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

Christina M. Bergey[1,2,3,*], Martin Lukindu[1,2], Rachel M. Wiltshire[1,2], Michael C. Fontaine[4,5], Jonathan K. Kayondo[6], and Nora J. Besansky[1,2,*]

[1]Department of Biological Sciences, University of Notre Dame, Notre Dame, IN 46556, USA

[2]Eck Institute for Global Health, University of Notre Dame, Notre Dame, IN 46556, USA

[3]Departments of Anthropology and Biology, Pennsylvania State University, University Park, PA 16802, USA.

[4]Groningen Institute for Evolutionary Life Sciences (GELIFES), University of Groningen, PO Box 11103 CC, Groningen, The Netherlands.

[5]MIVEGEC, IRD, CNRS, University of Montpellier, Montpellier, France.

[6]Department of Entomology, Uganda Virus Research Institute (UVRI), Entebbe, Uganda.

[*] *To whom correspondence should be addressed.*

February 28, 2019

Short title: Shared adaptation as connectivity proxy

Corresponding authors: C.M.B. (cxb585@psu.edu) and N.J.B. (nbesansk@nd.edu)

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 kilometers 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.

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 [1] and discernment of demographic independence from panmixia [2]. Population genetic methods for estimating migration using neutral markers may thus have limited utility when such a high proportion of diversity is shared 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 [3].

The major African malaria vector *Anopheles gambiae sensu stricto* (henceforth *An. gambiae*) is among the most genetically diverse eukaryotic species [4], with shallow population structure [4, 5] 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 [6]. 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 [7–9]. Such systems have recently been successfully engineered in the laboratory [10]. 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 genome-wide variation in *An. gambiae* mosquitoes living near and on the Ssese archipelago of Lake Victoria in Uganda (Fig. 1). We augment these analyses with a demonstration of our framework using selective sweep sharing as a proxy for connectivity when 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 [11]—is reduced. In addition to the high malaria prevalence of the islands (44% in children; 30% in children country-wide; [12]), 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 [13–16]. 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.

Results

The Ssese Islands are approximately 4-50 km from the mainland, and vary in size, infrastructure, and accessibility. Sampled islands range from Banda—a small, largely forested island of approximately 1 square kilometer with a single settlement—to Bugala—296 square kilometers, site of a 10,000 ha oil palm plantation [17], and linked to the mainland via ferry service [18]. To explore the partitioning of *An. gambiae* genetic variation in the Lake Victoria Basin (LVB), we sequenced the genomes of 116 mosquitoes from 5 island and 4 mainland localities (Fig. 1, Supplementary Table S1). We sequenced 10-23 individuals per site to an average depth of 17.6 $\pm$ 4.6 (Supplementary Table S2). After filtering (detailed in Methods), we identified 28.6 million high quality Single Nucleotide Polymorphisms (SNPs). We merged our dataset with that of the *An. gambiae* 1000 Genomes project (Ag1000G; [4]) for a combined dataset of 12.54 million SNPs (9.86 million after linkage disequilibrium pruning) in 881 individuals.

Figure 1: Map of Lake Victoria Basin study area.
Map of study area showing sampling localities on Ssese Islands (blue) and mainland localities (red) in Lake Victoria Basin. 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.

Genetic structure

We analyzed LVB population structure with context from continent-wide populations [4] of *An. gambiae* and sister species *Anopheles coluzzii* mosquitoes (formerly known as *An. gambiae* M molecular form [19]). Both Bayesian clustering ([20]; Fig. 2a) and principal component analysis (PCA; Supplementary Fig. 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; Fig. 1). With ≥ 6 clusters (which optimized predictive accuracy in the clustering analysis; Supplementary Fig. S2), island samples had distinct ancestry proportions (Fig. 2a), and with $k = 9$ clusters, we observed additional subdivision in LVB samples and the assignment of the majority of Ssese individuals' ancestry to a largely island-specific component (Figs. 2a, 2b, and Supplementary Fig. S3).

PCA of only LVB individuals 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 (Fig. 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 (Supplementary Fig. 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 and the rest to the island-specific component were present on all islands except Banda.

Differentiation concurred with observed population structure. Mean F_{ST} between sampling localities (range: 0.001-0.034) was approximately 0 (≤ 0.003) for mainland-mainland comparisons and was highest in comparisons involving the small island Banda (Fig. 2d). Geographic distances and F_{ST} were uncorrelated (Mantel $p = 0.88$; Supplementary Fig. S5). Island samples showed greater within- and between-locality sharing of genomic regions identi-

cal by descent (IBD), with sharing between nearby islands Sserinya, Banda and Bugala (Fig. 2e). Importantly, Banda Island shared no IBD regions with mainland sites, underscoring its contemporary isolation from the mainland.

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 (π; Wilcoxon rank sum test $p < 0.001$; Fig. 3a), a higher proportion of shared to rare variants (Tajima's D; $p < 0.001$; Fig. 3b), and more linkage among SNPs (LD; r^2; $p < 0.001$; Fig. 3f). They were however, similar in inbreeding coefficient (F; $p = 0.0719$; Fig. 3c), number of long runs of homozygosity (F_{ROH}; $p = 0.182$; Fig. 3d), and proportions of low frequency SNPs (Fig. 3e). The small island Banda was the most extreme in these measures.

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 [21] based on the site frequency spectrum (SFS; Fig. 4a) and patterns of IBD sharing ([22]; Fig. 4b), respectively. Short-term final mainland sizes were unrealistically high, likely due to low samples sizes for each locality, but island-mainland differences were nonetheless informative. In both, islands had consistently lower N_e compared to mainland populations extending back 500 generations ($\sim$ 50 years) and often severely fluctuated, particularly in the last 250 generations ($\sim$ 22 years). Mainland sites Wamala and Kaazi had island-like recent histories, with Wamala abruptly switching to an island-like pattern around 2005.

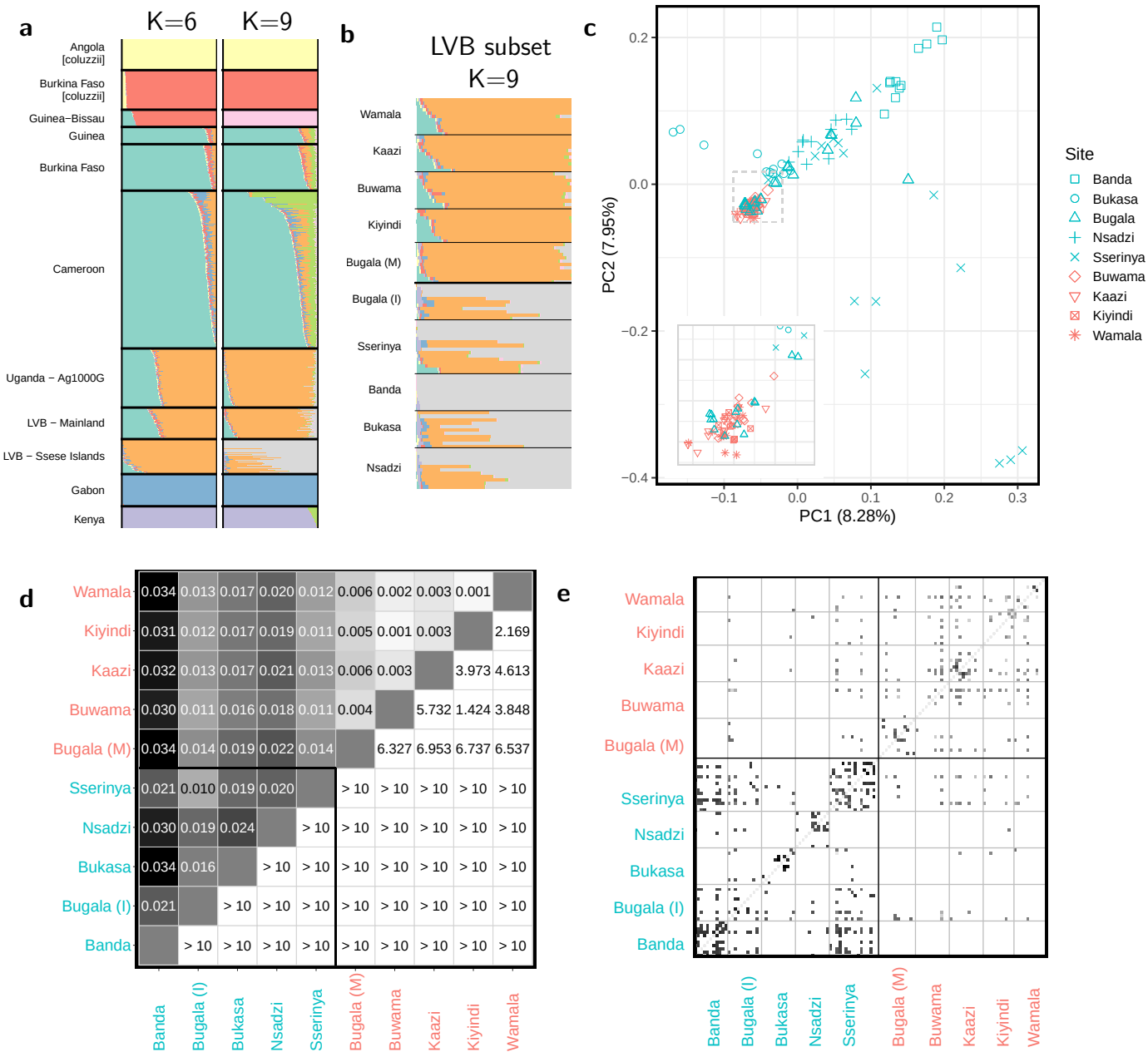

Figure 2: (Caption on next page.)

Figure 2: Population structure in the Lake Victoria Basin.
Analyses are based on chromosome 3 to avoid segregating inversions on other chromosome, unless otherwise noted. (A) ADMIXTURE-inferred ancestry of individuals in Lake Victoria Basin. Results based on analysis of LVB and Ag1000G merged dataset. Analysis is restricted to *A. 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 (Fig. 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 genome-wide 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.

To all pairs of LVB localities we fit an isolation-with-migration (IM) demographic model using δaδi, in which an ancestral population splits into two populations, allowing exponential growth and continuous asymmetrical migration between the daughter populations (Supplementary Figs. S6, S7). In all comparisons involving islands and some between mainland sites, the best fitting model as chosen via AIC had zero migration (Supplementary Tables S3, S4, and 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.

Selection

As beneficial 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

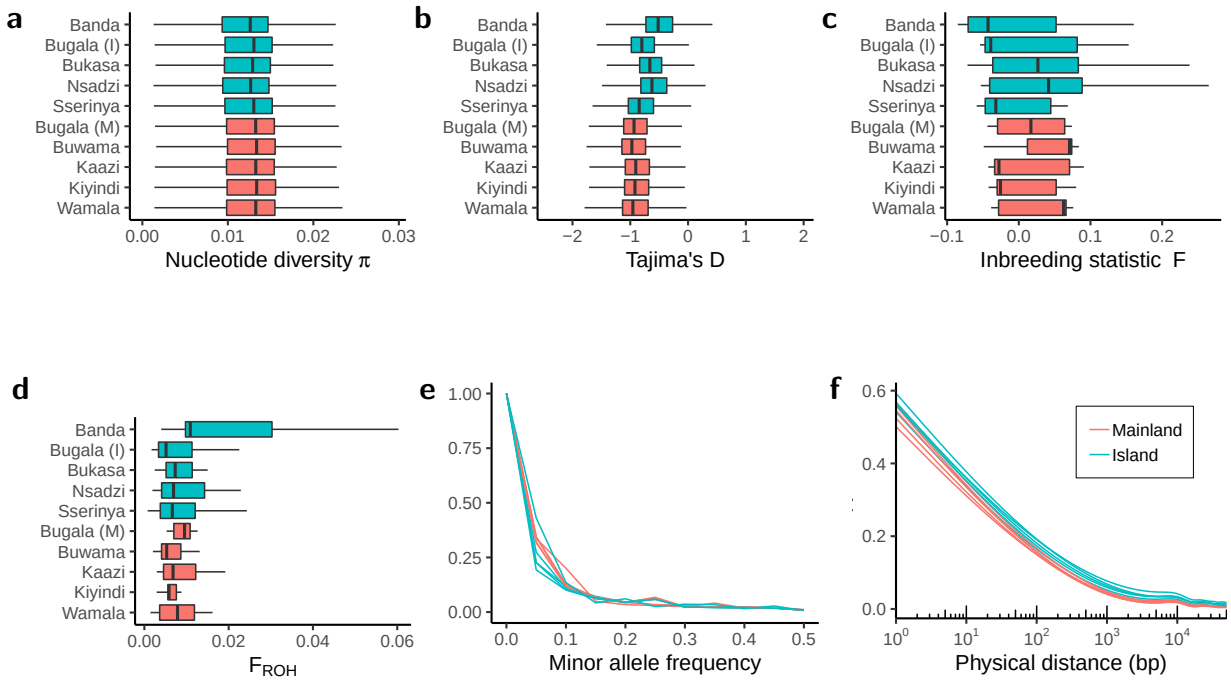

Figure 3: Diversity metrics in the Lake Victoria Basin samples.
Shown are a (A) boxplot of nucleotide diversity (π; in 10 kilobase windows), (B) boxplot of Tajima's D (in 10 kilobase windows), (C) boxplot of inbreeding statistic (F), (D) boxplot of length of runs of homozygosity (F_{ROH}), (E) histogram 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.

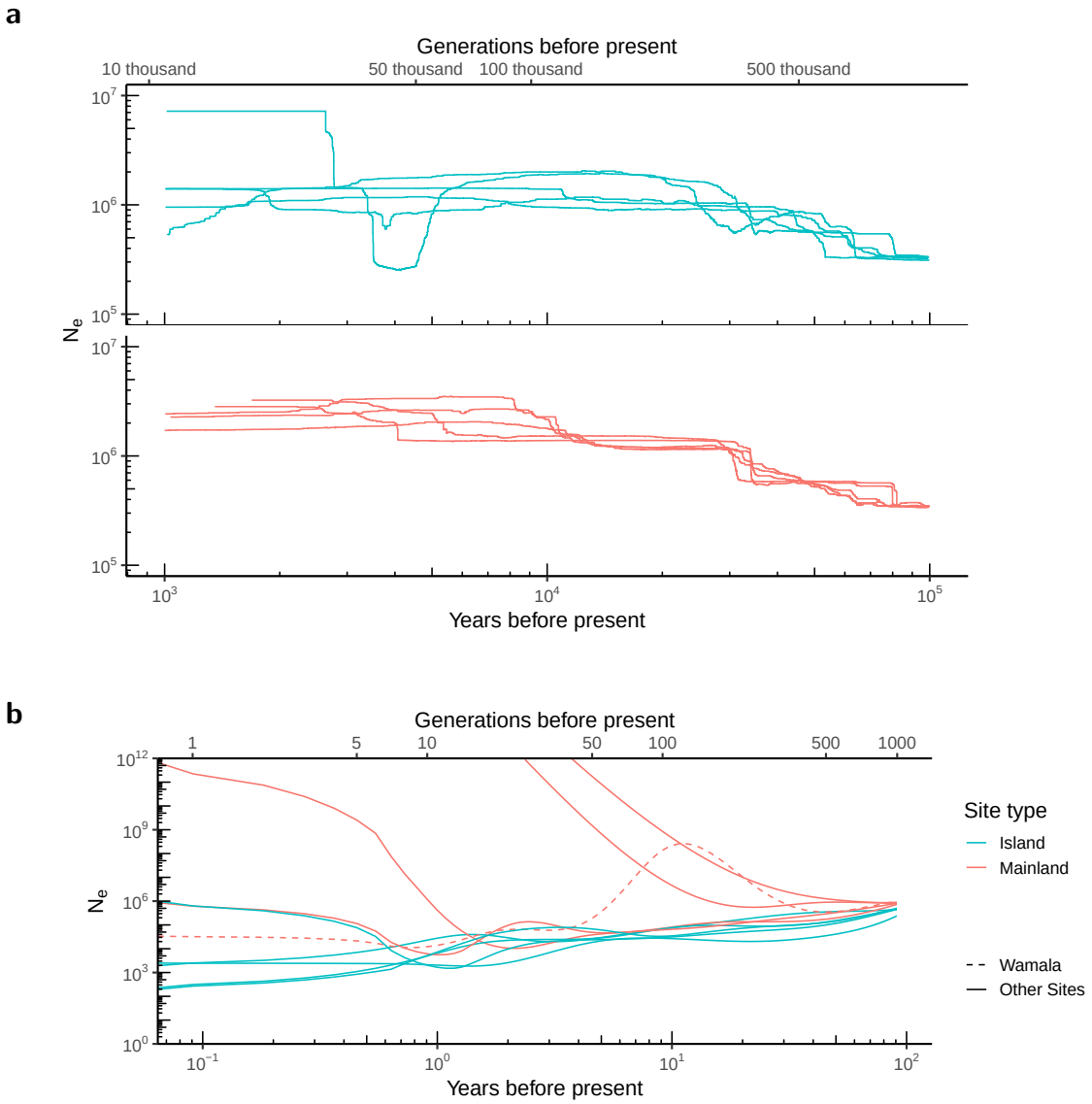

Figure 4: Population history of the Lake Victoria Basin samples.
(A) Long-term evolutionary population histories inferred via stairway plots for island and mainland samples. (B) Contemporary or short-term effective population size (N_e) history inferred using sharing of regions that are identical by 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.

selection in the same genomic region in populations with independent lineages would be consistent with several scenarios [23]: (*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*), but 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 Ssese Islands, reasoning that shared signatures of sweeps at a genomic location may indicate migration is occurring with the islands, while the absence was suggestive of isolation. We identified sweeps in the LVB using genome scans of between- and within-locality statistics, including F_{ST} ([24], Supplementary Fig. S8), Extended Haplotype Homozygosity (XP-EHH, [25], Supplementary Fig. S8), and haplotype

homozygosity (H12, [26], Supplementary Fig. 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 > 99th percentile). To add additional country-level context, we then intersected these regions with those under putative selection in a mainland Ugandan reference population (H12 > 95th percentile; [4]).

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 (Supplementary 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%; Supplementary 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, 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 (H12 > 95th percentile in Ag1000G; [4]).

As expected, outlier regions included known selective sweep targets from elsewhere in Africa ([4], Supplementary Table S9). All sweeps found in the reference Uganda population [4] were detected in at least some sampling localities in our LVB dataset, except the

sweep targeting *Vgsc*, which was likely excluded during filtration of the region adjacent to the centromere. 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 [27], exhibited low diversity, an excess of low frequency polymorphisms (Tajima's D), and elevated haplotype homozygosity (H12) within the LVB populations (Supplementary Figs. S9, S10). Pairwise statistics (F_{ST} and XP-EHH) indicated low differentiation between LVB localities, as expected for a continent-wide sweep (Supplementary Figs. S8). The signal was found in every LVB site, including all islands. Hierarchical clustering of LVB and Ag1000G haplotypes revealed clades with low inter-haplotype diversity, expected after selection rapidly increases the frequency of a haplotype containing adaptive variation (Supplementary Fig. S11). Consistent with previous results [4], 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 [4] 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 (Supplementary Table S8). As in previous studies [4], 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, ei-

ther 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 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 (Supplementary Figs. S8, S9, S12, and S13). The first, at 2L:34.1 Mb, contains many genes, including a cluster involved in chorion formation [28] near the signal peak. Haplotype clustering revealed a group of closely-related Ugandan individuals, consistent with a geographically bounded selective sweep (Supplementary Fig. 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, *e.g.*, 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, chrX:9,215,505-9,266,532). Low diversity haplotypes formed a single cluster including LVB haplotypes overwhelmingly from the islands (Supplementary Fig. 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 con-

sistent 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.

Discussion

Understanding the population genetics of island *Anopheles gambiae* has both evolutionary and practical importance. A limited number of genetic investigations have been conducted on oceanic [29–32] and lacustrine islands [33–36], though the latter have been limited in the type or count of molecular markers used. In contrast to shallow population structure across Africa [4, 5], partitioning of genetic variation on islands suggests varying isolation. Using a genome-wide dataset, we found differentiation between the Ssese 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, $F_{ST} = 0.007$ [4]). The Ssese 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 [31]). 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 "Ssese" derives from another arthropod vector, the tsetse fly (*Glossina* spp.) The tsetse-mediated arrival of sleeping sickness in 1902 brought "enormous mortality" [37,

pp. 332] to the 20 thousand residents, who were evacuated in 1909 [37, 38]. Though encouraged to return by 1920, the human population numbered only 4 thousand in 1941 [37] and took until 1980 to double [39], but has since rapidly risen to over 62 thousand (2015, projected; [18, 40]). 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 [34, 36]. 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 2000-2010 in the *Plasmodium falciparum* parasite rate ($Pf\mathrm{PR}_{2-10}$; a measure of malaria transmission intensity) in Mityana, the district containing Wamala [41].

Though previous *Anopheles* population genetic studies have inferred gene flow even among species [4, 42], we inferred 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; [4]). 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 polymor-

297 phism or the existence of inadequately sampled ancestral variation [43], rather than recent
298 admixture. This interpretation is consistent with the affinity we observed between the Ssese
299 Islands and West Africa in the structure of adaptive variation.

300 Discerning whether the absence of observed gene flow is due to lack of connectivity, the
301 opposition of selection, or the stochasticity of genetic drift is difficult. Instead we must rely
302 on estimates of the strength of selection in the two locales to inform our conclusions. For
303 example, we would expect that an insecticide sweep found all over Africa would spread in
304 island mosquito populations with insecticide treated bed nets, despite the considerable effect
305 of genetic drift in small populations. As insecticide treated bed net usage is present on the
306 islands [18], variation conferring a major selective advantage related to insecticides would
307 be expected to spread to and persist on the islands if migration allows the transfer, and the
308 strongest evidence of a lack of contemporary connectivity is therefore the absence of a sweep
309 on the islands that is widespread on the continent.

310 We found two sweeps on insecticide-related genes that are common targets of selection
311 elsewhere but which have incompletely colonized the Ssese Islands: one on cytochrome P450
312 monooxygenase *Cyp9K1* [44, 45] present on some islands, and another on glutathione S-
313 transferase genes (*Gste1-Gste7*; [46–49]) at extremely low frequency on the islands. That
314 the selective sweeps targeting these loci [4] have not fully colonized the islands despite the
315 advantage in detoxifying pyrethroids and DDT suggests a lack of contemporary exchange.
316 However, the sweep targeting the *Cyp6p* cluster was found on all islands, confirming past
317 gene flow had occurred at some point. Although these distributions confirm that past mi-
318 gration from the mainland to islands has occurred and we are unable to exclude low levels of
319 contemporary gene flow, taken together our data are consistent with potentially high degrees
320 of isolation on contemporary timescales for some islands of the Ssese archipelago.

321 Our investigation also identified two previously unknown signatures of selection. For
322 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 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 [50, 51], but also olfactory [52] and auditory [53] sensory processing. It has been recently implicated in nutritional homeostasis in *Drosophila* [54] and is known to interact with the TOR pathway [55], which has been identified as a target of ecological adaptation in *Drosophila* [56, 57] and *An. gambiae* [58]. 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, *e.g.* 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. Interestingly, the putative sweep coincides with a similar region of low diversity in a cryptic subgroup of *Anopheles gambiae sensu lato* (GOUNDRY; [42]), suggesting possible parallel selective events on independent mutations or adaptive introgression.

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 [59] particularly given the assumption of equilibrium between the evolutionary forces of

migration and drift [59–61], an unlikely state for huge *An. gambiae* populations [3]. 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 Ssese 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 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 [16], potentially up to hundreds of kilometers [11]. Although no island, lacustrine or oceanic, is completely isolated, such localities may still be ideal for initial gene drive field testing, as the geographical barriers maximize isolation to the extent possible [16], and absolute isolation on evolutionary timescales is unnecessary given the relatively short timeframe of small-scale field tests. Thus, the probability of contemporary migration may be sufficiently low to qualify some Ssese 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 Ssese Islands may be logical and tractable candidates for initial field tests of genetically modified *An. gambiae* mosquitoes, warranting further entomological study.

Materials and Methods

Experimental design Mosquitoes were sampled from 5 of the Ssese Islands in Lake Victoria, Uganda (Banda, Bukasa, Bugala, Nsadzi, and Sserinya) and 4 mainland sampling localities (Buwama, Kaazi, Kiyindi, and 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, U.S.A. for analysis.

DNA extraction, Library preparation, and Whole Genome Sequencing Animals were assigned to species level via a PCR-based assay [62] 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 [63] and then quantified via fluorometry (PicoGreen). Automated library preparation took place at the NYU 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.

Mapping and SNP calling, filtering Software version information is provided in Supplementary Table S10. After quality filtering and trimming using ea-utils' fastq-mcf (-l 15 -q 15 -w 4; [64]), reads were mapped to the *An. gambiae* reference genome (AgamP4 PEST; [65, 66]) using BWA aln and sampe with default parameters [67].

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 [4]) and filtered for quality [68],

excluding SNPs with QualByDepth (QD) < 2.0, RMSMappingQuality (MQ) < 40.0, FisherStrand (FS) > 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. Individuals and variants with high levels of missingness (> 10%) and variants that were not biallelic or exhibited values of HWE that were likely due to sequencing error ($p < 0.00001$) were excluded from further analysis. For use in population structure inference, the SNP dataset was further pruned for linkage disequilibrium by sliding a window of 50 SNPs across the genome in 5 SNP increments and recursively removing random SNPs in pairs with $r^2 > 0.5$ using PLINK [69, 70]. After filtration, the dataset contained 28,569,621 SNPs before LD pruning and 115 individuals. SNPs unpruned for linkage disequilibrium were phased with SHAPEIT2 [71] using an effective population size (N_e) of 1,000,000 (consistent with previous demographic modeling [4]), 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.

Population structure inference To explore population structure in a larger, continent-wide context, we merged our LVB SNP set with a recently published dataset of *Anopheles gambiae* individuals (from the Ag1000G project) [4]. Prior to filtering, biallelic SNPs from the LVB and Ag1000G datasets were merged using bcftools [72]. We excluded any SNP with greater than 10% missingness in either dataset, any SNPs that did not pass the accessibility filter of the Ag1000G dataset, and SNPs with 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 (leaving 881 individuals), we assigned individuals' genomes to ancestry components using ADMIXTURE [20]. We created 10 replicate samples of 100,000 SNPs from

chromosome 3 (prior to LD-pruning), 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 $r^2 > 0.01$ in sliding windows of size 500 SNPs and with a stepsize of 250 SNPs. For each replicate, we ran ADMIXTURE for 5 iterations in five-fold cross validation mode for values of k 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 [73]. 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 [69, 70], using the LVB-Ag1000G merged dataset (excluding the outlier Kenyan population; [4]) and 3,212,485 chromosome 3 SNPs (to avoid the well-known inversions on chromosome 2 and the X-chromosome) outside of heterochromatic regions (such as centromeric regions; [66]; Supplementary 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 this 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.

We computed the pairwise fixation index (F_{ST}) between locality samples for *An. gambiae* using the unbiased estimator of Hudson [74] as implemented in smartpca [75, 76]. To obtain overall values between sampling sites, per-SNP values were averaged across the genome excluding known 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

[77] as implemented in the R package ade4 [78], using these F_{ST} estimates and Euclidean geographic distances between localities.

To estimate fine-scale structure and relatedness between individuals, we estimated the proportion of pairs of individuals genomes that are identical by descent (IBD) using PLINK [69, 70]. We excluded heterochromatic and inversion regions, and retained informative pairs of SNPs within 500 kb in the pairwise population concordance test.

Diversity estimation Grouping individuals by site (except for Bugala, which was split based on the results of the PCA), we calculated nucleotide diversity (π) 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 (Danecek2011). 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. To visualize decay in LD, we plotted r^2 between SNPs against their physical distance in base pairs, first smoothing the data by fitting a generalized additive model (GAM) to them. We also inferred runs of homozygosity using PLINK [69, 70] to compare their length (F_{ROH}), requiring 10 homozygous SNPs spanning a distance of 100 kb and allowing for 3 heterozygous and 5 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.

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 [79] and analyzed them with IBDNe [22]. We restricted our analysis to SNPs from chromosome 3 to avoid 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 [4] and assumed a constant recombination rate of 2.0 cM/Mb (after [80]).

To estimate the long-term evolutionary demographic history of mosquitoes on and near the Ssese Islands, including a long-term estimate of N_e [81], 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 [21].

We also inferred a "two-population" isolation-with-migration (IM) demographic model with δaδi [82, 83] in which the ancestral population splits to form two daughter populations that are allowed to grow exponentially and exchange migrants asymmetrically, as described in the main text. For δaδi-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 *An. merus* [84]. 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 χ^2 distribution and rejected the null model if the p-value for the test statistic was < 0.05. 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 = 0.0.70%). We assessed the goodness-of-fit visually using the residuals of the comparison between model and data frequency spectra (Supplementary Fig. 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 $T \in (1e\text{-}8, 0.1)$, for migration were $m \in (1e\text{-}8, 10)$, and for genotyping uncertainty were $p_{misid} \in (1e\text{-}8, 1)$. Parameters were perturbed

before allowing up to 1000 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, ν, 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 [85], we excluded the population size change from our interpretation.

To translate δaδi- 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 [4].

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 [24] (as implemented in VCFtools [86]), and H12 (as implemented in SelectionHapStats [26]) and XP-EHH on a per-site basis (as implemented in selscan [87]) to detect long stretches of homozygosity in a given population considered alone or relative to another population [25]. 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 [80]). 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; [88, 89]].

We plotted these statistics across the genome to identify candidate regions with signa-

tures 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, we identified putative locality-specific sweeps (H12 over 99$^{\text{th}}$ percentile in one population), island-specific sweeps (H12 over 99$^{\text{th}}$ percentile in 4 or more of the 5 island localities but 0 or 1 mainland localities), or LVB mainland-specific sweeps (H12 over 99$^{\text{th}}$ percentile in 3 or more of the 4 island localities but 0 or 1 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 if the H12 values of each of the Ag1000G populations (excluding Kenya due to its signatures of admixture and recent population decline; [4]) 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; [4]). 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.

Acknowledgments

The authors would like to thank the UVRI field entomology team: Christine Babirye, Ronald Mayanja, Paul Mawejje, 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 subcontracts to J.K.K. and N.J.B. N.J.B. also received support from NIH R01 AI125360 and R21 AI123491. The NYU 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.

References

R. S. Waples, Separating the wheat from the chaff: Patterns of genetic differentiation in high gene flow species. *Journal of Heredity* **89**, 438–450 (1998).

R. S. Waples, A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci. *Conservation Genetics* **7**, 167 (2006).

P. A. Gagnaire, T. Broquet, D. Aurelle, F. Viard, A. Souissi, F. Bonhomme, S. Arnaud-Haond, N. Bierne, Using neutral, selected, and hitchhiker loci to assess connectivity of marine populations in the genomic era. *Evolutionary Applications* **8**, 769–786 (2015).

A. Miles, *et al.*, Genetic diversity of the African malaria vector *Anopheles gambiae*. *Nature* **552**, 96–100 (2017).

T. Lehmann, M. Licht, N. Elissa, B. T. a. Maega, J. M. Chimumbwa, F. T. Watsenga, C. S. Wondji, F. Simard, W. a. Hawley, Population structure of *Anopheles gambiae* in Africa. *Journal of Heredity* **94**, 133–147 (2003).

World Health Organization, *World Malaria Report 2017* (Geneva, 2017).

A. Burt, Heritable strategies for controlling insect vectors of disease. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **369**, 20130432 (2014).

577 J. Champer, A. Buchman, O. S. Akbari, Cheating evolution: engineering gene drives to
578 manipulate the fate of wild populations. *Nature Reviews Genetics* **17**, 146–159 (2016).

579 L. Alphey, Genetic control of mosquitoes. *Annual Review of Entomology* **59**, 205–224
580 (2014).

581 A. Hammond, R. Galizi, K. Kyrou, A. Simoni, C. Siniscalchi, D. Katsanos, M. Grib-
582 ble, D. Baker, E. Marois, S. Russell, A. Burt, N. Windbichler, A. Crisanti, T. Nolan, A
583 CRISPR-Cas9 gene drive system targeting female reproduction in the malaria mosquito
584 vector *Anopheles gambiae*. *Nature Biotechnology* **34**, 78–83 (2016).

585 A. Dao, A. S. Yaro, M. Diallo, S. Timbiné, D. L. Huestis, Y. Kassogué, A. I. Traoré,
586 Z. L. Sanogo, D. Samaké, T. Lehmann, Signatures of aestivation and migration in Sahelian
587 malaria mosquito populations. *Nature* **516**, 387–390 (2014).

588 Uganda Bureau of Statistics (UBOS), ICF, *Uganda Demographic and Health Survey 2016:*
589 *Key Indicators Report* (Kampala, Uganda: UBOS, and Rockville, Maryland, USA: UBOS
590 and ICF, 2017).

591 World Health Organization, *Guidance framework for testing genetically modified mosquitoes*
592 (2014).

593 L. Alphey, Malaria control with genetically manipulated insect vectors. *Science* **298**, 119–
594 121 (2002).

595 A. A. James, Gene drive systems in mosquitoes: rules of the road. *Trends in Parasitology*
596 **21**, 64–67 (2005).

597 S. James, F. H. Collins, P. A. Welkhoff, C. Emerson, H. C. J Godfray, M. Gottlieb, B. Green-
598 wood, S. W. Lindsay, C. M. Mbogo, F. O. Okumu, H. Quemada, M. Savadogo, J. A. Singh,
599 K. H. Tountas, Y. T. Toure, Pathway to deployment of gene drive mosquitoes as a poten-
600 tial biocontrol tool for elimination of malaria in sub-Saharan Africa: Recommendations of
601 a scientific working group. *American Journal of Tropical Medicine and Hygiene* **98**, 1–49
602 (2018).

603 I. Zeemeijer, Who Gets What, When and How?: New Corporate Land Acquisitions and
604 the Impact on Local Livelihoods in Uganda, Master's thesis, Utrecht University (2012).

605 Kalangala District Local Government District Management Improvement Plan 2012-2015,
606 *Tech. rep.* (2012).

607 M. Coetzee, R. H. Hunt, R. Wilkerson, A. della Torre, M. B. Coulibaly, N. J. Besan-
608 sky, *Anopheles coluzzii* and *Anopheles amharicus*, new members of the *Anopheles gambiae*
609 complex. *Zootaxa* **3619**, 246–274 (2013).

610 D. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated
611 individuals. *Genome Research* **19**, 1655–1664 (2009).

X. Liu, Y.-X. Fu, Exploring population size changes using SNP frequency spectra. *Nature Genetics* **47**, 555–559 (2015).

S. R. Browning, B. L. Browning, Accurate non-parametric estimation of recent effective population size from segments of identity by descent. *American Journal of Human Genetics* **97**, 404–418 (2015).

D. L. Stern, The genetic causes of convergent evolution. *Nature Reviews Genetics* **14**, 751–764 (2013).

B. Weir, C. Cockerham, Estimating F-statistics for the analysis of population structure. *Evolution* **38**, 1358–1370 (1984).

P. C. Sabeti, *et al.*, Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913–918 (2007).

N. R. Garud, P. W. Messer, E. O. Buzbas, D. A. Petrov, Recent selective sweeps in North American *Drosophila melanogaster* show signatures of soft sweeps. *PLoS Genetics* **11**, 1–32 (2015).

C. V. Edi, L. Djogbénou, A. M. Jenkins, K. Regna, M. A. Muskavitch, R. Poupardin, C. M. Jones, J. Essandoh, G. K. Kétoh, M. J. Paine, B. G. Koudou, M. J. Donnelly, H. Ranson, D. Weetman, CYP6 P450 Enzymes and ACE-1 Duplication Produce Extreme and Multiple Insecticide Resistance in the Malaria Mosquito *Anopheles gambiae*. *PLoS Genetics* **10** (2014).

D. A. Amenya, W. Chou, J. Li, G. Yan, P. D. Gershon, A. A. James, O. Marinotti, Proteomics reveals novel components of the *Anopheles gambiae* eggshell. *Journal of Insect Physiology* **56**, 1414–1419 (2010).

M. Moreno, P. Salgueiro, J. L. Vicente, J. Cano, P. J. Berzosa, A. de Lucio, F. Simard, A. Caccone, V. E. Do Rosario, J. Pinto, A. Benito, Genetic population structure of *Anopheles gambiae* in Equatorial Guinea. *Malaria Journal* **6**, 137 (2007).

J. C. Marshall, J. Pinto, J. D. Charlwood, G. Gentile, F. Santolamazza, F. Simard, A. della Torre, M. J. Donnelly, A. Caccone, 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 (2008).

C. D. Marsden, A. Cornel, Y. Lee, M. R. Sanford, L. C. Norris, P. B. Goodell, C. C. Nieman, S. Han, A. Rodrigues, J. Denis, A. Ouledi, G. C. Lanzaro, An analysis of two island groups as potential sites for trials of transgenic mosquitoes for malaria control. *Evolutionary Applications* **6**, 706–720 (2013).

D. Maliti, H. Ranson, S. Magesa, W. Kisinza, J. Mcha, K. Haji, G. Killeen, D. Weetman, 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**, e110910 (2014).

H. Chen, N. Minakawa, J. Beier, G. Yan, Population genetic structure of *Anopheles gambiae* mosquitoes on Lake Victoria islands, west Kenya. *Malaria Journal* **3**, 48 (2004).

J. K. Kayondo, L. G. Mukwaya, A. Stump, A. P. Michel, M. B. Coulibaly, N. J. Besansky, F. H. Collins, Genetic structure of *Anopheles gambiae* populations on islands in northwestern Lake Victoria, Uganda. *Malaria Journal* **4**, 59 (2005).

M. Lukindu, C. Bergey, R. Wiltshire, S. Small, B. Bourke, J. Kayondo, N. Besansky, Spatiotemporal genetic structure of *Anopheles gambiae* in the Northwestern Lake Victoria Basin, Uganda: Implications for genetic control trials in malaria endemic regions. *Parasites and Vectors* **11** (2018).

R. M. Wiltshire, C. M. Bergey, J. K. Kayondo, J. Birungi, L. G. Mukwaya, S. J. Emrich, N. J. Besansky, F. H. Collins, Reduced-representation sequencing identifies small effective population sizes of *Anopheles gambiae* in the north-western Lake Victoria basin, Uganda. *Malaria Journal* **17**, 285 (2018).

A. Thomas, The vegetation of the Sese Islands, Uganda: An illustration of edaphic factors in tropical ecology. *Journal of Ecology* **29**, 330–353 (1941).

G. D. Hale Carpenter, *A Naturalist on Lake Victoria* (E. P. Dutton and Company, New York, NY, 1920).

Uganda Bureau of Statistics, *2002 Uganda Population and Housing Census Analytical Report* (2002).

Uganda Bureau of Statistics, *The National Population and Housing Census 2014 - Main Report* (2016).

National Malaria Control Programme, Abt Associates, the INFORM Project, *An epidemiological profile of malaria and its control in Uganda* (2013).

J. E. Crawford, M. M. Riehle, K. Markianos, E. Bischoff, W. M. Guelbeogo, A. Gneme, N. Sagnon, K. D. Vernick, R. Nielsen, B. P. Lazzaro, Evolution of GOUNDRY, a cryptic subgroup of *Anopheles gambiae s.l.*, and its impact on susceptibility to *Plasmodium* infection. *Molecular Ecology* **25** (2016).

D. J. Lawson, L. van Dorp, D. Falush, A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nature Communications* **9**, 1–11 (2018).

J. Vontas, L. Grigoraki, J. Morgan, D. Tsakireli, G. Fuseini, L. Segura, J. Niemczura de Carvalho, R. Nguema, D. Weetman, M. A. Slotman, J. Hemingway, Rapid selection of a pyrethroid metabolic enzyme CYP9K1 by operational malaria control activities. *Proceedings of the National Academy of Sciences* p. 201719663 (2018).

B. Fossog Tene, R. Poupardin, C. Costantini, P. Awono-Ambene, C. S. Wondji, H. Ranson, C. Antonio-Nkondjio, 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** (2013).

A. A. Enayati, H. Ranson, J. Hemingway, Insect glutathione transferases and insecticide resistance. *Insect Molecular Biology* **14**, 3–8 (2005).

S. N. Mitchell, D. J. Rigden, A. J. Dowd, F. Lu, C. S. Wilding, D. Weetman, S. Dadzie, A. M. Jenkins, K. Regna, P. Boko, L. Djogbenou, M. A. T. Muskavitch, H. Ranson, M. J. I. Paine, O. Mayans, M. J. Donnelly, Metabolic and target-site mechanisms combine to confer strong DDT resistance in *Anopheles gambiae*. *PLoS ONE* **9** (2014).

C. M. Jones, H. K. Toé, A. Sanou, M. Namountougou, A. Hughes, A. Diabaté, R. Dabiré, F. Simard, H. Ranson, Additional selection for insecticide resistance in urban malaria vectors: DDT resistance in *Anopheles arabiensis* from Bobo-Dioulasso, Burkina Faso. *PLoS ONE* **7** (2012).

C. Fouet, C. Kamdem, S. Gamez, B. J. White, Genomic insights into adaptive divergence and speciation among malaria vectors of the *Anopheles nili* group. *Evolutionary Applications* **10**, 897–906 (2017).

R. Hardie, F. Martin, G. Cochrane, M. Juusola, P. Georgiev, P. Raghu, Molecular basis of amplification in *Drosophila* phototransduction: Roles for G protein, phospholipase C, and diacylglycerol kinase. *Neuron* **36**, 689–701 (2002).

Y. Huang, J. Xie, T. Wang, A fluorescence-based genetic screen to study retinal degeneration in *Drosophila*. *PLoS ONE* **10**, 1–19 (2015).

P. Kain, T. S. Chakraborty, S. Sundaram, O. Siddiqi, V. Rodrigues, G. Hasan, 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**, 4745–4755 (2008).

P. R. Senthilan, D. Piepenbrock, G. Ovezmyradov, B. Nadrowski, S. Bechstedt, S. Pauls, M. Winkler, W. M??bius, J. Howard, M. C. G??pfert, *Drosophila* auditory organ genes and genetic hearing defects. *Cell* **150**, 1042–1054 (2012).

C. S. Nelson, J. N. Beck, K. A. Wilson, E. R. Pilcher, P. Kapahi, R. B. Brem, Cross-phenotype association tests uncover genes mediating nutrient response in *Drosophila*. *BMC Genomics* **17**, 1–14 (2016).

Y. H. Lin, Y. C. Chen, T. Y. Kao, Y. C. Lin, T. E. Hsu, Y. C. Wu, W. W. Ja, T. J. Brummel, P. Kapahi, C. H. Yuh, L. K. Yu, Z. H. Lin, R. J. You, Y. T. Jhong, H. D. Wang, Diacylglycerol lipase regulates lifespan and oxidative stress response by inversely modulating TOR signaling in *Drosophila* and *C. elegans*. *Aging Cell* **13**, 755–764 (2014).

G. De Jong, Z. Bochdanovits, Latitudinal clines in *Drosophila melanogaster*: Body size, allozyme frequencies, inversion frequencies, and the insulin-signalling pathway. *Journal of Genetics* **82**, 207–223 (2003).

D. K. Fabian, M. Kapun, V. Nolte, R. Kofler, P. S. Schmidt, C. Schlötterer, T. Flatt, Genome-wide patterns of latitudinal differentiation among populations of *Drosophila melanogaster* from North America. *Molecular Ecology* **21**, 4748–4769 (2012).

C. Cheng, J. C. Tan, M. W. Hahn, N. J. Besansky, Systems genetic analysis of inversion polymorphisms in the malaria mosquito *Anopheles gambiae*. *Proceedings of the National Academy of Sciences* p. 201806760 (2018).

M. C. Whitlock, D. E. McCauley, Indirect measures of gene flow and migration: FST not equal to 1/(4Nm + 1). *Heredity* **82 (Pt. 2)**, 117–125 (1999).

A. J. Stow, W. E. Magnusson, Genetically defining populations is of limited use for evaluating and managing human impacts on gene flow. *Wildlife Research* **39**, 290–294 (2012).

A. Storfer, M. A. Murphy, S. F. Spear, R. Holderegger, L. P. Waits, Landscape genetics: Where are we now? *Molecular Ecology* **19**, 3496–3514 (2010).

J. A. Scott, W. G. Brogdon, F. H. Collins, Identification of single specimens of the *Anopheles gambiae* complex by the polymerase chain reaction. *The American Journal of Tropical Medicine and Hygiene* **49**, 520–529 (1993).

M. R. Green, J. Sambrook, *Molecular Cloning: A Laboratory Manual* (2012).

E. Aronesty, ea-utils: Command-line tools for processing biological sequencing data (2011).

R. Holt, *et al.*, The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* **298**, 129–149 (2002).

M. V. Sharakhova, M. P. Hammond, N. F. Lobo, J. Krzywinski, M. F. Unger, M. E. Hillenmeyer, R. V. Bruggner, E. Birney, F. H. Collins, Update of the *Anopheles gambiae* PEST genome assembly. *Genome Biology* **8**, R5 (2007).

H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, M. J. Daly, A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43**, 491–498 (2011).

S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, P. C. Sham, PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* **81**, 559–75 (2007).

C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, J. J. Lee, Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).

O. Delaneau, B. Howie, A. J. Cox, J. F. Zagury, J. Marchini, Haplotype estimation using sequencing reads. *American Journal of Human Genetics* **93**, 687–696 (2013).

H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–9 (2009).

N. M. Kopelman, J. Mayzel, M. Jakobsson, N. A. Rosenberg, I. Mayrose, CLUMPAK: a program for identifying clustering modes and packaging population structure inferences across K. *Molecular Ecology Resources* **15**, 1179–1191 (2015).

R. R. Hudson, M. Slatkin, W. P. Maddison, Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**, 583–589 (1992).

N. Patterson, A. L. Price, D. Reich, Population structure and eigenanalysis. *PLoS Genetics* **2**, 2074–2093 (2006).

A. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. a. Shadick, D. Reich, Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**, 904–9 (2006).

N. Mantel, The detection of disease clustering and a generalized regression approach. *Cancer Research* **27**, 209–220 (1967).

S. Dray, A. B. Dufour, The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software* **22**, 1–20 (2007).

B. L. Browning, S. R. Browning, Detecting identity by descent and estimating genotype error rates in sequence data. *American Journal of Human Genetics* **93**, 840–851 (2013).

C. S. Clarkson, A. Miles, N. J. Harding, D. Weetman, D. Kwiatkowski, M. Donnelly, The *Anopheles gambiae* 1000 Genomes Consortium, The genetic architecture of target-site resistance to pyrethroid insecticides in the African malaria vectors *Anopheles gambiae* and *Anopheles coluzzii*. *bioRxiv* p. Preprint at: https://www.biorxiv.org/content/early/2018/08/06/323980 (2018).

M. P. Hare, L. Nunney, M. K. Schwartz, D. E. Ruzzante, M. Burford, R. S. Waples, K. Ruegg, F. Palstra, Understanding and estimating effective population size for practical application in marine species management. *Conservation Biology* **25**, 438–449 (2011).

R. N. Gutenkunst, R. D. Hernandez, S. H. Williamson, C. D. Bustamante, Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics* **5**, e1000695 (2009).

A. J. Coffman, P. H. Hsieh, S. Gravel, R. N. Gutenkunst, Computationally efficient composite likelihood statistics for demographic inference. *Molecular Biology and Evolution* **33**, 591–593 (2016).

M. C. Fontaine, *et al.*, Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science* **347**, 1258524 (2014).

Z. Rosen, A. Bhaskar, Y. S. Song, Geometry of the sample frequency spectrum and the perils of demographic inference. *Genetics* **210**, 665–682 (2018).

P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin, The variant call format and VCFtools. *Bioinformatics* **27**, 2156–8 (2011).

Z. A. Szpiech, R. D. Hernandez, selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Molecular Biology and Evolution* **31**, 2824–2827 (2014).

I. A. Randhawa, M. Khatkar, P. Thomson, H. Raadsma, Composite selection signals can localize the trait specific genomic regions in multi-breed populations of cattle and sheep. *BMC Genetics* **15**, 34 (2014).

A. Wallberg, C. W. Pirk, M. H. Allsopp, M. T. Webster, Identification of multiple loci associated with social parasitism in honeybees. *PLoS Genetics* **12**, 1–30 (2016).

H. Li, Tabix: Fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics* **27**, 718–719 (2011).

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

E. Paradis, J. Claude, K. Strimmer, A{PE}: analyses of phylogenetics and evolution in {R} language. *Bioinformatics* **20**, 289–290 (2004).

E. Neuwirth, RColorBrewer: ColorBrewer Palettes (2014).

T. Galili, dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering. *Bioinformatics* (2015).

M. Gautier, R. Vitalis, rehh: An R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics* **28**, 1176–1177 (2012).

O. Tange, GNU Parallel - The Command-Line Power Tool. *;login: The USENIX Magazine* **36**, 42–47 (2011).

A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–2 (2010).

Data and materials availability

All scripts used in the analysis are available at 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.

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.

Conflict of Interest Statement

The authors declare no competing financial interests.

Supplemental Material for: Assessing connectivity despite high diversity in island populations of a malaria mosquito

Christina M. Bergey, Martin Lukindu, Rachel M. Wiltshire, Michael C. Fontaine, Jonathan K. Kayondo, and Nora J. Besansky

Tables

Table S1: Sampling sites and coordinates.

Location	Latitude	Longitude	Sample Count
Banda	-0.25893	32.39594	11
Bugala - Bugoma	-0.26697	32.07936	11
Bugala - Lutoboka	-0.31624	32.29246	7
Bugala - Mweena	-0.32806	32.31113	5
Bukasa	-0.48609	32.45091	11
Buwama	0.02077	32.10574	11
Kaazi	-0.31831	31.88183	11
Kiyindi	0.27558	33.14699	10
Nsadzi	-0.08632	32.58895	11
Sserinya	-0.26476	32.37228	16
Wamala	0.40811	31.99609	11

Table S2: List of individuals included in study with mean depth of sequencing coverage.

ID	Field ID	Island	Site	Mean depth
LVB2015-1	CM-KSB-J5	Nsadzi	Kansambwe	20.40
LVB2015-2	K-KSB-E1	Nsadzi	Kansambwe	24.50
LVB2015-3	RM-KSB-G1	Nsadzi	Kansambwe	17.90
LVB2015-4	NKG-F-G3	Bukasa	Nakibanga	4.90
LVB2015-6	NKG-F-H1	Bukasa	Nakibanga	15.80
LVB2015-7	NKG-K-I1	Bukasa	Nakibanga	15.80
LVB2015-8	NKG-K-K1	Bukasa	Nakibanga	19.90
LVB2015-9	NKG-M-C1	Bukasa	Nakibanga	22.30
LVB2015-10	NKG-M-D1	Bukasa	Nakibanga	20.30
LVB2015-11	NKG-M-F1	Bukasa	Nakibanga	23.90
LVB2015-14	MWN-K-A1	Bugala	Mweena	21.20
LVB2015-15	MWN-K-C2	Bugala	Mweena	18.30
LVB2015-16	MWN-P-D1	Bugala	Mweena	5.34
LVB2015-17	MWN-R-E1	Bugala	Mweena	22.60
LVB2015-18	MWN-R-F1	Bugala	Mweena	17.10
LVB2015-19	BDA-K-B1	Banda	Banda	14.40
LVB2015-20	BDA-K-B2	Banda	Banda	18.00

LVB2015-21	BBS-C-M1	Sserinya	Bbosa	19.70
LVB2015-22	BBS-F-F1	Sserinya	Bbosa	20.80
LVB2015-24	BBS-K-J3	Sserinya	Bbosa	22.80
LVB2015-25	BBS-K-J8	Sserinya	Bbosa	17.30
LVB2015-26	BBS-K-K2	Sserinya	Bbosa	18.20
LVB2015-27	BBS-M-L1	Sserinya	Bbosa	22.40
LVB2015-28	BBS-P-I4	Sserinya	Bbosa	23.30
LVB2015-29	BBS-R-A2	Sserinya	Bbosa	19.60
LVB2015-30	BBS-R-C1	Sserinya	Bbosa	16.10
LVB2015-32	KSS-F-E2	Sserinya	Kasisa	21.00
LVB2015-33	LBK-C-F1	Bugala	Lutoboka	21.00
LVB2015-34	LBK-C-F6	Bugala	Lutoboka	18.60
LVB2015-35	LBK-C-G6	Bugala	Lutoboka	20.90
LVB2015-36	LBK-K-E2	Bugala	Lutoboka	22.00
LVB2015-37	LBK-M-A1	Bugala	Lutoboka	18.50
LVB2015-39	LBK-R-O1	Bugala	Lutoboka	20.20
LVB2015-42	BGM-F-D1	Bugala	Bugoma	16.70
LVB2015-43	BGM-F-E2	Bugala	Bugoma	24.50
LVB2015-45	BGM-K-M2	Bugala	Bugoma	23.80
LVB2015-46	BGM-M-G1	Bugala	Bugoma	18.40
LVB2015-47	BGM-M-H2	Bugala	Bugoma	14.10
LVB2015-48	BGM-M-J1	Bugala	Bugoma	20.90
LVB2015-50	BGM-P-F9	Bugala	Bugoma	18.90
LVB2015-51	BGM-R-O2	Bugala	Bugoma	16.80
LVB2015-52	KZI-F-F001	Kaazi	Nabugabo	19.10
LVB2015-53	KZI-F-G001	Kaazi	Nabugabo	19.50
LVB2015-54	KZI-F-H001	Kaazi	Nabugabo	19.40
LVB2015-55	KZI-P-A001	Kaazi	Nabugabo	10.30
LVB2015-56	KZI-P-B005	Kaazi	Nabugabo	16.50
LVB2015-59	KZI-R-C003	Kaazi	Nabugabo	18.10
LVB2015-60	KZI-R-D007	Kaazi	Nabugabo	15.90
LVB2015-61	BWM-C-G001	Buwama	Buwama	16.10
LVB2015-62	BWM-C-H001	Buwama	Buwama	11.10
LVB2015-63	BWM-F-A001	Buwama	Buwama	20.40
LVB2015-64	BWM-F-B001	Buwama	Buwama	21.50
LVB2015-65	BWM-P-J001	Buwama	Buwama	14.80
LVB2015-66	BWM-R-C002	Buwama	Buwama	19.10

LVB2015-67	BWM-R-F005	Buwama	Buwama	22.30
LVB2015-68	NMA-C-E003	Wamala	Naama	20.30
LVB2015-69	NMA-C-F002	Wamala	Naama	17.80
LVB2015-70	NMA-F-A001	Wamala	Naama	13.10
LVB2015-71	NMA-K-B001	Wamala	Naama	22.10
LVB2015-72	NMA-K-C002	Wamala	Naama	18.00
LVB2015-73	NMA-P-G001	Wamala	Naama	18.20
LVB2015-74	NMA-P-H003	Wamala	Naama	16.60
LVB2015-76	KYD-C-G001	Kiyindi	Kiyindi	16.10
LVB2015-77	KYD-C-H001	Kiyindi	Kiyindi	11.80
LVB2015-78	KYD-C-I001	Kiyindi	Kiyindi	16.40
LVB2015-79	KYD-C-J002	Kiyindi	Kiyindi	11.50
LVB2015-80	KYD-F-A003	Kiyindi	Kiyindi	10.30
LVB2015-81	KYD-F-B004	Kiyindi	Kiyindi	21.50
LVB2015-82	KYD-K-D002	Kiyindi	Kiyindi	18.40
LVB2015-84	KYD-R-K001	Kiyindi	Kiyindi	16.80
LVB2015-89	BDA-K-E2	Banda	Banda	15.10
LVB2015-90	BDA-K-F1	Banda	Banda	25.10
LVB2015-91	BDA-M-N1	Banda	Banda	25.60
LVB2015-92	BDA-M-O4	Banda	Banda	17.60
LVB2015-93	BDA-M-Q1	Banda	Banda	39.20
LVB2015-96	CM-KSB-J2	Nsadzi	Kansambwe	9.22
LVB2015-97	CM-KSB-J3	Nsadzi	Kansambwe	10.10
LVB2015-98	CM-KSB-J6	Nsadzi	Kansambwe	16.90
LVB2015-100	K-KSB-D1	Nsadzi	Kansambwe	6.05
LVB2015-101	ML-KSB-M1	Nsadzi	Kansambwe	4.27
LVB2015-102	ML-KSB-M2	Nsadzi	Kansambwe	19.90
LVB2015-103	RM-KSB-G2	Nsadzi	Kansambwe	14.20
LVB2015-104	RM-KSB-G3	Nsadzi	Kansambwe	17.50
LVB2015-105	NKG-R-A12	Bukasa	Nakibanga	15.30
LVB2015-106	NKG-C-E1	Bukasa	Nakibanga	16.20
LVB2015-108	NKG-K-C5	Bukasa	Nakibanga	18.50
LVB2015-109	NKG-M-A1	Bukasa	Nakibanga	12.80
LVB2015-112	BDA-K-D4	Banda	Banda	12.70
LVB2015-113	BDA-K-E3	Banda	Banda	12.20
LVB2015-114	BDA-M-N5	Banda	Banda	15.00
LVB2015-115	BDA-M-P1	Banda	Banda	16.80

LVB2015-116	BBS-C-M3	Sserinya	Bbosa	16.60
LVB2015-117	BBS-K-J1	Sserinya	Bbosa	18.80
LVB2015-118	BBS-K-J11	Sserinya	Bbosa	14.60
LVB2015-120	BBS-K-K6	Sserinya	Bbosa	18.10
LVB2015-121	BBS-P-I8	Sserinya	Bbosa	15.00
LVB2015-122	BBS-R-A19	Sserinya	Bbosa	15.50
LVB2015-125	LBK-R-A5	Bugala	Lutoboka	18.10
LVB2015-126	BGM-K-K1	Bugala	Bugoma	15.20
LVB2015-128	BGM-M-H4	Bugala	Bugoma	20.30
LVB2015-129	BGM-P-F4	Bugala	Bugoma	19.00
LVB2015-130	KZI-F-G005	Kaazi	Nabugabo	18.60
LVB2015-131	KZI-P-A007	Kaazi	Nabugabo	15.80
LVB2015-132	KZI-R-C012	Kaazi	Nabugabo	15.10
LVB2015-133	KZI-R-E011	Kaazi	Nabugabo	16.30
LVB2015-134	BWM-P-I001	Buwama	Buwama	18.20
LVB2015-135	BWM-P-K002	Buwama	Buwama	19.30
LVB2015-136	BWM-R-D001	Buwama	Buwama	14.40
LVB2015-137	BWM-R-F002	Buwama	Buwama	19.90
LVB2015-138	NMA-C-E006	Wamala	Naama	21.90
LVB2015-139	NMA-C-F003	Wamala	Naama	20.40
LVB2015-140	NMA-P-G003	Wamala	Naama	18.90
LVB2015-141	NMA-R-I001	Wamala	Naama	14.10
LVB2015-142	KYD-F-B006	Kiyindi	Kiyindi	18.10
LVB2015-143	KYD-K-E003	Kiyindi	Kiyindi	14.20

Table S3: Results of two population demographic inference with IM model in δaδi when comparing island to island localities. Numbers in parentheses are bounds of 95% confidence interval computed using Fisher information matrix and 100 bootstrap replicates of 1 Mb from the dataset.

Localities	N_a	% Pop. 1 in Split	Pop. 1 ν_F	Pop. 2 ν_F	Time since split	m_{12}	m_{21}
Banda - Bugala (I)	531,000	0.603	1.94	9,800	3,290	None	None
	(530,000, 532,000)	(0.596, 0.609)	(1.8, 2.09)	(7,060, 12,500)	(3,190, 3,390)		
Banda - Bukasa	526,000	0.568	14.7	9,080	7,580	None	None
	(525,000, 527,000)	(0.556, 0.581)	(13.7, 15.6)	(7,240, 10,900)	(7,470, 7,690)		
Banda - Nsadzi	527,000	0.518	47.1	9,880	9,340	None	None
	(526,000, 528,000)	(0.502, 0.534)	(39.9, 54.4)	(7,310, 12,400)	(9,160, 9,530)		
Banda - Sserinya	531,000	0.489	10.5	9,840	4,430	None	None
	(530,000, 532,000)	(0.464, 0.514)	(8.69, 12.4)	(7,390, 12,300)	(4,250, 4,610)		
Bugala (I) - Bukasa	527,000	0.49	9,840	536	5,290	None	None
	(526,000, 528,000)	(0.471, 0.509)	(7,850, 11,800)	(447, 624)	(5,170, 5,410)		
Bugala (I) - Nsadzi	526,000	0.56	8,980	128	6,680	None	None
	(525,000, 527,000)	(0.536, 0.583)	(6,920, 11,000)	(110, 146)	(6,440, 6,910)		
Bugala (I) - Sserinya	530,000	0.61	5,850	49	2,130	None	None
	(529,000, 531,000)	(0.571, 0.65)	(3,610, 8,090)	(38.3, 59.6)	(1,950, 2,300)		
Bukasa - Nsadzi	527,000	0.499	3,420	335	9,340	None	None
	(526,000, 528,000)	(0.49, 0.507)	(2,860, 3,980)	(293, 377)	(9,190, 9,500)		
Bukasa - Sserinya	525,000	0.503	9,840	6,760	8,930	None	None
	(524,000, 526,000)	(0.495, 0.51)	(7,490, 12,200)	(5,630, 7,880)	(8,780, 9,080)		
Nsadzi - Sserinya	540,000	0.538	893	9,900	9,540	None	None
	(539,000, 541,000)	(0.521, 0.554)	(612, 1,170)	(7,810, 12,000)	(9,280, 9,790)		

Table S4: Results of two population demographic inference with IM model in $\delta a \delta i$ when comparing island to mainland localities. Numbers in parentheses are bounds of 95% confidence interval computed using Fisher information matrix and 100 bootstrap replicates of 1 Mb from the dataset.

Localities	N_a	% Pop. 1 in Split	Pop. 1 ν_F	Pop. 2 ν_F	Time since split	m_{12}	m_{21}
Banda - Bugala (M)	522,000	0.599	3.83	9,900	5,400	None	None
	(521,000, 523,000)	(0.586, 0.612)	(3.59, 4.07)	(8,410, 11,400)	(5,300, 5,500)		
Banda - Buwama	522,000	0.51	1.09	8,520	3,040	None	None
	(521,000, 523,000)	(0.491, 0.529)	(1.01, 1.16)	(6,830, 10,200)	(2,940, 3,150)		
Banda - Kaazi	522,000	0.563	3.99	9,960	5,890	None	None
	(521,000, 523,000)	(0.55, 0.575)	(3.71, 4.28)	(8,370, 11,500)	(5,760, 6,030)		
Banda - Kiyindi	510,000	0.568	1.72	9,910	3,910	None	None
	(509,000, 511,000)	(0.56, 0.577)	(1.69, 1.75)	(8,170, 11,600)	(3,880, 3,950)		
Banda - Wamala	521,000	0.562	3.9	7,970	5,510	None	None
	(520,000, 522,000)	(0.554, 0.57)	(3.66, 4.15)	(6,790, 9,140)	(5,390, 5,620)		
Bugala (I) - Bugala (M)	522,000	0.555	1,440	9,710	4,170	None	None
	(521,000, 523,000)	(0.543, 0.566)	(954, 1,930)	(7,420, 12,000)	(4,040, 4,300)		
Bugala (I) - Buwama	523,000	0.499	0.215	130	190	None	None
	(522,000, 524,000)	(0.496, 0.503)	(0.209, 0.222)	(2.12, 258)	(188, 192)		
Bugala (I) - Kaazi	523,000	0.366	1,420	8,510	5,060	None	None
	(522,000, 524,000)	(0.348, 0.384)	(1,210, 1,620)	(6,610, 10,400)	(4,890, 5,230)		
Bugala (I) - Kiyindi	508,000	0.363	1,350	5,910	3,580	None	None
	(507,000, 509,000)	(0.342, 0.383)	(1,070, 1,620)	(4,470, 7,350)	(3,420, 3,740)		
Bugala (I) - Wamala	520,000	0.486	2,060	9,270	3,700	None	None
	(519,000, 521,000)	(0.462, 0.51)	(1,610, 2,500)	(6,790, 11,800)	(3,570, 3,830)		
Bugala (M) - Bukasa	521,000	0.53	9,000	66.3	4,830	None	None
	(520,000, 522,000)	(0.508, 0.552)	(6,720, 11,300)	(52.2, 80.5)	(4,670, 4,990)		
Bugala (M) - Nsadzi	535,000	0.439	9,220	7.59	4,500	None	None

	(534,000, 536,000)	(0.423, 0.455)	(7,500, 10,900)	(6.91, 8.26)	(4,390, 4,620)		
Bugala (M) - Sserinya	534,000	0.516	9,590	48.8	4,040	None	None
	(533,000, 535,000)	(0.497, 0.535)	(7,640, 11,500)	(42.4, 55.2)	(3,930, 4,160)		
Bukasa - Buwama	522,000	0.501	1.43	9,960	1,690	None	None
	(521,000, 523,000)	(0.496, 0.506)	(1.4, 1.46)	(7,250, 12,700)	(1,670, 1,710)		
Bukasa - Kaazi	522,000	0.501	41.5	9,410	5,490	None	None
	(521,000, 522,000)	(0.491, 0.51)	(33.8, 49.3)	(6,990, 11,800)	(5,320, 5,650)		
Bukasa - Kiyindi	508,000	0.361	18.6	9,220	3,360	None	None
	(507,000, 509,000)	(0.336, 0.386)	(15.8, 21.4)	(6,790, 11,600)	(3,190, 3,540)		
Bukasa - Wamala	520,000	0.39	304	8,060	5,200	None	None
	(519,000, 521,000)	(0.372, 0.409)	(257, 351)	(6,270, 9,850)	(5,050, 5,360)		
Buwama - Nsadzi	524,000	0.54	9,810	1.25	1,930	None	None
	(523,000, 525,000)	(0.502, 0.578)	(6,950, 12,700)	(1.07, 1.44)	(1,800, 2,070)		
Buwama - Sserinya	523,000	0.493	54.2	0.137	187	None	None
	(522,000, 524,000)	(0.488, 0.498)	(24.2, 84.2)	(0.134, 0.14)	(186, 189)		
Kaazi - Nsadzi	524,000	0.489	9,540	19.4	5,730	None	None
	(523,000, 525,000)	(0.47, 0.507)	(7,690, 11,400)	(16, 22.8)	(5,540, 5,910)		
Kaazi - Sserinya	523,000	0.598	6,980	104	4,550	None	None
	(522,000, 524,000)	(0.579, 0.618)	(5,440, 8,510)	(89.7, 119)	(4,410, 4,680)		
Kiyindi - Nsadzi	511,000	0.499	9,940	2.62	2,840	None	None
	(510,000, 512,000)	(0.491, 0.507)	(7,340, 12,500)	(2.26, 2.98)	(2,690, 2,990)		
Kiyindi - Sserinya	509,000	0.5	1,870	0.129	185	None	None
	(508,000, 510,000)	(0.496, 0.503)	(-5,010, 8,750)	(0.126, 0.132)	(183, 187)		
Wamala - Nsadzi	523,000	0.473	8,490	9.72	4,420	None	None
	(522,000, 524,000)	(0.454, 0.492)	(6,720, 10,200)	(8.57, 10.9)	(4,290, 4,560)		
Wamala - Sserinya	521,000	0.649	9,840	39.1	2,790	None	None
	(520,000, 522,000)	(0.643, 0.655)	(7,070, 12,600)	(38.1, 40.1)	(2,760, 2,820)		

Table S5: Results of two population demographic inference with IM model in $\delta a \delta i$ when comparing mainland to mainland localities. Numbers in parentheses are bounds of 95% confidence interval computed using Fisher information matrix and 100 bootstrap replicates of 1 Mb from the dataset.

Localities	N_a	% Pop. 1 in Split	Pop. 1 ν_F	Pop. 2 ν_F	Time since split	m_{12}	m_{21}
Bugala (M) - Buwama	523,000	0.498	9,590	1,090	368	None	None
	(523,000, 524,000)	(0.49, 0.506)	(802, 18,400)	(573, 1,620)	(320, 416)		
Bugala (M) - Kaazi	521,000	0.483	7,570	8,450	2,710	None	None
	(520,000, 522,000)	(0.458, 0.509)	(5,740, 9,400)	(5,860, 11,000)	(2,590, 2,830)		
Bugala (M) - Kiyindi	507,000	0.361	1.88	3,870	355	0.0000169	0
	(506,000, 508,000)	(0.318, 0.404)	(1.74, 2.01)	(1,810, 5,920)	(348, 362)	(-14.3, 14.3)	(-31,100, 31,100)
Bugala (M) - Wamala	521,000	0.51	13.1	323	186	0	0.00226
	(520,000, 521,000)	(0.491, 0.53)	(9.03, 17.1)	(-168, 814)	(179, 194)	(-162, 162)	(-3,980, 3,980)
Buwama - Kaazi	523,000	0.553	706	186	181	3.16	2.12
	(522,000, 524,000)	(0.357, 0.749)	(-764, 2,180)	(-101, 472)	(122, 239)	(-5,690, 5,690)	(-836, 841)
Buwama - Kiyindi	511,000	0.322	9,110	9,810	369	761	178
	(510,000, 512,000)	(0.287, 0.357)	(5,390, 12,800)	(4,390, 15,200)	(302, 436)	(-80,200, 81,700)	(-81,200, 81,500)
Buwama - Wamala	523,000	0.365	31.7	19.7	124	9.27	0
	(522,000, 524,000)	(-1,000, 1,000)	(22, 41.3)	(13.6, 25.8)	(107, 141)	(-578, 596)	(-344, 344)
Kaazi - Kiyindi	510,000	0.448	9,900	1,670	438	923	718
	(509,000, 511,000)	(0.325, 0.571)	(1,100, 18,700)	(749, 2,600)	(365, 510)	(-64,900, 66,800)	(610, 825)
Kaazi - Wamala	521,000	0.581	9,200	4,820	974	None	None
	(520,000, 522,000)	(0.515, 0.647)	(3,680, 14,700)	(2,430, 7,210)	(795, 1,150)		
Kiyindi - Wamala	510,000	0.454	134	40.3	181	0	0.00363
	(509,000, 511,000)	(0.391, 0.518)	(57.9, 209)	(25.4, 55.3)	(166, 197)	(-663, 663)	(-3.69, 3.7)

Table S6: Locality-specific (in LVB) putative sweeps based on H12 statistic.

Site	Count	Chr.	Putative Sweeps	Other sites[1]
Banda	44	2L	28.6 Mb; 36 Mb; 36.4 Mb; 36.9 Mb; 37.6 Mb; 38.1 Mb; 39.1 Mb; 42.2 Mb; 43.4 Mb; 43.8 Mb; 44.3 Mb; 44.9 Mb; 45.4 Mb	1 also found in BFS, GNS
		2R	4.2 Mb; 12.3 Mb; 18.3 Mb; 23.6 Mb; 29.4 Mb; 30.3 Mb; 33.7 Mb; 34.8 Mb; 35.8 Mb; 36.5 Mb; 44.1 Mb; 44.6 Mb; 49.7 Mb	1 also found in BFM, BFS, CMS, GNS, GWA; 1 also found in BFM, GWA; 5 also found in GWA
		3L	18.5 Mb; 21.6 Mb; 23.4 Mb; 23.9 Mb; 32.8 Mb	
		3R	2.6 Mb; 7.9 Mb; 29.2 Mb; 30.5 Mb; 31.3 Mb; 32.1 Mb; 33.2 Mb; 45.3 Mb; 46.4 Mb; 47 Mb	1 also found in GNS
		X	0.5 Mb; 2.1 Mb; 4.3 Mb	1 also found in AOM
Bugala (I)	24	2L	2.5 Mb; 5.5 Mb; 7.1 Mb; 19 Mb; 31.1 Mb; 43 Mb; 45.7 Mb	1 also found in AOM, BFM, BFS, CMS, GAS, GNS, UGS; 1 also found in AOM, UGS
		2R	6.7 Mb; 21.1 Mb; 24 Mb; 24.6 Mb; 35.6 Mb; 37.1 Mb; 38.6 Mb; 39 Mb; 55.9 Mb	1 also found in BFM, GWA; 2 also found in GWA
		3L	17.2 Mb; 29.5 Mb	
		3R	26 Mb; 35.8 Mb; 37.5 Mb	
		X	3.5 Mb; 5.7 Mb; 10.8 Mb	
Bukasa	112	2L	12.6 Mb; 13.6 Mb; 17.7 Mb; 20.1 Mb; 20.9 Mb; 21.6 Mb; 22.7 Mb; 23.6 Mb; 24.7 Mb; 25.4 Mb; 26.2 Mb; 26.9 Mb; 27.3 Mb; 27.8 Mb; 28.4 Mb; 29.1 Mb; 30.1 Mb; 31.5 Mb; 32.3 Mb; 33.3 Mb; 35.8 Mb; 39.4 Mb; 39.8 Mb; 40.6 Mb; 41.4 Mb; 43.1 Mb; 45.6 Mb; 48.1 Mb; 49.3 Mb	1 also found in AOM, BFM, BFS, CMS, GAS, GNS; 1 also found in BFM, GAS; 1 also found in BFS, GAS, GNS; 1 also found in BFS, GNS; 2 also found in CMS; 2 also found in GAS; 1 also found in GWA

		2R	1.3 Mb; 4.7 Mb; 5.3 Mb; 7.2 Mb; 7.6 Mb; 8 Mb; 9.7 Mb; 10.5 Mb; 12 Mb; 12.4 Mb; 13.5 Mb; 14 Mb; 15.7 Mb; 16.9 Mb; 17.5 Mb; 19.4 Mb; 22.9 Mb; 24.9 Mb; 25.8 Mb; 26.6 Mb; 29.9 Mb; 30.8 Mb; 32.4 Mb; 33.4 Mb; 35.5 Mb; 37.6 Mb; 43 Mb; 45.6 Mb; 47 Mb; 49.5 Mb; 54.8 Mb	1 also found in AOM, GAS, GWA; 2 also found in BFM; 1 also found in BFM, GWA; 3 also found in GAS, GWA; 6 also found in GWA
		3L	7.3 Mb; 11.6 Mb; 13.1 Mb; 15.6 Mb; 18.1 Mb; 19.1 Mb; 19.8 Mb; 20.6 Mb; 24.2 Mb; 25.2 Mb; 27.3 Mb; 28 Mb; 28.7 Mb; 29.7 Mb; 30.6 Mb; 33.7 Mb; 34.7 Mb; 35.3 Mb; 36.1 Mb; 38.7 Mb; 39.9 Mb; 40.3 Mb; 41.2 Mb	2 also found in BFM; 1 also found in GAS
		3R	5.1 Mb; 5.9 Mb; 7.2 Mb; 8.9 Mb; 12.7 Mb; 13.3 Mb; 14.1 Mb; 14.9 Mb; 15.9 Mb; 17.2 Mb; 22.3 Mb; 23.3 Mb; 23.8 Mb; 24.9 Mb; 26.8 Mb; 27.9 Mb; 31.4 Mb; 33 Mb; 35.9 Mb; 36.9 Mb	1 also found in GWA
		X	1.7 Mb; 2.8 Mb; 4.9 Mb; 6 Mb; 7 Mb; 11.5 Mb; 12.5 Mb; 13.6 Mb; 16.7 Mb	1 also found in BFM, GAS, GWA; 4 also found in GAS
Buwama	27	2L	14.9 Mb; 15.9 Mb; 25.1 Mb; 26.5 Mb; 31.6 Mb	1 also found in BFM, GAS; 1 also found in GWA
		2R	24.4 Mb; 39.5 Mb; 44.5 Mb; 46.3 Mb; 49.1 Mb; 53.7 Mb; 55.3 Mb	1 also found in AOM, BFM, CMS; 1 also found in BFS, CMS, GNS; 1 also found in CMS; 1 also found in GWA
		3L	2.4 Mb; 3.1 Mb; 3.6 Mb; 4.1 Mb; 10.6 Mb; 16.1 Mb; 21.7 Mb; 29.8 Mb	
		3R	18 Mb; 29.1 Mb; 35.5 Mb; 37.7 Mb; 38.4 Mb; 38.9 Mb; 40.6 Mb	1 also found in GNS

Kaazi	15	2L	8.5 Mb; 34.6 Mb	1 also found in AOM
		2R	8.3 Mb; 23 Mb	1 also found in AOM; 1 also found in GAS
		3L	3.5 Mb; 4.8 Mb; 8.6 Mb; 11.8 Mb; 13 Mb; 15.8 Mb; 25 Mb; 26.8 Mb	1 also found in BFM; 1 also found in BFM, GNS
		3R	14.7 Mb; 15.6 Mb; 46 Mb	1 also found in GNS
Kiyindi	40	2L	2 Mb; 10.6 Mb; 17.8 Mb; 22.1 Mb; 23.9 Mb; 26 Mb; 28.7 Mb; 29.9 Mb; 34.8 Mb	1 also found in AOM, BFM, BFS, CMS, GNS, UGS; 1 also found in BFS, GNS; 3 also found in GAS
		2R	19.1 Mb; 20.2 Mb; 25.9 Mb; 35.3 Mb; 36.6 Mb; 38.1 Mb; 40 Mb; 41.7 Mb; 42.4 Mb; 45.3 Mb; 48.2 Mb; 48.6 Mb; 50.1 Mb; 52.2 Mb; 53.6 Mb; 54.7 Mb; 55.1 Mb	1 also found in AOM; 1 also found in AOM, BFS, CMS, GNS, GWA; 1 also found in BFM; 2 also found in GWA
		3L	1.2 Mb; 8.9 Mb; 12.1 Mb; 12.6 Mb; 13.5 Mb; 14.8 Mb; 15.4 Mb; 16 Mb; 16.8 Mb; 19.7 Mb; 26.7 Mb	1 also found in BFM; 1 also found in GWA
		3R	38 Mb; 41.9 Mb; 48.3 Mb	
Nsadzi	47	2L	23.2 Mb; 27 Mb; 45.5 Mb	
		2R	1.6 Mb; 2.3 Mb; 3.2 Mb; 4 Mb; 8.8 Mb; 10.2 Mb; 13.2 Mb; 16.1 Mb; 20 Mb; 21.3 Mb; 24.7 Mb; 30.5 Mb; 34.2 Mb; 37.3 Mb; 41.2 Mb; 43.5 Mb; 52 Mb	1 also found in BFM, GAS; 1 also found in BFM, GWA; 1 also found in BFS, CMS, GNS; 1 also found in CMS; 2 also found in GWA
		3L	10.5 Mb; 11 Mb; 24.3 Mb; 35 Mb; 35.4 Mb; 36.8 Mb; 37.6 Mb	1 also found in GAS
		3R	3.8 Mb; 6 Mb; 7.4 Mb; 19.9 Mb; 20.5 Mb; 21.4 Mb; 23 Mb; 24.2 Mb; 27.7 Mb; 41.6 Mb; 42.2 Mb; 48.2 Mb; 49.8 Mb; 50.4 Mb	1 also found in BFS, GNS
		X	0.7 Mb; 2.3 Mb; 5.2 Mb; 7.7 Mb; 11.9 Mb; 17.9 Mb	1 also found in BFM, GAS, GWA; 1 also found in GAS

Sserinya	35	2L	22.2 Mb; 24.2 Mb; 25.7 Mb; 33.2 Mb; 34.9 Mb; 35.4 Mb; 40.2 Mb; 41.1 Mb; 45.1 Mb; 45.9 Mb; 46.8 Mb	1 also found in BFM, GNS; 1 also found in CMS, GAS; 1 also found in GAS, GNS
		2R	0.4 Mb; 7.7 Mb; 21.5 Mb; 30 Mb; 32 Mb; 36.1 Mb	3 also found in GWA
		3L	10.1 Mb; 10.9 Mb; 14.6 Mb; 34.5 Mb; 41.8 Mb	1 also found in BFS, CMS, GNS, GWA, UGS
		3R	1.9 Mb; 10 Mb; 15 Mb; 24.8 Mb; 26.2 Mb; 27 Mb; 29 Mb	1 also found in GAS
		X	5.8 Mb; 12.7 Mb; 13.1 Mb; 18.1 Mb; 18.8 Mb; 21.3 Mb	1 also found in BFM, CMS, GAS, GWA; 1 also found in BFM, GAS, GWA; 1 also found in CMS, GNS, GWA; 2 also found in GAS
Wamala	25	2L	13.4 Mb; 15.5 Mb; 17.1 Mb; 19.1 Mb; 20 Mb	2 also found in GAS
		2R	21.2 Mb; 22.2 Mb; 29.6 Mb; 38.8 Mb; 39.7 Mb; 47.6 Mb; 48.3 Mb; 48.9 Mb	2 also found in AOM; 2 also found in GWA
		3L	3.3 Mb; 7.6 Mb; 8.2 Mb	
		3R	5 Mb; 39.2 Mb; 43.2 Mb; 46.5 Mb; 47.5 Mb; 48.5 Mb; 50.5 Mb; 50.9 Mb; 51.8 Mb	

[1] Ag1000G site codes: AOM: Angola [*coluzzii*]; BFM: Burkina Faso [*coluzzii*]; BFS: Burkina Faso [*gambiae*]; CMS: Cameroon [*gambiae*]; GAS: Gabon [*gambiae*]; GNS: Guinea [*gambiae*]; GWA: Guinea-Bissau; UGS: Uganda [*gambiae*]

Table S7: Putative sweeps based on H12 statistic present on islands but rare or absent on LVB mainland.

Chr.	Region Start	Region End	Island Sites with Putative Sweep	Mainland Sites with Putative Sweep	Outlier Island Localities	Outlier Mainland Localities	Ag1000G Populations with Putative Sweep
2R	16,200,000	16,300,000	4 / 5	0 / 4	Banda; Bukasa; Nsadzi; Sserinya	None	Guinea-Bissau
2R	17,300,000	17,500,000	4 / 5	1 / 4	Banda; Bugala (I); Bukasa; Sserinya	Buwama	Guinea-Bissau
2R	21,000,000	21,100,000	5 / 5	1 / 4	Banda; Bugala (I); Bukasa; Nsadzi; Sserinya	Buwama	None
2R	40,400,000	40,800,000	4 / 5	1 / 4	Bugala (I); Bukasa; Nsadzi; Sserinya	Wamala	Burkina Faso [*gambiae*], Cameroon [*gambiae*], Gabon [*gambiae*], Guinea-Bissau
2R	41,100,000	41,200,000	4 / 5	1 / 4	Banda; Bukasa; Nsadzi; Sserinya	Wamala	Cameroon [*gambiae*]
2R	55,800,000	55,900,000	4 / 5	1 / 4	Banda; Bugala (I); Bukasa; Sserinya	Kiyindi	Angola [*coluzzii*]
2L	7,700,000	7,800,000	4 / 5	1 / 4	Banda; Bugala (I); Bukasa; Nsadzi	Buwama	Guinea-Bissau, Uganda [*gambiae*]
2L	8,100,000	8,200,000	4 / 5	0 / 4	Banda; Bugala (I); Nsadzi; Sserinya	None	None
2L	42,400,000	42,500,000	4 / 5	0 / 4	Banda; Bugala (I); Nsadzi; Sserinya	None	None
2L	43,500,000	43,600,000	5 / 5	1 / 4	Banda; Bugala (I); Bukasa; Nsadzi; Sserinya	Buwama	None
2L	49,000,000	49,100,000	4 / 5	1 / 4	Banda; Bugala (I); Bukasa; Sserinya	Wamala	None

3R	26,600,000	26,700,000	5 / 5	0 / 4	Banda; Bugala (I); Bukasa; Nsadzi; Sserinya	None	None
3R	36,700,000	36,800,000	4 / 5	0 / 4	Banda; Bugala (I); Bukasa; Nsadzi	None	None
3R	44,200,000	44,300,000	4 / 5	1 / 4	Banda; Bukasa; Nsadzi; Sserinya	Kiyindi	Angola [*coluzzii*]
3R	46,200,000	46,300,000	4 / 5	0 / 4	Banda; Bukasa; Nsadzi; Sserinya	None	None
X	6,600,000	7,000,000	4 / 5	0 / 4	Banda; Bugala (I); Bukasa; Nsadzi; Sserinya	None	Burkina Faso [*coluzzii*], Gabon [*gambiae*], Guinea-Bissau
X	8,100,000	10,700,000	4 / 5	0 / 4	Banda; Bugala (I); Bukasa; Nsadzi; Sserinya	Kiyindi	Burkina Faso [*coluzzii*], Burkina Faso [*gambiae*], Gabon [*gambiae*]
X	11,300,000	11,800,000	5 / 5	0 / 4	Banda; Bugala (I); Bukasa; Nsadzi; Sserinya	None	Gabon [*gambiae*]
X	12,900,000	13,000,000	4 / 5	0 / 4	Banda; Bugala (I); Bukasa; Sserinya	None	Gabon [*gambiae*]
X	14,300,000	14,400,000	5 / 5	1 / 4	Banda; Bugala (I); Bukasa; Nsadzi; Sserinya	Kaazi	Gabon [*gambiae*]
X	16,200,000	16,300,000	4 / 5	1 / 4	Banda; Bugala (I); Bukasa; Sserinya	Kaazi	Burkina Faso [*coluzzii*], Gabon [*gambiae*]

Table S8: Putative sweeps based on H12 statistic present on LVB mainland but rare or absent on islands.

Chr.	Region Start	Region End	Island Sites with Putative Sweep	Mainland Sites with Putative Sweep	Outlier Island Localities	Outlier Mainland Localities	Ag1000G Populations with Putative Sweep
2R	27,600,000	27,700,000	1 / 5	3 / 4	Nsadzi	Buwama; Kiyindi; Wamala	None
2R	38,000,000	38,100,000	1 / 5	3 / 4	Bugala (I)	Buwama; Kiyindi; Wamala	None
2R	42,700,000	42,800,000	0 / 5	3 / 4	None	Buwama; Kiyindi; Wamala	None
2R	45,400,000	45,500,000	1 / 5	3 / 4	Sserinya	Buwama; Kiyindi; Wamala	None
2R	46,800,000	46,900,000	1 / 5	3 / 4	Banda	Buwama; Kiyindi; Wamala	Cameroon [*gambiae*]
2R	48,000,000	48,100,000	1 / 5	3 / 4	Bukasa	Buwama; Kaazi; Wamala	Angola [*coluzzii*], Cameroon [*gambiae*]
2R	48,800,000	48,900,000	1 / 5	3 / 4	Nsadzi	Buwama; Kaazi; Wamala	None
2R	50,900,000	51,000,000	1 / 5	3 / 4	Bukasa	Kaazi; Kiyindi; Wamala	Burkina Faso [*gambiae*], Guinea [*gambiae*]
2R	51,500,000	51,600,000	0 / 5	3 / 4	None	Kaazi; Kiyindi; Wamala	None
2R	57,500,000	57,600,000	1 / 5	3 / 4	Banda	Buwama; Kaazi; Kiyindi	Angola [*coluzzii*], Guinea-Bissau
2L	2,900,000	3,000,000	1 / 5	4 / 4	Sserinya	Buwama; Kaazi; Kiyindi; Wamala	Angola [*coluzzii*], Burkina Faso [*coluzzii*], Burkina Faso [*gambiae*], Cameroon [*gambiae*], Gabon [*gambiae*], Guinea [*gambiae*], Uganda [*gambiae*]
2L	4,200,000	4,300,000	1 / 5	4 / 4	Bugala (I)	Buwama; Kaazi; Kiyindi; Wamala	Angola [*coluzzii*], Cameroon [*gambiae*], Gabon [*gambiae*], Uganda [*gambiae*]

2L	5,700,000	5,800,000	1 / 5	3 / 4	Bugala (I)	Buwama; Kaazi; Kiyindi	Angola [*coluzzii*], Guinea [*gambiae*], Uganda [*gambiae*]
2L	6,200,000	6,300,000	1 / 5	3 / 4	Bugala (I)	Kaazi; Kiyindi; Wamala	Uganda [*gambiae*]
2L	6,600,000	6,800,000	1 / 5	3 / 4	Bugala (I)	Kaazi; Kiyindi; Wamala	Angola [*coluzzii*], Cameroon [*gambiae*], Gabon [*gambiae*], Guinea [*gambiae*], Uganda [*gambiae*]
2L	10,000,000	10,100,000	1 / 5	3 / 4	Sserinya	Kaazi; Kiyindi; Wamala	None
2L	10,800,000	10,900,000	0 / 5	3 / 4	None	Kaazi; Kiyindi; Wamala	None
2L	11,300,000	11,400,000	1 / 5	3 / 4	Bugala (I)	Kaazi; Kiyindi; Wamala	Guinea-Bissau
2L	12,000,000	12,100,000	1 / 5	3 / 4	Bugala (I)	Kaazi; Kiyindi; Wamala	None
2L	12,400,000	13,000,000	0 / 5	3 / 4	Bukasa	Buwama; Kaazi; Kiyindi; Wamala	None
2L	14,500,000	14,900,000	1 / 5	3 / 4	Sserinya	Buwama; Kiyindi; Wamala	Gabon [*gambiae*], Uganda [*gambiae*]
2L	16,000,000	16,300,000	1 / 5	3 / 4	Bukasa	Buwama; Kaazi; Wamala	Gabon [*gambiae*]
2L	16,600,000	16,700,000	1 / 5	4 / 4	Bugala (I)	Buwama; Kaazi; Kiyindi; Wamala	None
2L	18,700,000	18,800,000	1 / 5	3 / 4	Nsadzi	Kaazi; Kiyindi; Wamala	None
2L	33,600,000	33,700,000	1 / 5	3 / 4	Bugala (I)	Buwama; Kaazi; Kiyindi	Angola [*coluzzii*]
2L	34,400,000	34,500,000	1 / 5	3 / 4	Sserinya	Buwama; Kaazi; Wamala	None

3R	28,500,000	28,700,000	1 / 5	4 / 4	Sserinya	Buwama; Kaazi; Kiyindi; Wamala	Burkina Faso [*coluzzii*], Burkina Faso [*gambiae*], Cameroon [*gambiae*], Gabon [*gambiae*], Guinea [*gambiae*], Uganda [*gambiae*]
3R	36,500,000	36,900,000	0 / 5	3 / 4	Nsadzi	Buwama; Kiyindi; Wamala	None
3R	43,000,000	43,100,000	0 / 5	3 / 4	None	Buwama; Kiyindi; Wamala	None
3R	43,700,000	44,100,000	0 / 5	3 / 4	Nsadzi	Buwama; Kiyindi; Wamala	Angola [*coluzzii*], Burkina Faso [*gambiae*], Guinea [*gambiae*], Uganda [*gambiae*]
3R	48,800,000	48,900,000	0 / 5	3 / 4	None	Buwama; Kiyindi; Wamala	None
3R	50,000,000	50,100,000	1 / 5	3 / 4	Sserinya	Kaazi; Kiyindi; Wamala	None
3L	7,000,000	7,100,000	1 / 5	4 / 4	Sserinya	Buwama; Kaazi; Kiyindi; Wamala	None
3L	11,500,000	11,600,000	1 / 5	3 / 4	Sserinya	Buwama; Kiyindi; Wamala	Burkina Faso [*coluzzii*]
3L	12,200,000	12,300,000	0 / 5	3 / 4	None	Kaazi; Kiyindi; Wamala	None
3L	13,400,000	13,500,000	0 / 5	3 / 4	None	Kaazi; Kiyindi; Wamala	None
3L	16,300,000	16,400,000	1 / 5	3 / 4	Sserinya	Buwama; Kiyindi; Wamala	Uganda [*gambiae*]

Table S9: Signatures of selective sweeps on known insecticide genes by site based on H12 statistic.

Chr.	Location	Insecticide Gene	Island Sites with Putative Sweep	Mainland Sites with Putative Sweep	Outlier Island Localities	Outlier Mainland Localities	Ag1000G Populations with Putative Sweep
2R	28,497,407	Cyp6p	5 / 5	4 / 4	Banda; Bugala (I); Bukasa; Nsadzi; Sserinya	Buwama; Kaazi; Kiyindi; Wamala	Angola [*coluzzii*], Burkina Faso [*coluzzii*], Burkina Faso [*gambiae*], Cameroon [*gambiae*], Guinea [*gambiae*], Uganda [*gambiae*]
3R	28,598,038	Gste	1 / 5	4 / 4	Sserinya	Buwama; Kaazi; Kiyindi; Wamala	Burkina Faso [*coluzzii*], Burkina Faso [*gambiae*], Cameroon [*gambiae*], Gabon [*gambiae*], Guinea [*gambiae*], Uganda [*gambiae*]
X	15,241,718	Cyp9k1	3 / 5	4 / 4	Banda; Bukasa; Sserinya	Buwama; Kaazi; Kiyindi; Wamala	Burkina Faso [*coluzzii*], Burkina Faso [*gambiae*], Gabon [*gambiae*], Guinea [*gambiae*]

Table S10: Software and versions used for major parts of analysis.

Software	Version	Citation
ea-utils	-	[64]
BWA	0.7.16a	[67]
GATK	3.8	[68]
PLINK	1.90b4.6	[69, 70]
SHAPEIT2	2.837	[71]
SAMtools/BCFtools	1.5	[72, 90]
ADMIXTURE	1.3.0	[20]
CLUMPAK	-	[73]
VCFtools	0.1.15	[86]
$\delta a \delta i$ (python package)	1.7.0	[82, 83]
Stairway plot - Jpopgen	2-beta	[21]
selscan	1.2.0a	[87]
adegenet (R package)	2.1.0	[91]
ape (R package)	5.0	[92]
RColorBrewer (R package)	1.1-2	[93]
dendextend (R package)	1.6.0	[94]
rehh (R package)	2.0.2	[95]
eigensoft	7.2.1	[75, 76]
GNU parallel	20170422	[96]
tabix	1.5	[90]
bedtools	2.26.0	[97]

Table S11: Genomic coordinates of heterochromatic and inverted regions.

Chromosome arm	Start	End	Information
2L	20,524,058	42,165,532	2La inversion [66]
2R	18,575,300	26,767,588	2Rb inversion [66]
2L	1	2,431,617	Heterochromatic region [66]
2L	5,078,962	5,788,875	Heterochromatic region [66]
2R	58,984,778	61,545,105	Heterochromatic region [66]
3L	1	1,815,119	Heterochromatic region [66]
3L	4,264,713	5,031,692	Heterochromatic region [66]
3R	38,988,757	41,860,198	Heterochromatic region [66]
3R	52,161,877	53,200,684	Heterochromatic region [66]
X	20,009,764	24,393,108	Heterochromatic region [66]

Figures

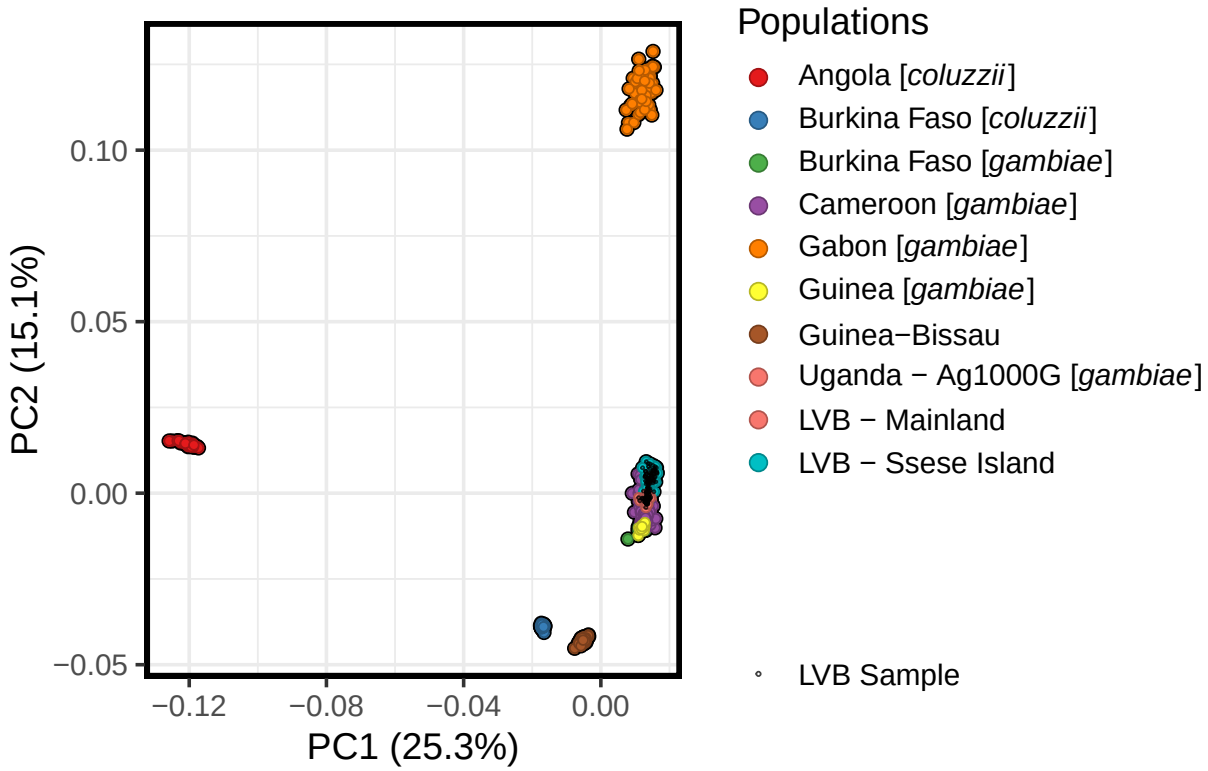

Figure S1: PCA plot of study individuals and *A. gambiae* and *A. coluzzii* individuals from reference Ag1000G populations, showing the first and second components. Kenyan population is not included, and analysis is based on chromosome 3.

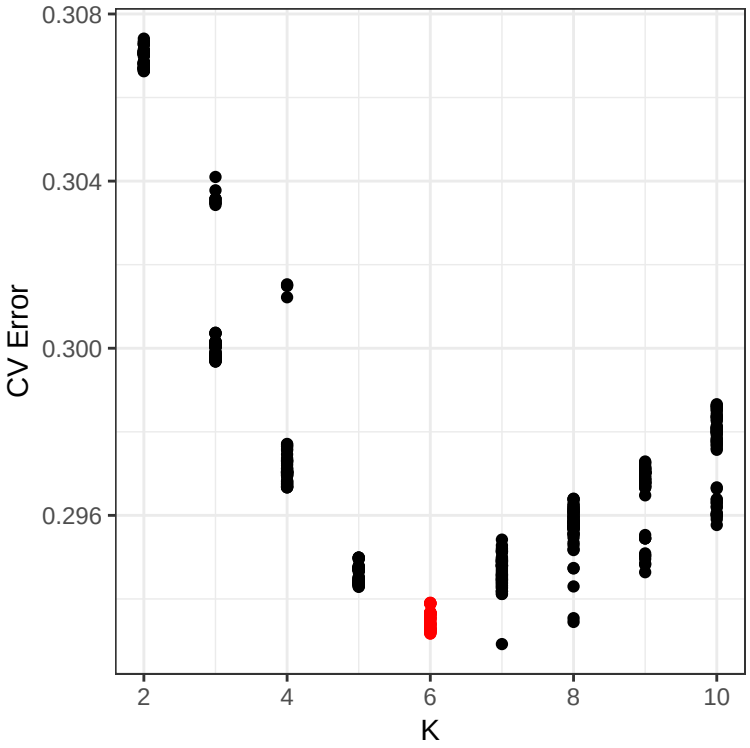

Figure S2: ADMIXTURE cross-validation error.
Cross-validation error for range of k values for ADMIXTURE analysis of Lake Victoria Basin individuals and *A. gambiae* and *A. coluzzii* Ag1000G reference populations.

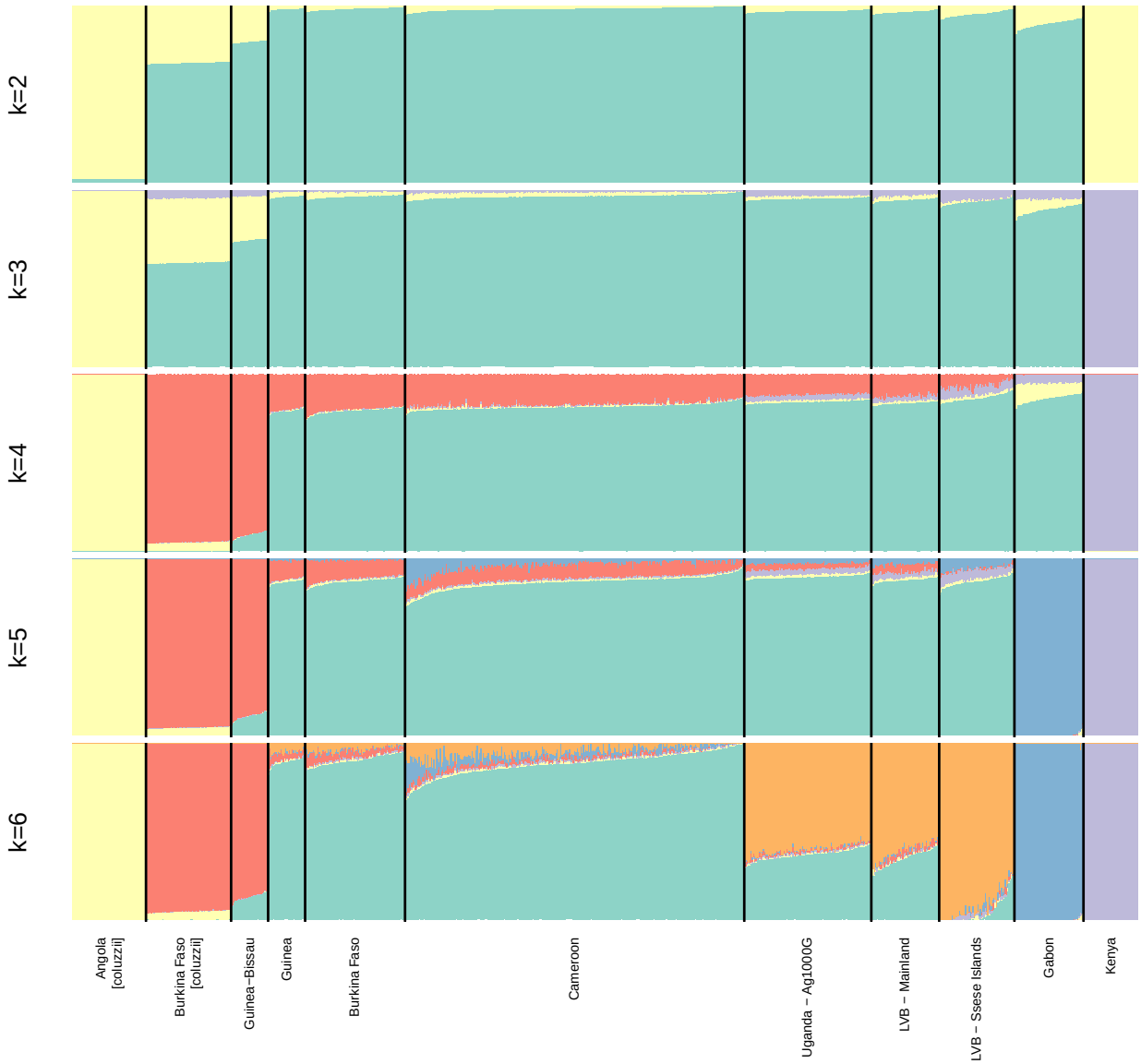

Figure S3: (Caption on next page.)

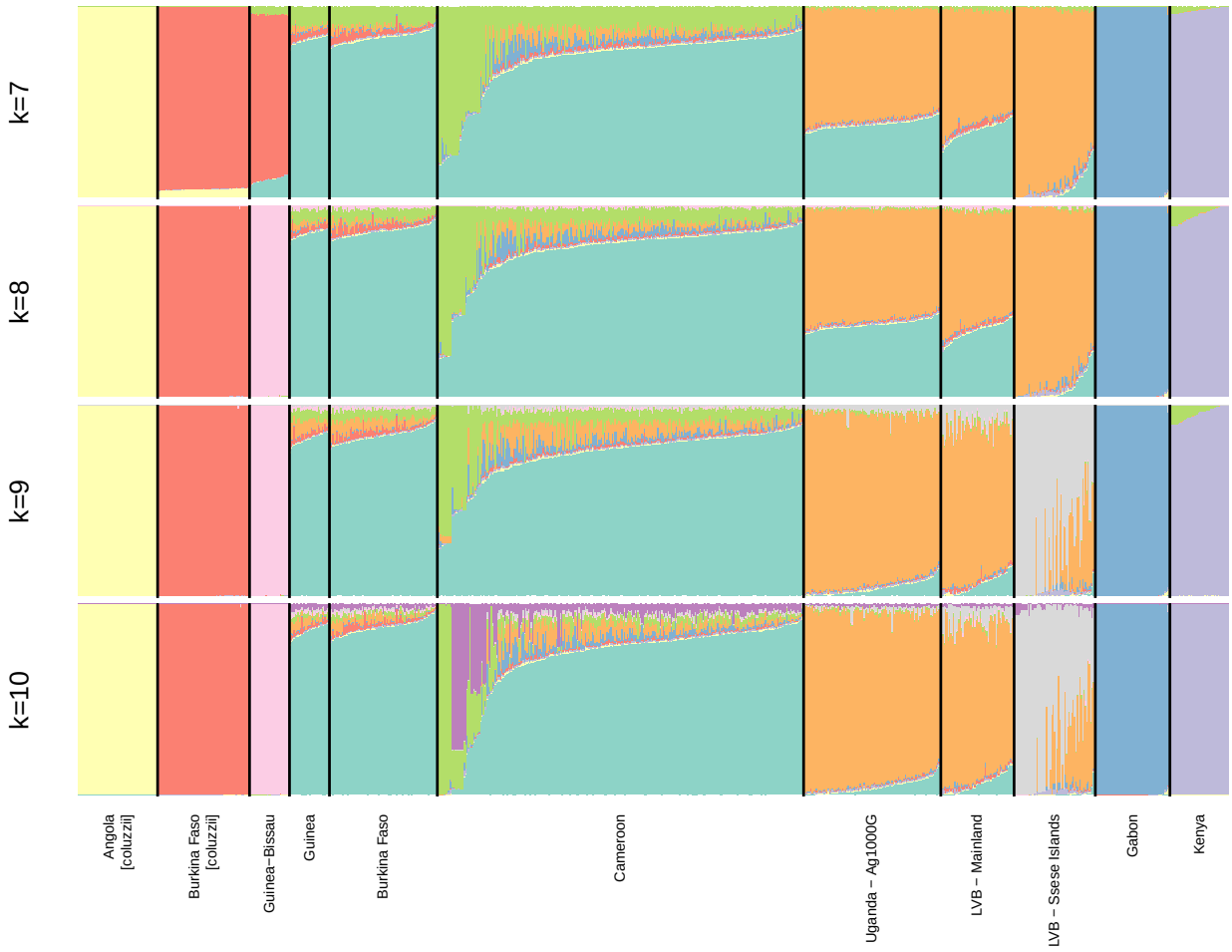

Figure S3: ADMIXTURE-inferred ancestry.
Ancestry of individuals in Lake Victoria Basin and of Ag1000G reference populations as inferred by ADMIXTURE clustering method. Samples are *A. gambiae* unless noted, and analysis is based on chromosome 3. Using $k = 6$ clusters minimizes cross validation error (Fig. S2).

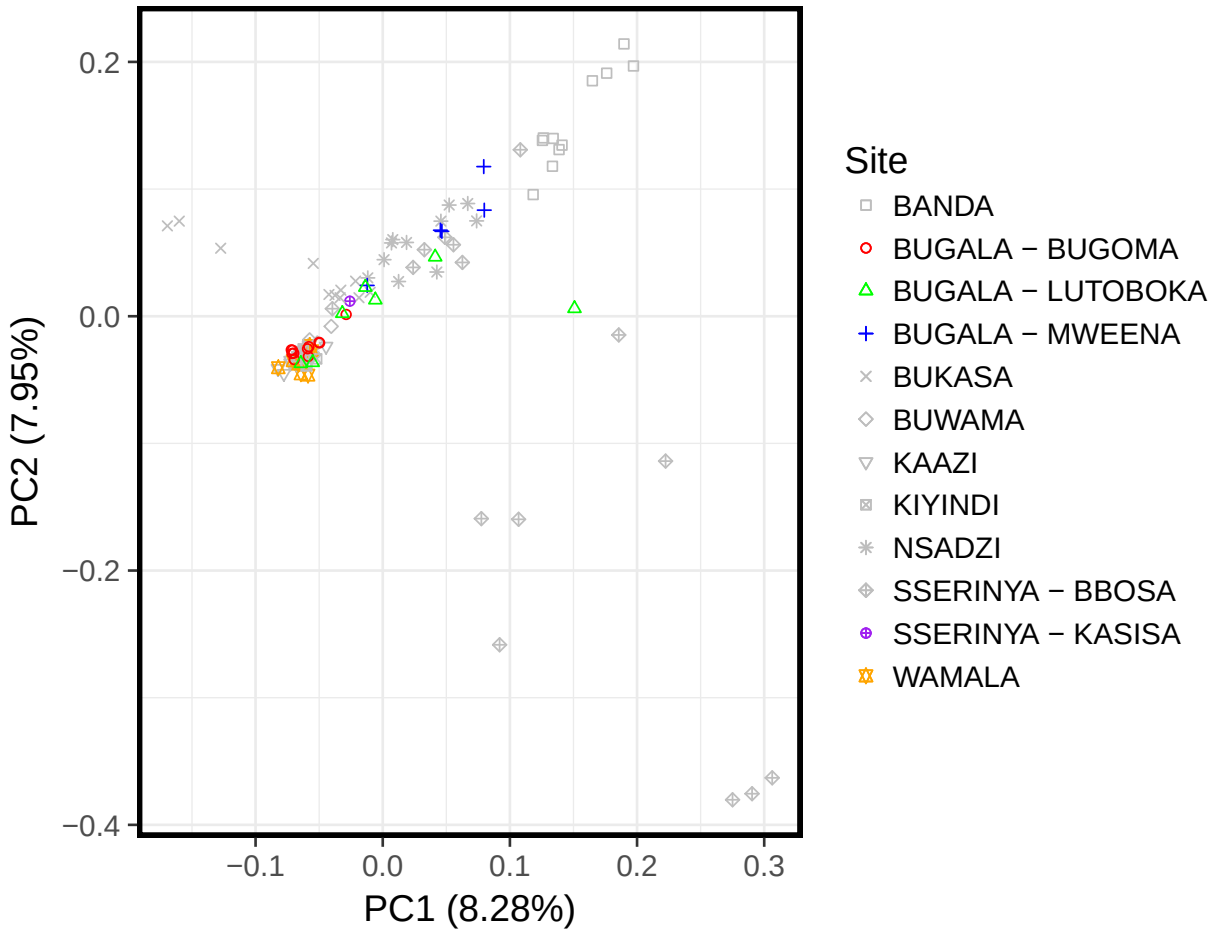

Figure S4: PCA showing Bugala subdivision.
PCA colored by sampling locations. Based on this analysis, individuals from Bugala were split into mainland- and island-like subpopulations. Samples from Sserinya Island, though sampled from two localities, were not split.

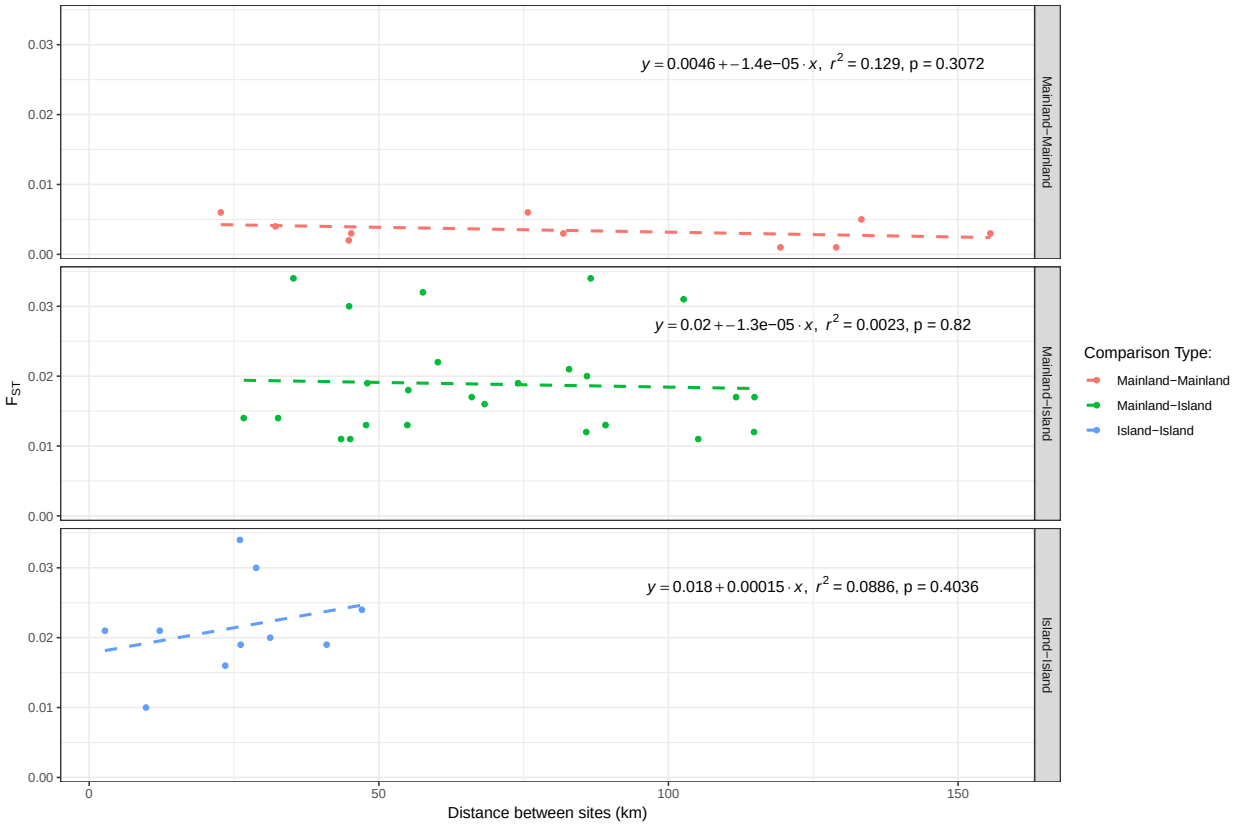

Figure S5: Correlations between genetic distance (F_{ST}) and geographic distance between localities. The p-values are for the test that the slope is significantly different from zero.

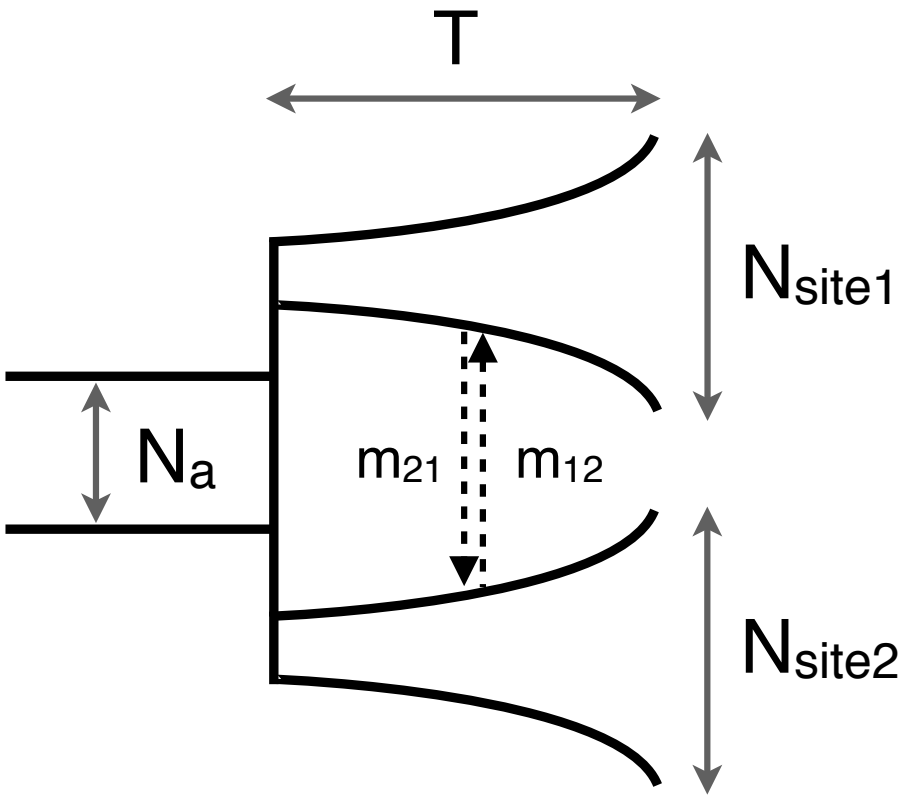

Figure S6: IM model schematics.
Schematic of model fit to data with $\delta a \delta i$ for population history inference between all pairs of sampled sites using IM model.

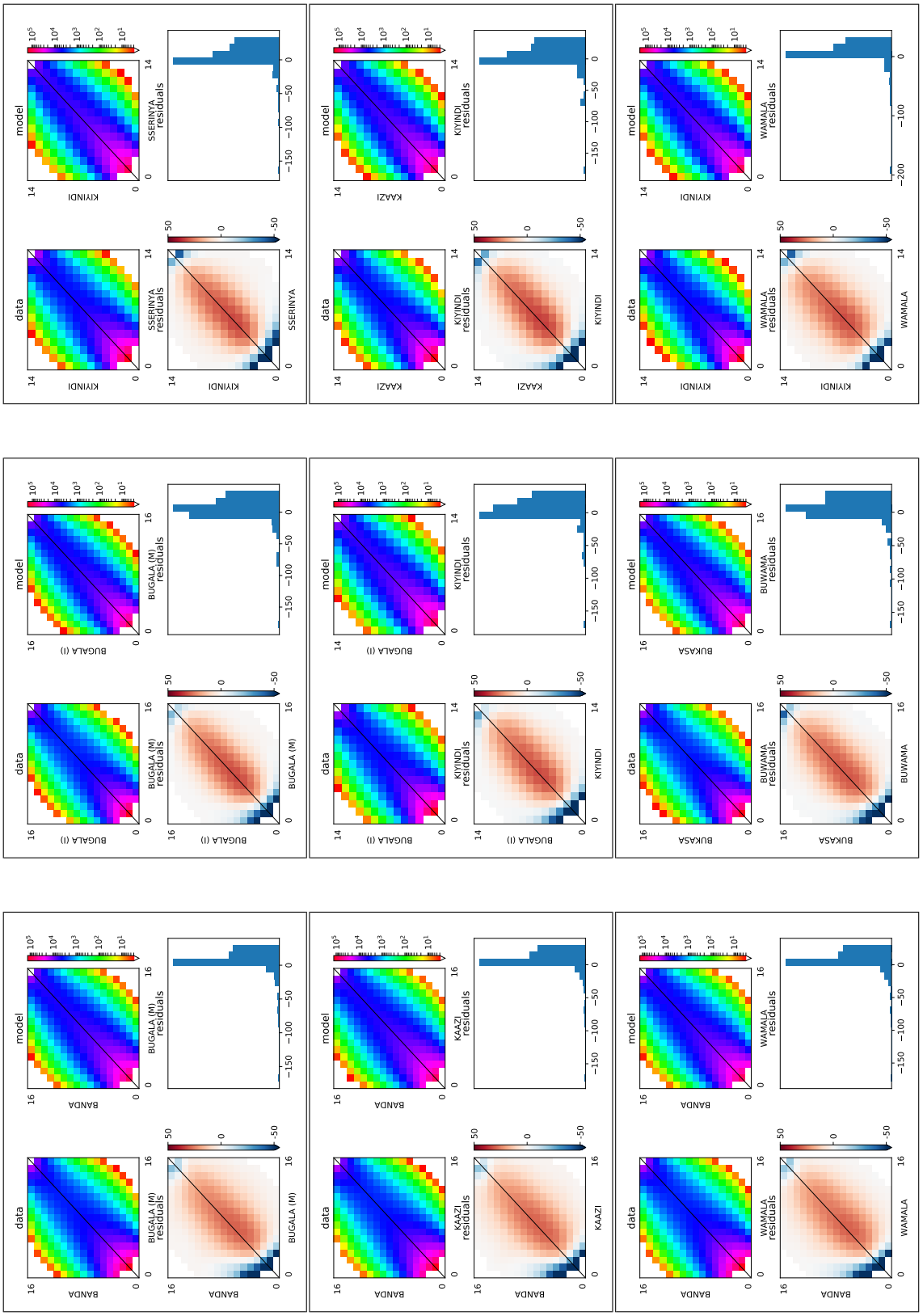

Figure S7: (Caption on next page.)

Figure S7: Two population $\delta a \delta i$ optimization results.
Comparison between best fitting model and data frequency spectra for two population $\delta a \delta i$ inference. Of the pairwise comparisons for which the best model included migration, a randomly selected set of nine are shown here. Two-dimensional frequency spectra are plotted as logarithmic colormaps for the data (upper left) and model (upper right), and the bottom row plots show the residuals between model and data. Positive residuals in red indicate the model predicts too many SNPs in that entry while negative residuals in blue indicate the model predicts too few.

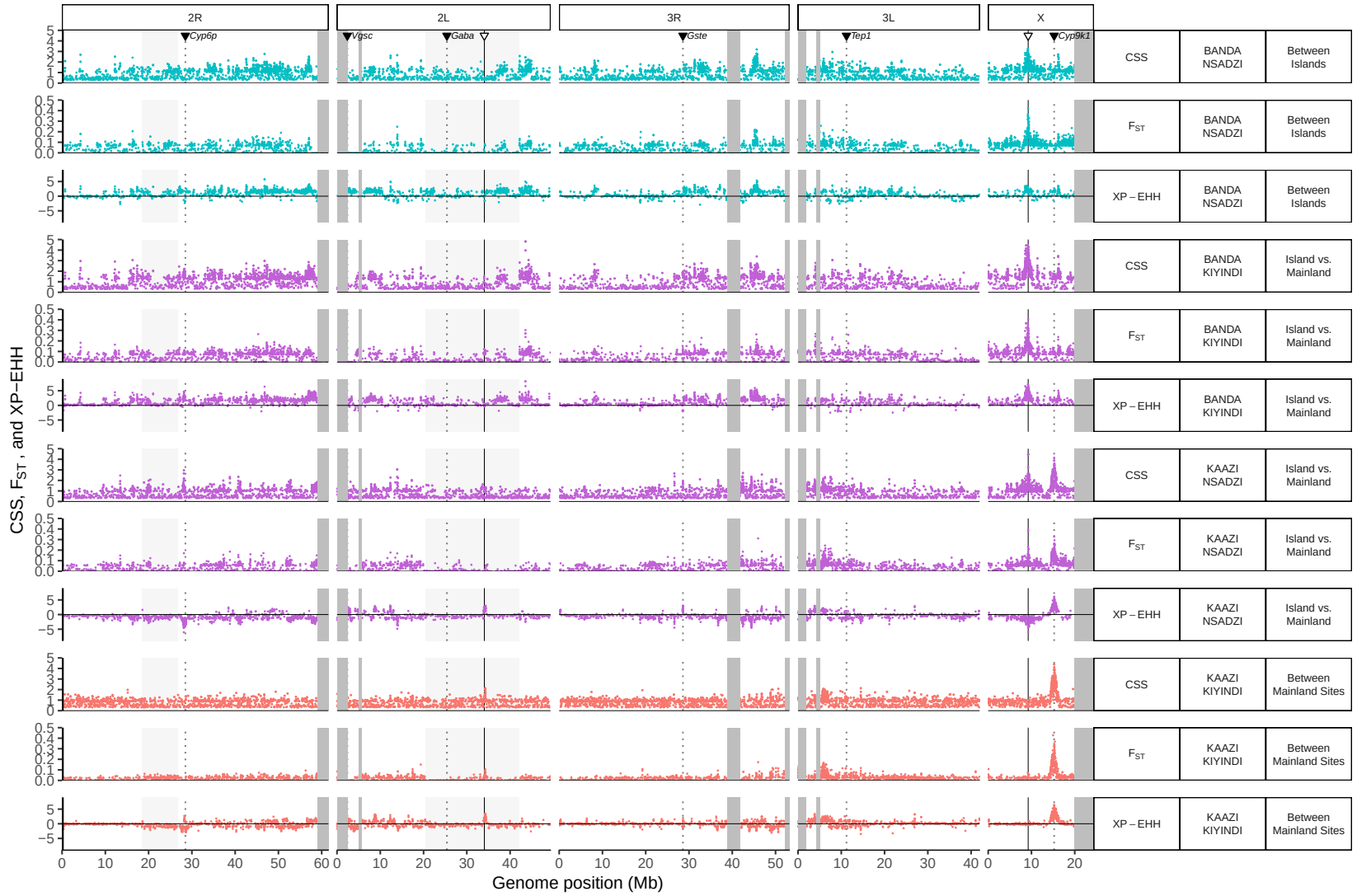

Figure S8: F_{ST}, XP-EHH, and Composite Selection Score (CSS) across genome.
F_{ST}, XP-EHH, and CSS averaged in windows of size 10 kb plotted across genome for pairwise comparisons of island and mainland localities. Shaded regions indicate inversions or heterochromatic regions (excluded from analysis) and dotted lines indicate known insecticide genes while solid lines indicate the two putative sweeps identified in the present study. Only several exemplar pairs of populations shown.

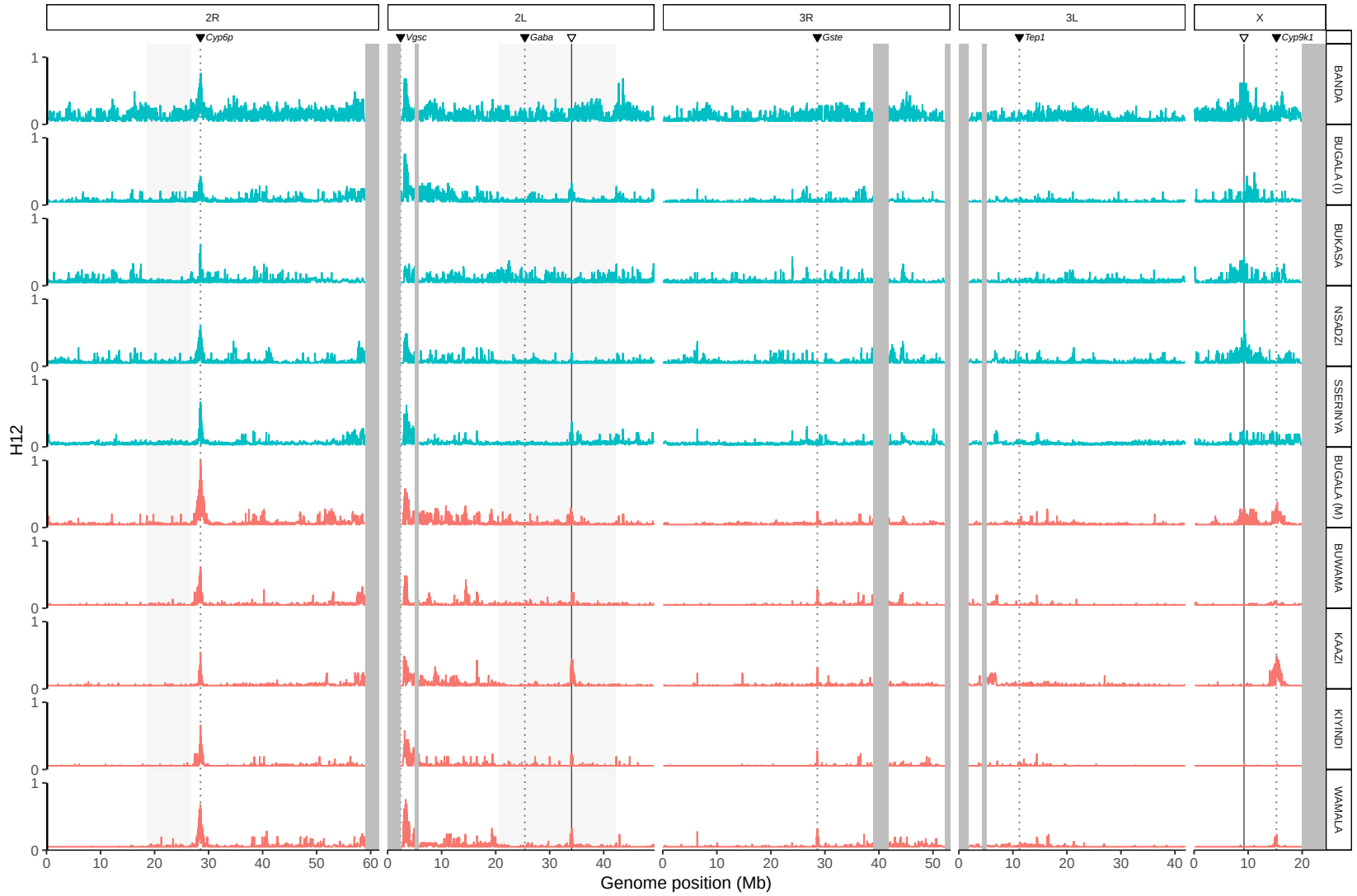

Figure S9: H12 across genome.

Values of H12, a measure of haplotype homozygosity, plotted across genome. Shaded regions indicate inversions or heterochromatic regions (excluded from analysis) and dotted lines indicate known insecticide genes while solid lines indicate the two putative sweeps identified in the present study.

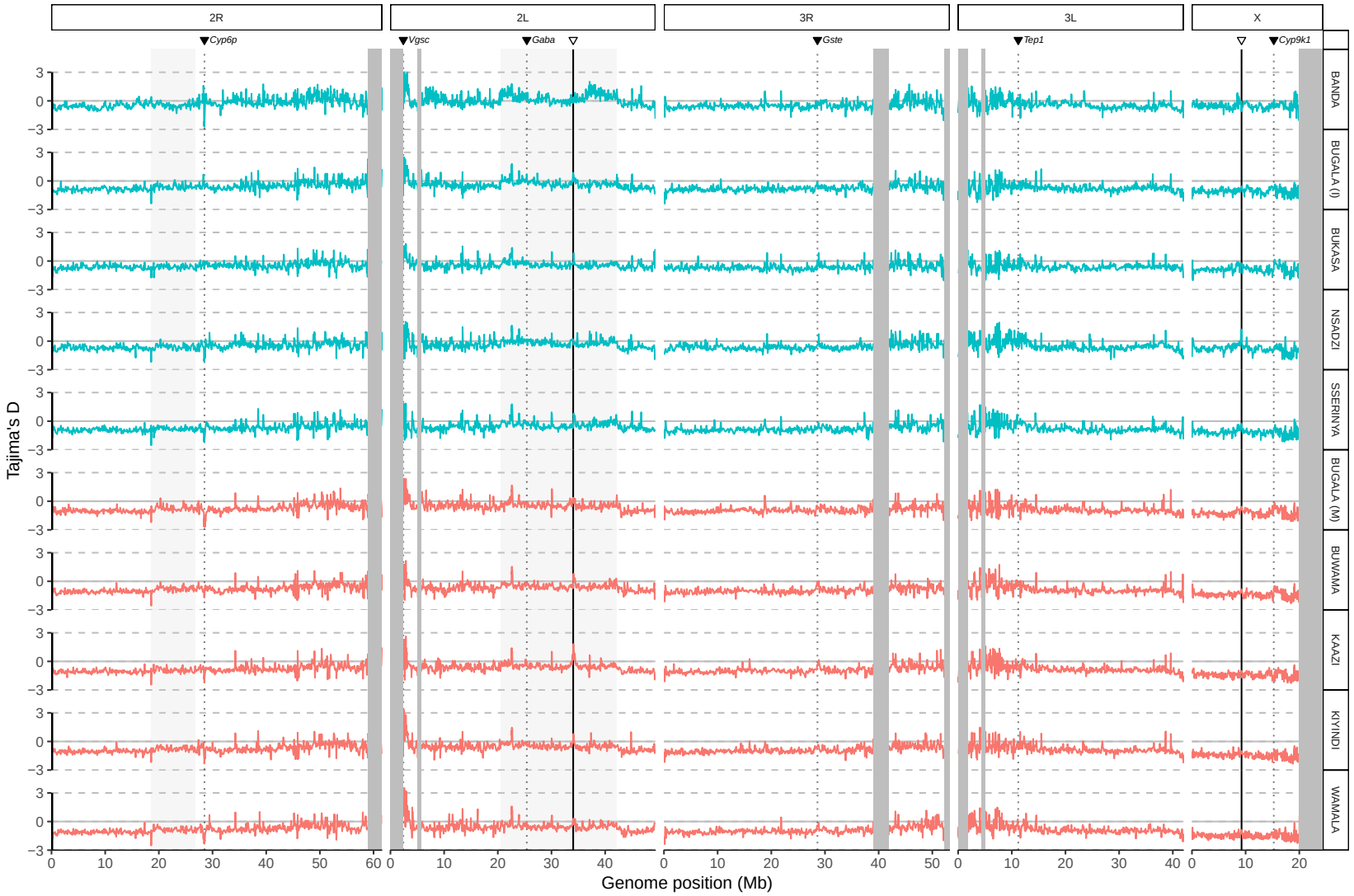

Figure S10: Tajima's D across genome.

Tajima's D plotted across genome. Shaded regions indicate inversions or heterochromatic regions (excluded from analysis) and dotted lines indicate known insecticide genes while solid lines indicate the two putative sweeps identified in the present study.

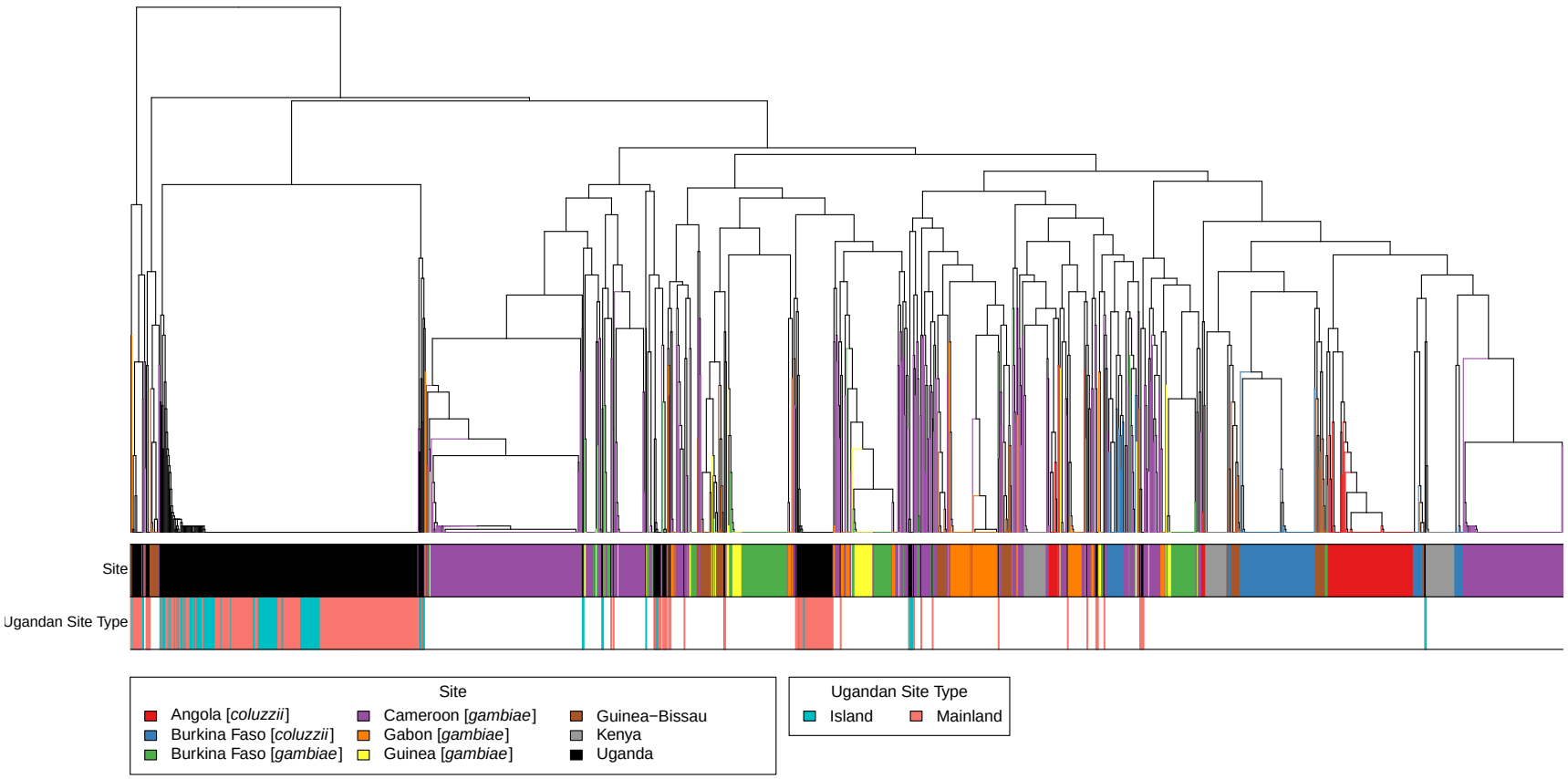

Figure S11: Tree for sweep on *Cyp6p* gene cluster on chromosome arm 2R.

Distance-based tree of haplotypes near sweep at *Cyp6p* gene cluster on chromosome arm 2R. Region shown is 10 kb up- and downstream of sweep target, centered at chr2R:28,501,972 (the approximate location of the peaks in pairwise statistics). Top color bar indicates locality, with all Ugandan individuals, from both the Ag1000G reference population and the LVB, in black. The bottom color bar differentiates the Ugandan individuals into mainland (red) and island (blue) individuals.

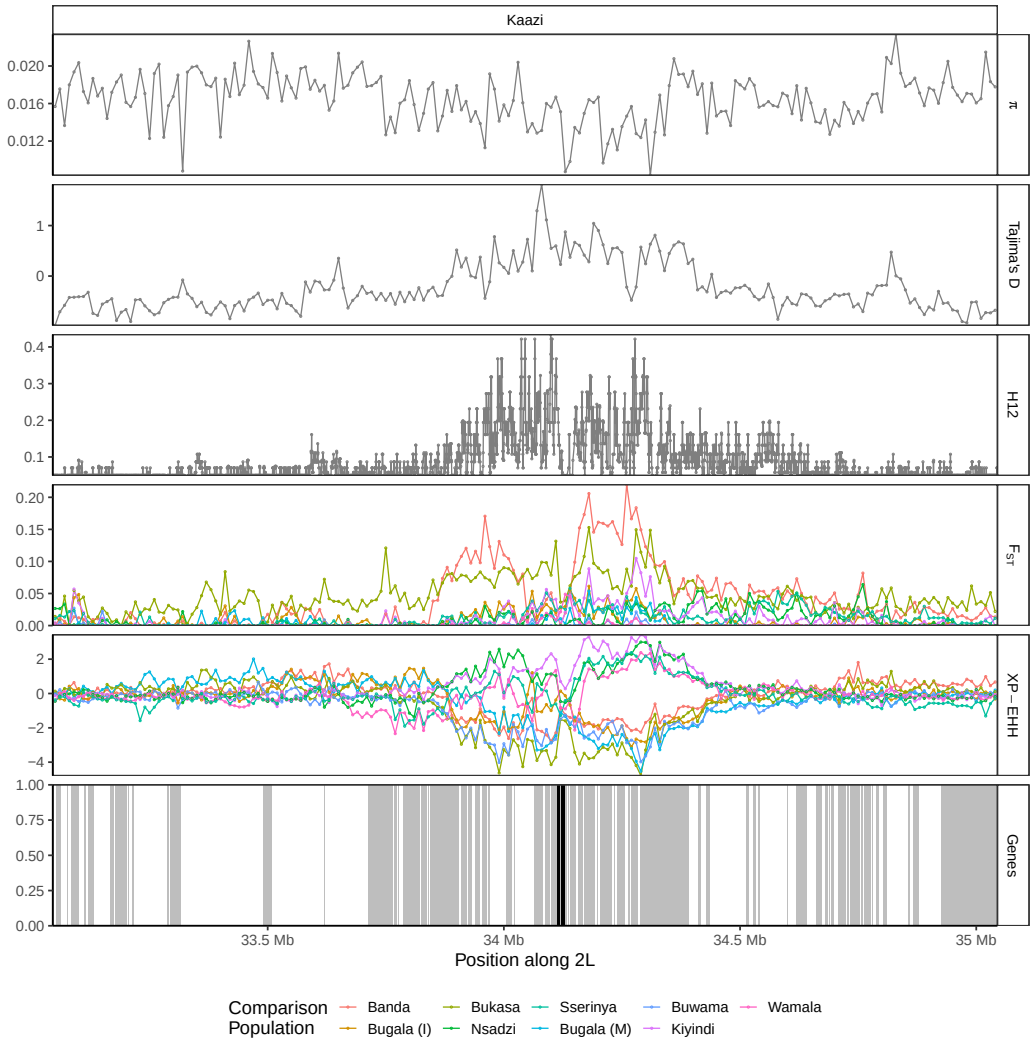

Figure S12: Selective sweep signal on chromosome 2L.
Population genetic statistics plotted near putative sweep on chromosome 2L. Focus population for all pairwise F_{ST} and XP-EHH comparisons is mainland site Kaazi, chosen to maximize peak height in these statistics. Region shown is 1 Mb up- and downstream of sweep target, centered at chr2L:34,044,820. Several genes involved in chorion formation (AGAP006549, AGAP006550, AGAP006551, AGAP006553, AGAP006554, AGAP006555 and AGAP006556) are highlighted with black vertical lines, while other genes are indicated with gray vertical lines.

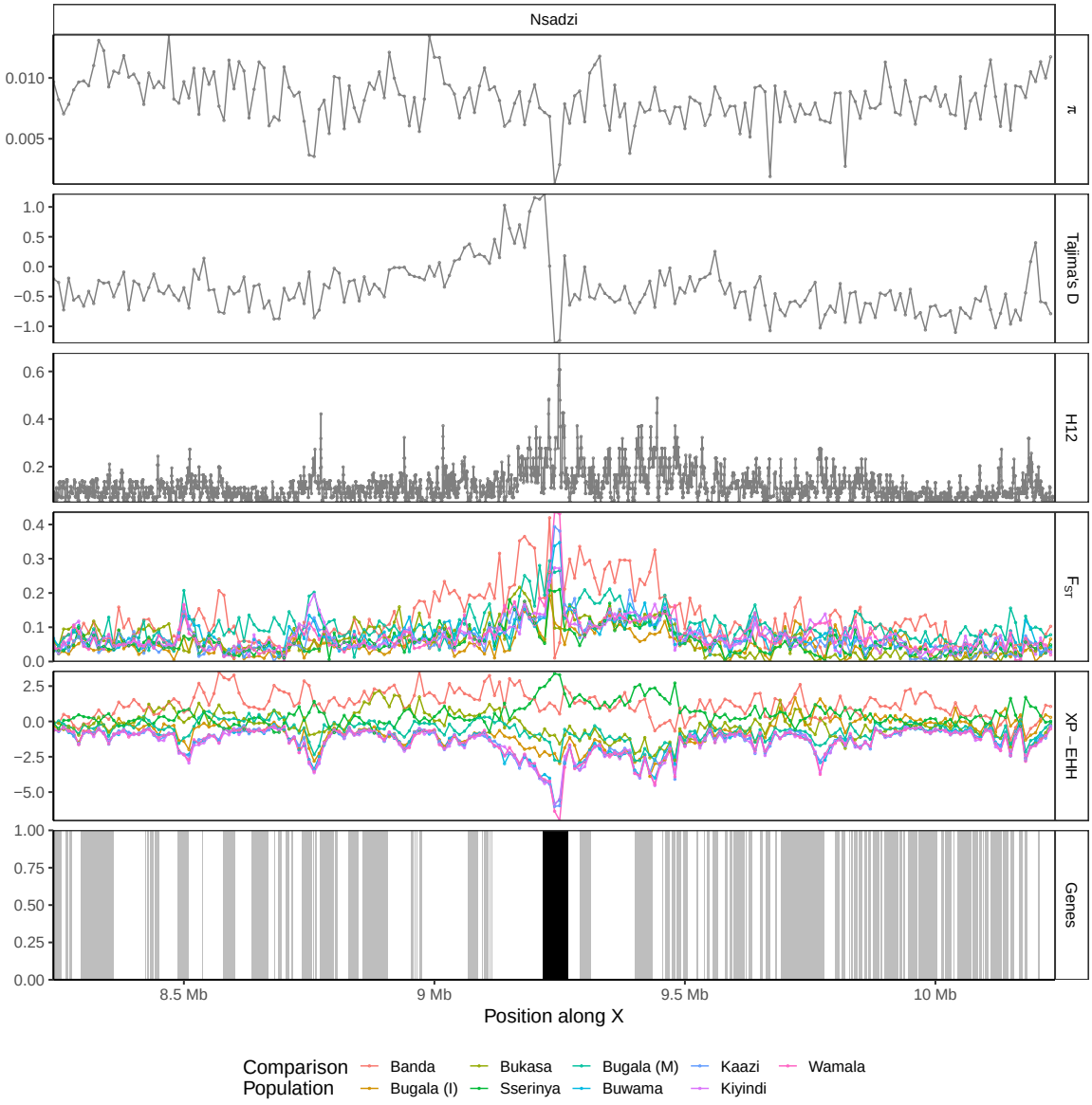

Figure S13: (Caption on next page.)

Figure S13: Selective sweep signal on X-chromosome near *rdgA* ortholog. Population genetic statistics plotted near putative sweep on X-chromosome. Focus population for all pairwise F_{ST} and XP-EHH comparisons is island site Nsadzi, chosen to maximize peak height in these statistics. Region shown is 1 Mb up- and downstream of sweep target, centered at chrX:9,238,942 (the approximate peak in pairwise statistics). The gene eye-specific diacylglycerol kinase (AGAP000519, chrX:9,215,505-9,266,532) is highlighted with a black vertical line, while other genes are indicated with gray vertical lines.

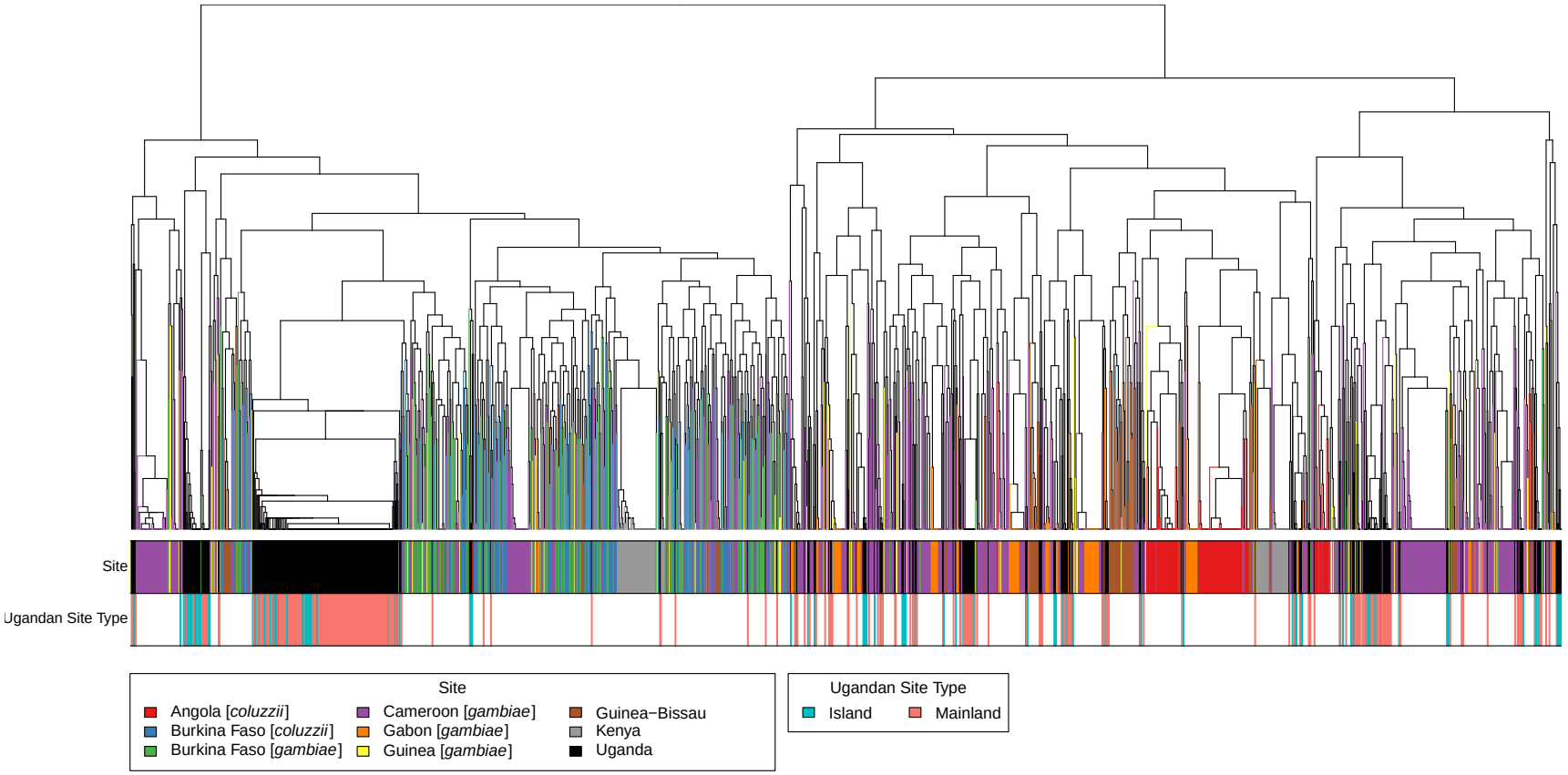

Figure S14: Tree for putative sweep on chromosome 2L.

Distance-based tree of haplotypes near putative sweep on chromosome 2L. Region shown is 10 kb up- and downstream of sweep target, centered at chr2L:34,044,820 (the approximate location of the peaks in pairwise statistics). Top color bar indicates locality, with all Ugandan individuals, from both the Ag1000G reference population and the LVB, in black. The bottom color bar differentiates the Ugandan individuals into mainland (red) and island (blue) individuals.

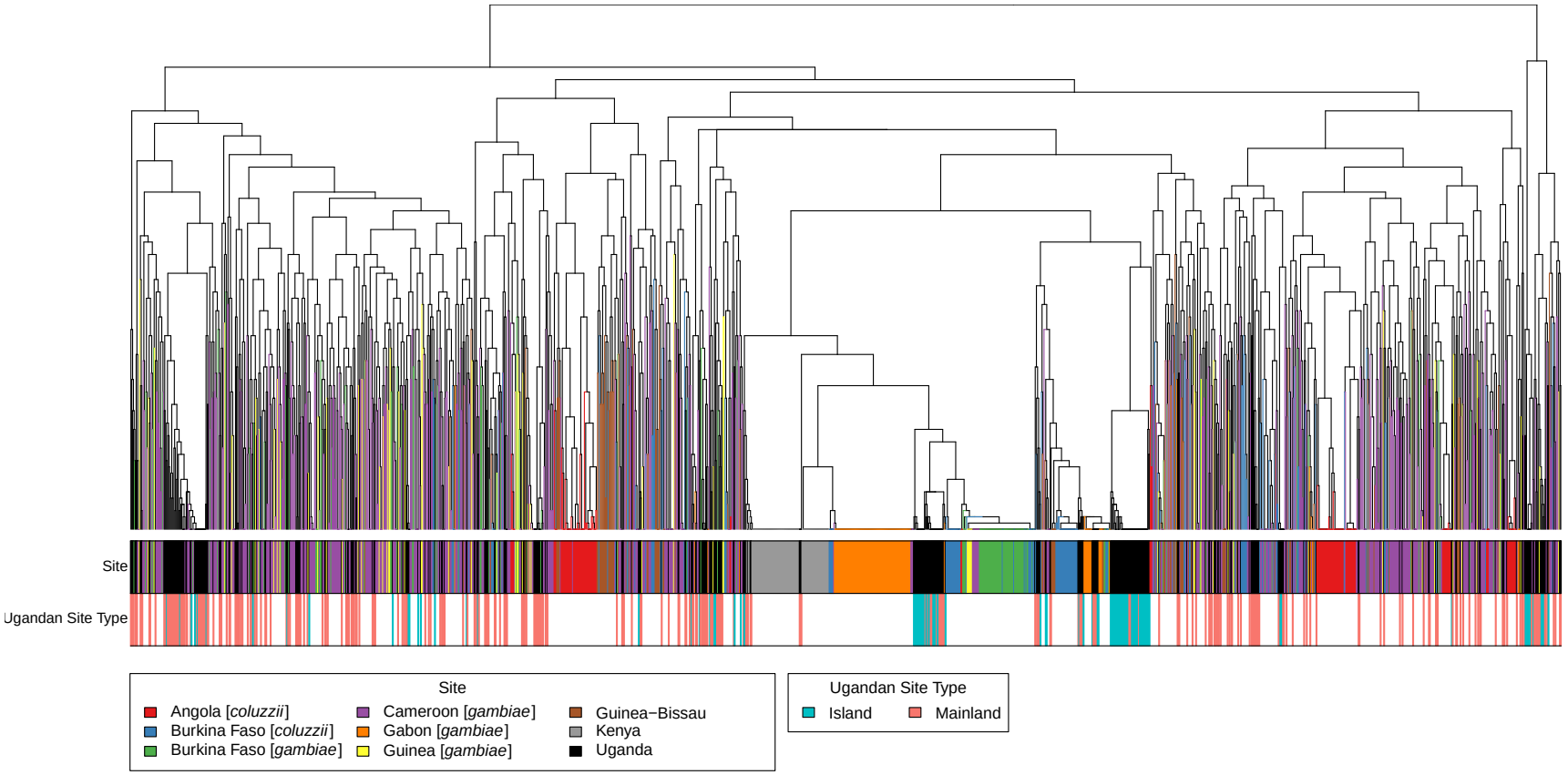

Figure S15: Tree for putative sweep on X-chromosome near *rdgA* ortholog.
Distance-based tree of haplotypes near putative sweep on X-chromosome. Region shown is 10 kb up- and downstream of sweep target, centered at chrX:9,238,942 (the approximate location of the peaks in pairwise statistics). Top color bar indicates locality, with all Ugandan individuals, from both the Ag1000G reference population and the LVB, in black. The bottom color bar differentiates the Ugandan individuals into mainland (red) and island (blue) individuals.