The Impact of Antimalarial Resistance on the Genetic Structure of *Plasmodium falciparum* in the DRC

Robert Verity[1,#], Ozkan Aydemir[2,#], Nicholas F. Brazeau[3,#], Oliver J. Watson[1], Nicholas J. Hathaway[4], Melchior Kashamuka Mwandagalirwa[5], Patrick W. Marsh[2], Kyaw Thwai[3], Travis Fulton[6], Madeline Denton[6], Andrew P. Morgan[6], Jonathan B. Parr[6], Patrick K. Tumwebaze[7], Melissa Conrad[8], Philip J. Rosenthal[8], Deus S. Ishengoma[9], Jeremiah Ngondi[10], Julie Gutman[11], Modest Mulenga[12], Douglas E. Norris[13], William J. Moss[14], Benedicta A Mensah[15], James L Myers-Hansen[15], Anita Ghansah[15], Antoinette K Tshefu[5], Azra C. Ghani[1], Steven R. Meshnick[3], Jeffrey A. Bailey[2,*], Jonathan J. Juliano[3,6,16,*,+]

1: Medical Research Council Centre for Global Infectious Disease Analysis, Department of Infectious Disease Epidemiology, Imperial College London, UK
2: Department of Pathology and Laboratory Medicine, Brown University, Providence, RI, USA
3: Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, USA
4: Program in Bioinformatics and Integrative Biology, University of Massachusetts, Worcester, MA, USA
5: Kinshasa School of Public Health, Hôpital Général Provincial de Référence de Kinshasa, Kinshasa, Democratic Republic of Congo
6: Division of Infectious Diseases, Department of Medicine, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
7: Infectious Disease Research Collaboration, Kampala, Uganda
8: Department of Medicine, University of California- San Francisco, San Francisco, CA, USA
9: National Institute for Medical Research, Tanga, Tanzania
10: RTI International, Dar es Salaam, Tanzania
11: Malaria Branch, Center for Global Health, Centers for Disease Control, Atlanta, GA, USA
12: Tropical Disease Research Centre, Ndola, Zambia
13: Department of Molecular Microbiology and Immunology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
14: Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
15: Noguchi Memorial Institute of Medical Research, University of Ghana, Accra, Ghana
16: Curriculum in Genetics and Molecular Biology, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

#: Co-first authors
*: Co-last authors
+corresponding

ABSTRACT

The Democratic Republic of the Congo (DRC) harbors 11% of global malaria cases, yet little is known about the spatial and genetic structure of the parasite population in that country. We sequenced 2537 *Plasmodium falciparum* infections, including a nationally representative population sample from DRC and samples from surrounding countries, using molecular inversion probes - a novel high-throughput genotyping tool. We identified an east-west divide in haplotypes known to confer resistance to chloroquine and sulfadoxine-pyrimethamine. Furthermore, we identified highly related parasites over large geographic distances, indicative of gene flow and migration. Our results were consistent with a background of isolation by distance combined with the effects of selection for antimalarial drug resistance. This study provides a high-resolution view of parasite genetic structure across a large country in Africa and provides a baseline to study how implementation programs may impact parasite populations.

BACKGROUND

Malaria remains one of the largest global public health challenges, with an estimated 219 million cases worldwide in 2017[1]. Despite decades of scale-up in control, there has been a recent resurgence, particularly in high transmission countries in sub-Saharan Africa[1]. In addition, the emergence of antimalarial resistance poses a major threat to current control and elimination efforts worldwide, and new tools are needed to quantify the changing landscape of drug resistance on timescales relevant to malaria control programmes. Genomics has emerged as a useful method for better understanding parasite populations that can be leveraged to support the design of effective interventions against a continually evolving parasite.

Data from genomic studies provides information that is complementary to epidemiological data[2], and can help to answer several key questions, including how parasites are transmitted, how drug resistance spreads, and how malaria control efforts impact the diversity of the parasite population. However, to date, efforts to use genomics to inform malaria control efforts have suffered from three major limitations. First, much of the work has been conducted in low transmission regions, such as Asia and transmission fringe regions of Africa, leaving it unclear how useful information can be gathered in the highest transmission settings. Some of these high burden regions have experienced increasing malaria prevalence in recent years and are now the center of strategic plans for control efforts[3,4]. Second, most genomic studies in Africa have relied upon convenience sampling from a few sites usually collected for other purposes, rather than population-representative samples. Lastly, studies have either relied on relatively few genetic markers, providing limited insight into the complete genome, or on expensive whole genome sequencing, limiting the number of samples studied. Overcoming these limitations is essential for genomics to have broader impacts on malaria control.

Within Africa, parasite populations have been shown to vary significantly between East and West, as demonstrated by their distinct antimalarial drug susceptibilities and population genetics[5,6]. However, few genomic studies have incorporated samples from central Africa, limiting our understanding of the connectivity of parasite populations across the continent. The Democratic Republic of the Congo (DRC) is the largest malaria-endemic country in Africa, borders nine countries and harbors approximately 11% of global *P. falciparum* malaria cases[1]. The DRC harbors a large, understudied parasite population that likely serves as a bridge between African parasite populations. Limited previous work has shown that the DRC represents a watershed between East and West African drug resistant parasite populations for sulfadoxine-pyrimethamine and chloroquine resistance[7–9]. More recently, parasite population structuring due to mutations at these and other loci associated with antimalarial resistance has been confirmed within the DRC[10]. However, studies focusing on hypervariable surface antigen diversity or neutral microsatellites have been unable to detect significant structure in the parasite population[10,11], likely due to a lack of high-quality genome-wide signal. A better understanding of parasite populations and the spread of antimalarial resistance in the DRC will allow for the design of more effective interventions accounting for evolutionary forces.

To address this knowledge gap, we leveraged a recent advance in malaria genomics, high-throughput molecular inversion probe (MIP) capture and sequencing, to characterize and map parasite population structure and antimalarial resistance profiles in the DRC and to define the connections of parasites within the DRC to East and West African parasite populations[12]. This approach provides a cost-effective and scalable method of genome interrogation, without the expense or informatic complexities of whole genome sequencing. We previously employed MIPs to comprehensively genotype known antimalarial resistance genes in several hundred samples from the DRC[10]. Here, we introduce an expanded MIP panel targeted at 1834 single nucleotide polymorphisms (SNPs) distributed throughout the *P. falciparum* genome, and designed to quantify differentiation and relatedness between samples. Using this panel of genome-wide SNP MIPs, in combination with the previous drug resistance MIP panel, we evaluated the parasite population diversity in 2537 parasite isolates from the DRC and surrounding countries in East and West Africa. We used this information to quantify relatedness of and gene-flow between parasites over large geographic scales and to assess the origins of antimalarial resistance mutations.

RESULTS

Sample quality and filtering: We obtained 2537 samples collected in 2013-2015 from the DRC and surrounding countries (DRC=2039, Ghana=194, Tanzania=120, Uganda=63, Zambia=121). All samples were sequenced using two separate MIP panels: a genome-wide panel designed to capture overall levels of differentiation and relatedness, and a drug resistance panel designed to target polymorphic sites known to be associated with antimalarial resistance[10]. The genome-wide panel included 739 ostensibly geographically informative SNPs, chosen on the basis of high differentiation (F_{ST}) between surrounding African countries in publicly available genomic sequences made available by the Pf3K project (see **Supplemental Text 1** and **Supplemental Table 1**), and 1151 putatively neutral SNPs distributed throughout the genome, with an overlap of 56 SNPs that were both neutral and geographically informative. The drug resistance panel included SNPs in known and putative drug resistance genes and has been described elsewhere [10]. The median number of unique molecular identifiers (UMIs) per MIP was 31 (range: 1-8,490) for the genome-wide panel, and 10 (range: 1-32,511) for the drug resistance panel. Complete UMI depth distributions are shown in **Supplemental Figure 1**. After filtering for samples and loci with sufficient UMI coverage, we were left with 1382 samples and 1079 loci from the genome-wide panel, and 674 samples and 1000 loci from the drug resistance panel, with an overlap of 452 samples between both panels. In addition to these samples, 114 controls consisting of known mixtures were sequenced and used to assess the accuracy of allele calls and frequencies. Expected versus measured allele frequencies for each SNP, calculated from these controls, are shown in **Supplemental Figure 2**.

Complexity of infection: Initial analyses focused on the genome-wide MIP panel only. Complexity of infection (COI) for each sample was estimated using THE REAL McCOIL[13] (**Supplemental Figure 3**). The mean COI was estimated at 2.2 (range 1 - 8) for the study as a whole. We observed significant differences in COI between countries (Ghana: 1.55 (non-parametric bootstrap 95% CI: 1.39 - 1.73), DRC: 2.23 (2.15 - 2.31), Tanzania: 2.17 (1.83 - 2.51), Zambia: 2.68 (2.39 - 3.00), Uganda 2.18 (1.87 - 2.51), and within the DRC we observed a statistically significant relationship between COI and *P. falciparum* prevalence by microscopy at both the province and cluster levels (**Supplemental Figure 4**), with higher COIs observed at higher prevalences.

Population structure: We explored population structure through principal component analysis (PCA) evaluated on within-sample allele frequencies at all 1079 genome-wide loci. We found the same separation between East and West Africa described in previous studies (**Figure 1**) as well as finer structure between regions within East Africa. DRC samples comprised a continuum between the East and West African clusters.

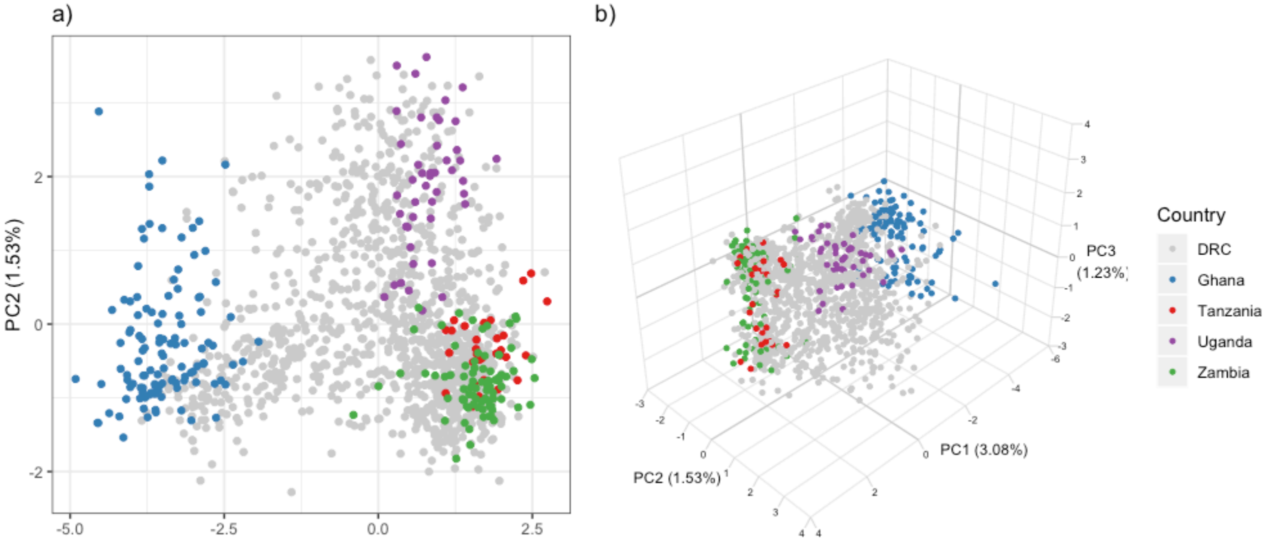

Figure 1 The first two (a) and three (b) principal components calculated from within-sample allele frequencies using the genome-wide MIP panel. Colors indicate country of origin of each sample.

The relative contribution of each locus to each principal component was quantified through normalized loading values. Relative contributions to the first four principal components are shown in **Figure 2**. After the fourth principal component the percent variance explained by subsequent components plateaued (**Supplemental Figure 5**). For principal component 1 (PC1) large contributions came from loci distributed throughout the genome, and a relatively larger contribution (65.2%) came from putatively geographically informative SNPs (non-parametric bootstrap, p<0.001). In contrast, contributions to PC2 were concentrated in a region on chromosome seven in close proximity to *P. falciparum* chloroquine resistance transporter (*pfcrt*), a known drug resistance locus, suggesting that resistance to chloroquine or amodiaquine may be driving differentiation along this secondary axis. For PC3, locus contributions were concentrated in three genic regions: PF3D7_0215300 (8.5%), PF3D7_0220300 (5.0%), and PF3D7_1127000 (4.3%). The first and largest of these encodes an acyl-CoA synthetase and is part of a diverse gene family known to undergo extensive gene conversion and recombination[14]. For PC4 we observed a region of high locus contribution on chromosome eight in close proximity to the known antifolate drug resistance gene dihydropteroate synthase (*dhps*). Combined, these results suggest that geography and drug resistance are both contributors to the observed population structure.

The relationship between the PCA results and the spatial distribution of parasites was explored by plotting raw principal component values against the geographic location of samples (**Figure 3a-3d**). For PC1 this revealed a complex pattern of spatial variation, containing both north-south and east-west clines. For PC2 and PC4 the maps essentially recapitulate the known geographic distribution of *pfcrt* and *dhps* resistance mutations, respectively (**Figure 3e-3f**). For PC3 the map indicates some east-west spatial structuring that is not explained by known markers of antimalarial resistance and warrants further investigation.

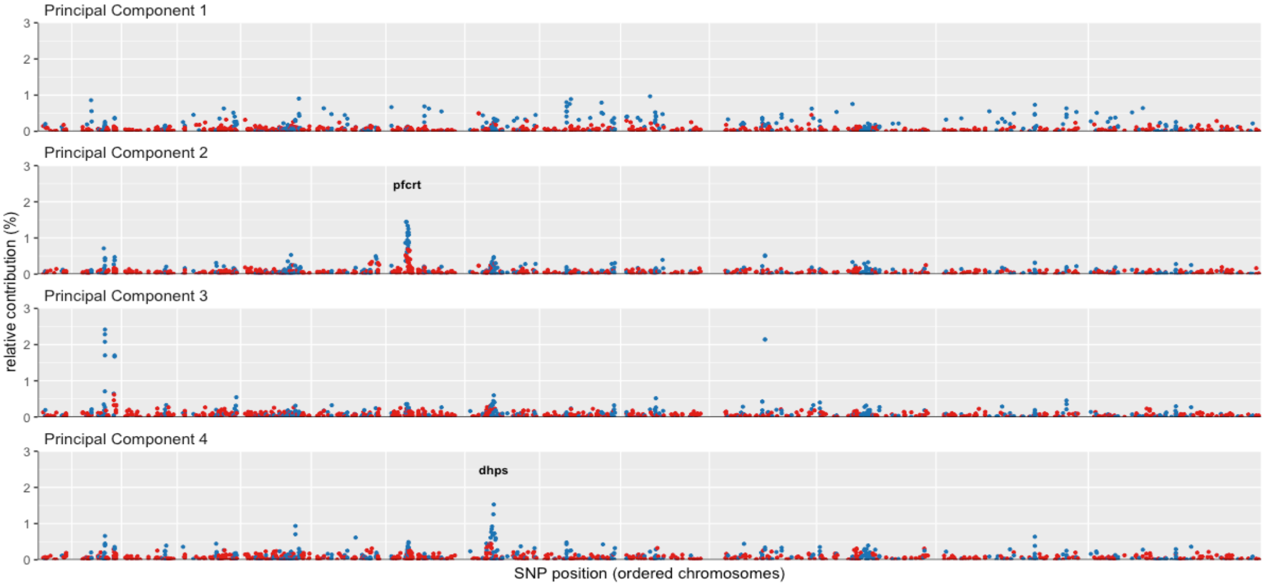

Figure 2 The relative contribution (%) of each locus to the first four principal components. Chromosomes are plotted in order, separated by vertical white gridlines. Point colors indicate sites that were chosen in the design based on F_{ST} values to be geographically informative (blue) or not (red).

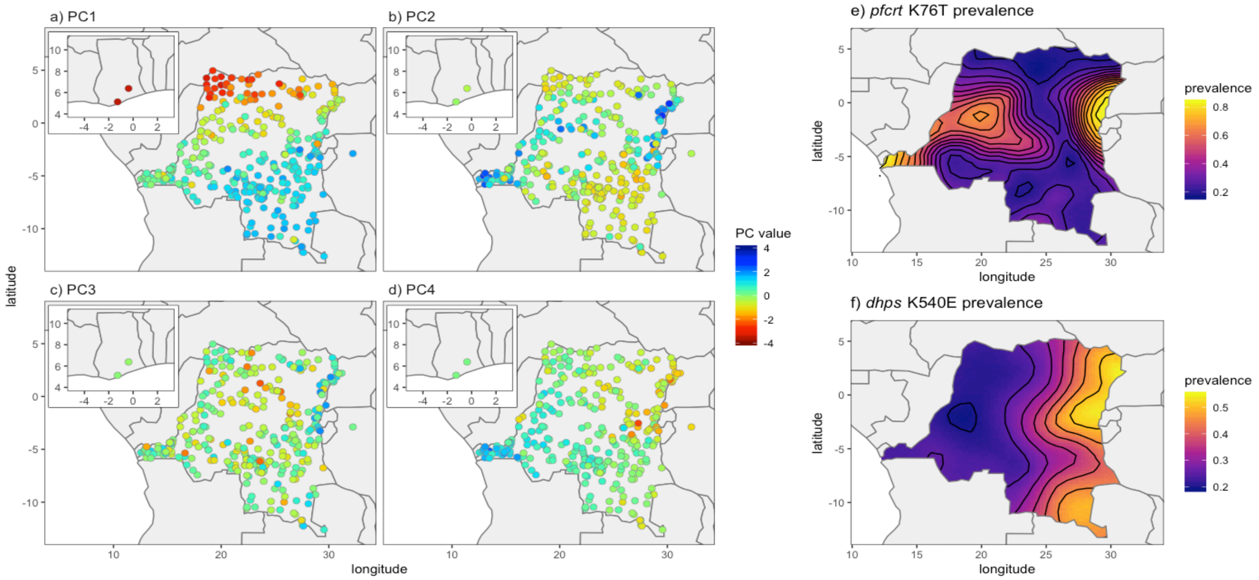

Figure 3 Panels (a) to (d) show the mean principal component value per DHS cluster. Panels (e) and (f) show estimated distributions of the prevalence of molecular markers of resistance for *pfcrt* and *pfdhps*.

Identity by Descent: The relatedness of all pairs of samples was explored through pairwise identity by descent (IBD), estimated using a maximum likelihood approach. IBD has advantages over simpler statistics like identity by state (IBS) in that it takes account of allele frequency distributions, and so provides an objective measure of relatedness that can be compared between studies[15]. The overall distribution of pairwise IBD was found to be heavy-tailed, consisting of a large body of weakly related samples and a tail of very highly related samples (**Figure 4**).

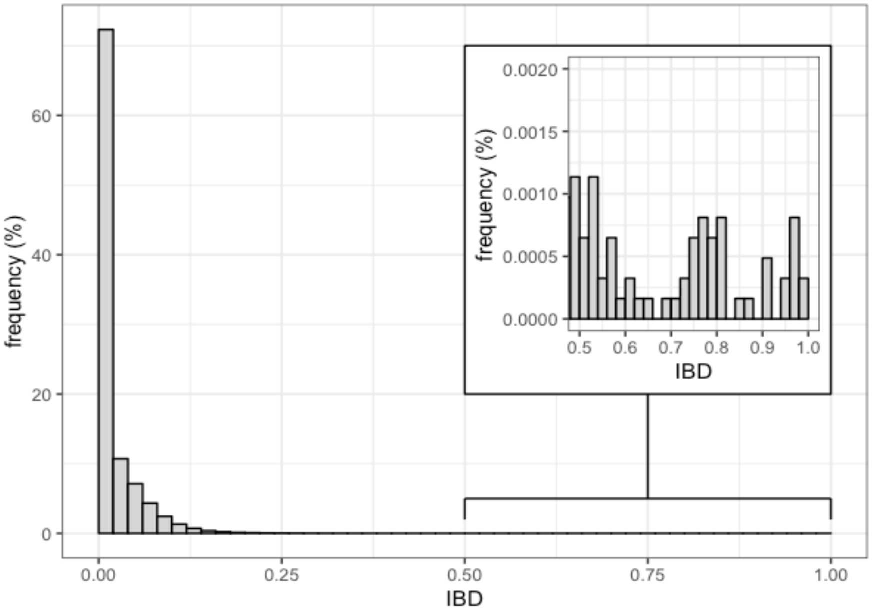

Figure 4 A histogram of pairwise identity by descent (IBD) between all samples, estimated by maximum likelihood. Inset shows the heavy tail of the distribution, with some pairs of samples having IBD > 0.9.

Mean IBD was significantly higher within clusters compared to between clusters (0.06 vs. 0.02, two-sample t-test, p<0.001). When plotted against geographic separation there was a clear fall-off of IBD with distance (**Figure 5a**), consistent with the classical pattern expected under isolation-by-distance[16,17]. Focussing on the tail of highly related samples, which includes the major strain in complex infections, there were 12 sample pairs with a relatedness greater than IBD=0.9. Comparison of raw allele frequency distributions confirmed that these were likely clones (**Supplemental Figure 6**). These highly related pairs were found more often within the same cluster than in different clusters (7 vs. 5 respectively, chi-squared test, p<0.001), suggesting the presence of local clonal transmission chains. The five between-cluster highly related pairs (**Figure 5b**) were spread over large geographic distances (281-1331 km), far beyond the normal expected scale of the breakdown in genetic relatedness (**Figure 5a**), suggesting recent long distance migration.

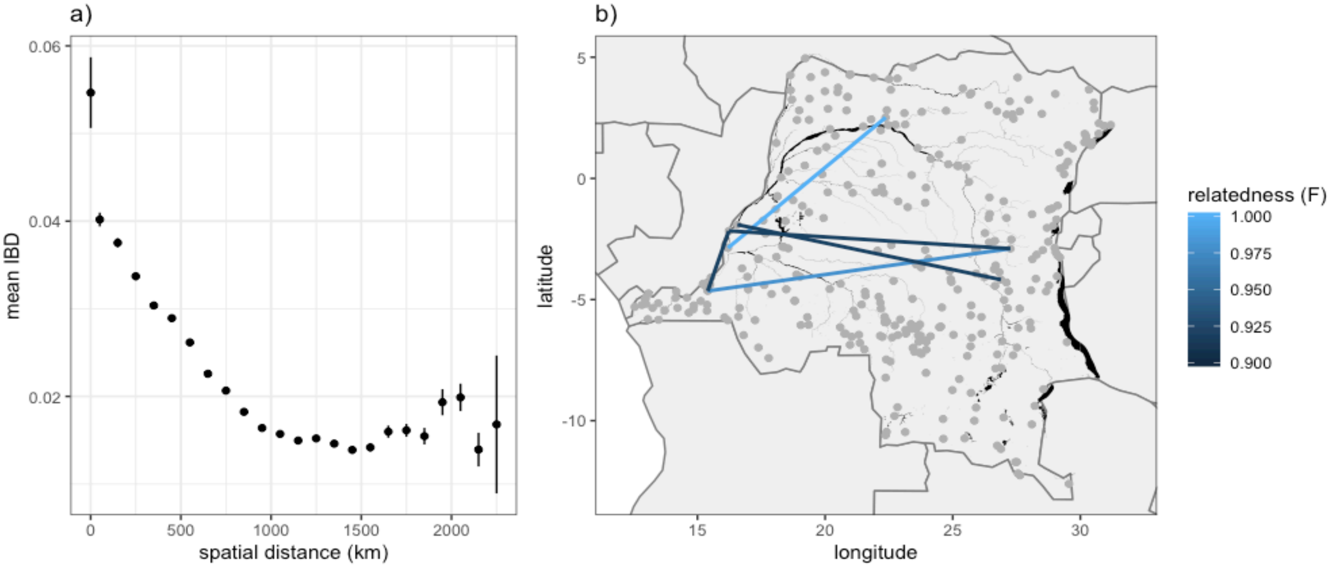

Figure 5 Panel (a) shows the mean IBD between clusters, binned by the spatial distance between clusters. Vertical lines show 95% confidence intervals. Panel (b) shows the spatial distribution of highly related (IBD>0.9) parasite pairs. Black areas indicate major water bodies, including the Congo River.

Prevalence of markers of resistance: Based on previous findings of an east-west divide in molecular markers of antimalarial resistance in the DRC[8,9], all samples in the DRC were divided by geographically-weighted K-means clustering into two populations (**Supplemental Figure 7**). The prevalence of every mutation identified by the drug resistance MIP panel was calculated in eastern and western DRC, as well as at the country level. **Table 1** gives a summary of all mutations that reached a prevalence >5% in any geographic unit, and a complete list of all identified mutations along with their prevalence is given in **Supplemental Table 2**. Note that in the *dhps* mutation **G**437A the reference is resistant, hence this is re-coded as A437**G** and prevalence values indicate the prevalence of the reference allele. Estimated prevalences of these alleles in the DRC as a whole were broadly similar to previously published estimates[10]. However, we did identify several polymorphisms in known and putative resistance genes not previously reported in the DRC, including *kelch* K189**T** and *pfatp6* N569**K**, both of which have been described at appreciable frequencies elsewhere in Africa[18–20].

Geographic distribution of haplotypes: Previous studies have demonstrated that mutations associated with antimalarial resistance are clustered into east-west groupings within DRC[8,10]. Focusing on the 107 samples from DRC that were identified as monoclonal from The REAL McCOIL analysis, we explored the joint distribution of all combinations of mutant haplotypes in both the *dhps* and *crt* genes. Raw combinations of mutations were visualized using the UpSet package in R[21], and the spatial distribution of haplotypes in the DRC was explored by plotting these same mutant combinations against their corresponding DHS cluster locations (**Figure 6**). Our results for *dhps* recapitulate those found previously, showing a clear east-west divide with the K540**E** and A581**G** mutants concentrated in the east, and S436**A** and A437**G** concentrated in the west. For *crt* we also find evidence of an east-west divide, with haplotypes containing N326**S** and F325**C** concentrated in the east and those containing I356**T** concentrated in the west.

				prevalence						
gene	chromosome	position	mutation name	overall	DRC	DRC West	DRC East	Ghana	Uganda	Zambia
atp6	chr1	267007	I723V	1.1	0.3	0.7	0.0	4.2	7.3	0.0
atp6	chr1	267257	G639D	2.0	1.8	2.9	1.0	0.0	7.3	0.0
atp6	chr1	267467	N569K	24.1	21.9	18.8	24.0	16.7	41.5	28.9
atp6	chr1	267882	E431K	15.3	17.0	18.8	15.7	16.7	9.8	6.7
atp6	chr1	267970	L402V	7.1	8.2	10.1	6.9	12.5	0.0	2.2
dhfr-ts	chr4	748239	N51I	83.0	79.5	81.2	78.4	75.0	100.0	97.8
dhfr-ts	chr4	748262	C59R	71.2	63.2	63.0	63.2	95.8	95.1	97.8
dhfr-ts	chr4	748410	S108N	97.8	97.1	97.1	97.1	100.0	100.0	100.0
dhfr-ts	chr4	748577	I164L	3.1	0.6	0.0	1.0	0.0	29.3	0.0
mdr1	chr5	958145	N86Y	12.4	14.3	18.8	11.3	16.7	7.3	0.0
mdr1	chr5	958440	Y184F	37.4	36.5	39.9	34.3	58.3	31.7	37.8
mdr1	chr5	958484	T199S	1.3	0.0	0.0	0.0	0.0	14.6	0.0
mdr1	chr5	958584	S232Y	2.7	3.5	5.1	2.5	0.0	0.0	0.0
mdr1	chr5	961625	D1246Y	4.4	2.9	3.6	2.5	0.0	24.4	0.0
crt	chr7	403620	M74I	30.3	28.7	37.7	22.5	16.7	85.4	0.0
crt	chr7	403621	N75E	30.3	28.7	37.7	22.5	16.7	85.4	0.0
crt	chr7	403625	K76T	30.3	28.7	37.7	22.5	16.7	85.4	0.0
crt	chr7	404407	A220S	28.1	24.6	31.9	19.6	8.3	100.0	0.0
crt	chr7	405600	I356T	7.1	9.4	21.0	1.5	0.0	0.0	0.0
dhps	chr8	549681	S436A	15.0	17.3	28.3	9.8	37.5	0.0	0.0
dhps	chr8	549685	A437G	73.2	67.3	72.5	63.7	95.8	100.0	82.2
dhps	chr8	549993	K540E	25.4	17.0	9.4	22.1	0.0	85.4	48.9
dhps	chr8	550117	A581G	8.2	6.1	2.2	8.8	0.0	34.1	4.4
k13	chr13	1726431	K189T	14.8	14.9	18.8	12.3	54.2	0.0	6.7
mdr2	chr14	1956202	I492V	23.2	21.3	22.5	20.6	20.8	31.7	31.1
mdr2	chr14	1956408	F423Y	31.4	30.1	28.3	31.4	29.2	36.6	37.8

Table 1 Prevalence (%) of mutations identified by the drug resistance MIP panel. Includes all mutations that reached a prevalence >5% in any given geographic unit.

Selective sweep and haplotype analysis: Using the drug resistance MIPs and genome-wide SNP MIPs combined, the extended haplotypes of the monoclonal infections were determined for 200kb upstream and downstream of each putative drug resistance allele that had at least 5% overall prevalence in the DRC. The CV**IET** haplotype within the *crt* gene showed a signal of positive selection, with longer haplotype blocks in western DRC as compared to eastern DRC (**Figure 7**; p'XP-EHH$_D$ < 0.05). In the east, patterns of haplotype homozygosity are consistent with positive selection for the derived I356**T** haplotype (**Supplemental Figure 8**), although a XP-EHH$_D$ statistic could not be calculated for this locus because the derived haplotype was absent in western DRC, supporting the geographic localization of the I356**T** mutation in the east (**Figure 6**).

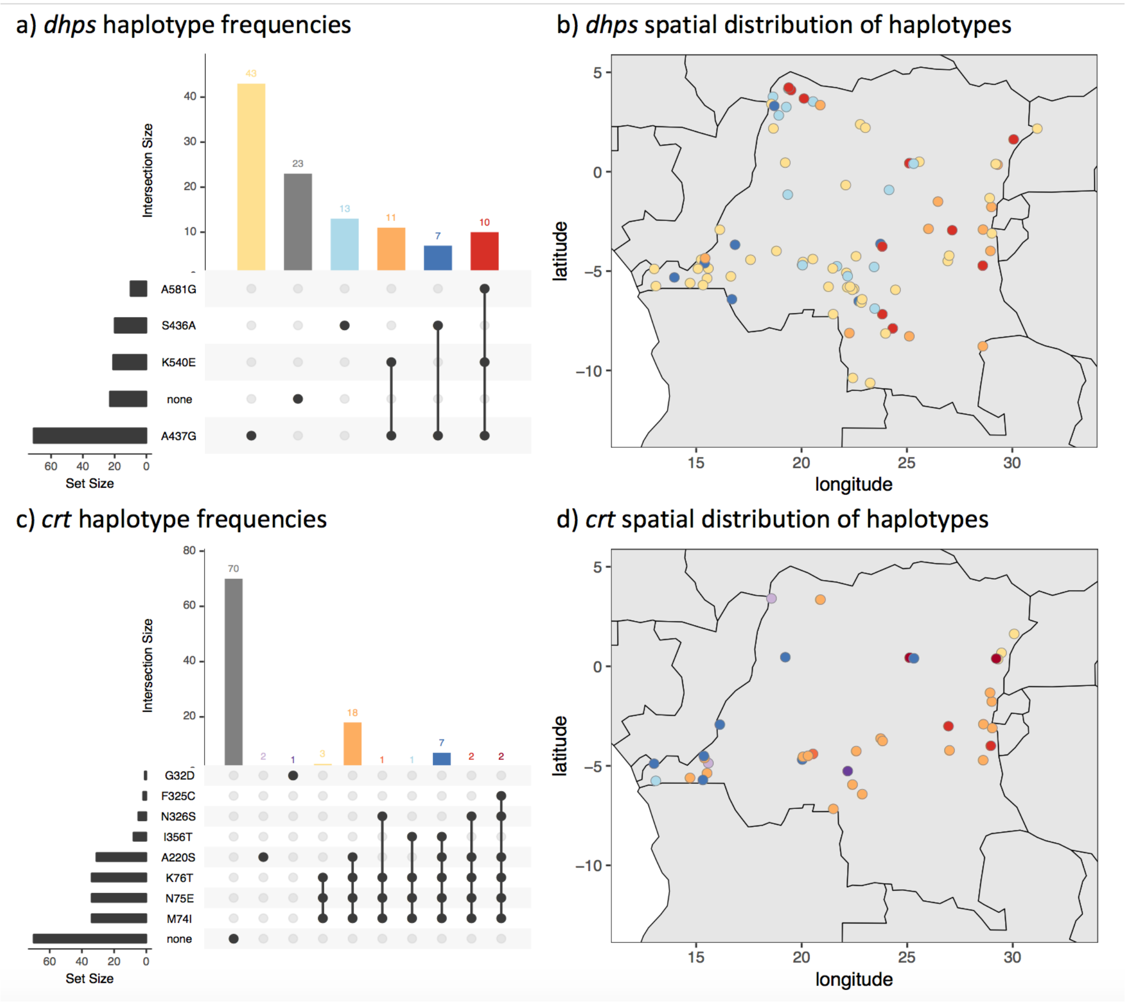

Figure 6 The spatial distribution of all combinations of mutant haplotypes for *dhps* and *crt* from the monoclonal DRC samples. Panels (a) and (c) are UpSet plots showing the number of times each combination of mutations was seen for *dhps* and *crt*, respectively. Panels (b) and (d) show these same haplotypes on a map of DRC. Colours correspond horizontally between panels, i.e. between (a) and (b), and between (c) and (d), with the exception of wild-type haplotypes (grey) which are not shown in panels (b) and (d).

Mutations in *dhps* were more difficult to interpret. This gene has undergone multiple selective sweeps associated with increasing drug resistance. The most recently introduced mutation into the DRC, *dhps* A581**<u>G</u>**, showed relatively conserved local haplotypes around the mutation in both eastern and western DRC (**Supplemental Figure 9**). Extended haplotypes around the other mutations (**Supplemental Figures 10 and 11**) are inconsistent with a classical hard sweep, perhaps due to selection on multiple independent haplotypes or to interference between A581**<u>G</u>** and other linked alleles. Finally, we did not detect any strong signals of differing patterns of recent positive selection between the eastern and western DRC among the *dhfr* and *mdr2* genes (**Supplemental Table 3, Supplementary Figure 12**).

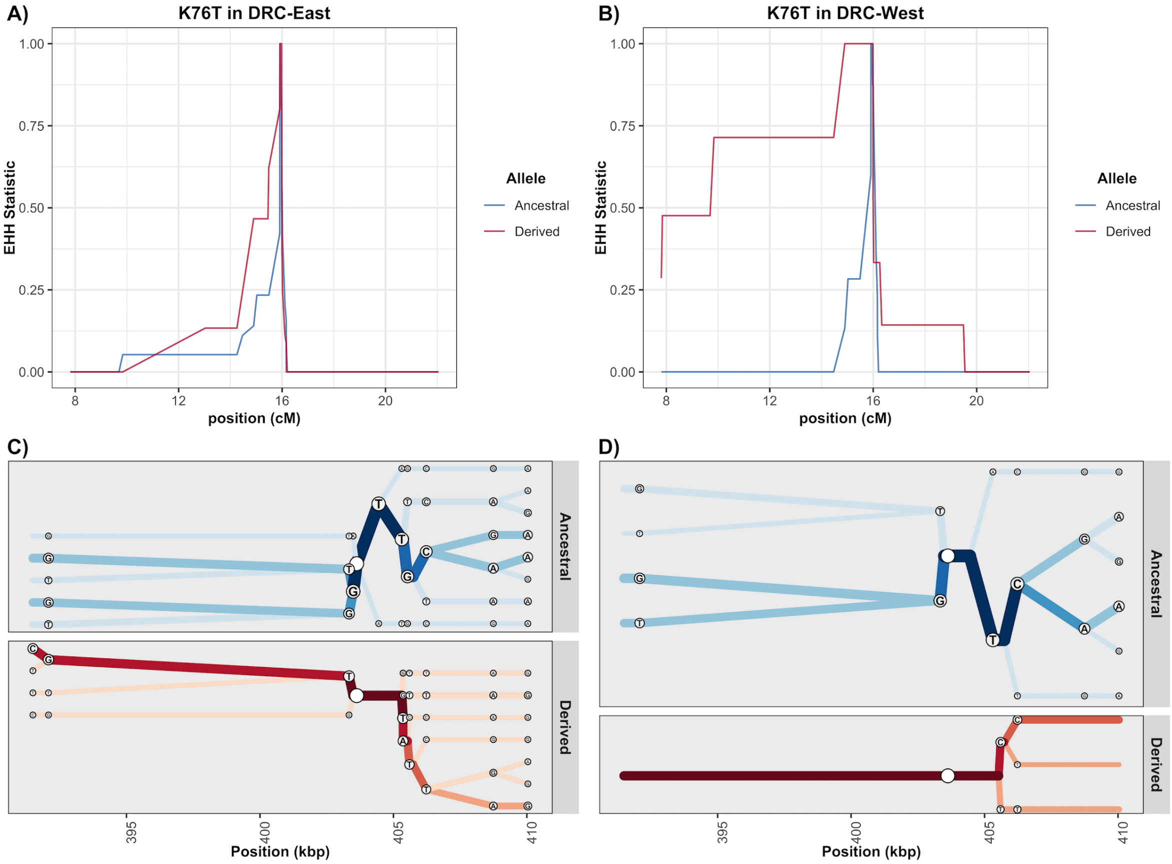

Figure 7. EHH and Bifurcation Plots for *pfcrt* K76T from the monoclonal samples with no missing genotype data. Panels (a) and (b) display EHH curves 200 kb upstream and downstream from the K76T core SNP in centimorgans among the samples from the eastern DRC and western DRC. Panels (c) and (d) show haplotype bifurcation plots with respect to the core allele ancestry and the eastern DRC and western DRC for a subsetted region. Position is considered in kilobases, and segregating sites for each haplotype are displayed at the nodes. Overall, there is strong evidence for recent positive selection of the *pfcrt* CV**IET** haplotype in the west that is mitigated in the east.

DISCUSSION

Here we provide the first large-scale, robustly sampled study of *falciparum* malaria in central Africa using MIP capture and sequencing, a novel high-throughput genotyping approach that is appropriate for large population based surveys. Using a panel of probes designed to detect genome-wide SNPs, combined with a second panel targeting drug resistance genes, we were able to show that the parasite population in the DRC contains a signal of differentiation by geographic separation, consistent with the classical pattern of isolation by distance. This background population structure is overlaid with the clear impacts of drug resistance mutations, which cause distinct structure between East and West African parasite populations. Additionally, the use of relatively dense genome-wide SNPs allowed us to carry out relatedness analysis, revealing a handful of cases where human hosts separated by many hundreds of kilometers were infected by essentially identical clones. Given the rapid breakdown of distinct genotypes by recombination in high transmission areas, it is highly likely that these events represent relatively recent infection and migration events. With this in mind, it is interesting to note that pairwise links of high relatedness tend to fall along the Congo River, an important route of transportation in DRC. Lastly, the combination of the two MIP panels allowed us to examine extended haplotypes surrounding drug resistance genes, revealing rapid breakdown of haplotypes in the population and different signals of selection in East vs. West DRC.

We previously investigated population structure using MIPs targeting 20 microsatellites in the DRC[10], failing to detect a strong signal of population structure based upon these markers. Here we leveraged the same 552 samples as the previous study, plus additional samples from the DRC and neighboring countries, to identify clear structure with an improved SNP-based genotyping method. Our ability to detect population structure in the present study is likely due to several factors. First, the new SNP panel contains nearly two orders of magnitude more markers than the previous panel. While this new SNP MIP panel expanded the number of loci interrogated, we have yet to achieve the full potential of MIPs. Specifically, massively increased, multiplexed probe sets that target additional portions of the genome are feasible. MIPs have now been used in human studies to detect as many as 55,000 markers in a single reaction[22]. Second, a large number of genome-wide SNPs in this study were chosen based on high F_{ST} values in publicly available samples from surrounding countries. This increases our power to detect geographic differentiation, but comes at the cost of not being able to comment on the relative importance of geography vs. drug resistance, which would require random genetic sampling or alternatively whole genomes. Similarly, we should be cautious when interpreting spatial clines in population structure from our data, as we may have greater power to detect structure along some axes than others due to the unequal distribution of surrounding countries in publicly available samples, although in general we have good representation in both the East-West and North-South directions.

The flexible nature of MIP panels allows for multiplex detection of SNPs associated with drug resistance in any known or putative resistance loci for which they are designed. This allowed for a more detailed evaluation of molecular markers associated with antimalarial resistance than has previously been possible in the DRC. To date, studies of antimalarial resistance markers in the DRC have focused primarily on *pfcrt* (K76T), *dhfr* (N51I, C59R, S108N, I164L), *dhps* (I431V, S436A, A437G, K540E, A581G, A613S), *pfmdr* (N86Y, F184Y, D1246Y), and a few *kelch* mutations[23–29]. The data suggests that mutations associated with artemisinin resistance remained absent in the country as of 2014. The World Health Organization identified 9 mutations within the K13 propeller region that are validated in terms of their clinical phenotype of artemisinin resistance, and a further 11 mutations that are candidates associated with the phenotype of delayed clearance.[30] We identified 14 mutations within the K13 gene (**Supplemental Table 2**), although none of these correspond to validated or candidate artemisinin resistance mutations.

Beyond looking at mutations within drug resistance genes, differences in extended haplotypes around drug resistance genes have been used to understand evolution and spread[31]. Though not originally designed for this purpose, the genome wide MIP panel can be leveraged for conducting similar analyses. For example, the differences in CV**IET** EHH between the West and East suggests that the CV**IET** haplotype in the West has potentially been more recently introduced, has experienced less breakdown through recombination, or has undergone stronger recent positive selection as compared to the East. Redesign of the selected targets with denser sampling around known drug resistance genes will allow for more robust assessment of these selected regions.

DRC's location in central Africa and the enormous number of malaria cases in the country means that malaria control in Africa likely depends on improving our understanding on Congolese malaria. This represents the largest study of falciparum population genetics in the DRC and, unlike other large population genetic studies of malaria in Africa, leverages a nationally representative sampling approach. Thus, this study provides the first data on fine-scale genetic structure of parasites at a national scale in Africa, and provides a baseline that can be used to study how implementation programs impact parasite populations in the region. The newly implemented MIP platform represents a highly scalable and cost-effective means of providing genome-wide genetic data, relative to whole genome sequencing [10]. The highly flexible nature of the platform allows it to be rapidly scaled in terms of targets and samples leading it to be applicable across malaria endemic countries.

METHODS

Study Populations: Chelex-extracted DNA from dried blood spots, collected as part of the 2013-2014 DRC Demographic Health Survey (DHS), was tested using quantitative real-time PCR as described previously[32,33]. Previously published DRC samples[10] were included (n=589), and used to set a Ct threshold of <30 which was applied to the remaining DRC samples (n=1450), resulting in a total of 2039 DRC samples sent for sequencing. These samples represented 369 of the overall 539 DHS clusters. In addition, dried blood spot samples from 4 further counties were used: Ghana (n=194), Tanzania (n=120), Uganda (n=63) and Zambia (n=121). Samples from Ghana were collected in 2014 from symptomatic RDT and/or microscopy positive individuals presenting at health care facilities in Begoro (n=94) and Cape Coast (n=98)[34]. Samples from Tanzania were collected in 2015 from symptomatic RDT-positive patients of all ages at Kharumwa Health Center in Northwest Tanzania[35]. Samples from Uganda were collected in 2013 from RDT-positive symptomatic patients at Kanungu in Southwest Uganda[36]. Finally, samples from Zambia were collected in 2013 from RDT positive individuals from a community survey of all ages in Nchelenge District in northeast Zambia on the border with the DRC. All non-DRC samples were Chelex extracted, except for the Ghanaian samples which were extracted using QiaQuick per protocol (Qiagen, Hilden, Germany).

MIP Design: We used two distinct MIP panels - a genome-wide panel designed to capture overall levels of differentiation and relatedness, and a drug resistance panel designed to target polymorphic sites known to be associated with antimalarial resistance. The drug resistance MIP panel has been described previously[10]. When selecting targets for the genome-wide panel, we used the publicly available *P. falciparum* whole genome sequences provided by the Pf3k and *P. falciparum* Community projects from the MalariaGEN Consortium. This consisted of sample sets from Cameroon (n=134), DRC (n=285), Kenya (n=52), Malawi (n=369), Nigeria (n=5), Tanzania (n=66) and Uganda (n=12) (**Supplemental Table 1**). The genomic sequence from these samples underwent alignment, variant calling, and variant-filtering following the Pf3k strategy consistent with the Genome Analysis Toolkit (GATK) Best Practices with minor modifications[37–40]. Full details of the bioinformatic pipeline used in MIP design are given in the **Supplemental Text**. Samples from Nigeria and Uganda were dropped after variant calling due to small sample sizes, and the final filtered sequences were used to calculate Weir and Cochran's F_{ST}[41] with respect to country for each biallelic locus. The 1,000 loci with the highest F_{ST} values were considered for MIP design as phylogeographically informative loci. Of these 1,000 potential loci, 739 were identified as regions that were suitable for MIP-probe design. Separately, from the combined SNP file, we identified 1595 loci that had a minor-allele frequency greater than 5%, had an F_{ST} value between 0.005 and 0.2, and were annotated by SNPEff as functionally silent mutations. These loci were identified as putatively neutral SNPs, and 1151 were found to be suitable for MIP design. The distribution of MIPs is shown in **Supplemental Figure 13** and MIP sequences and targets are shown in **Supplemental Table 4**.

Capture and Sequencing: In addition to patient samples, control samples were known mixtures of 4 strains of genomic DNA from malaria at the following ratios: 67% 3D7 (MRA-102, BEI Resources, Manasas, VA), 14% HB3 (MRA-155), 13% 7G8 (MRA-154) and 6% DD2 (MRA-156). They were also represented at two different parasite densities (29 and 467 parasites/µl). MIP capture and sequencing library preparation were carried out as previously described[10]. Drug resistance libraries were sequenced on Illumina MiSeq instrument using 250 bp paired end sequencing with dual indexing using MiSeq Reagent Kit v2. Genome-wide libraries were sequenced on Illumina Nextseq 500 instrument using 150 bp paired end sequencing with dual indexing using Nextseq 500/550 Mid-output Kit v2. Sequencing reads have been deposited into the NCBI SRA (Accession numbers: pending).

Variant Calling and filtering: Variant calling was performed as described previously[10]. Within each sample, variants were dropped if they had a Phred-scaled quality score of <20. Across samples, variant sites were dropped if they were observed only in one sample, or if they had a total UMI count of less than 5 across all samples. This data set was considered the final raw data used for additional filtering.

Additional filters were applied to both genome-wide and drug resistance datasets prior to carrying out analysis. Sites were restricted to SNPs, and in the case of the genome-wide panel these were filtered to the pre-designed biallelic target SNP sites. Any variant that was represented by a single UMI in a sample, or that had a within-sample allele frequency (WSAF = UMI count/coverage) less than 1%, was eliminated. Any site that was invariant across the entire dataset after this procedure was dropped. Samples were assessed for quality in terms of the proportion of low-coverage sites, where low-coverage was defined as fewer than 10 supporting UMIs. Samples with >50% low-coverage loci were dropped. Variant sites were then assessed by the same means in terms of the proportion of low-coverage samples, and sites with >50% low-coverage samples were dropped. Samples were then combined with metadata, including geographic information, and were only retained if there were at least 10 samples in a given country. This resulted in dropping Tanzanian samples from the drug resistance dataset, but no other countries were dropped. Post-filtering, genome-wide data consisted of 1382 samples (DRC = 1111, Ghana = 114, Tanzania = 30, Uganda = 45, Zambia = 82) and 1079 loci, and drug resistance data consisted of 674 samples (DRC = 557, Ghana = 29, Uganda = 43, Zambia = 45) and 1000 loci.

Complexity of Infection: We applied THE REAL McCOIL categorical method to the SNP genotyped samples to estimate the COI of each individual[13]. Details of the analysis are in the **Supplementary Text**.

Analysis of population structure: WSAFs were calculated for all genome-wide SNPs, with missing values imputed as the mean per locus. Principal component analysis (PCA) was carried out on WSAFs using the *prcomp* function in R version 3.5.1. The relative contribution of each locus was calculated from the loading values as $|l_i|/\sum_{i=1}^{L} |l_i|$, where $|l_i|$is the absolute value of the loading at locus i, and L is the total number of loci. PCA results were explored in a spatial context by taking the mean of the raw principal component values over all samples in a given DHS cluster, and plotting this against the geoposition of the cluster.

Identity by descent analysis: Pairwise identity by descent (IBD) was calculated between all samples from the genome-wide SNPs. We used Malécot's[42] definition of f as the probability of identity by descent, where f_{uv} can be defined as the probability of a randomly chosen locus being IBD between samples u and v. At locus i, let A denote the reference allele, which occurs at population allele frequency p_i, and let a denote the non-reference allele, which occurs at population allele frequency $q_i = 1 - p_i$. Assuming that both samples u and v are monoclonal, let X_{ui} denote the observed allele at locus i in sample u, and equivalently let X_{vi} denote the observed allele in sample v. Then the probabilities of all possible observed allele combinations between the two samples can be written:

$$Pr(X_{ui} = A, X_{vi} = A \mid f_{uv}) = f_{uv}p_i + (1 - f_{uv})p_i^2 \qquad \textbf{(eq1)}$$
$$Pr(X_{ui} = A, X_{vi} = a \mid f_{uv}) = (1 - f_{uv})p_i q_i$$
$$Pr(X_{ui} = a, X_{vi} = A \mid f_{uv}) = (1 - f_{uv})p_i q_i$$
$$Pr(X_{ui} = a, X_{vi} = a \mid f_{uv}) = f_{uv}q_i + (1 - f_{uv})q_i^2$$

from which we can calculate the likelihood of a given value of f_{uv} over all loci as:

$$L(f_{uv} \mid X_u, X_v) = \prod_{i=1}^{L} Pr(X_{ui}, X_{vi} \mid f_{uv}). \qquad \textbf{(eq2)}$$

In practice, population allele frequencies (p_i) were calculated using the mean WSAF for that locus over all samples. Samples were then coerced to monoclonal by calling the dominant allele at every locus. The likelihood was evaluated using **eq2** in log-space for a range of values of f_{uv} distributed between 0 and 1 in equal increments of 0.02. The maximum likelihood estimate $\hat{f}_{uv} = argmax_f \; L(f \mid X_u, X_v)$ was calculated between all sample pairs. Hereafter the terms "IBD" and $\hat{f}_{uv}$ are used interchangeably.

Mean IBD was calculated within and between DHS clusters, and compared using a two-sample t-test. Sample pairs were also binned into groups based on geographic separation (great circle distance) in 100km bins, with an additional bin at distance 0km to capture within-cluster comparisons. Mean and 95% confidence intervals of IBD ware calculated for each group. Finally, sample pairs with IBD>0.9 were identified, and explored in terms of their WSAFs and their spatial distribution.

Estimating mutation prevalence from drug resistance panel: Given previous findings of an East-West divide in molecular markers of antimalarial resistance in the DRC[8,9], all samples in the DRC were divided by geographically-weighted K-means clustering into two populations. The prevalence of every mutation identified by the drug resistance MIP panel was then calculated in East and West DRC, as well as at the country level. Prevalences in each DHS cluster were used to produce smooth prevalence maps using PrevMap version 1.4.2 in R[43], using the method described in Aydemir et. al. (2018)[10].

Analysis of monoclonal haplotypes: Results of the previous COI analysis on the genome-wide SNPs with THE REAL McCOIL were used to identify samples that were monoclonal with a high degree of confidence. Samples were defined as monoclonal if the upper 95% credible interval did not include any COI greater than one. This resulted in 408 monoclonal samples, of which 143 overlapped with the drug resistance MIP dataset and therefore could be used to explore the joint distribution of mutations in drug resistance genes. 107 of these were from DRC. Analysis focussed on the *dhps* and *crt* genes. Raw combinations of mutations were visualized using the UpSet package in R[21], and the spatial distribution of haplotypes was explored by plotting these same mutant combinations against DHS cluster geoposition.

Extended haplotype homozygosity analysis: In order to improve our power to detect hard-sweeps and capture patterns of linkage-disequilibrium with EHH statistics among putative drug resistance SNPs, we combined the genome-wide and the drug resistance filtered biallelic SNPs into a single dataset. Details of this analysis are described in the **Supplemental Text**.

All associated EHH calculations were carried out using the R-package rehh, and were truncated when fewer than two haplotypes were present or the EHH statistic fell below 0.05[44,45]. In addition, we allowed EHH integration calculations to be made without respect to “borders,” which were frequent due to the MIP-probe design. Although this would result in an inflated integration statistic if the EHH statistic had not yet reached 0 within the region of investigation, this problem was mitigated by only comparing between subpopulations, and not between loci. EHH decay, bifurcation plots, and haplotype plots were adapted from the rehh package objects and modified using ggplot[46]**.**

Acknowledgements: This work was supported by the National Institutes of Health (R01AI107949, R01AI139520, K24AI134990, R21AI121465, F30AI143172, U19AI089680). RV is funded by a Skills Development Fellowship: this award is jointly funded by the UK Medical Research Council (MRC) and the UK Department for International Development (DFID) under the MRC/DFID Concordat agreement and is also part of the EDCTP2 programme supported by the European Union. The authors would also like to thank everyone who participated in the studies and all members of the study teams in Ghana, Zambia, Uganda and Tanzania. The authors would like to thank the DHS Program and USAID for the collection and access to samples from the DRC Demographic Health Survey.

Ethics Approval: This study was approved by the Internal Review Board at UNC and the Ethics Committee of the Kinshasa School of Public Health.

Authors Contributions: R.V., O.A., N.F.B., J.A.B. and J.J.J contributed data analysis, writing and experimental design. O.J.W. contributed data analysis and writing. N.J.H. and A.P.M. contributed software design. M.K.M, J.P., M.C., P.J.R., P.T., D.S.I., J.G., J.N., D.E.N., W.M., M.M., J.L.M.H., A.G., B.M., and A.K.T. contributed samples from studies conducted at their sites and reviewed the manuscript. A.C.G. contributed to analysis design and reviewed the manuscript. K.T., P.K.M., T.F., and M.D. contributed laboratory analysis. S.R.M. contributed coordination with DRC investigators, experimental design and writing.

Competing Interests: None

REFERENCES

1. WHO | World malaria report 2017. (2018).
2. Neafsey, D. E. & Volkman, S. K. Malaria Genomics in the Era of Eradication. *Cold Spring Harb. Perspect. Med.* **7**, (2017).
3. Malaria. *Bill & Melinda Gates Foundation* Available at: https://www.gatesfoundation.org/What-We-Do/Global-Health/Malaria. (Accessed: 2nd May 2019)
4. World Health Organization. *WHO: High Burden to High Impact. A targeted malaria response 2019.*
5. Pearce, R. J. *et al.* Multiple origins and regional dispersal of resistant dhps in African Plasmodium falciparum malaria. *PLoS Med.* **6**, e1000055 (2009).
6. Ocholla, H. *et al.* Whole-genome scans provide evidence of adaptive evolution in Malawian Plasmodium falciparum isolates. *J. Infect. Dis.* **210**, 1991–2000 (2014).
7. Carrel, M. *et al.* The geography of malaria genetics in the Democratic Republic of Congo: A complex and fragmented landscape. *Soc. Sci. Med.* **133**, 233–241 (2015).
8. Taylor, S. M. *et al.* Plasmodium falciparum sulfadoxine resistance is geographically and genetically clustered within the DR Congo. *Sci. Rep.* **3**, 1165 (2013).
9. Antonia, A. L. *et al.* A cross-sectional survey of Plasmodium falciparum pfcrt mutant haplotypes in the Democratic Republic of Congo. *Am. J. Trop. Med. Hyg.* **90**, 1094–1097 (2014).
10. Aydemir, O. *et al.* Drug Resistance and Population Structure of Plasmodium falciparum Across the Democratic Republic of Congo using high-throughput Molecular Inversion Probes. *J. Infect. Dis.* (2018). doi:10.1093/infdis/jiy223
11. Verity, R. *et al.* Plasmodium falciparum genetic variation of var2csa in the Democratic Republic of the Congo. *Malar. J.* **17**, 46 (2018).
12. O'Roak, B. J. *et al.* Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science* **338**, 1619–1622 (2012).
13. Chang, H.-H. *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.* **13**, e1005348 (2017).
14. Bethke, L. L. *et al.* Duplication, gene conversion, and genetic diversity in the species-specific acyl-CoA synthetase gene family of Plasmodium falciparum. *Mol. Biochem. Parasitol.* **150**, 10–24 (2006).
15. Taylor, A. R., Jacob, P. E., Neafsey, D. E. & Buckee, C. O. Estimating relatedness between malaria parasites. doi:10.1101/575985
16. Rousset, F. Genetic differentiation and estimation of gene flow from F-statistics under isolation by

distance. *Genetics* **145**, 1219–1228 (1997).

17. J., A., Crow, J. F. & Kimura, M. An Introduction to Population Genetics Theory. *Population (French Edition)* **26**, 977 (1971).
18. Talundzic, E. *et al.* Molecular Epidemiology of Plasmodium falciparum kelch13 Mutations in Senegal Determined by Using Targeted Amplicon Deep Sequencing. *Antimicrob. Agents Chemother.* **61**, (2017).
19. Torrentino-Madamet, M. *et al.* Limited polymorphisms in k13 gene in Plasmodium falciparum isolates from Dakar, Senegal in 2012–2013. *Malaria Journal* **13**, 472 (2014).
20. Dahlström, S. *et al.* Diversity of the sarco/endoplasmic reticulum Ca(2+)-ATPase orthologue of Plasmodium falciparum (PfATP6). *Infect. Genet. Evol.* **8**, 340–345 (2008).
21. Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R. & Pfister, H. UpSet: Visualization of Intersecting Sets. *IEEE Trans. Vis. Comput. Graph.* **20**, 1983–1992 (2014).
22. Turner, E. H., Lee, C., Ng, S. B., Nickerson, D. A. & Shendure, J. Massively parallel exon capture and library-free resequencing across 16 genomes. *Nature Methods* **6**, 315–316 (2009).
23. Mvumbi, D. M. *et al.* Falciparum malaria molecular drug resistance in the Democratic Republic of Congo: a systematic review. *Malaria Journal* **14**, (2015).
24. Leroy, D. *et al.* African isolates show a high proportion of multiple copies of the Plasmodium falciparum plasmepsin-2 gene, a piperaquine resistance marker. *Malar. J.* **18**, 126 (2019).
25. Nkoli Mandoko, P. *et al.* Prevalence of Plasmodium falciparum parasites resistant to sulfadoxine/pyrimethamine in the Democratic Republic of the Congo: emergence of highly resistant pfdhfr/pfdhps alleles. *J. Antimicrob. Chemother.* **73**, 2704–2715 (2018).
26. Baraka, V. *et al.* Impact of treatment and re-treatment with artemether-lumefantrine and artesunate-amodiaquine on selection of Plasmodium falciparum multidrug resistance gene-1 polymorphisms in the Democratic Republic of Congo and Uganda. *PLoS One* **13**, e0191922 (2018).
27. Ruh, E., Bateko, J. P., Imir, T. & Taylan-Ozkan, A. Molecular identification of sulfadoxine-pyrimethamine resistance in malaria infected women who received intermittent preventive treatment in the Democratic Republic of Congo. *Malar. J.* **17**, 17 (2018).
28. Mvumbi, D. M. *et al.* Molecular surveillance of Plasmodium falciparum resistance to artemisinin-based combination therapies in the Democratic Republic of Congo. *PLoS One* **12**, e0179142 (2017).
29. Taylor, S. M. *et al.* Absence of putative artemisinin resistance mutations among Plasmodium falciparum in Sub-Saharan Africa: a molecular epidemiologic study. *J. Infect. Dis.* **211**, 680–688 (2015).

30. World Health Organization. *Status report on artemisinin and ACT resistance*. (2017).
31. Project, M. P. F. C. & MalariaGEN Plasmodium falciparum Community Project. Genomic epidemiology of artemisinin resistant malaria. *eLife* **5**, (2016).
32. et Suivi, M. du P. de la Mise en œuvre de la Révolution de la Modernité (MPSMRM), Ministere de la Santé Publique (MSP) and ICF International. 2014. *Enquête Démographique et de Santé en République Démocratique du Congo* **2014**, (2013).
33. Pickard, A. L. *et al.* Resistance to antimalarials in Southeast Asia and genetic polymorphisms in pfmdr1. *Antimicrob. Agents Chemother.* **47**, 2418–2423 (2003).
34. Abuaku, B. K. *et al.* Efficacy of Artesunate/Amodiaquine in the Treatment of Uncomplicated Malaria among Children in Ghana. *Am. J. Trop. Med. Hyg.* **97**, 690–695 (2017).
35. Ngondi, J. M. *et al.* Surveillance for sulfadoxine-pyrimethamine resistant malaria parasites in the Lake and Southern Zones, Tanzania, using pooling and next-generation sequencing. *Malar. J.* **16**, 236 (2017).
36. Tumwebaze, P. *et al.* Changing Antimalarial Drug Resistance Patterns Identified by Surveillance at Three Sites in Uganda. *J. Infect. Dis.* **215**, 631–635 (2017).
37. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
38. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10.1–33 (2013).
39. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
40. The Pf3K Project. *www.malariagen.net/data/pf3k-5* (2016).
41. Weir, B. S. & Cockerham, C. C. ESTIMATING F-STATISTICS FOR THE ANALYSIS OF POPULATION STRUCTURE. *Evolution* **38**, 1358–1370 (1984).
42. Malécot, G. *The Mathematics of Heredity*. (W.H. Freeman, 1970).
43. Giorgi, E. & Diggle, P. J. PrevMap: An R Package for Prevalence Mapping. *Journal of Statistical Software* **78**, (2017).
44. Gautier, M., Klassmann, A. & Vitalis, R. rehh 2.0: a reimplementation of the R package rehh to detect positive selection from haplotype structure. *Mol. Ecol. Resour.* **17**, 78–90 (2017).
45. Gautier, M. & Vitalis, R. rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics* **28**, 1176–1177 (2012).
46. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer New York, 2009).