RESEARCH

Evaluating the performance of *Plasmodium falciparum* genetic metrics for inferring National Malaria Control Programme reported incidence in Senegal

Wesley Wong[1], Stephen F. Schaffner[2], Julie Thwing[3], Mame Cheikh Seck[4], Jules Gomis[4], Younouss Diedhiou[4], Ngayo Sy[5], Medoune Ndiop[6], Fatou Ba[6], Ibrahima Diallo[6], Doudou Sene[6], Mamadou Alpha Diallo[4], Yaye Die Ndiaye[4], Mouhamad Sy[4], Aita Sene[4], Djiby Sow[4], Baba Dieye[4], Abdoulaye Tine[4], Jessica Ribado[7], Joshua Suresh[7], Albert Lee[7], Katherine E. Battle[7], Joshua L. Proctor[7], Caitlin A. Bever[7], Bronwyn MacInnis[2], Daouda Ndiaye[4], Daniel L. Hartl[8], Dyann F. Wirth[1,2] and Sarah K. Volkman[1,2,9*]

Abstract

Background Genetic surveillance of the *Plasmodium falciparum* parasite shows great promise for helping National Malaria Control Programmes (NMCPs) assess parasite transmission. Genetic metrics such as the frequency of polygenomic (multiple strain) infections, genetic clones, and the complexity of infection (COI, number of strains per infection) are correlated with transmission intensity. However, despite these correlations, it is unclear whether genetic metrics alone are sufficient to estimate clinical incidence.

Methods This study examined parasites from 3147 clinical infections sampled between the years 2012–2020 through passive case detection (PCD) across 16 clinic sites spread throughout Senegal. Samples were genotyped with a 24 single nucleotide polymorphism (SNP) molecular barcode that detects parasite strains, distinguishes polygenomic (multiple strain) from monogenomic (single strain) infections, and identifies clonal infections. To determine whether genetic signals can predict incidence, a series of Poisson generalized linear mixed-effects models were constructed to predict the incidence level at each clinical site from a set of genetic metrics designed to measure parasite clonality, superinfection, and co-transmission rates.

Results Model-predicted incidence was compared with the reported standard incidence data determined by the NMCP for each clinic and found that parasite genetic metrics generally correlated with reported incidence, with departures from expected values at very low annual incidence (< 10/1000/annual [‰]).

Conclusions When transmission is greater than 10 cases per 1000 annual parasite incidence (annual incidence > 10‰), parasite genetics can be used to accurately infer incidence and is consistent with superinfection-based hypotheses of malaria transmission. When transmission was < 10‰, many of the correlations between parasite genetics and incidence were reversed, which may reflect the disproportionate impact of importation and focal transmission on parasite genetics when local transmission levels are low.

*Correspondence:
Sarah K. Volkman
svolkman@hsph.harvard.edu
Full list of author information is available at the end of the article

Background

Pathogen genomics is revolutionizing public health by providing a rich data source for informing real-time, actionable recommendations for public health programmes [1]. Each pathogen genome is a unique record of its previous transmission history that can be used to study the origin and spread of infectious diseases in real time. Genetic surveillance of pathogen populations provides an opportunity to characterize pathogen transmission structure and provide data-informed recommendations to public health programmes to decrease transmission. In the past two decades alone, breakthroughs in genomic technologies and analytical techniques [2–4] have expanded pathogen genetic surveillance to a wide variety of viral and bacterial pathogens. Recent examples include the SARS-CoV-2 pandemic [5, 6], the 2013–2016 West African Ebola outbreaks [7, 8], and the Middle East Respiratory Syndrome (MERS) outbreaks in the Middle East [9].

Despite this success, extending genetic surveillance to more complex pathogens, such as the eukaryotic *Plasmodium falciparum* parasite that is the causative agent for the deadliest form of malaria, has been challenging. Unlike viral or bacterial pathogens, *P. falciparum* has a complex, 23-megabase genome with over 5000 genes whose genomic architecture is heavily influenced by meiotic recombination [10]. *Plasmodium falciparum* must undergo sexual reproduction within a mosquito vector to complete its life cycle prior to being transmitted to a new human host. The sexual nature of the *P. falciparum* parasite complicates many of the phylogenetic and phylodynamic techniques used in viral and bacterial genetic surveillance studies [2–4].

Malaria genetic surveillance has instead relied on identifying genetic epidemiology metrics that summarize the changes in parasite genetics observed from the empirical sampling of parasite genomes from malaria endemic regions. These genetic epidemiology metrics include the frequency of multiple strain (polygenomic) infections [11, 12], the number of strains per infection (complexity of infection, COI), the genetic relatedness of parasite strains [13, 14], and the frequency of clonal parasites in the population [15, 16].

Many of these genomic epidemiology metrics were identified by comparing sites with different levels of transmission intensity, whose measurement includes prevalence (frequency of infections), incidence (rate of new infections), and the entomological inoculation rate (EIR, number of infectious mosquito bites per individual). As such, malaria genetic epidemiology metrics tend to be associated with transmission intensity. Regions with high transmission intensity are expected to have high frequencies of polygenomic infections and high COIs because individuals are more likely to be superinfected with multiple infectious bites [17] and there is greater opportunity for parasite outcrossing. Conversely, regions with low transmission intensity are expected to have high frequencies of clonal or genetically related parasites [18, 19] due to increased levels of inbreeding associated with declining transmission and smaller effective parasite population sizes.

However, recent genomic analyses of polygenomic infections show that a large fraction of polygenomic infections are not the result of superinfection, but instead from the cotransmission of multiple parasite strains from a single infectious bite [13, 20–23]. Cotransmitted polygenomic infections do not represent multiple infectious bites and their presence suggests that superinfection-based predictions of transmission intensity from metrics such as the frequency of polygenomic infections and COI may be inaccurate. Despite this, lower rates of cotransmission were previously observed in Kedougou, a high transmission site of Senegal, than in Thies or Richard Toll, which are both sites with very low transmission [13]. These results suggest that the frequency of cotransmitted polygenomic infections could also be used to infer transmission intensity.

Mechanistically, cotransmission and superinfection are important drivers of parasite genetics, but how cotransmission and superinfection define the relationship between genetic metrics and epidemiological measures of transmission intensity is unknown, and it is unclear whether these relationships are consistent across the range of transmission from low to high intensity. It is also unclear to what extent other epidemiological factors, such as transmission heterogeneity (e.g., focal transmission) and importation, affect these genetic epidemiology metrics. A major goal for this study was to characterize the relationship between parasite genetics and transmission intensity using metrics that measure the impact of both superinfection and cotransmission, and to determine whether predictions of transmission intensity can be improved if both factors are considered.

In this study, the relationship between parasite genetics and malaria incidence as reported by the National Malaria Control Programme (NMCP) was examined. Malaria transmission in Senegal is highly heterogeneous and dependent on geographic location, ranging from < 1‰ to > 1000‰ annual incidence. This geographic disparity was ideal for evaluating the relationship between parasite genetics and incidence across a range of transmission intensities in a limited geographic area where reported incidence and genetic epidemiology metrics could be measured consistently across study sites and years. A series of mathematical models were used to quantify the relationship between parasite genetics and

Wong *et al. Malaria Journal*   (2024) 23:68

Page 3 of 13

incidence and identify transmission regimes (regions within the incidence parameter space) where the relationships between parasite genetics and incidence differ. Identifying these transmission regimes is important because they can arise from fundamental changes in transmission structure that affect how parasite genetics can be used to study transmission and develop data-informed public health recommendations.

Methods

Sampling strategy
Samples from health facilities were collected as previously described [13]. Samples were collected through PCD from febrile patients reporting to health posts or clinics during the malaria transmission season in Senegal (September to December), or actively detected in households in response to a case detected at the Bandafassi, Ndoga Babacar, Sessene, and Richard Toll clinics. Patients over 6 months of age presenting to the clinic with fever within the past 24 h and no history of anti-malarial use were diagnosed with malaria using microscopy or rapid diagnostic tests (RDTs). Parasite genomic DNA was extracted from filter papers spotted with blood samples collected from malaria-positive patients at all sites except Richard Toll. For Richard Toll, parasite genomic DNA was extracted from malaria positive RDT cassettes.

Ethical approval for these studies was obtained from the Ministry of Health and Social Action in Senegal (Avis Protocol SEN1949) and the Harvard T.H. Chan School of Public Health Institutional Review Board (Protocol 16330).

Barcoding with a 24 SNP molecular barcode
The 24 SNP (single nucleotide polymorphism) molecular barcode [24] provides a high-level snapshot of genetic diversity that trades genomic resolution for epidemiological sampling breadth. The barcode consists of 24 neutral SNPs spread throughout the malaria genome that are genotyped using a panel of TaqMan-based quantitative PCR (polymerase chain reaction) genotyping assays. Nucleic acid material was extracted from either filter paper or RDT and preamplified using previously established methods [25, 26]. Barcodes that had fewer than two missing sites were retained for analysis. Barcodes with two or more heterozygous sites were classified as polygenomic [13].

Full details regarding the criteria for calling homozygous and heterozygous sites are presented in a previous publication [13]. Briefly, homozygous and heterozygous sites were based on the difference in cycle threshold (the ΔCT) for the alleles present at each barcode position. The ΔCT threshold used to differentiate homozygous

and heterozygous sites was based on maximizing the true homozygous and true heterozygous rates observed in a series of DNA mixtures from lab-adapted 3D7, Dd2, and TM90C6B *P. falciparum* strains. These mixtures contained different pairwise combinations of the three laboratory-adapted strains mixed at different ratios ranging from 1:1 to 1:10. The true homozygous call rates in these mixtures were > 0.95. The detection limit for heterozygous sites depended on the mixture. For 1:1 and 1:3 mixtures, the true heterozygous call rate was > 0.95. For 1:5 and 1:10 mixtures, the true heterozygous call rates were 0.83 and 0.53, respectively.

Parasite genetic epidemiology quantification
Each of the five genetic epidemiology metrics were calculated for each site-year according to previously established methods. Polygenomic infections were identified as those infections whose barcode had two or more heterozygous sites. Polygenomic fraction was calculated by dividing the number of polygenomic infections by the total number of samples with a usable barcode (fewer than two missing sites). The fraction of non-unique monogenomic clones was defined as the proportion of monogenomic infections with a barcode genotype that is also found in one or more other monogenomic infections in the population. It was calculated as $1 - p_{mono,unique}$, where $p_{mono,unique}$ is the proportion of monogenomic infections with a unique barcode genotype. R_H and the cotransmission fraction were calculated as previously described [13]. Briefly, R_H quantifies the deviation in the observed, intra-host heterozygosity of polygenomic infections with the simulated expectations for a COI = 2 superinfection. Polygenomic infections with an $R_H > 0.30$ are inferred to be the result of cotransmission. *THE REAL McCOIL* [27] COI was calculated independently for each study site using the categorical method with the following parameter values: maxCOI = 25, threshold_ind = 20, threshold_site = 20, and err_method = 3. All other parameters used the default values. The median value estimated by *THE REAL McCOIL* was used as the point estimate of COI for each sample.

NMCP-reported incidences
When possible, NMCP-reported, district-level incidences for each collection site were used. However, this data was only readily available for the data collected in and after 2019. For older data, the region-level incidences reported in the annual NMCP reports were used [28–34].

Poisson generalized linear model
Model predictions were made with a Poisson Generalized Linear Model (GLM):

Wong *et al. Malaria Journal*   (2024) 23:68

Page 4 of 13

$$\log(\lambda_i) \sim x_i^T \beta \tag{1}$$

where λ_i is the predicted incidence for a given site-year, x_i is the vector containing the values of the covariates used in the model, and β is the vector of coefficients to be estimated. x_i includes the polygenomic fraction, the fraction of non-unique monogenomic clones, R_H, COI, and cotransmission fraction for each site-year and two categorical variables for (1) the region of origin and (2) whether incidences were measured at the district-level or at actual health facility catchment area. Models were fit using the GLM function in the Python 3 package *statsmodel* (v0.13.5).

Leave-one-out cross-validation was performed by splitting the dataset by sampling site and using all but one of the sites during model fitting. All site-years associated with the chosen site were removed from model fitting and was repeated for each of the 16 studied sampling sites. The estimates reported in this study used the average value obtained from leave-one-out cross-validation.

Akaike information criterion

The Akaike information criterion (AIC) was calculated as:

$$AIC = 2k - 2L \tag{2}$$

where k is the number of free parameters and L is the log likelihood. For the GLM trained on all the data, $k=7$ (one for each of the parameters used in the GLM). For the piecewise GLM, $k=8$, to include the additional incidence threshold parameter that separates $GLM_{below10}$ from $GLM_{above10}$.

Results

Study design overview

3147 *P. falciparum* clinical infections collected using PCD at 16 health facility clinic sites in Senegal were genotyped with a 24 SNP molecular barcode [24] (Fig. 1, Additional file 2: Fig. S1). Five genetic metrics (polygenomic fraction, COI, the fraction of non-unique monogenomic clones, R_H [13], and the cotransmission fraction) were calculated from the molecular barcode data for each site-year (Table 1, Additional file 2: Table S1, Fig. 2). Polygenomic
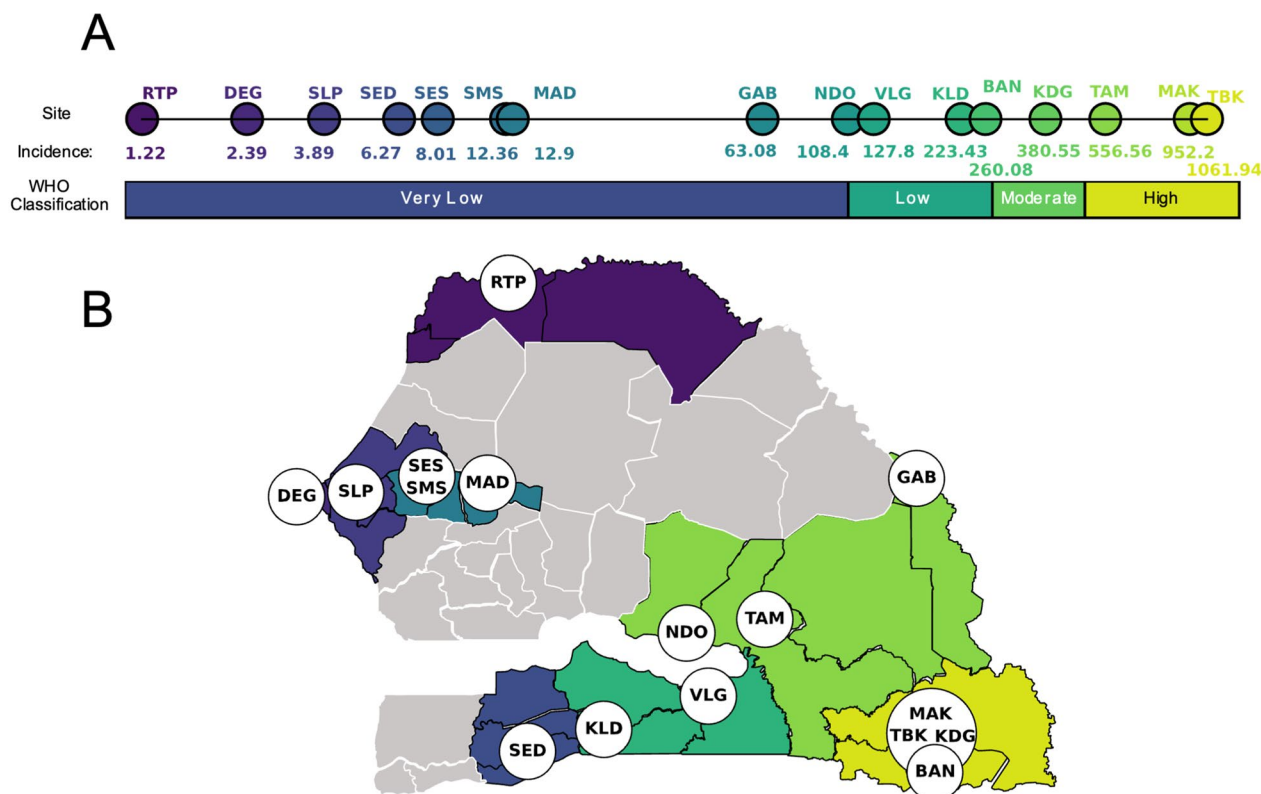
Fig. 1 Study design overview. Samples were collected from 16 sites throughout Senegal. (**A**) The average, NMCP-reported annual incidence values (‰) for each site during the time of sampling and their transmission classification according to the WHO. (**B**) The locations of each of the sampling sites and the district they are located in. The coloration used in the map is based on the average, district-level incidences, with yellow indicating the district with the highest average incidence and dark purple indicating the lowest

Wong *et al. Malaria Journal*   (2024) 23:68

Page 5 of 13

Table 1 Summary of the genetic metrics used in this study and their expected relationship with transmission intensity

Genetic metric	Definition	Transmission reflected	Expectation
Polygenomic fraction	Proportion of infections that are polygenomic	Superinfection and cotransmission	Increase with transmission intensity
Complexity of infection (COI)	Number of strains per infection	Superinfection and cotransmission	Increase with transmission intensity
The fraction of non-unique monogenomic clones	Proportion of monogenomic infections carrying a clone that was sampled elsewhere in the population	Clonal transmission	Decrease with transmission intensity
R_H	Estimate of polygenomic inbreeding based on intra-host heterozygosity	Superinfection and cotransmission	Decrease with transmission intensity
Cotransmission fraction	Proportion of polygenomics resulting from cotransmission	Cotransmission	Decrease with transmission intensity

Clones are defined as a parasite whose barcode is observed more than once in the population and is calculated only for monogenomic infections. COI was calculated using *THE REAL McCOIL*
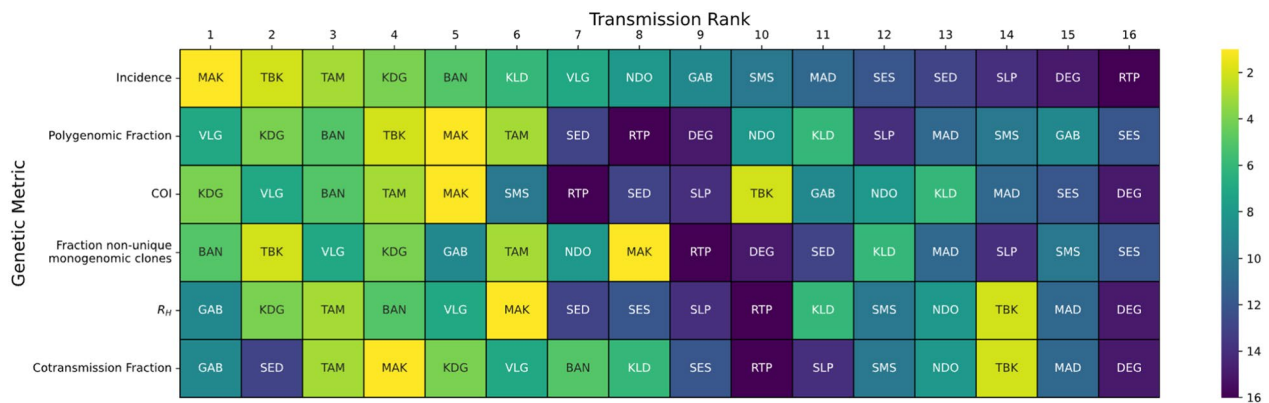
Fig. 2 Rank Analysis for each sample site based on the genetic metrics used in this study is shown. Transmission rank is defined as the rank one would use when using the metric to infer transmission intensity (Table 1), with 1 being the highest and 16 being the lowest. Assigned transmission ranks were assigned based on the average value calculated across sample years. For polygenomic fraction, *The REALMcCOI* [27] COI, and R_H [13], the transmission rank is positively correlated with the metric. For cotransmission fraction and the fraction of non-unique monogenomic clones, the transmission rank is negatively correlated with the metric

fraction and COI were chosen because they are among the most reported genetic metrics for assessing transmission intensity from malaria genetics. R_H is an estimate of intrahost heterozygosity that is designed to determine whether a polygenomic infection is the result of cotransmission or superinfection. The cotransmission fraction is the proportion of polygenomic infections with an R_H estimate greater than 0.3, which was previously identified to be the threshold for distinguishing cotransmitted from superinfected infections. The fraction of non-unique monogenomic clones represents the frequency of clonal transmission in the population, which both empirical and simulation-based studies suggest becomes more frequent as transmission intensity declines [14]. Polygenomic fraction and COI were expected to be positively correlated with incidence, while R_H, the fraction of non-unique monogenomic clones, and the cotransmission fraction

were expected to be negatively correlated with incidence (Table 1).

Malaria transmission in Senegal is highly heterogeneous, and the annual incidences reported by the NMCP for the clinic sites ranged from < 1 to > 1000‰ (Fig. 1A). Based on the definitions established by the World Health Organization (WHO) [35, 36], eight sites were "very low" transmission settings (< 100‰), three were "low" transmission settings (100–250‰), two were "moderate" transmission settings (250–450‰), and three were "high" transmission settings (> 450‰) (Fig. 1A). Five sites were sampled for multiple years: Dalaba (KDG, 2015–2020), Touba (MAD, 2019–2020), Sessene (SES, 2018–2020), Thies (SLP, 2015–2020), and Richard Toll (RTP, 2012–2015). The remaining 11 sites were sampled in either 2019 or 2020 or both (Table 2). NMCP estimates of annual incidence for each site were based on either the

Wong *et al. Malaria Journal*   (2024) 23:68

Page 6 of 13

Table 2 The regions, district, three-letter code, and sampling years for each site.

Region (RM)	District (DS)	Site (PS)	Code	Sample years
Kedougou	Kedougou	Tomboronkoto	TBK	2019
Kedougou	Kedougou	Mako	MAK	2019
Tambacounda	Tambacounda	Dialocoto	TAM	2019
Kedougou	Kedougou	Dalaba	KDG	2015–2020
Kedougou	Kedougou	Bandafassi	BAN*	2019
Kolda	Kolda	Bagadadji	KLD	2019
Kolda	Velingara	Ouassadou	VLG	2019
Tambacounda	Makacoulibantang	Ndoga Babacar	NDO*	2019
Tambacounda	Bakel	Gabou	GAB	2019
Diourbel	Diourbel	Madiyana 2	MAD	2019–2020
Diourbel	Diourbel	Keur Serigne Mbaye Sarr	SMS	2019
Diourbel	Diourbel	Sessene	SES*	2018–2020
Sedhiou	Sedhiou	Centre De Santé Goudomp	SED	2019
Thies	Thies	SLAP (Service de Lutte Antiparasitaire Clinic)	SLP	2015–2020
Dakar	Pikine	Deggo	DEG	2019
St-Louis	Richard Toll	Multiple	RTP*	2012–2015

Sites with ACD of household communities are asterisked (*)

reported district-level or region-level incidences for each clinical site ("Methods").

Transmission rank analyses reveal limitations of using individual genetic metrics to infer incidence

Whether parasite genetic metrics could be used to reliably rank sites by transmission intensity was evaluated by assigning each of the 16 sites examined in this study a transmission rank based on the expectations [37–39] listed in Table 1 and comparing them with the rank assigned by incidence (Fig. 2). This approach was designed to mimic inferences made using data collected from a single parasite genetic metric. A major goal of this rank analysis was to evaluate the consistency of the transmission ranks assigned by each genetic metric. For sites with multiple years, the rank was based on the average value. Overall, the consistency between the two was weak, with small to moderate amounts of concordance (Kendall rank correlation coefficient < 0.38, Additional file 2: Table S2). Of the genetic metrics examined, only the rank correlation between the fraction of non-unique monogenomic clones and incidence was statistically significant (Kendall rank correlation = 0.38, p-value = 0.04).

However, transmission ranks tended to be consistent when grouping the data into categories meant to represent those with "higher" and "lower" transmission sites. Transmission ranks were organized into 2×2 contingency tables (incidence *versus* each of the genetic metrics) with two categories: high transmission ranking (transmission rank < 8) and low transmission ranking

(transmission rank ≥ 8). The threshold at 8 was determined by visual analysis of the consistency in transmission ranks observed in the data (Additional file 2: Fig. S1). These thresholds roughly correspond to the WHO-defined categorization of "very low" transmission but are not exactly the same. Splitting the transmission rank data improved the correlation between each of the genetic metrics and incidence (Yule's Q between 0.47 and 0.96, Additional file 2: Table S2). The fraction of non-unique monogenomic clones, polygenomic fraction, and cotransmission fraction were correlated with these incidence groupings, but only the fraction of non-unique monogenomic clones was statistically significant (Yule's $Q = 0.96$, Pearson's chi-square correlation p-value = 0.01).

Multi-variate regression analyses identify different correlations between parasite genetics and NMCP-reported incidence

The transmission rank analyses showed that, while some genetic metrics could be used to broadly differentiate sites with annual incidence greater versus less than < 100‰, the exact transmission rank assigned by each genetic metric conflicted with one another and with incidence. One way of resolving these conflicts would be to use a multi-variate regression model to generate a single, site-specific incidence prediction using all the information collected from each genetic metric. A series of Poisson mixed-effects generalized linear models (GLM, "Methods") were constructed to quantify the relationship between genetics and incidence based on different

combinations of the five genetic metrics used in this study. This approach allowed one to evaluate the predictive power of each genetic metric alone and in combination with one another.

Model predictions from a GLM utilizing all five genetic metrics revealed two parameter regions with opposing model bias corresponding to regions with annual incidence > 10‰ and those with < 10‰. (Fig. 3). Overall, model predictions for regions with annual incidence > 10‰ were consistent with the reported data (Fig. 3A). However, the model tended to underestimate incidence relative to the official, NMCP-reported incidence values (Additional file 2: Fig. S4A). When annual incidence is > 10‰, increasing incidence was associated with increasing polygenomic fraction (Pearson correlation coefficient $r = 0.77$, p-value = 1.6e−4) and COI ($r = 0.60$, p-value = 8.6e−3), but decreasing the fraction of non-unique monogenomic clones ($r = 0.60$, p-value = 9.8e−6) (Additional file 2: Fig. S5). R_H and cotransmission fraction were also negatively associated but their correlations were not statistically significant ($r = -0.27$, p-value = 0.29 and $r = -0.17$, p-value = 0.496, respectively).

The patterns observed in regions with annual incidence < 10‰ differed from those observed in regions with > 10‰. Unlike in the higher transmission regions, the model predictions were not consistent with the NMCP-reported incidences and were consistently overestimated. The inability of the model to accurately predict incidence in regions with < 10‰ was because many of the correlations between parasite genetics and incidence differed from those observed in higher transmission regions (Fig. 4). When annual incidence < 10‰, sites with higher incidence were associated with decreasing polygenomic fraction ($r = -0.75$, p-value = 2.0e−3) and COI ($r = -0.60$, p-value = 0.04), but increasing the fraction of non-unique monogenomic clones ($r = 0.77$, p-value = 1.38e−3). The correlations between R_H and cotransmission fraction were also reversed but statistically insignificant ($r = 0.10$, p-value = 0.73 and $r = 0.12$, p-value = 0.69, respectively).

Of the metrics examined, the detection of non-unique clones had the greatest potential to be impacted by sample size. To account for this, model predictions were also made using down-sampled estimates of the fraction of non-unique monogenomic clones to control for differences in monogenomic infection sample size (Additional file 2: Fig. S6). The down-sampled fraction of non-unique monogenomic clones estimates were lower (Additional file 2: Fig. S6A, B) but otherwise had no major effect on model predictions (Additional file 2: Fig. S6C).

Based on these observations, a piecewise GLM model that splits the data into two groups, one for annual incidence < 10‰ ($GLM_{below10}$) and one for annual incidence > 10‰ ($GLM_{above10}$), was generated. The inflection point that separates $GLM_{below10}$ and $GLM_{above10}$ was determined by performing a parameter sweep over different incidence values (Additional file 2: Fig. S7). However, the dataset had a noticeable gap in sampling for sites with annual incidences between 10 and 100‰, which made it unclear how reliably we could identify the inflection point. For the purposes of this study, a conservative threshold of 10‰ was used because it represents the minimum value where splitting the data caused the model fits to improve. The resulting piecewise model significantly improved model fits (AIC [Akaike information criterion] = 1291.34) relative to the
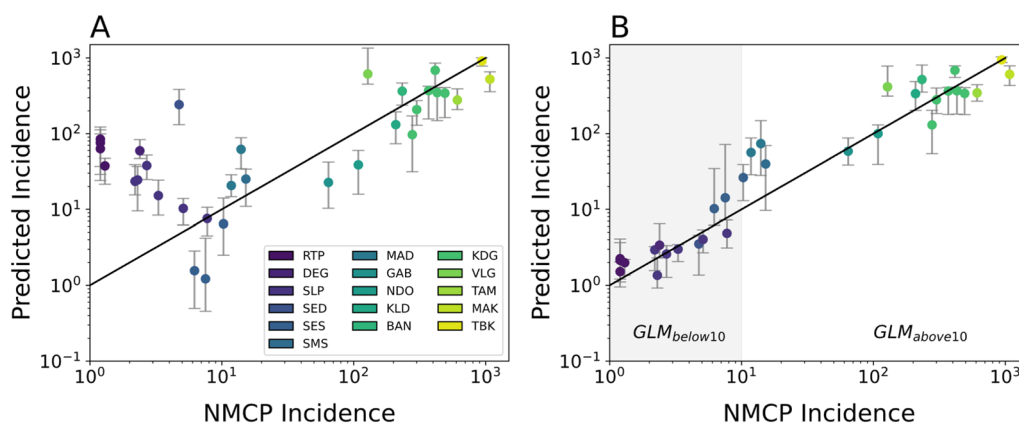
Fig. 3 Poisson mixed-effects GLM model predictions based on all five genetic metrics used in this study using **A** all the site-years and **B** from a piecewise GLM that splits the data in to low ($GLM_{below10}$, shaded in grey) and high transmission sites ($GLM_{above10}$). Each dot is the predicted incidence for all the examined site-years. Error bars represent the 95% confidence interval generated through iterative leave-one-out analyses where all the site-years associated with a chosen site were left out of the analysis. The legend applies to both **A** and **B**

Wong *et al. Malaria Journal*   (2024) 23:68

Page 8 of 13
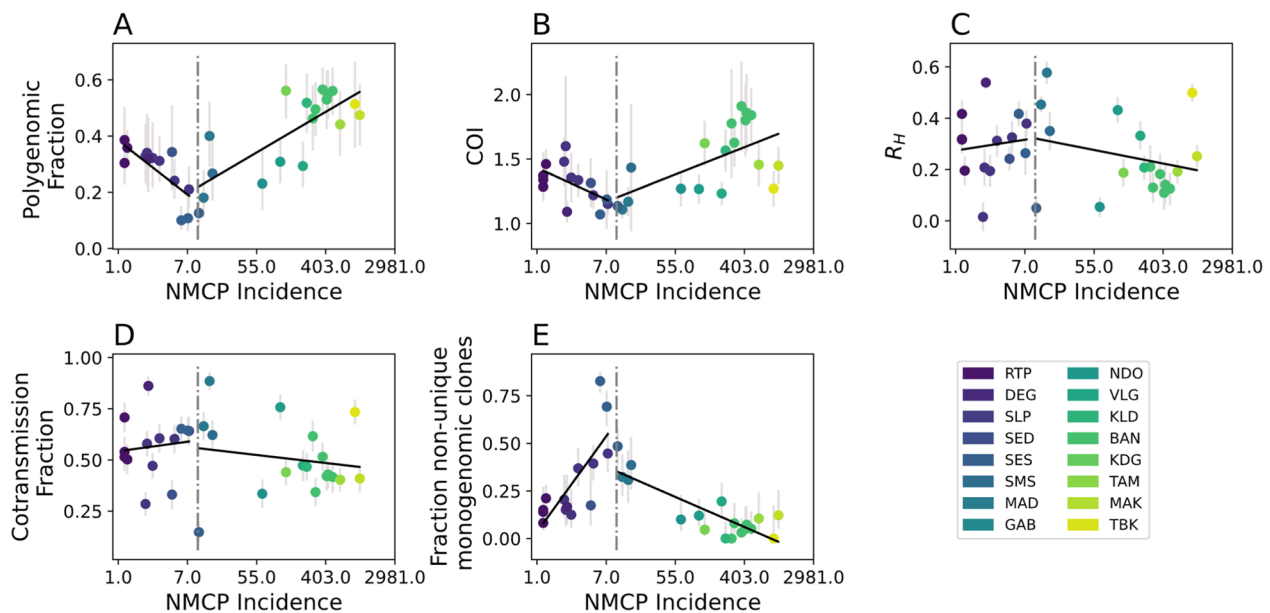
Fig. 4 Relationship between incidence and **A** polygenomic fraction, **B** THE REAL McCOIL COI, **C** R_H, **D** cotransmission fraction, and **E** the fraction of non-unique monogenomic clones. Error bars represent the 95% confidence interval for each examined site-year. The black lines represent model ordinary least squares regression model fits. A vertical dotted line is drawn at an NMCP-reported incidence of 10‰

original GLM (AIC = 3065.78) but tended to overestimate incidence when the reported values were close to the inflection point at 10‰ (Fig. 3B, Additional file 2: Fig. S4B).

### Parasite genetics can be used to estimate incidence in low to high transmission settings

Whether each genetic metric was sufficiently powered to estimate incidence by itself or whether there was a minimum subset of genetic metrics that could be used to accurately estimate incidence was evaluated. Because the piecewise GLM model strongly suggests that parasite genetics reflect different epidemiological dynamics when annual incidence < 10‰, whether the predictive power of these genetic epidemiology metrics was the same in regions with annual incidence < 10‰ and those with > 10‰ was also assessed. To answer these questions, a series of $GLM_{below10}$ and $GLM_{above10}$ models (collectively referred to as {$GLM_{below10}$} and {$GLM_{above10}$}) were trained with different combinations of genetic epidemiology metrics and their relative goodness-of-fit to the data evaluated using AIC (Fig. 5). The goal was to identify the set of genetic epidemiology metrics that resulted in the best-fitting model for both {$GLM_{below10}$} and {$GLM_{above10}$}.

For the {$GLM_{above10}$} models (Fig. 5A), the best fitting model was the one that included all five genetic metrics ($GLM_{above10}$, AIC = 1245.86). When examining each genetic metric individually, the polygenomic fraction

was the best predictor of incidence (AIC = 2342.57), followed by the fraction of non-unique monogenomic clones (AIC = 2624.32) (Additional file 2: Fig. S8). R_H and cotransmission were the worst individual predictors (AIC = 4501.90 and 4501.93 respectively) but including them in combination with other metrics improved model fit. In general, increasing the number of genetic metrics improved model fit. The average AIC for models with one, two, three, and four genetic metrics was 3560.88, 2858.25, 2178.32, and 1719.56 respectively.

For the {$GLM_{below10}$} models, there was no single, statistically best fitting model (Fig. 5B). The model with the lowest AIC was the one trained on the fraction of non-unique monogenomic clones alone (AIC = 51.05), followed by the one trained on polygenomic fraction alone (AIC = 51.16). However, based on standard definitions used to determine statistical significance (AIC difference > 2), it was not possible to determine which of the eight models with AIC < 53.05 was the best. On average, including additional genetic metrics did not improve model fit. The average AIC for models with one, two, three, and four genetic metrics was 55.68, 54.10, 54.44, and 55.86. The AIC for $GLM_{below10}$, the original model trained on all five genetic metrics, was 57.52.

Wong *et al. Malaria Journal*   (2024) 23:68

Page 9 of 13
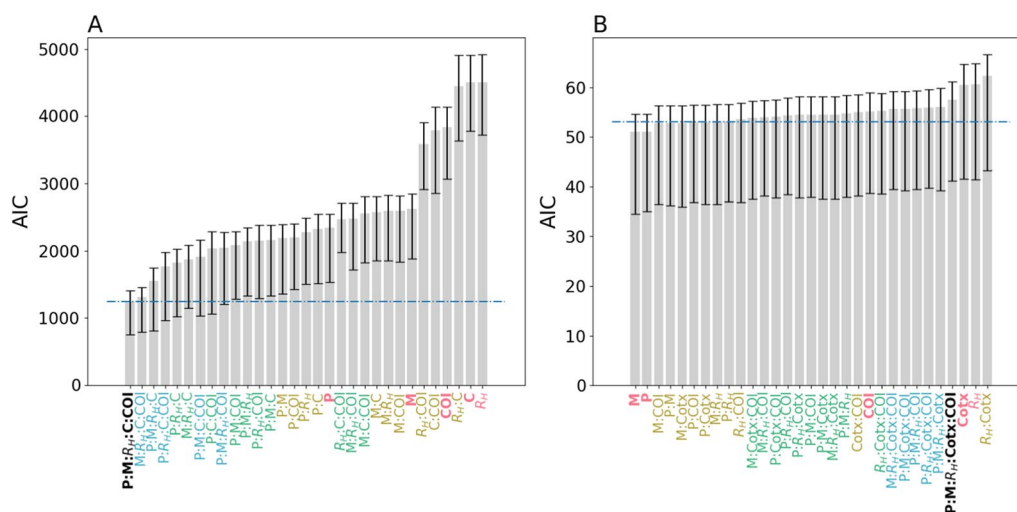
Fig. 5 AIC values for **A** GLM$_{above10}$ and **B** GLM$_{below10}$ models calibrated with different combinations of genetic epidemiology metrics. P refers to polygenomic fraction, Cotx refers to cotransmission fraction, COI refers to *THE REAL McCOIL* COI, and M refers to the fraction of non-unique monogenomic clones. For convenience, models trained with all five genetic epidemiology metrics are labelled in black, four in blue, three in green, two in yellow, and one in red. Error bars indicate the 95% confidence interval obtained from leave-one-out cross-validation. The dotted blue line indicates the line of statistical significance relative to the best fitting model. Bars whose average is above the dotted line are statistically worse than the best-fitting model. Bars whose average is below the dotted line perform equally well as the best-fitting model

Extrapolation of parasite genetic epidemiology based on trends seen in transmission settings with annual incidence > 10‰ overestimates transmission in transmission settings with annual incidence < 10‰

Based on these results, whether inappropriately applying the trends in parasite genetics observed in moderate-to-high transmission settings would cause erroneous incidence predictions in very low transmission settings was evaluated. To test this, GLM$_{above10}$ was used to generate incidence predictions for the sites with annual incidence < 10‰ (Additional file 2: Fig. S9). This approach reflects the current state of malaria genomic epidemiology analyses, where incidence is inferred based on the expectations stated in Table 2. The incidence values predicted by GLM$_{above10}$ were similar (Additional file 2: Fig. S9A) to those seen from the GLM that was trained on all the data (Fig. 3A) and settings with annual incidence < 10‰ were consistently overestimated.

Discussion

Parasite genetics has the potential to enable public health officials to evaluate changes in transmission in settings where the corresponding epidemiological data are either missing or difficult to collect. However, the utility of parasite genetic surveillance will depend on how informative genetics is for studying malaria transmission and whether the inclusion of genetics can enhance the confidence of estimates based on standard epidemiological measures of transmission. The major goals of this study were to (1) characterize the relationship between five malaria genetic epidemiology metrics that collectively assess the impact of superinfection, cotransmission, and clonal transmission and incidence to determine whether these relationships were constant across transmission strata, and (2) test the predictive power of five malaria genetic epidemiology metrics for inferring transmission intensity, which in this study was measured as the NMCP-reported incidence for the catchment health facility.

Senegal was an ideal setting for this analysis due to its extensive range of transmission intensities in a localized geographic region. By utilizing data collected across 16 health facilities located throughout the country, this study found that the relationship between parasite genetics and annual incidence changed in very low transmission settings with < 10‰. Based on these results, parasite genetics could be used to evaluate changes in incidence when the annual incidence is > 10‰ and used to assess potential sources of importation and other forms of transmission heterogeneity when transmission is low and falls below an annual incidence of 10‰.

When transmission is above an annual incidence > 10‰, the relationship between parasite genetics and reported incidence were consistent with previously established superinfection-based hypotheses that predict higher rates of multiple infections as transmission intensity increases [37–39]. Under these conditions, parasite genetics can be used to accurately infer incidence and increasing transmission intensity is associated with an increase in polygenomic fraction and COI and a decrease in the frequency of clonal parasites in the population.

Wong *et al. Malaria Journal*   (2024) 23:68

Page 10 of 13

These results suggest that NMCPs can utilize these correlations to quantify and compare the incidences of regions where transmission is high enough to be explained by superinfection, which in Senegal occurs when annual incidence > 10‰. However, accurately inferring NMCP-reported incidence in these moderate-to-high transmission areas required the incorporation of all five of the metrics used in this study, including those designed to measure cotransmission (R_H and cotransmission fraction). Thus, while superinfection is likely the dominant driver of parasite genetics when incidence > 10‰, the impact of cotransmission should not be discounted.

This study suggests that genomic epidemiological inference can be made with as few as 24 SNPs and a relatively small number of samples, which could be advantageous for assessing changing levels of transmission in high transmission settings with limited resources. These 24 SNPs can be genotyped from the genomic material extracted from discarded RDTs, which greatly reduces the technical and logistical complexities involved with collecting appropriate genetic material from clinical populations [25]. Collectively, this sampling strategy allowed us to develop a regression model to characterize and predict the NMCP-reported incidences when incidence > 10‰. In this study, the average number of samples per site year was 98.34, ranging from a minimum of 35 to 243 samples. These results suggest that a relatively small number of samples are needed to infer NMCP-reported incidence, but additional work is needed to assess the effect of sample size and sampling bias on genomic epidemiology inference more thoroughly.

However, the relationship between parasite genetics and incidence was not consistent across all transmission strata. When transmission falls below annual incidence < 10‰, many of the relationships between parasite genetics and incidence observed in higher transmission regions were reversed; increasing transmission intensity resulted in a decrease in polygenomic fraction and COI and an increase in the frequency of clonal parasites in the population. These results are difficult to explain under superinfection-based hypotheses, especially as the study sites with the lowest incidence, such as Richard Toll, had polygenomic fractions that were consistent with those seen in study sites with annual incidence > 400‰.

One possibility for the unusual trends in parasite genetics in very low transmission settings is that accurate quantification of the NMCP-reported incidence values is more difficult as transmission declines because infected individuals are infrequent and difficult to identify. Superinfection in very low transmission settings may also be more difficult to detect as parasite populations become more clonal and genetically related [40]. While the problems associated with sample ascertainment or measurement error in low transmission settings cannot be discounted, it is difficult to attribute these observations to sample ascertainment bias alone given the tight, but reversed, correlation observed between polygenomic fraction, the fraction of non-unique monogenomic clones, and COI in this transmission regime.

Instead, these changes could be driven by fundamental changes in transmission structure that affects the parasite genetics of very low transmission settings [41]. In Senegal, the reversed relationships between parasite genetics and transmission intensity could reflect the disproportionate impact of importation as local transmission declines. The 2013 Senegal census estimated that 14.6% of the population were internal lifetime migrants, meaning that their current area of residence differs from their birthplace [42]. The most popular destinations of internal lifetime migrants are in the low and very low transmission regions, such as Dakar, Diourbel, and Thies. Richard Toll also experiences seasonal influxes of migrant workers due to the presence of the Senegalese Sugar Company, and identical parasite clones were previously detected between Dakar and Richard Toll [42]. Anecdotal evidence obtained from the health facilities in the low transmission regions of this study suggest that patients with recent travel history were more likely to be tested and diagnosed with malaria. Regions with moderate to high levels of transmission, such as Kaolack and Kedougou, reported a net loss in population in the 2013 census. Thus, while it is possible to infer incidence from parasite genetics in the very low transmission settings of Senegal (Fig. 3B, $GLM_{below10}$), this is possibly due to the importation of parasites from the moderate- to high-transmission regions to the lower transmission regions of Senegal. Parasite genetics in very low transmission settings should be combined with data regarding travel history or other indicators of human movement [41, 43] to evaluate the potential impact of importation or focal transmission.

Overall, these results suggest that there are two distinct regimes where parasite genetics could be used to inform public health decision-making. When transmission is sufficiently high such that superinfection dominates, changes in parasite genetics can be used to infer incidence and quantify the transmission intensity in different regions. Parasite genetics could be especially valuable for evaluating the efficacy of public health interventions in reducing transmission in moderate to high transmission settings. However, when transmission falls below a certain threshold, these results suggest that parasite genetics should instead be used to begin evaluating the impact of importation or other heterogeneous transmission processes whose effects are masked by local mixing and transmission in high transmission settings but whose contributions are proportionally greater in low

Wong *et al. Malaria Journal*   (2024) 23:68

Page 11 of 13

transmission settings. The exact incidence threshold for distinguishing between these two paradigms is uncertain, but likely lies between an annual incidence of 10 and 100‰. Until additional study sites with annual incidence between 10 and 100‰ can be examined, the current WHO guidelines for defining sites with very low transmission sites (annual incidence < 100‰) could be used to determine when parasite genetics can be used to infer transmission intensity and when it can be used to study subtler effects associated with importation and other sources of transmission heterogeneity.

The careful examination of parasite genetics in very low transmission sites could help NMCPs address long standing issues regarding the role of importation and focal transmission in low transmission settings. Very low transmission sites should be identified prior to using parasite genetics, as inappropriately applying the trends in parasite genetics from higher transmission settings risks over-estimating incidence and ignores possible sources of infections in low transmission areas. This phenomenon is most clearly seen in the model predictions from the GLM trained on all the sites (Fig. 3A) and the out-of-sample extrapolations made with the GLM trained on only sites with annual incidence > 10‰ ($GLM_{above10}$, Additional file 2: Fig. S9). In practice, very low transmission sites can be distinguished from higher transmission sites using standard epidemiological metrics of incidence. The advantage of parasite genetics in very low transmission settings is that it could potentially allow NMCPs to identify source-sink populations or focal transmission sites that require targeted intervention for elimination. Additionally, parasite genetics in very low transmission sites could potentially be used to help countries confirm the absence of locally sustained transmission when applying for WHO certification of elimination.

Conclusion

In conclusion, this study evaluated the relationship between several parasite genetic epidemiology metrics and transmission intensity measured as annual incidence in Senegal. This study clearly shows that parasite genetic metrics behave differently under diverse transmission strata. When transmission is sufficiently high, changes in parasite genetics were consistent with different rates of superinfection and outcrossing as transmission intensity changes. However, when transmission intensity falls too low, changes in parasite genetics cannot be explained by current superinfection-based hypotheses; it is likely that changes in parasite genetics in very low transmission settings reflects importation or other forms of transmission heterogeneity. These results highlight the multidimensionality of parasite transmission and demonstrate the utility and limitations of parasite genetics for inferring

transmission intensity in Senegal. Future studies will continue to investigate the relationship between parasite genetics and other epidemiological metrics of transmission and importation as well as incorporate other genetic epidemiology metrics that were not explored in this study (*e.g.*, clonal barcode persistence [14, 44]).

Abbreviations

AIC	Akaike information criterion
COI	Complexity of infection
EIR	Entomological inoculation rate
GLM	Generalized linear model
$GLM_{above10}$	GLM trained on data with incidence > 10‰
$GLM_{below10}$	GLM trained on data with incidence < 10‰
MERS	Middle East Respiratory Syndrome
NMCP	National Malaria Control Programme
PCD	Passive case detection
PCR	Polymerase chain reaction
RDT	Rapid diagnostic test
SNP	Single nucleotide polymorphism
WHO	World Health Organization
ΔCT	Delta (difference) in cycle threshold

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12936-024-04897-z.

Additional file 1. Barcode data used in this study

Additional file 2. Additional tables and figures

Author contributions
WW led the analyses and generated all figures with support from SFS, JR, JS, AL, KEB, JLP, and CAB. WW, JT, and SKV wrote the main manuscript text. DN, DLH, DFW, and SKV supervised data collection and analysis. All other authors were involved with data collection. All authors reviewed and edited the manuscript. All authors read and approved the final manuscript.

Availability of data and materials
All the barcode data associated with this study are included in Additional file 1.

Declarations

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Immunology and Infectious Diseases, Harvard T. H. Chan School of Public Health, Boston, MA, USA. [2]Infectious Disease and Microbiome Program, The Broad Institute, Cambridge, MA, USA. [3]Malaria Branch, Division of Parasitic Diseases and Malaria, Global Health Center, Centers for Disease

Wong *et al. Malaria Journal* (2024) 23:68

Page 12 of 13

Control and Prevention, Atlanta, GA, USA. [4]Centre International de recherche, de formation en Genomique Appliquée et de Surveillance Sanitaire (CIGASS), Dakar, Senegal. [5]Section de Lutte Anti-Parasitaire (SLAP) Clinic, Thies, Senegal. [6]Programme National de Lutte contre le Paludisme (PNLP), Dakar, Senegal. [7]Institute for Disease Modeling at the 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.

References

1. Stockdale JE, Liu P, Colijn C. The potential of genomics for infectious disease forecasting. Nat Microbiol. 2022;7:1736–43.
2. Featherstone LA, Zhang JM, Vaughan TG, Duchene S. Epidemiological inference from pathogen genomes: a review of phylodynamic models and applications. Virus Evol. 2022;8:veac045.
3. Rife BD, Mavian C, Chen X, Ciccozzi M, Salemi M, Min J, et al. Phylodynamic applications in 21st century global infectious disease research. Glob Health Res Policy. 2017;2:13.
4. Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D, et al. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput Biol. 2014;10: e1003537.
5. Tegally H, Moir M, Everatt J, Giovanetti M, Scheepers C, Wilkinson E, et al. Emergence of SARS-CoV-2 omicron lineages BA.4 and BA.5 in South Africa. Nat Med. 2022;28:1785–90.
6. Campbell F, Archer B, Laurenson-Schafer H, Jinnai Y, Konings F, Batra N, et al. Increased transmissibility and global spread of SARS-CoV-2 variants of concern as at June 2021. Euro Surveill. 2021;26:2100509.
7. Holmes EC, Dudas G, Rambaut A, Andersen KG. The evolution of Ebola virus: insights from the 2013–2016 epidemic. Nature. 2016;538:193–200.
8. Dudas G, Carvalho LM, Bedford T, Tatem AJ, Baele G, Faria NR, et al. Virus genomes reveal factors that spread and sustained the Ebola epidemic. Nature. 2017;544:7650.
9. Dudas G, Carvalho LM, Rambaut A, Bedford T. MERS-CoV spillover at the camel-human interface. Elife. 2018;7: e31257.
10. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, et al. Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature. 2002;419:6906.
11. Arnot D. Unstable malaria in Sudan: the influence of the dry season: clone multiplicity of *Plasmodium falciparum* infections in individuals exposed to variable levels of disease transmission. Trans R Soc Trop Med Hyg. 1998;92:580–5.
12. Atuh NI, Anong DN, Jerome FC, Oriero E, Mohammed NI, D'Alessandro U, et al. High genetic complexity but low relatedness in *Plasmodium falciparum* infections from Western Savannah Highlands and coastal equatorial lowlands of Cameroon. Pathog Glob Health. 2022;116:428–37.
13. Wong W, Volkman S, Daniels R, Schaffner S, Sy M, Ndiaye YD, et al. RH: a genetic metric for measuring intrahost *Plasmodium falciparum* relatedness and distinguishing cotransmission from superinfection. PNAS Nexus. 2022;1:pgac187.
14. Daniels RF, Schaffner SF, Wenger EA, Proctor JL, Chang H-H, Wong W, et al. Modeling malaria genomics reveals transmission decline and rebound in Senegal. Proc Natl Acad Sci USA. 2015;112:7067–72.
15. 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 Genet. 2013;14:2.
16. Daniels R, Chang H-H, Séne PD, Park DC, Neafsey DE, Schaffner SF, et al. Genetic surveillance detects both clonal and epidemic transmission of malaria following enhanced intervention in Senegal. PLoS ONE. 2013;8: e60780.
17. Kelly-Hope LA, McKenzie FE. The multiplicity of malaria transmission: a review of entomological inoculation rate measurements and methods across sub-Saharan Africa. Malar J. 2009;8:19.
18. Gauthier C, Tibayrenc M. Population structure of malaria parasites: the driving epidemiological forces. Acta Trop. 2005;94:241–50.
19. Koepfli C, Mueller I. Malaria epidemiology at the clone level. Trends Parasitol. 2017;33:974.
20. Wong W, Griggs AD, Daniels RF, Schaffner SF, Ndiaye D, Bei AK, et al. Genetic relatedness analysis reveals the cotransmission of genetically related *Plasmodium falciparum* parasites in Thiès, Senegal. Genome Med. 2017;9:5.
21. Nkhoma SC, Banda RL, Khoswe S, Dzoole-Mwale TJ, Ward SA. Intra-host dynamics of co-infecting parasite genotypes in asymptomatic malaria patients. Infect Genet Evol. 2018;65:414–24.
22. Nkhoma SC, Nair S, Cheeseman IH, Rohr-Allegrini C, Singlam S, Nosten F, et al. Close kinship within multiple-genotype malaria parasite infections. Proc Biol Sci. 2012;279:2589–98.
23. 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;27:93-103.e4.
24. Daniels R, Volkman S, Milner D, Mahesh N, Neafsey D, Park D, et al. A general SNP-based molecular barcode for *Plasmodium falciparum* identification and tracking. Malar J. 2008;7:223.
25. Daniels RF, Schaffner SF, Dieye Y, Dieng G, Hainsworth M, Fall FB, et al. Genetic evidence for imported malaria and local transmission in Richard Toll, Senegal. Malar J. 2020;19:276.
26. Mharakurwa S, Daniels R, Scott A, Wirth DF, Thuma P, Volkman SK. Preamplification methods for tracking low-grade *Plasmodium falciparum* populations during scaled-up interventions in Southern Zambia. Malar J. 2014;13:89.
27. Chang H-H, Worby CJ, Yeka A, Nankabirwa J, Kamya MR, Staedke SG, et al. THE REAL McCOIL: a method for the concurrent estimation of the complexity of infection and SNP allele frequency for malaria parasites. PLoS Comput Biol. 2017;13: e1005348.
28. Programme National de lutte contre le Paludisme. Bulletin epidemiologique annuel 2021 du paludisme au Sénégal. 2021.
29. Programme National de lutte contre le Paludisme. Bulletin epidemiologique annuel 2018 du paludisme au Sénégal. 2019.
30. Programme National de lutte contre le Paludisme. Bulletin epidemiologique annuel 2019 du paludisme au Sénégal. 2020.
31. Programme National de lutte contre le Paludisme. Bulletin epidemiologique annuel 2017 du paludisme au Sénégal. 2018.
32. Programme National de lutte contre le Paludisme. Bulletin epidemiologique annuel 2016 du paludisme au Sénégal. 2017.
33. Programme National de lutte contre le Paludisme. Bulletin epidemiologique annuel 2015 du paludisme au Sénégal. 2016.
34. Programme National de lutte contre le Paludisme. Bulletin epidemiologique annuel 2014 du paludisme au Sénégal. 2015.
35. Ashton RA, Prosnitz D, Andrada A, Herrera S, Yé Y. Evaluating malaria programmes in moderate- and low-transmission settings: practical ways to generate robust evidence. Malar J. 2020;19:75.
36. WHO. Malaria surveillance, monitoring & evaluation: a reference manual. Geneva: World Health Organization; 2018.
37. Neafsey DE, Volkman SK. Malaria genomics in the era of eradication. Cold Spring Harb Perspect Med. 2017;7: a025544.
38. Neafsey DE, Taylor AR, MacInnis BL. Advances and opportunities in malaria population genomics. Nat Rev Genet. 2021;22:502–17.
39. Volkman SK, Neafsey DE, Schaffner SF, Park DJ, Wirth DF. Harnessing genomics and genome biology to understand malaria biology. Nat Rev Genet. 2012;13:315–28.
40. Schaffner SF, Badiane A, Khorgade A, Ndiop M, Gomis J, Wong W, et al. Malaria surveillance reveals parasite relatedness, signatures of selection, and correlates of transmission across Senegal. Nat Commun. 2023;14:7268.
41. 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. Sci Rep. 2020;10:3756.
42. National Agency of Statistics and Demography (Senegal). Senegal Population and Housing Census 2013. 2013.
43. Chang HH, Wesolowski A, Sinha I, Jacob CG, Mahmud A, Uddin D, et al. Mapping imported malaria in Bangladesh using parasite genetic and human mobility data. Elife. 2019;8: e43481.

Wong *et al. Malaria Journal*   (2024) 23:68

Page 13 of 13

44. Taylor AR, Schaffner SF, Cerqueira GC, Nkhoma SC, Anderson TJC, Sriprawat K, et al. Quantifying connectivity between local *Plasmodium falciparum* malaria parasite populations using identity by descent. PLoS Genet. 2017;13: e1007065.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.