Assessing the population genetic structure and demographic history of *Anopheles gambiae* and *An. arabiensis* at island and mainland populations in Uganda: Implications for testing novel malaria vector control approaches.

Rita Mwima [1,2], Tin-Yu J. Hui [3], Edward Lukyamuzi [1], Marilou Bodde [4], Alex Makunin [4], Krystal Birungi[1], Martin Lukindu[1], Ann Nanteza[2], Dennis Muhanguzi[2], Mara Lawniczak [4], Austin Burt [3], Jonathan K. Kayondo [1]

1 Department of Entomology, Uganda Virus Research Institute (UVRI), Entebbe, Uganda.
2 Department of Biotechnical and Diagnostic Sciences, College of Veterinary Medicine, Animal Resources and Biosecurity (COVAB), Makerere University, Uganda.
3 Silwood Park Campus, Department of Life Sciences, Imperial College London, Ascot, UK.
4 Tree of Life, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire, UK.

Abstract

Even with enormous effort put into malaria control, it remains a global burden, disproportionally affecting people residing in areas such as Uganda and many other countries in sub-Saharan Africa. Innovative tools for malaria control are being developed, such as a gene drive system to suppress the vectors' populations, requiring knowledge of the population genetics of malaria vectors, including their population structure, size and genetic diversity to ensure effective and targeted implementation.

This study collected 2918 *Anopheles gambiae* and 173 *Anopheles. arabiensis* across six populations from both the islands on Lake Victoria and mainland Uganda for amplicon sequencing. Large pairwise F_{ST} values were observed between the two species, indicating their divergence. We observed low but often significant F_{ST} values between the 6 *An. gambiae* populations, while between the *An. arabiensis* mainland populations, F_{ST} values were not significant. Principal Component Analysis also revealed strong genetic structure between the two species but did not provide a clear picture between populations within ~~a~~ each species. We also found that mainland *An. gambiae* populations had higher within population genetic diversity than the islands', while *An. arabiensis* had the lowest nucleotide diversity. Tajima's D values were all negative, suggesting a recent population expansion. The islands *An. gambiae* populations had very low contemporary effective population sizes in the tens and hundreds, as estimated from linkage disequilibrium, while the mainland population sizes were consistently higher, in the thousands.

This study revealed that there is significant genetic differentiation between *An. gambiae* and *An. arabiensis*, as well as between island and mainland populations of *An. gambiae*, with mainland populations showing higher connectivity and island populations exhibiting greater isolation. In contrast, *An. arabiensis* mainland populations displayed no significant differentiation, suggesting panmixia. These findings thus highlight the influence of geographic and ecological factors on population structure and provide critical insights for selecting candidate sites and designing field trials for genetic-based malaria control strategies.

Keywords: *An. gambiae*, *An. arabiensis*, population genetics, genetic structure, Single Nucleotide Polymorphisms (SNPs).

Background

Even with enormous progress over the years, malaria remains a global burden, especially in the Sub-Saharan Africa because of the exceptionally adaptive vector dynamics of the *Anophelines* among other factors such as insecticide resistance, drug resistance, limited access to healthcare, and environmental and socio-economic challenges. [1,2]. The World Health Organisation (WHO) reported about 200 million cases and 600,000 in 2023 [3] 249 million malaria cases in 2022 and 608,000 deaths from 85 malaria-endemic countries in 2022 [4], and similarly 247 million cases and 619,000 deaths in 2021 [5]. Sub-Saharan Africa carries the greatest burden globally and accounts for 94% of all malaria cases and about 96% of all deaths [3,4]. Uganda is one of the most highly burdened countries in sub-Saharan Africa and ranked third in the number of malaria cases worldwide in 2022, contributing 5% of the global burden [4].

The fight against malaria is increasingly hampered by the evolution of resistance in both the parasite and vectors and the partial efficacy of current control methods [3,6–9]. To achieve malaria elimination and eradication will thus require new products and interventions to be used in tandem with the already existing tools [10]. One promising strategy is a mosquito gene drive system, capable of spreading into an entire population from low initial frequencies [11,12] suppressing or replacing mosquito populations [13]. Significant advancements have been made so far in target gene identification, gene construct development, and genome sequencing of the major malaria mosquito populations [14–16].

Despite the proof-of-concept mosquito transformation studies done in laboratories that show promising results and have resulted in idententification of candidate gene drive systems [17–19], there is need to advance for further field trials such that product feasibility and efficacy can be evaluated. Therefore, detailed population genetic assessments must be done systematically before deploying any proposed trial or intervention. This involves understanding the relationship between malaria vectors and the environment they occupy, which provides information about gene flow and in turn aids in understanding species evolution [20]. Information on population genetic structure and gene flow within a vector species provides a basis for modeling the impact of trial intervention and in trial design and helps select suitable locations for field trials, given that, Ideally, these locations should have limited gene flow with surrounding populations [21,22].

In Uganda, the geographic and genetic isolation of lacustrine *An. gambiae* island populations from mainland renders them suitable candidate sites for initial testing of mosquito gene drive systems [15]. However, this isolation could eventually limit the spread of a gene drive during broader implementation [16].

Knowledge of A*n. gambiae* and *An. arabiensis* genetic diversity, structure, and population sizes at island and mainland sites will certainly inform potential variations at the target field sites and efficacy monitoring [21,23,24]. Several studies have examined the structure of *An. gambiae* at the island and mainland populations in Uganda [25–27], but there are no studies that compare the population structure of both *An. gambiae* and *An. arabiensis*, despite their known sympatric occurrence in many regions [28,29]. *An. arabiensis* from the lacustrine islands is often excluded from further analysis due to their limited numbers in the mosquito collections done to date [23,24,30], and anecdotal observations suggest that the islands on Lake Victoria may have minimal complexity of malaria vector species, with *An. gambiae* as the predominant vector (personal communication).

Also, microclimate variation is said to be an important ecological factor that affects mosquito survival and fitness and could therefore affect population numbers and hence malaria transmission potential [31,32]. Mosquito survivorship is strongly influenced by microclimatic conditions in a given area [33] and could result in increased density of a given species in an area whose numbers were previously very low [31,32,34].

With these points in mind, this study aims to provide an up-to-date overview of the population structure, size, and diversity of the two malaria vector species on the islands sites and selected mainland sites of Uganda using an amplicon panel designed to work across the whole *Anopheles* genus to identify species.

Methods

Mosquito sampling

Mosquito collections were conducted from six sites; three mainland sites (Kayonjo, Katuuso, Kibbuye) and three island (Bugiri, Kiimi, Kansambwe) populations (Figure 1). Kibbuye and Katuuso villages are situated in Mukono district and Kayonjo village in Kayunga district (Figure 1), all about 50km North West of the capital city, Kampala. While the lacustrine island sites are situated within Lake Victoria in Uganda and are connected by ship and ferry service to the Ugandan mainland.

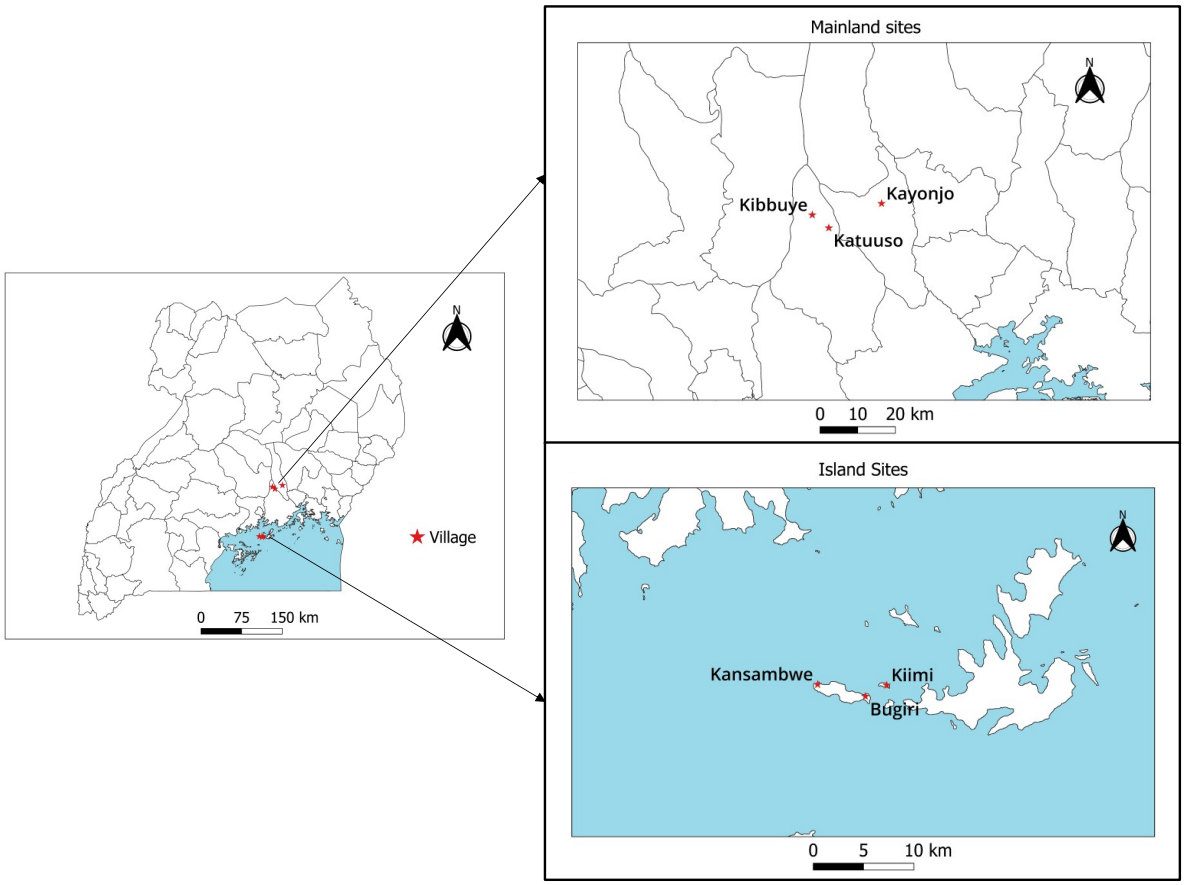


Figure 1: Map showing the location of the 3 mainland and 3 island study sites in Uganda.

On the islands, collections were done from April 2013 to April 2016, and on the mainland from January 2016 to December 2018. Adult mosquitoes were sampled indoors and outdoors using human landing catches (HLC), pyrethroid spray catches (PSC), indoor and outdoor aspirators, and catch basin traps (CBT). All mosquitoes were morphologically identified using the *Anopheles* morphological identification keys [35] and then stored in 80% ethanol in a -20 ^{0}C refrigerator for subsequent molecular analysis. For this study we kept only *An. gambiae* and

131 *An. arabiensis* mosquitoes, although some *An. funestus*, *An. maculipalpis*, and *An. coustani*
132 were also collected (but excluded from this analysis).

133 **Genomic DNA extraction**

134 A total of 3515 whole mosquito samples, each stored in a 96-well plate of 80% ethanol, were
135 shipped to the Wellcome Sanger Institute, United Kingdom, for DNA extraction and amplicon
136 sequencing. For details regarding the sequencing techniques used and the amplicon regions
137 considered, refer to the original publication [16]. In brief, DNA extraction was completed by
138 removing the ethanol from each specimen, adding 100 ul of lysis buffer C to each, and
139 incubating plates at 56 ^{0}C for overnight [16]. DNA from each mosquito was then subjected to
140 a single Polymerase Chain Reaction (PCR) reaction containing a 64 primer pair plex [16]. Each
141 reaction then went into a second PCR to add indexing primers. Eight plates of mosquitoes
142 were pooled for a single MiSeq library [16].

143 Using the **ANO**pheles **SP**ecies and **P**lasmodium (ANOSPP) panel of 64 phylogenetically
144 informative and highly variable "amplicon loci" of which 62 target the *Anopheles* nuclear
145 genome and the other 2 targeting the *Plasmodium* mitochondria (note: these amplicon loci
146 are of ~160 base pairs or "sites" long in downstream analyses), the generated sequence data
147 enabled species identification, detection of *Plasmodium*, hybrid or contaminated samples,
148 and identification of cryptic species [16] (For all other details regarding the sequencing
149 techniques, and the functions of the 62 amplicon loci, please refer to the original publication
150 [16,36]).

151 Using a Variational Autoencoder (VAE), a machine learning approach that identifies patterns
152 in high-dimentional data by compressing it into a lower-dimensional representation [37] to
153 distinguish between closely related species [36].

154 **Bioinformatics processing**

155 The raw reads from the generated sequence data for the confirmed *An. gambiae* and *An*.
156 *arabiensis* were aligned to the reference genome file (VectorBase-
157 59_AgambiaePEST_Genome.fasta). They were genotyped to produce variation data using
158 AmpSeeker, a computational pipeline used to analyse amplicon-sequenced data [38]. The
159 pipeline is built on the Snakemake workflow management system, is highly customisable and

allows users to specify the kinds of data operations, visualizations and analyses they want to be performed on their data (Quality control, Sequence Alignment visualizations, PCA plots, allele frequency calculations).

The resulting raw variant call format (VCF) file contains variant calls across all samples and comprises a total of 12,412 sites (inclusive of insertions and deletions (indels), primer regions, fixed and polymorphic sites). Subsequent subsetting was performed on the VCF file based on the metadata information, which provided details about the mosquito samples, including individual sample IDs, species, collection location, collection period, and season among others. In the filtering process, indels were removed, resulting in 9,890 remaining sites. These sites were distributed across the genome as follows: X (n = 952), 2L (n = 1881), 2R (n = 2815), 3L (n = 1615), 3R (n = 2627). The minor allele frequency (maf) spectrum between 0 and 2% is shown in Figure 2, and the number of alleles per locus (from fixed to quadriallelic) is in Figure 3. Our downstream analyses were based on these 9,890 sites with potential further filtering, such as according to maf and missing values.

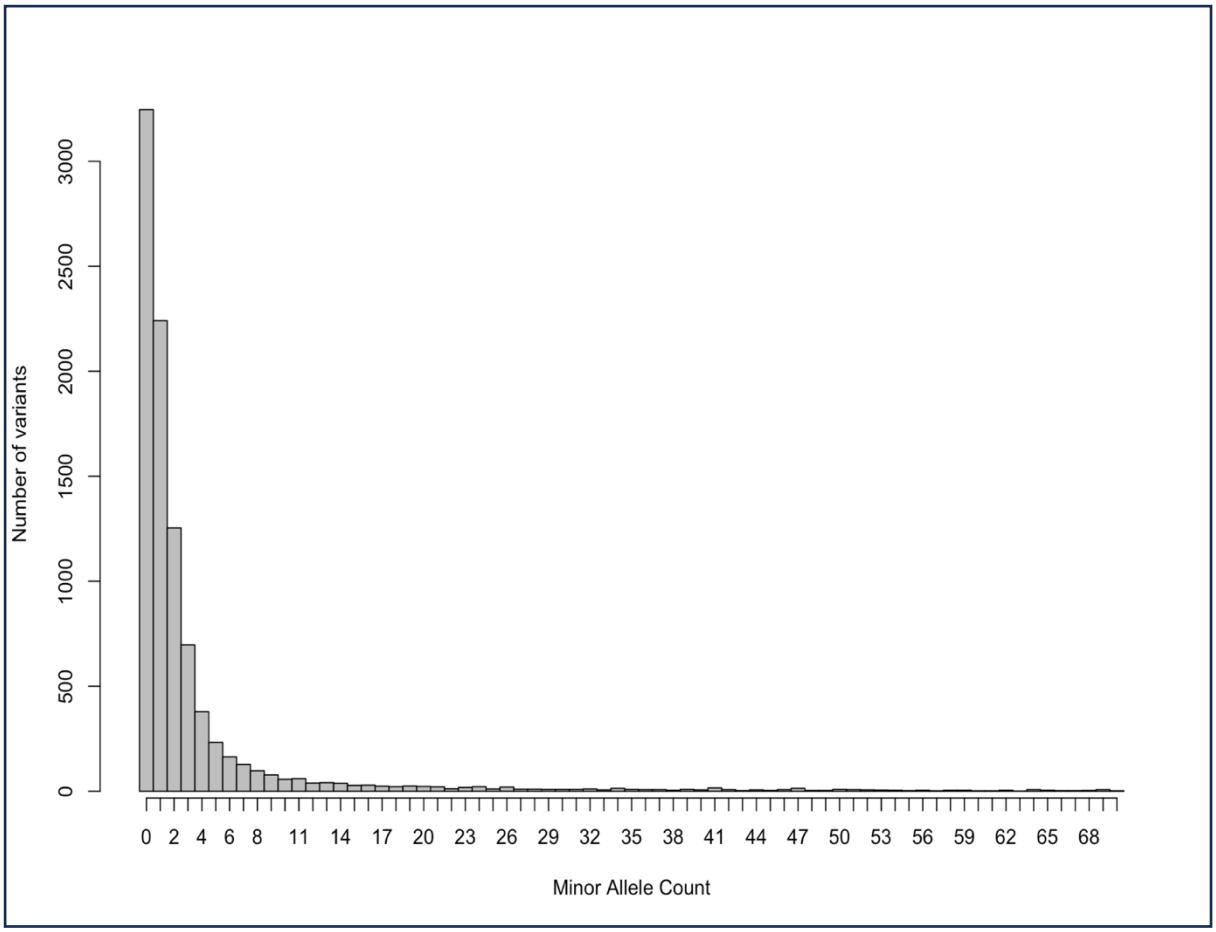


Figure 2: Minor allele frequency (MAF) spectrum between 0 and 2% (for the whole dataset).

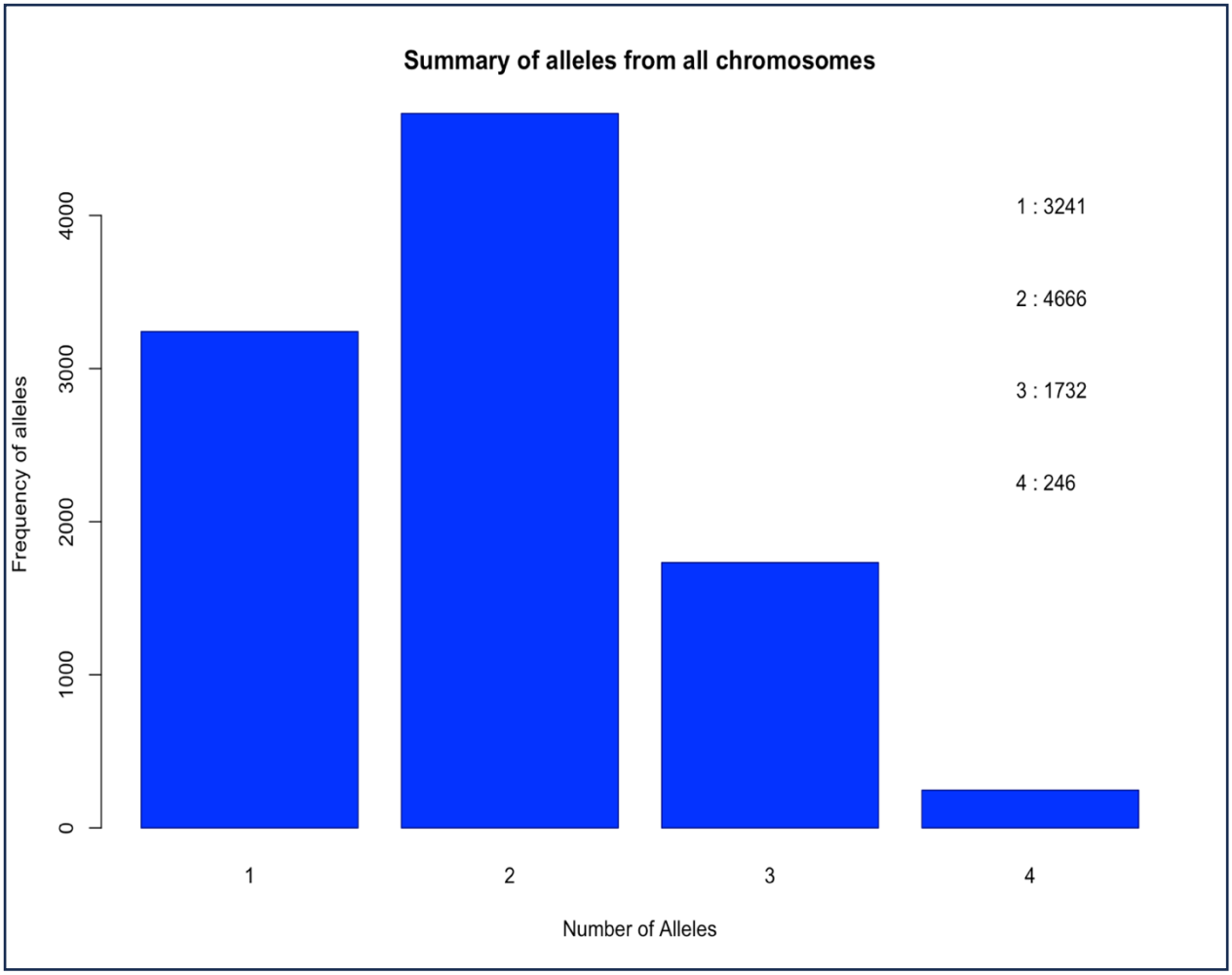


Figure 3: Summary of the number of alleles per site from all chromosomes. 9890 sites are included. The 5 sites with 0 alleles are missing values in all our *An. gambiae* and *An. arabiensis* samples and were excluded from the graph.

Population genomic analysis

Population structure

The following methods were used to describe population structure: (1) F_{ST} between and within pairs of populations using version 1.3.23 of the Popkin package in R [39,40], (2) Principal Components Analysis (PCA) using version 3.6.2 of the prcomp package in R [39,40], and (3) Bayesian clustering analysis with STRUCTURE v2.3.4 [41]. Average pairwise Hudson's F_{ST} was calculated using the fst_hudson_pairwise function of the Popkinsuppl package [42], after which standard errors were computed using the leave-one-out approach that is a specific form of the jackknife resampling method [43–45]. Z-scores were then computed by dividing the

225 estimated FST by its standard error. These Z-scores were used to assess the significance (P-
226 values) of the F_{ST} values. The P-values are derived from the Z_scores by comparing them
227 against the standard normal distribution, which allows for determining the probability of
228 observing such F_{ST} values under the null hypothesis of no differentiation between populations
229 [45]. All the *An. gambiae* collected from the same location were considered as a population,
230 while all 173 *An. arabiensis* were grouped as the seventh population.

231 For pairwise F_{ST}, we used biallelic sites with 10% missing data from within the 3L chromosome
232 with maf >=1%, while for PCA, we first considered all chromosomes and thereafter only the
233 3L arm. To investigate the role of geographic distance in shaping genetic differentiation, we
234 ran a Mantel test to test for isolation by distance (IBD) between populations by combining
235 Euclidean geographic distances (calculated from geographic coordinates) and genetic
236 distances based on 9999 permutations using GenAlEx 6.51b2 [46,47]. For *An. arabiensis*
237 populations, the same F_{ST} and Mantel tests were also conducted using the 3L chromosome
238 arm biallelic SNPs but only for the three mainland populations. The islands were left out
239 because of their small sample sizes (Table 1). PCA was performed in R v4.3.1 with the prcomp()
240 function [48,49], first for biallelic SNPs from all chromosomes and then for the 3L chromosome
241 arm.

242 **Table 1.** The number of confirmed *An. gambiae* and *An. arabiensis* samples across six
243 collection sites

	Bugiri †	Kiimi †	Kansambwe †	Kayonjo	Katuuso	Kibbuye	Total
An. gambiae	28	258	294	1178	450	710	2918
An. arabiensis	2	0	3	51	39	78	173
Total	30	258	297	1229	489	788	3091

244 † Denotes island populations.

245 We looked at doubletons, defined as loci with exactly two copies of the minor allele in a
246 biallelic site. A contingency table of observed counts of heterozygous doubleton pairs across
247 populations was constructed, and the expected counts were calculated based on sample size
248 proportions to assess population structure. Populations were grouped into three categories,
249 *arabiensis* (*An. arabiensis*), *mainland_gam* (mainland *An. gambiae*), and *Island_gam* (island

250 *An. gambiae*), and a chi-square test was conducted to compare observed and expected counts
251 to detect significant population structure.

252 To find the optimal number of clusters (K) in STRUCTURE, we conducted 20 independent runs
253 for each K (from 2-8), using a burn-in value of 100,000 iterations followed by 100,000
254 repetitions [41]. Then K was determined using the Delta K method of Evanno et al. [50],
255 following which CLUMPAK was used to construct a graphical representation of the genetic
256 structure of the 3091 (both *An. gambiae* and *An. arabiensis*) mosquito samples [51]. In this
257 analysis, we randomly chose one site per amplicon locus from the 2R, 3R, and 3L chromosome
258 arms to minimise genetic linkage, with the constraint that maf≥1%.

259 ***Genetic diversity and Ne***

260 To assess the genetic diversity per population, we calculated the overall nucleotide diversity
261 (p) as well as individually for the 6 *An. gambiae* and 3 mainland *An. arabiensis* populations for
262 each chromosome arm [52] using VCF files without missing data. Deviation from the standard
263 neutral model was tested using Tajima's D, which was performed for each population and
264 chromosome to examine population expansion [53]. Tajima's D and nucleotide diversity were
265 computed using the Pegas package in R [54]. Estimates of contemporary Ne were attained
266 using the linkage disequilibrium (LD) based method LDNe [55] of NeEstimator v.2.01 [56]. We
267 randomly chose one site per amplicon locus from chromosomes 3R and 3L with maf≥1% to
268 avoid tight linkage. We similarly individually computed nucleotide diversity, Tajima's D and
269 contemporary Ne for individual mainland *An. arabiensis* populations.

270

271

272

273

274

275

276

277

278

Results

Genetic differentiation (F_{ST})

2918 samples were confirmed as *An. gambiae* and 173 as *An. arabiensis*. These individuals and their resulting amplicon data were our subjects for downstream population genetic analyses, and Table 1 summarises their distribution across the six collection sites. The average pairwise Hudson's F_{ST} is found in Table 2A. All pairwise comparisons produced low F_{ST} values (0.00054-0.028), except between the "outgroup" *An. arabiensis* and any *An. gambiae* populations. Among *An. gambiae* mainland populations, pairwise F_{ST} values were very low (0.00054-0.0016, indicating minimal genetic differentiation, and similarly, low F_{ST} values were observed between Bugiri and the other two islands (Kiimi and Kansambwe), suggesting significant gene flow. The highest F_{ST} values occurred between Kansambwe and all the mainland sites (island vs mainland), and between Kiimi and Kansambwe (island vs island), implying restricted gene flow in these comparisons (Table 2A). The pairwise F_{ST} values between Bugiri and all the other sites, Kayonjo and Katuuso, Kiimi and Katuuso and Kiimi and Kibbuye were low and non-significant ($z < 1.96$, $P > 0.05$). We then calculated pairwise F_{ST} values among the mainland *An. arabiensis* populations (Table 2B), and low and statistically nonsignificant pairwise F_{ST} values were mostly recorded. Also to note, the F_{ST} value between Katuuso and Kibbuye was negative.

Table 2: Pairwise F_{ST} statistics

A. *An. gambiae* & *An. arabiensis*

	BUGIRI †	KIIMI †	KANSAMBWE †	KAYONJO	KATUUSO	KIBBUYE	*An. arabiensis*
BUGIRI †	0	0.6796089	0.5914296	1.762527	1.628254	1.47858	4.361302*
KIIMI †	0.00287	0	4.631849*	1.977574*	1.943631	1.78073	4.360677*
KANSAMBWE †	0.00272	0.02123	0	3.189523*	3.041091*	3.916723*	4.041615*
KAYONJO	0.00995	0.01199	0.02715	0	1.204753	1.991724*	3.779667*
KATUUSO	0.01201	0.01483	0.02841	0.00054	0	2.439927*	3.826917*
KIBBUYE	0.01033	0.01740	0.02433	0.00109	0.00163	0	3.737203*
An. arabiensis	0.21710	0.20209	0.26279	0.22654	0.23330	0.20902	0

B. *An. arabiensis*

	KAYONJO	KATUUSO	KIBBUYE
KAYONJO	0	1.817574	0.09362245
KATUUSO	0.00785541	0	-1.070926
KIBBUYE	0.00018109	-0.0014994	0

(A) Pairwise FST from the 3L chromosome arm for the six *An. gambiae* populations, with the 173 *An. arabiensis* grouped as the seventh population. (B) Pairwise FST among the three *An. arabiensis* populations. For both tables, the lower left triangle in each table shows average Hudson's F_{ST} values between each population pair, while the upper right triangle shows the z-scores for each F_{ST} value calculated via a block-jackknife method. The values with * indicate significant F_{ST} values ($z > 1.96$, $p < 0.05$). † Denotes island populations.

The pairwise F_{ST} values for *An. gambiae* were plotted against geographic distances to investigate the strength of isolation by distance [47,57], and a marginally significant correlation was detected (Figure 4, P = 0.052), suggesting a potential trend towards significance. Therefore, according to the IBD hypothesis, there may be a meaningful association between the genetic and geographical distances among populations.

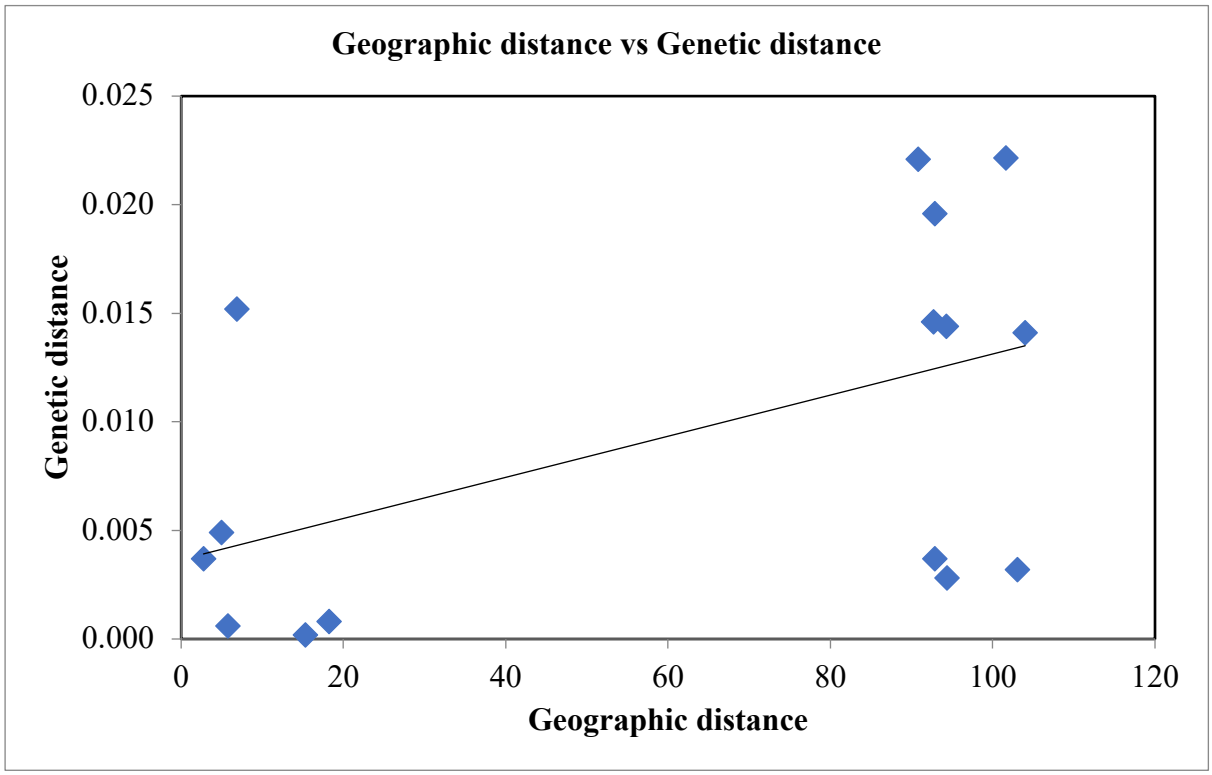
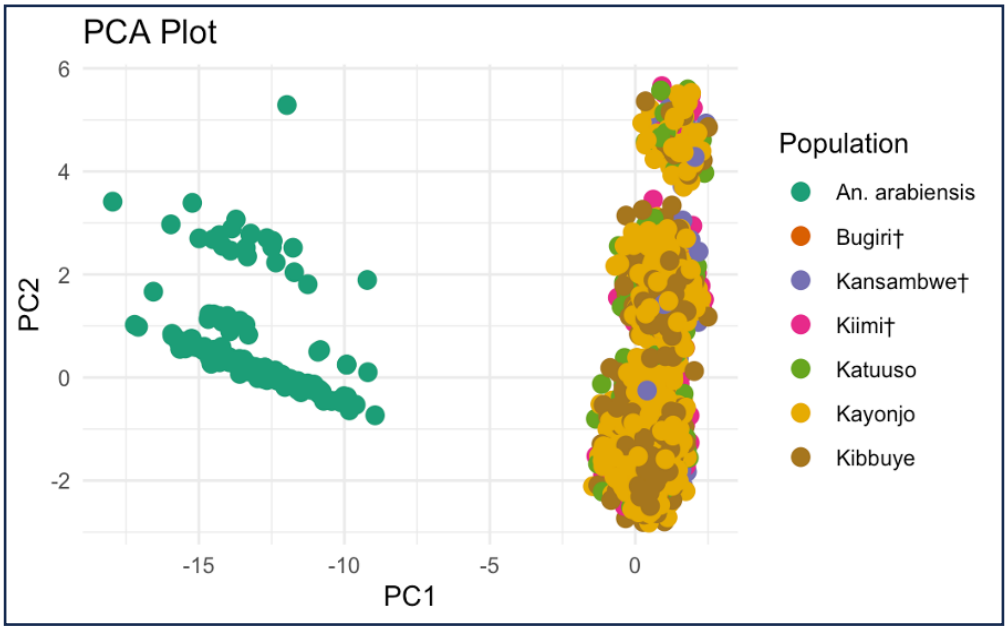


Figure 4: F_{ST} versus geographical distance (km) for *An. gambiae*. FST values were extracted from Table 2. The figure shows a positive slope with no significant correlation between the genetic and geographic distances, P = 0.063

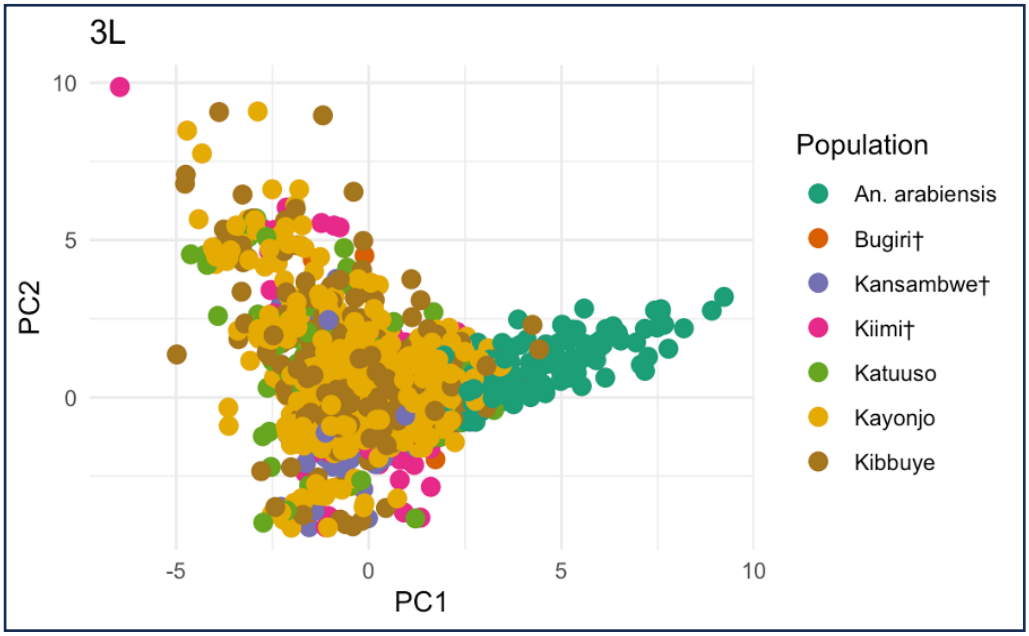
PCA

PCA visualises how the individuals were genetically clustered within and between collection sites. Analysis using all biallelic sites of all chromosome arms of all *An. gambiae* and *An. arabiensis* samples showed clustering into two distinct groups in accordance with each species' genetic difference (Figure 5A). The 3L chromosome arm from both *An. gambiae* and *An. arabiensis* also showed species-specific clustering (Figure 5B) and using chromosome 3L sites of only *An. gambiae*, no geographical clustering of individuals was observed (Figure 5C).

A



B



C

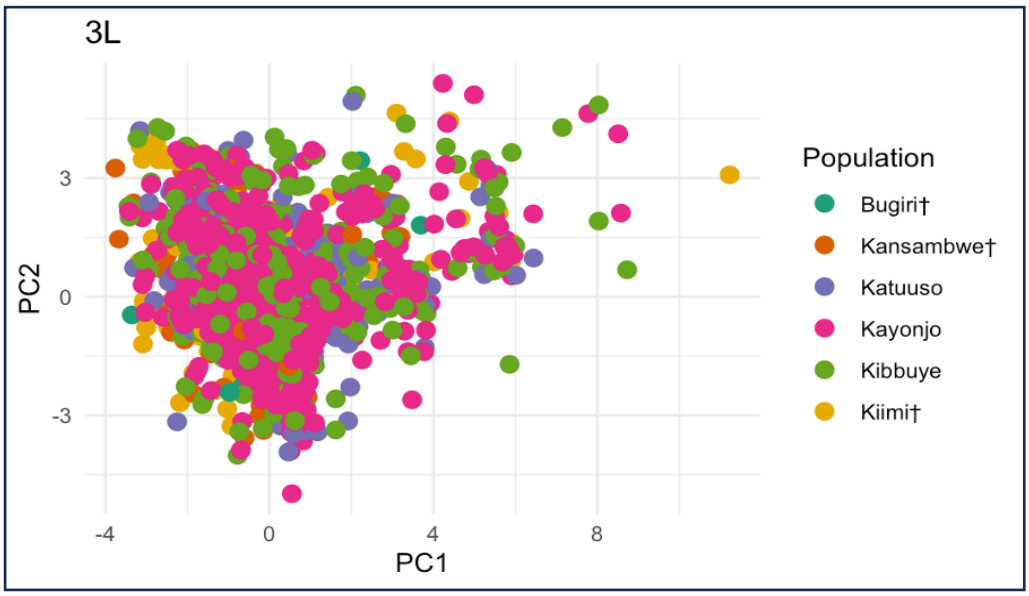


Figure 5 A: Principal component analysis of all chromosome arms of *An. gambiae* (island and mainland populations) and *An. arabiensis*, **B:** The PCA was calculated from the 3L chromosome arm of *An. gambiae* and *An. arabiensis* individuals, and **C**: The PCA was calculated from the 3L chromosome arm of only *An. gambiae* individuals. † Denotes island populations.

Allele sharing in doubleton variants

839 pairs of doubletons were discovered. We calculated the expected counts based on sample size proportions to account for the distribution of the observed heterozygote doubleton pairs under a null hypothesis assuming no population structure. Because of the limited resolution available to distinguish between individual populations in our dataset, we reduced the original 6+1 (Bugiri, Kiimi, Kansambwe, Kayonjo, Katuuso, Kibbuye, and *An. arabiensis*) population groups into three broader categories: *arabiensis*, *island_gam*, and *mainland_gam*. In this way, more robust statistical comparisons and interpretations of allele sharing across major population segments could be made. The reduction from 6+1 populations to three groups improved statistical power and provided insights into gene flow dynamics across the major geographic and species categories of interest.

There were significant deviations between observed and expected pairwise counts of doubleton pairs in the three groups: *arabiensis*, *island_gam*, and *mainland_gam*. The observed pairwise counts of heterozygotes within and between population groups highlight actual genetic similarities, while the expected pairwise counts reflect the hypothetical distribution of pairings in the absence of genetic structuring.

When we considered within-group genetic structuring (Table 3), the *mainland_gam* group showed a significantly higher within-group count of heterozygous pairs than expected (682 observed vs. 480 expected), which suggests a higher-than-anticipated allele sharing or pairing frequency within this group. The *island_gam* group showed a lower-than-expected level of allele sharing (2 observed vs. 29.54 expected), suggesting limited gene flow with other populations, while, the *arabiensis* group showed moderated within-group genetic isolation (15 observed versus 2.63 expected counts), suggesting that it is a relatively closed genetic pool with reduced gene flow to other populations.

Table 3: Allele sharing in doubleton variants

A. Observed counts

	arabiensis	*island_gam*	*mainland_gam*
arabiensis	15	4	102
island_gam	NA	2	34
mainland_gam	NA	NA	682

B. Expected counts

	arabiensis	*island_gam*	*mainland_gam*
arabiensis	2.628186	17.62252	71.03699
island_gam	NA	29.54064	238.1587
mainland_gam	NA	NA	480.01296

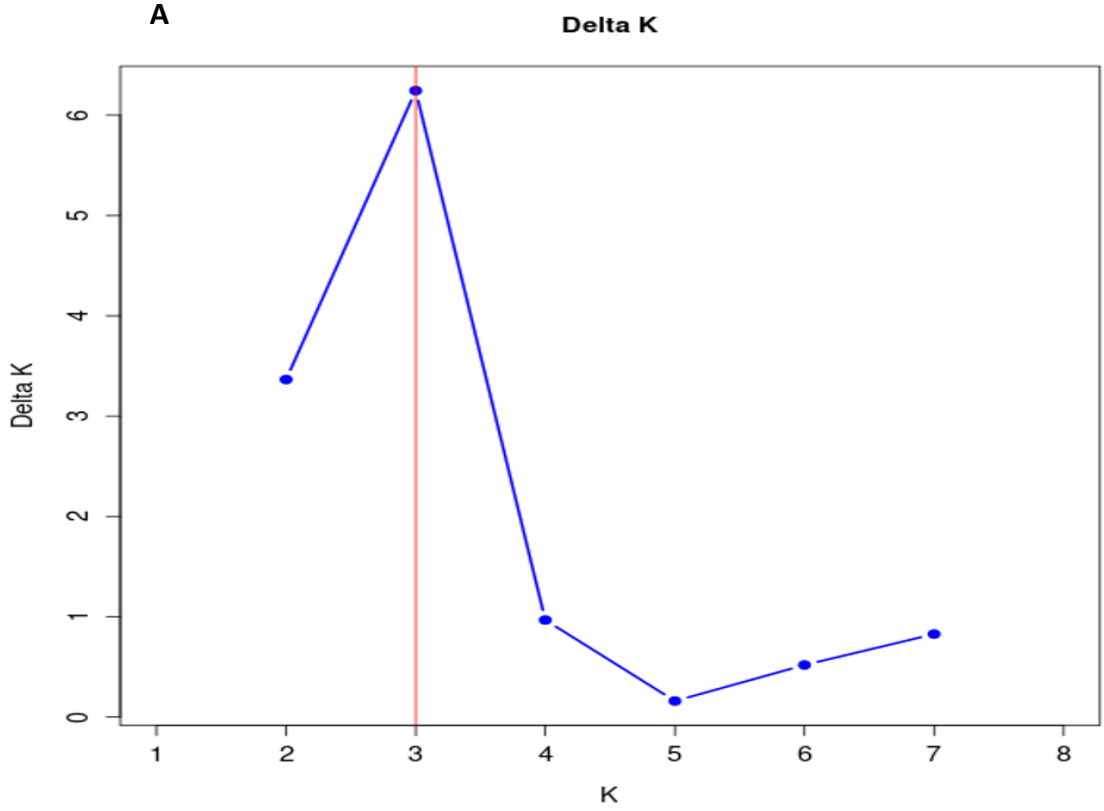
Within and between group counts

When we considered between-group genetic structuring, the observed gene flow between *arabiensis* and *mainland_gam* groups (102 observed versus 71.04 expected counts) indicates occasional genetic exchange, which could be as a result of ancestral similarities. *island_gam* versus *mainland_gam* have very significantly fewer pairs than expected (34 pairs versus an expectation of 238.16 pairs), suggesting restricted gene flow between the two geographic populations. The *arabiensis* versus *island_gam* groups had very significantly fewer pairs than expected (4 pairs versus an expectation of 17.62252 pairs), suggesting restricted gene flow between the two groups.

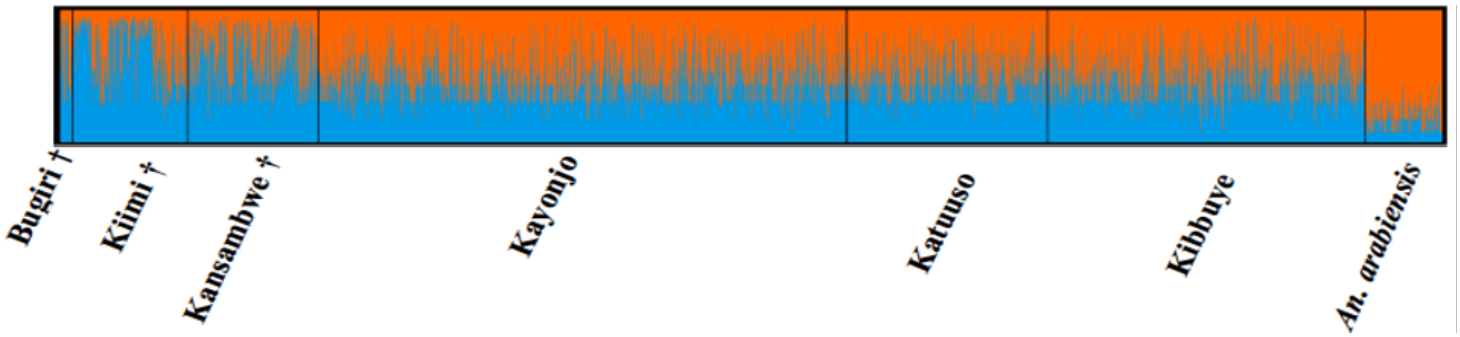
Bayesian clustering analysis

Based on the log-likelihood values and the DeltaK plot, K=3 was the optimal number of evolutionary clusters (Figure 6A). Therefore, the six *An. gambaie* populations and pooled *An. arabiensis* were substantially admixed at K = 3 (K = 3 best explains the genetic variance present between all the populations). Upon close inspection of the results from the Bayesian cluster analysis(Figure 6B), three significant genetic clusters were identified: (1) pooled *An. arabiensis* individuals, (2) island *An. gambiae* individuals, (3) mainland *An. gambiae* individuals. *An. arabiensis* were assigned to the orange population with the highest probabilities, separating them from all *An. gambiae* mosquitoes. The island *An. gambiae* were mostly assigned to the blue but with slightly more uncertainly. Much of the variation was recorded between populations (*An. gambiae* island and mainland individuals and *An. arabiensis* individuals) instead of among individuals within a population. These results from STRUCTURE (Figure 6B) were also largely consistent with those from PCA (Figure 5A-C) with clear separation at an interspecific level and no clear separation at the intraspecific level.

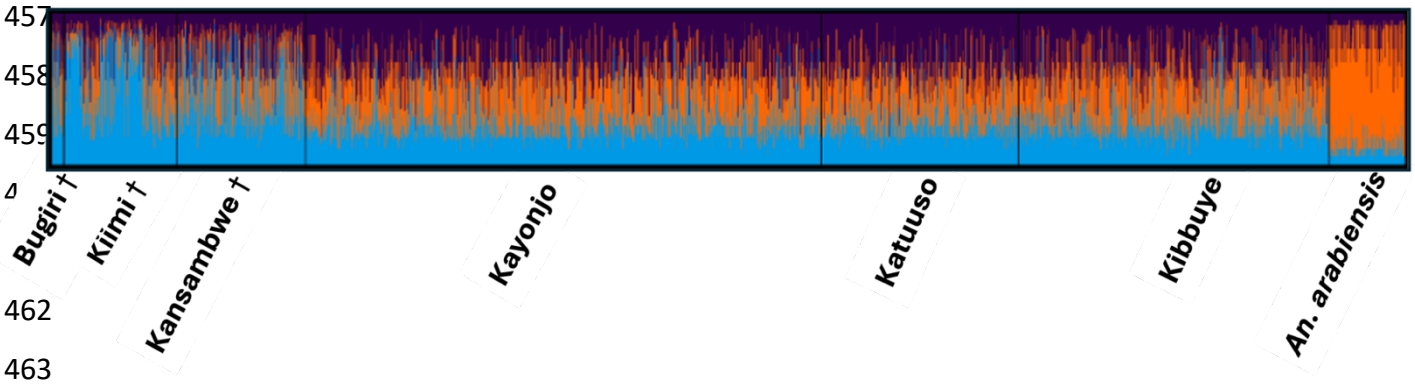
A


Delta K
Delta K
0
1
2
3
4
5
6
1
2
3
4
5
6
7
8
K

B


Bugiri †
Kiimi †
Kansambwe †
Kayonjo
Katuuso
Kibbuye
An. arabiensis

C

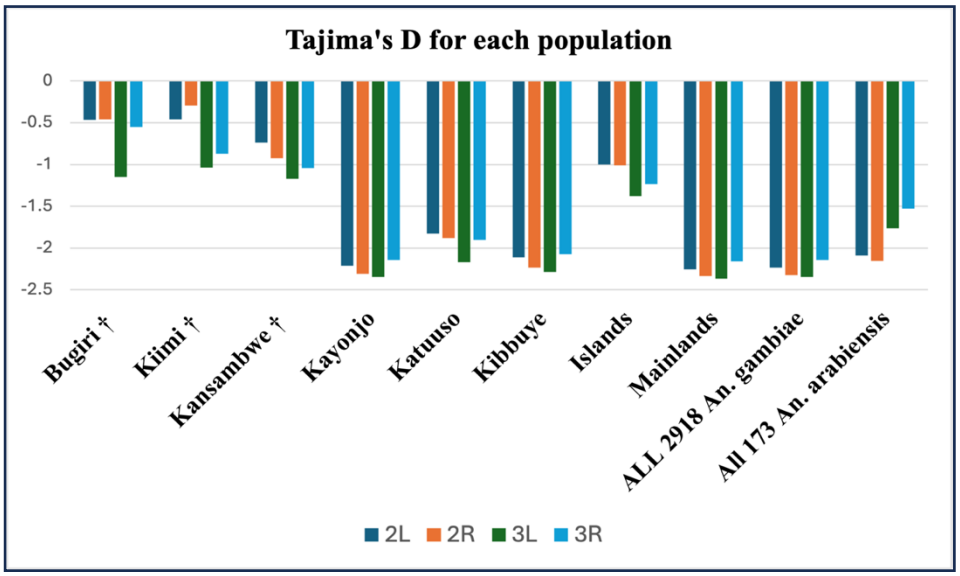

Bugiri †
Kiimi †
Kansambwe †
Kayonjo
Katuuso
Kibbuye
An. arabiensis

464 **Figure 6:** Outputs from Bayesian clustering analysis for the 2918 *An. gambiae* and 173 *An.*
465 *arabiensis*. **(A)** The optimal number of clusters (*K=3*) based on the Evanno et al [26] method
466 showing delta$K$ **(B)** Bar plot representing K=2 of 3091 mosquitoes, each represented by a thin
467 vertical bar colored in proportion to their estimated ancestry within each cluster (**C**) Bar plot
468 representing K=3 of 3091 mosquitoes. A vertical black line bounds each population and the
469 labels with † for islands while those without are for mainland samples.

470 **Genetic diversity and neutrality tests**

471 Nucleotide diversity (π) and Tajima's D were calculated for each *An. gambiae* population (SI
472 Table S1), then further for each chromosome arm (Figure 7A). The overall nucleotide diversity
473 for *An. gambiae* was 0.0115 (range from 0.0107 to 0.01169, SI Table S1) and that of *An.*
474 *arabiensis* was 0.00859 (ranging from 0.00829 to 0.00859; SI Table 2). Generally, higher
475 diversity values were reported for mainland *An. gambiae* populations (0.01143 to 0.01169; SI
476 Table 1) compared to those for island populations (0.01071 to 0.01103; SI Table 1). On the
477 other hand, the diversity values for *An. arabiensis* mainland populations were lower compared
478 to the mainland *An. gambiae* populations. The pattern of nucleotide diversity compared
479 between chromosomes for different populations (Figure 7B) was consistent across all
480 populations and chromosome arms (mainland sites' p > island sites' p). In each population,
481 mean chromosomal nucleotide diversity was highest on the 3R chromosome and lowest on
482 the 2R arm. However, pooled *An. arabiensis* (from all locations) had an overall low diversity
483 compared to the different *An. gambiae* populations, with the lowest diversity in the 2R
484 chromosome arm (P = 0.031, Figure 7B). Overall, the pooled island and mainland diversity falls
485 between the values for each sample site across chromosome arms (Figure 7B).
486 Mean Tajima's D was negative for all populations, an indication of an excess of rare alleles and
487 thus a deviation from the neutral model of a well-mixed population of constant size. However,
488 unlike the *An. gambiae* mainland populations, Tajima's D values for *An. gambiae* island
489 populations and all *An. arabiensis* populations were not statistically significant from from zero
490 (SI Table S1, S2). When tested for each chromosome and population, all Tajima's D values were
491 negative. The same was done for the *An. arabiensis* mainland populations (SI Table S2). The
492 Tajima's D value of the pooled *An. arabiensis* chromosome 2R arm was significantly more
493 negative than other arms of the same population (P = 0.031), and the 2R arm of the individual
494 and pooled populations.

A



B

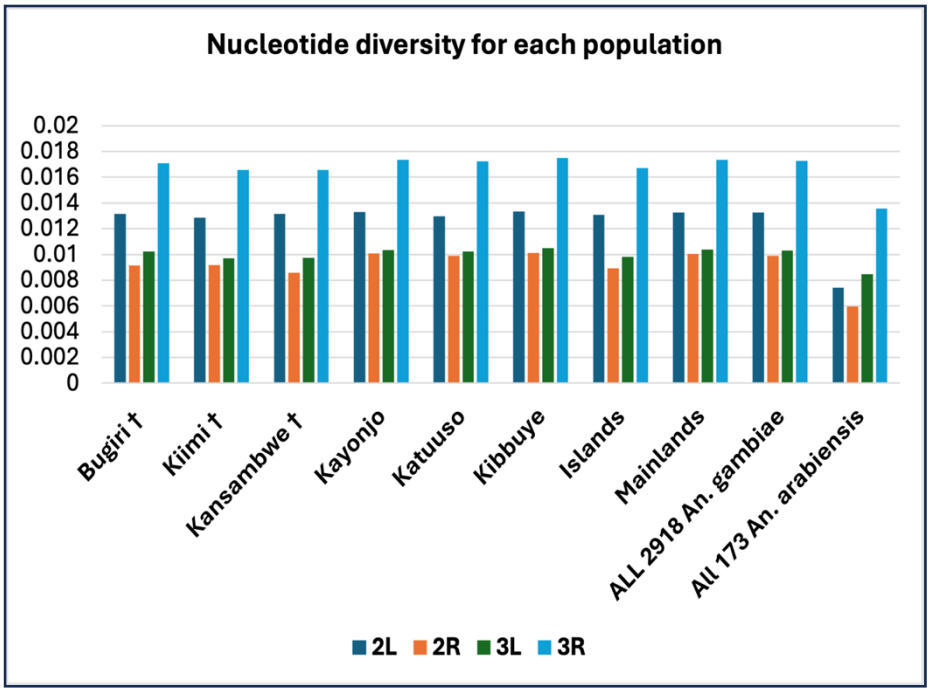


Figure 7: (A) Tajima's D for the six *An. gambiae* populations per chromosome arm. In addition, we calculated the same statistic for the following subsets: pooled *An. gambiae* from the island and the mainland sites, then all *An. gambiae*, and all *An. arabiensis* (**B)** Nucleotide diversity (π) for the six *An. gambiae* populations per chromosome arm. † Denotes an island population.

527 **Contemporary effective population size (*Ne*)**

528 Generally, for *An. gambiae* populations, smaller Ne values were observed for the island
529 populations (between 146 and 249), compared to mainland populations (between 4054 and
530 8190) (Table 4A). The point estimate for Bugiri was inconclusive but came with a very small
531 lower bound. For the *An. arabiensis* populations, we only estimated Ne for the mainland sites,
532 and their estimates were comparable to the mainland *An. gambiae* populations. (Table 4B).

533 **Table 4: Contemporary Ne estimates**

534 **A. *An. gambiae***

Population	No. of individuals	Number of sites	Overall r^2	Estimated Ne
Bugiri†	28	37	0.035383	∞ (15.1 – ∞)
Kiimi†	258	32	0.006168	146.4 (61.7–640.2)
Kansambwe†	294	32	0.004764	249.3 (146.7– 538.5)
Kayonjo	1178	15	0.000901	6723.1 (515.9 - ∞)
Katuuso	450	27	0.002181	4054.1 (640.9 - ∞)
Kibbuye	710	21	0.001416	8190.9 (521.5 - ∞)

535

536 ***B.* An*. arabiensis***

Population	No. of individuals	Number of sites	Overall r^2	Estimated Ne
Kayonjo	51	31	0.018910	∞ (49.4- ∞)
Katuuso	39	35	0.027836	3402.8 (36.3 - ∞)
Kibbuye	78	31	0.013438	3587.4 (95.9 – ∞)

537 Site choice is described in the main text. † Denotes island populations. 0.01 was the lowest
538 allele frequency used. The values in parentheses represent the jackknife range.

539

540

541

542

543

544 **Discussion**

545 **Polymorphic sites and population structure**

546 From the 62 amplicon loci we successfully mapped 9890 sites onto the reference genome,
547 many of which were fixed sites (Figure 3) or with low polymorphism (Figure 2). We recognise
548 that the primary objectives of ANOSPP are for species identification of any *Anopheline* and
549 parasite detection. Primer binding sites were selected to be extremely conserved because the
550 panel was designed to work across the entire genus, representing more than 100 million years
551 of divergence. Thus, these 62 loci lack highly polymorphic sites outside of the inversions and
552 this posed some challenges to the downstream population genetic analyses. Nonetheless, this
553 study provided a lot of useful insights into the population structure and diversity among the
554 chosen island and mainland *An. gambaie* and *An. arabiensis* populations in Uganda.

555 In general, F_{ST} values below 0.05 imply low genetic differentiation, while those between 0.05
556 and 0.15, 0.15 to 0.25, and above 0.25 correspondingly indicate moderate, high, and very
557 high genetic differentiation [58].

558 In this study, the level of genetic differentiation among the *An. gambiae* populations was
559 mostly low, regardless of the types of populations ($F_{ST} < 0.05$). Many of these low F_{ST} values,
560 however, were statistically larger than zero with small P-values, meaning that these
561 populations were not entirely panmictic but rather segregated while maintaining certain levels
562 of connectivity. All pairwise F_{ST} values between Bugiri and the rest of the *An. gambiae*
563 populations were very low and insignificant, which could either be the consequence of high
564 gene flow to/from Bugiri [59] or the small sample size of Bugiri that could have culminated in
565 insufficient power to detect true genetic differentiation (due to large sampling variance) [60].

566 High genetic differentiation was observed between *An. gambiae* and *An. arabiensis*
567 populations, concurring with the fact that they are two distinct species. When the *An.*
568 *arabiensis* mainland populations were independently considered, all the pairwise F_{ST} values
569 were generally low and non-significant. It is noteworthy that the F_{ST} value between Katuuso
570 and Kibbuye was negative, implying extremely low levels of genetic differentiation. The
571 Hudson's method allows negative F_{ST} values when the observed heterozygosity within
572 populations exceeds heterozygosity between populations, therefore, negative values indicate

573 that populations are more similar than expected, reflecting low or zero differentiation, with
574 no detectable genetic structure between them [43,61,62].

575 The genetic differentiation between *An. gambiae* populations was varied with inter-island
576 pairwise comparisons having slightly higher values compared to the inter-mainland pairwise
577 values. Higher differentiation values between island populations were comparable to that
578 reported in the Kayondo et al [25,30] study and could be because of their separation by water,
579 small populations and local adaptation. The highest pairwise F_{ST} values were recorded
580 between Kansambwe and all three mainland sites (P<0.05), while the lowest values were
581 recorded for Katuuso and Kayonjo (0.00054 p>0.05). These pairwise F_{ST} results attained in this
582 study were consistent with previous studies which showed lower differentiation within
583 mainland populations compared to differentiation within island populations and between
584 island and mainland populations [23,24,30].

585 In comparison, looking at the pairwise F_{ST} values for mainland *An. arabiensis* populations, low
586 and non-significant genetic differentiation was observed, implying unrestricted gene flow
587 between them. The negative F_{ST} values (between Katuuso and Kibbuye) could be because of
588 an excess of within-population genetic diversity over the between-population genetic
589 diversity [63].

590 The genetic differences between the island and mainland *An. gambiae* populations and within
591 island populations could be a result of their physical separation by water, small population size
592 and local site-specific ecological adaptation [25]. The low and non-significant F_{ST} values
593 between Bugiri and all the other sites, Kayonjo and Katuuso, Kiimi and Katuuso and Kiimi and
594 Kibbuye could be indicative of high levels of gene flow between them (mainland vs mainland),
595 or low sample size of the Bugiri population (low statistical power) [64].

596 We noted that low point estimates of pairwise fst yielded high Z_scores, suggesting the
597 presence of subtle yet statistically significant population structure despite low apparent
598 differentiation. This occurrence can largely be attributed to the statistical power afforded by a
599 large sample size, which can detect even minor genetic differences across populations [65,66].
600 For example, when sample sizes are substantial, there is an increased sensitivity to detect
601 structure, enabling the identification of subtle allele frequency shifts that might otherwise be

602 masked in smaller datasets [66]. Between Kiimi and Katuuso, the low pairwise fst value and
603 non-significant P-value could be because of temporal sampling variability, which could result
604 in capturing different allele frequencies due to temporal fluctuations [67–69]. Comparing the
605 relative amounts of gene flow taking place between populations, in turn, aids in predicting
606 the trajectory of the alleles introduced by gene drive into the wild-type population [21].

607 Given that the P-value from the Mantel test is slightly above the conventional threshold of
608 0.05 (P=0.052), it is very close to the threshold and therefore suggests a potential trend
609 toward a significant correlation between pairwise fst values and geographical distances
610 among populations. This could, therefore suggest a relevant relationship that deserves further
611 investigation into the relationship between genetic structure and geographical distribution.
612 This value could, however, indicate a weak or suggestive association between genetic
613 differentiation and spatial separation, potentially hinting at IBD in the population [70].
614 Therefore, the geographic distance between sites could be responsible for part of the
615 differentiation observed.

616 Although *An. gambiae* F_{ST} results showed reasonable differentiation between mainland and
617 island populations, such patterns were not observed from PCA. The first PC separated the two
618 species, but this could be the only definitive conclusion from the plots (Figures 4A, 4B). We
619 further observed a clustering that is not geographically distinct, with individuals from different
620 populations intermixed and showing no clear separation or clustering by location (Figure 4C).
621 This could suggest that these mosquito populations are not genetically distinct from one
622 another on the 3L chromosomal arm, thus implying minimal genetic differentiation between
623 these populations and that individuals from different locations share very similar genetic
624 profiles for the loci analyzed on this chromosome arm.

625 PCA generally captures the largest sources of genetic variation, which may not reveal distinct
626 clusters if other factors dominate the variation represented by the PCs [71]. Therefore,
627 ecological and seasonal influences that contribute to the observed structure warrant further
628 investigation into assessing the relative contribution of these factors to the differentiation
629 observed between these populations.

630 We investigated population structure using doubleton analysis that showed considerable
631 genetic connectivity within the mainland populations compared to island populations (Table
632 3). This suggests that the *island_gam* group is somewhat genetically isolated from other
633 groups and aligns with observed patterns in many malaria vector studies where isolated
634 populations exhibit unique genetic signatures due to limited allelic exchange with surrounding
635 groups [72,73]. The moderate within-group isolation of the *arabiensis* group (Table 3)
636 indicates more frequent genetic interactions within this group than expected, potentially due
637 to localized genetic structures, ecological factors, or other selective pressures driving allele
638 frequencies. This could reflect a localized population structure and its adaptability to arid
639 environments, which fosters more frequent allele pairings within the group than across more
640 widely distributed populations [74].

641 In the between-group comparisons, the *island_gam* versus *mainland_gam* counts showed a
642 rather marked deficit in inter-group pairwise counts (Table 3), suggestive of distinct separation
643 or limited interaction between the island and mainland populations, probably brought about
644 by geographical or ecological barriers. The counts between *arabiensis* and *mainland_gam*
645 groups indicate a degree of gene flow, implying occasional genetic exchange, which could have
646 direct implications for vector control, especially given the differential patterns of gene flow
647 and inbreeding detected across these populations.

648 This pattern of observed versus expected doubleton pairs sides with the broader view of
649 population structure in malaria vectors, where poor gene flow across ecological boundaries
650 frequently results in discrete genetic clusters [75], and has implications on both gene drive
651 strategies and management of insecticide resistance.

652 The strong genetic differentiation between *An. gambiae* and *An. arabiensis* populations was
653 also observed from the outputs from STRUCTURE (Figure 6). STRUCTURE software was used
654 to detect clustering between individuals using a Bayesian algorithm, based on local Hardy-
655 Weinberg equilibrium of genotypes at multiple loci [76], and found that K= 3 clusters were the
656 optimum assignment (Figure 6B). This means that three clusters are sufficient to represent
657 the major genetic differentiation within the dataset, and adding more clusters only subdivides
658 existing groups but does not add biologically meaningful sub-structure. This is often observed

659 when there's low genetic differentiation, where minor allele frequency differences can create
660 additional "clusters" in STRUCTURE without reflecting true population separation [77].

661 STRUCTURE output for the *An. gambiae* and *An. arabiensis* species of each population showed
662 sharing of ancestry between each individual at differing proportions (with differing
663 consistency between all island, mainland *An. gambiae* and *An. arabiensis* individuals). The
664 three genetic clusters corresponded to first species (*An. arabiensis*) then to geographical
665 (mainland and island *An. gambiae*) differences, with relatively more similarity between island
666 compared to mainland *An. gambiae* individuals (Figure 6B) These results are comparable to
667 the F_{ST} and PCA results that show less diversity between island and mainland *An. gambiae*
668 individuals and more diversity (clear separation) between *An. gambiae* and *An. arabiensis*
669 individuals [55,56].

670 Despite using very few polymorphic sites, STRUCTURE was able to outline the three clusters
671 with visible differences. Therefore, using more sites could produce clearer and more distinct
672 clustering. It is also notable that because pairwise F_{ST} is sensitive to allele frequency
673 differences between populations and works best when populations are distinct and
674 genetically isolated, it may not always capture subtle population differentiation, while
675 STRUCTURE, on the other hand, assigns individuals to clusters based on their multilocus
676 genotypes, and is thus more sensitive to subtle population differentiation [41,78–80].

677 **Genetic diversity and neutrality tests**

678 The Tajima's D test is used to distinguish between DNA sequences evolving neutrally from the
679 ones that are evolving under a non-random process or non-constant demography. The
680 statistic compares two measures of genetic diversity; the number of segregating sites (S) and
681 the average number of pairwise differences (π) in a population, and infers evolutionary
682 processes like selection, population expansion, or contraction [81]. Tajima's D values were
683 generally negative (SI Tables S1&S2), in line with previous findings [75]. More negative
684 Tajima's D values reported at mainland sites compared to the island sites was an indication
685 of an abundance of an excess of low-frequency alleles at the mainland sites, potentially due
686 to [66,67] more recent or stronger evolutionary events such as purifying selection, more
687 recent population expansion, or selective sweep that affected the genetic diversity in the
688 populations at the mainland [81–83]. Given that *An. gambiae* thrives in human-dorminated

689 environments, its population and range will thus expand most in areas where human
690 population density is high and growing, whether on island or mainland sites [84].

691 Tajima's D for *An. gambiae* island populations and all *An. arabiensis* populations were not
692 statistically significantly different from zero (SI Table S1, S2); which supported the fact that
693 these populations are in genetic and demographic equilibrium with no population bottleneck
694 or expansion, an occurrence consistent with the neutral mutation hypothesis [85]. This
695 suggests that these populations are at mutation-drift equilibrium, maintaining a stable
696 population [86,87]. This non-deviation from the neutral mutation hypothesis could be
697 because of the relatively small sample sizes at the island sites (SI Table S1) and *An. arabiensis*
698 (SI Table S2) populations compared to those at mainland sites, a quantity that Tajima's D is
699 said to be highly sensitive to [88].

700 Nucleotide diversity (π) is a useful measure of genetic diversity since it directly quantifies the
701 number of differences in genotypes between all possible pairs of alleles in a population, and
702 its variation is influenced by a species' life history [89]. Using the patterns of nucleotide
703 diversity, we can infer aspects of long-term population demography by considering deviations
704 from the null model of neutral variation in isolated populations of constant size [84]. It was
705 known that *An. arabiensis* has a lower diversity than *An. gambiae* [90]. The estimates of the
706 overall nucleotide diversity for each *An. gambiae* population were within the range of values
707 reported in previous studies which utilised four loci (white, tox, G6pd, xdh) across several
708 African locations [90] and immune-related (LRIM1, CTL4, CTLMA2, APL2) and housekeeping
709 genes [91], and were consistent with the fact that island populations display marginally lower
710 nucleotide diversity compared to the mainland [92,93]. The explanation for this could be
711 inbreeding, or higher degree of isolation with small neighborhood size [92,94–96]. Small
712 population sizes experiencing stochastic genetic losses due to genetic drift. Also limited
713 migration from meta-populations from the mainland due to the geographical barriers.

714 On chromosome arm 2R, the pooled *An. arabiensis* population has lower nucleotide diversity
715 than individual or pooled *An. gambiae* populations. This could be because chromosome 2R
716 contains many segregating inversions, some of which are polymorphic in *An. gambiae* and
717 fixed in *An. arabiensis,* which inversions can limit recombination in certain regions, leading to
718 reduced genetic diversity over time [97,98]. This phenomenon is especially pronounced in

species like *An. arabiensis*, where these inversions (2La, 2Rb and 3Ra) could be fixed, and further decrease genetic variation along the 2R chromosome relative to *An. gambiae* populations [99,100].

Additionally, the frequencies for inversions segregating in both species could be/are different for the two species [98]. This could also suggest that there is a decline in population size and/or balancing selection, a feature of the maintenance of segregating 2Rb, c, d, u and j inversions [84,101].

Contemporary effective population size

Estimates of Ne attained using the LDNe method [55] in NeEstimator v.2.01 [56] were based on its performance that is superior to other single-sample estimators [102,103]. While nucleotide diversity contains demographic information over a much longer timeframe, our Ne estimates from unlinked LD are local and recent, focusing only on one to a few generations ago. Ne estimates of the *An. gambiae* Island populations were generally smaller compared to mainland populations and were comparable to those reported in the Kayondo et al. and Wiltshire et al. studies [25–27]. These low Ne estimates recorded for island sites indicate their small neighbourhood size and could be the cause of higher levels of genetic differentiation (presented as pairwise F_{ST} estimates) shown in Table 4A [56], given that random genetic drift which affects the stability of allele frequencies in a population is more variable in smaller populations [27].

The low Ne values reported for island populations could also be influenced by factors such as the pronounced effect of seasonality on islands, leading to fluctuations in vector abundance, particularly during dry seasons when population size is suppressed, resulting in increased genetic drift. Furthermore, anthropogenic events that result in habitat fragmentation and patchy populations together with isolation through geographical barriers such as lakes restrict gene flow between subpopulations, leading to reduced genetic diversity and lower effective population size (Ne) within each isolated group [104]. The infinity values of point estimates such as those recorded for Bugiri, was almost certainly due to insufficient sample size compared to the magnitude of genetic drift. The infinite jackknife upper bound in the mainland populations suggests that based on the LD observed, it's statistically feasible that Ne could be extremely large, although the exact size cannot be pinpointed with certainty [105–107].

749 Larger Ne values observed in the mainland *An. gambiae* populations suggest greater genetic
750 diversity and better connectivity, making them more resilient to control measures and more
751 difficult to eliminate compared to smaller island populations [108–112]. High Ne thus implies
752 slower genetic drift which allows for sustained genetic variation that enhances adaptation to
753 environmental changes, moreover, studies on dry season persistence indicate that seasonality
754 influences Ne estimation, as dry season surviving populations can rebound and maintain
755 genetic diversity across seasons, further complicating efforts to control malaria vector
756 populations [113–117].

757 This knowledge of the size of the host population informs the introduction threshold for gene
758 drive releases (minimum number of mosquitoes required to transform a population), and
759 therefore, in this case, a small and reproductively isolated population could effectively reduce
760 this introduction threshold and enable the successful integration of the introduced gene [21]

761 **Conclusion**

762 The results from this study show that the mainland *An. gambiae* populations were more
763 diverse and differentiated compared to the mainland *An. arabiensis* populations. On the other
764 hand, the *An. gambiae* populations collected from the three islands considered in this study
765 were relatively genetically differentiated from the mainland populations, and even from one
766 another, we suspect primarily driven by water as a physical barrier to gene flow, smaller
767 effective population sizes and potentially local ecological adaptations. Due to the ease of
768 access to the islands which allows easy monitoring, the fact they comprised a smaller
769 proportion of non-target malaria vectors in Uganda (very few *An. arabiensis*), higher genetic
770 differentiation, genetic structure, smaller effective population sizes compared to mainland
771 sites, and their small geographical size, islands are promising sites for field trials to test the
772 effectiveness of mosquito gene drive systems.

773 As mentioned at the start, *An. arabiensis* individuals are often excluded from further analysis
774 due to their limited numbers in the mosquito collections done to date [23,24,30], which could
775 be because of the sampling methods used. Therefore, studies that target the exophilic and
776 zoophagic characteristics of this species should be carried out to give a better picture of its
777 density, specifically at the islands which are the potential sites for field trials to test mosquito

gene drive systems. This will help us to confirm the proportion of the non-target malaria vector species at the island sites.

As much as we would like to provide a comprehensive perspective of the overall spatial and temporal variation of the two species in Uganda, the latter cannot be investigated here because of the different sampling times in the island and mainland sampling. We welcome further studies on other biological factors, such as the mosquitoes' seasonal dynamics and dry season persistence mechanisms [118,119], as these contribute to our observed population structure.

Future works should include a dataset using whole genome sequenced data, exploring temporal sampling, further investigations into the specific contribution of the suggested factors we have used, as well as the influence of other suggested biotic (e.g., passive transport) and abiotic factors (e.g., seasonality and chromosome inversions) that may contribute to the observed differentiation. These will add a valuable body of information required for evaluating these Islands as potential release sites for mosquito gene drive systems.

Acknowledgements

We are grateful to Cindy Sewon, a former undergrad student at Imperial College, London for her valuable contributions, John, B. Connolly, Silke, Fuchs and Samantha, O'Loughlin for their support and helpful comments towards this work, and the field team that did the mosquito sample collection.

Conflict of interest

The authors declare no conflicts of interest. The funding bodies had do direct role in the design of the study nor in the collection, analysis, interpretation of data or in the writing of the manuscript

Author contributions:

JK conceived the research topic, hosted lab and field work. ML hosted the targeted amplicon sequencing. MB and AM participated in amplicon sequencing. RM analysed the data and wrote the first draft. TH supported in data analysis. EL supported in bioinformatics processing.

All authors edited and commented on the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the Bill & Melinda Gates Foundation and the Open Philanthropy Project, an advised fund of the Silicon Valley Community Foundation.

References

1. Bhatt S, Weiss DJ, Cameron E, Bisanzio D, Mappin B, Dalrymple U, et al. The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015. 2015;526:207–11.
2. Lehmann T, Dao A, Yaro AS, Adamou A, Kassogue Y, Diallo M, et al. Aestivation of the African Malaria Mosquito, Anopheles gambiae in the Sahel. American Journal of Tropical Medicine & Hygiene. 2010;83:601–6.
3. WHO. World Malaria Report. World Health Organization; 2024.
4. WHO. World Malaria Report. World Health Organization. 2023.
5. WHO. World Malaria Report. World Health Organization. 2022.
6. Hemingway J. The role of vector control in stopping the transmission of malaria: Threats and opportunities. Philosophical Transactions of the Royal Society B: Biological Sciences. Royal Society; 2014.
7. Mushtaq I, Sarwar MS, Chaudhry A, Shah SAH, Ahmad MM. Updates on traditional methods for combating malaria and emerging Wolbachia-based interventions. Front Cell Infect Microbiol. Frontiers Media SA; 2024.
8. Plowe C V. Malaria chemoprevention and drug resistance: a review of the literature and policy implications. Malar J. BioMed Central Ltd; 2022.
9. Thu AM, Phyo AP, Landier J, Parker DM, Nosten FH. Combating multidrug-resistant Plasmodium falciparum malaria. FEBS Journal. Blackwell Publishing Ltd; 2017. p. 2569–78.
10. Kaddumukasa MA, Wright J, Muleba M, Stevenson JC, Norris DE, Coetzee M. Genetic differentiation and population structure of Anopheles funestus from Uganda and the southern African countries of Malawi, Mozambique, Zambia and Zimbabwe. Parasit Vectors [Internet]. 2020;13:1–13. Available from: https://doi.org/10.1186/s13071-020-3962-1
11. Hammond A, Pollegioni P, Persampieri T, North A, Minuz R, Trusso A, et al. Gene-drive suppression of mosquito populations in large cages as a bridge between lab and field. Nat Commun. 2021;12.
12. Marshall JM, Akbari OS. Gene Drive Strategies for Population Replacement. Genetic Control of Malaria and Dengue. 2016;169–200.
13. Burt A. Heritable strategies for controlling insect vectors of disease. Philosophical Transactions of the Royal Society B: Biological Sciences. 2014;369.
14. Gantz VM, Jasinskiene N, Tatarenkova O, Fazekas A, Macias VM, Bier E, et al. Highly efficient Cas9-mediated gene drive for population modification of the malaria vector mosquito Anopheles stephensi. Proc Natl Acad Sci U S A. 2015;112:E6736–43.
15. Hammond MA, Galizi R. Gene drives to fight malaria: current state and future directions. Pathog Glob Health. 2017;111:412–23.
16. Makunin A, Korlević P, Park N, Goodwin S, Waterhouse RM, von Wyschetzki K, et al. A targeted amplicon sequencing panel to simultaneously identify mosquito species and Plasmodium presence across the entire Anopheles genus. Mol Ecol Resour. 2022;22:28–44.
17. Galizi R, Doyle LA, Menichelli M, Bernardini F, Deredec A, Burt A, et al. A synthetic sex ratio distortion system for the control of the human malaria mosquito. Nat Commun. 2014;5.
18. Gantz VM, Bier E. The mutagenic chain reaction: A method for converting heterozygous to homozygous mutations. Science (1979). 2015;348:442–4.
19. Hammond A, Galizi R, Kyrou K, Simoni A, Siniscalchi C, Katsanos D, et al. A CRISPR-Cas9 gene drive system targeting female reproduction in the malaria mosquito vector Anopheles gambiae. Nat Biotechnol. 2016;34:78–83.

20. Ng'habi, R., K., Knols, J., G., B., Lee, Y., Ferguson, M., H. and Lanzaro, C. G. Population genetic structure of Anopheles arabiensis and Anopheles gambiae in a malaria endemic region of southern Tanzania. Malar J. 2011;10.
21. Lanzaro GC, Tripet F. Gene flow among populations of Anopheles gambiae: a critical review. 2003;1–24. Available from: http://edepot.wur.nl/136907%5Cnpapers3://publication/uuid/F232B8BC-F74F-456E-A0DB-11E8198EE5B1
22. Lanzaro GC, Touré YT, Carnahan J, Zheng L, Dolo G, Traoré S, et al. Complexities in the genetic structure of Anopheles gambiae populations in west Africa as revealed by microsatellite DNA analysis. Proc Natl Acad Sci U S A. 1998;95:14260–5.
23. Wiltshire MR, Bergey MC, Kayondo KJ, Birungi J, Mukwaya GL, Emrich JS, et al. Reduced-representation sequencing identifies small effective population sizes of Anopheles gambiae in the north-western Lake Victoria basin, Uganda. Malar J. 2018;17.
24. Kayondo, K., J., Mukwaya, G., L., Stump, A., Michel, P., A., Coulibaly, B., M., Besansky, J., N. and Collins, H. F. Genetic structure of Anopheles gambiae populations on islands in northwestern Lake Victoria, Uganda. Malar J. 2005;4.
25. Kayondo, K., J., Mukwaya, G., L., Stump, A., Michel, P., A., Coulibaly, B., M., Besansky, J., N. and Collins, H. F. Genetic structure of Anopheles gambiae populations on islands in northwestern Lake Victoria, Uganda. Malar J. 2005;4.
26. Lukindu, M., Bergey, M., C., Wiltshire, M., R., Small, T., S., Bourke, P., B., Kayondo, K., J. and Besansky, J. N. Spatio-temporal genetic structure of Anopheles gambiae in the Northwestern Lake Victoria Basin, Uganda: implications for genetic control trials in malaria endemic regions. Parasit Vectors. 2018;11.
27. Wiltshire, M., R., Bergey, M., C., Kayondo, K., J. et al. Reduced-representation sequencing identifies small effective population sizes of Anopheles gambiae in the north-western Lake Victoria basin, Uganda. Malar J. 2018;17.
28. M. DJ, Cuamba N., Charlwood JD, F. CH, Townson H. Population structure in the malaria vector, Anopheles arabiensis Patton, in East Africa. Heredity (Edinb). 1999;83:408–17.
29. Della Torre A, Tu Z, Petrarca V. On the distribution and genetic differentiation of Anopheles gambiae s.s. molecular forms. Insect Biochem Mol Biol. 2005;35:755–69.
30. Lukindu M, Bergey MC, Wiltshire MR, Small TS, Bourke PB, Kayondo KJ, et al. Spatio-temporal genetic structure of Anopheles gambiae in the Northwestern Lake Victoria Basin, Uganda: implications for genetic control trials in malaria endemic regions. Parasit Vectors. 2018;11.
31. Ngowo HS, Kaindoa EW, Matthiopoulos J, Ferguson HM, Okumu FO. Variations in household microclimate affect outdoor-biting behaviour of malaria vectors [version 1; referees: 2 approved, 1 approved with reservations]. Wellcome Open Res. 2017;2:1–18.
32. Okech BA, Gouagna LC, Killeen GF, Knols BGJ, Kabiru EW, Beier JC, et al. Influence of sugar availability and indoor microclimate on survival of Anopheles gambiae (Diptera: Culicidae) under semifield conditions in western Kenya. J Med Entomol. 2003;40:657–63.
33. Zhong D, Wang X, Xu T, Zhou G, Wang Y, Lee MC, et al. Effects of microclimate condition changes due to land use and land cover changes on the survivorship of malaria vectors in China-Myanmar border region. PLoS One. 2016;11.
34. Zhong D, Wang X, Xu T, Zhou G, Wang Y, Lee MC, et al. Effects of microclimate condition changes due to land use and land cover changes on the survivorship of malaria vectors in China-Myanmar border region. PLoS One. 2016;11.

35. Gillies MT, and, Coetzee M. A supplement to the Anopheline of Africa South of the Sahara (Afro-Tropical region). South African Institute for Medical Research. 1987;55.
36. Boddé M, Makunin A, Ayala D, Bouafou L, Diabaté A, Ekpo UF, et al. High-resolution species assignment of Anopheles mosquitoes using k-mer distances on targeted sequences. Elife. 2022;11:1–40.
37. Kingma DP, Welling M. Auto-Encoding Variational Bayes. 2013; Available from: http://arxiv.org/abs/1312.6114
38. Nagi SC. AmpSeeker. 2023.
39. Ochoa A, Storey JD. FST and kinship for arbitrary population structures I: Generalized definitions [Internet]. 2016. Available from: http://biorxiv.org/lookup/doi/10.1101/083915
40. Ochoa A, Storey JD. Estimating FST and kinship for arbitrary population structures. PLoS Genet. 2021;17.
41. Pritchard KJ, Stephens M, Donnelly P. Inference of Population Structure Using Multilocus Genotype Data. Genetics. 2000;155:945–959.
42. Ochola A. SJD. popkinsuppl: supplement to popkin package [Internet]. 2022. Available from: https://github.com/OchoaLab/popkinsuppl.
43. Bhatia G, Patterson N, Sankararaman S, Price AL. Estimating and interpreting F ST : The impact of rare variants. Genome Res. 2013;23:1514–21.
44. Arnold B, Corbett-Detig RB, Hartl D, Bomblies K. RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. Mol Ecol. 2013;22:3179–90.
45. Efron B, Tibshirani JR. An introduction to the Bootstrap. Cox DR, Hinkley D V, Reid N, Rubin DB, Silverman BW, editors. Regression Analysis with Applications B.G. Wetherill. Taylor & Francis Group; 1994.
46. Peakall R, Smouse PE. GenALEx 6.5: Genetic analysis in Excel. Population genetic software for teaching and research-an update. Bioinformatics. 2012;28:2537–9.
47. Diniz-Filho JAF, Soares TN, Lima JS, Dobrovolski R, Landeiro VL, Telles MP de C, et al. Mantel test in population genetics. Genet Mol Biol. 2013;36:475–85.
48. Patterson N, Price AL, Reich D. Population Structure and Eigenanalysis. PLoS Genet. 2006;2.
49. Team RC. R: A language and environment for statistical computing. R Foundation for statistical computing, Vienna, Austria. 2023.
50. Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. Mol Ecol. 2005;14:2611–20.
51. Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA, Mayrose I. Clumpak: A program for identifying clustering modes and packaging population structure inferences across K. Mol Ecol Resour. 2015;15:1179–91.
52. Nei M, Miller JC. A simple method for estimating average number of nucleotide substitutions within and between populations from restriction data. Genetics. 1990;125:873–9.
53. Tajima F. Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. Genetics Society of America. 1989;123:585–95.
54. Paradis E. Pegas: An R package for population genetics with an integrated-modular approach. Bioinformatics. 2010;26:419–20.
55. Waples, S. , R. & Do C. LDNE: a program for estimating effective population size from data on linkage disequilibrium. Mol Ecol Resour. 2008;8:753–6.

56. Do C, Waples RS, Peel D, Macbeth GM, Tillett BJ, Ovenden JR. NeEstimator v2: Re-implementation of software for the estimation of contemporary effective population size (Ne) from genetic data. Mol Ecol Resour. 2014;14:209–14.
57. F. R. Genetic Differentiation and Estimation of Gene Flow from FStatistics Under Isolation by Distance. Genetics. 1997;34:399–405.
58. N. CR. Variability within and among Natural populations. Evolution and the Genetics of Populations, Volume 4. Biometric Society; 1979. p. 359.
59. Slatkin M. Gene Flow and the Geographic Structure of Natural Populations [Internet]. 1987. Available from: https://www.science.org
60. Kalinowski ST. Do polymorphic loci require large sample sizes to estimate genetic distances? Heredity (Edinb). 2005;94:33–6.
61. Charlesworth B. Measures of Divergence Between Populations and the Effect of Forces that Reduce Variability. Mol Biol Evol [Internet]. 1998;15:538–43. Available from: https://academic.oup.com/mbe/article/15/5/538/987854
62. Hudson RR, Slatkint M, Maddison WP. Estimation of Levels of Gene Flow From DNA Sequence Data. 1992.
63. Smaragdov MG, Kudinov AA, Uimari P. Assessing the genetic differentiation of holstein cattle herds in the Leningrad region using Fst statistics. Agricultural and Food Science. 2018;27:96–101.
64. Bortolotto E, Bucklin A, Mezzavilla M, Zane L, Patarnello T. Gone with the currents: Lack of genetic differentiation at the circum-continental scale in the Antarctic krill Euphausia superba. BMC Genet. 2011;12:32.
65. Das S, Máquina M, Phillips K, Cuamba N, Marrenjo D, Saúte F, et al. Fine-scale spatial distribution of deltamethrin resistance and population structure of Anopheles funestus and Anopheles arabiensis populations in Southern Mozambique. Malar J. 2023;22.
66. Fumagalli M. Assessing the effect of sequencing depth and sample size in population genetics inferences. PLoS One. 2013;8.
67. Bi K, Linderoth T, Singhal S, Vanderpool D, Patton JL, Nielsen R, et al. Temporal genomic contrasts reveal rapid evolutionary responses in an alpine mammal during recent climate change. PLoS Genet. 2019;15.
68. Machado HE, Bergland AO, Taylor R, Tilk S, Behrman E, Dyer K, et al. Broad geographic sampling reveals the shared basis and environmental correlates of seasonal adaptation in drosophila. Elife. 2021;10.
69. Sandoval-Castellanos E. Testing temporal changes in allele frequencies: A simulation approach. Genet Res (Camb). 2010;92:309–20.
70. Diniz-Filho JAF, Soares TN, Lima JS, Dobrovolski R, Landeiro VL, Telles MP de C, et al. Mantel test in population genetics. Genet Mol Biol. 2013;36:475–85.
71. Lenz M, Muller FJ, Zenke M, Schuppert A. Principal components analysis and the reported low intrinsic dimensionality of gene expression microarray data. Sci Rep. 2016;6:1–11.
72. Chen H, Minakawa N, Beier J, Yan G. Population genetic structure of Anopheles gambiae mosquitoes on Lake victoria islands, west Kenya. Malar J. 2004;3:1–8.
73. Marshall JC, Pinto J, Charlwood JD, Gentile G, Santolamazza F, Simard FR, et al. Exploring the origin and degree of genetic isolation of anopheles gambiae from the islands of São Tomé and Príncipe, potential sites for testing transgenic-based vector control. Evol Appl. 2008;1:631–44.

74. Lehmann T, Light M, Gimnig EJ, Hightower A, Vulule MJ, Hawley AW. Spatial and temporal variation in kinship among Anopheles gambiae (Diptera: Culicidae) mosquitoes. J Med Entomol. 2003;40:421–9.
75. The Anopheles gambiae 1000 Genomes consortium. Genetic diversity of the African malaria vector Anopheles gambiae. Nature. 2017;552:96–100.
76. Pritchard, K., J., Stephens, M. and Donnelly P, Pritchard JK, Stephens M, Donnelly P. Inference of Population Structure Using Multilocus Genotype Data. Genetics. 2000;155:945–959.
77. Guillot G. Inference of structure in subdivided populations at low levels of genetic differentiation - The correlated allele frequencies model revisited. Bioinformatics. 2008;24:2222–8.
78. Weir BS, Cockerham CC. Estimating F-statistics for the analysis of population structure. Evolution (N Y). 1984;38:1358–70.
79. Holsinger KE, Weir BS. Genetics in geographically structured populations: Defining, estimating and interpreting FST. Nat Rev Genet. 2009. p. 639–50.
80. Jombart T, Devillard S, Balloux F. Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. BMC Genet. 2010;11.
81. Tajima F. Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. 1989.
82. Nielsen R. Molecular signatures of natural selection. Annu Rev Genet. 2005. p. 197–218.
83. Kreitman M. Methods to detect selection in populations with applications to the human. Annu Rev Genomics Hum Genet. 2000;1:539–59.
84. O'Loughlin SM, Magesa S, Mbogo C, Mosha F, Midega J, Lomas S, et al. Genomic analyses of three malaria vectors reveals extensive shared polymorphism but contrasting population histories. Mol Biol Evol. 2014;31:889–902.
85. Wei Y, He S, Wang J, Fan P, He Y, Hu K, et al. Genome-wide SNPs reveal novel patterns of spatial genetic structure in Aedes albopictus (Diptera Culicidae) population in China. Front Public Health. 2022;10.
86. Parimittr P, Chareonviriyaphap T, Bangs MJ, Arunyawat U. Genetic variation of Aedes aegypti mosquitoes across Thailand based on nuclear DNA sequences. Agriculture and Natural Resources. 2018;52:596–602.
87. Dharmarathne HAKM, Weerasena OVDSJ, Perera KLNS, Galhena G. Genetic characterization of Aedes aegypti (Diptera: Culicidae) in Sri Lanka based on COI gene. J Vector Borne Dis. 2020;57:153–60.
88. Larsson H, Källman T, Gyllenstrand N, Lascoux M. Distribution of long-range linkage disequilibrium and tajima's d values in scandinavian populations of norway spruce (picea abies). G3: Genes, Genomes, Genetics. 2013;3:795–806.
89. Nei M, Li WH. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci U S A. 1979;76:5269–73.
90. Besansky NJ, Krzywinski J, Lehmann T, Simard F, Kernt M, Mukabayire O, et al. Semipermeable species boundaries between Anopheles gambiae and Anopheles arabiensis: Evidence from multilocus DNA sequence variation. Proc Natl Acad Sci U S A. 2003;100:10818–23.
91. Obbard DJ, Linton YM, Jiggins FM, Yan G, Little TJ. Population genetics of Plasmodium resistance genes in Anopheles gambiae: No evidence for strong selection. Mol Ecol. 2007;16:3497–510.

1042 92. Campos M, Hanemaaijer M, Gripkey H, Collier TC, Lee Y, Cornel AJ, et al. The origin of
1043 island populations of the African malaria mosquito, Anopheles coluzzii. Commun Biol.
1044 2021;4.
1045 93. Bergey CM, Lukindu M, Wiltshire RM, Fontaine MC, Kayondo JK, Besansky NJ. Assessing
1046 connectivity despite high diversity in island populations of a malaria mosquito. Evol Appl.
1047 2020;13:417–31.
1048 94. Lanzaro GC, Campos M, Crepeau M, Cornel A, Estrada A, Gripkey H, et al. Selection of
1049 sites for field trials of genetically engineered mosquitoes with gene drive. Evol Appl.
1050 2021;14:2147–61.
1051 95. Ellegren H, Galtier N. Determinants of genetic diversity. Nat Rev Genet. 2016;17:422–33.
1052 96. Fischer MC, Rellstab C, Leuzinger M, Roumet M, Gugerli F, Shimizu KK, et al. Estimating
1053 genomic diversity and population differentiation - an empirical comparison of microsatellite
1054 and SNP variation in Arabidopsis halleri. BMC Genomics. 2017;18:1–15.
1055 97. O'Loughlin SM, Magesa SM, Mbogo C, Mosha F, Midega J, Burt A. Genomic signatures of
1056 population decline in the malaria mosquito Anopheles gambiae. Malar J. 2016;15:1–10.
1057 98. Lobo NF, Sangaré DM, Regier AA, Reidenbach KR, Bretz DA, Sharakhova M V., et al.
1058 Breakpoint structure of the Anopheles gambiae 2Rb chromosomal inversion. Malar J.
1059 2010;9.
1060 99. White BJ, Cheng C, Sangaré D, Lobo NF, Collins FH, Besansky NJ. The population genomics
1061 of trans-specific inversion polymorphisms in Anopheles gambiae. Genetics. 2009;183:275–
1062 88.
1063 100. Petrarca V, Nugud AD, Elkarim Ahmed MA, Haridi AM, Di Deco MA, Coluzzi M.
1064 Cytogenetics of the Anopheles gambiae complex in Sudan, with special reference to An.
1065 arabiensis: Relationships with East and West African populations. Med Vet Entomol.
1066 2000;14:149–64.
1067 101. Peter Andolfatto, Frantz Depaulis AN. Inversion polymorhisms and nucleotide variability
1068 in Drosophila. Genetics Research Cambridge. 2001;77:1–8.
1069 102. Wang J. A comparison of single-sample estimators of effective population sizes from
1070 genetic marker data. Mol Ecol. 2016;25:4692–711.
1071 103. Gilbert KJ, Whitlock MC. Evaluating methods for estimating local effective population
1072 size with and without migration. Evolution (N Y). 2015;69:2154–66.
1073 104. Multini LC, de Souza AL da S, Marrelli MT, Wilke ABB. The influence of anthropogenic
1074 habitat fragmentation on the genetic structure and diversity of the malaria vector Anopheles
1075 cruzii (Diptera: Culicidae). Sci Rep. 2020;10:1–13.
1076 105. Hare MP, Nunney L, Schwartz MK, Ruzzante DE, Burford M, Waples RS, et al.
1077 Understanding and Estimating Effective Population Size for Practical Application in Marine
1078 Species Management. Conservation Biology. 2010. p. 438–49.
1079 106. Waples RS, Do C. Linkage disequilibrium estimates of contemporary Ne using highly
1080 variable genetic markers: A largely untapped resource for applied conservation and
1081 evolution. Evol Appl. 2009;3:244–62.
1082 107. Wang J. Estimation of effective population sizes from data on genetic markers.
1083 Philosophical Transactions of the Royal Society B: Biological Sciences. Royal Society; 2005. p.
1084 1395–409.
1085 108. Rousset F. Genetic differentiation and estimation of gene flow from F-statistics under
1086 isolation by distance. Genetics. 1997;145:1219–28.
1087 109. Neafsey DE, Waterhouse RM. Highly evolvable malaria vectors: The genomes of 16
1088 Anopheles mosquitoes. 2015 [cited 2024 Aug 12]; Available from: http://dx.doi.

110. Hoffmann AA, Willi Y. Detecting genetic responses to environmental change. Nat Rev Genet. 2008. p. 421–32.
111. Weetman D, Wilding CS, Neafsey DE, Müller P, Ochomo E, Isaacs AT, et al. Candidate-gene based GWAS identifies reproducible DNA markers for metabolic pyrethroid resistance from standing genetic variation in East African Anopheles gambiae. Sci Rep. 2018;8.
112. Athrey G, Hodges TK, Reddy MR, Overgaard HJ, Matias A, Ridl FC, et al. The Effective Population Size of Malaria Mosquitoes: Large Impact of Vector Control. PLoS Genet. 2012;8.
113. Charlesworth B. Effective population size and patterns of molecular evolution and variation. Nature Reviews|Genetics. 2009;10:195–205.
114. Yaro AS, Traoré AI, Huestis DL, Adamou A, Timbiné S, Kassogué Y, et al. Dry season reproductive depression of Anopheles gambiae in the Sahel. J Insect Physiol [Internet]. 2012;58:1050–9. Available from: http://dx.doi.org/10.1016/j.jinsphys.2012.04.002
115. White MT, Griffin JT, Churcher TS, Ferguson NM, Basáñez MG, Ghani AC. Modelling the impact of vector control interventions on Anopheles gambiae population dynamics. Parasit Vectors. 2011;4.
116. Kamdem C, Fouet C, Gamez S, White BJ. Pollutants and Insecticides Drive Local Adaptation in African Malaria Mosquitoes. Mol Biol Evol. 2017;34:1261–75.
117. Waples RS. Life-history traits and effective population size in species with overlapping generations revisited: The importance of adult mortality. Heredity (Edinb). 2016;117:241–50.
118. Mwima R, Hui TYJ, Kayondo JK, Burt A. The population genetics of partial diapause, with applications to the aestivating malaria mosquito Anopheles coluzzii. Mol Ecol Resour. 2024;24.
119. Mwima R, Hui TYJ, Nanteza A, Burt A, Kayondo JK. Potential persistence mechanisms of the major Anopheles gambiae species complex malaria vectors in sub-Saharan Africa: a narrative review. Malar J [Internet]. 2023 [cited 2024 Jun 24];22:1–12. Available from: https://malariajournal.biomedcentral.com/articles/10.1186/s12936-023-04775-0

Supplementary Information (SI)

Table S1. Nucleotide diversity and Tajima's D for the six *An. gambiae* populations. The last row is all *An. gambiae* (2918) combined for comparison. Site choice was as described in the main text. The asterisk * denotes statistical significance, ($p < 0.05$). † Denotes an island population.

Population	**Nucleotide diversity (π)**	**Tajima's D**
Kayonjo	0.01163	-2.3069*
Katuuso	0.01143	-2.0219*
Kibbuye	0.01169	-2.2356*
Kiimi †	0.01081	-0.8223
Kansambwe †	0.01071	-1.0827
Bugiri †	0.01103	-0.6954
All *An. gambiae* (Mainland & Island)	0.01150	-2.3073*

1121
1122

1123 **Table S2.** Nucleotide diversity and Tajima's D for the three *An. arabiensis* populations

1124 The last row is all *An. arabiensis* (173) combined for comparison. Site choice was as described

1125 in the main text.

Population	**Nucleotide diversity (π)**	**Tajima's D**
Kayonjo	0.00859	-1.17475
Katuuso	0.00829	-0.98993
Kibbuye	0.00847	-1.73386
All *An. arabiensis*	0.00859	-1.91156

1126
1127