Check for updates

RESEARCH ARTICLE

REVISED Geographic-genetic analysis of *Plasmodium falciparum* parasite populations from surveys of primary school children in Western Kenya [version 2; referees: 2 approved]

Irene Omedo [1], Polycarp Mogeni [1], Kirk Rockett[2], Alice Kamau [1], Christina Hubbart[2], Anna Jeffreys[2], Lynette Isabella Ochola-Oyier[1], Etienne P. de Villiers [1,3,4], Caroline W. Gitonga[5], Abdisalan M. Noor[4,5], Robert W. Snow [4,5], Dominic Kwiatkowski[2,6], Philip Bejon[1,7]

[1]KEMRI-Wellcome Trust Research Programme, Centre for Geographic Medicine Research-Coast, Kilifi, Kenya
[2]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, OX3 7BN, UK
[3]Department of Public Health, Pwani University, Kilifi, Kenya
[4]Centre for Tropical Medicine and Global Health, University of Oxford, Oxford, OX3 7LJ, UK
[5]Spatial Health Metrics Group, Kenya Medical Research Institute-Wellcome Trust Research Programme, Nairobi, Kenya
[6]Wellcome Trust Sanger Institute, Cambridge, CB10 1SA, UK
[7]Centre for Clinical Vaccinology and Tropical Medicine, Oxford, OX3 7LJ, UK

Abstract

Background. Malaria control, and finally malaria elimination, requires the identification and targeting of residual foci or hotspots of transmission. However, the level of parasite mixing within and between geographical locations is likely to impact the effectiveness and durability of control interventions and thus should be taken into consideration when developing control programs.

Methods. In order to determine the geographic-genetic patterns of *Plasmodium falciparum* parasite populations at a sub-national level in Kenya, we used the Sequenom platform to genotype 111 genome-wide distributed single nucleotide polymorphic (SNP) positions in 2486 isolates collected from children in 95 primary schools in western Kenya. We analysed these parasite genotypes for genetic structure using principal component analysis and assessed local and global clustering using statistical measures of spatial autocorrelation. We further examined the region for spatial barriers to parasite movement as well as directionality in the patterns of parasite movement.

Results. We found no evidence of population structure and little evidence of spatial autocorrelation of parasite genotypes (correlation coefficients <0.03 among parasite pairs in distance classes of 1km, 2km and 5km; p value<0.01). An analysis of the geographical distribution of allele frequencies showed weak evidence of variation in distribution of alleles, with clusters representing a higher than expected number of samples with the major allele being identified for 5 SNPs. Furthermore, we found no evidence of the existence of spatial barriers to parasite movement within the region, but observed directional movement of parasites among schools in two separate sections of the region studied.

Open Peer Review

Referee Status: ✔ ✔

	Invited Referees	
	1	**2**
REVISED **version 2** published 05 Sep 2017		✔ report
	↑	
version 1 published 20 Apr 2017	✔ report	✔ report

1 **Liwang Cui**, Pennsylvania State University, USA

2 **Lucy C. Okell** , Imperial College London, UK

Discuss this article

Comments (0)

Conclusions. Our findings illustrate a pattern of high parasite mixing within the study region. If this mixing is due to rapid gene flow, then "one-off" targeted interventions may not be currently effective at the sub-national scale in Western Kenya, due to the high parasite movement that is likely to lead to re-introduction of infection from surrounding regions. However repeated targeted interventions may reduce transmission in the surrounding regions.

KEMRI | Wellcome Trust

This article is included in the KEMRI | Wellcome Trust gateway.

Corresponding author: Irene Omedo (iomedo@kemri-wellcome.org)

How to cite this article: Omedo I, Mogeni P, Rockett K *et al.* **Geographic-genetic analysis of *Plasmodium falciparum* parasite populations from surveys of primary school children in Western Kenya [version 2; referees: 2 approved]** Wellcome Open Research 2017, **2**:29 (doi: 10.12688/wellcomeopenres.11228.2)

First published: 20 Apr 2017, **2**:29 (doi: 10.12688/wellcomeopenres.11228.1)

REVISED **Amendments from Version 1**

In response to the reviewer's comments, we have included a table showing the statistical tests performed on the data and have added a supplementary figure showing variation in *P. falciparum* parasite prevalence in schools across the study region. We have also added more detailed information on how spatial scan statistics, as implemented in SaTScan, were used to identify parasite clusters. As part of the recommendations for further work, we have suggested complementary analysis of human movement within the region based on mobile phone data, and have additionally recommended the need for separate analyses targeted at antigenic SNPs such as EBA175 and AMA1.

See referee reports

Introduction

Malaria incidence has markedly reduced in some parts of Africa[1–3]. In some instances, this has been associated with malaria control efforts[4], but has not been temporally associated with scaling-up malaria control in others[5]. In either case, transmission becomes more heterogeneous[2,6–12], leading to the emergence of hotspots of symptomatic and asymptomatic infections that may be targeted as part of a malaria control strategy[7,9,13–17]. In order to predict whether targeting hotspots is potentially an effective way of interrupting transmission, we need to understand the spatial and temporal scales over which parasite mixing can be observed. Limited genetic differentiation between malaria parasites has been shown to occur on a regional scale in sub-Saharan Africa[18–20], although we previously identified spatial structure at fine micro epidemiological scales within geographically defined regions in Kenya and The Gambia[21].

The level of parasite mixing is likely to impact the effectiveness of control interventions. In a recent randomized controlled trial of targeted integrated vector control in Rachuonyo south district in Western Kenya, an initial impact was seen within hotspot areas, but this did not reduce transmission outside the hotspot, and reductions within hotspots were not sustained[9]. This may have been due to rapid mixing of parasites from areas outside the intervention zones. Furthermore, declining malaria transmission is associated with increased risk of imported cases of infection and disease from high transmission to low transmission regions, hampering elimination efforts in the low transmission regions[22], and risking the spread of drug resistant malaria in higher transmission regions[23]. Thus, understanding parasite movement and gene flow will provide insights into novel, more targeted approaches to malaria elimination and combating the threats posed by re-introduction.

This study aimed to examine the geographic-genetic patterns of malaria parasite populations sampled at a sub-national level in Western Kenya, with a view of determining the extent of parasite mixing and genetic adaptation by the parasite population to its local environment. We genotyped 111 single nucleotide polymorphic (SNP) positions in *Plasmodium falciparum* isolates collected from children in 95 primary schools in two western Kenya provinces (Western and Nyanza) in order to analyse their genetic

relatedness. We then examined the parasite population structure based on principal component analysis (PCA), and used measures of local and global spatial autocorrelation to test for geographical relatedness among parasite genotypes. We further analysed the geographical distribution of allele frequencies to identify evidence of genetic adaptation of parasite populations to their local environment, and examined the region for spatial barriers to parasite movement, as well as for patterns in the direction of movement in either north/south or east/west directions.

Materials and methods

Study population

We sampled *P. falciparum* positive children from 95 primary schools in 20 districts in two provinces in Kenya (Western and Nyanza), located principally in the west of the country (Figure 1). Kenya has recorded a decline in malaria transmission in the past decade[2,24,25], and current country-wide malaria prevalence is estimated at 8%[26]. However, transmission is highly heterogeneous and the country can be divided into five malaria endemicity zones based on transmission intensity[26]. The highest malaria transmission intensity is currently experienced in western Kenya and is characterized by stable, endemic transmission along the lowlands, and unstable, epidemic transmission within the highlands[2,26,27], and despite scale up of interventions, malaria transmission has remained high, or even increased in certain parts of this region[28,29]. *P. falciparum* is the main causative agent of malaria, and is transmitted by different *Anopheles* mosquito species in different parts of this region[30,31].

Ethics statement

During initial sample collection, consent for participation in the surveys was based on passive, opt-out consent by parents rather than written, opt-in consent, due to the routine, low-risk nature of the surveys that were carried out under the mandate of the Ministry of Public Health and Sanitation to conduct disease surveillance[32]. Individual assent from the students was obtained before sample collection. Ethical approval for the current genotyping study was provided by Kenya Medical Research Institute (KEMRI) Ethical Review Committee (under SSC No. 2747) and study methods were carried out in accordance with approved guidelines.

Sample collection and DNA extraction

Finger prick blood spots were collected during a parasitological survey of primary school children across the country between October 2008 and March 2010[32]. An in-depth description of sample collection has been published previously[32,33]. Briefly, 480 primary schools were surveyed and 49,975 samples collected, with a maximum of 110 children being randomly surveyed in each school. Samples were collected by spotting 3 separate drops of 200µl finger prick blood onto Whatman filter papers. These samples were then air dried and stored, with desiccant, at 4°C. Additionally, each child was tested for *P. falciparum* parasite infection using rapid diagnostic tests. During sample collection, geospatial coordinates for each school were recorded. 2486 of the samples collected from 95 schools in western Kenya were found to be parasite positive. DNA was extracted from these parasite-positive samples.
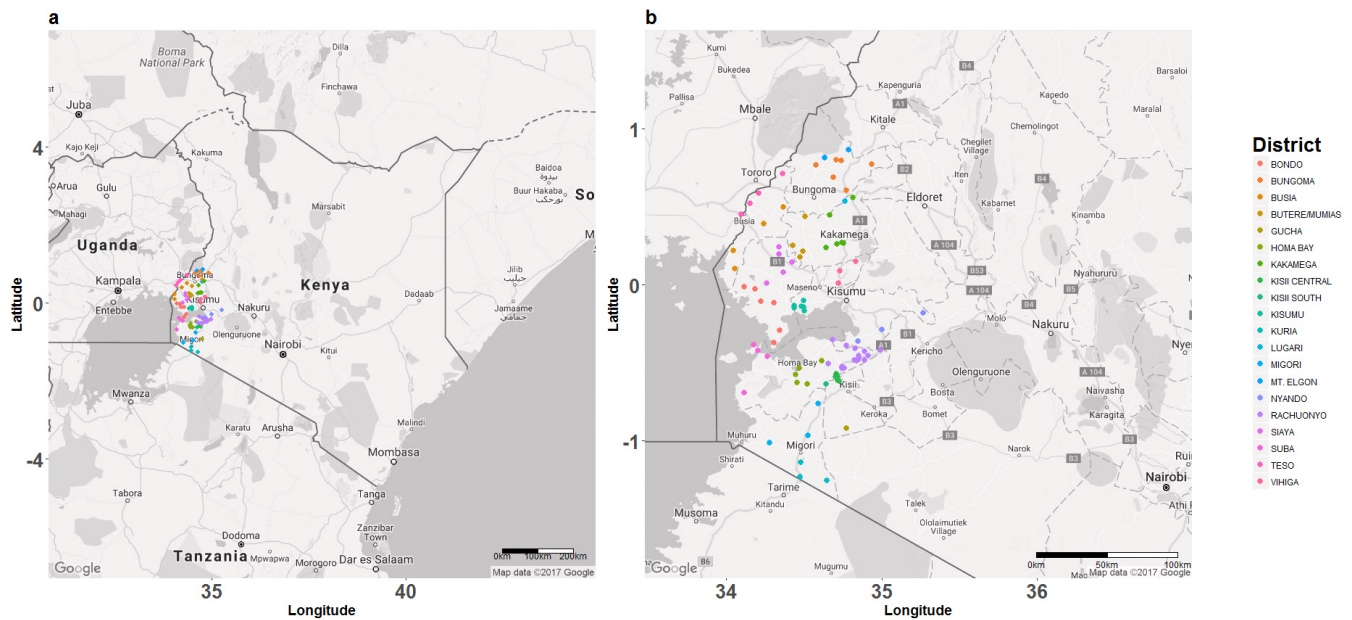
Figure 1. Spatial distribution of primary schools surveyed in a geographically defined region of western Kenya. 2486 *Plasmodium falciparum* positive samples were collected from children in 95 primary schools in 20 districts in this region. (**A**) Each dot represents an individual school, colour-coded by the administrative district in which it is located. (**B**) Map showing a close-up of the study region.

During DNA extraction, one of the blood spots was excised from each filter paper, cut into small pieces and placed into separate 1.5ml flip-top micro-centrifuge tubes (Eppendorf, Stevanage, UK). DNA was extracted using the QIAmp DNA investigator kit (Qiagen, UK), as per the manufacturer's instructions on the Qiagen BioRobot. Picogreen (Fisher Scientific-UK Ltd, Loghborough, UK) was used to determine DNA concentration in each sample.

SNP selection and genotyping

We typed 111 exonic SNP positions in 67 *P. falciparum* genes (Dataset 1 and Supplementary Table 1) in each of the parasite positive samples. These SNPs were a subset of those previously used to test the sensitivity and specificity of a customised Illumina GoldenGate genotyping platform and to identify a molecular barcode for distinguishing parasites from different geographical regions[19]. The 111 SNPs included those in well-known antigen-encoding genes, as well as housekeeping and hypothetical protein encoding genes. The SNPs were selected because they were genome-wide distributed, and polymorphic between at least two of three *P. falciparum* strains (3D7, HB3 and IT), which were selected because they are among the most studied and well characterised *P. falciparum* strains. Additionally, the SNPs had to be type-able on the genotyping platform (e.g. presence of a large enough conserved region around the SNP site that could be used to design locus specific primers for amplification). Genotyping was done on the Sequenom MassARRAY iPLEX platform[34]. This mass spectrometry-based genotyping platform allows multiplexing of up to 40 SNPs in a single reaction well and is suitable for typing 10s – 100s of SNPs in 100s – 1000s of samples. 3D7 (PlasmoDB release 24) was used

as a reference genome to design both locus specific PCR and iPLEX (SNP) extension primers using the Sequenom MassARRAY designer software (version 3.1). Locus-specific primers were pooled in a multiplexed PCR reaction and un-incorporated dNTPs were enzymatically dephosphorylated. The PCR products were then used in an iPLEX reaction where extension primers were bound immediately adjacent to target SNP sites and extended by a single nucleotide base into the SNP site using mass-modified dideoxynucleotides. Alleles were differentiated based on variations in their masses on the MALDI-TOF mass spectrometer.

Sample and SNP cut-off selection

To determine Sequenom genotyping success rates, we aggregated genotype data for individual samples and SNPs and applied pass/fail criteria to genotyping. Our positive control criteria were to include samples where at least 60% of SNP typing was successful and, among these, to include SNPs that were successfully typed in at least 60% of all samples. The selection criterion for successful typing was based on individually defined SNP intensity values (R) ranging from 0 to 1. SNPs with intensity values <0.1 were considered low quality and were categorized as failed and excluded from further analyses. In addition, allelic intensity ratios (θ) nearing 0 or 1 were used to classify SNP positions as homozygous, and intensity ratios of intermediate values were used to classify SNPs as heterozygous, representing mixed parasite populations in a single sample. Where mixed parasite populations were identified, we took the dominant genotype forward for further analysis, as represented by the majority SNP calls. Applying these inclusion criteria, we restricted our analyses to 83 SNPs and 1809 samples

collected from 88 schools in Western Kenya (Nyanza and Western provinces).

Statistical analysis

Several statistical analyses were carried out either using R statistical software (version 3.3.1)[35] or SaTScan software (version 9.3)[36] (Table 1).

Pairwise SNP and distance differences calculations. We computed parasite pairwise SNP and distance differences, comparing each parasite to every other parasite in the study population. For each parasite pair, we computed 1) the number of SNP differences at the 83 polymorphic positions analysed and 2) the distance between samples based on the geographical coordinates of the schools, taking 2.5km as an arbitrary distance between children attending the same school, assuming that on average schools were at least 5km apart, and taking 2.5km as the lower limit of detection of any two schools (Dataset 2). SNP differences between parasite pairs were then aggregated at the school level to determine the mean number of SNP differences among parasites per school. Nucleotide diversity (π) in the parasite population was computed as the average number of SNP differences per site between two parasites in all parasite pairwise comparisons. We further analysed how the number of SNP differences among parasites varied with distance between schools.

Analysis of parasite population genetics. Minor allele frequency distributions were computed for all 83 SNPs that had been successfully typed to determine the distribution of common and rare variants in the population. Population structuring was interrogated using principal component analysis (PCA) and computed based on singular value decomposition on a covariance matrix of pairwise SNP differences for all samples. SNPs that were included in the analysis, but were unsuccessfully typed in individual samples were replaced (in that sample) with the major allele in the population. Where there were mixed genotype infections in an individual sample, we took the major allele call to represent the dominant genotype at that position in that sample. PCA is a statistical technique widely used in analysis of genetic data where the number of genotypes interrogated is usually much higher than the number of samples analysed and is also useful when dealing with highly correlated data. The analysis transforms the original variables into new sets of variables that are linear combinations of the original variables, are uncorrelated and ordered based on the amount of variation in the original data that they explain. The first principal component explains most of the variation in the data, and subsequent components sequentially explain as much of the remaining variation as possible[37]. Scores (values of the transformed variables that correspond to a specific data point) representing individual parasite genotypes were computed for the first 3 principal components. Since geospatial positioning information was collected at the school level and samples from the same school were assigned the same geographical coordinates, principal component (PC) scores were later aggregated at the school level, and the mean PC score for each school plotted on a map of the study region.

Spatial autocorrelation analysis. To test the hypothesis that *P. falciparum* genetic structure has a spatial distribution component, we carried out an analysis of both global (Moran's *I*)[38] and local (scan statistics)[36] spatial autocorrelation among parasite genotypes. Moran's *I* simultaneously measures the correlation between feature attributes (parasite genotypes, represented here by the scores of the first 3 PCs) and locations (geospatial positioning of schools) to determine whether feature attributes are clustered, dispersed or randomly distributed in space. For each PC, Moran's *I* correlation coefficients were computed for parasite pairs falling within 3 increasing distance classes of 1km, 2km and 5km. We used 100 bootstrap replicates to determine the statistical significance of the Moran's *I* correlations observed for parasite pairs within each distance class.

Spatial scan statistics to detect statistically significant local clustering of genetically related parasites was carried out in

Table 1. Statistical tests carried out on *P. falciparum* parasites collected from primary school children in Western Kenya.

Statistical analysis	Function
Pairwise SNP and distance differences calculation	Determining the number of SNP differences and distance differences between parasite pairs in the dataset.
Minor allele frequency distribution	Analysis of the frequency and distribution of the minor alleles in the population.
Principal component analysis	Detecting *P. falciparum* population structure at the sub-national level.
Moran's *I* spatial autocorrelation analysis	Analysis of global spatial autocorrelation among parasite pairs at different geographical scales.
Spatial scan statistics	Identification of local, spatial clusters of parasite sub-populations.
Logistic regression	Analysis of the spatial pattern of distribution of allele frequencies among individual schools.
Bernoulli regression	Identification of spatial clusters of schools with similar frequencies at specific SNP loci.
Raster analyses by pixels	Detecting geographical regions that act as spatial barriers to parasite movement; Moran's *I* analysis was then used to determine if pixels acting as barriers to parasite movement were spatially autocorrelated.
Bearing regression analysis	Analysing directionality in parasite movement within the region.

SaTScan software (version 9.3)[36]. We carried out this analysis separately for genotypes represented by each of the first three PCs. For each PC, scores representing individual parasite genotypes were imported into SaTScan, together with the spatial coordinates of sample collection. We used the latitude/longitude coordinates to specify feature (sample) locations and the PC scores to represent feature attributes (parasite genotypes), and ran a purely spatial, retrospective analysis based on a normal probability distribution model implemented in SaTScan software[39]. We identified geographical regions with clusters containing parasite genotypes associated with high PC scores. In the normal probability distribution model, each observation or sample is associated with a single negative or positive continuous attribute (PC score) and the model uses a likelihood function based on the normal distribution. The spatial scan statistics employed here use a circular scanning window that is flexible both in location and size, with a radius that varies continuously from zero (including only a single sample location/school) to an upper limit set by the user (in this case 50% of the sample locations/schools). For each window location and size, a ratio of observed to expected PC scores was computed for samples found inside and outside the window. Clusters with high ratios were then noted and their statistical significance determined after accounting for multiple comparisons using random permutations. Shapefiles containing the spatial coordinates of statistically significant clusters were then imported into R and plotted on a map of western Kenya in order to identify locations of schools with genetically distinct parasite sub-populations.

Spatial distribution of allele frequencies. We used a logistic regression model to examine geographic variations in allele frequencies. Each school was included in the model as a categorical independent variable, with the binary (1/0) outcome set as the presence or absence of a specific SNP in parasites within that school. We compared this model to a null model in a likelihood ratio test to test the goodness of fit of the model containing the SNPs and to identify those SNPs that showed statistically significant ($p < 0.05$) variations in frequency. To keep from inflating allele frequencies, SNPs that failed genotyping in individual samples were excluded from analysis. For those SNPs that showed a significant variation in frequency between schools based on the logistic regression model, we computed the frequency of each SNP per school and plotted these on a map of the region to visualize the distribution pattern of SNP frequencies in schools within the region.

To examine whether there were geographical clusters of schools with similar allele frequencies, we ran the Bernoulli model in SaTScan, including only SNPs that were statistically significant in the logistic regression model. At each SNP position, samples were coded with a 1 if they contained the major allele, and 0 if they contained the minor allele. The Bernoulli model analyses the distribution of cases (major allele) and controls (minor allele) and tests the null hypothesis that the distribution of cases and controls is random within the geographical area. The spatial scan statistics based on this model involves scanning the geographical space for regions with higher than expected number of cases. The statistical

significance of identified clusters were computed as in the normal probability model above.

Analysis of spatial barriers to parasite movement. To test for the presence of spatial barriers to parasite movement and mixing within the region, we used raster analysis implemented in R statistical software to divide the study area into 192 10km-by-10km grids/pixels. Each pixel was then scored with either 1 or 0 depending on whether or not a straight line linking a school pair crossed the boundaries of that pixel. This was done for all 192 pixels and all school pairs in the region. These scores were then included in a multivariable linear regression analysis to test how the presence of a specific pixel affected the nucleotide diversity of parasites in schools separated by that pixel. To determine the statistical significance of observed differences in nucleotide diversity, the analysis was bootstrapped based on 10,000 resampling steps. This bootstrap value was chosen to allow us to obtain more precise estimates. We included the coefficient estimates derived from the pixel analysis in a Moran's I analysis to determine whether pixels acting as barriers to or gateways for parasite movement were spatially auto-correlated. Moran's I analysis was computed for schools falling within a 10km distance class, representing the same spatial extent as that used to generate the pixels. We used 100 bootstrap resampling steps to determine statistical significance in the Moran's I analysis.

We further carried out a regression analysis based on bearing to examine the parasite population for patterns of directional movement, either in the north/south or east/west directions. For each 10km-by-10km grid, school pairs that crossed its boundaries (scored as 1 in the previous pixel analysis) were selected. Each pair of schools was then individually coded as 1 if the absolute difference in latitude between them was greater than the absolute difference in longitude, indicating a north/south direction of movement, and coded as 0 if the reverse was true, indicating a west/east direction of movement. These new variables were then included in a multivariable linear regression analysis to test the effect of north/south or east/west directional movement on parasite genetic diversity. To determine the statistical significance of each pixel in acting as a force in directional movement, we ran a bootstrap analysis with 10,000 resampling steps. We then carried out Moran's I analysis, using coefficient estimates derived from the bearings' regression analysis as feature attributes, to examine the region for spatially auto-correlated directional movement. Moran's I analysis was once again carried out for schools falling within 10km distance classes, and with 100 resampling steps to determine statistical significance.

Results
Sequenom assay performance
We genotyped 111 genome-wide distributed exonic SNPs in 2486 *P. falciparum* positive samples collected from 95 primary schools in Western Kenya (Figure 1 and Figure 2). 83 of the 111 SNPs were successfully typed in 1809 samples from 88 schools in Western Kenya (1097 from Nyanza province and 712 from Western province). Subsequent analyses were carried out on these 83 SNPs and 1809 samples. Variation in parasite prevalence was observed
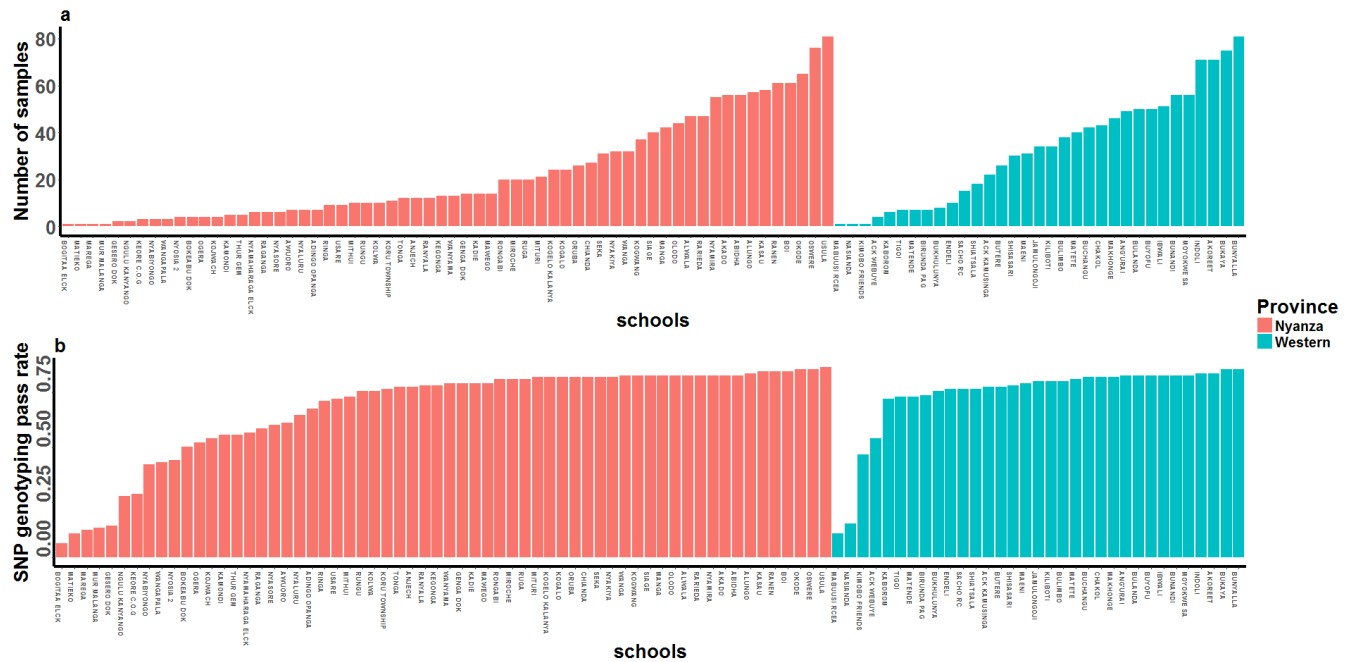
Figure 2. Distribution of *Plasmodium falciparum* positive samples and their associated genotyping success rates. (**A**) 2486 samples were collected from 95 schools (33 in western province and 62 in Nyanza province) in western Kenya. The total number of samples varied from 1 to 81 per school. (**B**) 111 single nucleotide polymorphic (SNP) positions were genotyped in all parasite positive samples. Mean genotyping success rates per school ranged from 6–86%.

in the region, with areas north and west of Lake Victoria having higher infection prevalence than areas south and east of the lake (Supplementary figure 1).

P. falciparum population genetics
Analysis of the minor allele frequency distribution showed that most of the genotyped SNPs were present at medium to high frequencies in the parasite population, with 51 of the 83 successfully typed SNPs having minor allele frequencies of 5% or higher. We also observed a high level of within-population genetic diversity in the parasite population, with an average nucleotide diversity (π) of 0.184 per SNP site.

Furthermore, using PCA, we found that the first three PCs cumulatively accounted for only 10.78% of the variation in the genotype data (PC1=3.74%, PC2=3.54%, PC3=3.5%), indicating high diversity among the parasites. At the sub-national level, we were unable to resolve the parasite population into distinct sub-populations based on PCA, and there was no difference in population structure in lower versus higher transmission intensity areas (Figure 3).

Spatial autocorrelation analysis
To examine structure to the PC scores, we analysed both local (spatial scan statistics) and global (Moran's *I*) spatial autocorrelation analysis. Spatial autocorrelation measures the extent to which geographical features and their associated data values are clustered, dispersed or randomly distributed in space. Moran's *I* analysis showed no statistically significant trends of spatial autocorrelation among parasite pairs that were close to each other in any of the three distance classes (1km, 2km and 5km) analysed. We found significant autocorrelations (p<0.01) among parasites that were on average at least 20km apart in space (Figure 4), but these autocorrelations were associated with very low correlation coefficients (< 0.03) and were not consistently seen in adjacent distance categories. Thus, the overall pattern seen from this analysis was that of little or no spatial autocorrelation in genotypes, even among parasite pairs that were very close to each other in space.

However, we identified one statistically significant (p=0.001) cluster based on PC2 when we analysed the data for local geographical clustering of distinct parasite genotypes using spatial scan statistics (Figure 5). This cluster was relatively large, with a radius of 67.84km, and included 852 of the 1809 samples. We identified no significant clusters when we analysed the first and third PCs.

Allele frequency distribution
We used a logistic regression model to examine the distribution of allele frequencies at each SNP position using log likelihood ratio testing for the effect of school. We identified 18 out of
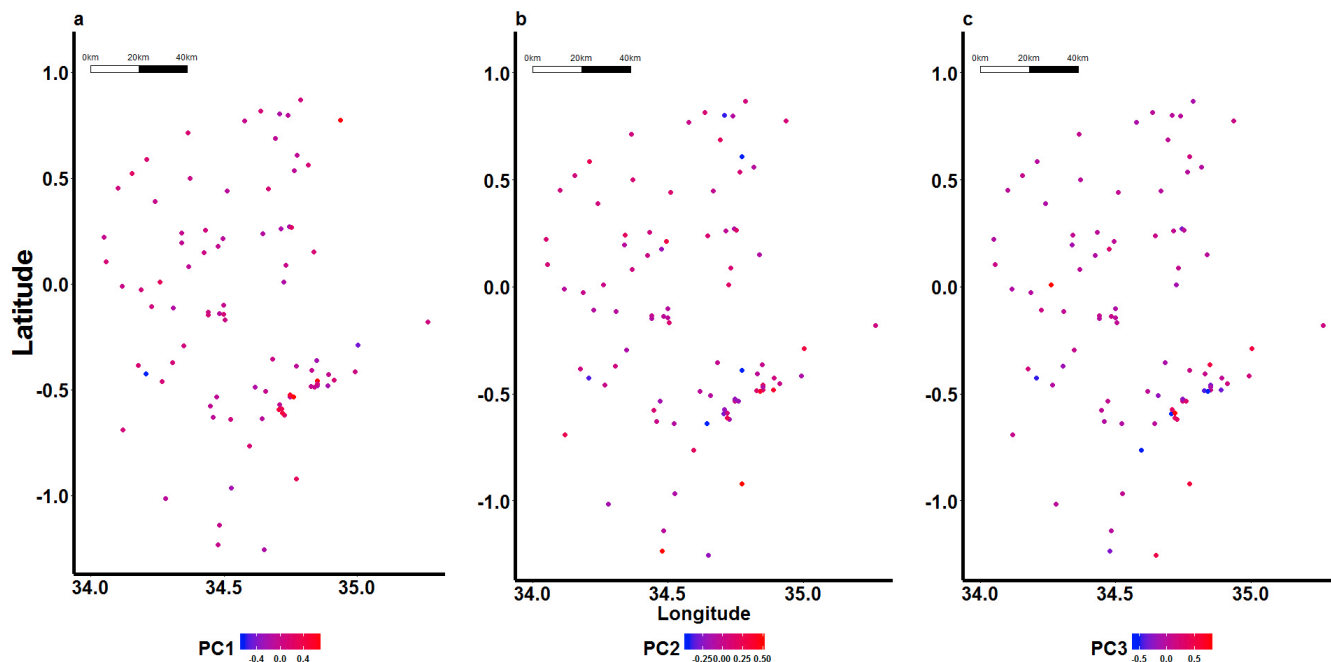
Figure 3. Spatial distribution of scores for the first 3 principal components (PCs) representing parasite genotypes. Geospatial positioning information was collected at the school level, thus PC scores (values of the transformed variables corresponding to a specific data point) were aggregated for all samples in an individual school. Here each dot represents a school, and has been colour-coded based on the mean genotype score of all parasite isolates collected in that school. Cumulatively, the first three PCs accounted for only 10.78% of the variation observed in the genotype data (PC1=3.74%, PC2=3.54%, PC3=3.5%).

83 SNPs that had statistically significant ($p < 0.05$) variations in frequencies among schools, although none of the SNPs were significant after Bonferroni adjustment for multiple testing.

We included these 18 SNPs in a spatial scan statistics analysis using the Bernoulli probability regression model and ran a purely spatial analysis to determine the geographic pattern of allele frequency distribution within the parasite population. For each SNP position, cases were represented by the major allele in the population while controls were represented by the minor allele. 5 of the 18 SNPs produced statistically significant geographical clusters containing schools with a higher than expected number of samples with the major allele (Supplementary Table 2).

Spatial variations in genetic differences between *P. falciparum* parasites

Using a linear regression model to examine the effect of distance on parasite genetic relatedness, we found that the number of

SNP differences between parasite pairs was positively correlated with distance between the parasites (effect size = 1.85×10^{-3}) (Figure 6). However, bootstrapping the analysis (to take into account the linked nature of pairwise observations) gave no statistically significant effects of distance on genetic relatedness ($p=0.347$; 95% CI = -0.012 – 0.017). These results provide no evidence for genetic isolation by distance in this parasite population.

Spatial barriers to parasite movement and mixing

We carried out raster analysis using 192 pixels to examine the study area for spatial barriers to parasite movement. Most of the pixels were found to have a non-significant influence on the number of SNP differences among parasites, and none were significant after correcting for multiple testing using the Bonferroni correction method (Figure 7a). Furthermore, a histogram of p values showed a null (uniform) distribution (Figure 7b), and an analysis of spatial relationships among pixels based on coefficient estimates derived from the pixel regression showed no evidence of autocorrelation
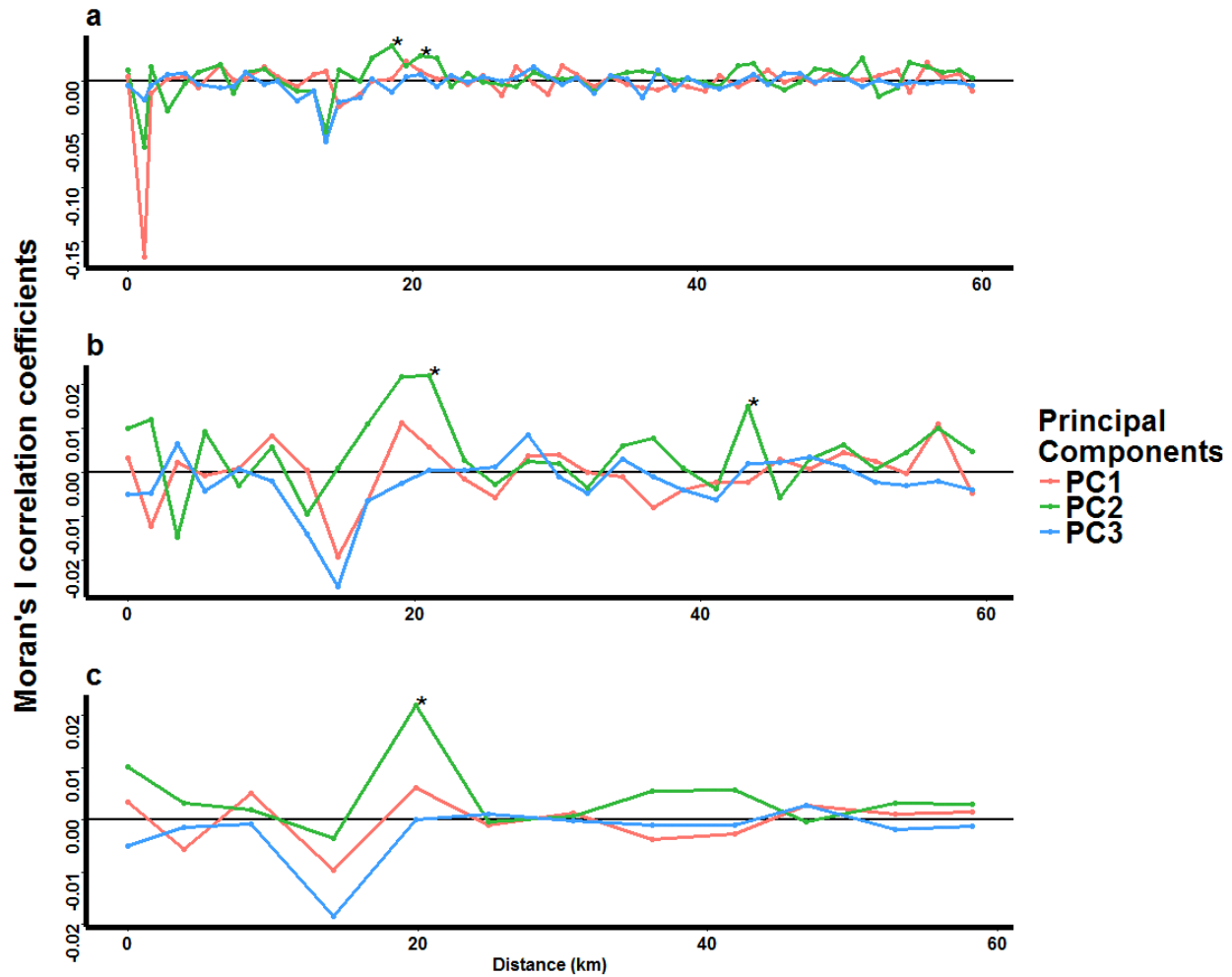
Figure 4. Moran's *I* correlation coefficients describing the spatial autocorrelation of genotypes of *Plasmodium falciparum* parasite pairs. Spatial autocorrelation was tested separately for parasites grouped into three distance classes of **a**) 1km, **b**) 2km and **c**) 5km. Within each distance class, correlations were computed for each of the first 3 principal components (PCs). The asterisks represent those distances at which statistically significant (p<0.01) correlation coefficients were found for parasite pairs within each distance class, indicative of possible clustering of specific parasite genotypes.

among pixels acting as either barriers to or gateways for parasite movement (Figure 7c).

We further carried out raster analysis by pixels to examine the bearing (direction of movement) of parasites in either the east/west or north/south directions. Individually, most of the 192 pixels were not significant factors in determining directional movement of parasites over the region (Figure 8a). However, some of the pixels were statistically significant (p<0.0003) in representing regions with greater east/west movement, even after accounting for multiple testing. When we included the regression coefficient

estimates derived from analysis of bearing in a Moran's *I* analysis to examine the parasite population for spatially auto-correlated direction of movement, we found evidence of statistically significant (p<0.01) autocorrelation for school pairs that were separated by up to 40km (Figure 8c).

Additionally, we identified two separate clusters of pixels within the region that showed patterns of specific directional movement, one in the north east (indicative of greater north/south movement) and another in the west (indicative of greater east/west movement) (Figure 9).
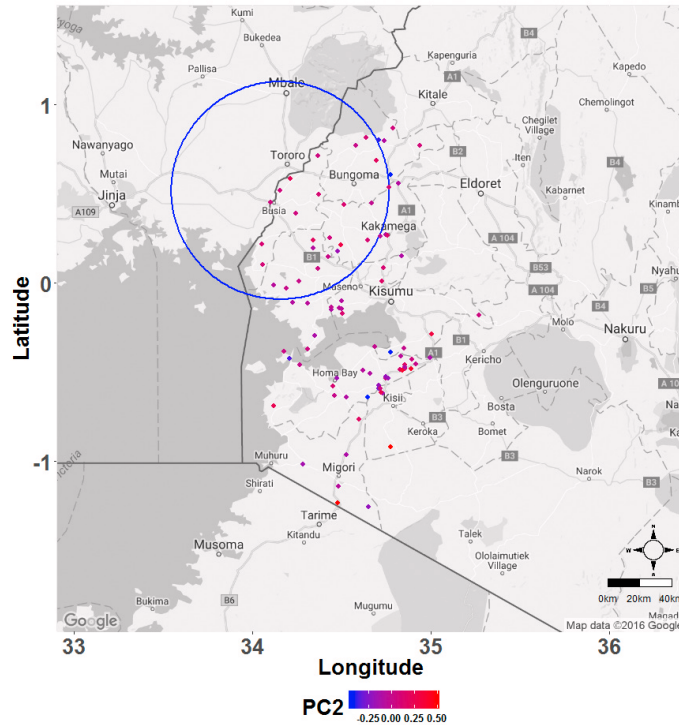
Figure 5. Spatial scan statistics to identify local spatially autocorrelated clusters of genetically distinct *Plasmodium falciparum* parasite sub-populations in western Kenya. Spatial scan statistics employing the use of multiple circular windows of varying sizes (ranging from covering only 1 sample up to 50% of the sample population) around samples geographically defined regions was used to compute the ratio between expected and observed number of genotypes within each window. Each window with higher than expected number of similar genotypes was noted down as a cluster, and its statistical significance determined after accounting for the multiple comparisons. Genotypes for individual parasites were assigned based on scores of the first 3 principal components. Here, each school is colour-coded based on the mean principal component score for all parasite genotypes found within it. One cluster of highly related parasite genotypes (blue circle) was identified when analysing the second principle component.
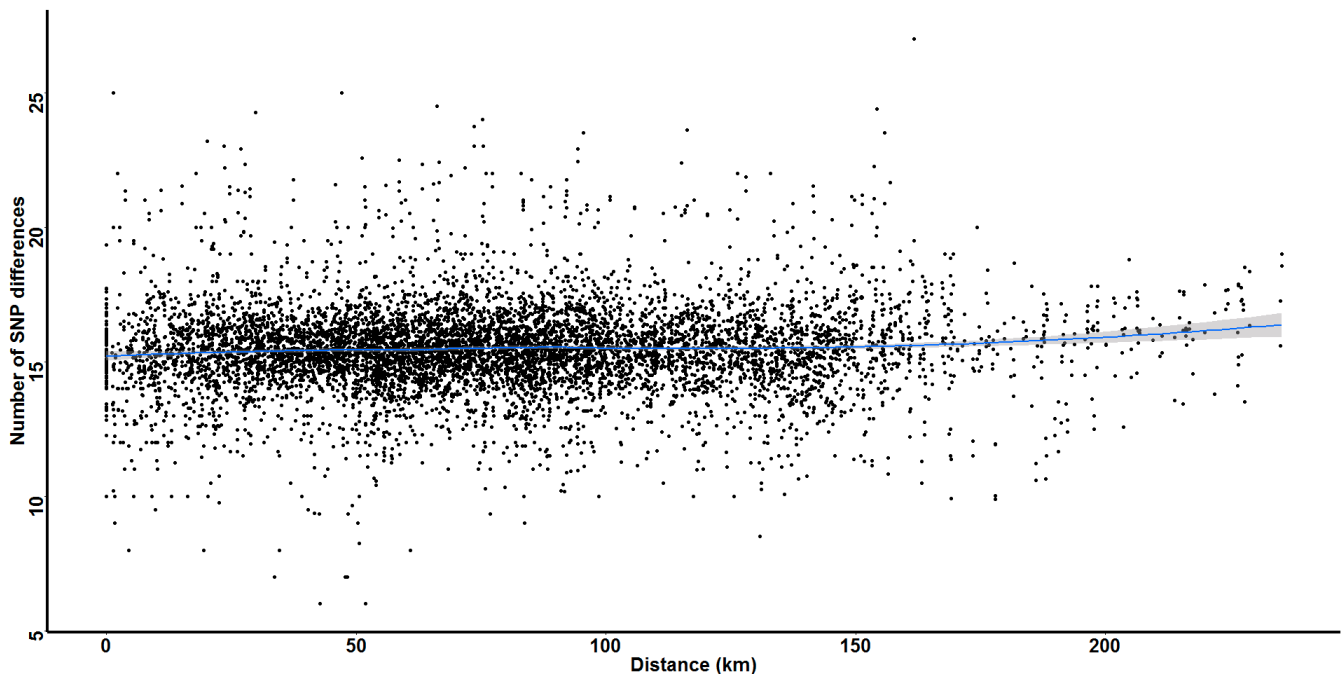
Figure 6. Variation in *Plasmodium falciparum* parasites' genetic diversity over distance. Genetic diversity was defined as the average number of single nucleotide polymorphism (SNP) differences between parasites in each pairwise school comparison, and was plotted against the distance between the corresponding school pair. The blue line represents loss-fitted smoothing with 95% confidence intervals (grey area).
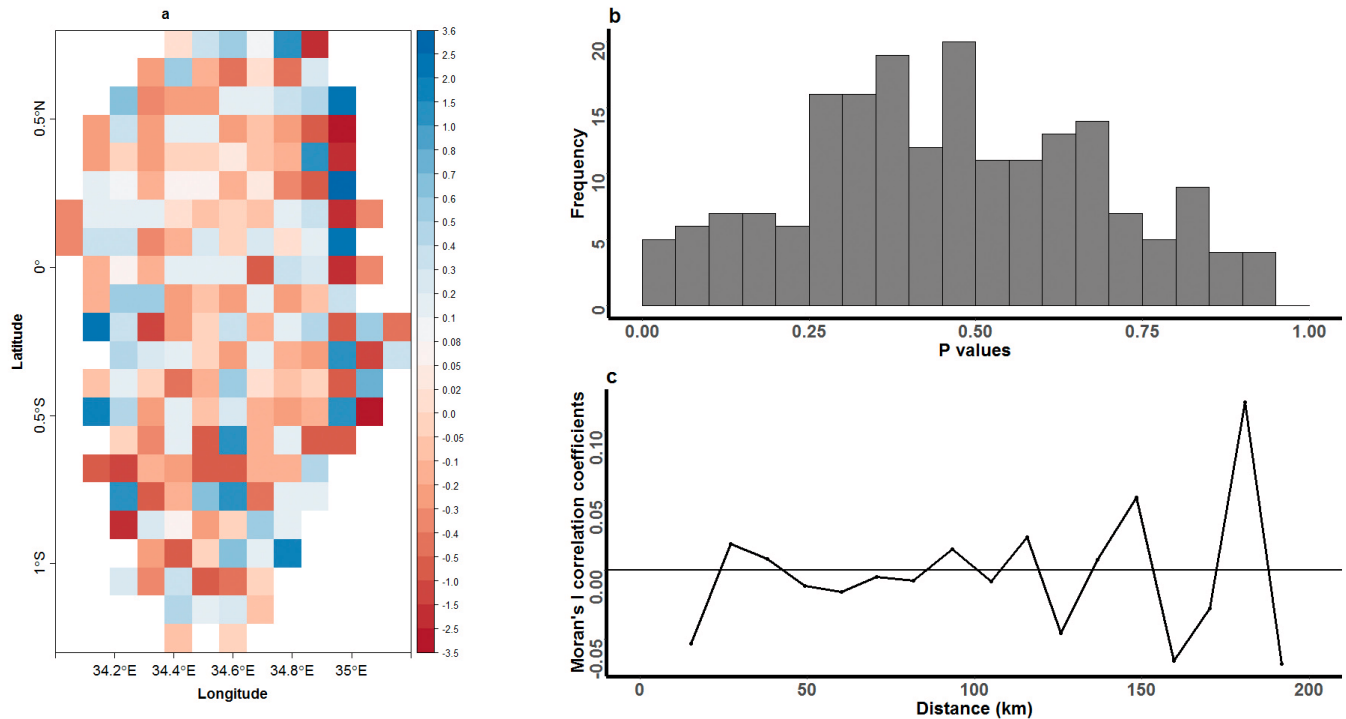
Figure 7. Raster analysis by pixels to examine the presence of spatial barriers to *Plasmodium falciparum* movement in a geographically defined region of western Kenya. (A) Each pixel represents a 10km-by-10km area of the region, and is colour-coded based on the coefficient estimates derived from a linear regression analysis that was used to test the impact of each pixel in acting as either a barrier (blue pixels) or gateway (red pixels) to parasite movement with the region. No pixels were significant barriers or gateways to parasite movement after Bonferroni correction to account for multiple comparisons. **(B)** Distribution of p values observed after bootstrapping the regression analysis (with 10,000 resampling steps) to determine the level of significance of pixels in acting as barriers to parasite movement. **(C)** Moran's *I* analysis describing the spatial autocorrelation between geographical locations of pixels and their associated coefficient estimates. Autocorrelation was calculated for parasites grouped in 10km distance bands, and the analysis was bootstrapped 100 times to determine significance.

Discussion

We have previously used SNP genotype data to examine the level of genetic relatedness among *P. falciparum* parasites on a micro-epidemiological scale within three regions with varying transmission intensities in Kenya and the Gambia, and found evidence of spatial sub-structure over short distances (i.e. <10km), despite a high level of parasite mixing[21]. In the present analysis, we examined the level of parasite mixing at a sub-national scale in Western Kenya, using parasitological data from primary school surveys to describe the patterns of parasite mixing at a larger geographical scale.

We selected primary school children as the study population because they are easy to sample. Furthermore, infection diversity peaks at 3 – 14 years and then declines in older age groups in high
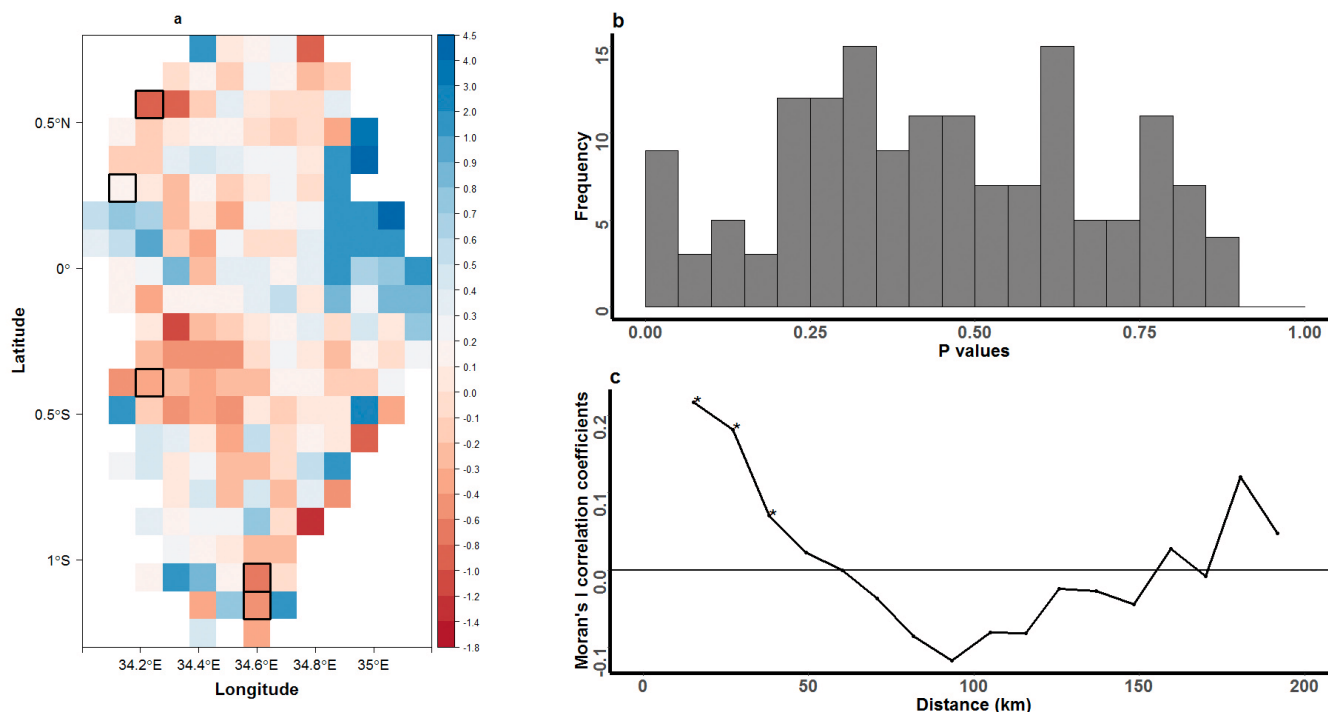
Figure 8. Raster analysis by pixels to examine patterns of north/south versus east/west directional movement of *Plasmodium falciparum* parasites in western Kenya. (**A**) Each pixel represents a 10km-by-10km area of the region, and is colour-coded based on coefficient estimates describing the effect size of each pixel in influencing directional movement. Pixels that were statistically significant after correcting for multiple testing are highlighted with black borders. Grids were colour-coded to represent east/west (red) or north/south (blue) movement. (**B**) Distribution of p values observed after bootstrapping the regression analysis (with 10,000 resampling steps) to determine the level of significance of pixels in influencing parasite directional movement. (**C**) Moran's *I* analysis to describe the spatial autocorrelation of movement within the region. The analysis was computed using geographical coordinates of individual pixels to represent feature locations and coefficient estimates derived from the bearing regression analysis to represent the associated feature values. Autocorrelation was computed for parasites grouped in 10km distance bands. Significant positive correlation coefficients (p<0.01; marked by asterisks) were observed for schools separated by up to 40 km within the 10km distance bands.

transmission settings[40–42]; hence our study sample is likely to contain a diverse genetic pool representative of parasites circulating in the region. Sampling only asymptomatic infections in schools may not give the whole range of genetic diversity within the region, as one study identified specific polymorphisms in AMA1 that could have been more frequent in symptomatic infections compared with asymptomatic infections[43]. Young children with symptomatic infections would be absent from school and away from the sampling frame. However, the sampling strategy was consistent across the different schools and the evidence of genomic variation in parasites according to clinical outcome is limited.

We found evidence of high genetic diversity in the Western Kenya parasite population, consistent with the high malaria transmission intensity experienced in this region[2,26,27,44]. Of the five malaria transmission zones in Kenya, Western Kenya currently experiences the highest transmission intensity[24,26], despite efforts to scale up various control interventions, such as long lasting insecticide nets, indoor residual spraying and artemisinin combination therapy, in this region[45–47].

Using PCA, we did not identify any genetic structure through inspection of the PC plots derived from SNP genotype data. This indicates an absence of discrete sub-populations within this *P. falciparum* parasite population, and is in agreement with our previous analysis of parasites from the same region[21], and with whole genome data from different African countries[18,20]. In South-East Asia, distinct sub-populations associated with antimalarial drug resistance have been described[48]. Previous analyses of *P. falciparum* population structure in western Kenya have also shown high genetic diversity and little population differentiation in this parasite population[49–51].

In contrast with a previous study of ours[21], analysis of trends in spatial relationships among parasite genotypes identified no significant autocorrelation using Moran's *I* spatial autocorrelation analysis. Overall, the consistent pattern observed across all distance classes was that of no autocorrelation among parasites in schools at all distances, with occasional inconsistent associations that we considered likely to be spurious. Using the spatial scan statistics, we identified only a single cluster of genetically related
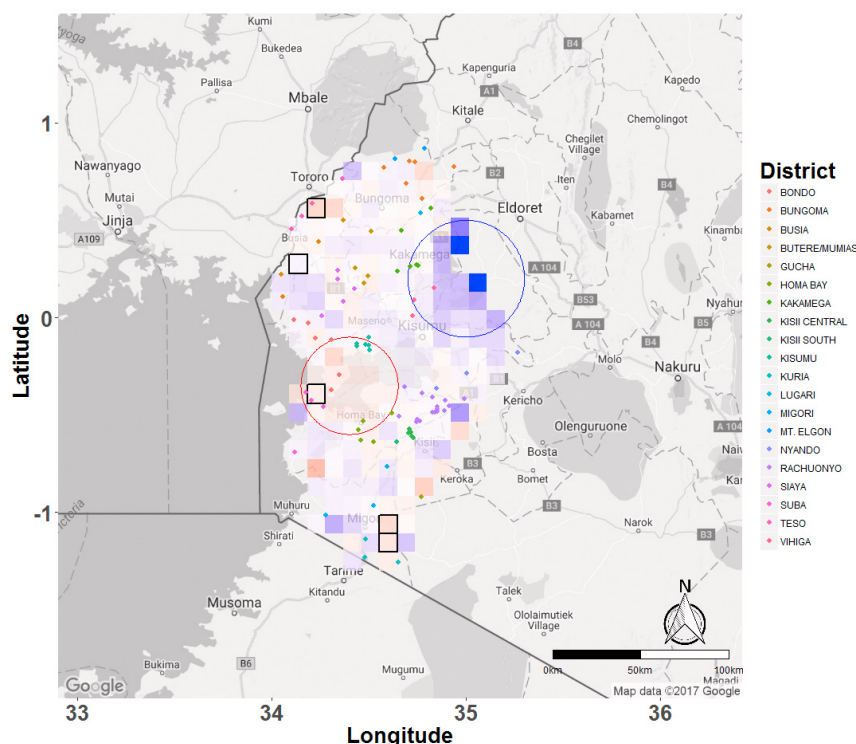
Figure 9. Map of the western Kenya study area with raster grids representing bearing analyses superimposed on top of it. Multivariable linear regression analysis was carried out to determine bearing (directionality of movement) of *Plasmodium falciparum* parasites among schools in the region. Grids are colour coded based on the coefficient estimates describing the effect size of that grid in influencing directional movement. Red represents east/west movement, while blue represents north/south movement. The grids with black borders represent those areas that were significant in east/west movement, even after Bonferroni-correction for multiple testing. The blue circle shows the region of the study site that had predominantly north/south movement, while the red circle represents that region that had predominantly east/west movement. Each dot represents a school, colour-coded based on the district in which the school is located.

parasites based on the second PC. This limited genetic clustering at both local and global scales, and weak evidence of genetic isolation by distance, are indicative of a parasite population that is well mixed at the sub-national geographical scale. This finding is in contrast to our micro-epidemiological study, which showed spatial structure to genetic relatedness over short distances[21].

In that previous study, however, we noted that the gradient between spatial separation and genetic relatedness was non-linear, and became less steep with distance such that past 10km there was little differentiation. This observation was hypothesized to be as a result of the rapid parasite movement and mixing observed within the study sites, with no geographical areas acting as spatial barriers to parasite movement, combined with a process operating at micro-geographical scales that results in a selection disadvantage to the autochthonous parasite population. In theory, the acquisition of parasite genotype-specific immunity or the impact of superinfection of incoming parasite types displacing existing parasites could meet these criteria. It is therefore consistent that we only identified a weak relationship in our study of schools where most pairs of schools were more than 10km apart.

We further examined the geographical distribution of allele frequencies for all 83 SNPs in our study population. Studies of allele frequency distribution have been used to determine parasite population structure and identify patterns of local adaptation of *P. falciparum* isolates[52–54]. Such local adaptation may be due to various selection pressures, including environmental pressure and immune selection, and may occur at individual, population or regional scales[55,56]. Of the 83 SNPs examined, we identified 18 SNPs that had statistically significant variations in allele frequencies among schools from the logistic regression analysis, although none of these 18 SNPs were significant after accounting for multiple testing. These findings suggest that although we were not sufficiently powered to distinguish any individual SNPs as likely to be significant beyond a Bonferroni correction, on the other hand the fact that 18 SNPs showed a value of p<0.05 when only 4 would be expected by chance suggests that there may be some genuine differences in frequency within the group of SNPs.

We reasoned that genuine geographical variation would be likely to show spatial clustering as well as variation by school, and so we

measured the scan statistic for those showing significant variation among schools. 5 of the 18 SNPs that were identified showed local clustering among schools. However, these SNPs were not entirely private to a sub-population and occurred in schools inside and outside the clusters. This finding provides weak support for the existence of variable local genetic selection pressures in this parasite population. The identification of SNPs with significant geographical variation in allele frequencies could indicate adaptation of *P. falciparum* parasite populations to their local environment, or more likely may indicate a temporary expansion of a parasite sub-population with a particular SNP simply due to random genetic drift.

An extensive analysis of the study area for spatial barriers to parasite movement using 10km-by-10km sized pixels provided little evidence for the existence of geographical regions that act as barriers to parasite movement at the sub-national scale, and is in agreement with our previous study which did not identify any barriers to parasite movement at a micro-epidemiological scale[21].

This observation of free movement over the western Kenya region is supported by a previous analysis of mobile phone data that was used to analyse patterns of human movement within the country[57], and which showed substantial movement of people within this region, and further supports our observation of little or no barriers to parasite movement within the region. However, we also observed a cluster of pixels representing predominantly north/south movement in the north east and another cluster representing predominantly east/west movement in the west of the study area, and when we analysed the site for spatial autocorrelation in the directionality of pixels, we found statistically significant autocorrelation for school pairs separated by up to 40km. This means that pixels with greater east/west movement were more frequently found next to other pixels with greater east/west movement, and similarly, pixels with greater north/south movement were more frequently found next to pixels with greater north/south movement. Furthermore, some individual pixels showed statistically significant directionality that met Bonferroni-adjusted significance criteria. Although spatial autocorrelation of directionality might simply be because the same data (the same school pairs) cross pixels that are physically close to each other, our observation of two large clusters of pixels with distinct directional patterns of parasite movement is unlikely to have been an artefact of the same school pairs being analysed when all pixels in the clusters were considered, suggesting that we are able to detect specific migration pathways of parasites.

The findings in this study have several implications for the outcomes of malaria control programmes. Since we show that parasite populations mix to high degrees within the region, with little evidence of geographical clustering, one might conclude that interventions targeting smaller geographical areas within the region are likely to reduce the flow of parasites to regions beyond the targeted region. However, the high degree of parasite mixing also means that parasites move relatively freely within the

region, and there is therefore a high likelihood of importation of infection from untargeted to targeted regions. This is strongly corroborated by evidence from a cluster-randomized controlled trial in a highland region of western Kenya that showed no impact in reducing transmission inside hotspots 16 weeks after applying interventions, possibly due to the importation of parasites from untargeted surrounding regions[9].

This study had some limitations. First, we cannot be definite about the time-scale over which gene flow has occurred. If the gene flow is rapid, this supports our conclusions regarding malaria control. On the other hand, it is possible that the well-mixed population emerged over a longer period of time and that gene flow, while resulting in complete mixing, could be less rapid, in which case targeted interventions would probably not have far reaching effects in the surrounding community. Our previous study showing spatial and temporal structure at a fine micro-epidemiological scale suggests rapid gene flow[21]. In that study, we showed that parasite pairs taken from nearby homesteads had fewer SNP differences between them than parasite pairs that were further apart. However, over the period of a month this distance gradient was attenuated, and was gone by one year. However, more definitive work will require an in-depth analysis of whole genome data to identify haplotypes and rare variants in the population, and infer variation over time.

Second, geospatial coordinate data was collected for schools as opposed to individual homesteads, and hence genotype data was aggregated at this level. We were therefore unable to detect structure at micro-epidemiological scales. Third, we analysed only a small number of SNPs. This made it impossible to detect relatively rare private SNPs. It is likely that a larger set of genetic markers will be required to identify private SNPs and evidence of local parasite adaptation. SNPs in genes previously shown to be under selection in the parasite genome may additionally be analysed to determine whether population structure is observed based on local variations in selection pressure. Our previous study showed no population structure when SNPs in EBA175 and AMA1 were analysed[21]. We therefore did not type and analyse separately SNPs in antigenic genes for the present study.

Fourth, we analysed genotype data collected from only one part of the country, thus we are unable to describe patterns of parasite flow across the country, or to generalize our findings to other geographical areas. Additional analyses of samples from other regions of the country that experience malaria transmission such as coast, eastern and north eastern provinces are recommended. Furthermore, over longer distances human movement becomes more important than mosquito movement in distributing parasites and therefore will need to be taken into consideration when analysing parasite genetic relatedness across large spatial scales. Information on travel distance can be obtained from travel history, or more objectively from mobile phone data, and can be used to track human movement between sources and sinks of parasite transmission. Concordance between spatial parasite genetic relatedness and human movement will further support our hypothesis of high parasite movement and mixing.

In conclusion, we have shown that parasites mix to high levels within the western Kenya region, with no evidence of parasite sub-populations and weak evidence of spatial autocorrelation of parasite genotypes at the local and global scales. We have also shown that directionality of parasite migration can be inferred based on genetic relatedness, and gene flow models, e.g. as implemented in Migrate-N software, can be used to determine the migration rates within the region, although such models are likely to prove more useful if distinct parasite populations exist and can be identified within the region.

Data availability

Figshare: Dataset 1: Genotyping results for 111 single nucleotide polymorphisms (SNPs) typed in 2486 *Plasmodium falciparum* samples collected from primary school children during a parasitological survey in western Kenya in 2009 and 2010.

Doi: http://dx.doi.org/10.6084/m9.figshare.4806619[58]

Figshare: Dataset 2: Single nucleotide polymorphisms (SNPs) and distance differences between *Plasmodium falciparum* parasite pairs sampled during a parasitological survey of primary school children in western Kenya.

Doi: http://dx.doi.org/10.6084/m9.figshare.4806631[59]

Supplementary material

Supplementary Table 1: Sequenom assay design information for *Plasmodium falciparum* samples collected from primary school children during a parasitological survey in Western Kenya. Data includes the locus and IPLEX specific primers used in the Sequenom reaction to amplify and genotype SNPs of interest. gene product, gene product name; gene_symbol, gene name; assay_code, name of SNP assay; chr_valid, chromosome number; coord_valid, SNP position on chromosome; reference_allele, 3D7 reference allele; non_reference_allele, alternative allele; sequence, 3D7 reference sequence spanning the SNP site; first_pcrp, first PCR primer sequence; second_pcrp, second PCR primer sequence; extension_primer, IPLEX extension primer sequence; ext1_call, extended IPLEX primer; ext1_mass, Mass of the extended IPLEX primer; ext1_seq, sequence of extended IPLEX primer; ext2_call, IPLEX primer with alternative extended allele; ext2_mass, Mass of the extended IPLEX primer with alternative allele; ext2_seq, sequence of extended IPLEX primer with alternative allele.

Click here to access the data.

Supplementary Table 2: SNPs that showed significant clusters based on similarities in allele frequencies. *Population size includes both cases (samples with the major allele) and controls (samples with the minor allele). Clusters were generated in SaTScan based on a Bernoulli probability model.

Click here to access the data.

Supplementary figure 1*: P. falciparum* parasite prevalence in primary schools across Western Kenya. Each dot represents an individual school, colour-coded based on parasite prevalence (%). Parasite prevalence ranged from 0.9% – 62%.

Click here to access the data.

References

1. Ceesay SJ, Casals-Pascual C, Erskine J, *et al.*: **Changes in malaria indices between 1999 and 2007 in The Gambia: a retrospective analysis.** *Lancet.* 2008; **372**(9649): 1545–1554.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

2. Noor AM, Kinyoki DK, Mundia CW, *et al.*: **The changing risk of *Plasmodium falciparum* malaria infection in Africa: 2000–10: a spatial and temporal analysis of transmission intensity.** *Lancet.* 2014; **383**(9930): 1739–1747.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

3. O'Meara WP, Bejon P, Mwangi TW, *et al.*: **Effect of a fall in malaria transmission on morbidity and mortality in Kilifi, Kenya.** *Lancet.* 2008; **372**(9649): 1555–1562.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

4. Bhattarai A, Ali AS, Kachur SP, *et al.*: **Impact of artemisinin-based combination therapy and insecticide-treated nets on malaria burden in Zanzibar.** *PLoS Med.* 2007; **4**(11): e309.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

5. Snow RW, Kibuchi E, Karuri SW, *et al.*: **Changing Malaria Prevalence on the Kenyan Coast since 1974: Climate, Drugs and Vector Control.** *PLoS One.* 2015; **10**(6): e0128792.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

6. Bejon P, Williams TN, Liljander A, *et al.*: **Stable and unstable malaria hotspots in longitudinal cohort studies in Kenya.** *PLoS Med.* 2010; **7**(7): e1000304.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

7. Bejon P, Williams TN, Nyundo C, *et al.*: **A micro-epidemiological analysis of febrile malaria in Coastal Kenya showing hotspots within hotspots.** *eLife.* 2014; **3**: e02130.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

8. Bousema T, Griffin JT, Sauerwein RW, *et al.*: **Hitting hotspots: spatial targeting of malaria for control and elimination.** *PLoS Med.* 2012; **9**(1): e1001165.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

9. Bousema T, Stresman G, Baidjoe AY, *et al.*: **The Impact of Hotspot-Targeted Interventions on Malaria Transmission in Rachuonyo South District in the Western Kenyan Highlands: A Cluster-Randomized Controlled Trial.** *PLoS Med.* 2016; **13**(4): e1001993.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

10. Machault V, Vignolles C, Pagès F, *et al.*: **Spatial heterogeneity and temporal evolution of malaria transmission risk in Dakar, Senegal, according to remotely sensed environmental data.** *Malar J.* 2010; **9**: 252.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

11. Oduro AR, Conway DJ, Schellenberg D, *et al.*: **Seroepidemiological and parasitological evaluation of the heterogeneity of malaria infection in the Gambia.** *Malar J.* 2013; **12**: 222.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

12. Oesterholt MJ, Bousema JT, Mwerinde OK, *et al.*: **Spatial and temporal variation in malaria transmission in a low endemicity area in northern Tanzania.** *Malar J.* 2006; **5**: 98.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

13. Bousema T, Drakeley C, Gesase S, *et al.*: **Identification of hot spots of malaria transmission for targeted malaria control.** *J Infect Dis.* 2010; **201**(11): 1764–1774.
    PubMed Abstract | **Publisher Full Text**

14. Carter R, Mendis KN, Roberts D: **Spatial targeting of interventions against malaria.** *Bull World Health Organ.* 2000; **78**(12): 1401–1411.
    PubMed Abstract | **Free Full Text**

15. Ruktanonchai NW, DeLeenheer P, Tatem AJ, *et al.*: **Identifying Malaria Transmission Foci for Elimination Using Human Mobility Data.** *PLoS Comput Biol.* 2016; **12**(4): e1004846.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

16. Woolhouse ME, Dye C, Etard JF, *et al.*: **Heterogeneities in the transmission of infectious agents: implications for the design of control programs.** *Proc Natl Acad Sci U S A.* 1997; **94**(1): 338–342.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

17. Dolgin E: **Targeting hotspots of transmission promises to reduce malaria.** *Nat Med.* 2010; **16**(10): 1055.
    PubMed Abstract | **Publisher Full Text**

18. Mobegi VA, Loua KM, Ahouidi AD, *et al.*: **Population genetic structure of *Plasmodium falciparum* across a region of diverse endemicity in West Africa.** *Malar J.* 2012; **11**: 223.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

19. Campino S, Auburn S, Kivinen K, *et al.*: **Population genetic analysis of *Plasmodium falciparum* parasites using a customized Illumina GoldenGate genotyping assay.** *PLoS One.* 2011; **6**(6): e20251.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

20. Manske M, Miotto O, Campino S, *et al.*: **Analysis of *Plasmodium falciparum* diversity in natural infections by deep sequencing.** *Nature.* 2012; **487**(7407): 375–379.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

21. Omedo I, Mogeni P, Bousema T, *et al.*: **Micro-epidemiological structuring**

of *Plasmodium falciparum* parasite populations in regions with varying transmission intensities in Africa. [version 1; referees: 4 approved]. *Wellcome Open Res.* 2017; **2**: 10.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

22. Patel JC, Taylor SM, Juliao PC, *et al.*: **Genetic Evidence of Importation of Drug-Resistant *Plasmodium falciparum* to Guatemala from the Democratic Republic of the Congo.** *Emerg Infect Dis.* 2014; **20**(6): 932–940.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

23. Klein EY: **Antimalarial drug resistance: a review of the biology and strategies to delay emergence and spread.** *Int J Antimicrob Agents.* 2013; **41**(4): 311–317.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

24. Bhatt S, Weiss DJ, Cameron E, *et al.*: **The effect of malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015.** *Nature.* 2015; **526**(7572): 207–211.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

25. Mogeni P, Williams TN, Fegan G, *et al.*: **Age, Spatial, and Temporal Variations in Hospital Admissions with Malaria in Kilifi County, Kenya: A 25-Year Longitudinal Observational Study.** *PLoS Med.* 2016; **13**(6): e1002047.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

26. Division of Malaria Control, M. o. P. H. a. S: **Kenya Malaria Indicator Survey.** 2015.
    Reference Source

27. Okiro EA, Alegana VA, Noor AM, *et al.*: **Malaria paediatric hospitalization between 1999 and 2008 across Kenya.** *BMC Med.* 2009; **7**: 75.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

28. Bayoh MN, Walker ED, Kosgei J, *et al.*: **Persistently high estimates of late night, indoor exposure to malaria vectors despite high coverage of insecticide treated nets.** *Parasit Vectors.* 2014; **7**: 380.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

29. Zhou G, Afrane YA, Vardo-Zalik AM, *et al.*: **Changing patterns of malaria epidemiology between 2002 and 2010 in Western Kenya: the fall and rise of malaria.** *PLoS One.* 2011; **6**(5): e20318.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

30. Stevenson J, St Laurent B, Lobo NF, *et al.*: **Novel vectors of malaria parasites in the western highlands of Kenya.** *Emerg Infect Dis.* 2012; **18**(9): 1547–1549.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

31. Olanga EA, Okombo L, Irungu LW, *et al.*: **Parasites and vectors of malaria on Rusinga Island, Western Kenya.** *Parasit Vectors.* 2015; **8**: 250.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

32. Gitonga CW, Karanja PN, Kihara J, *et al.*: **Implementing school malaria surveys in Kenya: towards a national surveillance system.** *Malar J.* 2010; **9**: 306.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

33. Gitonga CW, Edwards T, Karanja PN, *et al.*: ***Plasmodium* infection, anaemia and mosquito net use among school children across different settings in Kenya.** *Trop Med Int Health.* 2012; **17**(7): 858–870.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

34. Gabriel S, Ziaugra L, Tabbaa D: **SNP genotyping using the Sequenom MassARRAY iPLEX platform.** *Curr Protoc Hum Genet.* editorial board, Jonathan L. Haines ... [*et al.*]. 2009; **Chapter 2**: Unit 2.12.
    PubMed Abstract | **Publisher Full Text**

35. Team RCR: **A language and environment for statistical computing.** R Foundation for Statistical Computing, Vienna, Austria. 2016.

36. Kulldorff M: **SaTScan User Guide.** 2015.
    Reference Source

37. Ringnér M: **What is principal component analysis?** *Nat Biotechnol.* 2008; **26**(3): 303–304.
    PubMed Abstract | **Publisher Full Text**

38. Epperson BK, Li T: **Measurement of genetic structure within populations using Moran's spatial autocorrelation statistics.** *Proc Natl Acad Sci U S A.* 1996; **93**(19): 10528–10532.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

39. Kulldorf M: **SaTScan v9.3: Software for the spatila and space-time scan statistics.** 2014.

40. Owusu-Agyei S, Smith T, Beck HP, *et al.*: **Molecular epidemiology of *Plasmodium falciparum* infections among asymptomatic inhabitants of a holoendemic malarious area in northern Ghana.** *Trop Med Int Health.* 2002; **7**(5): 421–428.
    PubMed Abstract | **Publisher Full Text**

41. Smith, T, Beck HP, Kitua A, *et al.*: **Age dependence of the multiplicity of *Plasmodium falciparum* infections and of other malariological indices in an area of high endemicity.** *Trans R Soc Trop Med Hyg.* 1999; **93**(Suppl 1): 15–20.
    PubMed Abstract | **Publisher Full Text**

42. Konaté L, Zwetyenga J, Rogier C, *et al.*: **Variation of *Plasmodium falciparum* msp1 block 2 and msp2 allele prevalence and of infection complexity in two neighbouring Senegalese villages with different transmission conditions.** *Trans R Soc Trop Med Hyg.* 1999; **93**(Suppl 1): 21–28.
    PubMed Abstract | **Publisher Full Text**

43. Cortes A, Mellombo M, Mueller I, et al.: **Geographical structure of diversity and differences between symptomatic and asymptomatic infections for *Plasmodium falciparum* vaccine candidate AMA1.** *Infect Immun.* 2003; **71**(3): 1416–1426.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

44. Ingasia LA, Cheruiyot J, Okoth SA, et al.: **Genetic variability and population structure of *Plasmodium falciparum* parasite populations from different malaria ecological regions of Kenya.** *Infect Genet Evol.* 2016; **39**: 372–380.
PubMed Abstract | **Publisher Full Text**

45. Ototo EN, Mbugi JP, Wanjala CL, et al.: **Surveillance of malaria vector population density and biting behaviour in western Kenya.** *Malar J.* 2015; **14**: 244.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

46. Zhou G, Lee MC, Githeko AK, et al.: **Insecticide-Treated Net Campaign and Malaria Transmission in Western Kenya: 2003–2015.** *Front Public Health.* 2016; **4**: 153.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

47. Gatei W, Gimnig JE, Hawley W, et al.: **Genetic diversity of *Plasmodium falciparum* parasite by microsatellite markers after scale-up of insecticide-treated bed nets in western Kenya.** *Malar J.* 2015; **13**(Suppl 1): 495.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

48. Miotto O, Amato R, Ashley EA, et al.: **Genetic architecture of artemisinin-resistant *Plasmodium falciparum*.** *Nat Genet.* 2015; **47**(3): 226–234.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

49. Baliraine FN, Afrane YA, Amenya DA, et al.: **A cohort study of *Plasmodium falciparum* infection dynamics in Western Kenya Highlands.** *BMC Infect Dis.* 2010; **10**: 283.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

50. Bonizzoni M, Afrane Y, Baliraine FN, et al.: **Genetic structure of *Plasmodium falciparum* populations between lowland and highland sites and antimalarial drug resistance in Western Kenya.** *Infect Genet Evol.* 2009; **9**(5): 806–812.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

51. Zhong D, Afrane Y, Githeko A, et al.: ***Plasmodium falciparum* genetic diversity in western Kenya highlands.** *Am J Trop Med Hyg.* 2007; **77**(6): 1043–1050.
PubMed Abstract

52. Anderson TJ, Nair S, Sudimack D, et al.: **Geographical distribution of selected and putatively neutral SNPs in Southeast Asian malaria parasites.** *Mol Biol Evol.* 2005; **22**(12): 2362–2374.
PubMed Abstract | **Publisher Full Text**

53. Schlötterer C: **Towards a molecular characterization of adaptation in local populations.** *Curr Opin Genet Dev.* 2002; **12**(6): 683–687.
PubMed Abstract | **Publisher Full Text**

54. Gunther T, Coop G: **Robust identification of local adaptation from allele frequencies.** *Genetics.* 2013; **195**(1): 205–220.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

55. Ochola LI, Tetteh KK, Stewart LB, et al.: **Allele frequency-based and polymorphism-versus-divergence indices of balancing selection in a new filtered set of polymorphic genes in *Plasmodium falciparum*.** *Mol Biol Evol.* 2010; **27**(10): 2344–2351.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

56. Kaltz O, Shykoff JA: **Local adaptation in host-parasite systems.** *Heredity.* 1998; **81**: 361–370.
Publisher Full Text

57. Wesolowski A, Eagle N, Tatem AJ, et al.: **Quantifying the impact of human mobility on malaria.** *Science.* 2012; **338**(6104): 267–270.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

58. Omedo I, Mogeni P, Rockett K, et al.: **Dataset 1: Genotyping results for 111 single nucleotide polymorphisms (SNPs) typed in 2486 *Plasmodium falciparum* samples collected from primary school children during a parasitological survey in western Kenya in 2009 and 2010.** *Figshare.* 2017.
Data Source

59. Omedo I, Mogeni P, Rockett K, et al.: **Dataset 2: Single nucleotide polymorphisms (SNPs) and distance differences between *Plasmodium falciparum* parasite pairs sampled during a parasitological survey of primary school children in western Kenya.** *Figshare.* 2017.
Data Source

Open Peer Review

Current Referee Status: ✔ ✔

Version 2

Referee Report 06 September 2017

doi:10.21956/wellcomeopenres.13595.r25684

✔ **Lucy C. Okell** iD

MRC Centre for Outbreak Analysis & Modelling, Department of Infectious Disease Epidemiology, Imperial College London, London, UK

The authors have addressed my comments very thoroughly and I have no further additions.

Competing Interests: LCO declares grant funding from the World Health Organization (WHO), the Bill & Melinda Gates Foundation and Medicines for Malaria Venture (MMV), and has received consultancy contracts from MMV and WHO.

Referee Expertise: Mathematical modelling of malaria

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Referee Report 06 July 2017

doi:10.21956/wellcomeopenres.12114.r22048

✔ **Lucy C. Okell** iD

MRC Centre for Outbreak Analysis & Modelling, Department of Infectious Disease Epidemiology, Imperial College London, London, UK

This is a very interesting and well-conducted analysis of spatial structure in parasite populations in Western Kenya, as assessed by detailed SNP barcoding of parasites. This will be useful information for malaria control programmes in order to understand the appropriate spatial scale at which to target interventions, as well as of much interest to researchers in this field. It builds on a number of studies which assessed spatial structure of parasites in other locations on a smaller scale, and adds useful information to these other studies.

I have a few comments and suggestions:

- A number of different tests and analyses were conducted, which becomes slightly confusing. It would be helpful to have a list or table briefly indicating what each analysis is testing for.

- The authors mention this is an area of high transmission intensity – it would be interesting to additionally present a measure of malaria prevalence in the schools and roughly how much this varied across the area. Also, was there any difference in population structure in lower versus higher transmission areas? One might hypothesise that there could be.

- When an infection contained mixed parasite genotypes, the authors analysed only the dominant genotype. This is fine, but did they also consider using a method such as the REAL McCOIL[1]? This is a statistical method for estimating SNP frequencies, using data from mixed infections, which avoids discarding the additional information available from minority genotypes. This might make quite a difference in this area of higher transmission, where presumably a fairly high proportion of infections contain more than one parasite population. I am not sure whether the sample sizes would be sufficient to run this method for each school, and thus from there complete the other analyses, but it would be worth investigating.

- In the methods section, could the authors add some more detail on this analysis, or provide a reference? "We then ran a purely spatial analysis based on a normal probability distribution model and located geographical regions with clusters of high PC scores." I am not sure how easily others could reproduce it from its current description.

- The analysis looking at spatial barriers makes sense, but I wonder if the authors considered computing travel distance, and thus using more a priori information on potential barriers? Over these distances, presumably human movement may be more important than vector movement in distributing parasites. Although given the lack of spatial barriers identified by the pixel-based analysis, perhaps travel distance would be unlikely to show any different result.

- This section in the discussion: "Of the 83 SNPs examined, we identified 18 SNPs that had statistically significant variations in allele frequencies among schools from the logistic regression analysis. Although none of these 18 SNPs were significant after accounting for multiple testing, there was a clear excess of statistically significant SNPs, suggesting that the variations seen were not simply due to random chance." I found rather confusing. If I read it correctly, it suggests that after multiple testing, the differences in frequencies were not significant, and could be due to random chance?

- I found this sentence very interesting: "In that previous study, however, we noted that the gradient between spatial separation and genetic relatedness was non-linear, and became less steep with distance such that past 10km there was little differentiation."
Could the authors elaborate on possible hypotheses for this observation? (or summarise from the previous paper if already included there).

References

1. Chang HH, Worby CJ, Yeka A, Nankabirwa J, Kamya MR, Staedke SG, Dorsey G, Murphy M, Neafsey DE, Jeffreys AE, Hubbart C, Rockett KA, Amato R, Kwiatkowski DP, Buckee CO, Greenhouse B: THE REAL McCOIL: A method for the concurrent estimation of the complexity of infection and SNP allele

frequency for malaria parasites.*PLoS Comput Biol*. 2017; **13** (1): e1005348 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: I declare previous consultancies from Medicines for Malaria Venture and from the World Health Organization.

Referee Expertise: Mathematical modelling of malaria

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 29 Aug 2017
Irene Omedo, KEMRI- Wellcome Trust Research Programme, Kenya

We would like to thank you for reviewing our work and proving constructive advice on how to improve it as well as future studies. We have addressed your comments as shown below (in bold).

A number of different tests and analyses were conducted, which becomes slightly confusing. It would be helpful to have a list or table briefly indicating what each analysis is testing for.

A table showing the statistical tests and their functions has been added under the "statistical analysis" heading in the methods section.

The authors mention this is an area of high transmission intensity – it would be interesting to additionally present a measure of malaria prevalence in the schools and roughly how much this varied across the area. Also, was there any difference in population structure in lower versus higher transmission areas? One might hypothesise that there could be.

A map of the study region showing parasite prevalence for each school has been added as supplementary figure 1. The following accompanying text has been added to explain the observation from the map.

"Variation in parasite prevalence was observed in the region, with areas north and west of Lake Victoria having higher infection prevalence than areas south and east of the lake (Supplementary Figure 1)."

No differences in population structure were observed in lower versus higher transmission areas, and a statement to this effect has been included in the results section under 'P. falciparum population genetics':

"At the sub-national level, we were unable to resolve the parasite population into distinct sub-populations based on PCA, and there was no difference in population structure in lower versus higher transmission intensity areas."

When an infection contained mixed parasite genotypes, the authors analysed only the dominant genotype. This is fine, but did they also consider using a method such as the REAL McCOIL[1]? This is a statistical method for estimating SNP frequencies, using data from mixed infections, which avoids discarding the additional information available from minority genotypes. This might make quite a difference in this area of higher transmission, where presumably a fairly high proportion of infections contain more than one parasite population. I am not sure whether the sample sizes would be sufficient to run this method for each school, and thus from there complete the other analyses, but it would be worth investigating.

Although THE REAL McCOIL seems to be robust at estimating complexity of infections and allele frequencies in samples with mixed infections, its focused mainly on determining SNP frequencies. For this reason, THE REAL McCOIL requires parasites to be initially assigned into distinct populations. We do not believe that this statistical method is appropriate in our case because we wanted to focus our analyses on individual parasite genotypes and their relatedness in pairwise analyses, to determine whether we can use genetic relatedness at this level to identify distinct parasite populations, and then use measures of spatial autocorrelation to identify the location and size of any parasite sub-populations.

In the methods section, could the authors add some more detail on this analysis, or provide a reference? "We then ran a purely spatial analysis based on a normal probability distribution model and located geographical regions with clusters of high PC scores." I am not sure how easily others could reproduce it from its current description.

Additional information has been added to better explain the analysis. A reference has also been added. The additional text reads as follows:

"We used the latitude/longitude coordinates to specify feature (sample) locations and the PC scores to represent feature attributes (parasite genotypes), and ran a purely spatial, retrospective analysis based on a normal probability distribution model implemented in

SaTScan software[1]. We identified geographical regions with clusters containing parasite genotypes associated with high PC scores."

The analysis looking at spatial barriers makes sense, but I wonder if the authors considered computing travel distance, and thus using more a priori information on potential barriers? Over these distances, presumably human movement may be more important than vector movement in distributing parasites. Although given the lack of spatial barriers identified by the pixel-based analysis, perhaps travel distance would be unlikely to show any different result.

Travel distance will add more information and help to draw more concrete conclusions. Such information can be derived from travel history or mobile phone data. However, we did not have this data at the time of analysis and thus did not include it. We have however, recommended this as a logical next step in any future analysis of parasite gene flow when using genetic data to analyse parasite movement. We have included a sentence in the discussion recommending this that reads as follows:

"Additional analyses of samples from other regions of the country that experience malaria transmission such as coast, eastern and north eastern provinces are recommended. Furthermore, over longer distances human movement becomes more important than mosquito movement in distributing parasites and therefore will need to be taken into consideration when analysing parasite genetic relatedness across large spatial scales. Information on travel distance can be obtained from travel history, or more objectively from mobile phone data, and can be used to track human movement between sources and sinks of parasite transmission. Concordance between spatial parasite genetic relatedness and human movement will further support our hypothesis of high parasite movement and mixing."

This section in the discussion: "Of the 83 SNPs examined, we identified 18 SNPs that had statistically significant variations in allele frequencies among schools from the logistic regression analysis. Although none of these 18 SNPs were significant after accounting for multiple testing, there was a clear excess of statistically significant SNPs, suggesting that the variations seen were not simply due to random chance." I found rather confusing. If I read it correctly, it suggests that after multiple testing, the differences in frequencies were not significant, and could be due to random chance?

The statement has been clarified as follows:

"Of the 83 SNPs examined, we identified 18 SNPs that had statistically significant variations in allele frequencies among schools from the logistic regression analysis, although none of these 18 SNPs were significant after accounting for multiple testing. These findings suggest that although we were not sufficiently powered to distinguish any individual SNPs as likely to be significant beyond a Bonferroni correction, on the other hand the fact that 18 SNPs showed a value of p<0.05 when only 4 would be expected by chance suggests that there may be some genuine differences in frequency within the group of SNPs."

I found this sentence very interesting: "In that previous study, however, we noted that the gradient between spatial separation and genetic relatedness was non-linear, and became less steep with distance such that past 10km there was little differentiation."
Could the authors elaborate on possible hypotheses for this observation? (or summarise from the previous paper if already included there).

The following statement has been added to bring out the hypothesis of high parasite mixing drawn from the previous study. That study has also been referenced:

"This observation was hypothesized to be as a result of the rapid parasite movement and mixing observed within the study sites, with no geographical areas acting as spatial barriers to parasite movement, combined with a process operating at micro-geographical scales that results in a selection disadvantage to the autochthonous parasite population. In theory, the acquisition of parasite genotype-specific immunity or the impact of superinfection of incoming parasite types displacing existing parasites could meet these criteria."

Competing Interests: No competing interests were disclosed.

Referee Report 19 June 2017

doi:10.21956/wellcomeopenres.12114.r23564

Liwang Cui
Department of Entomology, Pennsylvania State University, University Park, PA, USA

This study analyzed the genetic structure of 2486 Plasmodium falciparum parasites collected from children in 95 primary schools in western Kenya. Using genotypes of 83 SNPs in 1809 samples collected from 88 schools, the authors did not identify clear genetic structuring of parasite populations using different spatial analysis methods. This study followed an earlier example, which compared parasite populations on a larger geographical scale. Similarly, no spatial barriers of parasite movement were identified, and two regions showed evidence of directional parasite movement. This is another nice piece of work, which provides useful information for the local malaria control programs.

Comments:
1. The authors may want to see whether separate analyses with antigenic SNPs (which may be under balancing selection) and other SNPs will reach similar conclusions.

2. The author may use a wider range colour coding for Figures 3 and 5 – the current colour scheme is difficult to see the differences.

3. Additional methods such as Migrate-n may provide further evidence on the directionality of parasite movements.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
I cannot comment. A qualified statistician is required.

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Referee Expertise: Molecular epidemiology

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

> Author Response 29 Aug 2017
> **Irene Omedo**, KEMRI- Wellcome Trust Research Programme, Kenya
>
> We would like to thank you for reviewing our work and proving constructive advice on how to improve it as well as future studies. We have addressed your comments as shown below (in bold).
>
> The authors may want to see whether separate analyses with antigenic SNPs (which may be under balancing selection) and other SNPs will reach similar conclusions.
>
> **The following statement has been added in the discussion section to address this point:**
>
> **"SNPs in genes previously shown to be under selection in the parasite genome may additionally be analysed to determine whether population structure is observed based on local variations in selection pressure. Our previous study showed no population structure when SNPs in EBA175 and AMA1 were analysed, we therefore did not type and analyse separately SNPs in antigenic genes for the present study."**
>
> The author may use a wider range colour coding for Figures 3 and 5 – the current colour scheme is difficult to see the differences.
>
> **Due to the high level of similarity among the parasite genotypes analysed, most of the PC scores are quite similar in value, and this would be reflected in any continuous colour scheme chosen. Furthermore, we plotted the mean PC score per school, which further reduced the variability of the individual data points. The colour range shown in the figures is therefore more a factor or the PC values, rather than the specific colours used.**

Additional methods such as Migrate-n may provide further evidence on the directionality of parasite movements.

Although Migrate-N is a useful software for inferring population size and migration rates, it seems to work best when parasites are grouped into distinct populations prior to estimating these parameters. Our analysis focused on using genetic relatedness and measures of spatial autocorrelation to identify population structure. Once such structure has been identified then Migrate-N could be used to determine migration rates between these different populations. Unfortunately, we did not find strong evidence for population structure among parasites in the western Kenya region, thus this analysis of gene flow may be less useful in this case. However, we have noted in the conclusion paragraph that such software is available and can be used to study gene flow, once distinct populations have been identified.

The statement in the conclusion paragraph reads as follows:

"We have also shown that directionality of parasite migration can be inferred based on genetic relatedness, and gene flow models, e.g. as implemented in migrate-n software, can be used to determine the migration rates within the region, although such models are likely to prove more useful if distinct parasite populations exist and can be identified within the region."

Competing Interests: No competing interests were disclosed.