Malaria surveillance reveals parasite relatedness, signatures of selection, and correlates of transmission across Senegal

Stephen F. Schaffner[1,4,8], Aida Badiane[2], Akanksha Khorgade[1], Medoune Ndiop[3], Jules Gomis[2], Wesley Wong[4], Yaye Die Ndiaye[2], Younouss Diedhiou[2], Julie Thwing[5], Mame Cheikh Seck[2], Angela Early[1], Mouhamad Sy[2], Awa Deme[2], Mamadou Alpha Diallo[2], Ngayo Sy[6], Aita Sene[2], Tolla Ndiaye[2], Djiby Sow[2], Baba Dieye[2], Ibrahima Mbaye Ndiaye[2], Amy Gaye[2], Aliou Ndiaye[2], Katherine E. Battle[7], Joshua L. Proctor[7], Caitlin Bever[7], Fatou Ba Fall[3], Ibrahima Diallo[3], Seynabou Gaye[3], Doudou Sene[3], Daniel L. Hartl[8], Dyann F. Wirth[1,2], Bronwyn MacInnis[1], Daouda Ndiaye[2], Sarah K. Volkman*[1,4,9]

*Correspondence: svolkman@hsph.harvard.edu

1 Infectious Disease and Microbiome Program, The Broad Institute, Cambridge, MA, USA
2 Centre International de recherche, de formation en Genomique Appliquee et de Surveillance Sanitaire (CIGASS), Dakar, Senegal
3 Programme National de Lutte contre le Paludisme (PNLP), Dakar, Senegal
4 Department of Immunology and Infectious Diseases, Harvard T. H. Chan School of Public Health, Boston, MA, USA
5 Centers for Disease Control and Prevention, Atlanta GA USA
6 Section de Lutte Anti-Parasitaire (SLAP) Clinic, Thies, Senegal
7 Institute for Disease Modeling in Global Health, Bill and Melinda Gates Foundation, Seattle. WA USA
8 Department of Organismic and Evolutionary Biology, Harvard University, Cambridge MA USA
9 College of Natural, Behavioral, and Health Sciences, Simmons University, Boston MA USA

Abstract

Parasite genetic surveillance has the potential to play an important role in malaria control. We describe here an analysis of data from the first year of an ongoing, nationwide program of genetic surveillance of *Plasmodium falciparum* parasites in Senegal, intended to provide actionable information for malaria control efforts. Looking for a good proxy for local malaria incidence, we found that the best predictor was the proportion of polygenomic infections (those with multiple genetically distinct parasites), although that relationship broke down at very low incidence. The proportion of closely related parasites in a site was more weakly correlated with incidence while the local genetic diversity was uninformative. Study of related parasites indicated their potential for discriminating local transmission patterns: two nearby study areas had similarly high fractions of relatives, but one area was dominated by clones and the other by outcrossed relatives. Throughout the country, most related parasites proved to belong to a single network of relatives, within which parasites were enriched for shared haplotypes at known and suspected drug resistance loci as well as at one novel locus, reflective of ongoing selection pressure.

Introduction

Despite progress over the past several decades toward malaria control and elimination, *Plasmodium falciparum* malaria remains a major global cause of human morbidity and mortality.

As countries seek to improve the targeting and effectiveness of malaria interventions, they require detailed information about the prevalence of drug resistance, about local epidemiology, and about how malaria burden is changing in response to interventions and other factors. Parasite genetic surveillance shows great promise as a source of information to support decision-making in all of these areas [2-5]. It already has a well-established role in tracking the spread of known drug resistance markers and detecting the appearance of new resistance alleles [6, 7]. In low transmission settings, genetic data can determine whether new cases arise from ongoing local transmission or from importation, and it can potentially identify sources of imported parasites [8-10].

Detecting changes in malaria burden is challenging, both because directly surveying parasite prevalence is difficult and because indirect estimates can be skewed by varying care-seeking behavior and access to health resources. Here again genetic surveys have the potential to provide important information, since parasite genetics has already been observed to reflect changes in transmission rates [11]. However, the relationship between malaria burden and genetics is complex, in part because the parasite life cycle includes both a haploid asexual stage in humans and a diploid sexual stage in the mosquito. Both self-fertilization and outcrossing can occur during the sexual cycle, with the latter requiring the presence of multiple genetically distinct parasites in the mosquito, something that occurs more frequently when transmission is higher. Thus, as malaria transmission drops, we can expect a decrease in the typical complexity of infection (COI, the number of distinct parasite genomes present in an individual) and an increasingly clonal parasite population, both of which have been observed [12]. The exact relationships between these aspects of parasite population genetics and transmission, and whether they are able to provide useful information for malaria control, are still being investigated. Even less is known about the utility of partially related parasites (close relatives that are not clones) in understanding transmission; these should become less common with declining transmission (and hence declining outcrossing) but also easier to detect in a smaller parasite population.

Here we describe an analysis of the relationship between parasite transmission dynamics and parasite genetics, based on the first year of *P. falciparum* genetic surveillance across Senegal. The project is intended as a testbed for learning how to apply parasite genetic surveys to inform malaria control strategy across the range of malaria transmission intensity seen in Senegal, which is stratified into three zones [Fig. 1a], with very low transmission in the north of the country (annual incidence < 5/1000/year), more moderate transmission in the middle (5 – 25/1000), and higher transmission in the southeast (> 25/1000). Much of Senegal is approaching pre-elimination status, but persistent high transmission in the south (> 500/1000 in some catchment zones) together with importation of infections into low transmission zones threaten these gains.

The dataset consists of genetic data from 1097 *P. falciparum* samples collected in 2019 from 23 health facilities in Senegal, primarily as part of the National Malaria Control Program's (NMCP's) sentinel surveillance network. All samples were subjected to barcode genotyping based on 24 single nucleotide polymorphisms (SNPs), which was used to identify clonal parasites and

polygenomic infections (those with COI > 1). A subset of primarily monogenomic samples was chosen for more detailed analysis using whole genome sequencing (WGS) and subsequent genome-wide SNP genotyping, which gave us the ability to detect partially related parasites and identify signatures of selection. Our goal was broadly to explore ways in which this genetic data could be informative for malaria control, and in particular to investigate the relationship between parasite genetics and malaria burden.

Results

We collected samples for genetic analysis during the malaria transmission season in Senegal (July to December 2019) from febrile individuals with uncomplicated malaria infections at 23 health facility sites across the country (Fig. 1, Table 1), selected to provide a broad representation of transmission intensity. For each site we also obtained data on local malaria incidence as determined by the NMCP's routine surveillance program [13]. The final dataset used for analysis comprised 1,097 samples: 310 with both barcode and genome-wide SNP data, 757 with barcode data only, and 30 with genome-wide SNP data only (their barcode data having failed our quality threshold for analysis). Samples chosen for genome-wide analysis were subjected to *P. falciparum* selective whole genome amplification and Illumina-based short-read whole genome sequencing. Monogenomic samples were analyzed at ~40,000 polymorphic genome-wide SNPs filtered to remove highly heterozygous regions and SNPs (see Materials and Methods and Supplemental Fig. 1 for details on filtering); when multiple clones were sequenced from a study site, only one was included in the analysis.

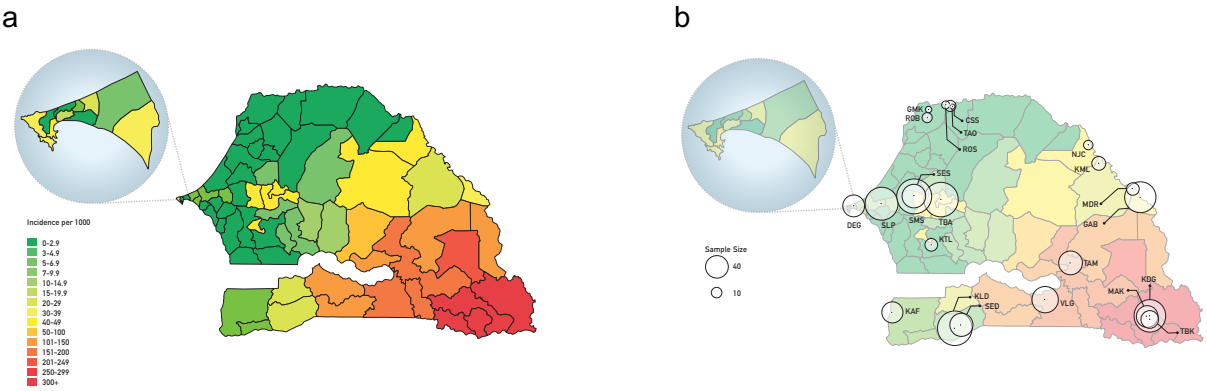


Figure 1: (a) Map of Senegal showing annual malaria incidence per 1000; (b) location of sample collection sites with relative sample size by site. Inset expands the Dakar region.

Table 1: Dataset and Site Information *Breakdown of samples from each site in the study. Each site has a three-letter code; in some cases, multiple sites are clustered into a single site with its own code for display on maps. Incidence: annual incidence per 1000; WGS: number of samples with whole genome sequence data; Barcode: samples with barcode genotyping data broken down by their classification as monogenomic (M) or polygenomic (P). Three sites with a single barcode sample have been omitted. Table shading matches the NMCP-designated malaria transmission zones: green (annual incidence <5/1000); yellow (incidence between 5 and 25/1000); red (annual incidence >25/1000).*

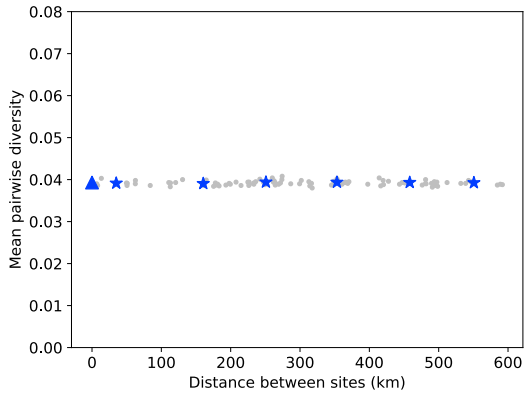
Codes		***Location***			***WGS***	***Barcode***	
Site	**Map**	**Region**	**District**	**Incidence**	**M**	**M**	**P**
GMK	STL	Saint Louis	Richard Toll	0.5	0	4	0
ROS	STL	Saint Louis	Richard Toll	0.8	0	5	2
KTL	KTL	Kaolack	Ndoffane	1.5	6	9	10
TAO	STL	Saint Louis	Richard Toll	1.6	0	2	0
NJC	STL	Matam	Matam	2	3	4	6
ROB	STL	Saint Louis	Richard Toll	2.1	0	9	3
CSS	STL	Saint Louis	Richard Toll	2.1	0	10	12
SLP	SLP	Thiès	Thiès	2.3	40	51	19
DEG	DEG	Dakar	Pikine	2.4	14	26	11
MDR	MDR	Tambacounda	Bakel	4.4	5	13	9
SED	SED	Sedhiou	Sedhiou	4.7	18	23	11
SES	DBL	Diourbel	Diourbel	7.5	13	90	12
KAF	KAF	Ziguinchor	Diouloulou	13.1	16	23	20
KML	KML	Matam	Kanel	13.2	7	10	11
SMS	DBL	Diourbel	Diourbel	15.2	3	39	15
TBA	TBA	Diourbel	Touba	19.1	41	61	17
GAB	GAB	Tambacounda	Bakel	64.3	33	47	25
VLG	VLG	Kolda	Velingara	127.8	21	41	53
KLD	KLD	Kolda	Kolda	208.3	32	70	40
KDG	KED	Kedougou	Kedougou	414.8	45	42	39
TAM	TAM	Tambacounda	Tambacounda	611.2	13	40	35
TBK	KED	Kedougou	Tomboronkoto	937.1	13	16	18
MAK	KED	Kedougou	Kedougou	1076.9	17	26	35
				TOTAL	**340**	**661**	**403**

Parasite relatedness varies with distance and reveals distinct patterns for two sites of similar incidence

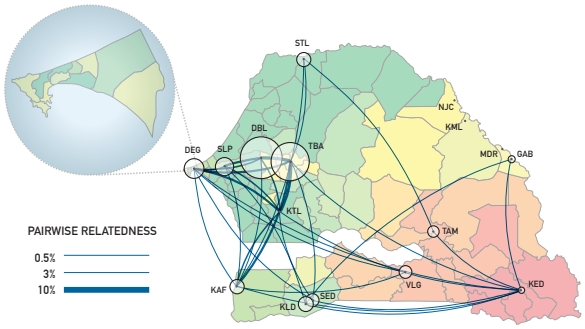
We first examined parasite genetic diversity, assessing in particular whether allele frequencies varied significantly throughout Senegal. As a measure we used the pairwise genetic distance between parasite genomes (i.e., the heterozygosity measured only at our trusted SNP loci). We found no evidence for systematic differences in allele frequencies and thus no evidence for parasite population structure within Senegal. Mean parasite diversity between study sites was independent of the distance between sites and was indistinguishable from diversity within sites (Fig. 2a).

Since allele frequencies provided no geographic information about parasites, we turned to parasite relatedness, which has elsewhere been shown to vary with distance and to be informative about parasite migration (e.g., [14]). Pairs of relatives were identified either as clones, based on barcode data, or as partial relatives, based on the fraction of their sequenced genomes that were identical by descent (IBD), that is, the fraction that consisted of chromosome segments with essentially identical alleles throughout [15-18] (see Methods). If the IBD fraction exceeded a threshold (4 or 5% of the genome, depending on sequencing depth — see Methods and Supplemental Fig. 2 for details), they were classified as relatives. Related parasites were recovered in 13 / 23 study sites (12 /15 sites with at least 25 samples) as well as between sites at all distance scales across the country (Fig. 2b). In total, 2,789 sample pairs were classified as related, amounting to 4.8% of all possible pairs. Unlike allele frequencies, the probability that a pair of parasites were related to each other (which we term ‘pairwise relatedness’) showed marked spatial structure, decreasing from 3.9% for parasites within the same site to less than 0.1% for separations greater than 300 km (Fig. 2c). While this trend conforms to expectations, we note that there is a large variance between sites separated by similar distances. Notably, all sites that have a high pairwise relatedness with other sites lie within the low transmission region in central Senegal and all also display high levels of within-site relatedness (Fig. 2b). This could reflect in part the fact that high within-site relatedness can make it easier to detect external relatives: descendants of imported parasites are more detectable in low transmission regions (because they make up a larger fraction of the local population) and persist for more generations (because there is less outcrossing).

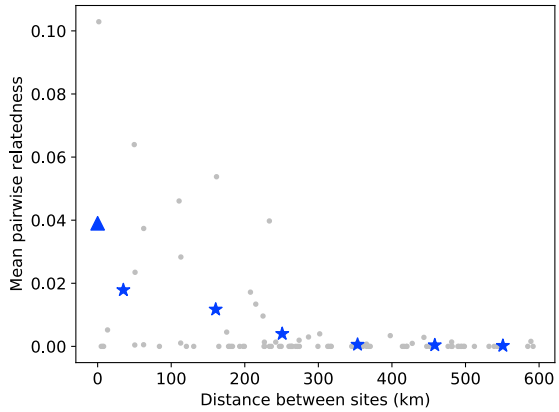
a



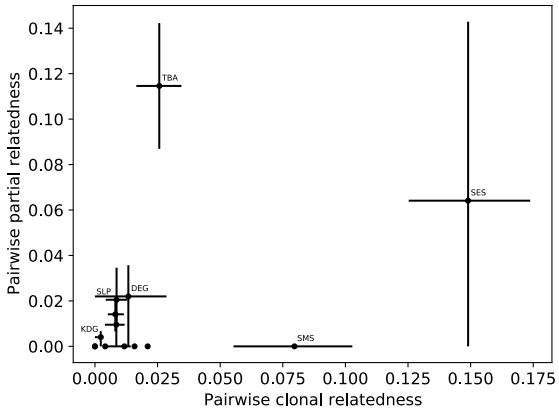
b



c



d



e

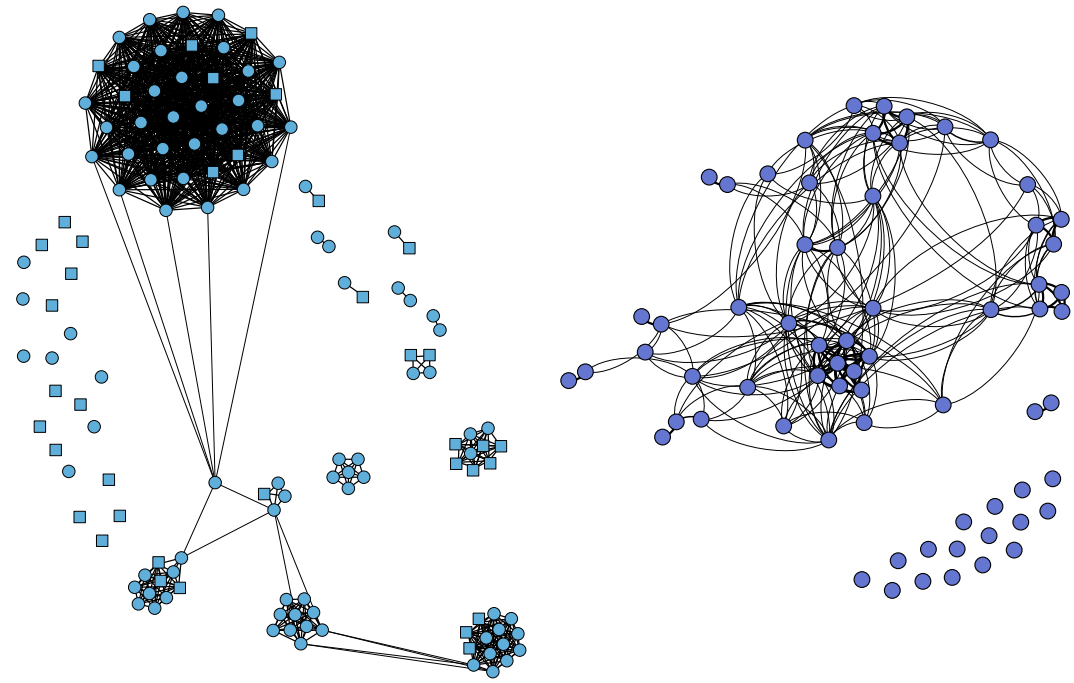


Figure 2. (a) Mean pairwise genetic diversity measured within (triangle) and between (stars) study sites as a function of distance (kilometers, km). Gray dots give values for individual site pairs, while triangles give weighted averages in 100 km bins. (b) Pairwise relatedness within and between sites, indicated by area of circles and thickness of lines, respectively. (c) Pairwise relatedness within (triangle) and between (stars) study sites as a function of distance (km). (d) Partial relatedness vs clonality for each study site (see Table 1 for site codes). (e) Networks of related parasites in Diourbel (left) and Touba (right). The two study sites in Diourbel are indicated by circles (SES) and squares (SMS). Line thickness is proportional to the degree of relatedness, with the thickest lines indicating clones.

The relatives identified here are of two distinct kinds, clones and partial relatives; as noted above, these have different relationships to transmission intensity, with partial relatedness more likely to occur where transmission is higher. We therefore investigated the two components separately, calculating *clonality* (the fraction of parasite pairs that are clones of each other) and *partial relatedness* (the fraction of non-clonal pairs that contain relatives) within each site (Fig. 2d). The two kinds of relatedness showed little correlation; in fact, the two cities with the highest overall relatedness had strikingly different patterns of relatedness. Touba (study site TBA, total relatedness = 0.137) had very high partial relatedness but unremarkable clonality. By contrast, Diourbel (study sites SES and SMS, total relatedness = 0.157), had clonality that was several-fold higher than in any other site and partial relatedness that was difficult to estimate because there were so few non-clonal parasites. The clonal parasites in Diourbel did not represent a single clonal expansion but instead occurred in numerous distinct clusters of varying size (Fig. 2e). Consistent with the finding of high clonality, the polygenomic fraction observed in the main study site in Diourbel (SES) is the lowest of any of our sites (0.12), implying that there is little opportunity for outcrossing to occur locally. The abundance of outcrossing evident in Touba, on the other hand, occurred despite the second lowest polygenomic fraction observed (0.22).

Transmission intensity is more correlated with the frequency of polygenomic infections than with relatedness

We next addressed one of our core goals, which was to investigate how well genetic measures can be used to estimate changing malaria burden. Previous work has shown that relatedness increases with decreasing transmission [11, 12], raising the possibility that measurements of relatedness could help guide malaria control efforts. To study the relationship of relatedness and transmission, we used a proxy for transmission, the reported malaria incidence at each site; this was based on case data routinely collected by the Senegal NMCP and calculated as the ratio of case counts per year to estimated catchment populations [13]. As predicted, pairwise relatedness was negatively correlated with incidence (Fig. 3a). However, the correlation was not strong (Pearson's r = -0.44) and the relationship was not at all linear. Partial relatedness and clonality calculate separately were even less correlated with incidence (r = -0.34 and r = -0.37, respectively). Primarily, the observed pattern suggests that at higher incidence (above roughly 50 per 1000 per year), relatedness is always low, while at lower incidence it can lie in a wide range.

Another, potentially more promising, aspect of genetics for tracking transmission [19] is COI, which has been proposed to offer both more accurate and more current information about transmission changes. The measure we investigated for this purpose was the frequency of polygenomic infections at each study site (termed the *polygenomic fraction* hereafter), based on the large number of samples for which we had barcode data. Given the small number of SNPs in the barcode, we treated COI as a dichotomous trait, with samples classified as "probable monogenomic" (2 heterozygous SNPs) or "probable polygenomic" (≥2 heterozygous SNPs, see Methods and Supplemental Fig. 3 for details).

Broadly, the polygenomic fraction correlated with transmission intensity across the country, decreasing from ~50% in the highest burden region in the southeast to 10–30% in the central region, but increasing again in the extremely low transmission region in the far north, where the fraction was 40% (Fig. 3b and, on a log scale, Supplemental Fig. 4). The correlation with measured incidence was moderately strong (Pearson's r = 0.77). The breakdown of the correlation at very low incidence is consistent with the previously reported behavior of COI [20]. The relationship was little changed when we adjusted the raw incidence estimates for differing levels of care seeking, testing of suspected cases, and reporting completeness at our study sites, using a previously described approach [21] (Supplemental Fig. 5); the most notable change was a seven-fold increase in the estimated incidence at the Velingara (VLG) site, making its estimated incidence more consistent with the observed polygenomic fraction.
The polygenomic fraction was more informative about incidence than relatedness, whether the latter was measured as total relatedness or separately as partial and clonal relatedness. When we regressed $\log_{10}$(incidence) on the polygenomic fraction and on relatedness, the former was the much better single predictor (R^2_{adj} = 0.527 for polygenomic fraction vs 0.127 for total relatedness, see Supplemental Table 1). Combining relatedness with the polygenomic fraction provided little additional information (R^2_{adj} = 0.550), as did replacing relatedness with its partial and clonal components (R^2_{adj} = 0.507).

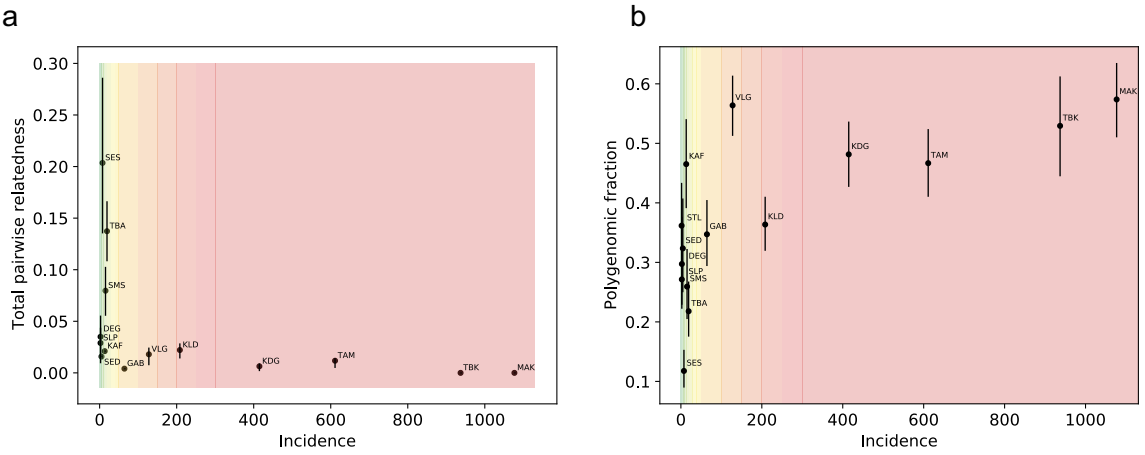
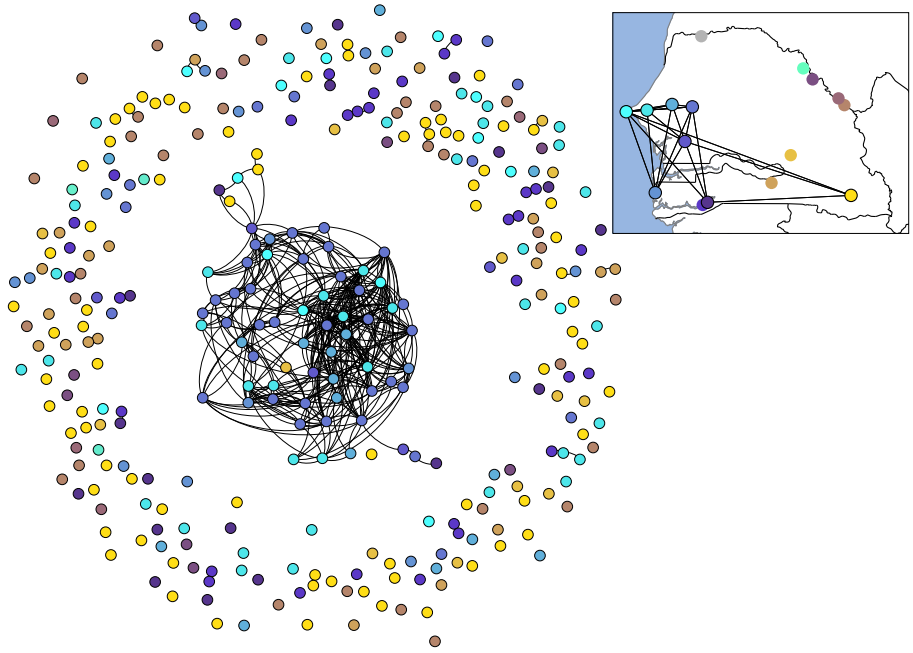


Figure 3. (a) Relationship between pairwise relatedness and reported incidence at study sites (site codes as in Table 1). (b) Relationship between polygenomic fraction and reported site incidence. In (a) and (b), error bars represent 68% confidence intervals arising from sampling error.

a



b

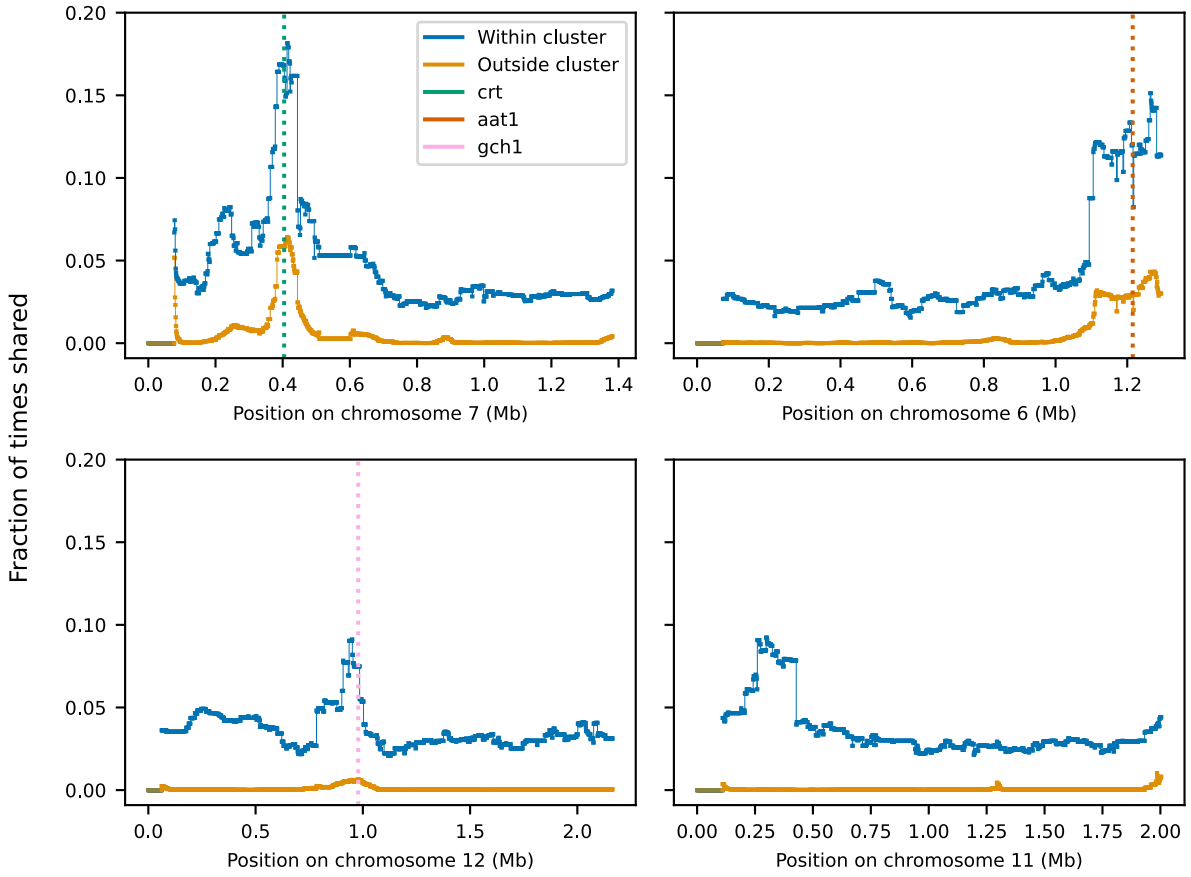


Figure 4. Parasite relatedness between study sites. (a) Network of relatedness between all sequenced samples (clones excluded). Samples are colored by study site as indicated on the inset map. (b) The fraction of sample pairs that are IBD as a function of genomic position, split into samples that are ("within cluster") or are not ("outside cluster") part of the large network of related parasites. The four examples shown are: a selective sweep around the known drug resistance locus at pfcrt (crt, green dotted line); a suspected sweep on chromosome 6 surrounding the aat1 locus (brown dotted line); a possible sweep on chromosome 12 containing the gch1 locus (pink dotted line); a possible unreported sweep on chromosome 11.

Parasite relatedness reveals network of connectedness across Senegal and loci under selection

When we grouped sequenced parasites from throughout the country into networks of partial relatives, we found a strikingly skewed distribution of network sizes. Omitting clones, there were 110 parasites with a relative in the dataset, or 32% of the total. These formed a total of 22 clusters of relatives. Twenty-one clusters contained two or three parasites and together accounted for 46 of the 110. The remaining single cluster contained the other 64 parasites (58% of partially related parasites) and included parasites from eight study sites across central and southern Senegal (Fig. 4a). This observation seems counterintuitive, since the large fraction of parasites in a single cluster suggests a small parasite effective population size, while the large number of partially related parasites requires frequent, repeated superinfection, something that does not normally occur in small populations.

One process that could contribute to this kind of clustering of relatives is ongoing positive selection, with increased relatedness driven by sharing of genomic segments under selection. This possibility would also be consistent with the observation that parasites within the large cluster were twice as likely to have a clone in our dataset as those from the same study sites but outside the cluster — 26.7% (16/60) vs. 12.5% (14/112) — although the difference was not statistically significant (p = 0.067, Fisher's exact test). To address this possibility, we looked for genome segments that were more widely shared (based on our IBD analysis) between cluster parasites than between non-cluster parasites. While we found no specific segments shared among the majority of cluster parasites, we did find multiple regions of increased sharing within the cluster (Fig. 4b). The most pronounced of these contain genetic loci known or suspected to be involved in drug resistance. For example, the *P. falciparum* chloroquine resistance transporter (*pfcrt*) locus (PF3D7_0709000) on chromosome 7 [22] is known to modulate drug resistance, the amino acid transporter (*aat1*) gene (PF3D7_0629500) on chromosome 6 has been implicated in drug resistance [23], and the GTP cyclohydrolase 1 (*gch1*) locus (3D7_1224000) on chromosome 12 has been implicated in antifolate resistance [24, 25]. In all of these cases, enhanced sharing was evident in both cluster and non-cluster associated parasites but was more pronounced among those in the cluster. The same was true of a region on chromosome 9 that has been reported previously to be under selection in Senegal and Gambia [26] but that has not been associated with drug resistance. A ~200 kb region, containing dozens of genes, on chromosome 11 showed excess sharing only among parasites

in the cluster, raising the possibility that it is an early signal of an emerging sweep (see Supplemental Materials for pileup plots for all chromosomes).

Discussion

While parasite genomic surveillance shows growing promise for informing national malaria control strategies, it is still limited by uncertainty about operationally informative metrics of genetic diversity. Here we describe a broad survey of *P. falciparum* genetic diversity across Senegal, collected through the NMCP's sentinel system, which offers an epidemiological framework for analysis of the genomic data. The data provide a countrywide baseline of circulating parasite diversity for ongoing surveillance, and a test set for identifying robust analytical approaches and meaningful metrics for understanding malaria transmission.

We found no evidence for parasite population structure within the country. In contrast, the distribution of related parasites, which reflects only very recent population history, showed clear geographic structure. Given the large historical population size of *P. falciparum* in the region, minimal population structure is not surprising. Less obvious is whether ongoing migration within the country will be enough to maintain that homogeneity despite currently small parasite population sizes across much of the country; this should become clearer with data from subsequent years.

We investigated the relationship of epidemiological incidence to statistics measuring different aspects of parasite genetics. This effort was motivated by the challenge of accurately monitoring malaria incidence [27], especially where transmission is low: direct measurement of the entomological inoculation rate becomes impractical when the rate is low, while estimates based on reported cases are biased by varying care-seeking behavior and access to health resources. Identifying genetic metrics that could serve as proxies for incidence would be of practical value for malaria control. Unfortunately, the same challenges also hinder efforts to assess the relationship between genetics and incidence, since we generally do not have firm estimates of the true incidence to act as a truth set. The size of the uncertainty is illustrated by one of our study sites, that in Velingara, for which correcting for biases in the reported incidence increased by estimated incidence by a factor of seven. Thus, a good understanding of the relationship between incidence and parasite genetics is likely to require multiple studies under different conditions as well as detailed modeling.

The statistics we compared to incidence were the frequency of relatedness (partial or complete) between parasites and the frequency of complex (polygenomic) infections in the population. We drew two main conclusions from the comparison. First, while both measures were correlated with local incidence, the polygenomic fraction was a much better predictor than relatedness, confirming modeling studies that identified COI-based statistics as superior to relatedness- or diversity-based ones [19, 28]. In particular, COI-based measures are predicted to respond more rapidly to changes in transmission [19], making them good candidates for monitoring the effect of interventions. Notably, the polygenomic fraction was the better predictor despite the limitations of low-resolution genetic data from the 24 SNP barcode, which did not allow us to estimate the number of genomes in polygenomic infections. This illustrates the continuing value

of low-resolution genetic data in capturing information about transmission and raises the hope that data with higher resolution for COI will be even more informative.

We also found that, while the correlation between either genetic measure and local incidence holds true for moderate and higher transmission regions, it breaks down where transmission is low (roughly, when incidence < 50/1000/year). In our low incidence sites, genetic measures suggest much higher incidence than is actually observed; similar behavior has been previously reported specifically for COI [20]. In the far north of Senegal, incidence is so low (~1/1000) that the same individual should rarely be infected twice simultaneously, and yet 40% of infections in the region are polygenomic. Based on the known epidemiology and a previous study of parasite genomics of this region, it seems likely that many parasites there represent recent importation from areas with higher burden [8, 28] and that polygenomic infections are either acquired elsewhere or are the result of local co-transmission of multiple parasite strains from a single imported polygenomic infection. It is thus not surprising that local parasite genetics in this part of Senegal reflect the characteristics of the source region more than the local transmission rate [29]. Similarly, relatedness should be low within low transmission regions because imported parasites are unlikely to be related to one another. When closely related parasites are detected in this kind of setting, they suggest at least some degree of ongoing local transmission. Better understanding of the transition of genetics into an importation-dominated regime will likely require both more in-depth study of particular regions and detailed modeling.

Despite its limitations as a predictor of transmission intensity, parasite relatedness proved to be illuminating about the heterogeneity of malaria transmission in Senegal. An unexpected finding was that two geographically close (~50 km apart) cities, Touba and Diourbel, with similarly high overall parasite relatedness differ markedly in the patterns of that relatedness, with partially related parasites common in Touba while clones dominate in Diourbel. Touba does have higher reported incidence than the main study site in Diourbel (15/1000 vs 8/1000), but the different transmission dynamics in the two cities may also reflect their differences as settings for malaria transmission. Touba, the second most populous city in Senegal, is located at the crossing of major roadways and is the site of the annual Grand Magal pilgrimage, during which more than 4 million individuals travel to the area [30] and mix with more than 1 million local inhabitants; in 2019, this took place on October 16 – 17, roughly in the middle of the sample collection period. Diourbel, on the other hand, is a smaller urban setting (estimated population around 100,000) that contains numerous large (500 – 1000 individuals) religious boarding schools called daaras, where up to several hundred school-aged boys, many from elsewhere in Senegal, live communally. Previous work has shown evidence of distinct, genetically identical parasites among infections from individual daaras elsewhere in Senegal, consistent with local and focal transmission [10].

Subsequent surveillance data should provide more fine-grained temporal data about transmission in Touba and the role that the Grand Magal pilgrimage plays in local transmission. This pilgrimage may serve as a natural experiment for studying the effect of malaria importation at different points in the transmission season; it draws millions of people from throughout Senegal and the broader region to a single city for a brief period, and that period shifts by ~11

days each year. Further data will also show whether the large number of clones and clonal clusters in Diourbel is a persistent feature and, if so, whether the same clones persist across the dry season or are replenished by new imported infections each year. Modeling and epidemiological investigation could both provide insights into the cause of the pattern in Diourbel, in particular investigating the possibility that it is driven by highly localized hot spots of transmission. Such hot spots could then be addressed with targeted, vector-based interventions. More broadly, these observations illustrate both the potential for genetic surveillance to provide actionable information about local transmission patterns and the need to better understand the implications of specific patterns.

Finally, another striking observation from the country-wide dataset was that clusters of partially related parasites had a highly skewed size distribution, consisting of 21 small clusters (2 – 3 parasites apiece) and a single large cluster connecting multiple sites and containing 64 parasites. Since this kind of pattern could signal the rapid spread of one or more positively selected alleles, we looked for genomic loci where IBD sharing was unusually common (a signature of positive selection [16]), and in particular was more common among related parasites. This revealed multiple signals of selection. Most of these were at loci known (chromosomes 4, 7, and 8) or suspected (chromosomes 6 and 12) to be under selection for drug resistance, while one was at a previously identified locus on chromosome 9 that is likely not associated with drug resistance. No single selected allele was responsible for all relatedness in the network of relatives, but several loci had markedly higher sharing within the network, and a signal at one candidate locus on chromosome 11 was seen only among samples in the network. These patterns could result either from enrichment for currently selected alleles among relatives or from higher frequency of selected alleles in the central part of Senegal where most of the related parasites were found; in either case; ongoing selection would be implicated. Since positive selection in *P. falciparum* is often driven by drug resistance alleles, we view these loci as targets for continued monitoring. More broadly, this finding suggests that searching for signals of selection among related parasites could be a valuable tool for surveillance of existing and new drug resistance loci, an approach that we are actively investigating.

Materials and Methods

Ethics Statement

Samples were obtained from febrile patients who presented at health facilities for care. Informed consent was obtained from all study participants. The study protocol was authorized by the Ministry of Health and Social Action in Senegal (SEN 19/49) and approved by the Institutional Review Board of the Harvard T.H. Chan School of Public Health (IRB protocol 16330). The CDC Human Research Protection Office reviewed the protocol and determined the CDC to be non-engaged.

Sample Collection

Blood samples were collected from clinics and health posts across Senegal from uncomplicated malaria infections detected by microscopy (clinics) or rapid diagnostic testing (RDT, sentinel sites). Collection sites were across the range of transmission intensity in Senegal, from very low

reported annual incidence (<1 per 1000) in the northwest to high reported annual incidence (>500 per 1000) in the southeast, including three main clinic sites, located in Thies, Diourbel, and Kedougou, that were augmented by health posts as part of the sentinel survey system of the National Malaria Control Program, with additional sites located in areas of interest. Blood was collected on Whatman 903 ProteinSaver filter paper material.

Incidence per thousand individuals in 2019 was calculated for each health facility using data provided by the Senegal NMCP [13]. We estimated the incidence from nearby sites for two sites (SLP and CSS) for which we did not have incidence data. The ROB site was used for CSS in Richard Toll, and a combination of incidence data from Medina Fall 1 and 2 were used as a proxy for the SLAP clinic in Thies.

DNA Extraction, Genotyping, Sequencing

Sample Workflow: Nucleic acid material was extracted from 1,353 blood spots collected on filter paper, with 1,066 of these samples subjected to pre-amplification, before molecular barcode genotyping of all samples. A total of 1,034 of the 1,353 samples passed initial barcode genotyping, with 636 samples classified as monogenomic infections. Of the 1,353 samples, 648 were subjected to whole genome sequencing (WGS).

DNA Extraction: Nucleic acid material was extracted from dried blood spots collected on filter paper obtained from all microscopy- and/or RDT-positive individuals presenting at clinics or health posts with fever according to routine Senegal protocols. Briefly, dried blood spots collected on ProteinSaver cards were extracted for genomic DNA from 2 to 3, 6-mm punches using the manufacturer protocol from the Promega Maxwell DNA IQ Casework Sample kit (Promega AS1210, Promega Corp., Madison, WI).

Genotyping: Extracted nucleic acid material was subjected to pre-amplification [31] and molecular barcode analysis [32]. The genotypes were called by their base designation (A, T, G, or C) with missing alleles identified by “X” and working alleles where both the major and minor alleles were designated by “N”. Only samples missing zero or one of the twenty-four assays were used for analysis, and monogenomic infections were called if there were zero or one of these assays called as “N”.

Sequencing: Genotyped samples with barcode calls of up to four missing positions and no more than two “N” calls were then used for WGS. Even though the criterion for monogenomic infections was no more than one “N” call, we included samples with two “N” calls as well to increase the likelihood of including all samples that could be analyzed for relatedness (described below). Samples were subjected to sWGA [33] involving multiple displacement amplification using Phi29 polymerase followed by magnetic bead clean up and quantification. Resultant material was subjected to fragmentation and NEBNext Ultra II FS DNA library construction according to manufacturer instructions (New England Biolabs, Beverly, MA).

Variant calling

Variant calling was performed in accordance with the best practices established as a part of the Pf3K project using GATK3.5.0 and *Plasmodium falciparum* 3D7v.3 reference assembly for read alignment with bwa-mem [34]. Picard toolkit was used to mark and remove duplicates and to assess quality control metrics. For genome sequence data, individual genotypes were discarded if the supporting read was < 5 reads. The downstream analysis was limited to a set of preferred SNP sites all located within the callable loci of *P. falciparum [1]*.

Selection of preferred SNP sites based on Pf3k data

To avoid artifacts caused by the high AT content and extensive low-complexity regions of the *P. falciparum* genome, we restricted our sequence analysis to a highly filtered set of 149,582 single nucleotide polymorphism (SNP) sites, where the filters (based on the global Pf3k dataset [34]) removed sites and regions with anomalously high heterozygosity.

To develop these filters, we downloaded the complete set of VCF files for Pf3k release 5 [34]. Polygenomic samples were identified with DEploid [35], and removed, along with duplicate samples from Malawi and the small number of samples from Nigeria (since there were too few of the latter for within-country analyses also performed with this dataset), leaving a total of 1328 samples.

Genotype calls for sites identified in the Pf3k VCF files as single nucleotide variants were extracted without restriction to biallelic sites, provided there were at least five copies of non-major alleles across the complete set of monogenomic samples. Heterozygous calls were tabulated for each site and were used to construct a series of filters to remove regions and sites with elevated rates of heterozygosity (Supplemental Fig. 1). The filters were as follows: 1) Sites were excluded unless they were in 'core' chromosome regions, as defined in Miles et al. [1] 2) Remaining regions were divided into non-overlapping 2 kb windows, which were filtered on the basis of mean heterozygosity in the window, with a threshold at 0.03 (Supplemental Fig. 1c). 3) SNPs within 7 base pairs of an indel were excluded (Supplemental Fig. 1d). 4) Individual SNPs with heterozygosity ≥ 0.04 were excluded. To the surviving set of SNPs we added the two sites from our 24-SNP barcode that were not already included, yielding 149,582 sites used in this study. Of these SNPS, approximately 40,000 proved to be polymorphic (minor allele frequency > 0.01) in our Senegal dataset.

Sample selection and filtering

Barcodes were excluded from further analysis if they had more than one missing assay. For the sequence data, all sites with a sequencing read depth < 5 in passing samples were masked out. Sequenced samples with a high rate of heterozygous calls (Supplemental Fig. 6) in the preferred SNP set were classified as possibly polygenomic and excluded from further analysis, with the threshold set at a heterozygous fraction of 0.0024. Sequenced samples for which less than 25% of sites had at least 5x sequencing depth were likewise excluded, leaving 340 samples in the dataset for analysis. Analysis of study sites was restricted to sites with at least 15 sequences.

Calculation of clonality and complexity of infection

Clonal pairs of parasites were identified from barcode data, with clones defined as pairs with zero mismatches between their barcodes and all other pairs defined as non-clones. See Supplemental Fig. 7 for a comparison with sequence data where sequence and barcode data overlapped.

Polygenomic samples were also identified from barcode data, as those having > 1 heterozygous genotype; since barcode genotypes were used to select monogenomic samples for sequencing, the barcode data was the primary source of information about COI. This classification was corrected with sequence data when available, based on the threshold on the heterozygous fraction given above. Based on comparison between barcode and sequence data for the same samples (Supplemental Fig. 6), we estimate that classification of a sample as monogenomic was correct 87% of the time when based on its barcode genotype alone and 91% of the time after applying overlapping sequencing information, while 97–98% of probable polygenomic calls were correct.

Calculation of relatedness

Clonality was calculated as the fraction of barcoded sample pairs that had zero genotype differences between them. 97% of samples identified as clones by barcode were confirmed by whole genome data in the 107 samples for which the comparison could be made.

IBD for estimating pairwise relatedness was calculated with hmmIBD [36], version 2.0.4, with options -m 20 -n 40 (maximum of 40 generations and 20 fit iterations). To study shared IBD segments in possible selective sweeps, a modified version of hmmIBD was used with the same parameters. In this version, the fraction of sites that are IBD, which is normally a free parameter in the model, was kept fixed at 50% while identifying IBD segments. This was done to remove a bias in the hidden Markov model, which is more likely to identify segments as IBD when the overall relatedness of the two samples is high. Pileup of IBD segments across sample pairs was then calculated in 1 kb windows; a sample pair was considered IBD within that window if an IBD segment overlapped the window at all.

IBD Threshold

To identify a minimum IBD fraction that would reliably signal close relatedness between parasites, we generated IBD distributions for two sets of sample pairs, one enriched for related pairs, based on time and location of sampling, and the other similarly depleted. For this purpose, we used an expanded dataset that included sequence data from Thiès (SLP) from prior years. The enriched sample set contained pairs of parasites where each parasite was from 2019 and from one of three sites with a high degree of relatedness: SLP, SES, and TBA. For the depleted set, the parasites in each pair had to have been sampled at least ten years apart and one of them had to come from a site other than the three listed above. We compared the distribution of IBD fractions between the two sample sets, making the comparison separately for pairs with high sequencing coverage (>50% of SNPs with at least 5x coverage) and for those with at least one parasite with lower coverage. Based on the comparison between the distributions (Supplemental Fig. 3), we defined all pairs with high coverage and with an IBD

fraction >4% to be related, while for pairs with lower coverage the threshold was set at 5%. Using these thresholds, depleted pairs were 0.6% as likely to be labeled 'related' as enriched pairs.

Total pairwise relatedness was calculated as (clonal relatedness) + (partial relatedness) * (1 - clonal relatedness), where clonal relatedness is the fraction of all pairs that are clones and partial relatedness is the fraction of non-clone pairs that are partially related. When calculating diversity and relatedness within sites and as a function of binned distance (Fig. 2 b and c), sites were weighted by the number of sequenced samples (within site) or the geometric mean of the number of sequenced samples from the two sites (between sites).

Estimating sampling error for relatedness measures

We empirically evaluated methods for calculating confidence intervals for pairwise partial relatedness within study sites. To do so, we took the expanded dataset mentioned above (this dataset plus earlier sequence data from Thiès), restricting sites to TBA, SLP, SES, SMS in order to approximate the relatedness seen within one site, and treated the dataset as a population with measured relatedness. From it we repeatedly drew sample sets of specified size to simulate the sample set for one of our sites and evaluated how often a particular method for calculating confidence intervals contained the true value. A standard bootstrap approach was anticonservative and biased, the latter because sampling with replacement introduces spurious relatedness. The method we identified as the most accurate was repeated downsampling of the test set without replacement, specifically, downsampling to 55% of the initial sample. Performing a similar test with the full set of barcodes, we identified 60% as the appropriate downsampled size for estimating CIs for pairwise clonality. When determining the final confidence intervals, downsampling was repeated 1000 times for each CI. Confidence intervals for relatedness between sites were determined by simple bootstrapping.

Network diagrams of related parasites were created with the software package Gephi, version 0.9.2.

Predictors of incidence

Ordinary least squares regression was performed with the Python package statsmodel.api. The incidence was log-transformed because of the wide range in incidence values, because fold changes at all incidence levels are relevant to malaria control efforts, and because of the highly nonlinear relationship between incidence and genetic measures. Goodness of fit was assessed by R^2_{adj}, which corrects for the number of explanatory variables in the regression.

Competing interests

The authors have no competing interests to declare.

Disclaimer

The findings and conclusions in this paper are those of the authors and do not necessarily represent the official position of the U.S. Centers for Disease Control and Prevention.

Data availability

Barcode data are provided as a supplemental file. Sequence data are currently being prepared for upload to the NCBI Sequence Read Archive; we will update this preprint when the BioProject number is assigned.

Acknowledgements

We would like to express gratitude to the people of Senegal, and to all the health care workers at the sentinel sites and clinics including medical directors, doctors, nurses, and clinic staff for their participation and support in the collection of samples required for this analysis. We thank the Senegal National Malaria Control Program in the Ministry of Health and Sanitation for their ongoing support and collaboration for these studies. We thank Kairon Shao for help in generating the genomic data. Funding for this work was provided by the Bill & Melinda Gates Foundation (OPP1156051) to DFW, and the National Institutes of Health (5R21AI141843-02) to SKV.

References:

1. Miles, A., et al., *Indels, structural variation, and recombination drive genomic diversity in Plasmodium falciparum.* Genome Res, 2016. **26**(9): p. 1288-99.
2. WHO, *Meeting report of the technical consultation on the role of parasite and anopheline genetics in malaria surveillance*. 2019, WHO: Geneva, Switzerland. p. 50.
3. Dalmat, R., et al., *Use cases for genetic epidemiology in malaria elimination.* Malar J, 2019. **18**(1): p. 163.
4. Jacob, C.G., et al., *Genetic surveillance in the Greater Mekong subregion and South Asia to support malaria control and elimination.* Elife, 2021. **10**.
5. Noviyanti, R., et al., *Implementing parasite genotyping into national surveillance frameworks: feedback from control programmes and researchers in the Asia-Pacific region.* Malar J, 2020. **19**(1): p. 271.
6. Moser, K.A., et al., *Describing the current status of Plasmodium falciparum population structure and drug resistance within mainland Tanzania using molecular inversion probes.* Mol Ecol, 2021. **30**(1): p. 100-113.
7. Ndiaye, Y.D., et al., *Genetic surveillance for monitoring the impact of drug use on Plasmodium falciparum populations.* Int J Parasitol Drugs Drug Resist, 2021. **17**: p. 12-22.
8. Daniels, R.F., et al., *Genetic evidence for imported malaria and local transmission in Richard Toll, Senegal.* Malar J, 2020. **19**(1): p. 276.
9. Liu, Y., et al., *Confirmation of the absence of local transmission and geographic assignment of imported falciparum malaria cases to China using microsatellite panel.* Malar J, 2020. **19**(1): p. 244.
10. Sy, M., et al., *Plasmodium falciparum genomic surveillance reveals spatial and temporal trends, association of genetic and physical distance, and household clustering.* Sci Rep, 2022. **12**(1): p. 938.
11. Neafsey, D.E., A.R. Taylor, and B.L. MacInnis, *Advances and opportunities in malaria population genomics.* Nat Rev Genet, 2021. **22**(8): p. 502-517.
12. Daniels, R.F., et al., *Modeling malaria genomics reveals transmission decline and rebound in Senegal.* Proc Natl Acad Sci U S A, 2015. **112**(22): p. 7067-72.

13. PNLP, *Malaria Annual Epidemiological Bulletin 2019*. 2020, National Malaria Control Program, Senegal.
14. Taylor, A.R., et al., *Quantifying connectivity between local Plasmodium falciparum malaria parasite populations using identity by descent.* PLoS Genet, 2017. **13**(10): p. e1007065.
15. Brown, T.S., et al., *Distinguishing gene flow between malaria parasite populations.* PLoS Genet, 2021. **17**(12): p. e1009335.
16. Henden, L., et al., *Identity-by-descent analyses for measuring population dynamics and selection in recombining pathogens.* PLoS Genet, 2018. **14**(5): p. e1007279.
17. Hill, W.G., et al., *Estimation of inbreeding coefficients from genotypic data on multiple alleles, and application to estimation of clonality in malaria parasites.* Genet Res, 1995. **65**(1): p. 53-61.
18. Wong, W., et al., *Genetic relatedness analysis reveals the cotransmission of genetically related Plasmodium falciparum parasites in Thies, Senegal.* Genome Med, 2017. **9**(1): p. 5.
19. Hendry, J.A., D. Kwiatkowski, and G. McVean, *Elucidating relationships between P.falciparum prevalence and measures of genetic diversity with a combined genetic-epidemiological model of malaria.* PLoS Comput Biol, 2021. **17**(8): p. e1009287.
20. Watson, O.J., et al., *Evaluating the Performance of Malaria Genetics for Inferring Changes in Transmission Intensity Using Transmission Modeling.* Mol Biol Evol, 2021. **38**(1): p. 274-289.
21. Thwing, J., et al., *A Robust Estimator of Malaria Incidence from Routine Health Facility Data.* Am J Trop Med Hyg, 2020. **102**(4): p. 811-820.
22. Djimde, A., et al., *A molecular marker for chloroquine-resistant falciparum malaria.* N Engl J Med, 2001. **344**(4): p. 257-63.
23. Tindall, S.M., et al., *Heterologous Expression of a Novel Drug Transporter from the Malaria Parasite Alters Resistance to Quinoline Antimalarials.* Sci Rep, 2018. **8**(1): p. 2464.
24. Heinberg, A. and L. Kirkman, *The molecular basis of antifolate resistance in Plasmodium falciparum: looking beyond point mutations.* Ann N Y Acad Sci, 2015. **1342**(1): p. 10-8.
25. Rocamora, F. and E.A. Winzeler, *Genomic Approaches to Drug Resistance in Malaria.* Annu Rev Microbiol, 2020. **74**: p. 761-786.
26. Duffy, C.W., et al., *Multi-population genomic analysis of malaria parasites indicates local selection and differentiation at the gdv1 locus regulating sexual development.* Sci Rep, 2018. **8**(1): p. 15763.
27. WHO, *World Malaria Report 2022*. 2022.
28. Albert Lee, Y.D.N., Aida Badiane, Awa Deme, Rachel F. Daniels, Stephen F. Schaffner, Fatou Ba Fall, Médoune Ndiop, Alioune Badara Gueye, Ibrahima Diallo, Katherine E. Battle, Edward A. Wenger, Caitlin A. Bever, Doudou Sene, Bronwyn MacInnis, Dyann F. Wirth, Daouda Ndiaye, Daniel L. Hartl, Sarah K. Volkman, Joshua L. Proctor, *Modeling the levels, trends, and connectivity of malaria transmission using genomic data from a health facility in Thiès, Senegal.* medRxiv, 2021.
29. Wong, W., et al., *R (H): a genetic metric for measuring intrahost Plasmodium falciparum relatedness and distinguishing cotransmission from superinfection.* PNAS Nexus, 2022. **1**(4): p. pgac187.
30. Gautret, P., et al., *The 2020 Grand Magal of Touba, Senegal in the time of the COVID-19 pandemic.* Travel Med Infect Dis, 2020. **38**: p. 101880.
31. Mharakurwa, S., et al., *Pre-amplification methods for tracking low-grade Plasmodium falciparum populations during scaled-up interventions in Southern Zambia.* Malar J, 2014. **13**: p. 89.

32. Daniels, R., et al., *A general SNP-based molecular barcode for Plasmodium falciparum identification and tracking.* Malar J, 2008. **7**: p. 223.
33. Oyola, S.O., et al., *Whole genome sequencing of Plasmodium falciparum from dried blood spots using selective whole genome amplification.* Malar J, 2016. **15**(1): p. 597.
34. MalariaGEN, *An open dataset of Plasmodium falciparum genome variation in 7,000 worldwide samples.* Wellcome Open Research, 2021. **6**(42).
35. Zhu, S.J., J. Almagro-Garcia, and G. McVean, *Deconvolution of multiple infections in Plasmodium falciparum from high throughput sequencing data.* Bioinformatics, 2018. **34**(1): p. 9-15.
36. Schaffner, S.F., et al., *hmmIBD: software to infer pairwise identity by descent between haploid genotypes.* Malar J, 2018. **17**(1): p. 196.