Microgeographic epidemiology of malaria parasites in an irrigated area of western Kenya by deep amplicon sequencing

Authors: Elizabeth Hemming-Schroeder[1,2], Daibin Zhong[1], Solomon Kibret [1], Amanda Chie[1], Ming-Chieh Lee[1], Guofa Zhou[1], Harrysone Atieli[3], Andrew Githeko[4], James W. Kazura[2], Guiyun Yan (guiyuny@hs.uci.edu)[1]*

1 Program in Public Health, University of California, Irvine, CA 92697, USA

2 Center for Global Health and Diseases, Case Western Reserve University, Cleveland, Ohio 44106, USA

3 School of Public Health and Community Development, Maseno University, Kenya

4 Center for Global Health Research, Kenya Medical Research Institute, Kisumu, Kenya

* Corresponding author

Summary: This study highlights the usefulness of targeted deep sequencing of the *cpmp* amplicon in revealing malaria transmission patterns at a microgeographic spatial scale (< 20 km diameter) and the impacts that irrigated agriculture has on the epidemiology of malaria parasites.

Abstract

To improve food security, investments in irrigated agriculture are anticipated to increase throughout Africa. However, the extent that environmental changes from water resource development will impact malaria epidemiology remains unclear. This study was designed to both compare the sensitivity of molecular markers used in deep amplicon sequencing for evaluating malaria transmission intensities; and assess malaria transmission intensity at various proximities to an irrigation scheme. Compared to *ama1, csp,* and *msp1* amplicons, *cpmp* required the smallest sample size to detect differences in infection complexity between transmission risk zones. Transmission intensity was highest within five kilometers of the irrigation scheme by PCR positivity rate, infection complexity, and linkage disequilibrium. The irrigated area provided a source of parasite infections for the surrounding 2–10 km area. This study highlights the suitability of the *cpmp* amplicon as a measure for transmission intensities and the impact of irrigation on microgeographic epidemiology of malaria parasites.

Keywords: *Plasmodium falciparum;* irrigation; transmission intensity; amplicon sequencing; microgeographic epidemiology; infection complexity

Introduction

Development of irrigation schemes has been linked to an increase in malaria risk by providing more habitats for mosquito breeding, enhancing habitat stability, and/or providing vectors to bridge transmission seasons [1]. For example, irrigated rice fields in a sub-arid region of Madagascar were found to have a transmission rate 150 times higher than in the original ecosystem; additionally, the region transformed from a seasonal transmission area to a perennial transmission area [2]. However, there are examples where there has been no observed impact or even a reduction in malaria prevalence following water resource development projects [3]. For instance, the Diama Dam in Senegal had no impact on malaria transmission despite an increase in mosquito vector abundance [4, 5]. In contrast, the construction of small dams in the Sundergarh district of India led to a decrease in local malaria prevalence [6], which was attributed to increased flow conditions that rendered mosquito vector breeding unfavorable in the river downstream of the dams. Thus, the impacts that water resource development projects have on malaria risk are likely context specific, depending on the local epidemiologic setting.

Malaria epidemiological studies can be enhanced by leveraging data from the genetic diversity of parasites. Individuals in malaria-endemic areas often harbor multiple, genetically distinct strains, known as polyclonal infections. Polyclonal infections can result from two processes: 1) superinfection, which occurs from subsequent bites by multiple, infected mosquitoes, each carrying a unique parasite genotype; or 2) co-transmission, when an individual is bitten by a single mosquito, in which the mosquito is carrying multiple parasite genotypes [7]. Furthermore, when multiple parasite genotypes are present in a mosquito host during the sexual stage, a high rate of genetic recombination occurs [8], generating novel local genotypes [7, 9], and thus leading to more opportunities for polyclonal infections to occur. Therefore, the number of parasite strains co-infecting a single host, called the multiplicity of infection (MOI), is expected to positively correlate with the intensity of malaria transmission because of increased opportunities for superinfection and co-transmission to occur in high transmission areas [7, 10].

While MOI may be a useful indicator for transmission intensity, in some studies, there has been no observed positive correlation between infection complexity and malaria parasite prevalence [11, 12]. However, inconclusive correlations between MOI and transmission intensity may be in part due to the choice of genotyping method and/or molecular marker(s), which can affect the ability to detect differences in MOI at varying transmission intensities. PCR-based genotyping, such as by size-polymorphic antigens or microsatellites, often underestimates MOI, as it is challenging to detect minority strains, which result in faint bands or weak fluorescent signals [13]. In contrast, deep amplicon sequencing has been shown to be more sensitive for detecting MOI [14-16]. Moreover, choice of molecular marker for amplicon sequencing, which can vary in terms of the number of single nucleotide polymorphisms (SNPs), can also impact the ability to assess MOI [14, 16], and thus potentially the sensitivity to assess differences in MOI between varying transmission intensities. Recently, Lerch et al. [17] demonstrated that *cpmp* (PF3D7_0104100, “conserved *Plasmodium* membrane protein”) detected a higher mean MOI and contained more SNPs than common molecular markers within parasites originating from Papua New Guinea.

Therefore, the objectives of this study were the following: 1) examine the level of sequence polymorphism found in *cpmp, ama1*, *csp,* and *msp1* markers in *Plasmodium falciparum* by deep amplicon sequencing; 2) evaluate *P. falciparum* malaria transmission intensity at various proximities to an irrigation scheme by PCR positivity rate, MOI, and other molecular indices; 3) compare the sensitivity of molecular markers in detecting differences in

MOI among the various transmission intensity zones; and 4) assess patterns of genetic connectivity among parasite isolates among irrigated transmission risk zones.

Methods

Ethics statement. Scientific and ethical clearance was obtained from the institutional scientific ethical review board of University of California, Irvine, USA and Maseno University, Kenya. Written informed consent/assent for study participation was obtained from all consenting heads of households, parents/guardians (for minors under age of 18), and each individual who was willing to participate in the study

Study design. This study was conducted in the Oluch Irrigation Scheme (latitude 0°26'44"S, longitude 34°31'28.0"E, elevation 1200 m) and its vicinity in Homa Bay County, western Kenya (Figure 1). This area falls within the lake-endemic zone, which has year-round malaria transmission with seasonal variability [18]. To build sustainable agriculture and reduce poverty, the Kimira-Oluch Smallholder Farm Improvement Project was funded by the African Development Bank. As part of this project, the Oluch irrigation scheme was initiated in 2007 and completed in 2015. Clusters from the Oluch irrigated scheme and surrounding areas were assigned to the following classes: “high”, i.e., within the irrigated area, “medium”, i.e., within 2 km area from the boundary of the irrigation scheme, and “low”, i.e., > 5 km from the boundary of the irrigation scheme. Cluster radii varied from 0.25 to 1.0 km.

Sample collection. Finger prick blood samples were collected during two cross-sectional surveys in Feb/Mar and Jun/Jul of 2018. The Feb/Mar collection period occurs after the short rains, whereas Jun/Jul occurs after the long rains.. A random subset of residents, ranging from 19–86 per cluster, were selected for participation in this study. Study participants reported having no malaria symptoms at the time of collection, so that only parasites from subclinical infections were examined. Whatman 3MM filter papers were used for collecting and storing ~50 µl of blood. DNA was extracted and purified from dried blood spots by the Saponin/Chelex method [19]. A multiplexed Taqman probe assay was performed to identify malaria parasite species [20, 21]. Of those, 91 samples that were infected with the *P. falciparum* parasite species only were randomly selected for evaluation by deep amplicon sequencing.

Amplicon deep sequencing. Amplification and sequencing of the molecular markers *ama1, cpmp, csp,* and *msp1* were performed based on previously published primers and protocols [14, 16, 17, 22, 23] with modifications. Briefly, the sequencing library was generated by three rounds of PCR. The primary PCR amplified each relevant molecular marker. The second round was a nested, marker-specific amplification primers that carried a 5' linker sequence. PCR products of each molecular marker were then pooled by sample, and the third round of PCR was completed using primers with sample-specific barcode sequences to allow for pooling of samples and subsequent de-multiplexing. *Plasmodium falciparum* laboratory strains HB3 (MRA-155G), DD2 (MRA-150G), and 3D7 (MRA-102G) were used as controls in duplicate at the ratios 1:0:0, 7:3:0, and 6:3:1. Thirty samples were amplified in duplicate to assess potential PCR or sequencing errors. Amplicons were cleaned and normalized to 1ng/µl concentration using the SequalPrep Normalization Plate Kit (ThermoFisher Scientific, Inc., Waltham, MA, USA). Sequencing was performed on an Illumina Miseq using a MiSeq reagent kit v3 PE300 (UCI Genomics High-Throughput Facility) with PhiX control (Illumina, PhiXControl v3). PCR protocols are described further in the supplementary material, including primer sequences (Supplementary Table 1), PCR reaction mixtures (Supplementary Table 2), and thermocycling conditions (Supplementary Table 3).

Bioinformatic analysis. Paired ends were joined with fastq-join (parameters: 8% maximum percent difference and 20 minimum overlap). Pooled reads were de-multiplexed by molecular marker with seqkit grep by matching ten base pairs at the 3' end of the forward marker-specific primers. Adapter and primer sequences, as well as low quality base pairs (parameter: error threshold 0.01) were subsequently trimmed by seqtk trimfq. Additionally, de-multiplexed samples by molecular marker yielding < 25 reads sequencing coverage were excluded. Haplotypes were determined with the use of SeekDeep software [24]. Specifically, Seekdeep qluster was run (parameter: illumina) to create haplotypes with relative abundances by collapsing reads on specific errors. Next, for replicate comparison and final results filtering, SeekDeep processClusters was run (parameters: strictErrors, illumina, fracCutOff 0.035, clusterCutOff 3, and hq 1), in which the parameters allowed for a few low quality mismatches but no indels and one high quality mismatch, cleared out low abundant clusters (3.5% and lower), and removed clusters of size three or less. These parameters resulted in consistent and accurate haplotype calling among positive controls and sample replicates for all molecular markers.

Data analysis. All data analyses were performed in R unless otherwise noted. The statistical significance of differences in PCR positivity rate among risk zones was assessed by chi-square test with a Bonferroni correction to limit the Type 1 error rate. The 95% confidence intervals were estimated by the modified Wald method. The significance of MOI differences among risk zones was evaluated by Wilcoxon rank-sum tests with a Bonferroni correction due to the non-normal distribution of MOI values. Sample size simulations were carried out by drawing from a Poisson distribution [25, 26] with lambda equal to the average MOI by molecular marker and risk zone. Each simulation was performed with 10,000 replicates. Nucleotide diversity (Pi) and linkage disequilibrium (LD) were assessed for *cpmp* amplicons in DNA Sequence Polymorphism v5 by coalescent simulations (parameters: free recombination and 1000 replicates) [27]. Differences in Pi and LD were assessed by the pairwise comparison method assuming normal distributions. Proportion of shared alleles was measured by *cpmp* amplicons using the R package *adegenet* [28]. Isolates which shared greater than 98% of polymorphic alleles (≤ 1 nucleotide difference), henceforth referred to as highly related pairs, were visualized and evaluated for patterns of allele sharing. Differences in proportions of highly related pairs among risk zones was evaluated by a chi-square test with a false discovery rate correction. Differences in geographic distances between highly related isolates and all isolates were assessed by fit of ANOVA followed by TukeyHSD post-hoc analysis. Finally, to assess the magnitude and directionality of parasite migration among risk zones, a Bayesian inference method based on coalescent theory was implemented in Migrate-N 3.7.2 [29, 30]. Parameters estimated from genetic data include mutation-scaled immigration rate (M), mutation-scaled population size (θ), and the number of effective migrants per generation (Nm, calculated as θM/2). Ten independent runs were conducted with a burn-in of 10^6 steps, sampling increment of 10 steps, and 10^6 recorded steps in each chain for a total of 10^7 visited parameter values.

Results

PCR positivity rate among transmission risk zones. In total, 2112 samples were evaluated for the presence of *Plasmodium* malaria parasites from 21 study sites within and surrounding the Oluch irrigation scheme (Figure 1). In Feb/Mar, the PCR positivity rate was significantly higher in the high risk zone (21.0% [74/353]), as compared to the medium (11.1% [45/407]) and low (11.0% [32/290]) zones (X^2 = 13.31, P = .0006 and X^2 = 10.69, P = .002, respectively) (Figure 2). In Jun/Jul, the PCR positivity rate was highest in the medium zone (18.5% [65/352]) followed by high (16.2% [81/500]) and low (14.8% [31/210]), but pairwise differences were not significant ($X^2 \leq 1.03$, $P \geq .93$ for all comparisons) (Figure 2).

Sequence reads and haplotype determination. A random subset of *P. falciparum* infected samples were selected for evaluation by deep amplicon sequencing. A total of 116 pooled PCR reactions (85 samples, 25 replicate PCR reactions, and 6 controls) were successfully amplified, sequenced, and joined, resulting in 515,952 reads. Average depth of reads per sample was 4634 (range: 206 – 9560). Of the four molecular markers, sequencing by *cpmp* detected the highest number of unique haplotypes (# hap. = 78), the highest percentage of polymorphic infections (49.1%), and highest average MOI (2.02) (Table 1; Supplementary Figure 1). These trends were consistent when comparing values amongst all samples and matched samples only (Supplementary Table 4). Average MOI was generally highest in the medium risk zone and lowest in the low risk zone across molecular markers (Table 1).

Sample size simulations. To assess the power of discriminating between transmission intensities by MOI estimated from sequencing of the four molecular markers, sample sizes (up to 5,000) were simulated for each molecular marker and compared. Only by sequencing of *cpmp* or *msp1,* was an average unadjusted P-value ≤ .0167 achieved among 10,000 simulations when comparing high vs. low and medium vs. low risk zones at sample size ≤ 66 per risk zone (Figure 3A). The smallest sample size required to obtain significant differences between the high and low risk zone in ≥ 80% of 10,000 simulations was lowest for *cpmp* (45 at α = .0167; 33 at α = .05) (Figure 3B). Likewise, sequencing by *cpmp* required the lowest sample size to achieve significant differences between the medium and low risk zone (25 at α = .0167; 19 at α = .05) (Figure 3B). Detecting significant differences between the high and medium risk zones required sample sizes exceeding 150 for all molecular markers, but was lowest for *csp* (261 at α = .0167; 196 at α = .05) (Figure 3B). Lastly, to evaluate how simulated P-values compared to empirical P-values from this study, we visualized the distribution of P-values derived from 10,000 simulations at sample sizes equivalent to the empirical sample sizes (Figure 3C). Eight of the twelve empirical P-values were within the middle 50% of P-values from respective simulations (Figure 3C).

Comparison of genetic indices among transmission risk zones by *cpmp*. Since amplicon sequencing by *cpmp* detected the highest average MOI and required the lowest sample sizes to detect significant differences in MOI in two of three comparisons by simulated data, subsequent analyses were based on *cpmp* amplicons exclusively. Average MOI was significantly higher in the medium risk zone than in the low risk zone (P = 0.004; Wilcoxon rank-sum test) (Figure 4A). MOI did not vary significantly between the high and medium risk zone or between the high and low risk zone (P = .74 and P = .06, respectively). Notably, MOI did not vary significantly among other risk factors, including gender (P = 1), age group (P ≥ .27 for all comparisons), and season of collection (P = 1) (Supplementary Figure 2). Average nucleotide diversity did not differ significantly among transmission risk zones based on confidence intervals obtained from coalescent simulations (P > .05) (Figure 4B). LD was significantly lower in the high and medium risk zones (0.020 and 0.022, respectively) as compared to the low risk zone (0.077) (P = .05) (Figure 4C).

Allele sharing and patterns of relatedness among parasite isolates by *cpmp*. To assess patterns of genetic connectivity, highly related isolate pairs (sharing > 98% of alleles) were identified and mapped by geographic locality (Figure 5A). Four pairs of highly related isolates originated from the same cluster, and so are not visible on the map: High (1 pair), Low (2 pairs), and Medium (1 pair). The most common allele sharing pattern was between isolates originating from the high and medium risk zones (Figure 5B). Moreover, geographic distance alone did not explain this allele sharing pattern (Figure 5C), as geographic distance did not significantly differ among highly related pairs, identical pairs, and all pairs ($P \geq .13$ for all comparisons; unpaired t-test with Bonferroni correction). Finally, an analysis of migration rates revealed that the highest magnitude of migration among risk zones occurred in the direction of the high to the medium risk zone (Nm = 25.4), which was followed by medium to low risk zone (Nm = 18.8), and then high to low risk zone (Nm = 17.4), indicating that areas closer to the irrigation scheme provide sources of parasites for surrounding areas (Figure 6).

Discussion

This study examined the utility of four molecular markers (*ama1*, *cpmp*, *csp*, and *msp1)* used for deep amplicon sequencing in evaluating malaria transmission intensities and assessed malaria transmission intensity and genetic connectivity among various proximities to an irrigation scheme. We found that amplicon sequencing by *cpmp* had the highest sensitivity to detect transmission intensity differences based on MOI. Additionally, we found that indicators of transmission intensity were highest within 5 km of the irrigation scheme, as compared to 5–10 km from the irrigation scheme. This finding is in agreement with previous studies in Africa that documented significantly higher malaria intensity at close proximity (< 5 km) to irrigation dams than those located further away (> 5 km) [31-34]. Furthermore, based on *cpmp* amplicon data, we demonstrated that the area within 2 km of the irrigation scheme provides a source of parasite infections for the surrounding 2–10 km area. These findings highlight the value of the *cpmp* amplicon in studying microgeographic epidemiology of malaria and that irrigated agriculture promotes a source of parasite infections for surrounding areas in this epidemiologic setting.

Quantifying malaria transmission intensity can be done through several indices, such as the traditional metrics of entomological inoculation rate (EIR) or clinical incidence. However, measuring EIR is particularly labor-intensive [35], and neither of these metrics account for polymorphic infections. Thus, genotyping-based metrics provide an appealing alternative or supplement to quantify malaria transmission intensity. In particular, MOI is an attractive indicator for assessing transmission intensity, as it requires a modest sample size [36] and can be measured from a single time point [26]. Consistent with a study in Papua New Guinea [17], we found that the *cpmp* amplicon detected the highest average MOI among common molecular markers by deep amplicon sequencing. Furthermore, we demonstrated that this enhanced sensitivity to detect MOI resulted in requiring a smaller sample size than alternative markers to observe differences among transmission risk zones in this study. Therefore, MOI assessed by *cpmp* amplicon sequencing is a relatively low cost metric, requiring a rather low sampling effort.

Additionally, we showed that deep sequencing by *cpmp* can be useful for revealing source-sink parasite dynamics [29, 30]. Identifying reservoirs or factors that promote reservoirs of malaria parasites can be used to plan effective interventions for malaria control and elimination [36]. Traditional molecular markers, such as microsatellites or SNP barcodes, are also useful for inferring patterns of genetic connectivity [37-39]. However, constructing haplotypes from these multilocus markers is not feasible for polymorphic infections. As a result, in areas of high transmission, where polymorphic infections are predominant, a large amount of data is loss or thrown out when using common multilocus markers. Thus, deep amplicon sequencing provides a solution to tracking malaria parasites and identifying source-sink patterns without losing information from polymorphic infections.

Across sub-Saharan Africa, irrigation has been blamed for intensifying malaria in areas with unstable disease transmission [31, 40]. There is evidence demonstrating that irrigation schemes create conducive microclimates for malaria mosquitoes to breed [41, 42] and promote longevity in adult mosquitoes [43]. Behaviorally, people living close to irrigated fields spend more time outside on their field during the early hours of the night where they are not protected by indoor-based control measures (e.g. bed nets, indoor residual spraying) while the mosquitoes are active [44, 45]. Increased malaria transmission intensity coupled with outdoor behavior could potentially compromise the efficacy of existing malaria intervention efforts in the irrigated areas that mainly rely on indoor-based control tools. As Africa is planning to achieve malaria elimination in many of its regions by 2030 [46], such localized transmission pockets could challenge intervention efforts in the region. Thus, additional vector control strategies are critically needed to address outdoor transmission. Integrated vector management by incorporating larval management through irrigation canal management may help reduce malaria around irrigated settings as suggested in previous field studies [47, 48].

This study had certain limitations. PCR positivity rate and MOI, like all indicators of malaria transmission, have inherent biases. For example, parasite positivity rate is affected by acquired immunity and antimalarial drug use, and it does not capture multiple infections [26]. MOI is also mediated by acquired immunity and antimalarial drug use. Additionally, MOI is limited by the diversity of parasite populations, i.e. where parasite populations are less diverse, MOI may be an underestimate of transmission because multiple clones are indistinguishable [49]. The ability to detect MOI is limited by the selected limit of detection, which was conservatively set at 3.5% for this study. As a result, it is likely we missed minority genotypes below this threshold, but this trade-off is balanced by filtering out potential noise from PCR and sequencing errors [13]. We found that decreasing this threshold increased variability among replicates. Lowering the limit of detection may be possible by decreasing the number of PCR cycles and increasing the amount of starting genetic material instead to limit potential PCR errors. With that said, the overall trends observed in this study should not be affected by these limitations.

To conclude, we demonstrated the importance of selecting a highly polymorphic molecular marker used for amplicon sequencing for detecting differences in malaria transmission intensities. We found that deep sequencing of *cpmp* required the smallest sample size of the four molecular markers to detect significant differences in MOI among transmission risk zones. Transmission intensity by all indices was lowest 5–10 km from irrigation scheme compared to < 5 km from the irrigation scheme. Moreover, the high migration rates from the irrigated area to surrounding areas suggests that the Oluch irrigation scheme is promoting a source of parasite infections for nearby areas. These findings highlight the usefulness of sequencing by *cpmp* in detecting patterns of malaria transmission at a microgeographic spatial scale (< 20 km study area), as well as the impacts that irrigated agriculture has on microgeographic epidemiology of malaria parasites.

Funding

This work was supported by the National Institutes of Health [grant numbers U19AI129326, R01AI050243, D43TW001505, and F32AI147460].

Acknowledgments

We thank the study participants and the field team of the ICEMR project in Kenya for their involvement in this study. We thank Nicole Hemming-Schroeder for her contribution in coding the sample size simulations.

Conflict of interests

The authors declare that they have no conflict of interests.

Data availability statement

The sequences for this project have been deposited at GenBank under accession numbers: MT818243 – MT818353.

Footnotes

Potential conflicts of interest. All authors: No reported conflicts of interest.

Funding statement. This work was supported by grants from the National Institutes of Health (NIH) (U19AI129326, R01AI050243, D43TW001505, and F32AI147460).

Correspondence author. Guiyun Yan (guiyuny@hs.uci.edu), Department of Ecology and Evolutionary Biology and Program in Public Health, University of California, Irvine, CA 92617, USA; phone: 1-949-824-0175; fax: 1-949-824-0249

References

1. Kibret S, Wilson GG, Ryder D, Tekie H, Petros B. The influence of dams on malaria transmission in sub-Saharan Africa. EcoHealth **2017**; 14:408-19.

2. Marrama L, Jambou R, Rakotoarivony I, et al. Malaria transmission in Southern Madagascar: influence of the environment and hydro-agricultural works in sub-arid and humid regions: part 1. Entomological investigations. Acta tropica **2004**; 89:193-203.

3. Baudon D, Robert V, Darriet F, Huerre M. Impact of building a dam on the transmission of malaria. Malaria survey conducted in southeast Mauritania. Bulletin de la Societe de pathologie exotique et de ses filiales **1986**; 79:123-9.

4. Sanchez-Ribas J, Parra-Henao G, Guimarães AÉ. Impact of dams and irrigation schemes in Anopheline (Diptera: Culicidae) bionomics and malaria epidemiology. Revista do Instituto de Medicina Tropical de São Paulo **2012**; 54:179-91.

5. Sow S, De Vlas S, Engels D, Gryseels B. Water-related disease patterns before and after the construction of the Diama dam in northern Senegal. Annals of Tropical Medicine & Parasitology **2002**; 96:575-86.

6. Sharma SK, Tyagi PK, Upadhyay AK, Haque MA, Adak T, Dash AP. Building small dams can decrease malaria: a comparative study from Sundargarh District, Orissa, India. Acta tropica **2008**; 107:174-8.

7. Nkhoma SC, Trevino SG, Gorena KM, et al. Co-transmission of Related Malaria Parasite Lineages Shapes Within-Host Parasite Diversity. Cell Host & Microbe **2020**; 27:93-103. e4.

8. Kolakovich KA, Ssengoba A, Wojcik K, et al. Plasmodium vivax: favored gene frequencies of the merozoite surface protein-1 and the multiplicity of infection in a malaria endemic region. Experimental parasitology **1996**; 83:11-8.

9. Mu J, Awadalla P, Duan J, et al. Recombination hotspots and population structure in Plasmodium falciparum. PLoS biology **2005**; 3.

10. Mideo N, Bailey JA, Hathaway NJ, et al. A deep sequencing tool for partitioning clearance rates following antimalarial treatment in polyclonal infections. Evolution, medicine, and public health **2016**; 2016:21-36.

11. Fola AA, Harrison GA, Hazairin MH, et al. Higher complexity of infection and genetic diversity of Plasmodium vivax than Plasmodium falciparum across all malaria transmission zones of Papua New Guinea. The American journal of tropical medicine and hygiene **2017**; 96:630-41.

12. Getachew S, To S, Trimarsanto H, et al. Variation in complexity of infection and transmission stability between neighbouring populations of Plasmodium vivax in Southern Ethiopia. PloS one **2015**; 10.

13. Zhong D, Koepfli C, Cui L, Yan G. Molecular approaches to determine the multiplicity of Plasmodium infections. Malaria journal **2018**; 17:172.

14. Zhong D, Lo E, Wang X, et al. Multiplicity and molecular epidemiology of Plasmodium vivax and Plasmodium falciparum infections in East Africa. Malaria journal **2018**; 17:185.

15. Lalremruata A, Jeyaraj S, Engleitner T, et al. Species and genotype diversity of Plasmodium in malaria patients from Gabon analysed by next generation sequencing. Malaria journal **2017**; 16:398.

16. Lerch A, Koepfli C, Hofmann NE, et al. Development of amplicon deep sequencing markers and data analysis pipeline for genotyping multi-clonal malaria infections. BMC genomics **2017**; 18:864.

17. Lerch A, Koepfli C, Hofmann NE, et al. Longitudinal tracking and quantification of individual Plasmodium falciparum clones in complex infections. Scientific reports **2019**; 9:1-8.

18. Kenya Malaria Operational Plan. Vol. FY 2019. pmi.gov: President's Malaria Initiative, **2019**.

19. Bereczky S, MÅrtensson A, Gil JP, FÄrnert A. Rapid DNA extraction from archive blood spots on filter paper for genotyping of Plasmodium falciparum. The American journal of tropical medicine and hygiene **2005**; 72:249-51.

20. Shokoples SE, Ndao M, Kowalewska-Grochowska K, Yanow SK. Multiplexed real-time PCR assay for discrimination of Plasmodium species with improved sensitivity for mixed infections. Journal of clinical microbiology **2009**; 47:975-80.

21. Veron V, Simon S, Carme B. Multiplex real-time PCR detection of P. falciparum, P. vivax and P. malariae in human blood samples. Experimental parasitology **2009**; 121:346-51.

22. Snounou G, Zhu X, Siripoon N, et al. Biased distribution of msp1 and msp2 allelic variants in Plasmodium falciparum populations in Thailand. Transactions of the Royal Society of Tropical Medicine and Hygiene **1999**; 93:369-74.

23. Neafsey DE, Juraska M, Bedford T, et al. Genetic diversity and protective efficacy of the RTS, S/AS01 malaria vaccine. New England Journal of Medicine **2015**; 373:2025-37.

24. Hathaway NJ, Parobek CM, Juliano JJ, Bailey JA. SeekDeep: single-base resolution de novo clustering for amplicon deep sequencing. Nucleic acids research **2018**; 46:e21-e.

25. Dietz K, Wernsdorfer W, McGregor I. Mathematical models for transmission and control of malaria. Principles and Practice of Malariology **1988**:1091-133.

26. Tusting LS, Bousema T, Smith DL, Drakeley C. Measuring changes in Plasmodium falciparum transmission: precision, accuracy and costs of metrics. Advances in parasitology. Vol. 84: Elsevier, **2014**:151-208.

27. Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics **2009**; 25:1451-2.

28. Jombart T, Ahmed I. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics **2011**; 27:3070-1.

29. Beerli P. How to use MIGRATE or why are Markov chain Monte Carlo programs difficult to use. Population genetics for animal conservation **2009**; 17:42-79.

30. Beerli P, Mashayekhi S, Sadeghi M, Khodaei M, Shaw K. Population Genetic Inference With MIGRATE. Current protocols in bioinformatics **2019**; 68:e87.

31. Kibret S, Lautze J, McCartney M, Wilson GG, Nhamo L. Malaria impact of large dams in sub-Saharan Africa: maps, estimates and predictions. Malaria journal **2015**; 14:339.

32. Lautze J, McCartney M, Kirshen P, Olana D, Jayasinghe G, Spielman A. Effect of a large dam on malaria risk: the Koka reservoir in Ethiopia. Tropical Medicine & International Health **2007**; 12:982-9.

33. Kibret S, Wilson GG, Tekie H, Petros B. Increased malaria transmission around irrigation schemes in Ethiopia and the potential of canal water management for malaria vector control. Malaria journal **2014**; 13:360.

34. Kyei-Baafour E, Tornyigah B, Buade B, et al. Impact of an Irrigation Dam on the Transmission and Diversity of Plasmodium falciparum in a Seasonal Malaria Transmission Area of Northern Ghana. Journal of Tropical Medicine **2020**; 2020.

35. Dye C. The analysis of parasite transmission by bloodsucking insects. Annual review of entomology **1992**; 37:1-19.

36. Neafsey DE, Volkman SK. Malaria genomics in the era of eradication. Cold Spring Harbor perspectives in medicine **2017**; 7:a025544.

37. Koepfli C, Mueller I. Malaria epidemiology at the clone level. Trends in parasitology **2017**; 33:974-85.

38. Lo E, Hemming-Schroeder E, Yewhalaw D, et al. Transmission dynamics of co-endemic Plasmodium vivax and P. falciparum in Ethiopia and prevalence of antimalarial resistant genotypes. PLoS neglected tropical diseases **2017**; 11:e0005806.

39. Lo E, Lam N, Hemming-Schroeder E, et al. Frequent spread of Plasmodium vivax malaria maintains high genetic diversity at the Myanmar-China border, without distance and landscape barriers. The Journal of infectious diseases **2017**; 216:1254-63.

40. Keiser J, de Castro MC, Maltese MF, et al. Effect of irrigation and large dams on the burden of malaria on a global and regional scale. The American journal of tropical medicine and hygiene **2005**; 72:392-406.

41. Muriuki JM, Kitala P, Muchemi G, Njeru I, Karanja J, Bett B. A comparison of malaria prevalence, control and management strategies in irrigated and non-irrigated areas in eastern Kenya. Malaria journal **2016**; 15:402.

42. Hawaria D, Demissew A, Kibret S, Lee M-C, Yewhalaw D, Yan G. Effects of environmental modification on the diversity and positivity of anopheline mosquito aquatic habitats at Arjo-Dedessa irrigation development site, Southwest Ethiopia. Infectious diseases of poverty **2020**; 9:9.

43. Lu Z-x, Yu X-p, Heong K-l, Cui H. Effect of nitrogen fertilizer on herbivores and its stimulation to major insect pests in rice. Rice Science **2007**; 14:56-66.

44. Kibret S, Wilson G. Increased outdoor biting tendency of Anopheles arabiensis and its challenge for malaria control in Central Ethiopia. public health **2016**; 141:143-5.

45. Yohannes M, Boelee E. Early biting rhythm in the afro- tropical vector of malaria, Anopheles arabiensis, and challenges for its control in Ethiopia. Medical and veterinary entomology **2012**; 26:103-5.

46. Organization WH. Global technical strategy for malaria 2016-2030. World Health Organization, **2015**.

47. Keiser J, Utzinger J, Singer BH. The potential of intermittent irrigation for increasing rice yields, lowering water consumption, reducing methane emissions, and controlling malaria in African rice fields. Journal of the American Mosquito Control Association **2002**; 18:329-40.

48. Van Den Berg H, Von Hildebrand A, Ragunathan V, Das PK. Reducing vector-borne disease by empowering farmers in integrated vector management. Bulletin of the World Health Organization **2007**; 85:561-6.

49. Mueller I, Schoepflin S, Smith TA, et al. Force of infection is key to understanding the epidemiology of Plasmodium falciparum malaria in Papua New Guinean children. Proceedings of the National Academy of Sciences **2012**; 109:10030-5.

Table and Figure Legends

Table 1. Comparison of infection complexity among molecular markers, Kenya.

Figure 1. Study site localities in relation to the Oluch irrigation scheme, Kenya.

Figure 2. PCR positivity rate among transmission risk zones and season, Kenya. Square indicates overall positivity rate among clusters by risk zone. Lines indicate 95% confidence intervals. Asterisks indicate statistical significance (chi-square test with Bonferroni correction; P ≤ .05). Feb/Mar occurs after the short rains (Oct-Dec), and Jun/Jul occurs after the long rains (Mar-May).

Figure 3. Simulated sample sizes to assess power of hypothesis testing by molecular markers, Kenya. (A) Mean unadjusted P-value by simulated sample size. Dashed red line indicates α = 0.167. Dotted red line indicates α = 0.05. (B) Minimum sample size for unadjusted P-value to be less than α in ≥ 80% of simulations. (C) Empirical P-value compared to distribution of simulated P-values at equivalent sample sizes. Diamonds indicate empirical values. Box plots indicate distribution of simulated values. Lower and upper hinges correspond to the first and third quartiles. All results are based on Wilcoxon rank-sum tests among 10,000 simulations per sample size. Sample sizes were simulated from Poisson distributions with λ = empirical average MOI for a given molecular marker and transmission zone.

Figure 4. Genetic indices among transmission risk zones by *cpmp* amplicon sequencing. (A) Multiplicity of infection (MOI) among risk zones. Dots indicate individual data points. Triangles

indicate average values. (B) Nucleotide diversity among risk zones. (C) Linkage disequilibrium among risk zones. Squares indicate average values among 1000 coalescent simulations. Lines indicate 95% confidence interval from simulations. Asterisks indicate statistical significance.

Figure 5. Allele sharing among parasite isolates by *cpmp* amplicon sequencing, Kenya. (A) Highly related isolates by geographic locality. Dots indicate individual parasite isolates. Lines connect isolates from different clusters which share > 98% of polymorphic alleles. Thin lines indicate 98– 99.9% allele sharing; thick lines indicate 100% allele sharing. (B) Highly related pairs by risk zones. Bar height indicates the proportion of highly related pairs among all isolate pairs for each risk zone combination. Labels indicate the total count number of highly related pairs for each risk zone combination. For x-axis labels, letters indicate risk zone combinations: H = High; M = Med.; L = Low. Asterisks indicate statistical significance (chi-square test with FDR correction, $P \leq .05$). Four pairs of highly related isolates originated from the same cluster, and so are not visible on the map: HH (1 pair), LL (2 pairs), and MM (1 pair). (C) Histogram of highly related pairs by geographic distance between pairs. Solid, green line indicates density plot. Vertical, dashed line indicates mean geographic distance between isolate pairs. “All” indicates geographic distances between all isolates pairs. Geographic distances were not significantly different between allele sharing categories (unpaired t-test with Bonferroni correction, $P \geq .13$ for all comparisons).

Figure 6. Migration rates among risk zones by *cpmp* sequencing, Kenya. Thickness of arrows is proportional to the estimated mutation-scaled migration rate (M). Diameter of circles is proportional to the estimated mutation-scaled population sizes (θ). Numbers indicate the

estimated effective number of migrants per generation (Nm, calculated as θM/2) between populations. Black arrows indicate predominant direction of migration.

Figure 1

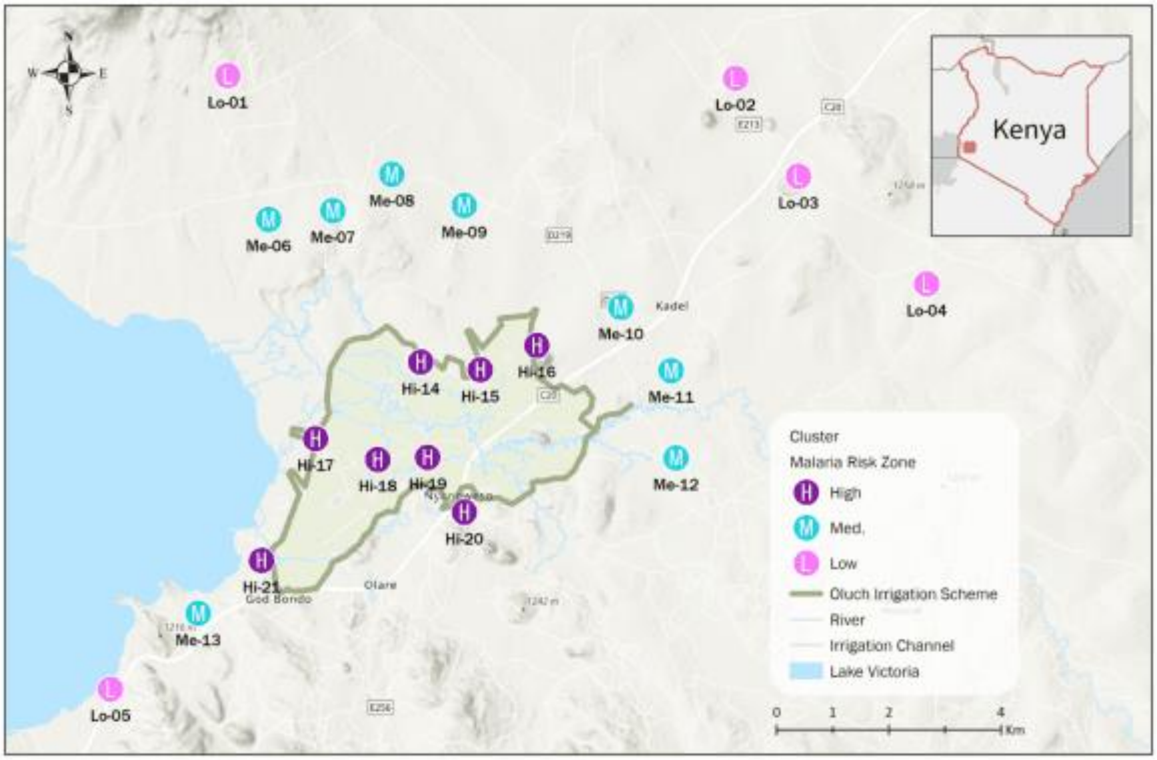
N
W
E
S
Lo-01
Lo-02
E213
C20
Lo-03
Me-06
Me-07
Me-08
Me-09
D219
Me-10
Kadel
Lo-04
Hi-14
Hi-15
Hi-16
C20
Me-11
Hi-17
Hi-18
Hi-19
Nyamwaso
Hi-20
Me-12
Hi-21
God Bondo
Olare
1242 m
Me-13
1216 m
Lo-05
E256
Kenya
Cluster
Malaria Risk Zone
H High
M Med.
L Low
Oluch Irrigation Scheme
River
Irrigation Channel
Lake Victoria
0
1
2
4
Km

Figure 2

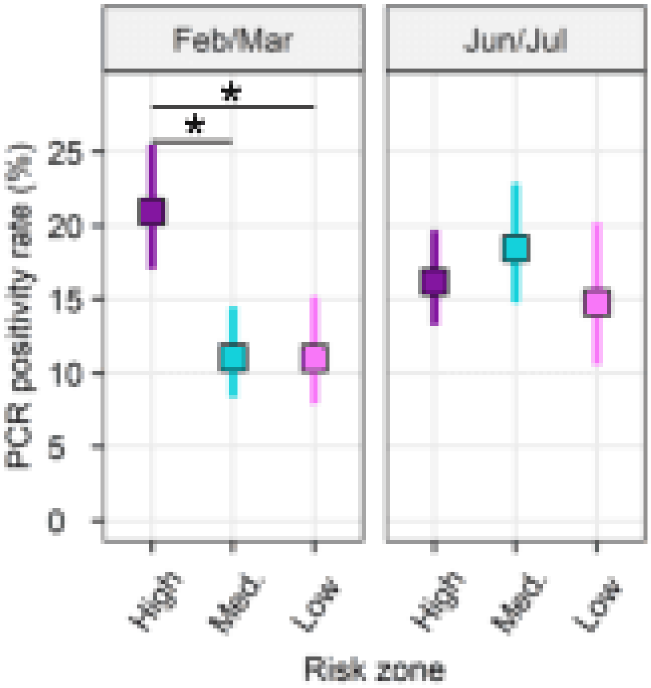
Feb/Mar
Jun/Jul
PCR positivity rate (%)
0
5
10
15
20
25
*
*
High
Med
Low
High
Med
Low
Risk zone

Figure 3

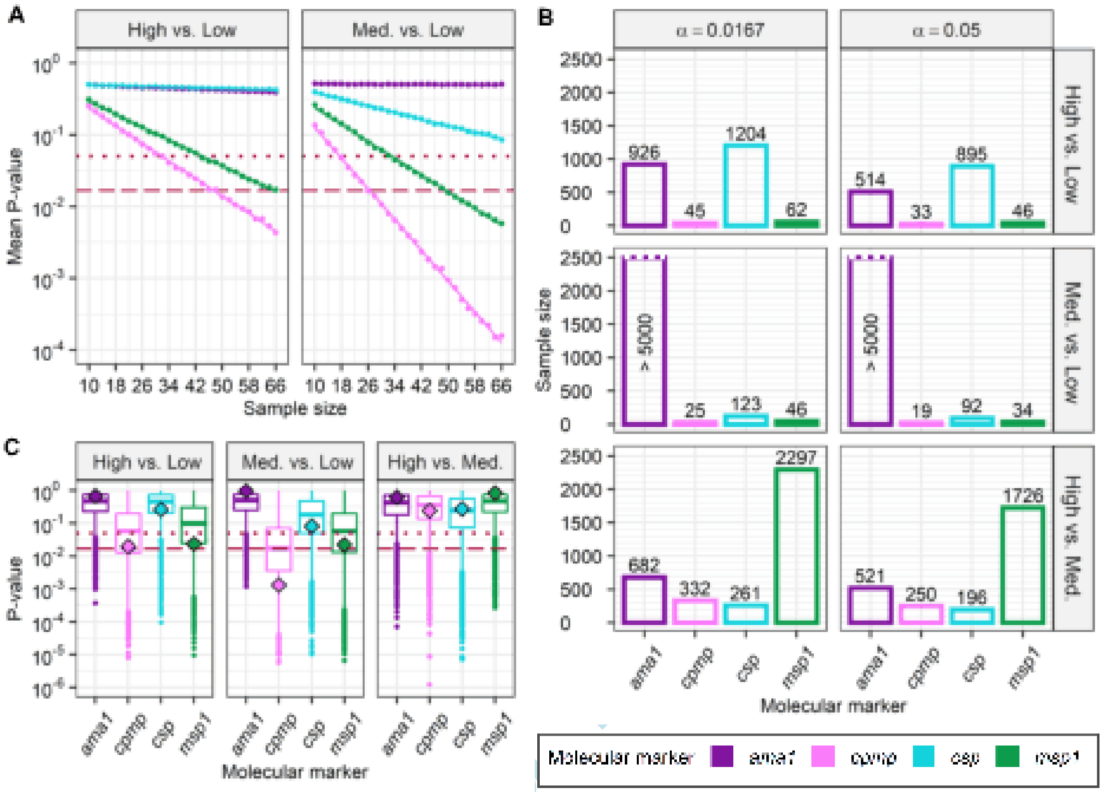
A
High vs. Low
Med. vs. Low
Mean P-value
10^0
10^-1
10^-2
10^-3
10^-4
10 18 26 34 42 50 58 66
Sample size
B
α = 0.0167
α = 0.05
High vs. Low
Med. vs. Low
High vs. Med.
Sample size
2500
2000
1500
1000
500
0
926
45
1204
62
514
33
895
46
> 5000
25
123
46
19
92
34
682
332
261
2297
521
250
196
1726
ama1
cpmp
csp
msp1
Molecular marker
C
High vs. Low
Med. vs. Low
High vs. Med.
P-value
10^0
10^-1
10^-2
10^-3
10^-4
10^-5
10^-6
ama1
cpmp
csp
msp1
Molecular marker
Molecular marker ■ ama1 ■ cpmp ■ csp ■ msp1

Figure 4

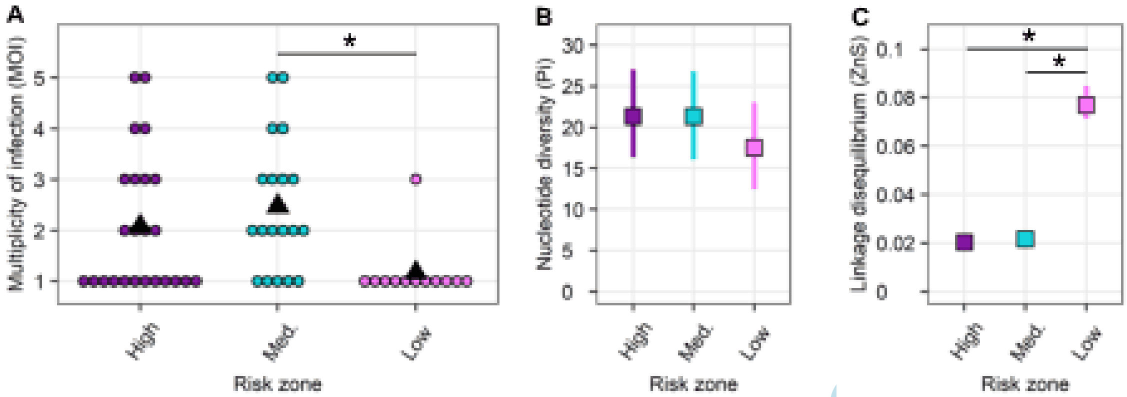
A
Multiplicity of infection (MOI)
1
2
3
4
5
*
High
Med
Low
Risk zone
B
Nucleotide diversity (Pi)
0
5
10
15
20
25
30
High
Med
Low
Risk zone
C
Linkage disequilibrium (ZnS)
0
0.02
0.04
0.06
0.08
0.1
*
*
High
Med
Low
Risk zone

Figure 5

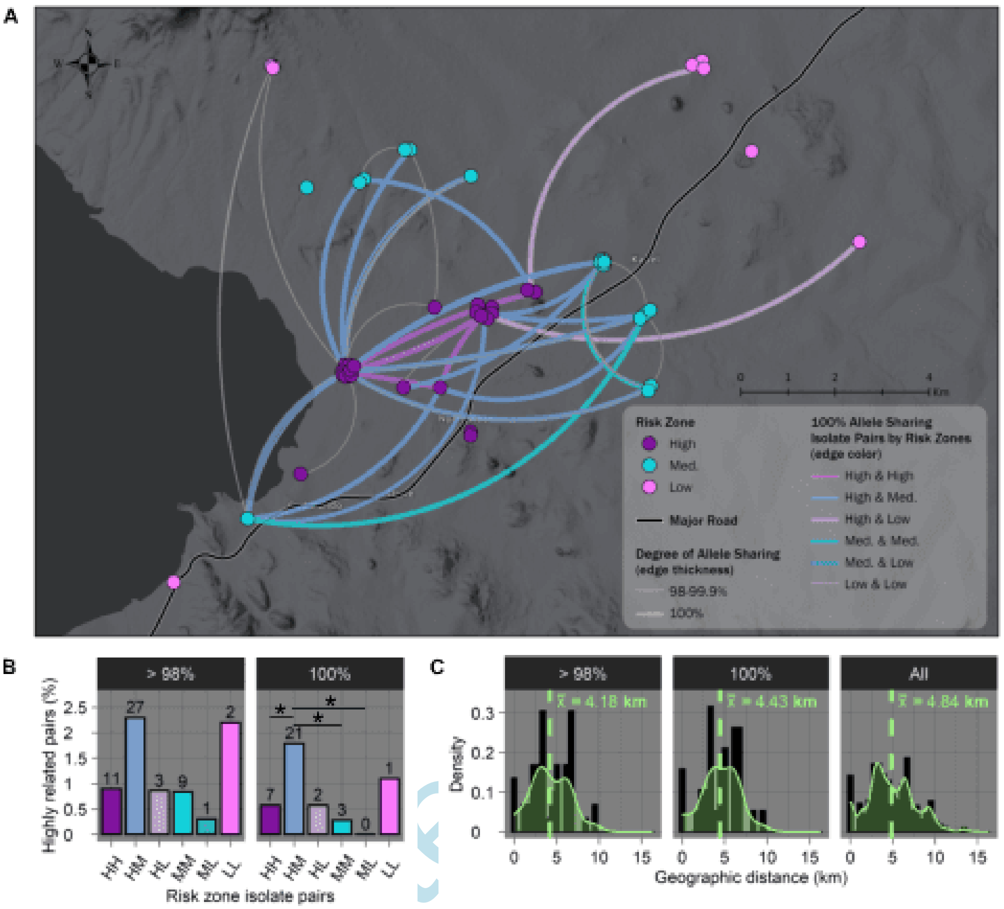
A
N
W
E
S
0
1
2
4
Km
Risk Zone
High
Med.
Low
Major Road
Degree of Allele Sharing (edge thickness)
98-99.9%
100%
100% Allele Sharing Isolate Pairs by Risk Zones (edge color)
High & High
High & Med.
High & Low
Med. & Med.
Med. & Low
Low & Low
B
> 98%
100%
Highly related pairs (%)
2.5
2
1.5
1
0.5
0
11
27
3
9
1
2
7
21
2
3
0
1
*
*
*
HH
HM
HL
MM
ML
LL
Risk zone isolate pairs
C
> 98%
100%
All
x̄ = 4.18 km
x̄ = 4.43 km
x̄ = 4.84 km
Density
0.3
0.2
0.1
0
0
5
10
15
Geographic distance (km)

Figure 6

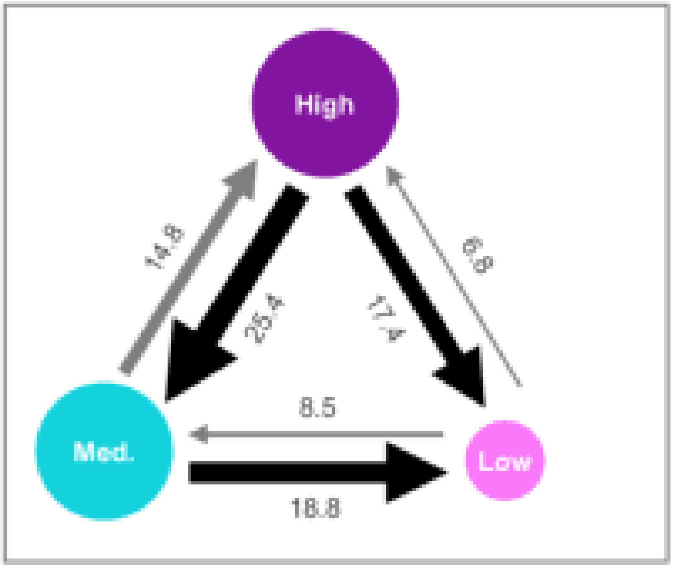

Table 1. Comparison of infection complexity among molecular markers and transmission risk zones, Kenya.

				Avg. MOI			
	n	# Hap.	% Poly.	High	Med.	Low	Overall
ama1	46	20	34.8	1.41	1.63	1.60	1.52
cpmp	55	78	49.1	2.08	2.47	1.17	2.02
csp	56	39	41.6	1.61	2.00	1.44	1.71
msp1	77	55	42.9	1.91	2.05	1.15	1.79
comb.	85	--	55.3	2.24	2.55	1.55	2.19

"n" indicates sample size. "# hap." indicates total number of unique haplotypes detected. "Avg. MOI" indicates average multiplicity of infection (MOI) among samples. "% poly." indicates the percentage of samples which had > 1 haplotype (polyclonal). "comb." indicates pooled values for all molecular markers, i.e. the maximum MOI per sample is obtained of the four molecular markers.