Assessing connectivity despite high diversity in island populations of the malaria mosquito *Anopheles gambiae*

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.*

October 3, 2018

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

Abstract:

Modeling and cage experiments suggest that mosquito gene drive systems will enable malaria eradication, but establishing safety and efficacy requires field-testing in isolated populations. Documenting genetic isolation is notoriously difficult for species with vast polymorphic populations like the principal African malaria vector *Anopheles gambiae*. Using genome-wide variation, we assess Lake Victoria islands as candidate field-testing sites. One island, 30 kilometers offshore, is as differentiated from mainland samples as populations from across the continent, and we confirm isolation using adaptive variation as a powerful assay of connectivity. Collectively, our results suggest sufficient contemporary isolation of these islands to warrant consideration as field-testing locations.

In efforts to reduce the approximately 445,000 annual deaths attributable to malaria [1], conventional vector control techniques may soon be augmented with releases 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 [2–4], and such systems have recently been successfully engineered in the laboratory for the major African malaria vector *Anopheles gambiae sensu stricto* (henceforth *An. gambiae*) [5]. Effective implementation of any genetic control method, not least a gene drive system designed to spread in a super-Mendelian fashion, will benefit from a detailed understanding of population structure and connectivity.

Several life history characteristics of *An. gambiae* complicate efforts to estimate connectivity from genetic data, however. High fecundity and dispersal potential result in large interconnected populations exhibiting shallow population structure [6, 7], manifested in the genome as high levels of polymorphism [6] that impede accurate estimation of connectivity [8]

and discernment of demographic independence from panmixia [9]. Population genetic methods for estimating migration between *An. gambiae* populations using neutral markers may 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 [10]: Adaptive introgression of beneficial haplotypes indicates migration occurred, while the absence of a selective sweep signature that is otherwise widespread would suggest barriers to gene flow.

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 evolutionary insight, investigations of island population structure have practical rationales. Geographically-isolated islands have been proposed as initial field sites to test the dynamics of transgene spread while limiting their movement beyond the study population [12–15]. Antecedent studies of population structure and connectivity of potential release sites are necessary 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.

We analyzed genome-wide variation in *An. gambiae* mosquitoes living near and on the Ssese archipelago of Lake Victoria in Uganda (Fig. 1) to understand the determinants of their genetic variation, recent and long-term connectivity and demographic history, and the spread of adaptive variants across the region. In addition to the high malaria prevalence of the islands (44% in children; 30% in children country-wide; [16]), we were motivated by the potential of such an island to be a field site for future tests of gene-drive vector control strategies.

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, 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; [6]) 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 [6] 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; Fig. 2d) 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, 2c, and Supplementary Fig. S1).

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. 2e). 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. S3). 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 small island Banda (Fig. 2f). Geographic distances and F_{ST} were uncorrelated (Mantel $p = 0.88$; $R^2 = 0.08$, $p = 0.048$; Supplementary Fig. S4). Island samples showed greater within- and between-locality shar-

ing of genomic regions identical by descent (IBD), with sharing between nearby islands Sserinya, Banda and Bugala (Fig. 2g). 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 lower nucleotide diversity (π), slightly higher proportion of shared to rare variants (Tajima's D), more variance in inbreeding coefficient (F), more linkage among SNPs (LD; r^2), longer runs of homozygosity (F_{ROH}), and longer IBD tracks (Fig. 3). 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 per-locality sample sizes, 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.

To all pairs of LVB localities we fit an isolation-with-migration (IM) demographic model using $\delta a\delta i$, in which an ancestral population splits into two populations, allowing exponential growth and continuous asymmetrical migration between the daughter populations (Supple-

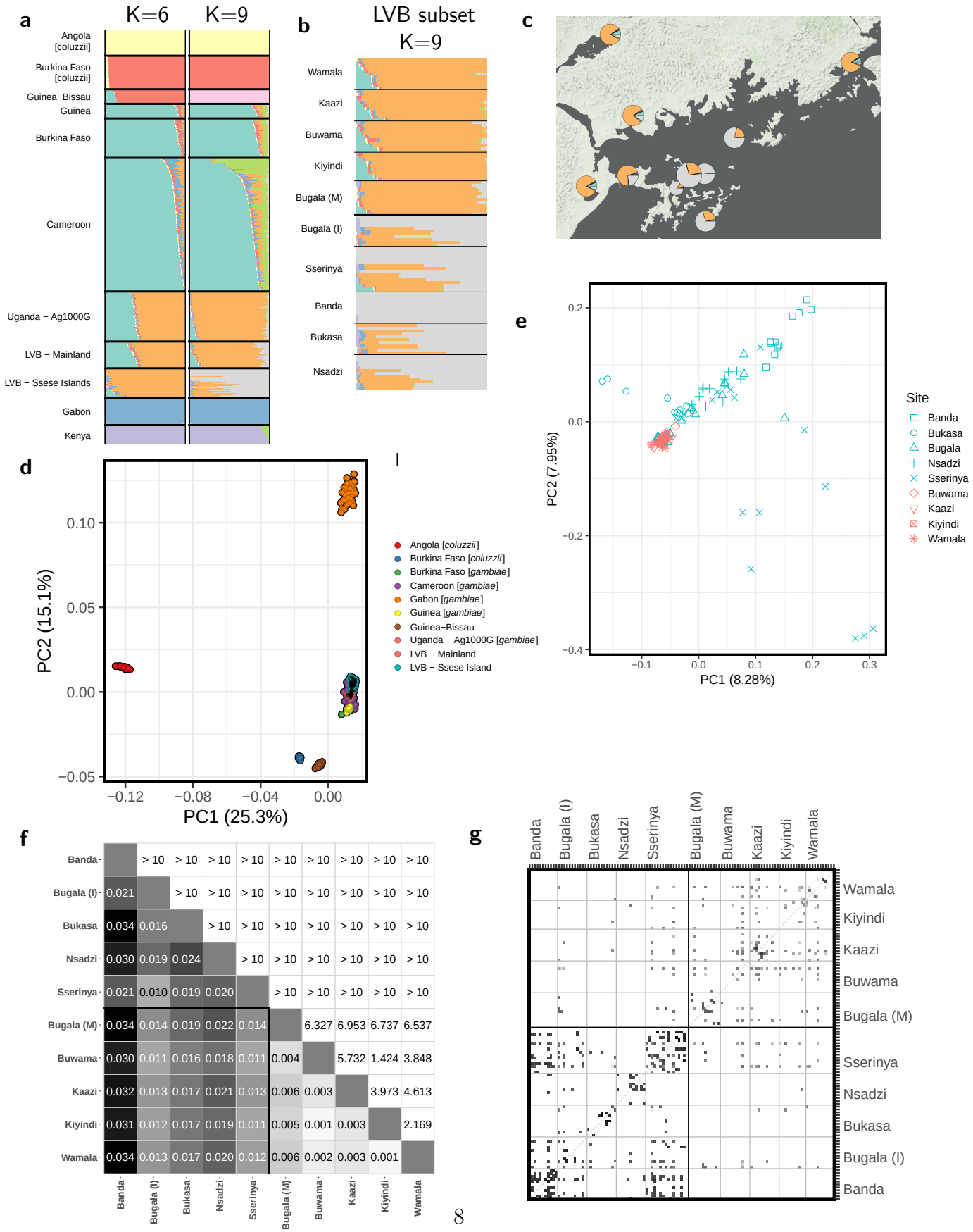

Figure 2: (Caption on next page.)

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, at 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) Ancestry of individuals in Lake Victoria Basin and of Ag1000G reference populations as inferred by clustering into $k = 9$ clusters. Samples are *A. gambiae* unless noted. (D) PCA plot of study individuals and *A. gambiae* and *A. coluzzii* individuals from reference Ag1000G populations. (E) 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, the island sample of Bugala was split into mainland- and island-like subpopulations ("Bugala (M)" and "Bugala (I)," respectively) for subsequent analyses (Fig. S3). (F) Heatmap of F_{ST} between sites (lower triangle) and associated z-score (upper triangle). "Bugala (M)" and "Bugala (I)" are the mainland- and island-like subpopulations of Bugala. (G) Genome-wide pairwise IBD proportions between individuals, based on the full genome, plotted on a logarithmic scale.

mentary Fig. S5). In all comparisons involving islands and some between mainland sites, the best fitting model as chosen via AIC had zero migration (Supplementary Tables S4, S5, and S6). Time since population split was much more recent for mainland-mainland comparisons (excluding Bugala, median: 511 years) than those involving islands (island-island median: 9,080 years; island-mainland median: 5,450 years). Island-island split time confidence intervals typically did not overlap those involving mainland sites.

Selection

We next investigated patterns of selection using genome scans of between- and within-locality statistics (Supplementary Figs. S8, S7, Supplemental Text), including F_{ST} [23], Extended Haplotype Homozygosity (XP-EHH, [24]), and haplotype homozygosity (H12, [25]). Outlier regions included known selective sweep targets [6], including insecticide resistance-associated cytochrome P450 *Cyp6P2* which exhibited low diversity (π), an excess of low frequency

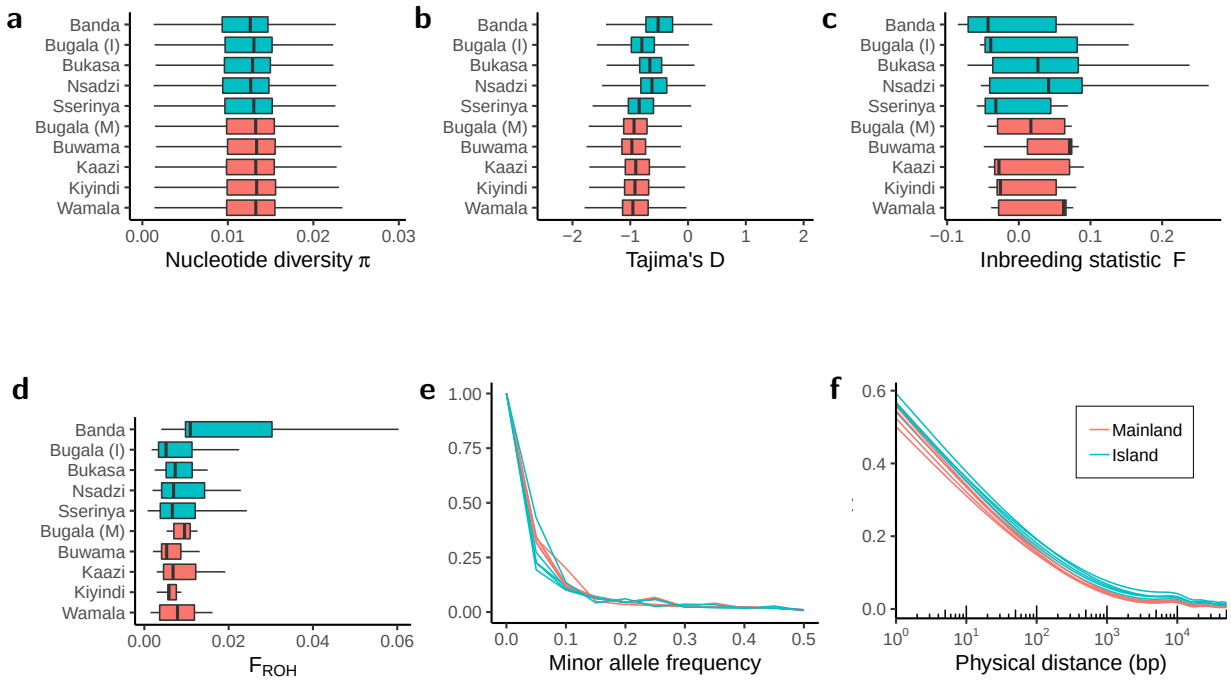

Figure 3: Diversity metrics in the Lake Victoria Basin samples.
Shown are a (A) boxplot of nucleotide diversity (π; in 10 kilobase windows), (B) boxplot of Tajima's D (in 10 kilobase windows), (C) boxplot of inbreeding statistic (F), (D) boxplot of length of runs of homozygosity (F_{ROH}), (E) histogram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

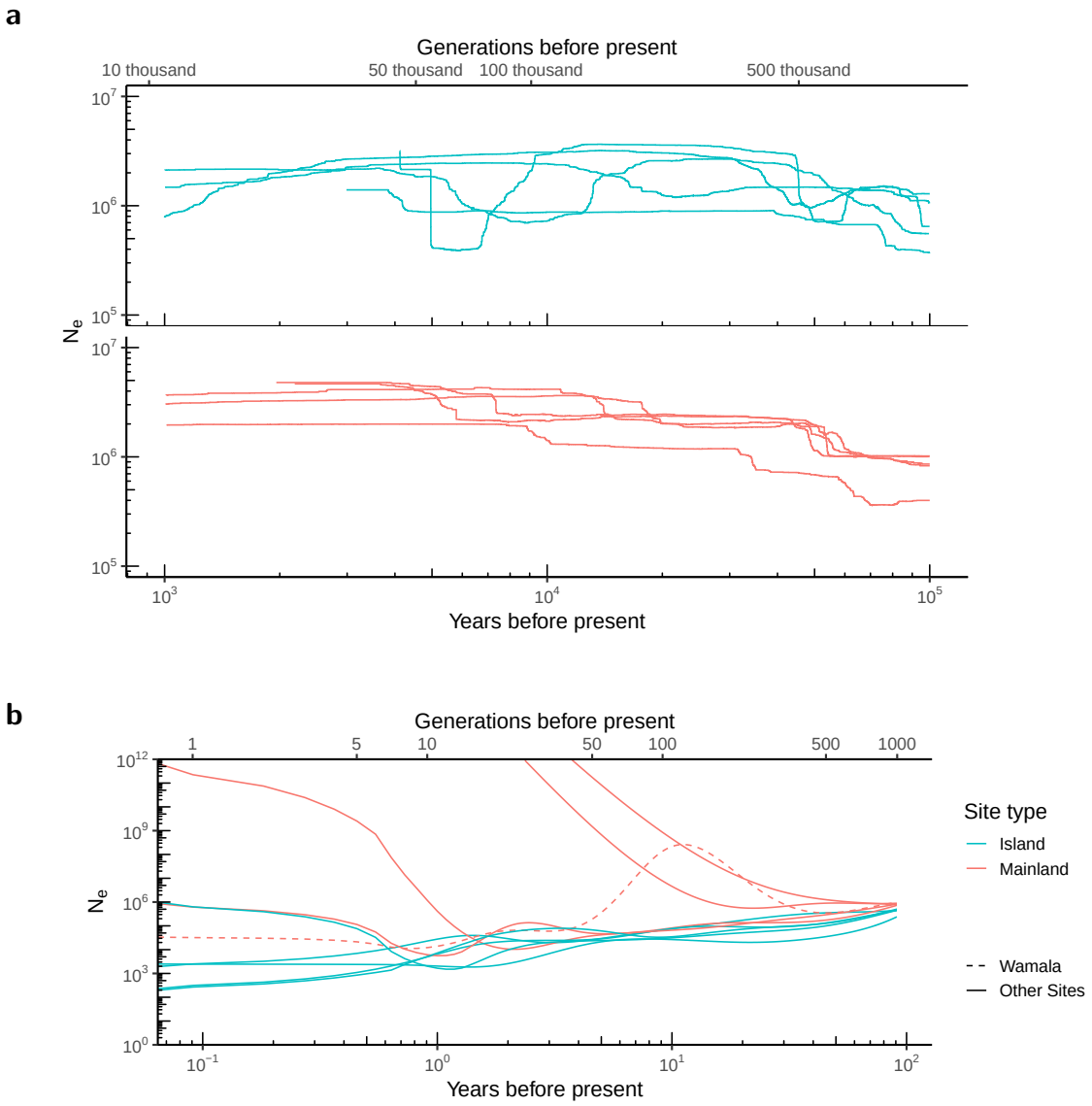

Figure 4: Population history of the Lake Victoria Basin samples.
(A) Long-term evolutionary population histories inferred via stairway plots for island and mainland samples. (B) Contemporary or short-term effective population size (N_e) history inferred using sharing of regions that are identical by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

polymorphisms (Tajima's D), and elevated haplotype homozygosity (H12), but low differentiation between LVB localities, as expected for a continent-wide sweep (Supplementary Fig. S11). Hierarchical clustering of LVB and Ag1000G haplotypes revealed clades with low inter-individual diversity, expected after selection rapidly increases the frequency of a haplotype containing adaptive variation (Supplementary Fig. S17).

Widespread selective sweeps that are absent or at extremely low frequency on the islands are strong evidence against contemporary connectivity. To test for such sweeps, we identified locality-, island-, or LVB mainland-specific sweeps (H12 > 99th percentile), and intersected these regions with those under putative selection across the continent (H12 > 95th percentile in Ag1000G; [6]). 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 (Supplementary Table S7). Some sweeps targeting insecticide genes with continent-wide prevalence ([6]; Supplementary Table S10) were found to have colonized the islands incompletely (*Cyp9K1*: Supplementary Figs. S16 and S12). For instance, the sweep on the cluster of genes encoding glutathione S-transferases (*Gste1-Gste7*) was present across the continent but confined largely to the mainland in the LVB (Supplementary Figs. S18, and S13).

Besides known insecticide-related loci, we identified two regions of elevated between-locality differentiation, low diversity, and extended homozygosity (Supplementary Figs. S8, S7, S10, and 5). The first, at 2L:34.1 Mb, contains many genes, including a cluster involved in chorion formation [26] near the signal peak. Haplotype clustering revealed a group of closely-related Ugandan individuals, consistent with a geographically bounded selective sweep (Supplementary Fig. S14), but the selected variation had not fully colonized the islands. Similar low-variation clades in distinct genetic backgrounds were also found in, *e.g.*, Cameroon and Angola, suggesting convergent selection. The second, at X:9.2 Mb, coincided precisely with eye-specific diacylglycerol kinase (AGAP000519, chrX:9,215,505-9,266,532). Suggestive of

a single sweep, low diversity haplotypes formed a single cluster including LVB haplotypes overwhelmingly from the islands and surprisingly most closely related to haplotypes from distant locations, primarily Gabon and Burkina Faso rather than Uganda (Supplementary Fig. S15). Other adaptive variation supports this surprising affinity between the islands and West Africa: While the LVB mainland-specific sweeps (Supplementary Table S9) were co-located more often with those of the nearby reference Ugandan population (24%) than those of Gabon (16%), far more island-specific sweeps (Supplementary Table S8) were also putative sweeps in the Gabon population (33%) than in the Ugandan population (4.8%).

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 [27–30] and lacustrine islands [31–34], though the latter have been limited in the type or count of molecular markers used. In contrast to shallow population structure across Africa [6, 7], 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 or from different species (*e.g.*, Banda vs. Wamala, $F_{ST} = 0.034$; mainland Uganda vs. Burkina Faso, $F_{ST} = 0.007$ [6]; *An. gambiae* vs. *An. coluzzii* in Burkina Faso: $F_{ST} = 0.031$ [6]). 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 [29]). 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 sharing

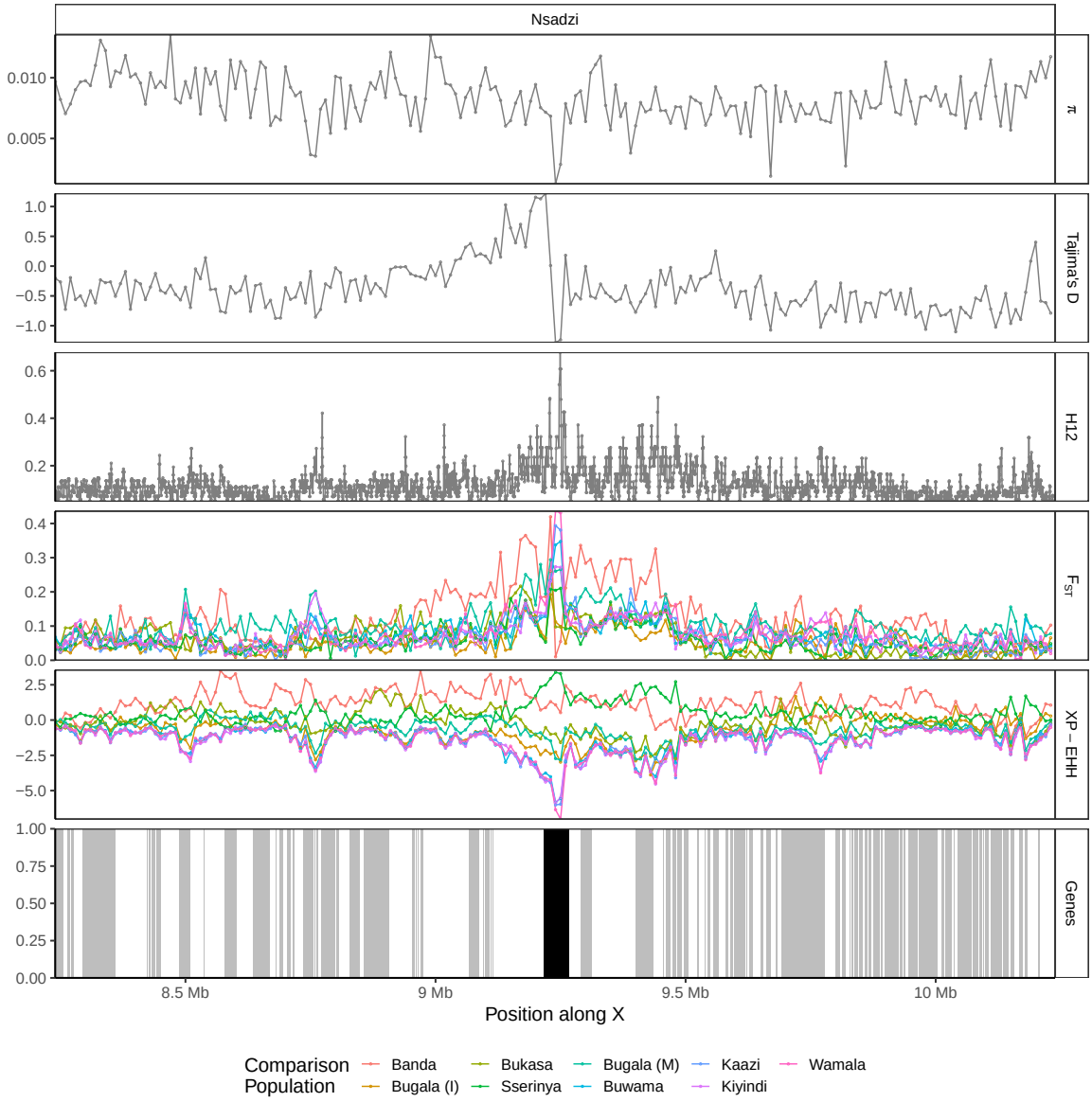

Figure 5: (Caption on next page.)

Figure 5: 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. Region shown is 1 Mb up- and downstream of sweep target, centered at chrX:9,238,942. The gene eye-specific diacylglycerol kinase (AGAP000519, chrX:9,215,505-9,266,532) is highlighted in black.

between Banda and the mainland.

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" [35, pp. 332] to the 20 thousand residents, who were evacuated in 1909 [35, 36]. Though encouraged to return by 1920, the human population numbered only 4 thousand in 1941 [35] and took until 1980 to double [37], but has since rapidly risen to over 62 thousand (2015, projected; [18, 38]). 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 [32, 34]. 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 ($\mathrm{P}f\mathrm{PR}_{2-10}$; a measure of malaria transmission intensity) in Mityana, the district containing Wamala [39].

Though previous *Anopheles* population genetic studies have inferred gene flow even among species [6, 40], 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; [6]). Although bootstrapping-derived confidence intervals permit some certainty, our model fit is not optimal and additional sampling is necessary to clarify population history. Our inferred lack of gene flow to the islands appears contradictory to

the presence of individuals who share ancestry with the mainland on all islands but Banda. We cannot dismiss the possibility that this indicates actual migration occurs. If so, effects of migration would have to be sufficiently countered by local selection to limit its effect on allele frequency spectra, rendering effective migration (as estimated in population history inference) zero. The apparent contradiction can also be resolved if shared ancestry between islands and mainland suggested by the clustering result is interpreted as retention of shared ancestral polymorphism or the existence of inadequately sampled ancestral variation [41], rather than recent admixture. This interpretation is consistent with the affinity we observed between the Ssese Islands and West Africa in the structure of adaptive variation.

As insecticide treated bed net usage is present on the islands [18], variation conferring a major selective advantage would be expected to spread to and persist on the islands if migration allows the transfer, and the strongest evidence of a lack of contemporary connectivity is therefore the absence of a sweep on the islands that is widespread on the continent. We found two sweeps on insecticide-related genes that are common targets of selection elsewhere but which have incompletely colonized the Ssese Islands: one on cytochrome P450 monooxygenase *Cyp9K1* [42, 43] present on some islands, and another on glutathione S-transferase genes (*Gste1-Gste7*; [44–47]) at extremely low frequency on the islands. That the selective sweeps targeting these loci [6] have not fully colonized the islands despite the advantage in detoxifying pyrethroids and DDT suggests a lack of contemporary exchange.

Our investigation also identified two previously unknown signatures of selection with similar uneven distributions. The first encompassed many genes, including a cluster involved in egg shell formation, and the confinement of the signal to Ugandan mosquitoes and limited distribution on the islands suggests a local origin and spread via short distance migration. Overlapping signals in distinct backgrounds suggest the region has been affected by multiple independent convergent sweeps. 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 [48, 49], but also olfactory [50] and auditory [51] sensory processing. It has been recently implicated in nutritional homeostasis in *Drosophila* [52] and is known to interact with the TOR pathway [53], which has been identified as a target of ecological adaptation in *Drosophila* [54, 55] and *An. gambiae* [56]. The sweep appears largely confined to island individuals in the LVB, but their most closely related haplotypes are primarily from Gabon, Burkina Faso, and Kenya. Shared extended haplotypes suggest a single sweep event, not convergence. Possible explanations include long distance migration of an adaptive variant persisting on only the islands, possibly due to a local selective advantage resisting the introgression of mainland haplotypes. 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; [40]), suggesting possible convergence.

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 [57] particularly given the assumption of equilibrium between the evolutionary forces of migration and drift [57–59], an unlikely state for huge *An. gambiae* populations [10]. 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 no island, lacustrine or oceanic, is completely isolated, the probability of contem-

porary migration may be sufficiently low to qualify some Ssese Islands as candidate field sites. Furthermore, 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 trials. 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 study.

Acknowledgments

The authors would like to thank the UVRI field entomology team: Christine Babirye, Ronald Mayanja, Paul Mawejje, Kevin Nakato, and Fred Ssenfuka. 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. We thank Nicholas Harding and Alistair Miles for helpful discussion.

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.

References

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

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

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

[4] Alphey, L. Genetic control of mosquitoes. *Annual Review of Entomology* **59**, 205–224 (2014).

[5] Hammond, A. *et al.* A CRISPR-Cas9 gene drive system targeting female reproduction in the malaria mosquito vector *Anopheles gambiae*. *Nature Biotechnology* **34**, 78–83 (2016).

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

[7] Lehmann, T. *et al.* Population structure of *Anopheles gambiae* in Africa. *Journal of Heredity* **94**, 133–147 (2003).

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

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

[10] Gagnaire, P. A. *et al.* Using neutral, selected, and hitchhiker loci to assess connectivity of marine populations in the genomic era. *Evolutionary Applications* **8**, 769–786 (2015).

[11] Dao, A. *et al.* Signatures of aestivation and migration in Sahelian malaria mosquito populations. *Nature* **516**, 387–390 (2014).

[12] World Health Organization. *Guidance framework for testing genetically modified mosquitoes* (2014).

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

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

[15] James, S. *et al.* Pathway to deployment of gene drive mosquitoes as a potential biocontrol tool for elimination of malaria in sub-Saharan Africa: Recommendations of a scientific working group. *American Journal of Tropical Medicine and Hygiene* **98**, 1–49 (2018).

[16] Uganda Bureau of Statistics (UBOS) & ICF. *Uganda Demographic and Health Survey 2016: Key Indicators Report* (Kampala, Uganda: UBOS, and Rockville, Maryland, USA: UBOS and ICF, 2017).

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

[18] Kalangala District Local Government District Management Improvement Plan 2012-2015. Tech. Rep. (2012).

[19] Coetzee, M. *et al.* *Anopheles coluzzii* and *Anopheles amharicus*, new members of the *Anopheles gambiae* complex. *Zootaxa* **3619**, 246–274 (2013).

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

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

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

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

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

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

[26] Amenya, D. A. *et al.* Proteomics reveals novel components of the *Anopheles gambiae* eggshell. *Journal of Insect Physiology* **56**, 1414–1419 (2010).

[27] Moreno, M. *et al.* Genetic population structure of *Anopheles gambiae* in Equatorial Guinea. *Malaria Journal* **6**, 137 (2007).

[28] Marshall, J. C. *et al.* Exploring the origin and degree of genetic isolation of *Anopheles gambiae* from the islands of São Tomé and Príncipe, potential sites for testing transgenic-based vector control. *Evolutionary Applications* **1**, 631–644 (2008).

[29] Marsden, C. D. *et al.* An analysis of two island groups as potential sites for trials of transgenic mosquitoes for malaria control. *Evolutionary Applications* **6**, 706–720 (2013).

[30] Maliti, D. *et al.* 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).

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

[32] Kayondo, J. K. *et al.* Genetic structure of *Anopheles gambiae* populations on islands in northwestern Lake Victoria, Uganda. *Malaria Journal* **4**, 59 (2005).

[33] Lukindu, M. *et al.* Spatio-temporal genetic structure of *Anopheles gambiae* in the Northwestern Lake Victoria Basin, Uganda: Implications for genetic control trials in malaria endemic regions. *Parasites and Vectors* **11** (2018).

[34] Wiltshire, R. M. *et al.* Reduced-representation sequencing identifies small effective population sizes of *Anopheles gambiae* in the north-western Lake Victoria basin, Uganda. *Malaria Journal* **17**, 285 (2018).

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

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

[37] Uganda Bureau of Statistics. *2002 Uganda Population and Housing Census Analytical Report* (2002).

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

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

[40] Crawford, J. E. *et al.* Evolution of GOUNDRY, a cryptic subgroup of *Anopheles gambiae s.l.*, and its impact on susceptibility to *Plasmodium* infection. *Molecular Ecology* **25** (2016).

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

[42] Vontas, J. *et al.* Rapid selection of a pyrethroid metabolic enzyme CYP9K1 by operational malaria control activities. *Proceedings of the National Academy of Sciences* 201719663 (2018).

[43] Fossog Tene, B. *et al.* 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).

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

[45] Mitchell, S. N. *et al.* Metabolic and target-site mechanisms combine to confer strong DDT resistance in *Anopheles gambiae*. *PLoS ONE* **9** (2014).

[46] Jones, C. M. *et al.* Additional selection for insecticide resistance in urban malaria vectors: DDT resistance in *Anopheles arabiensis* from Bobo-Dioulasso, Burkina Faso. *PLoS ONE* **7** (2012).

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

[48] Hardie, R. *et al.* Molecular basis of amplification in *Drosophila* phototransduction: Roles for G protein, phospholipase C, and diacylglycerol kinase. *Neuron* **36**, 689–701 (2002).

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

[50] Kain, P. *et al.* 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).

[51] Senthilan, P. R. *et al.* *Drosophila* auditory organ genes and genetic hearing defects. *Cell* **150**, 1042–1054 (2012).

[52] Nelson, C. S. *et al.* Cross-phenotype association tests uncover genes mediating nutrient response in *Drosophila*. *BMC Genomics* **17**, 1–14 (2016).

[53] Lin, Y. H. *et al.* Diacylglycerol lipase regulates lifespan and oxidative stress response by inversely modulating TOR signaling in *Drosophila* and *C. elegans*. *Aging Cell* **13**, 755–764 (2014).

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

[55] Fabian, D. K. *et al.* Genome-wide patterns of latitudinal differentiation among populations of *Drosophila melanogaster* from North America. *Molecular Ecology* **21**, 4748–4769 (2012).

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

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

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

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

Online Methods

Sample collection, morphological ID 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 [60] 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 [61] 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 S11. After quality filtering and trimming using ea-utils' fastq-mcf (-l 15 -q 15 -w 4; [62]), reads were mapped to the *An. gambiae* reference genome (AgamP4 PEST; [63, 64]) using BWA aln and sampe with default parameters [65].

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) and filtered for quality [66], excluding SNPs with QualByDepth (QD) < 2.0, RMSMappingQuality (MQ)

455 < 40.0, FisherStrand (FS) > 60.0, HaplotypeScore > 13.0, or ReadPosRankSum $<$ -8.0. All
456 bioinformatic steps for read mapping and variant identification are encapsulated in the NGS-
457 map pipeline (https://github.com/bergeycm/NGS-map). This yielded 33.1 million SNPs.
458 Individuals and variants with high levels of missingness ($> 10\%$) and variants that were
459 not biallelic or exhibited values of HWE that were likely due to sequencing error ($p <$
460 0.00001) were excluded from further analysis. For use in population structure inference, the
461 SNP dataset was further pruned for linkage disequilibrium by sliding a window of 50 SNPs
462 across the genome in 5 SNP increments and recursively removing random SNPs in pairs
463 with $r^2 > 0.5$ using PLINK [67, 68]. After filtration, the dataset contained 28,569,621 SNPs
464 before LD pruning and 115 individuals. SNPs unpruned for linkage disequilibrium were
465 phased with SHAPEIT2 [69] using an effective population size (N_e) of 1,000,000 (consistent
466 with previous demographic modeling [70]), default MCMC parameters (7 burn-in MCMC
467 iterations, 8 pruning iterations, and 20 main iterations), conditioning states for haplotype
468 estimation ($K = 100$), and window size of 2 Mb.

469 **Population structure inference** To explore population structure in a larger, continent-
470 wide context, we merged our LVB SNP set with a recently published dataset of *Anopheles*
471 *gambiae* individuals (from the Ag1000G project) [70]. Prior to filtering, biallelic SNPs from
472 the LVB and Ag1000G datasets were merged using bcftools [71]. We excluded any SNP with
473 greater than 10% missingness in either dataset, any SNPs that did not pass the accessibility
474 filter of the Ag1000G dataset, and SNPs with MAF $< 1\%$. After this filtration, our merged
475 SNP dataset contained 12,537,007 SNPs.

476 After pruning the merged dataset for LD (leaving 9,861,756 SNPs) and excluding labo-
477 ratory crosses (leaving 881 individuals), we assigned individuals' genomes to ancestry com-
478 ponents using ADMIXTURE [72]. We created 10 replicate samples of 100,000 SNPs from
479 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 [67, 68], using the LVB-Ag1000G merged dataset (excluding the outlier Kenyan population; [70]) 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; [64]; Supplementary Table S3). 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 [67, 68]. 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 [67, 68] 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.

Demographic history inference To estimate the long-term evolutionary demographic history of mosquitoes on and near the Ssese Islands, including a long-term estimate of N_e [79], we inferred population demographic history for each site via stairway plots using the full site frequency spectra from the same dataset [80].

To estimate the contemporary or short-term N_e for each site, we inferred regions of IBD from unphased data with IBDseq [81] and analyzed them with IBDNe [82]. 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 [70] and assumed a constant recombination rate of 2.0 cM/Mb [83].

We also inferred a "two-population" isolation-with-migration (IM) demographic model with δaδi [84, 85] 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 masked. We polarized the SNPs using outgroup information from *Anopheles merus* and *An. merus* [86]. 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 = 0.67%). We assessed the goodness-of-fit visually using the residuals of the comparison between model and data frequency spectra (Supplementary Fig. S6). 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.

To translate $\delta a\delta i$- and stairway plot-based estimates of N_e and time to individuals and years respectively, we assumed a generation time of 11 per year and a mutation rate of $3.5e$-9 per generation [70].

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 [87] (as implemented in VCFtools [88]), and H12 (as implemented in SelectionHapStats [89]) and XP-EHH on a per-site basis (as implemented in selscan [90]) to detect long stretches of homozygosity in a given population considered alone or relative to another population [91]. 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 [83]. 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; [92, 93]].

We plotted these statistics across the genome to identify candidate regions with signatures of selection, including high differentiation between samples from different localities, reduced variability within a sample, and extended haplotype homozygosity. To identify regions of the genome showing signatures of selection specific to certain geographic areas, we identified genomic regions with elevated H12 in a subset of localities, and confirmed both elevated differentiation (as inferred from F_{ST}) and evidence of differing selective sweep histories (as inferred from XP-EHH). Excluding the mainland-like portion of Bugala, we identified

putative locality-specific sweeps (H12 over 99th percentile in one population), island-specific sweeps (H12 over 99th percentile in 4 or more of the 5 island localities but 0 or 1 mainland localities), or LVB mainland-specific sweeps (H12 over 99th 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; [70]) 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; [70]). 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 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.

Script and data 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.

References

[60] Scott, J. A., Brogdon, W. G. & Collins, F. H. 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).

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

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

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

[64] Sharakhova, M. V. *et al.* Update of the *Anopheles gambiae* PEST genome assembly. *Genome Biology* **8**, R5 (2007).

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

[66] DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43**, 491–498 (2011).

[67] Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* **81**, 559–75 (2007).

[68] Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).

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

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

[71] Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–9 (2009).

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

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

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

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

[76] Price, A. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**, 904–9 (2006).

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

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

[79] Hare, M. P. *et al.* Understanding and estimating effective population size for practical application in marine species management. *Conservation Biology* **25**, 438–449 (2011).

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

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

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

[83] Clarkson, C. S. *et al.* The genetic architecture of target-site resistance to pyrethroid insecticides in the African malaria vectors *Anopheles gambiae* and *Anopheles coluzzii*. *bioRxiv* Preprint at: https://www.biorxiv.org/content/early/2018/08/06/323980 (2018).

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

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

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

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

670 [88] Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–8
671 (2011).

672 [89] Garud, N. R., Messer, P. W., Buzbas, E. O. & Petrov, D. A. Recent selective sweeps in
673 North American *Drosophila melanogaster* show signatures of soft sweeps. *PLoS Genetics*
674 **11**, 1–32 (2015).

675 [90] Szpiech, Z. A. & Hernandez, R. D. selscan: an efficient multithreaded program to
676 perform EHH-based scans for positive selection. *Molecular Biology and Evolution* **31**,
677 2824–2827 (2014).

678 [91] Sabeti, P. C. *et al.* Genome-wide detection and characterization of positive selection in
679 human populations. *Nature* **449**, 913–918 (2007).

680 [92] Randhawa, I. A., Khatkar, M., Thomson, P. & Raadsma, H. Composite selection signals
681 can localize the trait specific genomic regions in multi-breed populations of cattle and
682 sheep. *BMC Genetics* **15**, 34 (2014).

683 [93] Wallberg, A., Pirk, C. W., Allsopp, M. H. & Webster, M. T. Identification of multiple
684 loci associated with social parasitism in honeybees. *PLoS Genetics* **12**, 1–30 (2016).

Supplemental Material for: Assessing connectivity despite high diversity in island populations of the malaria mosquito *Anopheles gambiae*

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: Genomic coordinates of heterochromatic and inverted regions.

Chromosome arm	Start	End	Information
2L	20,524,058	42,165,532	2La inversion [94]
2R	18,575,300	26,767,588	2Rb inversion [94]
2L	1	2,431,617	Heterochromatic region [94]
2L	5,078,962	5,788,875	Heterochromatic region [94]
2R	58,984,778	61,545,105	Heterochromatic region [94]
3L	1	1,815,119	Heterochromatic region [94]
3L	4,264,713	5,031,692	Heterochromatic region [94]
3R	38,988,757	41,860,198	Heterochromatic region [94]

3R	52,161,877	53,200,684	Heterochromatic region [94]
X	20,009,764	24,393,108	Heterochromatic region [94]

Table S4: Results of two population demographic inference with IM model in $\delta a \delta i$ 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)	762,000	0.538	2.79	9,570	5,050	None	None
	(762,000, 763,000)	(0.522, 0.555)	(2.55, 3.03)	(8,000, 11,100)	(4,900, 5,200)		
Banda - Bukasa	755,000	0.595	8.53	9,850	9,760	None	None
	(754,000, 756,000)	(0.59, 0.599)	(8.22, 8.84)	(9,030, 10,700)	(9,660, 9,850)		
Banda - Nsadzi	760,000	0.557	40.3	8,550	13,800	None	None
	(759,000, 761,000)	(0.551, 0.562)	(37.6, 43)	(7,500, 9,590)	(13,600, 14,000)		
Banda - Sserinya	764,000	0.497	4.74	8,460	5,100	None	None
	(763,000, 765,000)	(0.489, 0.506)	(4.23, 5.25)	(6,590, 10,300)	(4,930, 5,280)		
Bugala (I) - Bukasa	759,000	0.588	9,330	2,960	8,400	None	None
	(758,000, 760,000)	(0.575, 0.6)	(7,690, 11,000)	(2,600, 3,320)	(8,210, 8,580)		
Bugala (I) - Nsadzi	759,000	0.499	9,350	30.2	8,380	None	None
	(758,000, 760,000)	(0.492, 0.505)	(7,740, 11,000)	(26.2, 34.2)	(8,170, 8,590)		
Bugala (I) - Sserinya	763,000	0.592	7,940	593	4,860	None	None
	(762,000, 764,000)	(0.574, 0.61)	(6,530, 9,350)	(513, 673)	(4,720, 5,010)		
Bukasa - Nsadzi	759,000	0.436	9,720	77.4	12,200	None	None
	(758,000, 760,000)	(0.427, 0.446)	(8,270, 11,200)	(67.2, 87.6)	(11,900, 12,400)		
Bukasa - Sserinya	755,000	0.493	9,960	5,090	12,700	None	None
	(754,000, 756,000)	(0.488, 0.497)	(8,190, 11,700)	(4,470, 5,710)	(12,600, 12,900)		
Nsadzi - Sserinya	777,000	0.615	53.9	9,210	10,500	None	None
	(776,000, 778,000)	(0.594, 0.635)	(33.2, 74.5)	(7,420, 11,000)	(9,940, 11,100)		

Table S5: 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)	751,000	0.522	4.38	8,580	7,610	None	None
	(750,000, 752,000)	(0.511, 0.532)	(4.15, 4.61)	(7,470, 9,690)	(7,460, 7,750)		
Banda - Buwama	751,000	0.457	1.92	7,470	5,160	None	None
	(751,000, 752,000)	(0.439, 0.476)	(1.68, 2.16)	(6,460, 8,480)	(4,960, 5,350)		
Banda - Kaazi	752,000	0.477	3.75	9,040	7,550	None	None
	(751,000, 753,000)	(0.466, 0.488)	(3.52, 3.99)	(7,680, 10,400)	(7,400, 7,690)		
Banda - Kiyindi	735,000	0.511	1.82	9,360	5,450	None	None
	(734,000, 736,000)	(0.501, 0.52)	(1.76, 1.87)	(8,160, 10,600)	(5,410, 5,490)		
Banda - Wamala	750,000	0.596	2.2	8,740	6,600	None	None
	(749,000, 751,000)	(0.586, 0.606)	(2.09, 2.3)	(7,610, 9,860)	(6,480, 6,720)		
Bugala (I) - Bugala (M)	752,000	0.496	3,090	8,940	6,580	None	None
	(751,000, 753,000)	(0.484, 0.508)	(2,490, 3,690)	(7,420, 10,500)	(6,450, 6,710)		
Bugala (I) - Buwama	753,000	0.5	0.198	9,330	274	None	None
	(753,000, 754,000)	(0.497, 0.502)	(0.194, 0.202)	(-63,800, 82,500)	(271, 276)		
Bugala (I) - Kaazi	753,000	0.401	1,500	9,240	7,780	None	None
	(752,000, 754,000)	(0.388, 0.414)	(1,350, 1,640)	(7,950, 10,500)	(7,630, 7,940)		
Bugala (I) - Kiyindi	735,000	0.478	0.14	8,260	254	None	None
	(734,000, 736,000)	(0.47, 0.487)	(0.137, 0.143)	(-92,500, 109,000)	(250, 259)		
Bugala (I) - Wamala	748,000	0.479	7,230	6,590	5,880	None	None
	(748,000, 749,000)	(0.462, 0.496)	(5,220, 9,240)	(5,120, 8,050)	(5,730, 6,040)		
Bugala (M) - Bukasa	751,000	0.497	9,080	47.2	6,870	None	None
	(750,000, 751,000)	(0.49, 0.504)	(7,630, 10,500)	(42.5, 52)	(6,750, 7,000)		
Bugala (M) - Nsadzi	771,000	0.381	9,030	3.95	5,610	None	None

	(770,000, 772,000)	(0.368, 0.394)	(7,690, 10,400)	(3.64, 4.26)	(5,480, 5,750)		
Bugala (M) - Sserinya	768,000	0.483	9,070	22.7	5,080	None	None
	(767,000, 769,000)	(0.471, 0.495)	(7,620, 10,500)	(21.5, 24)	(5,020, 5,140)		
Bukasa - Buwama	751,000	0.536	1.74	9,680	2,650	None	None
	(750,000, 752,000)	(0.527, 0.545)	(1.71, 1.77)	(7,860, 11,500)	(2,630, 2,670)		
Bukasa - Kaazi	751,000	0.533	16.2	7,730	6,690	None	None
	(750,000, 752,000)	(0.524, 0.543)	(15.7, 16.6)	(6,260, 9,200)	(6,620, 6,760)		
Bukasa - Kiyindi	733,000	0.549	4.64	9,170	4,120	None	None
	(732,000, 734,000)	(0.523, 0.575)	(3.93, 5.36)	(7,280, 11,100)	(3,900, 4,350)		
Bukasa - Wamala	748,000	0.361	281	6,790	7,320	None	None
	(747,000, 749,000)	(0.345, 0.377)	(222, 340)	(5,510, 8,070)	(7,110, 7,530)		
Buwama - Nsadzi	756,000	0.608	9,500	3.2	3,960	None	None
	(755,000, 756,000)	(0.593, 0.624)	(8,020, 11,000)	(2.98, 3.42)	(3,820, 4,090)		
Buwama - Sserinya	753,000	0.498	5,090	0.134	273	None	None
	(752,000, 754,000)	(0.495, 0.501)	(-12,800, 23,000)	(0.132, 0.136)	(271, 274)		
Kaazi - Nsadzi	756,000	0.516	9,050	11.6	7,350	None	None
	(755,000, 756,000)	(0.493, 0.539)	(7,510, 10,600)	(8.26, 14.9)	(6,970, 7,730)		
Kaazi - Sserinya	752,000	0.529	7,950	15	4,920	None	None
	(752,000, 753,000)	(0.511, 0.547)	(6,600, 9,300)	(13.4, 16.6)	(4,780, 5,060)		
Kiyindi - Nsadzi	738,000	0.534	9,990	3.79	4,800	None	None
	(738,000, 739,000)	(0.527, 0.541)	(8,580, 11,400)	(3.59, 3.99)	(4,750, 4,860)		
Kiyindi - Sserinya	735,000	0.5	5,400	0.129	267	None	None
	(734,000, 736,000)	(0.497, 0.502)	(-18,000, 28,800)	(0.127, 0.131)	(265, 269)		
Wamala - Nsadzi	755,000	0.485	9,160	5.71	5,890	None	None
	(754,000, 755,000)	(0.479, 0.491)	(7,870, 10,400)	(5.6, 5.81)	(5,860, 5,930)		
Wamala - Sserinya	751,000	0.636	9,240	152	5,290	None	None
	(750,000, 752,000)	(0.622, 0.65)	(7,720, 10,700)	(135, 169)	(5,140, 5,440)		

Table S6: 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	750,000	0.172	1.58	4.63	341	0	0.339
	(749,000, 751,000)	(0.164, 0.18)	(1.52, 1.63)	(4.35, 4.9)	(337, 344)	(-5.81, 5.81)	(-15.3, 16)
Bugala (M) - Kaazi	750,000	0.517	1,840	6,270	3,330	None	None
	(749,000, 751,000)	(0.484, 0.55)	(1,430, 2,250)	(4,610, 7,930)	(3,230, 3,430)		
Bugala (M) - Kiyindi	735,000	0.552	9,800	267	1,340	None	None
	(734,000, 736,000)	(0.412, 0.693)	(2,120, 17,500)	(160, 373)	(733, 1,960)		
Bugala (M) - Wamala	748,000	0.325	9.38	18.8	911	0	1.89
	(747,000, 749,000)	(0.308, 0.342)	(8.92, 9.85)	(17.4, 20.2)	(899, 922)	(-16.9, 16.9)	(-30.7, 34.5)
Buwama - Kaazi	753,000	0.352	1,230	614	180	0	236
	(752,000, 754,000)	(-10.2, 10.9)	(-1,800, 4,270)	(-900, 2,130)	(176, 183)	(-10,600, 10,600)	(204, 267)
Buwama - Kiyindi	738,000	0.677	12.5	4.56	91	0	0
	(737,000, 739,000)	(-1,560, 1,560)	(9.48, 15.5)	(3.48, 5.64)	(74.2, 108)	(-524, 524)	(-375, 375)
Buwama - Wamala	753,000	0.709	9,280	6,750	255	718	1,660
	(752,000, 754,000)	(0.621, 0.798)	(-510, 19,100)	(315, 13,200)	(177, 333)	(-36,500, 37,900)	(-74,700, 78,000)
Kaazi - Kiyindi	736,000	0.447	549	7,600	764	143	603
	(735,000, 737,000)	(0.371, 0.523)	(376, 722)	(4,230, 11,000)	(661, 868)	(83.5, 202)	(-25,400, 26,600)
Kaazi - Wamala	749,000	0.499	5,890	7,880	1,720	None	None
	(749,000, 750,000)	(0.493, 0.504)	(4,040, 7,730)	(5,300, 10,500)	(1,610, 1,830)		
Kiyindi - Wamala	736,000	0.5	7,750	6,520	511	3.98	47
	(735,000, 737,000)	(0.494, 0.505)	(3,650, 11,900)	(3,050, 10,000)	(463, 560)	(-1,690, 1,690)	(-20,600, 20,700)

Table S7: 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 S8: 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 S9: 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 S10: 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 S11: Software and versions used for major parts of analysis.

Software	Version	Citation
ea-utils	-	[95]
BWA	0.7.16a	[96]
GATK	3.8	[97]
PLINK	1.90b4.6	[98, 99]
SHAPEIT2	2.837	[100]
SAMtools/BCFtools	1.5	[101, 102]
ADMIXTURE	1.3.0	[103]
CLUMPAK	-	[104]
VCFtools	0.1.15	[105]
$\delta a \delta i$ (python package)	1.7.0	[106, 107]
Stairway plot - Jpopgen	2-beta	[108]
selscan	1.2.0a	[109]
adegenet (R package)	2.1.0	[110]
ape (R package)	5.0	[111]
RColorBrewer (R package)	1.1-2	[112]
dendextend (R package)	1.6.0	[113]
rehh (R package)	2.0.2	[114]
eigensoft	7.2.1	[115, 116]
GNU parallel	20170422	[117]
tabix	1.5	[101]
bedtools	2.26.0	[118]

Figures

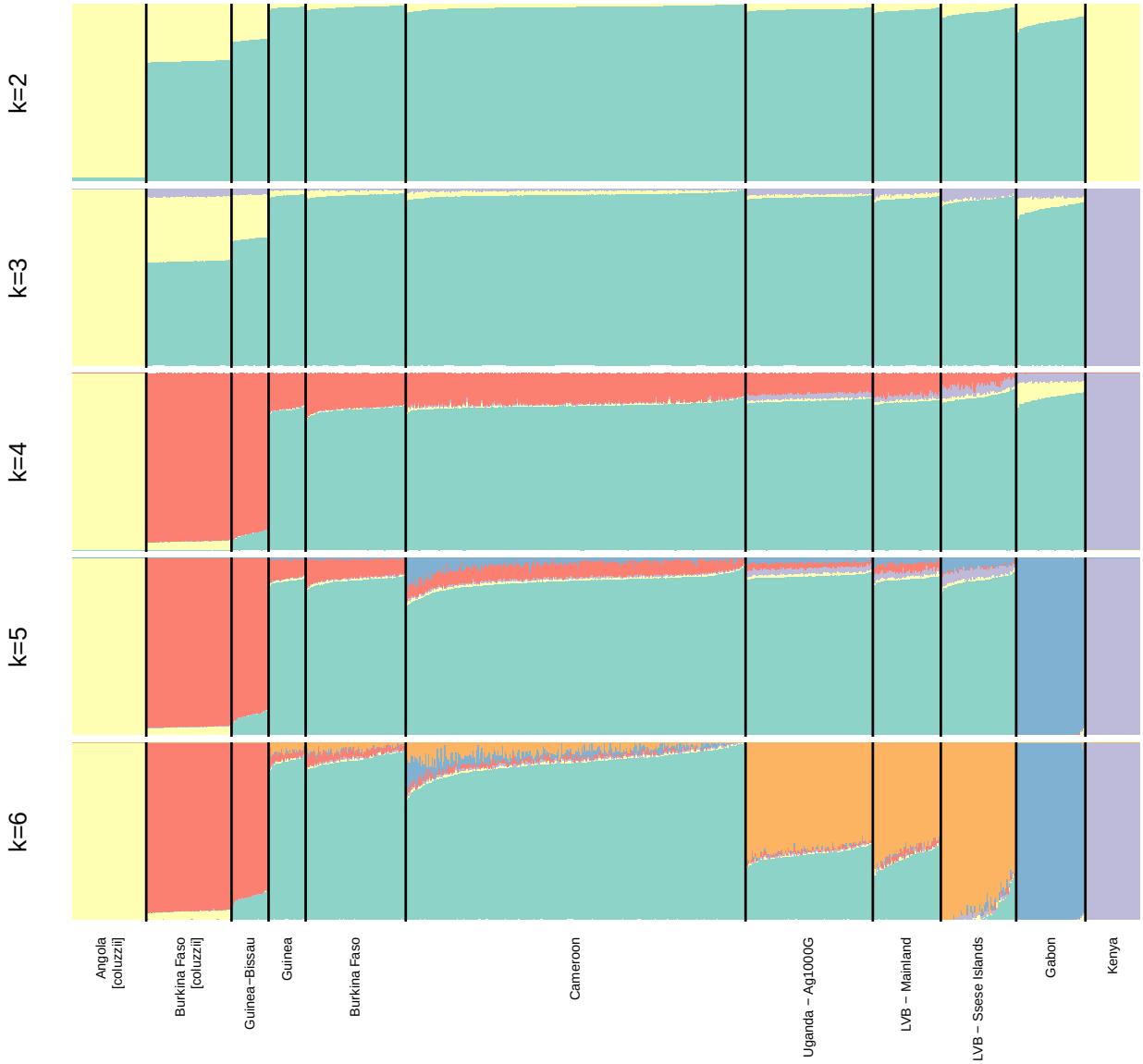

Figure S1: (Caption on next page.)

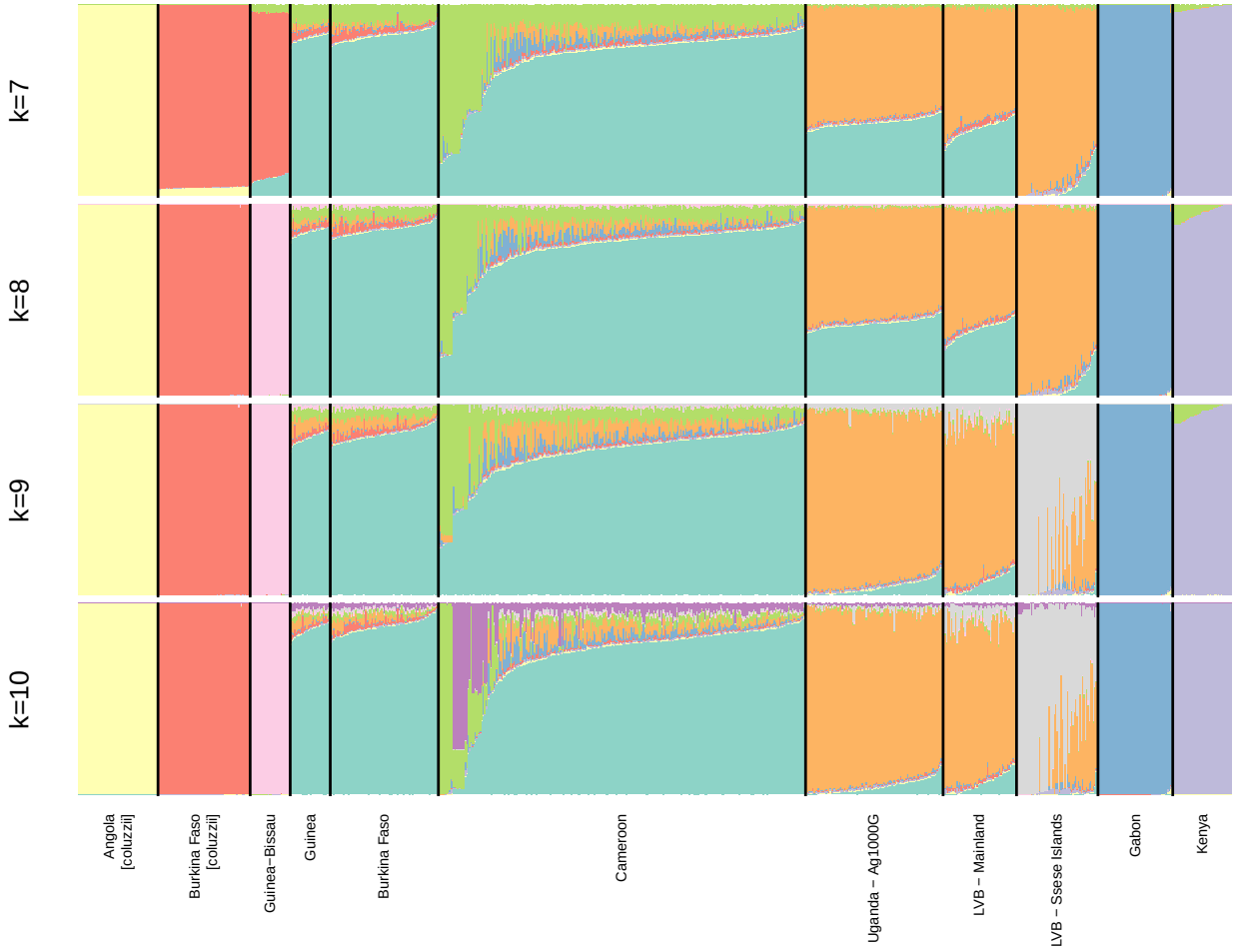

Figure S1: ADMIXTURE-inferred ancestry.
Ancestry of individuals in Lake Victoria Basin and of Ag1000G reference populations as inferred by ADMIXTURE clustering method. Samples are *A. gambiae* unless noted, and analysis is based on chromosome 3. Using $k = 6$ clusters minimizes cross validation error (Fig. S2).

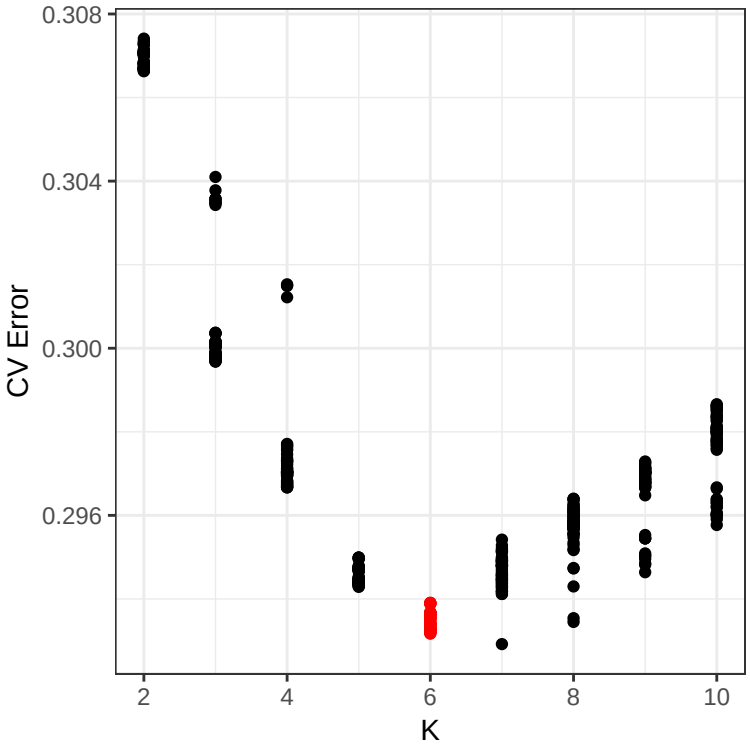

Figure S2: ADMIXTURE cross-validation error.
Cross-validation error for range of k values for ADMIXTURE analysis of Lake Victoria Basin individuals and *A. gambiae* and *A. coluzzii* Ag1000G reference populations.

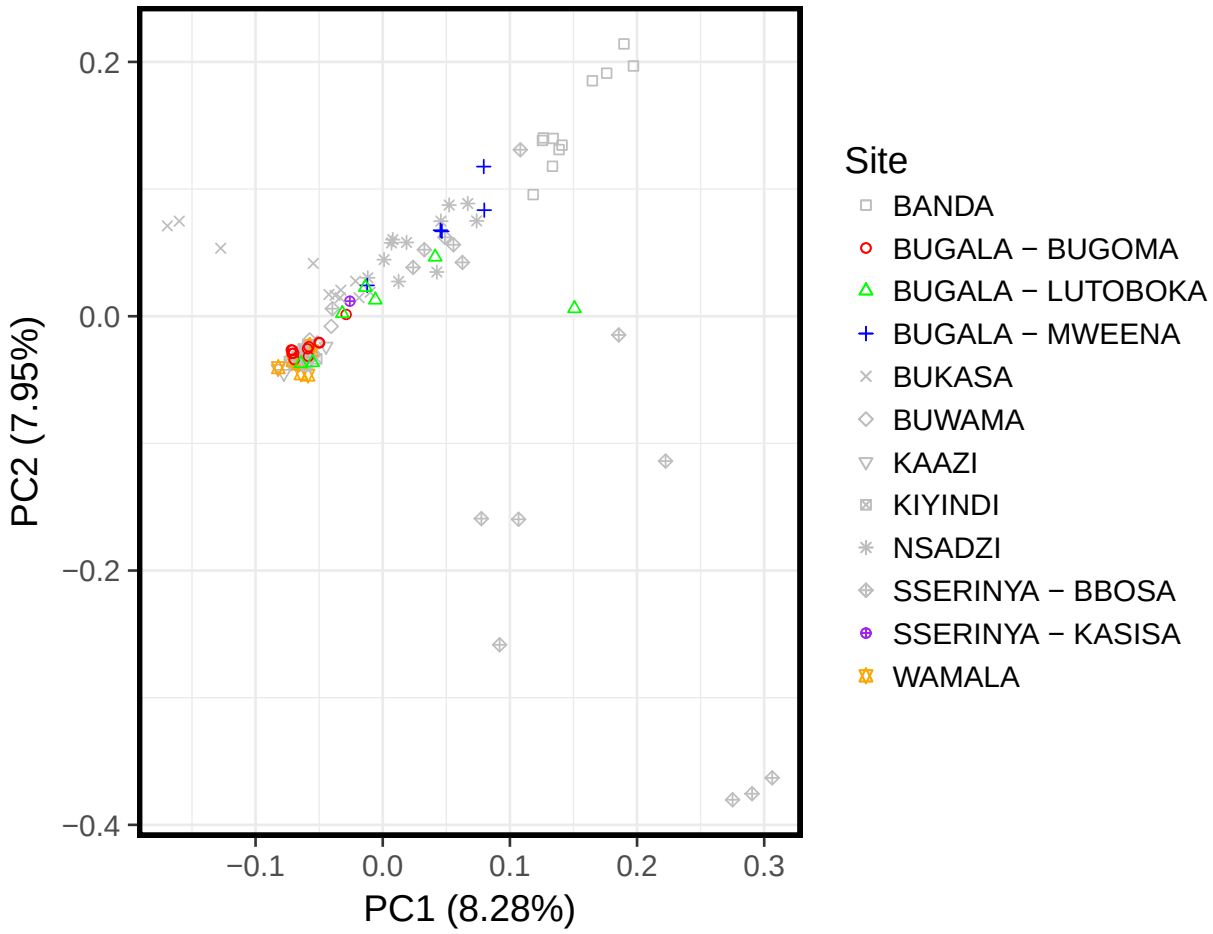

Figure S3: PCA showing Bugala subdivision.
PCA colored by sampling locations. Based on this analysis, individuals from Bugala were split into mainland- and island-like subpopulations. Samples from Sserinya Island, though sampled from two localities, were not split.

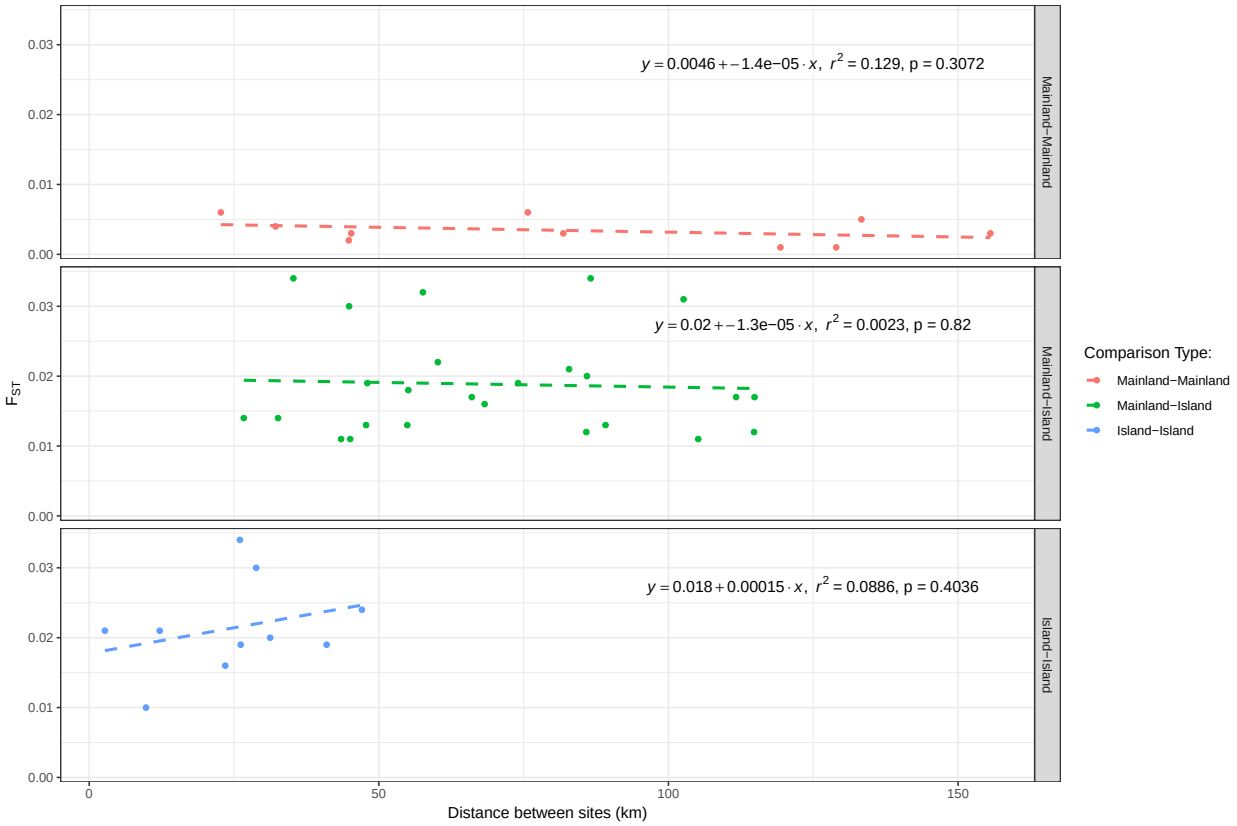

Figure S4: Correlations between genetic distance (F_{ST}) and geographic distance between localities. The p-values are for the test that the slope is significantly different from zero.

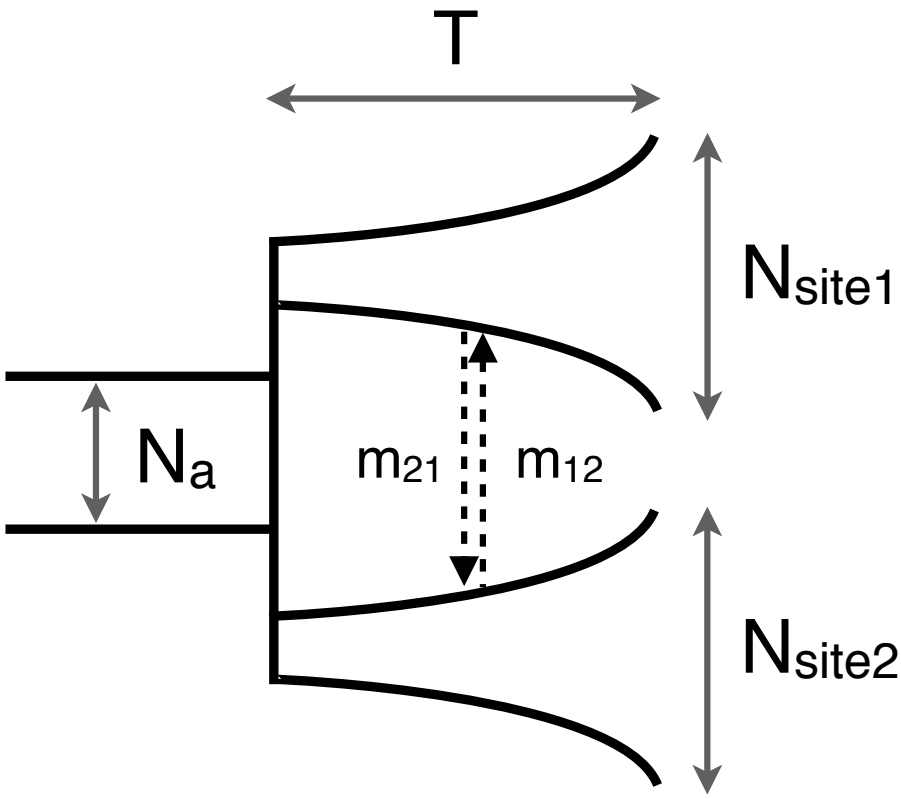

Figure S5: IM model schematics.
Schematic of model fit to data with $\delta a \delta i$ for population history inference between all pairs of sampled sites using IM model.

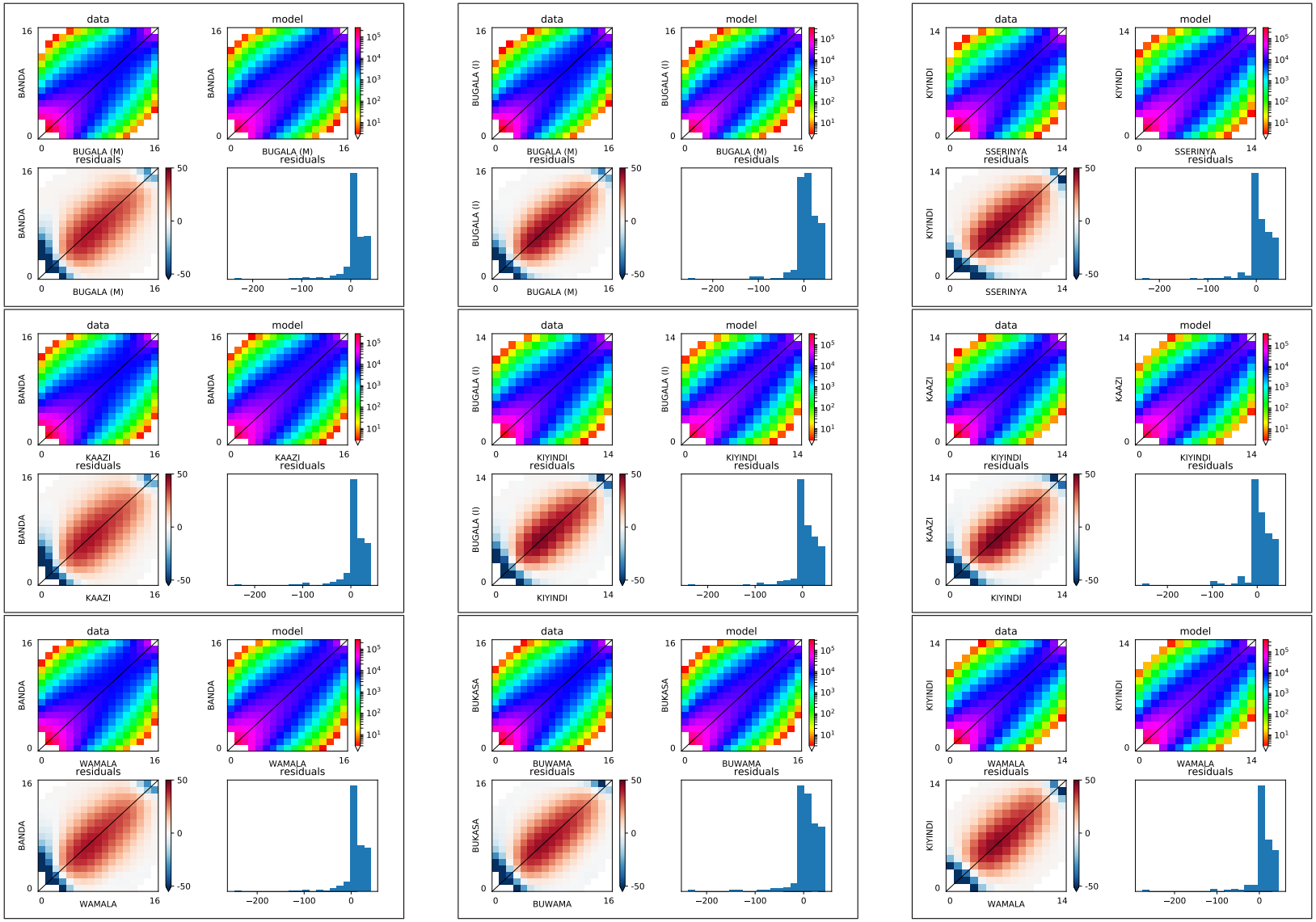

Figure S6: (Caption on next page.)

Figure S6: 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.

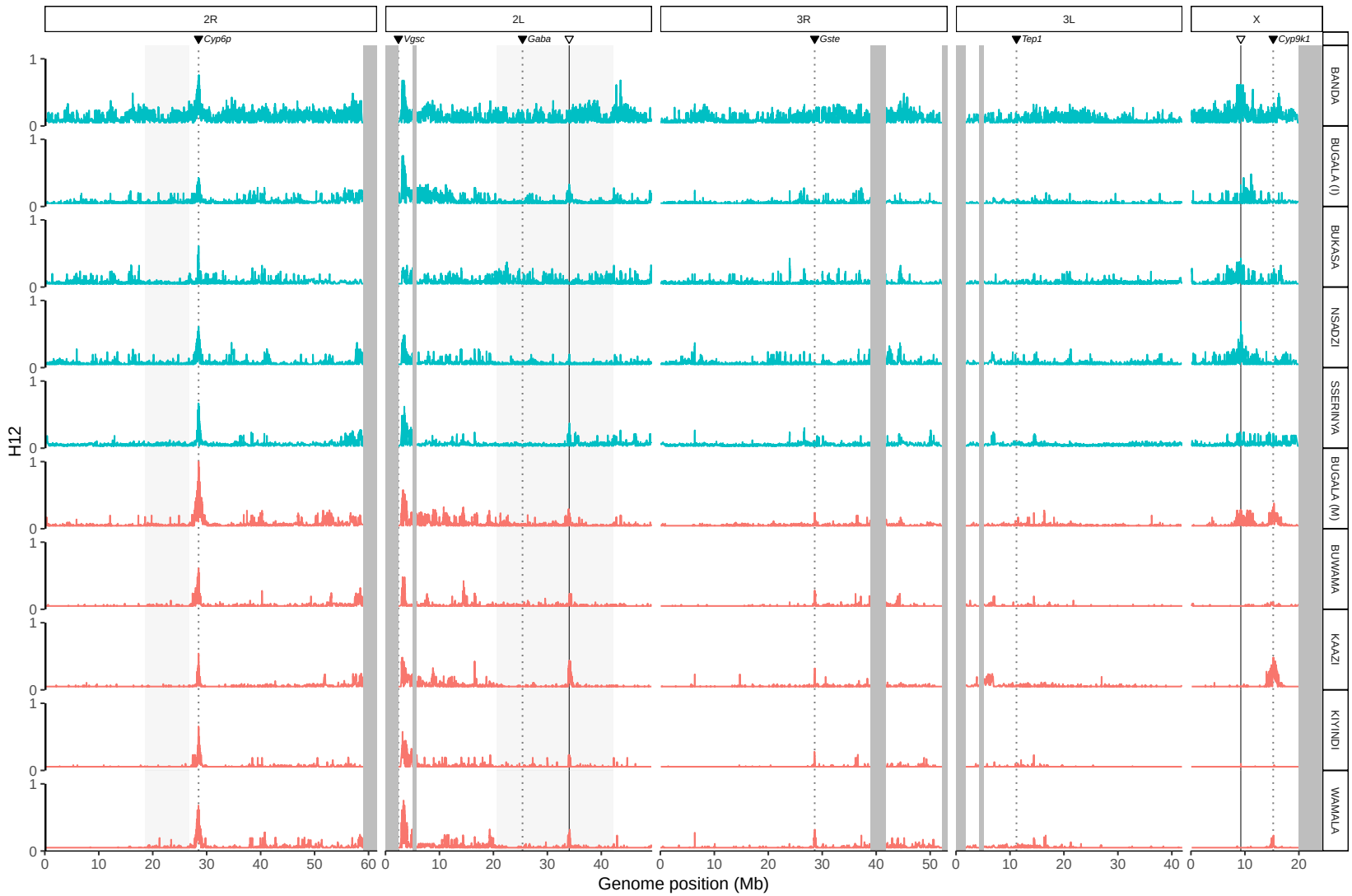

Figure S7: H12 across genome.

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 dashed lines indicate the two putative sweeps identified in the present study.

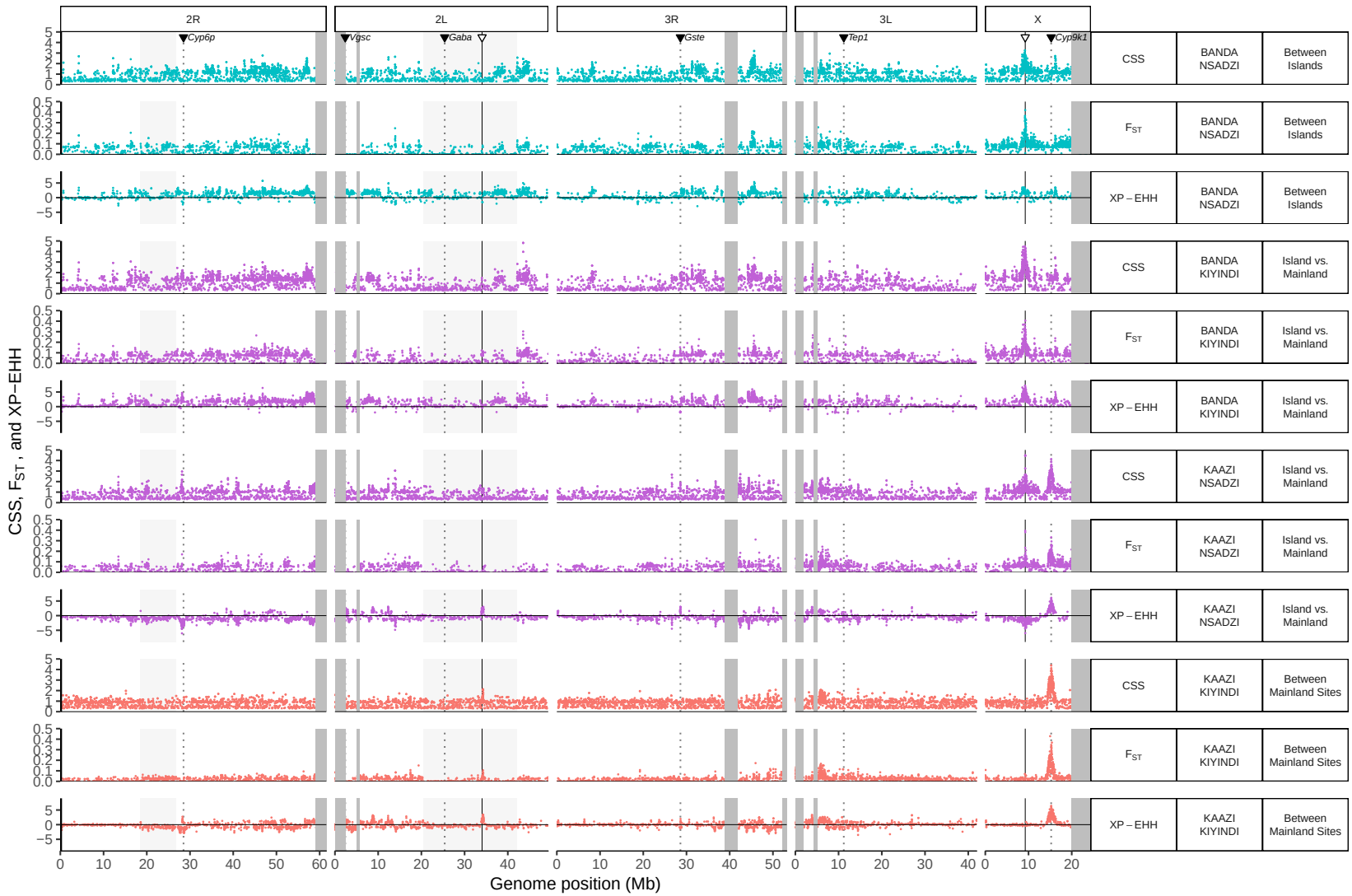

Figure S8: F_{ST} and XP-EHH across genome.

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 dashed lines indicate the two putative sweeps identified in the present study. Only several exemplar pairs of populations shown.

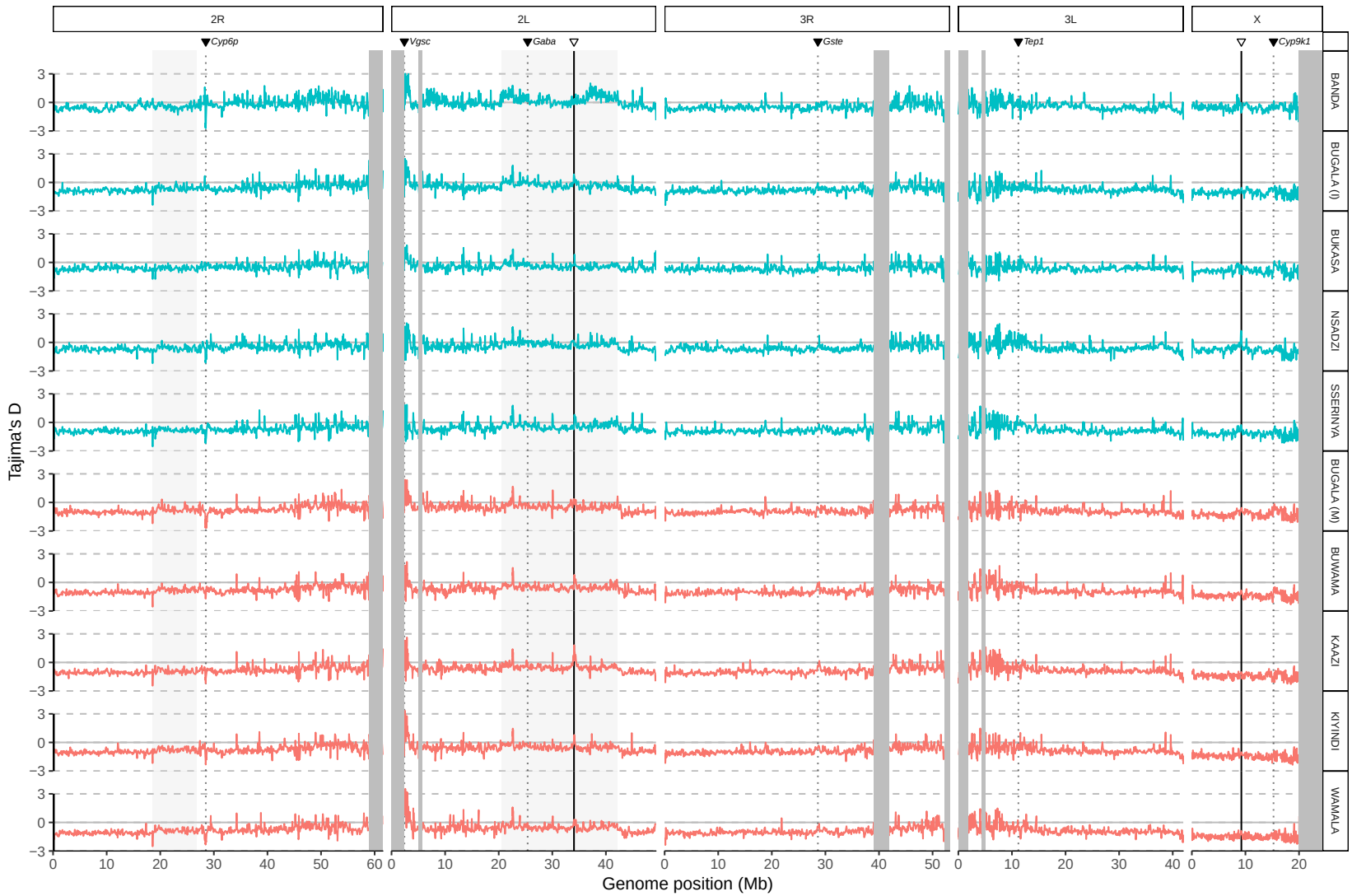

Figure S9: Tajima's D across genome.

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

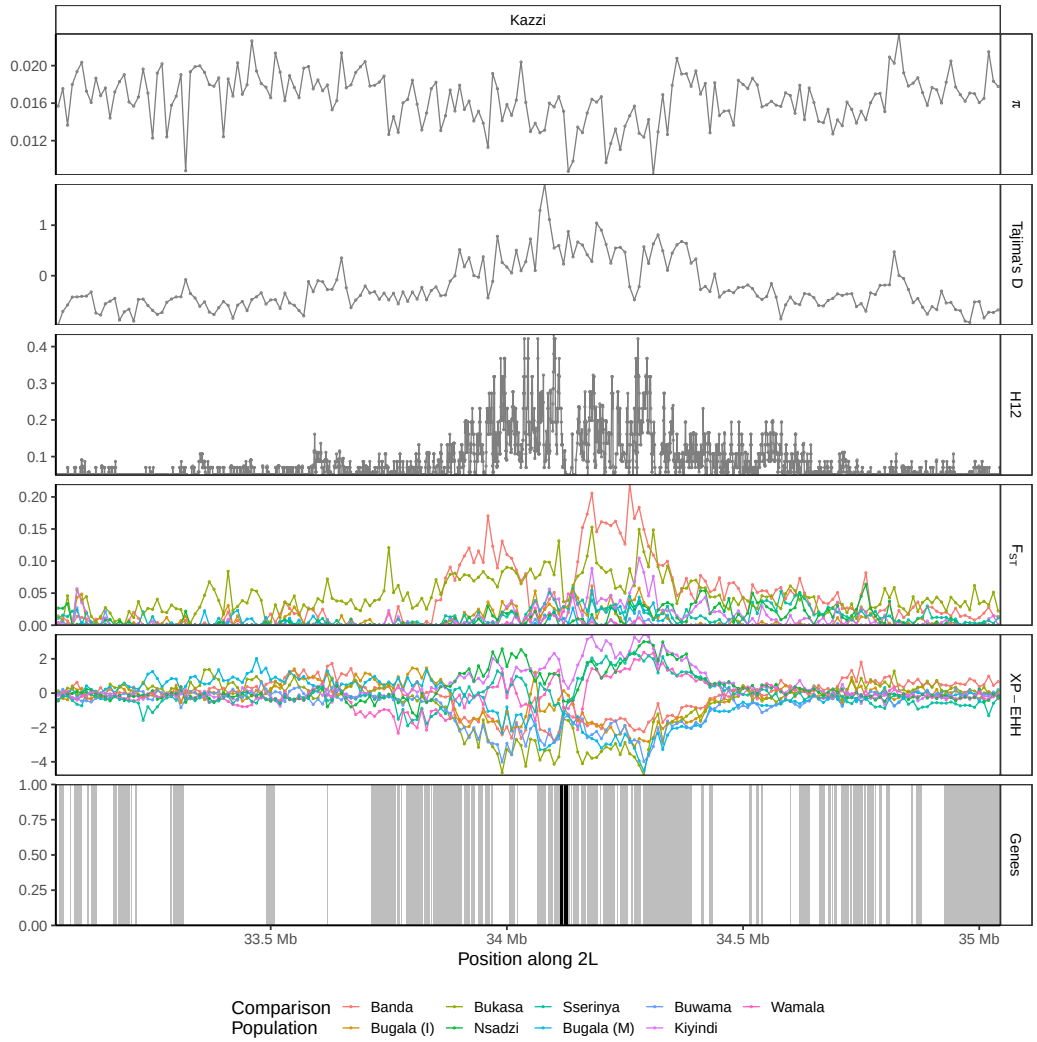

Figure S10: Selective sweep signal on chromosome 2L.
Population genetic statistics plotted near putative sweep on chromosome 2L. Focus population for all pairwise F_{ST} and XP-EHH comparisons is mainland site Kaazi. Region shown is 1 Mb up- and downstream of sweep target, centered at chr2L:34,044,820. Several genes involved in chorion formation (AGAP006549, AGAP006550, AGAP006551, AGAP006553, AGAP006554, AGAP006555 and AGAP006556) are shown in black.

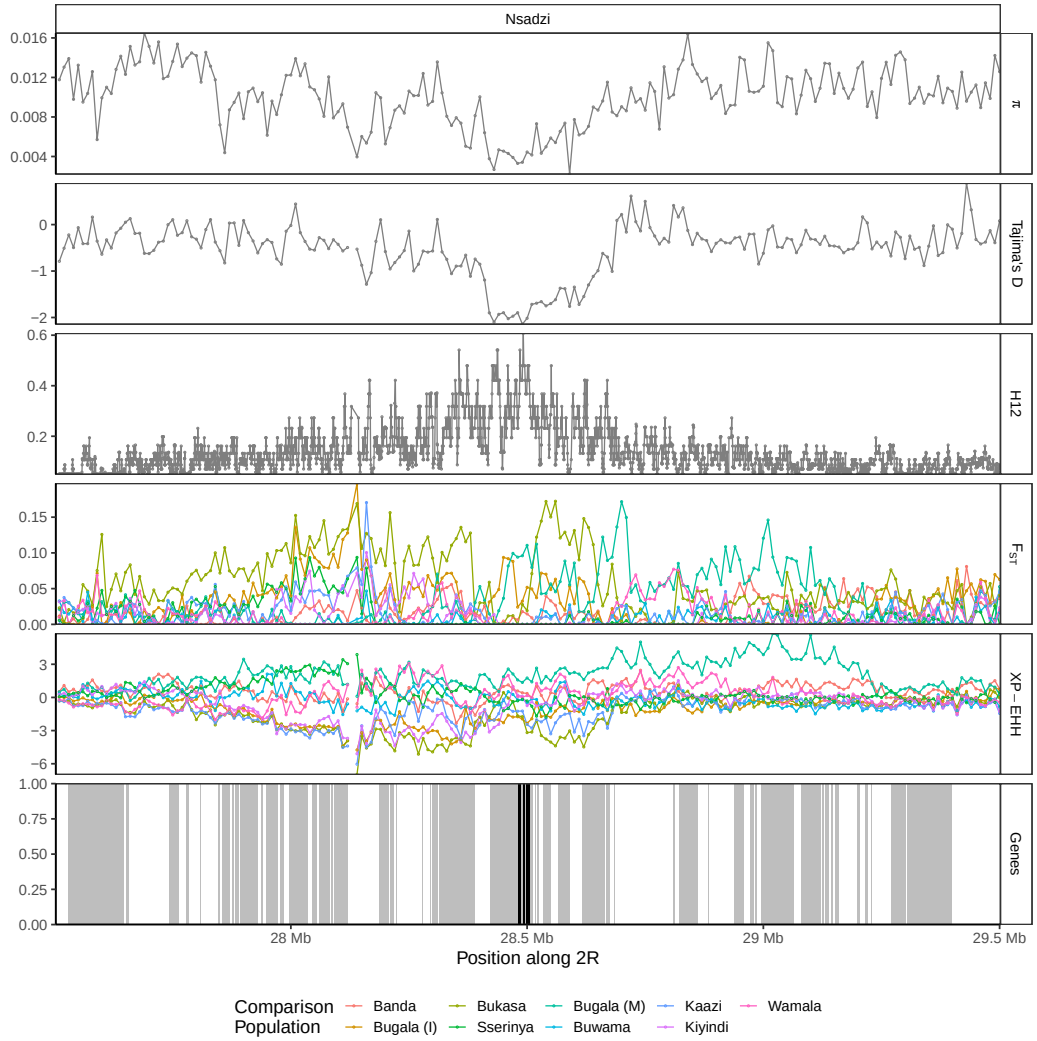

Figure S11: Selective sweep signal at *Cyp6* gene cluster on chromosome 2R. Population genetic statistics plotted near *Cyp6* gene cluster on chromosome 2R. Focus population for all pairwise F_{ST} and XP-EHH comparisons is island site Nsadzi. Region shown is 1 Mb up- and downstream of gene cluster, centered at chr2R:28,501,972, and *Cyp6* genes are highlighted in black.

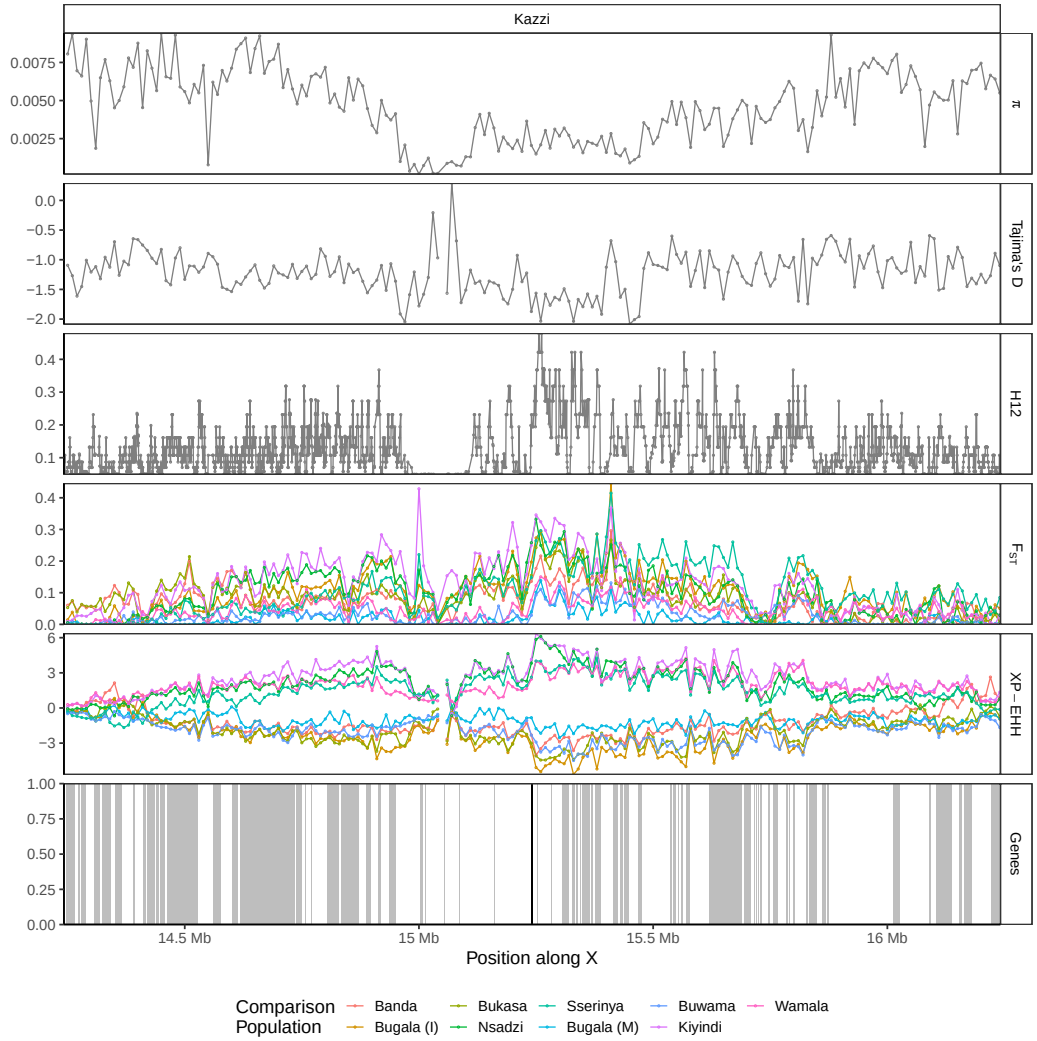

Figure S12: Selective sweep signal at gene *Cyp9K1* on X-chromosome. Population genetic statistics plotted near gene *Cyp9K1* on X-chromosome. Focus population for all pairwise F_{ST} and XP-EHH comparisons is mainland site Kaazi. Region shown is 1 Mb up- and downstream of gene cluster, centered at chrX:15,241,718, and *Cyp9K1* gene is highlighted in black.

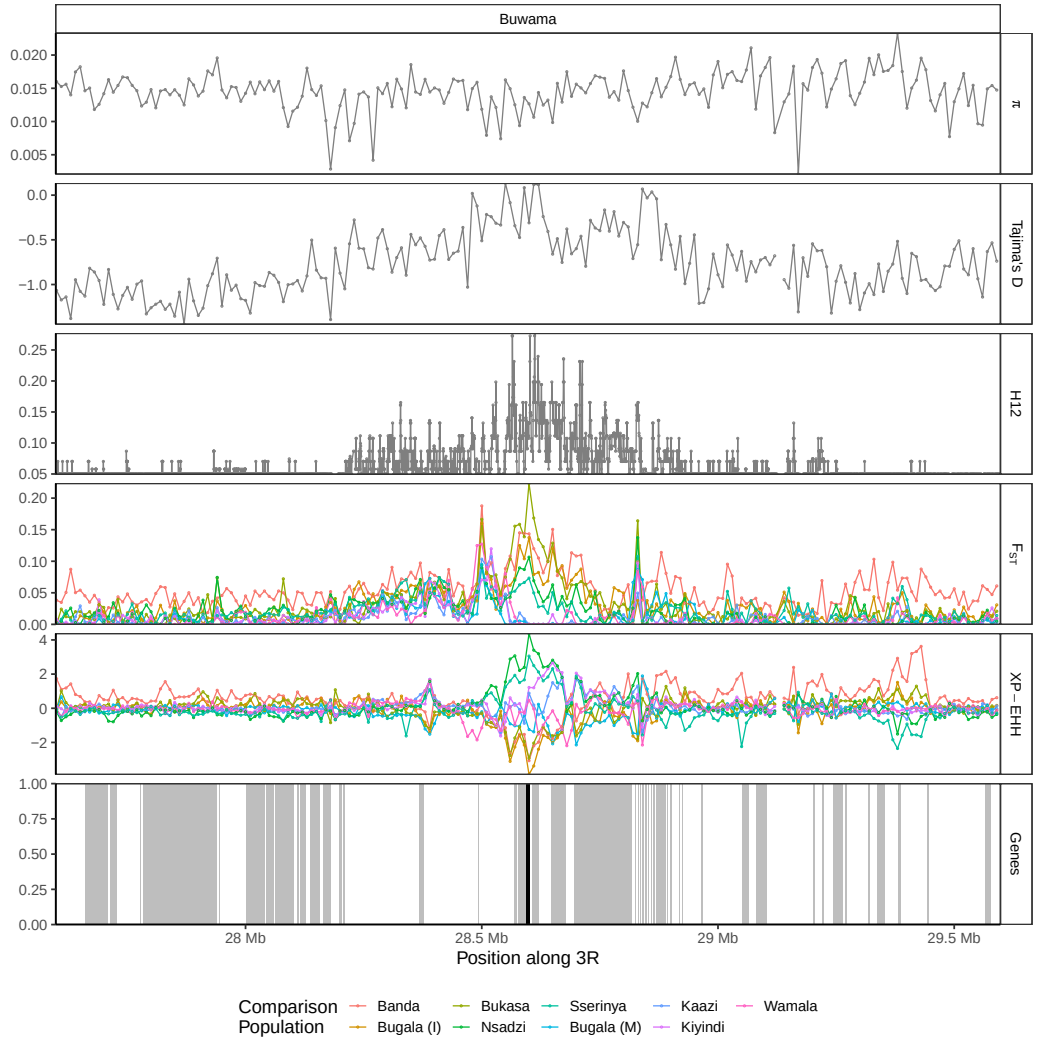

Figure S13: Selective sweep signal at *Gste* gene cluster on chromosome 3R. Population genetic statistics plotted near *Gste* gene cluster on chromosome 3R. Focus population for all pairwise F_{ST} and XP-EHH comparisons is mainland site Buwama. Region shown is 1 Mb up- and downstream of gene cluster, centered at chr3R:28,598,038, and *Gste* genes are highlighted in black.

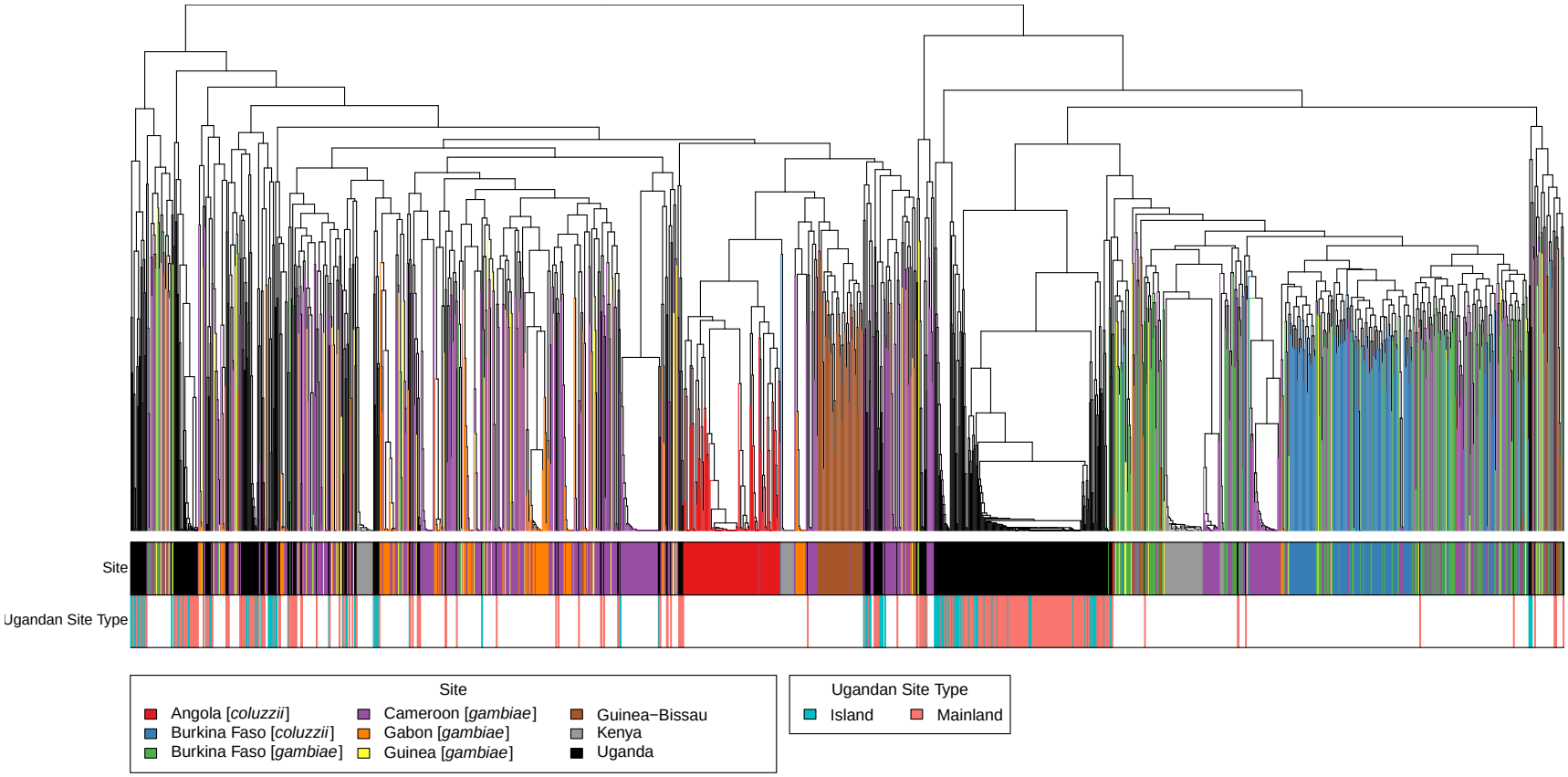

Figure S14: Tree for putative sweep on chromosome 2L.

Distance-based tree of haplotypes near putative sweep on chromosome 2L. Region shown is 100 kb up- and downstream of sweep target, centered at chr2L:34,044,820. 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.

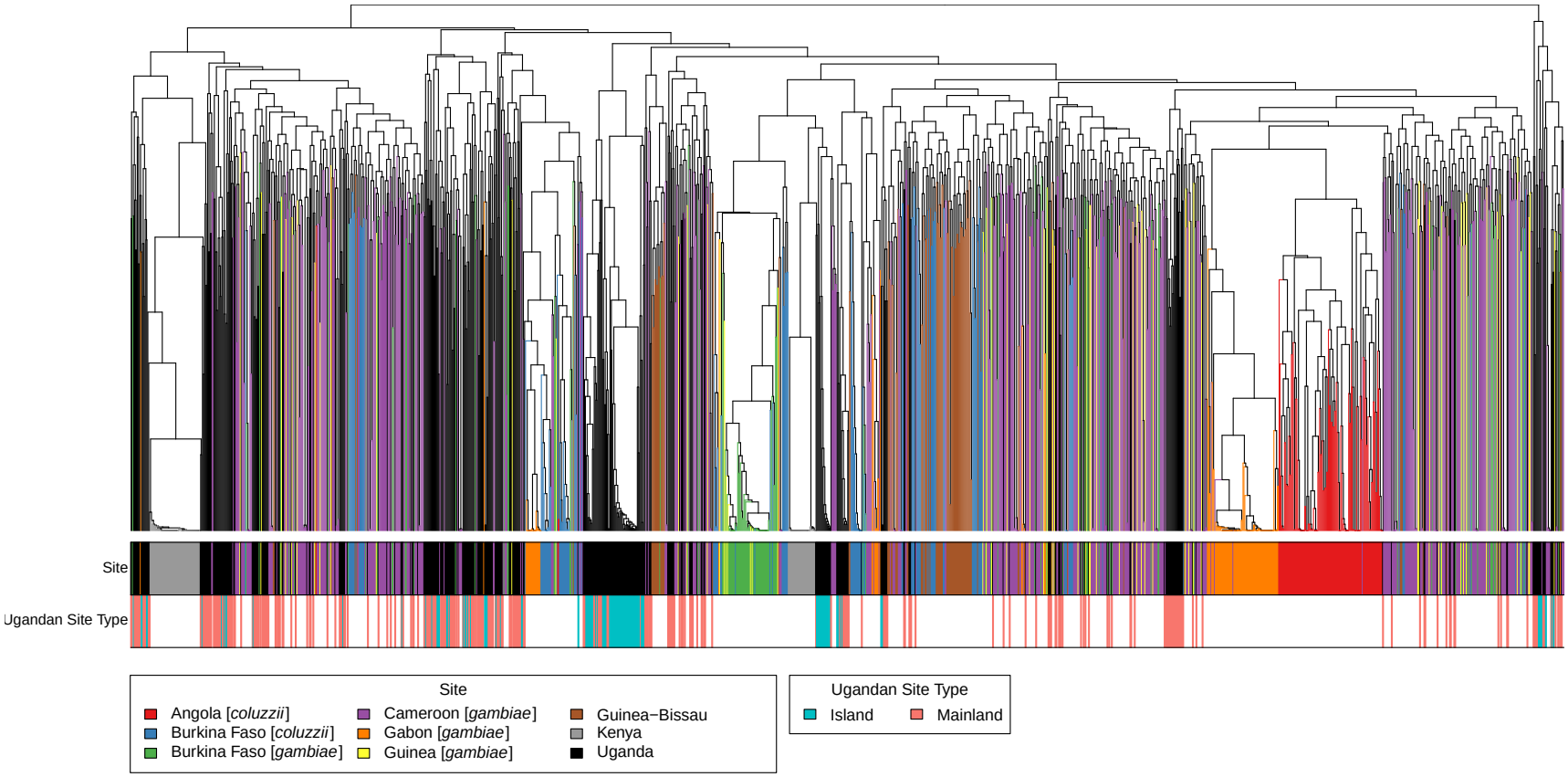

Figure S15: Tree for putative sweep on X-chromosome near *rdgA* ortholog.
Distance-based tree of haplotypes near putative sweep on X-chromosome. Region shown is 100 kb up- and downstream of sweep target, centered at chrX:9,238,942. Top color bar indicates locality, with all Ugandan individuals, from 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.

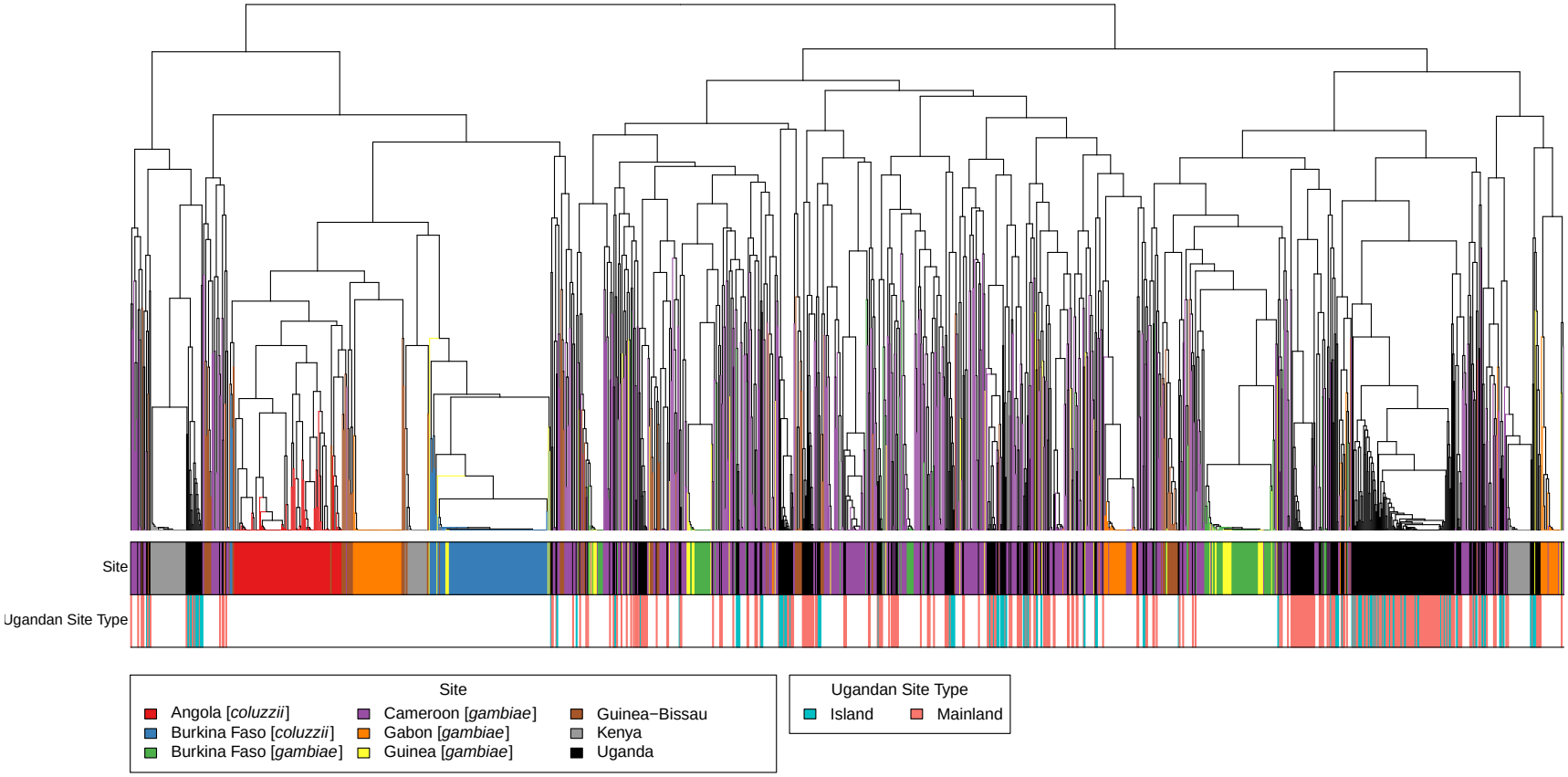

Figure S16: Tree for sweep at gene *Cyp9K1* on X-chromosome.

Distance-based tree of haplotypes near sweep at gene *Cyp9K1* on X-chromosome. Region shown is 100 kb up- and downstream of sweep target, centered at chrX:15,241,718. (Insufficient variants preclude inferring tree for region of width 20 kb.) 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.

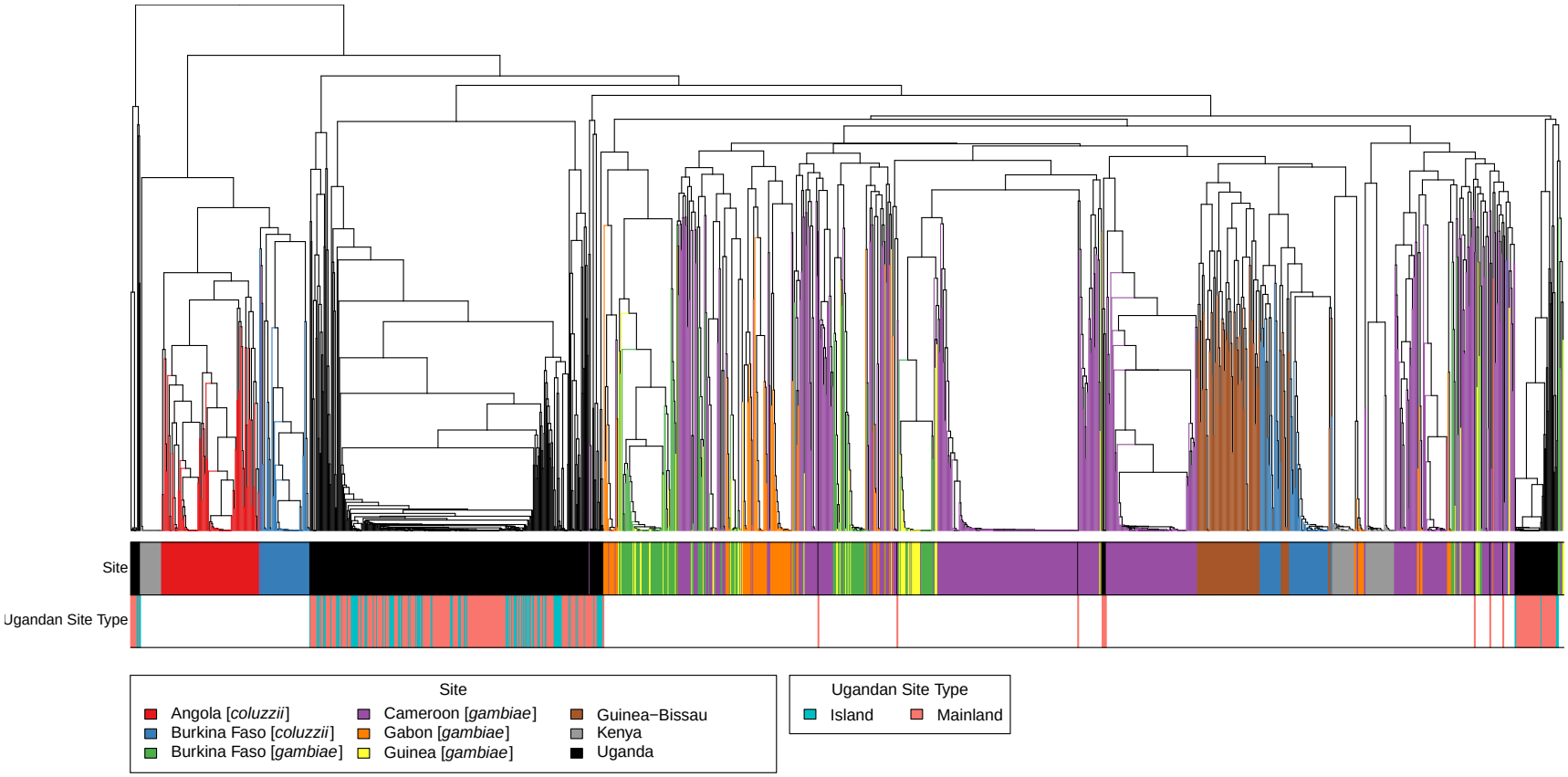

Figure S17: Tree for sweep on *Cyp6* gene cluster on chromosome 2R.

Distance-based tree of haplotypes near sweep at *Cyp6* gene cluster on chromosome 2R. Region shown is 100 kb up- and downstream of sweep target, centered at chr2R:28,501,972. 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.

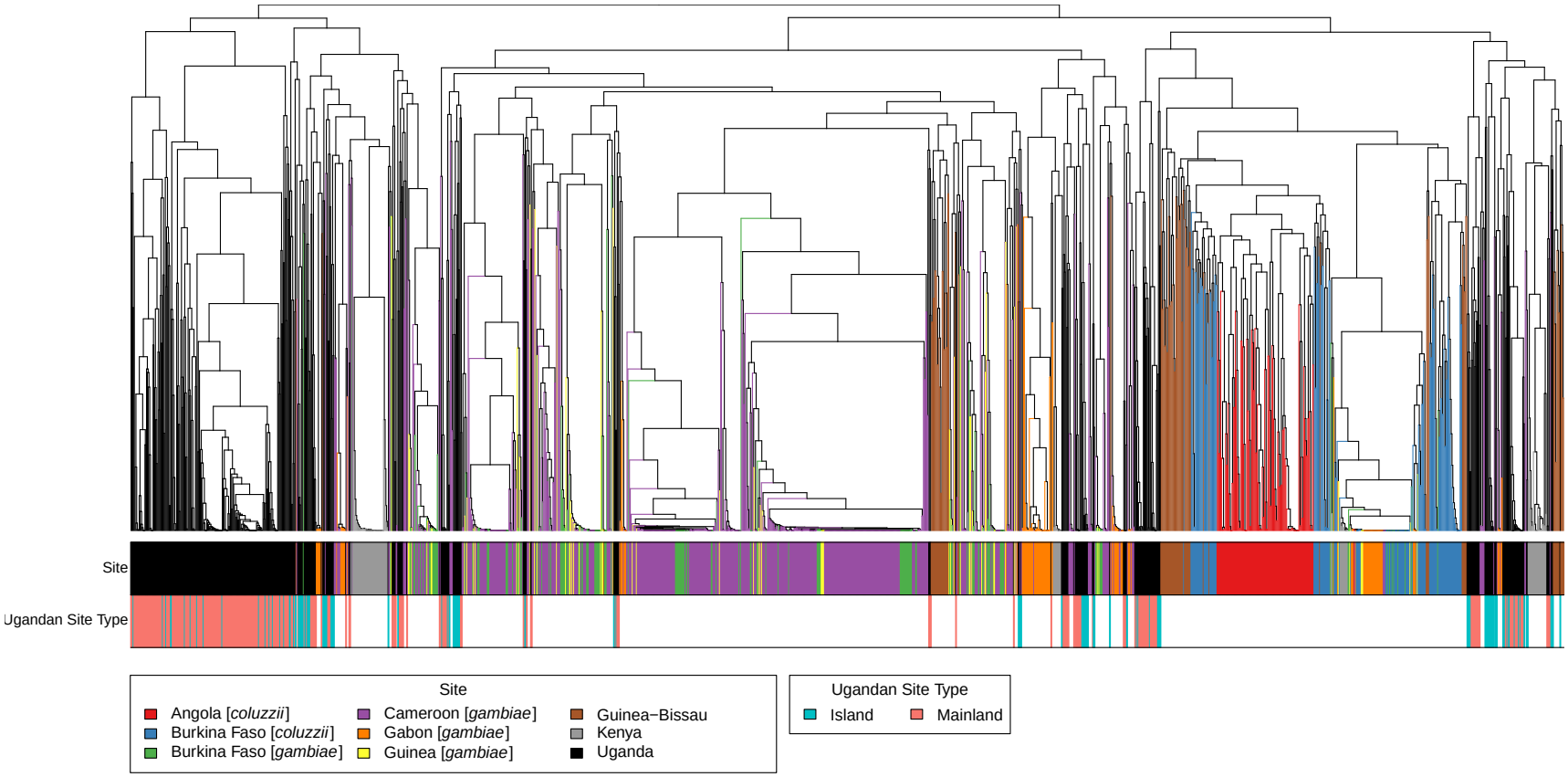

Figure S18: Tree for sweep on *Cyp6* gene cluster on chromosome 3R.

Distance-based tree of haplotypes near sweep at *Cyp6* gene cluster on chromosome 3R. Region shown is 100 kb up- and downstream of sweep target, centered at chr3R:28,598,038. 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.

References

[94] Sharakhova, M. V. *et al.* Update of the *Anopheles gambiae* PEST genome assembly. *Genome Biology* **8**, R5 (2007).

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

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

[97] DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43**, 491–498 (2011).

[98] Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* **81**, 559–75 (2007).

[99] Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).

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

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

[102] Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–9 (2009).

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

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

[105] Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–8 (2011).

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

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

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

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

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

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

[112] Neuwirth, E. *RColorBrewer: ColorBrewer Palettes* (2014).

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

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

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

[116] Price, A. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**, 904–9 (2006).

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

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