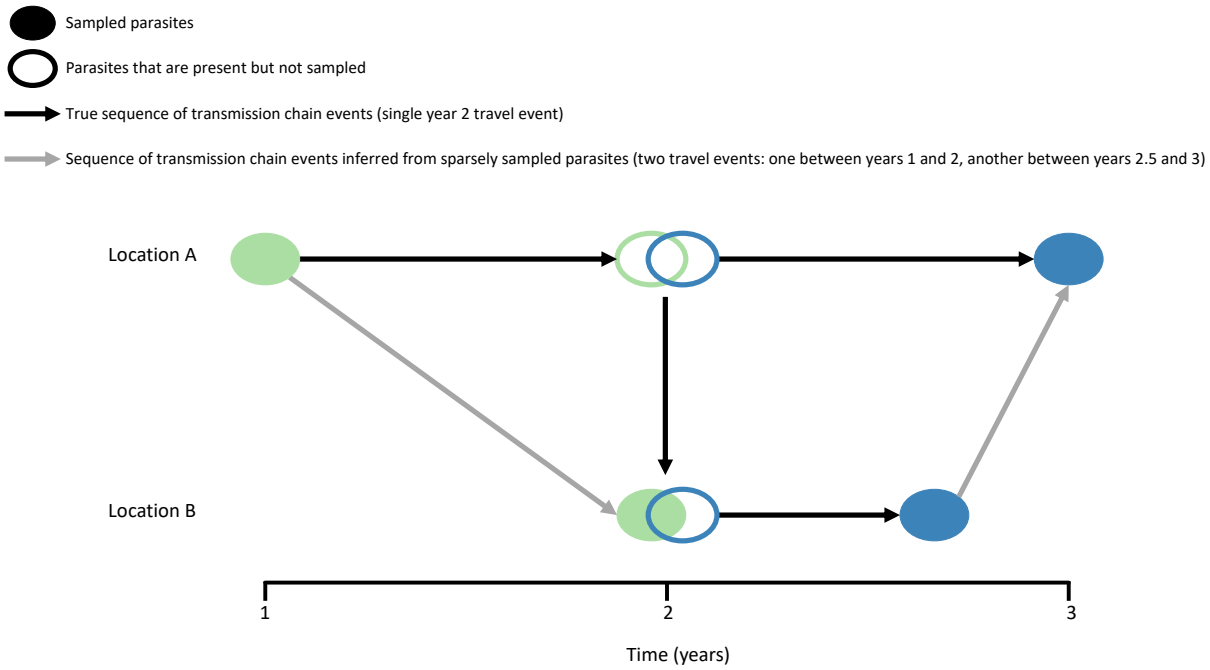


Figure 6: Schematic illustrating why sample collection chronology is not necessarily representative of the true sequence of transmission chain events when sampling is sparse and clonal propagation is frequent. The schematic shows two hypothetical locations A and B where malaria parasites have been sampled sparsely: solid ellipses represent sampled parasites, open ellipses represent parasites that were present but not sampled, different colours denote different parasite genotypes.

spatial scale: along the Colombian-Pacific coast where clonal propagation is frequent. While our approach largely confirms a previous report based on F_{ST} [11], estimates of relatedness provide more granularity while their confidence intervals account for uncertainty thus provide statistical rigor, e.g. when highly-related parasite sample pairs are classified. Our approach includes two additional contributions: concepts from optimal transport (1-Wasserstein distance) are used to compare parasite populations in an entirely threshold-free manner; and clonal components are identified using graph components and confidence intervals, thereby circumventing reliance on an arbitrary clonal threshold. Our overall approach could be adapted for analyses of viruses and bacteria that show recombination or reshuffling of segments as well as clonal propagation [28–31], to other protozoans (e.g. *Toxoplasma*, *Cryprosporidium* [32–34]), and to the many fungal pathogens [35], plants [6, 7], and animals with mixed mating systems.

IBD-based relatedness estimates recovered 1) a large fraction of highly-related parasite sample pairs within Guapi, a city on the Colombian-Pacific coast that is relatively isolated besides infrequent marine traffic; 2) a low fraction of highly-related parasite sample pairs within Buenaventura, the most important port on the Colombian-Pacific coast and thus the least isolated city in this study; and 3) a disproportionally large fraction of highly-related parasite pairs between Buenaventura and Tumaco (departure from isolation-by-distance), where Tumaco is the second largest port on the Colombian-Pacific coast. These observations accord with several published previously: 1) elevated LD in a *P. falciparum* subpopulation (identified using STRUCTURE [27,37]) predominant in Guapi; 2) rapid LD decay in a *P. falciparum* subpopulation predominant in Buenaventura; and 3) lowest genetic differentiation (based on F_{ST} estimates) between provinces Valle (Buenaventura) and Nariño (Tumaco) [11]. LD, STRUCTURE and F_{ST} analyses all rely on allelic variation. The concordance between results based on relatedness and allelic variation suggests that *P. falciparum* outbreeding on the Colombian-Pacific coast is infrequent enough that both types of analyses generate

insight on approximately the same time scale.

The aforementioned results generate hypotheses around the frequency of marine traffic and malaria transmission on the Colombian-Pacific coast. Notwithstanding long-range windborne dispersal, which may be critical for malaria transmission in Africa [38], anopheline flight range is generally small (around 3.5 km [39]). As such, long-range malaria parasite dispersal on the Colombian-Pacific coast is almost certainly human-mediated. A recent study of *P. vivax* proposed that human movement across a "malaria corridor" stretching from the northwest to the south of the Colombian-Pacific Coast likely promotes *P. vivax* gene flow, and that mining activities may provide transmission "contact zones" [40], similarly proposed for *P. falciparum* [20]. *P. falciparum* population connectivity is consistent with the human "malaria corridor" hypothesis, especially since it correlates with accessibility, not isolation-by-distance. Both infected humans and mosquitoes are compatible with this hypothesis, i.e. checks for infected *Anopheles spp.* on boats may be merited [41, 42]. However, relatively high differentiation between populations of *An. albimanus* (one of the three primary vectors of malaria in Colombia [43]) from Buenaventura and Tumaco [44], points towards human carriage.

The Colombian-Pacific coast has long been associated with the international trade of gold and narcotics, but until recently human migration in the region was largely domestic. The flow of international migrants infected with *Plasmodium spp.* has increased significantly in recent years. In 2019, 2190 of 2288 (95.7%) of non-domestic malaria cases reported in Colombia were from Venezuela; other sources included South America (Peru, Panama, French Guyana, Ecuador, Brazil) and some African countries (Uganda, Republic of the Congo, Nigeria, Ivory Coast, Cameroon, Angola) [45]. Some of the infected Venezuelan nationals are migrating southward to Ecuador and Peru [22]. Other non-domestic cases may be associated with the traffic of people who arrive at Colombian ports with a view towards northward travel e.g. to the USA via Central America and Panama [46]. Genetic surveillance of "international parasites" may help malaria control efforts in Colombia.

Considerable violence in the South Pacific region of Colombia between 1993 and 2007 combined with historically high malaria case counts [15] could have caused fleeting connectivity between *P. falciparum* populations. Based on relatedness between clonal components, we refute the hypothesis that connectivity between Buenaventura and Tumaco was due to a single individual with a multiclonal infection. We cannot reject a single travel event involving multiple individuals with independent infections, however. Contemporary data on more densely sampled cases and on mosquito and human movement are required to characterise extant connectivity, its reach beyond Colombia (see e.g. [47]), and to rule out alternative hypotheses.

Regarding alternative hypotheses, heterogeneous vectorial capacity and antimalarial drug pressure could selectively enhance parasite survival in such a way that generates apparent connectivity between Buenaventura and Tumaco, e.g. if parasites are adapted to local vectors whose distributions are more similar between Buenaventura and Tumaco than elsewhere. Although adult *An. albimanus* B and *An. neivai* `s.l.` have been detected in the vicinities of both cities [44, 48], the species distributions in the vicinities of Buenaventura and Tumaco differ more than those in the vicinities of Tumaco and Guapi [48]. As such, heterogeneous vectorial capacity seems an unlikely alternative hypothesis. Similarly, relatedness may be greater among parasites with comparable antimalarial resistance: a recent study of South East Asian *P. falciparum* parasites found greater relatedness in the recent past among parasites with artemisinin resistance mutations versus without [49]. This study used size-stratified IBD segments to date relatedness [49].[1] On the Colombian-Pacific coast, IBD segment size inference could help identify some recently related parasites. However, it requires whole genome sequence data and is hard (if not presently impossible) to interpret in the face of frequent selfing that is transmission dependent [25]. The development of an ancestral recombination model that incorporates transmission-dependent selfing is a research priority in malaria genetic epidemiology and would aid research on other organisms that show both outbreeding and clonal propagation.

[1] In malaria, IBD segments (genome segments that are descended from a common ancestor unbroken by recombination [4]) are broken down each time genetically distinct parasites outcross. As such, IBD segment size is distributed according to a number of out-crossed generations, which increase over time, albeit in a complex transmission-dependent manner, which is not yet fully understood [9, 10].

Methods

Data

This study relies entirely on previously published data that are publicly available [11, 25]. In the original study by Echeverry et al., finger-prick blood spot samples were obtained from patients with symptomatic uncomplicated malaria [11]. Samples were collected between 1993 and 2007 from five cities in four provinces: Tadó and Quibdó in Chocó, Buenaventura in Valle, Guapi in Cauca and Tumaco in Nariño (Table A.2) [11]. Informed consent was obtained from all the subjects enrolled, as approved by CIDEIM Institutional Review Board (IRB) [11]. The Colombian-pacific coast is one of the rainiest regions of the world [44, 50]. At that time, Colombia had approximately 100,000 malaria cases per year [11, 15]. Collectively Chocó, Valle, Cauca and Nariño accounted for up to 75% of the *P. falciparum* cases reported, with relatively high transmission in Chocó and relatively low transmission in Valle and Cauca [11].

The data that feature in this descriptive study also feature in a recent methodological study concerning data requirements for relatedness inference [25]. As in [25], we did not post-process the data in any way besides mapping SNP positions to the *P. falciparum* 3d7 v3 reference genome and recoding heteroallelic calls as missing (since all samples with fewer than 10 heteroallelic SNP calls were classified monoclonal previously [11]). The monoclonal data include 325 *P. falciparum* samples with data on 250 biallelic SNPs whose minor allele frequency estimates (the minor allele sample count divided by 325) range from 0.006 to 0.495 (Figure A.3).

Relatedness inference and classification of parasite sample pairs and groups

For each pairwise parasite sample comparison, we generated a relatedness estimate and 95% confidence interval using the hidden Markov model and parametric bootstrap described in [25]. Sample pairs were classified as unrelated, related, highly-related and clonal using confidence interval bounds as follows and summarised in Table 1. A pair was classified unrelated if its relatedness estimate, $\hat{r}$, was statistically indistinguishable from zero with lower confidence interval bound (LCI) less than ϵ. A pair was classified related if its relatedness estimate, $\hat{r}$, was statistically distinguishable from zero with LCI $> \epsilon$. A pair was considered highly-related if its relatedness estimate, $\hat{r}$, was statistically distinguishable from some specified threshold, τ, with LCI $> \tau$. A pair was considered clonal if its relatedness estimate, $\hat{r}$, was statistically indistinguishable from one with upper confidence interval bound (UCI) $> 1-\epsilon$. Note that these classifications are possible because all estimates are informative, i.e. no confidence intervals span the entire zero to one range (Figure 1. These classifications are neither necessarily exclusive nor conversely true: a clonal parasite pair is related, but a related parasite pair is not necessarily clonal. Throughout, $\epsilon = 0.01$. In the main analysis (Figure 2) $\tau = 0.25$, in the sensitivity analysis (Figure A.1a) $\tau \in \{0.25, 0.50\}$.

In addition to classifying parasite sample pairs, we classify groups of statistically indistinguishable parasite samples, which we call clonal components because they are defined using the simple concept of components from graph theory. First, we construct a super-graph whose vertices are parasite samples connected by edges that are weighted by relatedness estimates. Within the super-graph, a clonal component is a sub-graph within which all parasite samples are connected to one another (directly or not) via edges whose weights are statistically indistinguishable from one, while being connected to parasites samples outside the sub-graph via edges whose weights are not statistically indistinguishable from one. Clonal components tend to be fully connected (i.e. all parasite samples within the clonal component are directly connected to one another by edges whose weights are statistically indistinguishable from one). The `igraph` package [51] in `R` [52] was used to identify clonal components and to visualise them using the Fruchterman-Reingold layout algorithm [36].

Spatiotemporal trends in *P. falciparum* population connectivity

Spatiotemporal trends in population connectivity were explored visually by partitioning parasite sample pairs by their collection cities and dates, then plotting the per-partition fraction of highly-related pairs. Error bars were constructed by re-sampling per-partition parasite sample pairs 100 times with replacement and taking the 2.5th and 97.5th percentiles of the fraction of highly-related pairs as the lower and upper limits, respectively. Sensitivity to $\tau = 0.25$ (high relatedness threshold used in Figure 2) was explored using alternative $\tau \in \{0.25, 0.50\}$ (Figure A.1) and also by using a threshold-free approach (Figure 3) as follows.

To explore population connectivity using a threshold-free approach, we calculated 1-Wasserstein distances between groups of parasite samples from different cities using the `transport` [53] package in `R` [52]. Specifically, for a pair of cities a and b, we construct a $n_a \times n_b$ genetic distance matrix, G, of $1 - \hat{r}_{ij}$ (where n_a and n_b are the parasite sample counts from cities a and b, respectively, $i = 1, \ldots, n_a$ and $j = 1, \ldots, n_b$) and two vectors $w_a = (1/n_a, \ldots, 1/n_a)$ and $w_b = (1/n_b, \ldots, 1/n_b)$ of length n_a and n_b, respectively. We then calculate the 1-Wasserstein distance, which minimises the total cost of transporting w_a to w_b, where $1 - \hat{r}_{ij}$ is the cost of transporting a single unit, using `transport::transport(`w_a, w_b, `costm =` G, `method = "shortsimplex")`. This amounts to treating parasite samples from different cities as draws from different distributions, where the 1-Wasserstein distance can be interpreted as the effort required to transform a distribution of parasite samples from one city into a distribution from another [16]. City pairs with smaller 1-Wasserstein distances are interpreted as having greater connectivity between the *P. falciparum* populations collected from them. Error bars were constructed by re-sampling parasite sample pairs per inter-city partition 100 times with replacement and taking the 2.5th and 97.5th percentiles of the distribution 1-Wasserstein distances based on the re-sampled sample pairs as the lower and upper limits, respectively.

Data and code availability

All data analyses were performed in `R` [52]. The data are publicly available as a `.RData` files and the code is publicly available as `.R` scripts at `https://github.com/artaylor85/ColombianBarcode`.

Funding

A.R.T. and C.O.B. are supported by a Maximizing Investigators' Research Award for Early Stage Investigators (R35 GM-124715) (https://www.nih.gov/). D.F.E. received financial support from Colciencias, call 656-2014 "Es Tiempo de Volver" award FP44842-503-2014 (https://minciencias.gov.co/). T.J.C.A. is supported by funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health (R37 AI048071) (https://www.niaid.nih.gov/). This project was funded in part with federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under grant number U19 AI-110818 to the Broad Institute (D.E.N.) (https://www.niaid.nih.gov/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments

Thank you to Pierre Jacob for guidance on the calculation of the 1-Wasserstein distances and to James Watson, Manuela Carrasquilla and Vladimir Corredor for helpful comments and discussion.

Competing interests statement

The authors have declared that no competing interests exist.

References

[1] Dalmat R, Naughton B, Kwan-Gett TS, Slyker J, Stuckey EM. Use cases for genetic epidemiology in malaria elimination. Malaria journal. 2019;18(1):163.

[2] Holder M, Lewis PO. Phylogeny estimation: traditional and Bayesian approaches. Nature reviews genetics. 2003;4(4):275–284.

[3] Biek R, Pybus OG, Lloyd-Smith JO, Didelot X. Measurably evolving pathogens in the genomic era. Trends in ecology & evolution. 2015;30(6):306–313.

[4] Thompson EA. Identity by descent: variation in meiosis, across genomes, and in populations. Genetics. 2013;194(2):301–326.

[5] Taylor AR, Schaffner SF, Cerqueira GC, Nkhoma SC, Anderson TJ, Sriprawat K, et al. Quantifying connectivity between local Plasmodium falciparum malaria parasite populations using identity by descent. PLoS genetics. 2017;13(10):e1007065.

[6] Grant AG, Kalisz S. Do selfing species have greater niche breadth? Support from ecological niche modeling. Evolution. 2019;.

[7] Mattila TM, Laenen B, Slotte T. Population genomics of transitions to selfing in Brassicaceae model systems. In: Statistical Population Genomics. Springer; 2020. p. 269–287.

[8] Siegel SV, Rayner JC. Single cell sequencing shines a light on malaria parasite relatedness in complex infections. Trends in Parasitology. 2020;36(2):83–85.

[9] Nkhoma SC, Nair S, Cheeseman IH, Rohr-Allegrini C, Singlam S, Nosten F, et al. Close kinship within multiple-genotype malaria parasite infections. Proceedings of the Royal Society B: Biological Sciences. 2012;279(1738):2589–2598.

[10] Nkhoma SC, Trevino SG, Gorena KM, Nair S, Khoswe S, Jett C, et al. Co-transmission of Related Malaria Parasite Lineages Shapes Within-Host Parasite Diversity. Cell Host & Microbe. 2020;27(1):93–103.

[11] Echeverry DF, Nair S, Osorio L, Menon S, Murillo C, Anderson TJC. Long term persistence of clonal malaria parasite Plasmodium falciparum lineages in the Colombian Pacific region. BMC Genetics. 2013;14(2).

[12] Omedo I, Mogeni P, Rockett K, Kamau A, Hubbart C, Jeffreys A, et al. Geographic-genetic analysis of Plasmodium falciparum parasite populations from surveys of primary school children in Western Kenya. Wellcome open research. 2017;2.

[13] Omedo I, Mogeni P, Bousema T, Rockett K, Amambua-Ngwa A, Oyier I, et al. Micro-epidemiological structuring of Plasmodium falciparum parasite populations in regions with varying transmission intensities in Africa. Wellcome open research. 2017;2.

[14] Tessema S, Wesolowski A, Chen A, Murphy M, Wilheim J, Mupiri AR, et al. Using parasite genetic and human mobility data to infer local and cross-border malaria connectivity in Southern Africa. Elife. 2019;8:e43510.

[15] Rodríguez JCP, Uribe GÁ, Araújo RM, Narváez PC, Valencia SH. Epidemiology and control of malaria in Colombia. Memórias do Instituto Oswaldo Cruz. 2011;106:114–122.

[16] Feged-Rivadeneira A, Ángel A, González-Casabianca F, Rivera C. Malaria intensity in Colombia by regions and populations. PLoS ONE. 2018;13(9):e0203673. doi:10.6084/m9.figshare.6863780.

[17] Castellanos A, Chaparro-Narváez P, Morales-Plaza CD, Alzate A, Padilla J, Arévalo M, et al. Malaria in gold-mining areas in Colombia. Memorias do Instituto Oswaldo Cruz. 2016;111(1):59–66. doi:10.1590/0074-02760150382.

[18] 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(273):1–18. doi:10.1186/s12936-017-1925-6.

[19] Daniels JP. Increasing malaria in Venezuela threatens regional progress. The Lancet Infectious diseases. 2018;18(3):257. doi:10.1016/S1473-3099(18)30086-0.

[20] Knudson A, González-Casabianca F, Feged-Rivadeneira A, Pedreros MF, Aponte S, Olaya A, et al. Spatio-temporal dynamics of Plasmodium falciparum transmission within a spatial unit on the Colombian Pacific Coast. Scientific Reports. 2020;10(1):1–16.

[21] Grillet ME, Leopoldo V, Oletta JF, Tami A, Conn JE. Malaria in Venezuela requires response. Science. 2018;359(6375):528.

[22] Jaramillo-ochoa R, Sippy R, Farrell DF, Cueva-aponte C, Beltrán-ayala E, Gonzaga JL, et al. Effects of Political Instability in Venezuela on Malaria Resurgence at Ecuador–Peru Border, 2018. Emerging Infectious Diseases. 2019;25(4):834–836.

[23] Daniels JP. Venezuela in crisis. The Lancet Infectious Diseases. 2019;19(1):28. doi:10.1016/s1473-3099(18)30745-x.

[24] Rodríguez-Morales AJ, Suárez JA, Risquez A, Villamil-Gómez WE, Paniz-Mondolfi A. Consequences of Venezuela's massive migration crisis on imported malaria in Colombia, 2016–2018. Travel Medicine and Infectious Disease. 2019;28(February):98–99. doi:10.1016/j.tmaid.2019.02.004.

[25] Taylor AR, Jacob PE, Neafsey DE, Buckee CO. Estimating relatedness between malaria parasites. Genetics. 2019; p. genetics–302120.

[26] Neafsey DE, Schaffner SF, Volkman SK, Park D, Montgomery P, Milner DA, et al. Genome-wide SNP genotyping highlights the role of natural selection in Plasmodium falciparum population divergence. Genome biology. 2008;9(12):R171.

[27] Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155(2):945–959.

[28] Wille M, Holmes EC. The Ecology and Evolution of Influenza Viruses. Cold Spring Harbor Perspectives in Medicine. 2019; p. a038489.

[29] Katz EM, Esona MD, Betrapally NS, Lucia A, Neira YR, Rey GJ, et al. Whole-gene analysis of inter-genogroup reassortant rotaviruses from the Dominican Republic: Emergence of equine-like G3 strains and evidence of their reassortment with locally-circulating strains. Virology. 2019;534:114–131.

[30] Caugant DA, Brynildsrud OB. Neisseria meningitidis: using genomics to understand diversity, evolution and pathogenesis. Nature Reviews Microbiology. 2019; p. 1–13.

[31] Smith JM, Feil EJ, Smith NH. Population structure and evolutionary dynamics of pathogenic bacteria. Bioessays. 2000;22(12):1115–1122.

[32] Tibayrenc M, Ayala FJ. The clonal theory of parasitic protozoa: 12 years on. Trends in parasitology. 2002;18(9):405–410.

[33] Rajendran C, Su C, Dubey JP. Molecular genotyping of Toxoplasma gondii from Central and South America revealed high diversity within and between populations. Infection, Genetics and Evolution. 2012;12(2):359–368.

[34] Nader JL, Mathers TC, Ward BJ, Pachebat JA, Swain MT, Robinson G, et al. Evolutionary genomics of anthroponosis in Cryptosporidium. Nature microbiology. 2019;4(5):826–836.

[35] Nieuwenhuis BP, James TY. The frequency of sex in fungi. Philosophical Transactions of the Royal Society B: Biological Sciences. 2016;371(1706):20150540.

[36] Fruchterman TM, Reingold EM. Graph drawing by force-directed placement. Software: Practice and experience. 1991;21(11):1129–1164.

[37] Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics. 2003;164(4):1567–1587.

[38] Huestis DL, Dao A, Diallo M, Sanogo ZL, Samake D, Yaro AS, et al. Windborne long-distance migration of malaria mosquitoes in the Sahel. Nature. 2019;574(7778):404–408.

[39] Verdonschot PF, Besse-Lototskaya AA. Flight distance of mosquitoes (Culicidae): a metadata analysis to support the management of barrier zones around rewetted and newly constructed wetlands. Limnologica-Ecology and Management of Inland Waters. 2014;45:69–79.

[40] Pacheco MA, Schneider KA, Céspedes N, Herrera S, Arévalo-Herrera M, Escalante AA. Limited differentiation among Plasmodium vivax populations from the northwest and to the south Pacific Coast of Colombia: A malaria corridor? PLoS Neglected Tropical Diseases. 2019;13(3):e0007310. doi:10.1371/journal.pntd.0007310.

[41] Guagliardo SA, Morrison AC, Barboza JL, Requena E, Astete H, Vazquez-Prokopec G, et al. River boats contribute to the regional spread of the dengue vector Aedes aegypti in the Peruvian Amazon. PLoS neglected tropical diseases. 2015;9(4):e0003648.

[42] Lounibos LP. Invasions by insect vectors of human disease. Annual review of entomology. 2002;47(1):233–266.

[43] Montoya-Lerma J, Solarte YA, Giraldo-Calderón GI, Quiñones ML, Ruiz-López F, Wilkerson RC, et al. Malaria vector species in Colombia: a review. Memórias do Instituto Oswaldo Cruz. 2011;106:223–238.

[44] Gutiérrez LA, Naranjo NJ, Cienfuegos AV, Muskus CE, Luckhart S, Conn JE, et al. Population structure analyses and demographic history of the malaria vector Anopheles albimanus from the Caribbean and the Pacific regions of Colombia. Malaria journal. 2009;8(1):259.

[45] Instituto Nacional de Salud Colombia, Dirección de Vigilancia y Analisis del Riesgo en Salud Pública. Boletín Epidemiológico Semanal: semana epidemiológica 52. 2019;doi:10.33610/23576189.2019.52.

[46] Wabgou M, Vargas D, Carabalí JA. Las migraciones internacionales en Colombia. Investigación & Desarrollo. 2012;20(1):142–167.

[47] Vera-Arias CA, Castro LE, Gómez-Obando J, Sáenz FE. Diverse origin of Plasmodium falciparum in northwest Ecuador. Malaria journal. 2019;18(1):251.

[48] Ahumada ML, Orjuela LI, Pareja PX, Conde M, Cabarcas DM, Cubillos EFG, et al. Spatial distributions of Anopheles species in relation to malaria incidence at 70 localities in the highly endemic Northwest and South Pacific coast regions of Colombia. Malaria Journal. 2016;15(407):1–16. doi:10.1186/s12936-016-1421-4.

[49] Shetty AC, Jacob CG, Huang F, Li Y, Agrawal S, Saunders DL, et al. Genomic structure and diversity of Plasmodium falciparum in Southeast Asia reveal recent parasite migration patterns. Nature communications. 2019;10(1):2665.

[50] Naranjo-Díaz N, Altamiranda M, Luckhart S, Conn JE, Correa MM. Malaria vectors in ecologically heterogeneous localities of the Colombian Pacific region. PLoS ONE. 2014;9(8):e103769. doi:10.1371/journal.pone.0103769.

[51] Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal. 2006;Complex Systems:1695.

[52] R Core Team. R: A Language and Environment for Statistical Computing; 2018. Available from: https://www.R-project.org/.

[53] Schuhmacher D, Bähre B, Gottschlich C, Hartmann V, Heinemann F, Schmitzer B. transport: Computation of Optimal Transport Plans and Wasserstein Distances; 2019. Available from: https://cran.r-project.org/package=transport.

[54] Diaz G, Lasso AM, Murillo C, Montenegro LM, Echeverry DF. Evidence of self-medication with chloroquine before consultation for malaria in the southern pacific coast region of Colombia. American Journal of Tropical Medicine and Hygiene. 2019;100(1):66–71. doi:10.4269/ajtmh.18-0515.

[55] Valencia SH, Ocampo ID, Arce-Plata MI, Recht J, Arévalo-Herrera M. Glucose-6-phosphate dehydrogenase deficiency prevalence and genetic variants in malaria endemic areas of Colombia. Malaria Journal. 2016;15(291):1–9. doi:10.1186/s12936-016-1343-1.

[56] Vallejo AF, Chaparro PE, Benavides Y, Álvarez Á, Quintero JP, Padilla J, et al. High prevalence of sub-microscopic infections in Colombia. Malaria Journal. 2015;14(201):1–7. doi:10.1186/s12936-015-0711-6.

[57] Valero-Bernal MV, Tanner M, Muñoz-Navarro S, Valero-Bernal JF. Proportion of fever attributable to malaria in Colombia: Potential indicators for monitoring progress towards malaria elimination. Revista de Salud Pública. 2017;19:45–51.

[58] Pava Z, Echeverry DF, Díaz G, Murillo C. Large variation in detection of histidine-rich protein 2 in Plasmodium falciparum isolates from Colombia. The American journal of tropical medicine and hygiene. 2010;83(4):834–837.

[59] Solano CM, Okoth SA, 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.

[60] 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.

[61] Padilla JC, Chaparro PE, Molina K, Arevalo-Herrera M, Herrera S. Is there malaria transmission in urban settings in Colombia? Malaria Journal. 2015;14(453):1–9. doi:10.1186/s12936-015-0956-0.

Appendix A

Climate e.g. very heavy rainfall and flooding on the Pacific Coast [44, 50]
Rich vector diversity [48, 50]†
Parasite resistance to antimalarial drugs [18, 20]
Persistent drug pressure due to frequent self-medication [54]
Prevalence of glucose-6-phosphate dehydrogenase deficiency [18, 55]†
Prevalence of duffy-negative individuals resistant to *P. vivax* invasion [18, 56]†
Sub-microscopic and asymptomatic infections [18, 20, 56, 57]† and infections that can evade detection by some rapid diagnostic tests [18, 20, 58–60]
Per-urban and urban transmission [18, 61]
Human migration due to deforestation and gold mining [16–19]

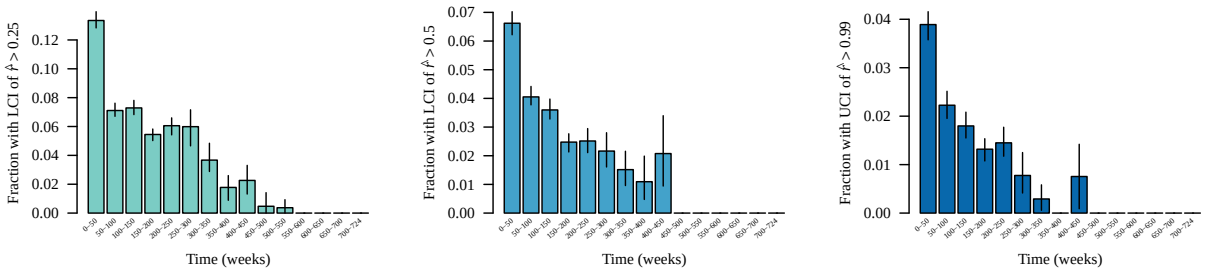
Table A.1: Factors associated with malaria epidemiology in Colombia: a non exhaustive list. †One or more citations are specific to Tierralta and the (South) Pacific coast.

City (Province)	1993	1994	1997	1999	2000	2001	2002	2003	2004	2005	2006	2007	Total
Tumaco (Nariño)	0	0	0	2	2	10	11	59	0	23	0	25	132
Guapi (Cauca)	0	0	0	1	1	0	0	66	0	0	0	0	68
Buenaventura (Valle)	4	1	0	5	0	0	0	0	12	15	10	0	47
Quibdó (Chocó)	0	0	2	0	6	1	0	0	14	6	13	22	64
Tadó (Chocó)	0	0	0	0	0	12	2	0	0	0	0	0	14
Total	4	1	2	8	9	23	13	125	26	44	23	47	325

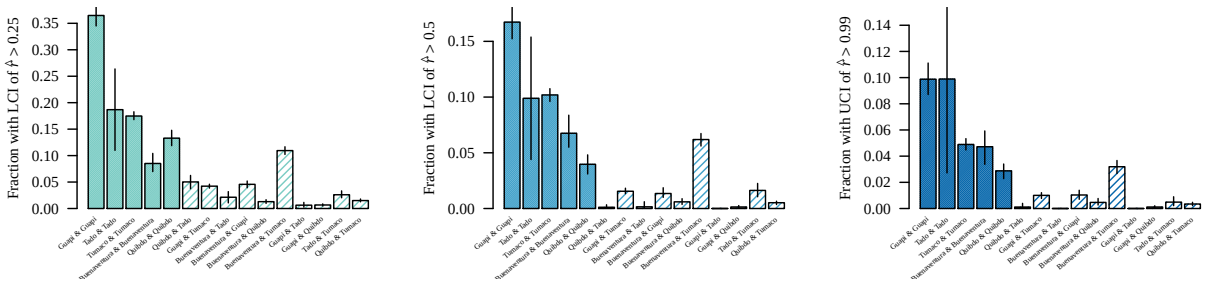
Table A.2: Yearly monoclonal *P. falciparum* sample counts per city.

	CC1	CC12	CC14	CC20
CC12	0.712 (0.635)			
CC14	0.733 (0.709)	0.648 (0.517)		
CC20	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	
CC40	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.079 (0.000)

Table A.3: Average relatedness to three decimal places between clonal components (CCs) 1, 12, 14, 20 and 40 with maximum 2.5% confidence interval bound in parentheses. The maximum 2.5% confidence interval bound indicates that relatedness between C20 and C40 is not statistically distinguishable from zero, for example.



(a) Partitioned by time between collection dates.



(b) Partitioned by collection city.

Figure A.1: Fractions of highly-related sample pairs partitioned in time and space: sensitivity to high-relatedness thresholds. highly-related samples pairs are defined as those with lower confidence interval bound (LCI) of relatedness estimate, $\hat{r}$, greater than thresholds 0.25 and 0.50; or with upper confidence interval bound (UCI) of $\hat{r} > 0.99$ (i.e. clonal parasite sample pairs; Table 1). Colours correspond to Figure 1.

Clonal component	Date	City
CC1	1999-03-15	Guapi
CC2	1999-04-13	Buenaventura
CC3	2000-04-13	Quibdo
CC4	2000-06-29	Quibdo
CC5	2000-11-23	Tumaco
CC6	2001-01-29	Tumaco
CC7	2001-02-08	Tumaco
CC8	2001-02-17	Tumaco
CC9	2001-06-07	Tado
CC10	2001-06-08	Tado
CC11	2001-12-03	Tado
CC12	2002-04-03	Tumaco
CC13	2002-04-03	Tumaco
CC14	2002-04-04	Tumaco
CC15	2002-04-05	Tumaco
CC16	2003-01-07	Guapi
CC17	2003-03-03	Guapi
CC18	2003-03-07	Guapi
CC19	2003-03-17	Tumaco
CC20	2003-03-22	Guapi
CC21	2003-04-11	Guapi
CC22	2003-05-15	Tumaco
CC23	2003-05-20	Tumaco
CC24	2003-05-29	Tumaco
CC25	2003-09-08	Tumaco
CC26	2003-10-01	Tumaco
CC27	2003-10-03	Guapi
CC28	2003-10-06	Guapi
CC29	2003-10-07	Guapi
CC30	2003-10-21	Tumaco
CC31	2003-11-05	Guapi
CC32	2004-05-31	Buenaventura
CC33	2004-07-23	Buenaventura
CC34	2004-08-24	Quibdo
CC35	2004-10-27	Quibdo
CC36	2004-11-10	Quibdo
CC37	2004-12-10	Quibdo
CC38	2005-02-28	Quibdo
CC39	2005-07-27	Tumaco
CC40	2005-07-28	Tumaco
CC41	2006-08-03	Quibdo
CC42	2006-08-22	Buenaventura
CC43	2006-09-05	Quibdo
CC44	2006-10-10	Quibdo
CC45	2007-03-22	Quibdo
CC46	2007-03-28	Quibdo

Table A.4: Date and city of collection of earliest parasite sample per clonal component.

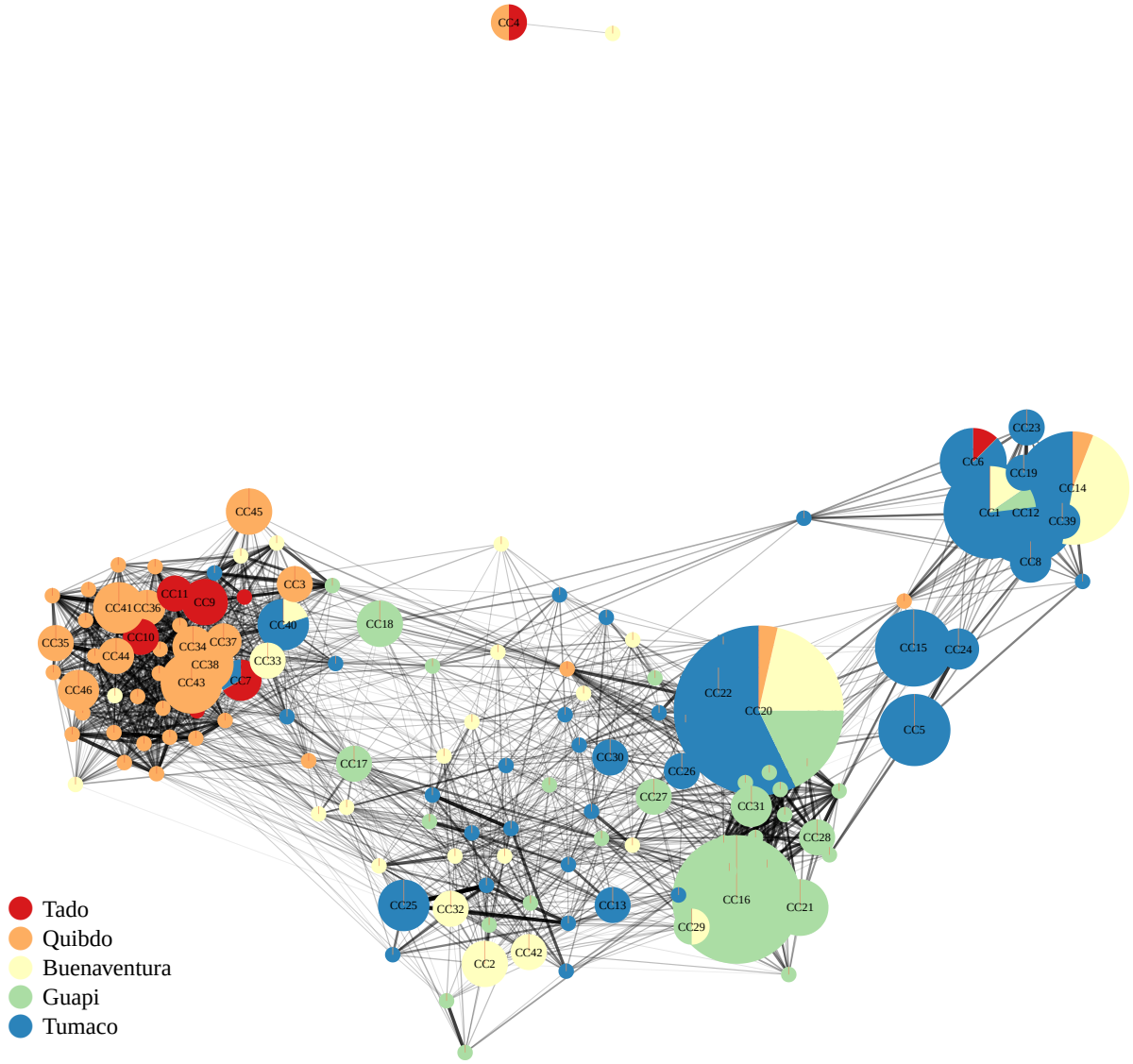


Figure A.2: Clonal components and singletons (vertices) and the average relatedness between them (edges). Clonal components (CCs) are groups of two or more statistically indistinguishable parasite samples. Singletons are individual parasite samples that do not belong to a CC. Vertices are plotted using the Fruchterman-Reingold layout algorithm [36], thereby clustering inter-related vertices. The size of each CC vertex is proportional to the number of parasite samples per CC, ranging from 2 to 28 statistically indistinguishable parasite samples. CCs are named in order of the collection date of the earliest parasite sample per CC (Table A.4). CCs with parasite samples from two or more sites are depicted as pie charts. Colour denotes the city of parasite sample collection. Edge transparency and weight is proportional to average relatedness, ranging from 0.003 to 0.912. Relatedness estimates that are indistinguishable from zero were set to zero. Edges whose average relatedness is zero are not plotted. Each CC besides CC4 is related to at least one other. CC4 is likely a contaminant; see main text. A singleton from Buenaventura, which is loosely related to CC4, may also be a contaminant.

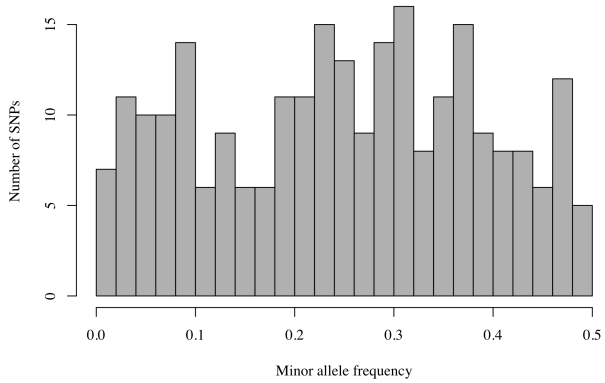


Figure A.3: Histogram of minor allele frequencies estimated using all 325 monoclonal *P. falciparum* samples genotyped at 250 biallelic SNPs.