






## ORIGINAL ARTICLE

# Assessing connectivity despite high diversity in island populations of a malaria mosquito

Christina M. Bergey<sup>1,2,3,4</sup>  | Martin Lukindu<sup>1,2</sup>  | Rachel M. Wiltshire<sup>1,2</sup>  |  
Michael C. Fontaine<sup>5,6</sup>  | Jonathan K. Kayondo<sup>7</sup> | Nora J. Besansky<sup>1,2</sup> 

<sup>1</sup>Department of Biological Sciences, University of Notre Dame, Notre Dame, IN, USA

<sup>2</sup>Eck Institute for Global Health, University of Notre Dame, Notre Dame, IN, USA

<sup>3</sup>Department of Genetics, Rutgers University, Piscataway, NJ, USA

<sup>4</sup>Departments of Anthropology and Biology, Pennsylvania State University, University Park, PA, USA

<sup>5</sup>Groningen Institute for Evolutionary Life Sciences (GELIFES), University of Groningen, Groningen, The Netherlands

<sup>6</sup>MIVEGEC, IRD, CNRS, University of Montpellier, Montpellier, France

<sup>7</sup>Department of Entomology, Uganda Virus Research Institute (UVRI), Entebbe, Uganda

## Correspondence

Christina M. Bergey and Nora J. Besansky, Department of Biological Sciences, University of Notre Dame, Notre Dame, IN 46556, USA.

Emails: christina.bergey@rutgers.edu (C.M.B.) and nbesansk@nd.edu (N.J.B.)

## Funding information

Open Philanthropy Project Fund; National Institute of Allergy and Infectious Diseases, Grant/Award Number: R01 AI125360 and R21 AI123491; Bill & Melinda Gates Foundation, Grant/Award Number: P30CA016087; Silicon Valley Community Foundation; NIH, Grant/Award Number: R01 AI125360 and R21 AI123491; New York University School of Medicine

## Abstract

Documenting isolation is notoriously difficult for species with vast polymorphic populations. High proportions of shared variation impede estimation of connectivity, even despite leveraging information from many genetic markers. We overcome these impediments by combining classical analysis of neutral variation with assays of the structure of selected variation, demonstrated using populations of the principal African malaria vector *Anopheles gambiae*. Accurate estimation of mosquito migration is crucial for efforts to combat malaria. Modeling and cage experiments suggest that mosquito gene drive systems will enable malaria eradication, but establishing safety and efficacy requires identification of isolated populations in which to conduct field testing. We assess Lake Victoria islands as candidate sites, finding one island 30 km offshore is as differentiated from mainland samples as populations from across the continent. Collectively, our results suggest sufficient contemporary isolation of these islands to warrant consideration as field-testing locations and illustrate shared adaptive variation as a useful proxy for connectivity in highly polymorphic species.

## KEYWORDS

*Anopheles gambiae*, gene drive technology, gene flow, malaria, migration

## 1 | INTRODUCTION

The difficulties in estimating migration with genetic methods are exacerbated for large, interconnected populations exhibiting shallow population structure. Large population sizes result in high levels

of polymorphism in the genome and impede accurate estimation of connectivity (Waples, 1998) and discernment of demographic independence from panmixia (Waples, 2006). Population genetic methods for estimating migration using neutral markers may thus have limited utility when such a high proportion of diversity is shared

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2019 The Authors. *Evolutionary Applications* published by John Wiley & Sons Ltd

between populations, a failing that is only partially redressed with the high quantity of markers available from massively parallel sequencing. The most powerful window into migration may instead be the distribution of selected variants (Gagnaire et al., 2015).

The major African malaria vector *Anopheles gambiae* Giles, 1902 sensu stricto (henceforth *An. gambiae*) is among the most genetically diverse eukaryotic species (Miles et al., 2017), with shallow population structure (Lehmann et al., 2003; Miles et al., 2017) that complicates efforts to estimate connectivity from genetic data. Overcoming these obstacles to infer migration accurately is crucial for control efforts to reduce the approximately 445,000 annual deaths attributable to malaria (World Health Organization, 2017). Such vector control efforts include novel methods involving the release of genetically modified mosquitoes. The most promising involve introducing transgenes into the mosquito genome or its endosymbionts that interrupt pathogen transmission coupled with a gene drive system to propagate the effector genes through a population (Alphey, 2014; Burt, 2014; Champer, Buchman, & Akbari, 2016). Such systems have recently been successfully engineered in the laboratory (Gantz et al., 2015; Hammond et al., 2015, 2017). A detailed understanding of population structure and connectivity is essential for effective implementation of any genetic control method, not least a gene drive system designed to spread in a super-Mendelian fashion.

Here, we analyze population structure, demographic history, and migration between populations from genomewide variation in *An. gambiae* mosquitoes living near and on the Ssese archipelago of Lake Victoria in Uganda (Figure 1). We augment these analyses with a demonstration of our framework using selective sweep sharing as a proxy for connectivity. We propose that our approach will be useful for inferring migration in taxa with high variation. Islands present natural laboratories for disentangling the determinants of population structure, as gene flow—likely important in post-dry season recolonization (Dao et al., 2014)—is reduced. In addition to the high malaria prevalence of the islands (44% in children; 30% in children country-wide; Uganda Bureau of Statistics (UBOS) and ICF, 2017), we were motivated by the potential of such an island to be a field site for future tests of gene drive vector control strategies: Geographically isolated islands have been proposed as locales to test the dynamics of transgene spread while limiting their movement beyond the study population (Alphey, 2002; James, 2005; James et al., 2018; World Health Organization, 2014). Antecedent studies of population structure and connectivity of potential release sites are crucial to evaluate the success of such field trials, as well as to quantify the chance of migration of transgenic insects carrying constructs designed to propagate across mosquito populations and country borders.

## 2 | MATERIALS AND METHODS

### 2.1 | Experimental design

Mosquitoes were sampled from five of the Ssese Islands in Lake Victoria, Uganda (Banda, Bukasa, Bugala, Nsadzi, and Sserinya), and four mainland sampling localities (Buwama, Kaazi, Kiyindi, and



**FIGURE 1** Map of Lake Victoria Basin (LVB) study area. Map of study area showing sampling localities on Ssese Islands (blue) and mainland localities (red) in LVB. The Ag1000G reference population, Nagongera, Tororo District, is not shown, but lies 111 km NE of Kiyindi, 57 km from the shore of Lake Victoria. Map data copyright 2018 Google

Wamala) at varying distances from the lake in May and June 2015. Sampling took place between 4:40 and 8:15 over a 30-day period as follows: Indoor resting mosquitoes were collected from residences via mouth or mechanical aspirators and subsequently identified morphologically to species group. Female mosquitoes assigned to the *An. gambiae* sensu lato complex based on morphology ( $N = 575$ ) were included in further analyses. All mosquitoes were preserved with silica desiccant and transported to the University of Notre Dame, Indiana, USA, for analysis.

### 2.2 | DNA extraction, library preparation, and whole-genome sequencing

Animals were assigned to species level via a PCR-based assay (Scott, Brogdon, & Collins, 1993) using DNA present in a single leg or wing. DNA from individual *An. gambiae* s. s.  $N = 116$  mosquitoes was extracted from the whole body via phenol–chloroform extraction (Green & Sambrook, 2012) and then quantified via fluorometry (PicoGreen). Automated library preparation took place at the New York University Langone Medical Center with the Biomek SPRIWorks HT system using KAPA Library Preparation Kits, and libraries were sequenced on the Illumina HiSeq 2500 with 100 paired end cycles.

### 2.3 | Mapping and SNP calling, filtering

Software version information is provided in Table S10. After quality filtering and trimming using ea-utils' fastq-mcf (-l 15 -q 15 -w 4;

Aronesty, 2011), reads were mapped to the *An. gambiae* reference genome (AgamP4 PEST; Holt et al., 2002; Sharakhova et al., 2007) using BWA aln and sampe with default parameters (Li & Durbin, 2009).

After realignment around indels with GATK's IndelRealigner, variants were called using GATK's UnifiedGenotyper (with `-stand_call_conf 50.0` and `-stand_emit_conf 10.0`; selected to be consistent with methods of recent comparison SNP dataset; Miles et al., 2017) and filtered for quality (DePristo et al., 2011), excluding SNPs with `QualByDepth < 2.0`, `RMSMappingQuality < 40.0`, `FisherStrand > 60.0`, `HaplotypeScore > 13.0`, or `ReadPosRankSum < -8.0`. All bioinformatic steps for read mapping and variant identification are encapsulated in the NGS-map pipeline (<https://github.com/bergeycm/NGS-map>). This yielded 33.1 million SNPs. Three individuals sequenced to lower coverage (4.3–5.3 $\times$ ) were included to maximize sample size, and the following filtering steps were applied to remove sequencing errors. Individuals and variants with high levels of missingness (>10%) and variants that were not biallelic or exhibited values of Hardy-Weinberg equilibrium (HWE) that were likely due to sequencing error ( $p < .00001$ ) were excluded from further analysis (as extreme departure from HWE is an indicator of a likely technical error in sequencing or genotyping.) For use in population structure inference, the SNP dataset was further pruned for linkage disequilibrium by sliding a window that is 50 SNPs long across the genome in 5 SNP increments and recursively removing random SNPs in any pairs with  $r^2 > .5$  using PLINK (Chang et al., 2015; Purcell et al., 2007). After filtration, the dataset contained 28,569,621 SNPs before LD pruning and 115 individuals. SNPs unpruned for linkage disequilibrium were phased with SHAPEIT2 (Delaneau, Howie, Cox, Zagury, & Marchini, 2013) using an effective population size ( $N_e$ ) of 1,000,000 (consistent with previous demographic modelling; Miles et al., 2017), default MCMC parameters (7 burn-in MCMC iterations, 8 pruning iterations, and 20 main iterations), conditioning states for haplotype estimation ( $K = 100$ ), and window size of 2 Mb.

## 2.4 | Population structure inference

To explore population structure in a larger, continent-wide context, we merged our Lake Victoria Basin (LVB) SNP set with a recently published dataset of *An. gambiae* individuals (from the Ag1000G project) collected between 2000 and 2012 from Angola, Burkina Faso, Guinea-Bissau, Guinea, Cameroon, Gabon, Uganda, and Kenya (Miles et al., 2017). Prior to filtering, biallelic SNPs from the LVB and Ag1000G datasets were merged using bcftools (Li et al., 2009). We excluded any SNP with >10% missingness in either dataset, any SNPs that did not pass the accessibility filter of the Ag1000G dataset, and SNPs with minor allele frequency (MAF) <1%. After this filtration, our merged SNP dataset contained 12,537,007 SNPs.

After pruning the merged dataset for LD (leaving 9,861,756 SNPs) and excluding laboratory crosses from the Ag1000G dataset (leaving 881 individuals), we assigned individuals' genomes to ancestry components using ADMIXTURE to better understand population structure in the LVB (Alexander, Novembre, & Lange,

2009). We created 10 replicate samples of 100,000 SNPs from chromosome 3 (prior to LD pruning and restricted to avoid the well-known inversions on other chromosomes), including only biallelic SNPs in euchromatic regions with MAF > 1%. These replicate datasets were pruned for LD by randomly selecting from pairs of SNPs with  $r^2 > .01$  in sliding windows of size 500 SNPs and with a stepsize of 250 SNPs. For each replicate, we ran ADMIXTURE for five iterations in fivefold cross-validation mode for values of  $k$  clusters from 2 to 10. This resulted in 50 estimates for each value of  $k$ . We assessed these results using the online version of CLUMPAK with default settings to ensure the stability of the resulting clustering (Kopelman, Mayzel, Jakobsson, Rosenberg, & Mayrose, 2015). CLUMPAK clusters the replicate runs' Q-matrices to produce a major cluster for each value of  $k$ , which we then visualized. The lowest cross-validation error was found for  $k = 6$  clusters, but we also display ancestry estimates with  $k = 9$  clusters to further explore patterns of structure with a level of subdivision at which the Ssese Island individuals are assigned a unique ancestry component.

We visualized population structure via principal components analysis (PCA) with PLINK (Chang et al., 2015; Purcell et al., 2007), using the LVB-Ag1000G merged dataset (excluding the outlier, highly inbred Kenyan population; Miles et al., 2017) and 3,212,485 chromosome 3 SNPs (to avoid the common inversions on chromosome 2 and the X chromosome) outside of heterochromatic regions (such as centromeric regions; Sharakhova et al., 2007; Table S11). We next performed a PCA on the LVB dataset alone, pruning for LD and low-MAF (<1%) SNPs on chromosome 3. Based on the results of these analyses, we split individuals from the large island of Bugala into two clusters for subsequent analyses: those that cluster with mainland individuals and those that cluster with individuals from the smaller islands.

For the LVB dataset, we computed the pairwise fixation index ( $F_{ST}$ ) between-locality samples for *An. gambiae* using the unbiased estimator of Hudson (Hudson, Slatkin, & Maddison, 1992) as implemented in smartpca (Patterson, Price, & Reich, 2006; Price et al., 2006). To obtain overall values between-sampling sites, per-SNP values were averaged across the genome excluding common inversions (2La, 2Rb, and 2Rc) and heterochromatic regions. We also computed z-scores via block jackknife, using 42 blocks of size 5 Mb. We tested for isolation by distance, or a correlation between genetic and geographic distances, with a Mantel test (Mantel, 1967) as implemented in the R package ade4 (Dray & Dufour, 2007), using these  $F_{ST}$  estimates and Euclidean geographic distances between localities.

To estimate fine-scale structure and relatedness between LVB individuals, we estimated the proportion of pairs of individuals genomes that are identical by descent (IBD) using PLINK (Chang et al., 2015; Purcell et al., 2007) and assuming a constant recombination rate of 2.0 cM/Mb (after Clarkson et al., 2018) since we lacked a recombination map. We excluded heterochromatic and inversion regions and retained informative pairs of SNPs within 500 kb in the pairwise population concordance test.

## 2.5 | Diversity estimation

We computed numerous diversity metrics to test the hypothesis that island and mainland sites differed in these key measures for vector control. Grouping individuals by site (except for Bugala, which was split based on the results of the PCA), we calculated nucleotide diversity ( $\pi$ ) and Tajima's  $D$  in nonoverlapping windows of size 10 kb, the inbreeding coefficient ( $F$ ) estimated with the method of moments, minor allele frequencies (the site frequency spectrum, SFS), and a measure of linkage disequilibrium ( $r^2$ ) using VCFtools (Danecek et al., 2011). The inbreeding coefficient ( $F$ ) was estimated with the method of moments as implemented in VCFtools (Danecek et al., 2011), as  $F_i = (O_i - E_i)/(N_i - E_i)$ , where, for individual  $i$ ,  $O_i$  is the total observed number of loci which are homozygous,  $E_i$  is the total expected number of loci homozygous by chance, and  $N_i$  is the total number of genotyped loci. For  $r^2$ , we computed the measure for all SNPs (unpruned for linkage) within 50 kb of a random set of 100 SNPs with MAF > 10% and corrected for differences in sample size by subtracting  $1/n$ , where  $n$  equaled the number of sampled chromosomes per site, after (Miles et al., 2017). To visualize decay in LD, we plotted  $r^2$  between SNPs against their physical distance in base pairs, first smoothing the data to aid in visualization by fitting a generalized additive model (GAM) to them. We also inferred runs of homozygosity using PLINK (Chang et al., 2015; Purcell et al., 2007) to compare their length ( $F_{ROH}$ ), requiring 10 homozygous SNPs spanning a distance of 100 kb and allowing for three heterozygous and five missing SNPs in the window. Runs of homozygosity were inferred using LD-pruned SNPs outside of inversions or heterochromatic regions. We tested the significance of differences in these statistics between island and mainland categories using a two-sided Wilcoxon rank-sum test.

## 2.6 | Demographic history inference

To estimate the contemporary or short-term  $N_e$  for each site, we inferred regions of IBD from unphased data with IBDseq (Browning & Browning, 2013) and analyzed them with IBDNe (Browning & Browning, 2015). We restricted our analysis to SNPs from chromosome 3 to avoid common inverted regions. We allowed a minimum IBD tract length of 0.005 cM (or 5 kb), scaling it down from the recommended length for human genomes due to mosquitoes' high level of heterozygosity (Miles et al., 2017) and assumed a constant recombination rate of 2.0 cM/Mb (after Clarkson et al., 2018).

To estimate the long-term evolutionary demographic history of mosquitoes on and near the Ssesse Islands, including a long-term estimate of  $N_e$ , we inferred population demographic history for each site via stairway plots using the full site frequency spectra based on SNPs on chromosome 3 with heterochromatic regions and regions within 5 kb of a gene excluded (Liu & Fu, 2015).

We also inferred a "two-population" isolation-with-migration (IM) demographic model with  $\delta a \delta i$  (Coffman, Hsieh, Gravel, & Gutenkunst, 2016; Gutenkunst, Hernandez, Williamson, & Bustamante, 2009) in which the ancestral population splits to form

two daughter populations that are allowed to grow exponentially and exchange migrants asymmetrically. This modeling allowed us to infer whether pairs of sites, including mainland and island pairs, were demographically independent and had deep split times, consistent with a greater degree of isolation. For  $\delta a \delta i$ -based analyses, we used the full dataset of SNPs on chromosome 3, not pruned for LD but with heterochromatic regions and regions within 5 kb of a gene masked. We polarized the SNPs using outgroup information from *Anopheles merus* and *A. merus* (Fontaine et al., 2014). We fit this two-population model and the same model without migration to all pairs of locality samples, choosing the optimal model using the Godambe information matrix and an adjusted likelihood ratio test to compare the two nested models. We compared the test statistic to a  $\chi^2$  distribution and rejected the null model if the  $p$ -value for the test statistic was <.05. For both, singletons and doubletons private to one population were masked from the analysis and a parameter encompassing genotype uncertainty was included in the models and found to be low (mean 0.70%). We assessed the goodness of fit visually using the residuals of the comparison between model and data frequency spectra (Figure S7). Using the site frequency spectrum, we projected down to 2–6 fewer chromosomes than the total for the smaller population to maximize information given missing data. We set the grid points to  $\{n, n + 10, n + 20\}$ , where  $n$  = the number of chromosomes. Bounds for  $N_e$  scalars were  $\nu \in (0.01, 10,000)$ , for time were  $T \in (1e-8, 0.1)$ , for migration were  $m \in (1e-8, 10)$ , and for genotyping uncertainty were  $p_{\text{misid}} \in (1e-8, 1)$ . Parameters were perturbed before allowing up to 1,000 iterations for optimization. We estimated parameter uncertainty using the Fisher information matrix and 100 bootstrap replicates of 1 Mb from the dataset. If the Hessian was found to be not invertible when computing the Fisher information matrix, the results of that iteration were excluded from the analysis. For population size change parameters,  $\nu$ , optimized values for one or both populations were often close to the upper limit. Due to this runaway behavior, common in analyses of the SFS (Rosen, Bhaskar, & Song, 2018), we excluded the population size change from our interpretation.

To translate  $\delta a \delta i$ - and stairway plot-based estimates of  $N_e$  and time to individuals and years respectively, we assumed a generation time of 11 per year and a mutation rate of  $3.5e-9$  per generation (Miles et al., 2017).

## 2.7 | Selection inference

To infer candidate genes and regions with selection histories that varied geographically, we compared allele frequencies and haplotype diversity between the sampling sites. To infer differing selection between-sampling sites, we computed  $F_{ST}$  between all populations in windows of size 10 kb using the estimator of Weir and Cockerham (1984) (as implemented in VCFtools; Danecek et al., 2011), and H12 (as implemented in SelectionHapStats; Garud, Messer, Buzbas, & Petrov, 2015) and XP-EHH on a per-site basis (as implemented in selscan; Szpiech & Hernandez, 2014) to detect long stretches of homozygosity in a given population considered alone or relative to

another population (Sabeti et al., 2007). For XP-EHH, EHH was calculated in windows of size 100 kb in each direction from core SNPs, allowing EHH decay curves to extend up to 1 Mb from the core, and SNPs with MAF < 0.05 were excluded from consideration as a core SNP. As we lacked a fine-scale genetic map for *Anopheles*, we assumed a constant recombination rate of 2.0 cM/Mb (after Clarkson et al., 2018). Scores were normalized within chromosomal arms and the X chromosome. The between-locality statistics,  $F_{ST}$  and XP-EHH, were summarized using the composite selection score (CSS; Randhawa, Khatkar, Thomson, & Raadsma, 2014; Wallberg, Pirk, Allsopp, & Webster, 2016).

We plotted these statistics across the genome to identify candidate regions with signatures of selection, including high differentiation between samples from different localities, reduced variability within a sample, and extended haplotype homozygosity. To identify regions of the genome showing signatures of selection specific to certain geographic areas, we identified genomic regions with elevated H12 in a subset of localities and confirmed both elevated differentiation (as inferred from  $F_{ST}$ ) and evidence of differing selective sweep histories (as inferred from XP-EHH). Excluding the mainland-like portion of Bugala (to focus on its putative ancestral island population as opposed to recent migrants from the mainland), we identified putative locality-specific sweeps (H12 over 99th percentile in one population), island-specific sweeps (H12 over 99th percentile in four or more of the five island localities but 0 or 1 mainland localities), or LVB mainland-specific sweeps (H12 over 99th percentile in three or more of the four mainland localities but zero or one island localities). To place these putative sweeps in their continental context, for the region of each putative locality-, island-, or LVB mainland-specific sweep, we determined whether the H12 values of each of the Ag1000G populations (excluding Kenya due to its signatures of potential admixture and recent population decline; Miles et al., 2017) were in the top 5% for that population, indicating a possible selective sweep at the same location.

We further explored the haplotype structure and putative functional impact of loci for which we detected signatures of potential selection to determine the count and geographic distribution of independent selective sweeps. To provide necessary context for the reconstruction of sweeps and quantify long-distance haplotype sharing between populations, we included data from several other *An. gambiae* populations across Africa (Burkina Faso, Cameroon, Gabon, Guinea, Guinea-Bissau, Kenya, and other Ugandan individuals; Miles et al., 2017). We computed the pairwise distance matrix as the raw number of base pairs that differed and grouped haplotypes via hierarchical clustering analysis (implemented in the hclust R function) in regions of size 100 kb centered on each peak in pairwise  $F_{ST}$  or XP-EHH, or the average of peaks, in the case for multiple nearby spikes. As short terminal branches can result from a beneficial allele and linked variants rising to fixation during a recent selective sweep, we identified such clusters by cutting the tree at a height of 0.4 SNP differences per kb.

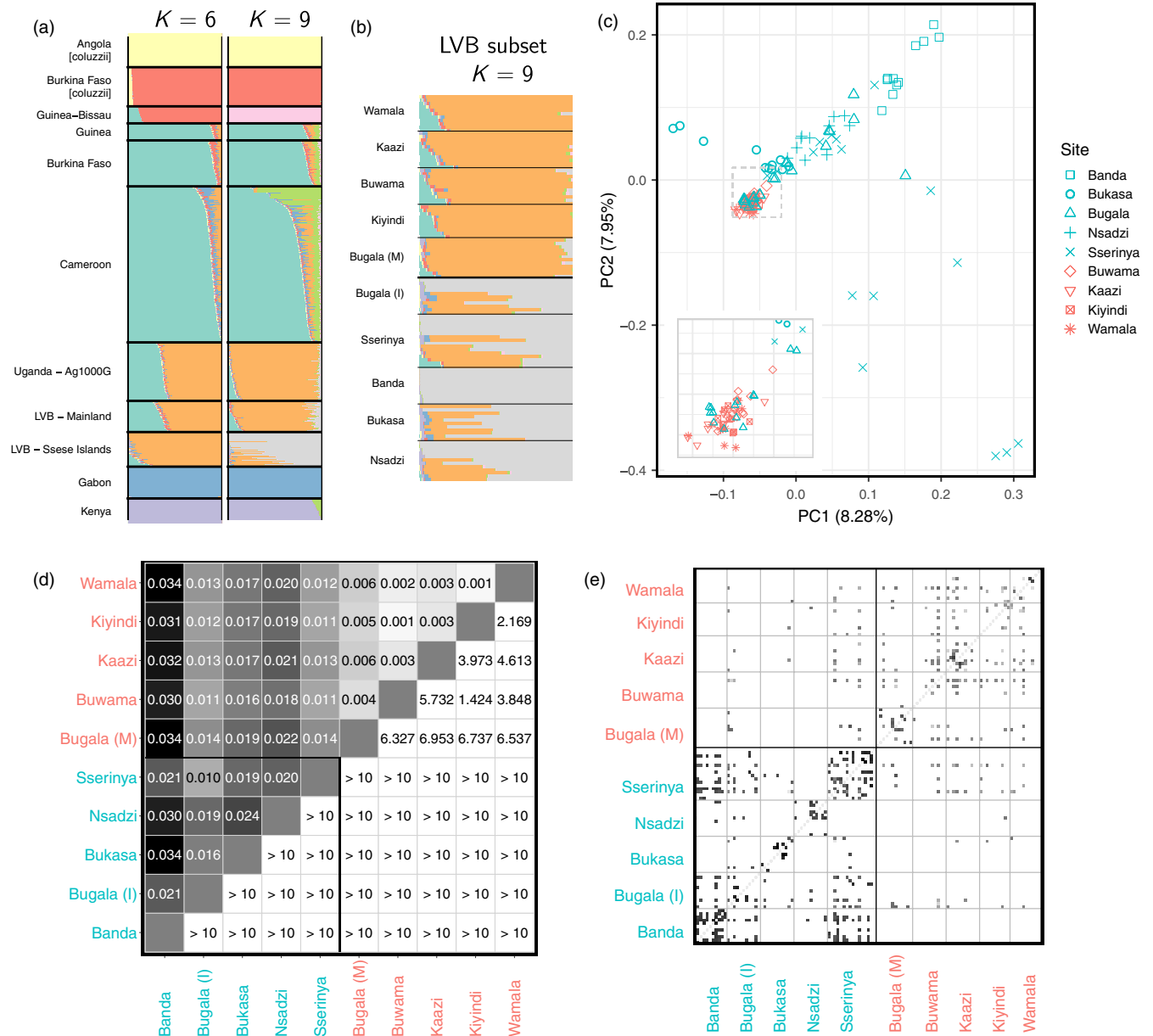
### 3 | RESULTS

The Ssesse Islands are approximately 4–50 km from the mainland (farther than the average flight distance of *An. gambiae*; Verdonschot & Besse-Lototskaya, 2014) and vary in size, infrastructure, and accessibility. Sampled islands range from Banda—a small, largely forested island of approximately 1 km<sup>2</sup> with a single settlement—to Bugala—296 km<sup>2</sup>, site of a 10,000 ha oil palm plantation (Zeemeijer, 2012), and linked to the mainland via ferry service (*Kalangala District Local Government District Management Improvement Plan 2012–2015*, 2012). To explore the partitioning of *An. gambiae* genetic variation in the LVB, we sequenced the genomes of 116 mosquitoes from five island and four mainland localities (Figure 1, Table S1). We sequenced 10–23 individuals per site to an average depth of  $17.6 \pm 4.6$  (Table S2). After filtering, we identified 28.6 million high-quality single nucleotide polymorphisms (SNPs). To provide additional, continent-wide context for the LVB population structure, we merged our dataset with that of the *An. gambiae* 1,000 Genomes project (Ag1000G; Miles et al., 2017) for a combined dataset of 12.54 million SNPs (9.86 million after linkage disequilibrium pruning) in 881 individuals.

#### 3.1 | Genetic structure

We analyzed LVB population structure with context from continent-wide populations (Miles et al., 2017) of *An. gambiae* and sister species *Anopheles coluzzii* mosquitoes (formerly known as *An. gambiae* M molecular form; Coetzee et al., 2013). Both Bayesian clustering (Alexander et al., 2009; Figure 2a) and principal component analysis (PCA; Figure S1) showed LVB individuals closely related to the Ugandan reference population (Nagongera, Tororo; 0°46′12.0″N, 34°01′34.0″E; ~57 km from Lake Victoria; Figure 1). With  $\geq 6$  clusters (which optimized predictive accuracy in the clustering analysis and had the lowest cross-validation error; Figure S2), island samples had distinct ancestry proportions (Figure 2a), and beginning with  $k = 9$  clusters, we observed additional subdivision in LVB samples and the assignment of the majority of Ssesse individuals' ancestry to a largely island-specific component, indicated here in gray (Figures 2a, 2, S3).

Principal components analysis of only LVB individuals (based on chromosome 3 to avoid the well-known inversions on chromosome 2 and the X chromosome) indicated little differentiation among mainland samples in the first two components and varying degrees of differentiation on islands, with Banda, Sserinya, and Bukasa the most extreme (Figure 2c). Twelve of 23 individuals from Bugala, the largest, most developed, and most connected island, exhibited affinity to mainland individuals instead of ancestry typical of the islands (Figure S4). As both PCA and clustering analyses revealed this differentiation, we split the Bugala sample into mainland- and island-like subsets for subsequent analyses (hereafter referenced as “Bugala (M)” and “Bugala (I),” respectively). Individuals with partial ancestry attributable to the component prevalent on the mainland

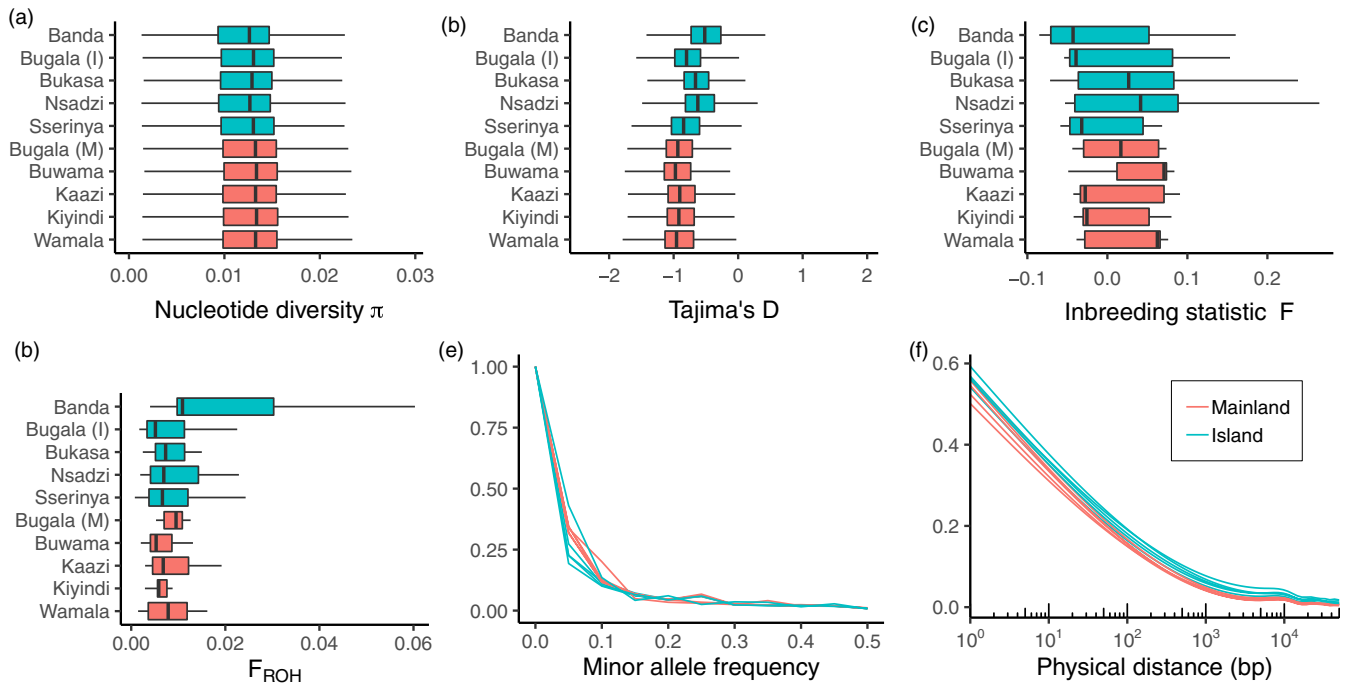


**FIGURE 2** Population structure in the Lake Victoria Basin (LVB). Analyses are based on chromosome 3 to avoid segregating inversions on other chromosomes, unless otherwise noted. (a) ADMIXTURE-inferred ancestry of individuals in LVB. Results based on analysis of LVB and Ag1000G merged dataset. Analysis is restricted to *Anopheles gambiae* s. s. Clustering shown for  $k = 6$  clusters, which minimizes cross-validation error, and  $k = 9$  clusters, the lowest  $k$  for which island individuals have the majority of their ancestry assigned to an island-specific cluster. (b) Results of the clustering analysis with  $k = 9$  clusters for LVB individuals, split by sampling locality. (c) Plot of first two components of PCA of Lake Victoria Basin individuals showing locality of origin. Mainland individuals are colored red, while island individuals are blue, and point shape indicates sampling locality. Based on these results and that of ADMIXTURE analysis, the island sample of Bugala was split into mainland- and island-like subpopulations ("Bugala (M)" and "Bugala (I)," respectively) for subsequent analyses (Figure S4). (d) Heatmap of  $F_{ST}$  between sites (lower triangle) and associated z-score computed via block jackknife (upper triangle). "Bugala (M)" and "Bugala (I)" are the mainland- and island-like subpopulations of Bugala. (e) Proportion of genomewide pairwise IBD sharing between individuals, based on the full genome. Each small square represents a comparison between two individuals, and darker colors indicate a higher proportion of the two genomes is in IBD, shaded on a logarithmic scale. Individuals are grouped by locality

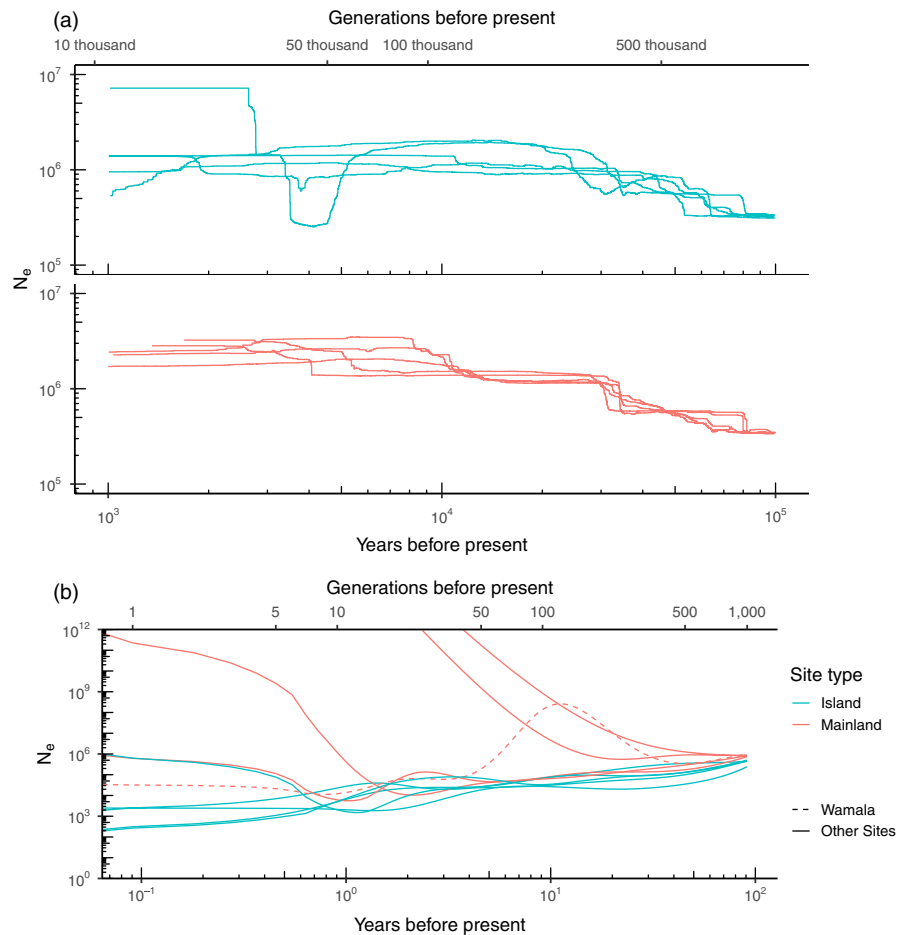
and the rest to the island-specific component were present on all islands except Banda.

Within the LVB, differentiation concurred with observed population structure. Mean  $F_{ST}$  between-sampling localities (range: 0.001–0.034) was approximately 0 ( $\leq 0.003$ ) for mainland-mainland comparisons and was highest in comparisons involving the

small island Banda (Figure 2d). Geographic distances and  $F_{ST}$  were uncorrelated (Mantel  $p = .88$ ; Figure S5). Island samples showed greater within- and between-locality sharing of genomic regions IBD (inferred using a constant recombination rate of 2.0 cM/Mb after [Clarkson et al., 2018] since we lacked a recombination map), with sharing between nearby islands Sserinya, Banda and Bugala



**FIGURE 3** Diversity metrics in the Lake Victoria Basin samples. Shown are (a) boxplot of nucleotide diversity ( $\pi$ ; in 10 kb windows), (b) boxplot of Tajima's  $D$  (in 10 kb windows), (c) boxplot of inbreeding statistic ( $F$ ), (d) boxplot of length of runs of homozygosity ( $F_{ROH}$ ), (e) histogram of minor allele frequency (MAF), and (f) decay in linkage disequilibrium ( $r^2$ ), all grouped by sampling locality. For all boxplots, outlier points are not shown



**FIGURE 4** Population history of the Lake Victoria Basin samples. (a) Long-term evolutionary population histories inferred via stairway plots for island and mainland samples. (b) Contemporary or short-term effective population size ( $N_e$ ) history inferred using sharing of regions that are identical by descent (IBD). Wamala, a mainland locality showing island-like fluctuations in population size, is indicated with a dashed line. Plot truncated to exclude implausibly high estimates that are likely an artifact of sample size

(Figure 2e). Importantly, Banda Island shared no IBD regions with mainland sites, underscoring its contemporary lack of gene flow with the mainland. However, we also detected shared IBD between Banda and nearby Sserinya Island, indicating the potential for gene flow from Banda to the mainland via an interisland route.

### 3.2 | Genetic diversity

Consistent with the predicted decrease in genetic variation for semi-isolated island populations due to inbreeding and smaller effective population sizes ( $N_e$ ), islands displayed slightly lower nucleotide diversity ( $\pi$ ; Wilcoxon rank-sum test  $p < .001$ ; Figure 3a), a higher proportion of shared to rare variants (Tajima's  $D$ ;  $p < .001$ ; Figure 3b), and more linkage among SNPs (LD;  $r^2$ ; Fig.  $p < .001$ ; Figure 3f). They were, however, similar in inbreeding coefficient ( $F$ ;  $p = .0719$ ; Figure 3c), number of long runs of homozygosity ( $F_{ROH}$ ;  $p = .182$ ; Figure 3d) and proportions of low-frequency SNPs (Figure 3e). The small island Banda was the most extreme in these measures.

### 3.3 | Demographic history

To test islands for isolation and demographic independence from the mainland, we inferred the population history of LVB samples by estimating long-term and recent trends in  $N_e$  using stairway plots (Liu & Fu, 2015) based on the site frequency spectrum (SFS; Figure 4a) and patterns of IBD sharing (Browning & Browning, 2015; Figure 4b), respectively. We assumed a generation time of 11 per year. Short-term final mainland sizes were unrealistically high, likely due to low sample sizes for each locality, and should be interpreted with caution given our use of a constant recombination rate for IBD inference in the absence of a recombination map. Nonetheless, differences between islands and mainland sites were informative. In both, islands had consistently lower  $N_e$  compared to mainland populations extending back 500 generations (~50 years) and often severely fluctuated, particularly in the last 250 generations (~22 years). Mainland sites Wamala and Kaazi had island-like recent histories, with Wamala abruptly switching to an island-like pattern around 2005.

To all pairs of LVB localities, we fit an IM demographic model using  $\delta a \delta i$ , in which an ancestral population splits into two populations, allowing exponential growth and continuous asymmetrical migration between the daughter populations (Figures S6, S7). In all comparisons involving islands and some between mainland sites, the best fitting model as chosen via AIC had zero migration (Tables S3–S5). Time since population split was much more recent for mainland-mainland comparisons (excluding Bugala, median: 361 years) than those involving islands (island–island median: 7,128 years; island-mainland median: 4,043 years). Island–island split time confidence intervals typically did not overlap those involving mainland sites.

### 3.4 | Selection

As adaptive variants would be the most likely signatures of past gene flow to persist, we next examined signatures of selective sweeps for

insight into migration. Identifying signatures of selection in the same genomic region in populations with independent lineages would be consistent with several scenarios (Stern, 2013): (i) independent parallel selective sweeps on de novo mutations, (ii) independent parallel selective sweeps on shared ancestral variation, or (iii) selective sweeps on variants transferred via gene flow. As we were most interested in the transfer of adaptive variants for its insight into migration (iii), we distinguished between the alternative scenarios as follows.

We would expect independent sweeps on novel mutations (i) to exhibit differences in genetic background between the two populations, evidenced by distinct haplotype clusters, each comprising near-identical haplotypes separated by individual haplotypes lacking signatures of a selective sweep. In both other scenarios (ii–iii), we would instead expect haplotypes with the sweep to group together when clustered by genetic distance. By itself, haplotype information does not differentiate sweeps targeting standing ancestral variation (ii) from those targeting adaptive variants spread through gene flow (iii). However, additional information such as geographic distance between the populations, estimates of gene flow inferred from other regions of the genome, and assessment of gene flow between other nearby populations, may suggest that one of these scenarios is the more likely.

While the sharing of a sweep may indicate migration between populations, the inverse would be suggestive—though not conclusive—of barriers to gene flow. A lack of sharing of a selective sweep signal between two populations may indicate no migration is occurring. However, it would also be consistent with the occurrence of migration that is subsequently countered by the local effects of selection or lost to genetic drift.

We first compared mainland Uganda and the Ssesse Islands, reasoning that shared signatures of sweeps at a genomic location may indicate migration is occurring with the islands, while the absence was suggestive—but not conclusive—of isolation. We identified sweeps in the LVB using genome scans of between- and within-locality statistics, including  $F_{ST}$  (Weir & Cockerham, 1984, Figure S8), extended haplotype homozygosity (XP-EHH, Sabeti et al., 2007, Figure S8), and haplotype homozygosity (H12, Garud et al., 2015, Figure S9). To test for sweeps that were variable within the LVB, we identified locality-specific sweeps (found at only one sampling site in the LVB), sweeps that were found in our island localities but not mainland LVB localities, and sweeps that were found only in our mainland LVB localities (all defined as  $H12 > 99$ th percentile). To add additional country-level context, we then intersected these regions with those under putative selection in a mainland Ugandan reference population ( $H12 > 95$ th percentile; Miles et al., 2017).

Some genomic locations had heterogeneous selection signals within the LVB and within Uganda, indicative of potential geographic barriers to gene flow or local variation in selective regimes. Locality-specific putative sweeps were more prevalent on island than LVB mainland localities (mean per locality: island = 52.4; mainland = 26.8), concordant with increased isolation of the islands (Table S6). Sweeps detected only or primarily in mainland LVB localities were shared



with the Ag1000G Ugandan reference population more often (8 of 37; 22%) than those found only or primarily on islands (1 of 21, 5%; Tables S7 and S8), again indicative of some barriers to gene flow with the islands.

We next reasoned that continent-wide selective sweeps, with broadly distributed selective advantage, would be the most likely to be shared via gene flow. Widespread sweeps that were absent or at extremely low frequency on the islands would be a strong suggestion against contemporary gene flow (e.g., since the advent of insecticide use), and those that conversely were present on the islands would be indicative that gene flow had occurred, if the alternative scenarios could be excluded as outlined above. To identify these regions, we intersected our set of sweeps with those under putative selection in populations across the continent ( $H_{12} > 95$ th percentile in Ag1000G; Miles et al., 2017).

As expected, outlier regions included known selective sweep targets from elsewhere in Africa (Miles et al., 2017, Table S9). All sweeps found in the reference Uganda population (Miles et al., 2017) were detected in at least some sampling localities in our LVB dataset, except the sweep targeting *Vgsc*, which was excluded during filtration of the heterochromatic region adjacent to the centromere (2L:1–2431617). For instance, the large genomic region spanning the cluster of insecticide resistance-associated cytochrome P450s (*Cyp6p*) on chromosome arm 2R, including *Cyp6p3* which is upregulated in mosquitoes with permethrin and bendiocarb resistance (Edi et al., 2014), exhibited low diversity, an excess of low-frequency polymorphisms (Tajima's  $D$ ), and elevated haplotype homozygosity ( $H_{12}$ ) within the LVB populations (Figures S9 and S10). Pairwise statistics ( $F_{ST}$  and XP-EHH) indicated low differentiation between LVB localities, as expected for a continent-wide sweep (Figure S8). The signal was found in every LVB site, including all islands. Hierarchical clustering of LVB and Ag1000G haplotypes revealed clades with low interhaplotype diversity, expected after selection rapidly increases the frequency of a haplotype containing adaptive variation (Figure S11). Consistent with previous results (Miles et al., 2017), these clusters of closely related haplotypes on independent lineages indicate that multiple parallel sweeps targeting the *Cyp6p* region have occurred in several genetic backgrounds at numerous localities across Africa. Within Uganda, since almost all mainland and island individuals carry haplotypes from a single cluster, the selected haplotype of this cluster likely spread to near-fixation via gene flow.

In contrast, some sweeps with continent-wide prevalence including the reference Ugandan population (Miles et al., 2017) were found at all mainland LVB sites but had colonized the islands incompletely. For example, a region on chromosome arm 2L (2L:2,900,000–3,000,000) was found in all assayed Ag1000G populations and LVB mainland sites, but found on no island but Sserinya (Table S8). As in previous studies (Miles et al., 2017), independent clusters of low-diversity haplotypes in varied genetic backgrounds suggest multiple sweeps targeting the cluster of genes encoding glutathione S-transferases (*Gste1-Gste7*), including one sweep specific to Uganda. This Ugandan sweep was similarly confined largely to the mainland in the LVB. These sweeps at targets of selection throughout the continent

that are largely restricted to the mainland are suggestive of strong barriers to gene flow to the islands, either due to lack of connectivity or the countering effects of selection or drift. Other sweeps had colonized the islands incompletely. The sweep targeting cytochrome P450 gene *Cyp9k1*, likely linked to pyrethroid resistance (Vontas et al., 2018), probably arose multiple times independently, since Ugandan haplotypes do not cluster with low-diversity clusters from elsewhere in Africa. Within the LVB, the sweep signature is found on some, but not all islands, suggesting some barrier to gene flow or local selection limiting the spread of the sweep.

Two regions exhibited selection signals similar in amplitude to known insecticide-related loci, with elevated between-locality differentiation, low diversity, and extended homozygosity (Figures S8, S9, S12, and S13). The first, at 2L:34.1 Mb, contains many genes, including a cluster involved in chorion formation (Amenya et al., 2010) near the signal peak. Haplotype clustering revealed a group of closely related Ugandan individuals, consistent with a geographically bounded selective sweep (Figure S14). The selected variation had not fully colonized the islands or the LVB mainland sites, however, suggesting some barriers to gene flow, loss due to drift at some localities, or local differences in selective pressure within the LVB. Elsewhere in Africa, clustering analysis revealed other low-variation clades in distinct genetic backgrounds in, for example, Cameroon and Angola, suggesting parallel selection on independent mutations at this locus.

The second putative sweep, at X:9.2 Mb, coincided precisely with eye-specific diacylglycerol kinase (AGAP000519, X:9,215,505–9,266,532). Low-diversity haplotypes formed a single cluster including LVB haplotypes overwhelmingly from the islands (Figure S15). Transfer via gene flow between islands but not to the mainland is reasonable, given the connectivity patterns we have inferred from neutral variation. Additionally, local selection may be countering the spread of the sweep to the mainland. However, more surprisingly, these island haplotypes with evidence of a selective sweep were most closely related to haplotypes from distant locations, primarily Gabon and Burkina Faso rather than Uganda. This sharing of extended haplotypes between islands and distant localities is consistent with either gene flow or independent sweeps targeting ancestral standing variation. Of these alternatives, extremely long-distance gene flow that persists only on islands seems less likely.

## 4 | DISCUSSION

Although the perfect field-testing site for gene drive mosquitoes would lack any outward migration, mosquitoes' massive dispersal potential renders the existence of such a site doubtful. However, a genetically modified construct created to induce population suppression would be expected to reduce the transgenic and overall mosquito population to low levels or absence in a period of years (James et al., 2018). Identification of a geographically bounded site with minimal expected migration over such a short period is a more tractable goal than finding a completely isolated population. Here,

we have used a combination of classical population genetic techniques and those relying on adaptive variants to assess islands in the LVB as possible field-testing sites for transgenic mosquitoes. We have found that the probability of contemporary migration (e.g., migration over the past several years) may be sufficiently low to qualify some Ssesse Islands as candidate field sites, worthy of more intensive sampling and scrutiny.

Understanding the population genetics of island *An. gambiae* has both evolutionary and practical importance. A limited number of genetic investigations have been conducted on oceanic (Maliti et al., 2014; Marsden et al., 2013; Marshall et al., 2008; Moreno et al., 2007; Salgueiro, Moreno, Simard, O'Brochta, & Pinto, 2013) and lacustrine islands (Chen, Minakawa, Beier, & Yan, 2004; Kayondo et al., 2005; Lukindu et al., 2018; Wiltshire et al., 2018), though most have been limited in the type or count of molecular markers used. Of the estimates of gene flow from previous studies of oceanic or island gene flow, for instance, only one relied on more than a few dozen SNPs (Wiltshire et al., 2018), with the rest based on fewer SNPs (Marsden et al., 2013) or markers such as microsatellites (Chen et al., 2004; Kayondo et al., 2005; Maliti et al., 2014; Moreno et al., 2007), transposable elements (Salgueiro et al., 2013), or mitochondrial or ribosomal loci (Lukindu et al., 2018; Marshall et al., 2008). In contrast to shallow population structure across Africa (Lehmann et al., 2003; Miles et al., 2017), partitioning of genetic variation on islands suggests varying isolation. Using a genomewide dataset, we found differentiation between the Ssesse Islands to be relatively high in the context of continent-wide structure, with the differentiation between Banda Island (only 30 km offshore) and mainland localities on par with or higher than for populations on opposite sides of the continent (e.g., Banda vs Wamala,  $F_{ST} = 0.034$ ; mainland Uganda vs Burkina Faso,  $F_{ST} = 0.007$ ; Miles et al., 2017). The Ssesse Islands are approximately as differentiated as all but the most outlying oceanic islands tested (e.g., mainland Tanzania vs Comoros, 690–830 km apart,  $F_{ST} = 0.199$ – $0.250$ ; however, note that the estimate is based on only 31 SNP loci; Marsden et al., 2013). Patterns of haplotype sharing did include direct evidence for the recent exchange of migrants between nearby islands, but analyses based on haplotype sharing, Bayesian clustering, and demographic reconstruction included no evidence of direct sharing between Banda and the mainland. Banda is nonetheless connected to other islands and thereby indirectly connected to the mainland, and additional sampling may reveal signs of admixture. Additional sampling on Banda and other islands that are disjunct from the rest of the archipelago would be prudent when assessing potential field-testing locations.

The name "Ssesse" derives from another arthropod vector, the tsetse fly (*Glossina* spp.) The tsetse-mediated arrival of sleeping sickness in 1902 brought "enormous mortality" (Thomas, 1941, p. 332) to the 20,000 residents, who were evacuated in 1909 (Hale Carpenter, 1920; Thomas, 1941). Though encouraged to return by 1920, the human population numbered only 4,000 in 1941 (Thomas, 1941) and took until 1980 to double (Uganda Bureau of Statistics, 2002), but has since rapidly risen to over 62,000 (2015, projected; Kalangala District Local Government District Management Improvement

Plan 2012–2015, 2012; Uganda Bureau of Statistics, 2016). The impacts on mosquito populations of this prolonged depression in human population size, coupled with water barriers to mosquito migration, are reflected in the distinctive demographic histories of island *An. gambiae* populations, which were smaller and fluctuated more than mainland localities, echoing previous results (Kayondo et al., 2005; Wiltshire et al., 2018). Two mainland sites had island-like recent population histories, with Wamala abruptly switching from a mainland-like to island-like growth pattern around 2005. This coincides precisely with a  $\geq 20\%$  reduction from 2000 to 2010 in the *Plasmodium falciparum* parasite rate (PfPR<sub>2–10</sub>; a measure of malaria transmission intensity) in Mityana, the district containing Wamala (National Malaria Control Programme et al., 2013).

Though previous *Anopheles* population genetic studies have inferred gene flow even among species (Crawford et al., 2016; Miles et al., 2017), the SFS-based demographic models with the best fit suggested that no genetic exchange had occurred since the split between island sites and between islands and the mainland. Island pairs were inferred to have split far deeper in the past (5,000–14,000 years ago) than mainland sites (typically < 500 years ago), on par with the inferred split time between Uganda and Kenya (approximately 4,000 years ago; Miles et al., 2017). Although bootstrapping-derived confidence intervals permit some certainty, our model fit is not optimal likely due to low sample sizes and high levels of shared ancestral variation, and additional sampling is necessary to clarify population history. Our inferred lack of gene flow to the islands appears contradictory to the presence of individuals who share ancestry with the mainland on all islands but Banda. We cannot dismiss the possibility that this indicates actual migration occurs. If so, effects of migration would have to be sufficiently countered by local selection to limit its effect on allele frequency spectra, rendering effective migration (as estimated in population history inference) zero. The apparent contradiction can also be resolved if shared ancestry between islands and mainland suggested by the clustering result is interpreted as retention of shared ancestral polymorphism or the existence of inadequately sampled ancestral variation (Lawson, Dorp, & Falush, 2018), rather than recent admixture. This interpretation is consistent with the affinity we observed between the Ssesse Islands and West Africa in the structure of adaptive variation.

Discerning whether the absence of observed gene flow is due to lack of connectivity, the opposition of selection (possibly differing between island and mainland sites), or the stochasticity of genetic drift is difficult. Instead, we must rely on estimates of the strength of selection in the two locales to inform our conclusions. For example, we would expect that an insecticide sweep found over a large region in Africa would spread in island mosquito populations with insecticide treated bed nets, despite the considerable effect of genetic drift in small populations. As insecticide treated bed net usage is present on the islands (Kalangala District Local Government District Management Improvement Plan 2012–2015, 2012), variation conferring a major selective advantage related to insecticides would be expected to spread to and persist on the islands if migration allows the transfer, and the strongest evidence of a lack of contemporary

connectivity is therefore the absence of a sweep on the islands that is widespread on the continent.

We found two sweeps on insecticide-related genes that are common targets of selection elsewhere but which have incompletely colonized the Ssesse Islands: one on cytochrome P450 monooxygenase *Cyp9K1* (Fossog Tene et al., 2013; Vontas et al., 2018) present on some islands, and another on glutathione S-transferase genes (*Gste1-Gste7*; Enayati, Ranson, & Hemingway, 2005; Fouet, Kamdem, Gamez, & White, 2017; Jones et al., 2012; Mitchell et al., 2014) at extremely low frequency on the islands. That the selective sweeps targeting these loci (Miles et al., 2017) have not fully colonized the islands despite the advantage in detoxifying pyrethroids and DDT suggests a lack of contemporary exchange (e.g., since the advent of insecticide use). However, the sweep targeting the *Cyp6p* cluster was found on all islands, confirming past gene flow had occurred at some point. (The insecticide resistance this likely indicates should be considered in planning potential field trials.) Although these distributions confirm that past migration from the mainland to islands has occurred and we are unable to exclude low levels of contemporary gene flow, taken together our data are consistent with potentially high degrees of gene flow restriction on contemporary timescales for some islands of the Ssesse archipelago.

Our investigation also identified two previously unknown signatures of selection. For the first, on chromosome arm 2L and encompassing many genes, haplotypes with sweeps in distinct genetic backgrounds across Africa suggest the region has been affected by multiple independent convergent sweeps. In Uganda, most individuals with the sweep are from the mainland, suggesting a local origin and spread via short-distance migration. The putative target of the second sweep is diacylglycerol kinase (AGAP000519) on the X chromosome, a homolog of retinal degeneration A (*rdgA*) in *Drosophila*. The gene is highly pleiotropic, contributing to signal transduction in the fly visual system (Hardie et al., 2002; Huang, Xie, & Wang, 2015), but also olfactory (Kain et al., 2008) and auditory (Senthilan et al., 2012) sensory processing. It has been recently implicated in nutritional homeostasis in *Drosophila* (Nelson et al., 2016) and is known to interact with the target of rapamycin (TOR) pathway (Lin et al., 2014), which has been identified as a target of ecological adaptation in *Drosophila* (De Jong & Bochdanovits, 2003; Fabian et al., 2012) and *An. gambiae* (Cheng, Tan, Hahn, & Besansky, 2018). The sweep appears largely confined to island individuals in the LVB, but the cluster of haplotypes also includes those from Gabon, Burkina Faso, and Kenya. Shared extended haplotypes suggest a single sweep event spread by gene flow or selection on standing ancestral variation, not independent selection on de novo mutations. Possible explanations include long-distance migration of an adaptive variant persisting on only the islands or, more reasonably, selection on standing ancestral variation. We have not found obvious candidate targets of selection, for example coding changes, which may be due to imperfect annotation of the genome or the likely possibility that the target is a non-coding regulator of transcription or was filtered from our dataset. Further functional studies would be needed to clarify the selective advantage that these haplotypes confer.

Population structure investigations are paramount for informing the design and deployment of control strategies, including field trials of transgenic mosquitoes. We demonstrate alternatives to simple extrapolation of migration rates from differentiation, which is fraught (Whitlock & McCauley, 1999) particularly given the assumption of equilibrium between the evolutionary forces of migration and drift (Storfer, Murphy, Spear, Holderegger, & Waits, 2010; Stow & Magnusson, 2012; Whitlock & McCauley, 1999), an unlikely state for huge *An. gambiae* populations (Gagnaire et al., 2015). We suggest that future assessments of connectivity include, as we have, the spatial distribution of adaptive variation, identification of recent migrants via haplotype sharing, and demographic history modeling, from which we have inferred the Ssesse Islands to be relatively isolated on contemporary time scales. Though we cannot exclude the possibility of a small amount of gene flow over evolutionary time between our most isolated islands and the mainland, the data are consistent with a sufficiently low amount of contemporary gene flow that it becomes reasonable to consider these islands as isolated on short time frames.

A completely isolated population of mosquitoes is not a reasonable expectation given mosquitoes' propensity for active and even passive (human-aided or windborne) dispersal (James et al., 2018), potentially up to hundreds of kilometers (Dao et al., 2014). Although no population of mosquitoes on an island, lacustrine or oceanic, is completely genetically isolated, such localities may still be ideal for initial gene drive field testing, as the geographic barriers maximize isolation to the extent possible (James et al., 2018), and absolute isolation on evolutionary timescales is unnecessary given the relatively short timeframe of small-scale field tests (e.g., several years). Thus, the probability of contemporary migration may be sufficiently low to qualify some Ssesse Islands as candidate field sites. Additionally, the assessment of the islands' suitability as potential sites for field trials of genetically modified mosquitoes must also consider the logistical ease of access and monitoring that the bounded geography of a small lacustrine island with low human population density affords initial field tests. Due consideration should be provided to these characteristics of small lake islands that may be appealing to regulators, field scientists, local communities, and other stakeholders. Given such features and the probable rarity of migration, the Ssesse Islands may be logical and tractable candidates for initial field tests of genetically modified *An. gambiae* mosquitoes, warranting further entomological study.

## ACKNOWLEDGEMENTS

The authors would like to thank the UVRI field entomology team: Christine Babirye, Ronald Mayanja, Paul Maweje, Kevin Nakato, and Fred Ssenfuka. We thank Nicholas Harding and Alistair Miles for helpful discussion. This study was supported by Target Malaria, which receives core funding from the Bill & Melinda Gates Foundation and from the Open Philanthropy Project Fund, an advised fund of Silicon Valley Community Foundation, through sub-contracts to J.K.K. and N.J.B. N.J.B. also received support from NIH

R01 AI125360 and R21 AI123491. The New York University School of Medicine's Genome Technology Center is partially supported by the Cancer Center Support Grant P30CA016087 at the Laura and Isaac Perlmutter Cancer Center.

## CONFLICT OF INTEREST

The authors declare no competing financial interests.

## AUTHOR CONTRIBUTIONS

C.M.B., J.K.K., and N.J.B. designed the study; C.M.B., M.L., R.M.W., and J.K.K. collected biological samples; C.M.B. analyzed the data; C.M.B., M.C.F., and N.J.B. wrote the manuscript; M.C.F., J.K.K., and N.J.B. supervised the research; C.M.B., M.L., R.M.W., M.C.F., J.K.K., and N.J.B. edited the manuscript.

## DATA AVAILABILITY STATEMENT

All scripts used in the analysis are available at [https://github.com/bergeycm/Anopheles\\_gambiae\\_structure\\_LVB](https://github.com/bergeycm/Anopheles_gambiae_structure_LVB) and released under the GNU General Public License v3. Sequencing read data for the LVB individuals are deposited in the NCBI Short Read Archive (SRA) under BioProject accession PRJNA493853.

## ORCID

Christina M. Bergey  <https://orcid.org/0000-0001-8336-8078>

Martin Lukindu  <https://orcid.org/0000-0002-1447-3525>

Rachel M. Wiltshire  <https://orcid.org/0000-0001-5199-4883>

Michael C. Fontaine  <https://orcid.org/0000-0003-1156-4154>

Nora J. Besansky  <https://orcid.org/0000-0003-0646-0721>

## REFERENCES

- Alexander, D., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19, 1655–1664. <https://doi.org/10.1101/gr.094052.109>
- Alphey, L. (2002). Malaria control with genetically manipulated insect vectors. *Science*, 298(5591), 119–121. <https://doi.org/10.1126/science.1078278>
- Alphey, L. (2014). Genetic control of mosquitoes. *Annual Review of Entomology*, 59(1), 205–224. <https://doi.org/10.1146/annurev-ento-011613-162002>
- Amenya, D. A., Chou, W., Li, J., Yan, G., Gershon, P. D., James, A. A., & Marinotti, O. (2010). Proteomics reveals novel components of the *Anopheles gambiae* eggshell. *Journal of Insect Physiology*, 56(10), 1414–1419. <https://doi.org/10.1016/j.jinsphys.2010.04.013>
- Aronesty, E. (2011). *ea-utils: Command-line tools for processing biological sequencing data*. [Computer software]. Retrieved from <https://expressionanalysis.github.io/ea-utils>
- Browning, B. L., & Browning, S. R. (2013). Detecting identity by descent and estimating genotype error rates in sequence data. *American Journal of Human Genetics*, 93(5), 840–851. <https://doi.org/10.1016/j.ajhg.2013.09.014>
- Browning, S. R., & Browning, B. L. (2015). Accurate non-parametric estimation of recent effective population size from segments of identity by descent. *American Journal of Human Genetics*, 97(3), 404–418. <https://doi.org/10.1016/j.ajhg.2015.07.012>
- Burt, A. (2014). Heritable strategies for controlling insect vectors of disease. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 369(1645), 20130432. <https://doi.org/10.1098/rstb.2013.0432>
- Champer, J., Buchman, A., & Akbari, O. S. (2016). Cheating evolution: Engineering gene drives to manipulate the fate of wild populations. *Nature Reviews Genetics*, 17(3), 146–159. <https://doi.org/10.1038/nrg.2015.34>
- Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, 4(1), 7. <https://doi.org/10.1186/s13742-015-0047-8>
- Chen, H., Minakawa, N., Beier, J., & Yan, G. (2004). Population genetic structure of *Anopheles gambiae* mosquitoes on Lake Victoria islands, west Kenya. *Malaria Journal*, 3, 48. <https://doi.org/10.1186/1475-2875-3-48>
- Cheng, C., Tan, J. C., Hahn, M. W., & Besansky, N. J. (2018). Systems genetic analysis of inversion polymorphisms in the malaria mosquito *Anopheles gambiae*. *Proceedings of the National Academy of Sciences*, 115(30), E7005–E7014. 201806760. <https://doi.org/10.1073/pnas.1806760115>
- Clarkson, C. S., Miles, A., Harding, N. J., Weetman, D., Kwiatkowski, D., Donnelly, M., & The *Anopheles gambiae* 1000 Genomes Consortium. (2018). The genetic architecture of target-site resistance to pyrethroid insecticides in the African malaria vectors *Anopheles gambiae* and *Anopheles coluzzii*. *bioRxiv*, <https://www.biorxiv.org/content/early/2018/08/06/323980>. <https://doi.org/10.1101/323980>
- Coetzee, M., Hunt, R. H., Wilkerson, R., della Torre, A., Coulbaly, M. B., & Besansky, N. J. (2013). *Anopheles coluzzii* and *Anopheles amharicus*, new members of the *Anopheles gambiae* complex. *Zootaxa*, 3619(3), 246–274. <https://doi.org/10.11646/zootaxa.3619.3.2>
- Coffman, A. J., Hsieh, P. H., Gravel, S., & Gutenkunst, R. N. (2016). Computationally efficient composite likelihood statistics for demographic inference. *Molecular Biology and Evolution*, 33(2), 591–593. <https://doi.org/10.1093/molbev/msv255>
- Crawford, J. E., Riehle, M. M., Markianos, K., Bischoff, E., Guelbeogo, W. M., Gneme, A., ... Lazzaro, B. P. (2016). Evolution of GOUNDRY, a cryptic subgroup of *Anopheles gambiae* s.l., and its impact on susceptibility to *Plasmodium* infection. *Molecular Ecology*, 25(7):1494–1510. <https://doi.org/10.1111/mec.13572>
- Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., ... Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15), 2156–2158. <https://doi.org/10.1093/bioinformatics/btr330>
- Dao, A., Yaro, A. S., Diallo, M., Timbiné, S., Huestis, D. L., Kassogué, Y., ... Lehmann, T. (2014). Signatures of aestivation and migration in Sahelian malaria mosquito populations. *Nature*, 516(7531), 387–390. <https://doi.org/10.1038/nature13987>
- De Jong, G., & Bochdanovits, Z. (2003). Latitudinal clines in *Drosophila melanogaster*: Body size, allozyme frequencies, inversion frequencies, and the insulin-signalling pathway. *Journal of Genetics*, 82(3), 207–223. <https://doi.org/10.1007/BF02715819>
- Delaneau, O., Howie, B., Cox, A. J., Zagury, J. F., & Marchini, J. (2013). Haplotype estimation using sequencing reads. *American Journal of Human Genetics*, 93(4), 687–696. <https://doi.org/10.1016/j.ajhg.2013.09.002>
- DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., ... Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5), 491–498. <https://doi.org/10.1038/ng.806>

- Dray, S., & Dufour, A. B. (2007). The ade4 package: Implementing the duality diagram for ecologists. *Journal of Statistical Software*, 22(4), 1–20.
- Edi, C. V., Djogbénou, L., Jenkins, A. M., Regna, K., Muskavitch, M. A., Poupardin, R., ... Weetman, D. (2014). CYP6 P450 enzymes and ACE-1 duplication produce extreme and multiple insecticide resistance in the malaria mosquito *Anopheles gambiae*. *PLoS Genetics*, 10(3), e1004236. <https://doi.org/10.1371/journal.pgen.1004236>
- Enayati, A. A., Ranson, H., & Hemingway, J. (2005). Insect glutathione transferases and insecticide resistance. *Insect Molecular Biology*, 14(1), 3–8. <https://doi.org/10.1111/j.1365-2583.2004.00529.x>
- Fabian, D. K., Kapun, M., Nolte, V., Kofler, R., Schmidt, P. S., Schlotterer, C., & Flatt, T. (2012). Genome-wide patterns of latitudinal differentiation among populations of *Drosophila melanogaster* from North America. *Molecular Ecology*, 21(19), 4748–4769. <https://doi.org/10.1111/j.1365-294X.2012.05731.x>
- Fontaine, M. C., Pease, J. B., Steele, A., Waterhouse, M., Neafsey, D. E., Sharakhov, I. V., ... Besansky, N. J. (2014). Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science*, 347(6217), 1258524. <https://doi.org/10.1126/science.1258522>
- Fossog Tene, B., Poupardin, R., Costantini, C., Awono-Ambene, P., Wondji, C. S., Ranson, H., & Antonio-Nkondjio, C. (2013). Resistance to DDT in an urban setting: Common mechanisms implicated in both M and S forms of *Anopheles gambiae* in the city of Yaoundé Cameroon. *PLoS ONE*, 8(4), e61408. <https://doi.org/10.1371/journal.pone.0061408>
- Fouet, C., Kamdem, C., Gamez, S., & White, B. J. (2017). Genomic insights into adaptive divergence and speciation among malaria vectors of the *Anopheles nili* group. *Evolutionary Applications*, 10(9), 897–906. <https://doi.org/10.1111/eva.12492>
- Gagnaire, P. A., Broquet, T., Aurelle, D., Viard, F., Souissi, A., Bonhomme, F., ... Bierne, N. (2015). Using neutral, selected, and hitchhiker loci to assess connectivity of marine populations in the genomic era. *Evolutionary Applications*, 8(8), 769–786. <https://doi.org/10.1111/eva.12288>
- Gantz, V. M., Jasinskiene, N., Tatarenkova, O., Fazekas, A., Macias, V. M., Bier, E., & James, A. A. (2015). Highly efficient Cas9-mediated gene drive for population modification of the malaria vector mosquito *Anopheles stephensi*. *Proceedings of the National Academy of Sciences*, 112(49), E6736–E6743. <https://doi.org/10.1073/pnas.1521077112>
- Garud, N. R., Messer, P. W., Buzbas, E. O., & Petrov, D. A. (2015). Recent selective sweeps in North American *Drosophila melanogaster* show signatures of soft sweeps. *PLoS Genetics*, 11(2), 1–32. <https://doi.org/10.1371/journal.pgen.1005004>
- Green, M. R., & Sambrook, J. (2012). *Molecular cloning: A laboratory manual*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.
- Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, 5(10), e1000695. <https://doi.org/10.1371/journal.pgen.1000695>
- Hale Carpenter, G. D. (1920). *A naturalist on Lake Victoria*. New York, NY: E. P. Dutton and Company. <https://doi.org/10.15713/ins.mmj.3>
- Hammond, A., Galizi, R., Kyrou, K., Simoni, A., Siniscalchi, C., Katsanos, D., ... Nolan, T. (2015). A CRISPR-Cas9 gene drive system targeting female reproduction in the malaria mosquito vector *Anopheles gambiae*. *Nature Biotechnology*, 34(1), 78–83. <https://doi.org/10.1038/nbt.3439>
- Hammond, A. M., Kyrou, K., Bruttini, M., North, A., Galizi, R., Karlsson, X., ... Nolan, T. (2017). The creation and selection of mutations resistant to a gene drive over multiple generations in the malaria mosquito. *PLoS Genetics*, 13(10), e1007039. <https://doi.org/10.1371/journal.pgen.1007039>
- Hardie, R., Martin, F., Cochrane, G., Juusola, M., Georgiev, P., & Raghu, P. (2002). Molecular basis of amplification in *Drosophila* phototransduction: Roles for G protein, phospholipase C, and diacylglycerol kinase. *Neuron*, 36(4), 689–701. [https://doi.org/10.1016/S0896-6273\(02\)01048-6](https://doi.org/10.1016/S0896-6273(02)01048-6)
- Holt, R., Subramanian, G., Halpern, A., Sutton, G., Charlab, R., Nusskern, D., ... Hoffman, S. (2002). The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science*, 298(October 2002), 129–149. <https://doi.org/10.1126/science.1076181>
- Huang, Y., Xie, J., & Wang, T. (2015). A fluorescence-based genetic screen to study retinal degeneration in *Drosophila*. *PLoS ONE*, 10(12), 1–19. <https://doi.org/10.1371/journal.pone.0144925>
- Hudson, R. R., Slatkin, M., & Maddison, W. P. (1992). Estimation of levels of gene flow from DNA sequence data. *Genetics*, 132(2), 583–589.
- James, A. A. (2005). Gene drive systems in mosquitoes: Rules of the road. *Trends in Parasitology*, 21(2), 64–67. <https://doi.org/10.1016/j.pt.2004.11.004>
- James, S., Collins, F. H., Welkhoff, P. A., Emerson, C., Godfray, H. C. J., Gottlieb, M., ... Toure, Y. T. (2018). Pathway to deployment of gene drive mosquitoes as a potential biocontrol tool for elimination of malaria in sub-Saharan Africa: Recommendations of a scientific working group. *American Journal of Tropical Medicine and Hygiene*, 98(Suppl 6), 1–49. <https://doi.org/10.4269/ajtmh.18-0083>
- Jones, C. M., Toé, H. K., Sanou, A., Namountougou, M., Hughes, A., Diabaté, A., ... Ranson, H. (2012). Additional selection for insecticide resistance in urban malaria vectors: DDT resistance in *Anopheles arabiensis* from Bobo-Dioulasso, Burkina Faso. *PLoS ONE*, 7(9), e45995. <https://doi.org/10.1371/journal.pone.0045995>
- Kain, P., Chakraborty, T. S., Sundaram, S., Siddiqi, O., Rodrigues, V., & Hasan, G. (2008). Reduced odor responses from antennal neurons of G(q)alpha, phospholipase Cbeta, and rdgA mutants in *Drosophila* support a role for a phospholipid intermediate in insect olfactory transduction. *The Journal of Neuroscience*, 28(18), 4745–4755. <https://doi.org/10.1523/JNEUROSCI.5306-07.2008>
- Kalangala District Local Government District Management Improvement Plan 2012–2015. (2012). *Technical Report*.
- Kayondo, J. K., Mukwaya, L. G., Stump, A., Michel, A. P., Coulibaly, M. B., Besansky, N. J., & Collins, F. H. (2005). Genetic structure of *Anopheles gambiae* populations on islands in northwestern Lake Victoria Uganda. *Malaria Journal*, 4, 59. <https://doi.org/10.1186/1475-2875-4-59>
- Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A., & Mayrose, I. (2015). CLUMPAK: A program for identifying clustering modes and packaging population structure inferences across K. *Molecular Ecology Resources*, 15(5), 1179–1191. <https://doi.org/10.1016/j.coviro.2015.09.001>
- Lawson, D. J., van Dorp, L., & Falush, D. (2018). A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nature Communications*, 9(3258), 1–11. <https://doi.org/10.1038/s41467-018-05257-7>
- Lehmann, T., Licht, M., Elissa, N., Maega, B. T. A., Chimumbwa, J. M., Watsenga, F. T., ... Hawley, W. A. (2003). Population structure of *Anopheles gambiae* in Africa. *Journal of Heredity*, 94(2), 133–147. <https://doi.org/10.1093/jhered/esg024>
- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14), 1754–1760. <https://doi.org/10.1093/bioinformatics/btp324>
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>
- Lin, Y. H., Chen, Y. C., Kao, T. Y., Lin, Y. C., Hsu, T. E., Wu, Y. C., ... Wang, H. D. (2014). Diacylglycerol lipase regulates lifespan and oxidative stress response by inversely modulating TOR signaling in *Drosophila* and *C. elegans*. *Aging Cell*, 13(4), 755–764. <https://doi.org/10.1111/ace1.12232>
- Liu, X., & Fu, Y.-X. (2015). Exploring population size changes using SNP frequency spectra. *Nature Genetics*, 47(5), 555–559. <https://doi.org/10.1038/ng.3254>

- Lukindu, M., Bergey, C. M., Wiltshire, R. M., Small, S. T., Bourke, B. P., Kayondo, J. K., & Besansky, N. J. (2018). Spatio-temporal genetic structure of *Anopheles gambiae* in the Northwestern Lake Victoria Basin, Uganda: Implications for genetic control trials in malaria endemic regions. *Parasites & Vectors*, *11*(1), 246. <https://doi.org/10.1186/s13071-018-2826-4>
- Maliti, D., Ranson, H., Magesa, S., Kisinza, W., Mcha, J., Haji, K., ... Weetman, D. (2014). Islands and stepping-stones: Comparative population structure of *Anopheles gambiae sensu stricto* and *Anopheles arabiensis* in Tanzania and implications for the spread of insecticide resistance. *PLoS ONE*, *9*(10), e110910. <https://doi.org/10.1371/journal.pone.0110910>
- Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, *27*(2), 209–220. <https://doi.org/10.1136/bmj.2.5524.1260-d>
- Marsden, C. D., Cornel, A., Lee, Y., Sanford, M. R., Norris, L. C., Goodell, P. B., ... Lanzaro, G. C. (2013). An analysis of two island groups as potential sites for trials of transgenic mosquitoes for malaria control. *Evolutionary Applications*, *6*, 706–720. <https://doi.org/10.1111/eva.12056>
- Marshall, J. C., Pinto, J., Charlwood, J. D., Gentile, G., Santolamazza, F., Simard, F., ... Caccone, A. (2008). 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. *Evolutionary Applications*, *1*, 631–644. <https://doi.org/10.1111/j.1752-4571.2008.00048.x>
- Miles, A., Harding, N. J., Bottà, G., Clarkson, C. S., Antão, T., Kozak, K., ... Kwiatkowski, D. P. (2017). Genetic diversity of the African malaria vector *Anopheles gambiae*. *Nature*, *552*, 96–100. <https://doi.org/10.1038/nature24995>
- Mitchell, S. N., Rigden, D. J., Dowd, A. J., Lu, F., Wilding, C. S., Weetman, D., ... Donnelly, M. J. (2014). Metabolic and target-site mechanisms combine to confer strong DDT resistance in *Anopheles gambiae*. *PLoS ONE*, *9*(3), e92662. <https://doi.org/10.1371/journal.pone.0092662>
- Moreno, M., Salgueiro, P., Vicente, J. L., Cano, J., Berzosa, P. J., de Lucio, A., ... Benito, A. (2007). Genetic population structure of *Anopheles gambiae* in Equatorial Guinea. *Malaria Journal*, *6*, 137. <https://doi.org/10.1186/1475-2875-6-137>
- National Malaria Control Programme, Abt Associates, & the INFORM Project. (2013). An epidemiological profile of malaria and its control in Uganda. <http://www.inform-malaria.org/wp-content/uploads/2014/05/Uganda-Epi-Report-060214.pdf>
- Nelson, C. S., Beck, J. N., Wilson, K. A., Pilcher, E. R., Kapahi, P., & Brem, R. B. (2016). Cross-phenotype association tests uncover genes mediating nutrient response in *Drosophila*. *BMC Genomics*, *17*(1), 1–14. <https://doi.org/10.1186/s12864-016-3137-9>
- Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, *2*(12), 2074–2093. <https://doi.org/10.1371/journal.pgen.0020190>
- Price, A., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904–909. <https://doi.org/10.1038/ng1847>
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., ... Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575. <https://doi.org/10.1086/519795>
- Randhawa, I. A., Khatkar, M., Thomson, P., & Raadsma, H. (2014). Composite selection signals can localize the trait specific genomic regions in multi-breed populations of cattle and sheep. *BMC Genetics*, *15*(1), 34. <https://doi.org/10.1186/1471-2156-15-34>
- Rosen, Z., Bhaskar, A., & Song, Y. S. (2018). Geometry of the sample frequency spectrum and the perils of demographic inference. *Genetics*, *210*(2), 665–682. <https://doi.org/10.1534/genetics.118.300733>
- Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., ... Stewart, J. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature*, *449*(7164), 913–918. <https://doi.org/10.1038/nature06250>
- Salgueiro, P., Moreno, M., Simard, F., O'Brochta, D., & Pinto, J. (2013). New Insights into the population structure of *Anopheles gambiae* s.s. in the Gulf of Guinea Islands revealed by Herves transposable elements. *PLoS ONE*, *8*(4), e62964. <https://doi.org/10.1371/journal.pone.0062964>
- Scott, J. A., Brogdon, W. G., & Collins, F. H. (1993). Identification of single specimens of the *Anopheles gambiae* complex by the polymerase chain reaction. *The American Journal of Tropical Medicine and Hygiene*, *49*(4), 520–529. <https://doi.org/10.4269/ajtmh.1993.49.520>
- Senthilan, P. R., Piepenbrock, D., Ovezmyradov, G., Nadrowski, B., Bechstedt, S., Pauls, S., ... Göpfert, M. C. (2012). *Drosophila* auditory organ genes and genetic hearing defects. *Cell*, *150*(5), 1042–1054. <https://doi.org/10.1016/j.cell.2012.06.043>
- Sharakhova, M. V., Hammond, M. P., Lobo, N. F., Krzywinski, J., Unger, M. F., Hillenmeyer, M. E., ... Collins, F. H. (2007). Update of the *Anopheles gambiae* PEST genome assembly. *Genome Biology*, *8*(1), R5. <https://doi.org/10.1186/gb-2007-8-1-r5>
- Stern, D. L. (2013). The genetic causes of convergent evolution. *Nature Reviews Genetics*, *14*(11), 751–764. <https://doi.org/10.1038/nrg3483>
- Storfer, A., Murphy, M. A., Spear, S. F., Holderegger, R., & Waits, L. P. (2010). Landscape genetics: Where are we now? *Molecular Ecology*, *19*(17), 3496–3514. <https://doi.org/10.1111/j.1365-294X.2010.04691.x>
- Stow, A. J., & Magnusson, W. E. (2012). Genetically defining populations is of limited use for evaluating and managing human impacts on gene flow. *Wildlife Research*, *39*, 290–294. <https://doi.org/10.1071/WR11150>
- Szpiech, Z. A., & Hernandez, R. D. (2014). selscan: An efficient multi-threaded program to perform EHH-based scans for positive selection. *Molecular Biology and Evolution*, *31*(10), 2824–2827. <https://doi.org/10.1093/molbev/msu211>
- Thomas, A. (1941). The vegetation of the Sese Islands, Uganda: An illustration of edaphic factors in tropical ecology. *Journal of Ecology*, *29*(2), 330–353. <https://doi.org/10.2307/2256396>
- Uganda Bureau of Statistics (UBOS), & ICF (2017). *Uganda demographic and health survey 2016: Key indicators report*. Kampala, Uganda: UBOS, and Rockville, MD: UBOS and ICF. <https://doi.org/10.2307/41329750>
- Uganda Bureau of Statistics (2002). *2002 Uganda population and housing census analytical report*. Kampala, Uganda.
- Uganda Bureau of Statistics. (2016). *The national population and housing census 2014 – Main report*. Kampala, Uganda.
- Verdonschot, P. F., & Besse-Lototskaya, A. A. (2014). Flight distance of mosquitoes (Culicidae): A metadata analysis to support the management of barrier zones around rewetted and newly constructed wetlands. *Limnologia*, *45*, 69–79. <https://doi.org/10.1016/j.limno.2013.11.002>
- Vontas, J., Grigoraki, L., Morgan, J., Tsakireli, D., Fouseini, G., Segura, L., ... Hemingway, J. (2018). Rapid selection of a pyrethroid metabolic enzyme CYP9K1 by operational malaria control activities. *Proceedings of the National Academy of Sciences*, *115*(18), 4619–4624. <https://doi.org/10.1073/pnas.1719663115>
- Wallberg, A., Pirk, C. W., Allsopp, M. H., & Webster, M. T. (2016). Identification of multiple loci associated with social parasitism in honeybees. *PLoS Genetics*, *12*(6), 1–30. <https://doi.org/10.1371/journal.pgen.1006097>
- Waples, R. S. (1998). Separating the wheat from the chaff: Patterns of genetic differentiation in high gene flow species. *Journal of Heredity*, *89*(5), 438–450. <https://doi.org/10.1093/jhered/89.5.438>
- Waples, R. S. (2006). A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene

- loci. *Conservation Genetics*, 7(2), 167. <https://doi.org/10.1007/s10592-005-9100-y>
- Weir, B., & Cockerham, C. (1984). Estimating  $F$ -statistics for the analysis of population structure. *Evolution*, 38(6), 1358–1370.
- Whitlock, M. C., & McCauley, D. E. (1999). Indirect measures of gene flow and migration:  $F_{ST}$  not equal to  $1/(4Nm + 1)$ . *Heredity*, 82 (Pt. 2) (November 1998), 117–125. <https://doi.org/10.1038/sj.hdy.6884960>
- Wiltshire, R. M., Bergey, C. M., Kayondo, J. K., Birungi, J., Mukwaya, L. G., Emrich, S. J., ... Collins, F. H. (2018). Reduced-representation sequencing identifies small effective population sizes of *Anopheles gambiae* in the north-western Lake Victoria basin, Uganda. *Malaria Journal*, 17(1), 285. <https://doi.org/10.1186/s12936-018-2432-0>
- World Health Organization-TDR and the Foundation for the National Institutes of Health. (2014). *Guidance framework for testing of genetically modified mosquitoes*. Geneva, Switzerland: WHO-TDR and FNHI. <http://www.who.int/tdr/publications/year/2014/guide-fmrk-gm-mosquit/en/>
- World Health Organization (2017). *World Malaria report 2017*. Geneva, Switzerland.

- Zeemeijer, I. (2012). Who gets what, when and how?: New corporate land acquisitions and the impact on local livelihoods in Uganda (Master's Thesis). Utrecht University.

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

**How to cite this article:** Bergey CM, Lukindu M, Wiltshire RM, Fontaine MC, Kayondo JK, Besansky NJ. Assessing connectivity despite high diversity in island populations of a malaria mosquito. *Evol Appl*. 2020;13:417–431. <https://doi.org/10.1111/eva.12878>