The origin of island populations of the African malaria mosquito, *Anopheles coluzzii*

Melina Campos[1], Mark Hanemaaijer[1], Hans Gripkey[1], Travis Collier[1], Yoosook Lee[1], Anthoney Cornel[1], Joao Pinto[2], Herodes Rompão[3], and Gregory Lanzaro[1]

[1]University of California Davis
[2]Universidade Nova de Lisboa
[3]Programa Nacional de Luta Contra o Paludismo

August 23, 2020

Abstract

Anopheles coluzzii is a major malaria vector throughout its distribution in west-central Africa. Here we present a whole-genome resequencing study of 77 specimens from eight localities, that covers a large part of this species' range, including three islands in the Gulf of Guinea: Bioko, São Tomé and Príncipe. Population genomic analyses encompassed structure of mainland populations, of island populations and connectivity between island and mainland populations. Three genetic clusters were found among mainland populations and genetic distances among all populations fit an isolation-by-distance model. Genomic analyses were applied to estimating the demographic history and ancestry (cross-coalescence) for each island. Taken together with the unique biogeography and history of human occupation for each island they present a coherent explanation underlying contemporary levels of genetic isolation between mainland and island populations and among island populations. We discuss the relation of our findings to the suitability of São Tomé and Príncipe islands as candidate sites for potential field trials of genetic-based malaria control strategies.

Keywords

gene flow, phylogeography, whole genome, remoteness, coalescence analysis

INTRODUCTION

From Darwin's early work to the present, oceanic islands have served as model systems for the development of evolutionary theory. Attributes such as small size, distinct boundaries and simplified biotas together with relative youth and geographical isolation make islands a centerpiece of scientific interest (Losos & Ricklefs, 2009). Island remoteness is an obvious barrier for migration and one of the key factors in the theory of island biogeography, which relates island size and distance from mainland with number of species (MacArthur & Wilson, 1967). Migration between related populations allows the exchange of heritable information and enhances genetic diversity, which is generally lower in island populations compared to mainland ones (Frankham, 1997).

From an applied perspective, geographically isolated sites are being considered for initial field trials of new genetic technologies applied to mosquito populations with the goal of malaria elimination. Confined field trials facilitate assessment of risk and effectiveness of genetically engineered mosquitoes (GEM) without the confounding influences of migration (A. A. James, 2005; S. James et al., 2018; Scott, Takken, Knols, & Boëte, 2002). Malaria is a life-threatening parasitic disease, that in 2018 resulted in an estimated 405,000 deaths of which 94% occurred in Africa (WHO, 2019). Anopheline mosquitoes are responsible for transmitting malaria parasites to humans. In Africa,*Anopheles gambiae* and *A. coluzzii* are among the principal vector

species (Coetzee et al., 2013; Sinka et al., 2012; Wiebe et al., 2017). Malaria elimination strategies and interventions greatly rely on vector control methods (WHO, 2017). However, modelling studies have shown that conventional vector control is insufficient for endemic malaria elimination (Griffin et al., 2010; Walker, Griffin, Ferguson, & Ghani, 2016), which reinforces the conclusion that new methods, which may include GEM, are urgently needed (Gantz et al., 2015; Hammond et al., 2016; Kyrou et al., 2018; Macias et al., 2020).

A thorough study of islands off the coast of Africa with the aim of identifying candidate sites for initial field trials of GEM has identified São Tomé and Príncipe (STP) as strong candidates (Vector Genetics Lab., 2018). This archipelago consists of two small oceanic islands in the Gulf of Guinea (West Africa), about 250 and 225 km, respectively, off the coast of Gabon. *Anopheles coluzzii* is thought to be the only malaria vector present on these islands (Moreno et al., 2007; Pinto et al., 2000). Previous studies have shown genetic isolation between *A. coluzzii* populations from São Tomé and Príncipe islands, as well as between the island and mainland (Marshall et al., 2008; Moreno et al., 2007; Pinto et al., 2000; Salgueiro, Moreno, Simard, O'Brochta, & Pinto, 2013) reinforcing the choice of STP as a suitable location for initial release of GEM. STP has recently reached the pre-elimination malaria level, due to the success of a combination of interventions, including indoor residual spraying, insecticide treated nets and artemisinin-based combination therapy (Chen et al., 2019; P. Lee, Liu, Rampao, Rosário, & Shaio, 2010; Teklehaimanot, Teklehaimanot, Kiszewski, Rampao, & Sachs, 2009). Sustainability of these malaria vector control methods are challenged by limited financial support and decreased mosquito susceptibility to insecticides (Chen et al., 2019).

Here we extend earlier studies describing genetic isolation between island and mainland *A. coluzzii* populations by applying analyses of 77 individual mosquito genomes. We present genome resequencing data from three island populations: São Tomé, Príncipe and Bioko (Equatorial Guinea); and continental populations from five African countries (Angola, Benin, Gabon, Mali and Cameroon). This sampling scheme covers the majority of *A. coluzzii* predicted distribution (Lehmann & Diabate, 2008). This is the first study using whole-genome sequencing to assess connectivity of conspecific populations on islands in the Gulf of Guinea with populations on the mainland. Results of this mosquito population genomics study are discussed in relation to the evolution of divergence among island and mainland populations. In addition, we consider our results as they relate to current and future vector control methods on the islands.

MATERIAL & METHODS

Population sampling

We used individual *A. coluzzii* specimens (N =77) archived at the Vector Genetics Laboratory, UC Davis, and at the Instituto de Higiene e Medicina Tropical (IHMT), Universidade Nova de Lisboa, Portugal for this study (Figure 1; Table S1). This set of specimens included samples from eight localities: three islands in the Gulf of Guinea (N = 3 from Bioko, N = 14 from São Tomé and$N$ = 17 from Príncipe), four Gulf of Guinea coastal mainland sites (N = 8 from Angola, N =11 from Benin, N =9 from Cameroon, and N =5 from Gabon) and one inland site (Mali$N$ =10). Species diagnostic was performed using species-specific markers included in the DIS assay as described in Y. Lee, Weakley, Nieman, Malvick, and Lanzaro (2015).

Whole genome sequencing

DNA from individual mosquitoes was extracted using a Qiagen Biosprint following our established protocol (Nieman, Yamasaki, Collier, & Lee, 2015). DNA yield was measured using a dsDNA high sensitivity assay kit on a Qubit instrument (Thermo Fisher Scientific, Waltham, MA, USA). The KAPA HyperPlus Kit (Roche Sequencing Solutions, Indianapolis, Indiana, USA) was used for individual genomic DNA libraries using 10 ng DNA as input, as described in Yamasaki et al. (2016). Size selection of the libraries and clean-up was performed using AMPure SPRI beads (Beckman Coulter Life Sciences, Indianapolis, Indiana, USA). Individual library concentrations were measured using Qubit and then pooled in equal amounts for sequencing using an Illumina HiSeq 4000 instrument at the UC Davis DNA Technologies Core facility.

Data processing, mapping and variant calling

After filtering and trimming demultiplexed raw reads using Trimmomatic v0.36 (Bolger, Lohse, & Usadel, 2014), reads were mapped to the reference AgamP4 (Holt et al., 2002; Sharakhova et al., 2007) using BWA-MEM v0.7.15 (Li, 2013) with default settings. Duplicate reads were removed using Sambamba markdup (Tarasov, Vilella, Cuppen, Nijman, & Prins, 2015). Freebayes v1.2.0 (Garrison & Marth, 2012) was used for variant calling, with standard filters and the "-no-population-priors", "theta = 0.01", and "max-comple-gap = 0" options. Variants were normalized with *vt normalize* v0.5 (Tan, Abecasis, & Kang, 2015) and those without support from both overlapping forward and reverse reads were removed using vcffilter v1.0.0rc2 (https://github.com/vcflib/vcflib). Only biallelic SNPs with minimum depth of 8 and maximum of 2% of missing data were used for further analysis.

Mitochondria

Mitogenome variant calling were generated assuming single ploidy using Freebayes v1.2.0. Singletons and SNPs in an AT-rich region were removed from further analysis due to low coverage (Hanemaaijer et al., 2018). A neighbor-joining tree was constructed with Nei's distance matrix and 1,000 bootstrap replicates using the R package *ape* 5.4 (Paradis & Schliep, 2019).

Genetic structure of populations

Analysis of population structure was based on chromosome 3 SNPs only. This was done to avoid confounding signals from polymorphic inversions on chromosomes 2 and X (Sharakhova et al., 2007). Heterochromatic regions on chromosome 3 were also filtered out (Sharakhova et al., 2007). After this filtering, pruning for LD and singletons was performed for principal component analysis (PCA) using scikit-allel v1.2.0 (Alistair & Harding, 2017). Hudson's estimator (Bhatia, Patterson, Sankararaman, & Price, 2013; Hudson, Slatkin, & Maddison, 1992) was used for pairwise fixation indices F_{ST}calculation implemented in scikit-allel v1.2.0. TheseF_{ST} estimates were used for correlation between genetic and geographic distances with a Mantel test (Mantel, 1967) using the R package ade4 (Dray & Dufour, 2007). Population structure was also explored by assignment of individual genomes to ancestry components using ADMIXTURE v1.3.0 (Alexander, Novembre, & Lange, 2009). A total of three replicate samples of 100,000 SNPs from chromosome 3 were submitted for admixture analysis. For each replicate, five iterations were performed for values of K clusters from 1 to 15. The results were compiled using the online version of CLUMPAK and plotted in R. Best-fitting K was determined by the lowest cross-validation value based on the 15 ADMIXTURE runs.

Nucleotide diversity (π) and Tajima's D were calculated in nonoverlapping windows of 10 kb on autosomal chromosomes using VCFtools (Danecek et al., 2011). VCFtools was also used for the calculation of inbreeding statistics (F_{IS}) using the method-of-moments approach. Runs of homozygosity (ROH) were inferred outside inversions and heterochromatic regions and LD-pruned SNPs set using PLINK v1.9 (Purcell et al., 2007). The results were grouped by population and significance tests performed between the islands of São Tomé and Príncipe and the remaining populations under study using a Wilcoxon test for the means in R.

Population sizes and cross-coalescence analysis

Population size estimation and cross-coalescence analysis were performed using the multiple sequentially Markovian coalescent pipeline MSMC2 v2.0.2 (Schiffels & Durbin, 2014) following the author's protocol (https://github.com/stschiff/msmc2). For this analysis, SNPs were phased with SHAPEIT2 v2.9 (Delaneau, Marchini, Genomes Project, & Genomes Project, 2014) using an *A. gambiae* recombination map (Anopheles gambiae 1000 Genomes et al., 2017). Four samples per population (except Bioko; N =3) were randomly selected and used for population size and two per population for cross-coalescence inter-population, as described in Schmidt et al., 2019. The results were plotted in R, assuming 10 generations per year and mutation rate of 2.85 x 10^{-9} (median between mutation rates in insects as used in Schmidt et al. (2019).

RESULTS

Mosquito sampling and sequencing

We sequenced the genomes of 77 *Anopheles coluzzii* samples originating from island or mainland populations.

In total 2.3 billion reads were sequenced with a mean genome coverage of 11.3× per sample (Table S1). On average, 94.6% of reads were mapped to the reference genome AgamP4. We identified 19.9 million high-quality biallelic single-nucleotide polymorphisms (SNPs) across the genome.

Population structure

We investigated the genetic structure of *A. coluzzii* from island and mainland populations in West and Central Africa (Figure 1). Only biallelic SNPs on chromosome 3 were used to avoid confounding factors from common paracentric inversions on other chromosomes. Both principal component and Bayesian clustering analyses suggest that populations on São Tomé and Príncipe Islands (STP) are genetically differentiated from mainland populations (Figure 2a, b). On the other hand, Bioko Island clusters with mainland populations from Gabon and Cameroon (Figure 2a). With $K = 2$ (lower cross-validation value for $1 < K < 15$, Figure S1), STP samples form a unique cluster distinct from mainland and Bioko samples which are contained within two additional groups. When K is set to 5, mainland populations plus Bioko Island are distributed among three genetic clusters which are differentiated from São Tomé and Príncipe samples (Figure 2). Among mainland populations, a geographic latitudinal clustering was found i.e. Mali and Benin forming a north-western group, followed by Cameroon, Gabon and Bioko Island in central Africa, and a third cluster consisting of the samples from Angola (Figure 2a, b). The same clustering pattern was derived at using biallelic SNPs on the mitochondrial genome of these samples (Figure S2).

Mean F_{ST} between STP populations and mainland populations was significantly higher than F_{ST}among mainland populations only (Figure 2c, d). Príncipe island presented the highest F_{ST} values (Figure 2c). Mantel tests for geographic distances and F_{ST}were uncorrelated if all population comparisons were included (p= .81; Figure S3) but significantly correlated when STP were excluded (p = .005; Figure S3).

Population diversity

Mean nucleotide diversity (π) was measured over the whole genome of*A. coluzzii.* Specimens from STP populations carried significantly less diversity compared with mainland populations (Figure 3a). Tajima's D statistic for *A. coluzzii* sequence from STP was D > 0, indicating a scarcity of rare alleles consistent with population bottleneck at the initial introduction of this species into the islands (founder event) and continued maintenance of small population size relative to populations on the mainland. STP populations also showed longer runs of homozygosity (F_{ROH} ; Fig. 3c) and higher inbreeding statistics (F_{IS} ; Fig. 3d). These results are consistent with the hypothesis of reduced genetic diversity in island populations due to inbreeding and smaller population sizes. Bioko island was grouped with mainland populations, consistent with it's geological history and suggesting continued contact with mainland populations.

Population demographic history and cross-coalescence

A reconstruction of the demographic history of *A. coluzzii* was created using Multiple sequentially Markovian coalescence (MSMC) analysis applied to our genome sequence data. Prior to about 900,000 years ago effective population size of a putative ancestral population was relatively large (N_e˜10^7) and stable (Fig. 4A). From that point forward there appear numerous splits from a central Africa group (Bioko, Cameroon, Gabon, São Tomé and Príncipe) with different population groups following distinct demographic trajectories. The first of these splits appears to have occurred at roughly the same time (˜1M years ago) forming western populations (Mali and Benin) and a distinct population in Angola. Inland western populations experienced initial size expansion subsequentially stabilizing and remaining relatively large. The Angola population followed a declining trajectory stabilizing at an intermediate level about 50,000 years ago. Populations of *A. coluzzii* became established in São Tomé and Príncipe around ˜90,000 years ago experiencing a dramatic population bottleneck in the process of colonizing the islands (founder effect). A second split at roughly the same time distinguished Cameroon from Bioko and Gabon. While Bioko and Gabon population sizes continually decreased until the most recent reliable estimation (˜30,000 years ago; gray bar Fig 4a), Cameroon increased reaching similar levels to west Africa.

Shared history of populations was estimated for pairs using cross-coalescence. A higher relative cross-

coalescence (RCC), indicates less time to the last common ancestor shared by the two populations in a specific comparison (Figure 4b). Heuristically, RCC < 0.5 (mid-point) denotes an isolation between populations. All populations shared common ancestors in the deep past, reflecting high connectivity. About ˜200,000 years ago, RCC between west African populations (Mali and Benin) and all other populations started decreasing considerably, reaching the mid-point or below (RCC < 0.4 for Angola comparisons) (Figure 4b). RCC between São Tomé and Príncipe populations and Angola and central Africa populations declines over time. The islands became isolated about ˜90,000 years ago (RCC < 0.3) (Figure 4c-top). However, high RCC was found between São Tomé and Príncipe, comparable with the pair Mali and Benin, suggesting some degree of historical gene flow (Figure 4b-bottom).

DISCUSSION

Dispersal of malaria vector species has been extensively explored because it directly affects disease transmission, the spread of insecticide resistance and strategies for controlling mosquitoes (Clarkson et al., 2018; Huestis et al., 2019). Mosquito dispersal can be measured by conventional mark–recapture experiments for short range (Service, 1997; Touré et al., 1998), directly by air-borne insect sampling for long distances (Huestis et al., 2019) or through estimation of gene flow between populations applied at various scales. Here we describe important aspects of *A. coluzzii* dispersal and historical phylogeography using a population genomics approach. This is the first whole-genome resequencing study covering a large part of this species' range focusing on island as well as mainland populations (Calzetta et al., 2008; della Torre, Tu, & Petrarca, 2005).

Mainland Populations

Anopheles coluzzii samples from mainland populations were consistently divided into three geographically related genetic clusters: i) Mali and Benin in West Africa; ii) Cameroon and Gabon in Central Africa and iii) Angola (Figure 2a; Figure 2b). *Anopheles coluzzii*overlaps with its sister species *A. gambiae* over 90% of its geographical range, which comprises Central and West Africa (Lehmann & Diabate, 2008). However, in contrast to the shallow population structure found in *A. gambiae* across the African continent (Anopheles gambiae 1000 Genomes et al., 2017; Lehmann et al., 2003; Weetman, Wilding, Steen, Pinto, & Donnelly, 2012), in this study *A. coluzzii* populations presented positive isolation by distance (Figure S3b), corroborating previous reports (Clarkson et al., 2020; Pinto et al., 2013). In the Sahel, where there is a pronounced dry season, recent studies have suggested that *A. coluzzii* persists through dry seasons via dormancy (aestivation), whereas *A. gambiae*populations experience local extinctions followed by reestablishment via long-distance migration (Arcaz et al., 2016; Dao et al., 2014; Lehmann et al., 2017). In addition, *A. gambiae* has a far broader geographical distribution (across most of sub-Saharan Africa). These observations suggest that *A. gambiae* has a greater capacity for dispersal compared with its sister species *A. coluzzii* .

Based on our analyses of historical population size and cross-coalescence, we hypothesize that, like *A. gambiae* (Schmidt et al. 2019) the geographical origin of *A. coluzzii* was from a west African ancestral population represented by the current Mali and Benin groups (Figure 4). West African and Cameroonian populations have no sign of strong historical fluctuations in population size, whereas Gabon and Angola populations each experienced a decrease in effective population size. Concerning the cross-coalescence analysis, we observed a split that occurred ˜200,000 years ago separating west African populations (Mali and Benin) firstly from Angola, then from all others (RCC$<$ .5; Figure 4), consistent with observations of*A. gambiae* which were correlated with the Congo River basin as a geological barrier to dispersal (Schmidt et al., 2019; Voelker et al., 2013).

Island Populations

Anopheles coluzzii from São Tomé and Príncipe islands presented patterns typical for remote oceanic island populations i.e. reduced genetic diversity, signs of inbreeding and low population size, whereas Bioko island population was similar to those on the mainland (Figure 3; Figure 4a). All three are located in the Gulf of Guinea as part of the chain of volcanoes of the Cameroonian line (Burke, 2001). Bioko is only 32 km off the coast of Gabon whilst São Tomé and Príncipe are 250 and 225 km from Gabon respectively. Beyond simply distance from the African mainland, important biogeographic aspects differentiate Bioko from São Tomé and

Príncipe islands and these are reflected by the biology of the organisms inhabiting these islands.

Bioko is a land-bridge island, lying on the continental shelf in shallow seas only 60m deep (Jones, 1994; Juste & Ibanez, 1994). Sea levels were historically lowered sufficiently to connect Bioko to the Africa mainland during the last glaciation (Jones, 1994). In contrast, São Tomé and Príncipe are oceanic islands that have never been connected with the mainland nor with each other, and are separated by seas over 1,800m deep (Jones, 1994). Reflecting its continental origin, Bioko's fauna and flora are relatively species-rich, but with low levels of species endemism, suggesting its former connection to the mainland and short period of isolation (Jones, 1994; Juste & Fa, 1994; Juste & Ibanez, 1994). São Tomé and Príncipe islands present an inverted pattern i.e. high endemism, that includes mosquito' species (Jones, 1994; Loiseau et al., 2019), and low species richness. In STP islands only two species of anopheline mosquitoes have been reported, *A. coluzzii* and*A. coustani* (Pinto et al., 2000), whilst on Bioko there are at least five: *A. gambiae s.s* (hereafter *A. gambiae*),*A. coluzzii* , *A. melas* , *A. funestus* , *A. smithii* (Molina et al., 1993; Reimer et al., 2005).

Relationship Between Island and Mainland Populations

Regarding connectivity between island and mainland populations, we found strong evidence that *A. coluzzii* from STP islands are isolated from mainland populations. While samples from Bioko are closely related to central African populations from Cameroon and Gabon. These results were supported by population structure analysis using PCA (Figure 2a), admixture (Figure 2b) and pairwise F_{ST} (Figure 2c). Considerably higher genetic divergence was found between São Tomé or Príncipe island populations and those on the mainland than among mainland populations ($p<0.001$; Figure 2d). In addition, divergence was high between the two islands ($F_{ST}= 0.11$); and admixture analysis ($K =5$) assigned each island to a distinct genetic cluster (Figure 2b).

The results we report here are consistent with and extend earlier work on the genetic structure of mainland and island populations of *A. coluzzii* around the Gulf of Guinea. This earlier work described the genetic structure of populations using microsatellite markers (Moreno et al., 2007; Pinto et al., 2002), mitochondrial ND5, rDNA internal transcribed spacer sequences (Marshall et al., 2008) and transposable element (*Herves*) insertion site polymorphisms (Salgueiro et al., 2013). Collectively these works revealed genetic isolation between populations in STP and Gabon and little isolation between populations on Bioko and the mainland.

Our analysis shows clear isolation of STP *A. coluzzii* populations from those on the African mainland and suggests that these populations were established about ˜90,000 years ago (RCC $< .3$). Analysis of populations on Bioko indicate recent (RCC $> .6$; ˜30,000 years ago; Figure 4b) and perhaps contemporary gene flow with those in Cameroon and Gabon. This observation agrees with the geological history of Bioko which became isolated from the mainland by rising sea levels only ˜11,000 years ago (Jones, 1994). São Tomé and Príncipe island populations experienced a sharp decrease in size, suggesting that a small portion of the ancestral population became established there (founder effect), whereas the population size trajectory on Bioko is similar to the mainland (Figure 4a). Of note, exact time in years could change if the assumptions for mutation rate and number of generations per year are revised, however relative separation remain relevant.

São Tomé and Príncipe as Sites for Novel GEM Mosquito Field Trials

In this study we used a population genomics analysis to explore the relationship between malaria mosquito populations on the remote oceanic islands of São Tomé and Príncipe with mainland populations bordering the Gulf of Guinea in west Africa. Our results are consistent with studies of mosquitoes on other oceanic islands (Schmidt et al., 2019; Clarkson et al., 2020). Similar work analyzing anopheline mosquitoes on lacustrine islands in Lake Victoria suggest a much lower degree of genetic isolation (Bergey et al., 2020), which is not surprising considering their geographic proximity to the mainland.

Population genetic studies are vital for improving the design and organization of vector control strategies, including field trials of GEM mosquitoes. Genetic control methods offer potential for low-cost, sustainable malaria elimination in highly endemic areas where conventional methods have shown to be insufficient (Griffin et al., 2010; Walker et al., 2016). In order to best evaluate the performance of modified mosquitoes, a confined

field trial site is required, defined by minimal gene flow between neighboring populations. Here we show that populations of *A. coluzzii* from the islands of São Tomé and Príncipe are genetically isolated, both from each other and from nearest mainland populations. Previous studies have reported isolation of these islands using fewer genetic markers (Pinto et al., 2002; Marshall et al., 2008; Salgueiro et al., 2013). We have expanded this work by analyzing individual mosquito whole genome sequences from a wide range of *A. coluzzii* populations on the mainland for comparison. Future work on populations of malaria vectors in São Tomé and Príncipe will focus on the relationship among *A. coluzzii* sub-populations within each island.

ACKNOWLEDGEMENTS

This work was supported by grants from the UC Irvine Malaria Initiative Program, Open Philanthropy and NIH R56 grant (R56AI130277). We thank the National Malaria Control Program personnel from São Tomé and Príncipe and, Ministry of Health in São Tomé and Príncipe who facilitated our field collections in São Tomé.

REFERENCES

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res, 19* (9), 1655-1664. doi:10.1101/gr.094052.109

Alistair, M., & Harding, N. (2017). cggh/scikit-allel: v1.2.0 (Version v1.2.0). *Zenodo* .

Anopheles gambiae 1000 Genomes, C., Data analysis, g., Partner working, g., Sample, c.-A., Burkina, F., Cameroon, . . . Project, c. (2017). Genetic diversity of the African malaria vector Anopheles gambiae. *Nature, 552* (7683), 96-100. doi:10.1038/nature24995

Arcaz, A. C., Huestis, D. L., Dao, A., Yaro, A. S., Diallo, M., Andersen, J., . . . Lehmann, T. (2016). Desiccation tolerance in Anopheles coluzzii: the effects of spiracle size and cuticular hydrocarbons. *J Exp Biol, 219* (Pt 11), 1675-1688. doi:10.1242/jeb.135665

Bergey, C. M., Lukindu, M., Wiltshire, R. M., Fontaine, M. C., Kayondo, J. K., & Besansky, N. J. (2020). Assessing connectivity despite high diversity in island populations of a malaria mosquito. *Evol Appl, 13* (2), 417-431. doi:10.1111/eva.12878

Bhatia, G., Patterson, N., Sankararaman, S., & Price, A. L. (2013). Estimating and interpreting FST: the impact of rare variants. *Genome Res, 23* (9), 1514-1521. doi:10.1101/gr.154831.113

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. . *Bioinformatics* .

Burke, K. (2001). Origin of the Cameroon line of volcano-capped swells. *The Journal of Geology* .

Calzetta, M., Santolamazza, F., Carrara, G. C., Cani, P. J., GFortes, F., Di Deco, M. A., . . . Petrarca, V. (2008). Distribution and Chromosomal Characterization of the Anopheles gambiae Complex in Angola. *Am J Trop Med Hyg* .

Chen, Y. A., Lien, J. C., Tseng, L. F., Cheng, C. F., Lin, W. Y., Wang, H. Y., & Tsai, K. H. (2019). Effects of indoor residual spraying and outdoor larval control on Anopheles coluzzii from São Tomé and Príncipe, two islands with pre-eliminated malaria. *Malar J, 18* (1), 405. doi:10.1186/s12936-019-3037-y

Clarkson, C. S., Miles, A., Harding, N. J., Lucas, E. R., Battey, C. J., Amaya-Romero, J. E., . . . Kwiatkowski, D. P. (2020). Genome variation and population structure among 1,142 mosquitoes of the African malaria vector species Anopheles gambiae and Anopheles coluzzii. *bioRxiv preprint doi: https://doi.org/10.1101/864314* . doi:10.1101/864314

Clarkson, C. S., Miles, A., Harding, N. J., Weetman, D., Kwiatkowski, D., Donnelly, M., & Consortium, T. A. g. G. (2018). The genetic architecture of target-site resistance to pyrethroid insecticides in the African malaria vectors Anopheles gambiae and Anopheles coluzzii. *bioRxiv* . doi:10.1101/323980

Coetzee, M., Hunt, R. H., Wilkerson, R., Torre, A. D., Coulibaly, M. B., & Besansky, N. J. (2013). Anopheles coluzzii and Anopheles amharicus, new members of the Anopheles gambiae complex. *Zootaxa, 3619* (3), 246-274. doi:10.11646/zootaxa.3619.3.2

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., . . . Genomes Project Analysis, G. (2011). The variant call format and VCFtools. *Bioinformatics, 27* (15), 2156-2158. doi:10.1093/bioinformatics/btr330

Dao, A., Yaro, A. S., Diallo, M., Timbine, S., Huestis, D. L., Kassogue, Y., . . . Lehmann, T. (2014). Signatures of aestivation and migration in Sahelian malaria mosquito populations. *Nature, 516* (7531), 387-390. doi:10.1038/nature13987

Delaneau, O., Marchini, J., Genomes Project, C., & Genomes Project, C. (2014). Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nat Commun, 5* , 3934. doi:10.1038/ncomms4934

della Torre, A., Tu, Z., & Petrarca, V. (2005). On the distribution and genetic differentiation of Anopheles gambiae s.s. molecular forms.*Insect Biochem Mol Biol, 35* (7), 755-769. doi:10.1016/j.ibmb.2005.02.006

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

Frankham, R. (1997). Do island populations have less genetic variation than mainland populations? *Heredity (Edinb), 78* .

Gantz, V. M., Jasinskiene, N., Tatarenkova, O., Fazekas, A., Macias, V. M., Bier, E., & James, A. A. (2015). Highly efficient Cas9-mediated gene drive for population modification of the malaria vector mosquito Anopheles stephensi. *Proc Natl Acad Sci U S A, 112* (49), E6736-6743. doi:10.1073/pnas.1521077112

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. . *arXiv preprint arXiv:1207.3907 [q-bio.GN].*

Griffin, J. T., Hollingsworth, T. D., Okell, L. C., Churcher, T. S., White, M., Hinsley, W., . . . Ghani, A. C. (2010). Reducing Plasmodium falciparum malaria transmission in Africa: a model-based evaluation of intervention strategies. *PLoS Med, 7* (8). doi:10.1371/journal.pmed.1000324

Hammond, A., Galizi, R., Kyrou, K., Simoni, A., Siniscalchi, C., Katsanos, D., . . . Nolan, T. (2016). A CRISPR-Cas9 gene drive system targeting female reproduction in the malaria mosquito vector Anopheles gambiae. *Nat Biotechnol, 34* (1), 78-83. doi:10.1038/nbt.3439

Hanemaaijer, M. J., Houston, P. D., Collier, T. C., Norris, L. C., Fofana, A., Lanzaro, G. C., . . . Lee, Y. (2018). Mitochondrial genomes of Anopheles arabiensis, An. gambiae and An. coluzzii show no clear species division. *F1000Res, 7* , 347. doi:10.12688/f1000research.13807.2

Holt, R. A., Subramanian, G. M., Halpern, A., Sutton, G. G., Charlab, R., Nusskern, D. R., . . . Loftus, B. (2002). The Genome Sequence of the Malaria Mosquito Anopheles gambiae. *Science* .

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

Huestis, D. L., Dao, A., Diallo, M., Sanogo, Z. L., Samake, D., Yaro, A. S., . . . Lehmann, T. (2019). Windborne long-distance migration of malaria mosquitoes in the Sahel. *Nature, 574* (7778), 404-408. doi:10.1038/s41586-019-1622-4

James, A. A. (2005). Gene drive systems in mosquitoes: rules of the road. *Trends in Parasitology* .

James, S., Collins, F. H., Welkhoff, P. A., Emerson, C., Godfray, H. C. J., Gottlieb, M., . . . Touré, Y. T. (2018). Pathway to deployment of gene drive mosquitoes as a potential biocontrol tool for elimination of malaria in sub-aharan Africa: recommendations of a scientific working group *Am J Trop Med Hyg* .

Jones, P. J. (1994). Biodiversity in the Gulf of Guinea: an overview.*Biodiversity and Conservation* .

Juste, J. B., & Fa, J. E. (1994). Biodiversity conservation in the Gulf of Guinea islands: taking stock and preparing action. *Biodiversity and Conservation* .

Juste, J. B., & Ibanez, C. (1994). Bats of the Gulf of Guinea islands: fauna1 composition and origins. *Biodiversity and Conservation* .

Kyrou, K., Hammond, A. M., Galizi, R., Kranjc, N., Burt, A., Beaghton, A. K., . . . Crisanti, A. (2018). A CRISPR-Cas9 gene drive targeting doublesex causes complete population suppression in caged Anopheles gambiae mosquitoes. *Nat Biotechnol, 36* (11), 1062-1066. doi:10.1038/nbt.4245

Lee, P., Liu, C., Rampao, H. S., Rosário, V. E., & Shaio, M. (2010). Pre-elimination of malaria on the island of Príncipe. *Malar J* .

Lee, Y., Weakley, A. M., Nieman, C. C., Malvick, J., & Lanzaro, G. C. (2015). A multi-detection assay for malaria transmitting mosquitoes.*J Vis Exp* (96), e52385. doi:10.3791/52385

Lehmann, T., & Diabate, A. (2008). The molecular forms of Anopheles gambiae: a phenotypic perspective. *Infect Genet Evol, 8* (5), 737-746. doi:10.1016/j.meegid.2008.06.003

Lehmann, T., Licht, M., Elissa, N., Maega, B. T., Chimumbwa, J. M., Watsenga, F. T., . . . Hawley, W. A. (2003). Population Structure of Anopheles gambiae in Africa. *J Hered, 94* (2), 133-147. doi:10.1093/jhered/esg024

Lehmann, T., Weetman, D., Huestis, D. L., Yaro, A. S., Kassogue, Y., Diallo, M., . . . Dao, A. (2017). Tracing the origin of the early wet-season Anopheles coluzzii in the Sahel. *Evol Appl, 10* (7), 704-717. doi:10.1111/eva.12486

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. . *arXiv:1303.3997v1 [q-bio.GN].*

Loiseau, C., Melo, M., Lee, Y., Pereira, H., Hanemaaijer, M. J., Lanzaro, G. C., . . . Gilbert, F. (2019). High endemism of mosquitoes on São Tomé and Príncipe Islands: evaluating the general dynamic model in a worldwide island comparison. *Insect Conservation and Diversity, 12* (1), 69-79. doi:10.1111/icad.12308

Losos, J. B., & Ricklefs, R. E. (2009). Adaptation and diversification on islands. *Nature, 457* (7231), 830-836. doi:10.1038/nature07893

MacArthur, R. H., & Wilson, E. O. (1967). The Theory of Island Biogeography *Princeton University Press* .

Macias, V. M., McKeand, S., Chaverra-Rodriguez, D., Hughes, G. L., Fazekas, A., Pujhari, S., . . . Rasgon, J. L. (2020). Cas9-Mediated Gene-Editing in the Malaria Mosquito Anopheles stephensi by ReMOT Control. *G3 (Bethesda), 10* (4), 1353-1360. doi:10.1534/g3.120.401133

Mantel, N. (1967). The detection of diases clustering and generalized regression approach. *Cancer Research* .

Marshall, J. C., Pinto, J., Charlwood, J. D., Gentile, G., Santolamazza, F., Simard, F., . . . Caccone, A. (2008). Exploring the origin and degree of genetic isolation of Anopheles gambiae from the islands of São Tomé and Príncipe, potential sites for testing transgenic-based vector control. *Evol Appl, 1* (4), 631-644. doi:10.1111/j.1752-4571.2008.00048.x

Molina, R., Benito, A., Roche, J., Blanca, F., C., A., Sanchez, A., & Alvar, J. (1993). Baseline entomological data for a pilot malaria control program in Equatorial Guinea. *J Med Entomol, 30* , 622-624.

Moreno, M., Salgueiro, P., Vicente, J. L., Cano, J., Berzosa, P. J., de Lucio, A., . . . Benito, A. (2007). Genetic population structure of Anopheles gambiae in Equatorial Guinea. *Malar J, 6* , 137. doi:10.1186/1475-2875-6-137

Nieman, C. C., Yamasaki, Y., Collier, T. C., & Lee, Y. (2015). A DNA extraction protocol for improved DNA yield from individual mosquitoes.*F1000Res, 4* , 1314. doi:10.12688/f1000research.7413.1

Paradis, E., & Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics, 35* (3), 526-528. doi:10.1093/bioinformatics/bty633

Pinto, J., Donnelly, M. J., Sousa, C. A., Ferreira, G. C., Elissa, N., Rosário, V. E., & Charlwood, J. D. (2002). Genetic structure of Anopheles gambiae (Diptera: Culicidae) in Sao Tome and Principe (West Africa): implications for malaria control. *Molecular Ecology* .

Pinto, J., Egyir-Yawson, A., Vicente, J., Gomes, B., Santolamazza, F., Moreno, M., . . . Della Torre, A. (2013). Geographic population structure of the African malaria vector Anopheles gambiae suggests a role for the forest-savannah biome transition as a barrier to gene flow.*Evol Appl, 6* (6), 910-924. doi:10.1111/eva.12075

Pinto, J., Sousa, C. A., Gil, V., Ferreira, C., Gonçalves, L., Lopes, D., . . . Rosário, V. E. (2000). Malaria in São Tomé and Príncipe parasite prevalences and vector densities. *Acta Tropica 76* , 185-193.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., . . . Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet, 81* (3), 559-575. doi:10.1086/519795

Reimer, L. J., Tripet, F., Slotman, M., Spielman, A., Fondjo, E., & Lanzaro, G. C. (2005). An unusual distribution of the kdr gene among populations of Anopheles gambiae on the island of Bioko, Equatorial Guinea. *Insect Mol Biol, 14* (6), 683-688. doi:10.1111/j.1365-2583.2005.00599.x

Salgueiro, P., Moreno, M., Simard, F., O'Brochta, D., & Pinto, J. (2013). New insights into the population structure of Anopheles gambiae s.s. in the Gulf of Guinea Islands revealed by Herves transposable elements. *PLoS One, 8* (4), e62964. doi:10.1371/journal.pone.0062964

Schiffels, S., & Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nat Genet, 46* (8), 919-925. doi:10.1038/ng.3015

Schmidt, H., Lee, Y., Collier, T. C., Hanemaaijer, M. J., Kirstein, O. D., Ouledi, A., . . . Lanzaro, G. C. (2019). Transcontinental dispersal of Anopheles gambiae occurred from West African origin via serial founder events. *Commun Biol, 2* , 473. doi:10.1038/s42003-019-0717-7

Scott, T. W., Takken, W., Knols, B. G., & Boëte, C. (2002). The ecology of genetically modified mosquitoes. *Science, 298* .

Service, M. W. (1997). Mosquito (Diptera: Culicidae) dispersal–the long and short of it. . *Journal of medical entomology* . doi:https://doi.org/10.1093/jmedent/34.6.579

Sharakhova, M. V., Hammond, M. P., Lobo, N. F., Krzywinski, J., Unger, M. F., Hillenmeyer, M. E., . . . Collins, F. H. (2007). Update of the Anopheles gambiae PEST genome assembly. *Genome Biol, 8* (1), R5. doi:10.1186/gb-2007-8-1-r5

Sinka, M. E., Bangs, M. J., Manguin, S., Rubio-Palis, Y., Chareonviriyaphap, T., Coetzee, M., . . . Hay, S. I. (2012). A global map of dominant malaria vectors. *Parasit Vectors* .

Tan, A., Abecasis, G. R., & Kang, H. M. (2015). Unified representation of genetic variants. *Bioinformatics, 31* (13), 2202-2204. doi:10.1093/bioinformatics/btv112

Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J., & Prins, P. (2015). Sambamba: fast processing of NGS alignment formats.*Bioinformatics, 31* (12), 2032-2034. doi:10.1093/bioinformatics/btv098

Teklehaimanot, H. D., Teklehaimanot, A., Kiszewski, A., Rampao, H. S., & Sachs, J. D. (2009). Malaria in São Tomé and Príncipe: on the brink of elimination after three years of effective antimalarial measures.*Am J Trop Med Hyg* .

Touré, Y. T., Dolo, G., Petrarca, V., Traore, S. F., Bouare, M., Dao, A., . . . Taylor, C. E. (1998). Mark–release–recapture experiments with Anopheles gambiae s.l. in Banambani Village, Mali, to determine population size and structure. *Med Vet Entomol* .

Voelker, G., Marks, B. D., Kahindo, C., A'Genonga, U., Bapeamoni, F., Duffie, L. E., . . . Light, J. E. (2013). River barriers and cryptic biodiversity in an evolutionary museum. *Ecol Evol, 3* (3), 536-545. doi:10.1002/ece3.482

Walker, P. G. T., Griffin, J. T., Ferguson, N. M., & Ghani, A. C. (2016). Estimating the most efficient allocation of interventions to achieve reductions in Plasmodium falciparum malaria burden and transmission in Africa: a modelling study. *The Lancet Global Health, 4* (7), e474-e484. doi:10.1016/s2214-109x(16)30073-0

Weetman, D., Wilding, C. S., Steen, K., Pinto, J., & Donnelly, M. J. (2012). Gene flow-dependent genomic divergence between Anopheles gambiae M and S forms. *Mol Biol Evol, 29* (1), 279-291. doi:10.1093/molbev/msr199

WHO. (2017). A framework for malaria elimination. *Geneva: World Health Organization* .

WHO. (2019). Malaria Report. *Geneva: World Health Organization* .

Wiebe, A., Longbottom, J., Gleave, K., Shearer, F. M., Sinka, M. E., Massey, N. C., . . . Moyes, C. L. (2017). Geographical distributions of African malaria vector sibling species and evidence for insecticide resistance. *Malar J, 16* (1), 85. doi:10.1186/s12936-017-1734-y

Yamasaki, Y. K., Nieman, C. C., Chang, A. N., Collier, T. C., Main, B. J., & Lee, Y. (2016). Improved tools for genomic DNA library construction of small insects. *F1000Res.* doi:doi: 10.7490/f1000research.1111322.1

DATA ACCESSIBILITY

New whole genome sequence data included in this study are deposited in NCBI GenBank with accession numbers SAMN15641374- SAMN15641426 under BioProject ID PRJNA648422.

AUTHOR CONTRIBUTIONS

M.C. designed research, analyzed data, wrote the paper. M.H. designed research, analyzed data. H.G. processed samples. T.C.C. contributed with analytical tools. Y.L. designed research, wrote the paper. A.J.C. field collection. J.P. designed research, field collection. H.R. field collection. G.C.L. designed research, wrote the paper.

SUPPORTING INFORMATION

Supplementary Table S1

Figures S1-S3