RESEARCH ARTICLE SUMMARY

MOSQUITO GENETICS

Genomic diversity of the African malaria vector *Anopheles funestus*

Marilou Boddé†, Joachim Nwezeobi†, *et al.*

Full article and list of author affiliations: https://doi.org/10.1126/science.adu3596

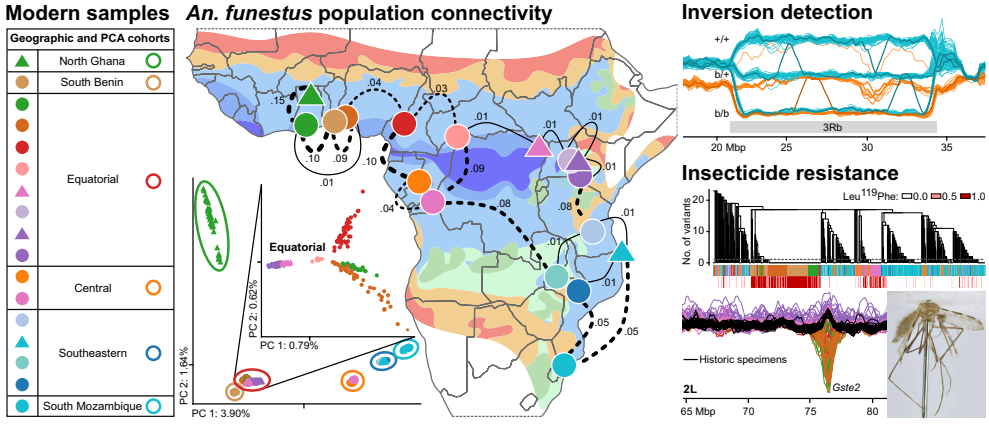
INTRODUCTION: The mosquito species *Anopheles funestus* is a major contributor to human malaria transmission across its vast sub-Saharan African range. Vector control of the other three major malaria-transmitting species in the Gambiae Complex has benefited from a deep understanding of genetic diversity, population structure, and the emergence and spread of insecticide resistance through the whole-genome sequencing of hundreds of individuals from many African countries. We completed whole-genome sequencing of 656 modern samples collected since 2014 and 45 historic samples collected between 1927 and 1967 to create a foundational understanding of genomic diversity in *An. funestus* across the continent.

RATIONALE: Since large scale deployment of insecticides began in the 1950s, *An. funestus* has rapidly evolved resistance throughout much of its range. However, it is an open question whether resistance alleles have evolved independently in multiple locations, whether they are shared between different populations through gene flow, or whether resistant populations have entirely replaced historically susceptible populations. A clearer genomic view on continental population structure is crucial for implementing strategic use of insecticides, taking into account the potential emergence and spread of insecticide resistance alleles. Additionally, with the implementation of gene drive release for vector control likely in the coming years, we need to be able to predict the spread of gene drive under different release scenarios, which is only possible if detailed knowledge of population connectivity across the continent, and how it varies along the genome, is in place.

RESULTS: We found that the 17 geographic regions from which our samples originated form six population clusters with varying degrees of genome-wide differentiation. One of these populations, the Equatorial cohort, spans more than 4000 km and comprises individuals from seven countries. In close geographic proximity to this cohort, we found two genetically distinct ecotypes that appear to have a restricted range and distinct chromosomal karyotypes. Using a windowed principal components analysis (PCA) approach, we explored structure across the genome. We used this approach to identify segregating inversions and classify every individual into its specific inversion karyotype. We also identified genomic regions that have exceptional levels of divergence in comparison to other collinear parts of the genome. Some of these outlier regions are clearly driven by selection for insecticide resistance, as they contain loci with excessive haplotype sharing, often centered on genes known to play a role in insecticide resistance in many insect species. We show that the *Gste2* resistance allele has at least two independent origins and that, despite reports of DDT resistance emerging in the 1950s, none of the historic samples in this study carry DDT resistance alleles found in modern-day populations.

CONCLUSION: Variable structure—such as that observed in this work, with some populations readily sharing alleles across the continent, and others clearly geographically proximal but genetically distinct—is a challenge for vector control. Even if the Gambiae Complex disappeared today, malaria would still rage through Africa until *An. funestus* is also effectively targeted. The greater understanding of the high levels of genetic diversity and the complex population structure of *An. funestus* presented in this study will underpin smarter surveillance and targeted vector control. □

Corresponding author: Mara K. N. Lawniczak (mara@sanger.ac.uk) †These authors contributed equally to this work. Cite this article as M. Boddé *et al.*, *Science* **389**, eadu3596 (2025). DOI: 10.1126/science.adu3596



Continental population structure among 656 modern *An. funestus* specimens. PCA, fixation index (F_{ST}), and a windowed PCA approach reveal how population structure varies across the continent and along the genome. The windowed PCA shows distinct patterns for segregating inversions (including double recombinants) and loci under positive selection. Insecticide resistance mutations were found in selective sweeps on different haplotypic backgrounds, suggesting multiple independent origins, but most resistance mutations were not observed in 45 historic specimens. SNP, single-nucleotide polymorphism; L119F, Leu^{119}Phe mutation.

MOSQUITO GENETICS

Genomic diversity of the African malaria vector *Anopheles funestus*

Marilou Boddé[1,2,3]†, Joachim Nwezeobi[1]†, Petra Korlević[1], Alex Makunin[1], Ousman Akone-Ella[4], Sonia Barasa[5], Mahamat Gadji[6], Lee Hart[1], Emmanuel W. Kaindoa[7], Katie Love[1], Eric R. Lucas[8], Ibra Lujumba[5], Mara Máquina[9], Sanjay C. Nagi[8], Joel O. Odero[7], Brian Polo[5], Claire Sangbakembi[10], Samuel Dadzie[11], Lizette L. Koekemoer[12], Dominic Kwiatkowski[1]‡, Erica McAlister[13], Eric Ochomo[5,8], Fredros Okumu[7], Krijn Paaijmans[14], David P. Tchouassi[15], Charles S. Wondji[6,8], Diego Ayala[2,16], Richard Durbin[1,17], Alistair Miles[1], Mara K. N. Lawniczak[1]*

Anopheles funestus s.s. is a major human malaria vector across Africa. To study its evolution, especially under vector control pressure, we sequenced 656 modern specimens (collected 2014 to 2018) and 45 historic specimens (collected 1927 to 1967) from 16 African countries. Despite high genetic diversity, the species shows stable but considerable continental population structure. Although one population showed little differentiation over a century and 4000 kilometers, nearby, we found two genetically distinct ecotypes. Vector control has resulted in strong signals of selection, with some resistance alleles shared across populations through gene flow and others arising independently. Fortunately, we found that a promising gene drive target in *Anopheles gambiae* is highly conserved in *An. funestus*. These insights will enable more strategic insecticide usage and gene drive deployment, supporting malaria elimination.

The mosquito species *Anopheles funestus* (Giles, 1900) has an extraordinary adaptive potential, demonstrated by its vast geographic range that spans sub-Saharan Africa (*1*) (shaded region, Fig. 1A). *An. funestus* is highly anthropophilic (*2*), has a substantially longer lifespan than other African malaria vectors (*3*), and, in some areas, is also reported to have extremely high *Plasmodium* infection rates (*4*–*6*). In much of eastern and southern Africa, the species is the major malaria vector (*7*). Like the vectors in the Gambiae Complex, the species is contributing to outdoor early evening and late morning biting in response to the use of indoor-based interventions, including bed nets (*8*–*10*). Although it can be difficult to find as larvae, it may have an extended transmission season in some locations, probably owing to a preference for larval habitats that persist in the dry months (*11*). *An. funestus* harbors several polymorphic chromosomal inversions that play an important role in adaptation (*12*). Similar to the three other major African human malaria vectors, which are all members of the Gambiae Complex, *An. funestus* exhibits insecticide resistance across its range (*13*) and, in some locations, even higher tolerance to insecticides than *An. gambiae* (*14*). Alleles that confer resistance can be shared across great distances or can differ locally, indicating a complicated mix of population connectivity and selective pressures (*15*). These characteristics, combined with its widespread distribution, make *An. funestus* a critical target for malaria control efforts. Substantial genomic data resources exist for the three major malaria vector species of the Gambiae Complex (*16*), and these have become foundational for the study and implementation of control efforts (*17*). In this study, we establish a baseline understanding of genetic diversity, population structure, inversion frequencies, and historic and current selection in *An. funestus* across the African continent at a whole genome level.

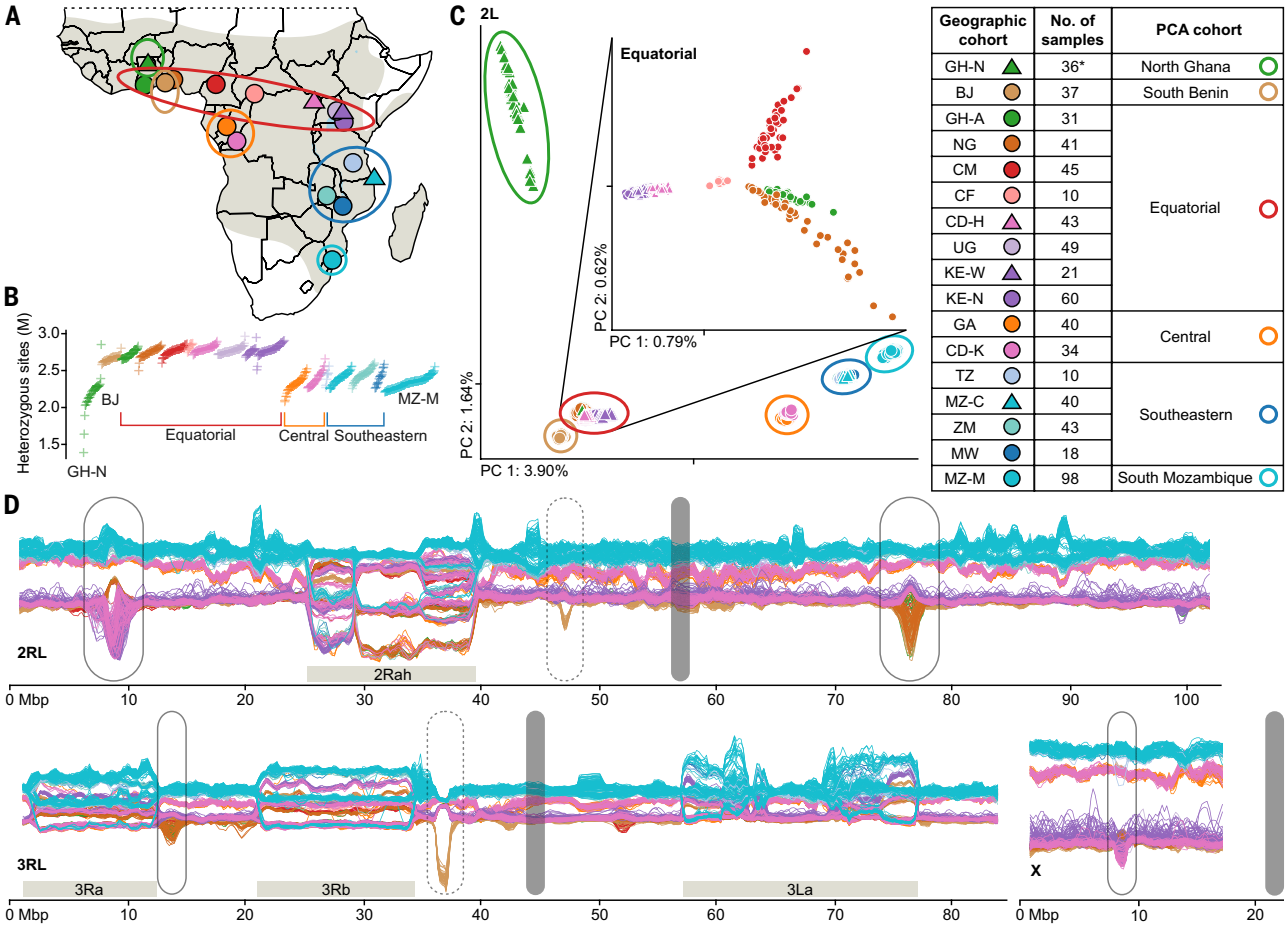
Sample acquisition, sequencing, and genetic diversity

In 2017, we put out an open call to join the project through contributing wild *An. funestus* samples collected from 2014 onwards. We carried out short-read sequencing at ~35x coverage depth for >800 wild-caught specimens, of which 656 individuals from 13 African countries [Benin (BJ); Cameroon (CM); Central African Republic (CF); Democratic Republic of the Congo, Kinshasa (CD-K) and Haut-Uélé (CD-H); Gabon (GA); Ghana, Northern Region (GH-N) and Ashanti (GH-A); Kenya, Western Province (KE-W) and Nyanza (KE-N); Malawi (MW); Mozambique, Cabo Delgado (MZ-C) and Maputo (MZ-M); Nigeria (NG); Tanzania (TZ); Uganda (UG); and Zambia (ZM)] passed all quality control (QC) evaluations (Fig. 1A and fig. S1A; materials and methods). Sequencing reads were aligned to a 251-Mbp high-quality chromosomal reference genome created from a wild-caught individual from Gabon (*18*), and each successfully sequenced sample was assigned to a geographic cohort based on its original collection location (Fig. 1A and table S1). Using a static-cutoff (sc) site filter, 73 million out of 162 million (45%) accessible sites were segregating among the sequenced samples (fig. S1, C and D; materials and methods). Disregarding singletons, 49 million single-nucleotide polymorphisms (SNPs) were present on two or more chromosomes, 17.3% of which had more than two alleles. For some analyses, we used a more stringent decision tree (dt) site filter that resulted in 48 million out of 114 million (42%) segregating accessible sites, with 31 million present on two or more chromosomes, and 15.2% having more than two alleles. Genome-wide nucleotide diversity was between 1.4 and 1.7% in each geographic cohort (fig. S2C), with most individuals having between 2 and 3 million heterozygous sites, and there was higher diversity among Equatorial individuals and cohorts (Fig. 1B and fig. S2, A and C). The Watterson θ estimator, which is proportional to effective population size under a simple coalescent model with constant mutation rate, was also higher in the Equatorial cohorts (fig. S2D). Mosquitoes from the North Ghana cohort have lower heterozygosity due to long runs of homozygosity (ROH; figs. S2B and S5D); the three mosquitoes with <2 million heterozygous sites had the highest fractions of their genome in ROH, even compared with other individuals from the North Ghana cohort.

Population structure and ecotypes

Principal components analysis (PCA) on variants from chromosome arm 2L, which has no known common chromosomal inversions (*19*), showed that PC1 is correlated with latitude (Fig. 1C). Population structure is similar across other autosomal arms, but the patterns are more complicated because of segregating inversions (fig. S1E). We used the observed clusters in the PCA plot to define six PCA cohorts, five of which (North Ghana, South Benin, Central, Southeastern, and South Mozambique) cover a regional area, whereas the sixth Equatorial cohort includes samples from seven countries spanning a 4000-km range from Ghana to Kenya (Fig. 1, A and C), suggesting a high degree of genetic connectivity across the equatorial region of Africa. Further PCA on the Equatorial cohort alone showed that the samples are clustered by geographic location (Fig. 1C, inset). Samples from the Equatorial cohort display more negative Tajima's *D* values than those from other

[1]Wellcome Sanger Institute, Hinxton, UK. [2]Institut Pasteur de Madagascar, Antananarivo, Madagascar. [3]Leibniz Institute for the Analysis of Biodiversity Change, Bonn, Germany. [4]CIRMF, Franceville, Gabon. [5]Kenya Medical Research Institute, Nairobi, Kenya. [6]Centre for Research In Infectious Disease, Yaounde, Cameroon. [7]Ifakara Health Institute, Ifakara, Tanzania. [8]Liverpool School of Tropical Medicine, Liverpool, UK. [9]Centro de Investigação em Saúde de Manhiça (CISM), Maputo, Mozambique. [10]Institut Pasteur de Bangui, Bangui, Gabon. [11]Noguchi Memorial Institute for Medical Research, Legon, Ghana. [12]University of the Witwatersrand, Johannesburg, South Africa. [13]Natural History Museum, London, UK. [14]Arizona State University, Tempe, Arizona, US. [15]International Centre of Insect Physiology and Ecology, Nairobi, Kenya. [16]MIVEGEC, University of Montpellier, CNRS, IRD, Montpellier, France. [17]Department of Genetics, University of Cambridge, Cambridge, UK. ***Corresponding author. Email: mara@sanger.ac.uk** †These authors contributed equally to this work. ‡Deceased.



Geographic cohort	No. of samples	PCA cohort
GH-N	36*	North Ghana
BJ	37	South Benin
GH-A	31	Equatorial
NG	41	
CM	45	
CF	10	
CD-H	43	
UG	49	
KE-W	21	
KE-N	60	
GA	40	Central
CD-K	34	
TZ	10	Southeastern
MZ-C	40	
ZM	43	
MW	18	
MZ-M	98	South Mozambique

Fig. 1. Population structure of 656 *An. funestus* specimens collected across Africa. (**A**) Map showing sequenced individuals from sub-Saharan Africa grouped into 17 geographic cohorts based on their collection location (filled circular and triangular markers) and 6 PCA cohorts consisting of one or more geographic cohorts (open elliptical markers), with the gray shaded area showing the *An. funestus* range across the continent [adapted from Sinka *et al.* (*1*)]. The asterisk represents one sample collected in GH-N that clusters with Equatorial on the PCA and has therefore been removed in analyses using PCA cohorts. (**B**) Number of heterozygous sites (in millions) per individual, with each cross representing one individual in subset_2 (females only; for subset_1, see fig. S2A) and ordered along the horizontal axis by geographic cohort (color, same order as the legend) and number of heterozygous sites. (**C**) Projection along the first two PCs computed on chromosome arm 2L, with the percentage of explained variance in the axis labels and ticks denoting the 0 position along each axis. The PCA cohorts are indicated with open elliptical markers. The inset PC plot was computed only on samples from the Equatorial PCA cohort. (**D**) Sliding window PCA along the genome. Each line corresponds to one individual, and individuals from GH-N were excluded. The genome was split in overlapping windows of approximately 1-Mbp and 200-kbp steps, a PCA was performed for each window, PC1 values are plotted along the *y* axis, and the window centers in genomic coordinates are plotted along the *x* axis. (the corresponding plot displaying the PC2 values for the same windows is shown in fig. S7A). Inversion regions are indicated by horizontal gray bars, centromeres are indicated by vertical grey bars, and lined ellipses indicate regions of divergence across several cohorts (Fig. 3 and figs. S8 to S10), whereas the two dotted ellipses showcase divergence specific to the South Benin PCA cohort (fig. S5, B and C). GH-N, Ghana Northern Region; BJ, Benin; GH-A, Ghana Ashanti; NG, Nigeria; CM, Cameroon; CF, Central African Republic; CD-H, Democratic Republic of the Congo, Haut-Uélé; UG, Uganda; KE-W, Kenya Western Province; KE-N, Kenya Nyanza; GA, Gabon; CD-K, Democratic Republic of the Congo Kinshasa; TZ, Tanzania; MZ-C, Mozambique Cabo Delgado; ZM, Zambia; MW, Malawi; MZ-M, Mozambique Maputo.

PCA cohorts supporting an expanding population (fig. S2E). Pairwise fixation indices (F_{ST}) and patterns of doubleton sharing support the PCA cohort groupings (fig. S2, F and G). PCA was computed on common variants (minor allele frequency > 0.01) that tend to be older, whereas doubleton sharing focused on rare variants that tend to be more recent (*20*), suggesting that the observed structure does not differ considerably between historic and recent events.

Although the North Ghana and South Benin cohorts are *An. funestus* s.s. (fig. S3A) and are geographically close to the Equatorial cohort (Fig. 1A), they appear to be genetically distinct in the PCA and thus may represent new ecotypes. In *Anopheles*, the concept of ecotype is somewhat interchangeable with the concept of a "chromosomal form," which has an inversion karyotype that is either fixed and different from neighboring or sympatric populations or at very different frequencies. The broader concept of "ecotype" is complex and only variably supported by genomic data (*21*). In this work, we use "ecotype" to mean a population that is more genetically distinct than expected based on geographic distance. Previous work has defined two ecotypes in *An. funestus*: the Folonzo ecotype, which has varying inversion frequencies and a preference to breed in vegetative ditches and swamps, and the Kiribina ecotype, which has a nearly fixed standard karyotype and has adapted to breed in rice fields (*22*, *23*). Although Folonzo is more widespread and Kiribina has only so far been found in Burkina Faso, these ecotypes are sympatric in Burkina Faso and thought to be semireproductively isolated (*22*, *23*). Previous population genomic sequencing of Kiribina and Folonzo individuals from Burkina Faso showed genome-wide differentiation between these sympatric ecotypes that extended beyond the genomic regions harboring the inversions that segregate at different frequencies (*22*, *23*). Our sampling did not contain known Kiribina or Folonzo, so to

better explore how these previously defined ecotypes relate to the geographic and PCA cohorts that we identified in this work, we integrated published Kiribina and Folonzo data into our dataset. Consistent with previous work, we found that Folonzo and Kiribina form distinct clusters, but, unexpectedly, both ecotypes clustered with our Equatorial cohort (fig. S3B) (*23*). Therefore, our samples from North Ghana and South Benin appear to be more genetically distinct from the Equatorial population than Kiribina and Folonzo, even though they were collected from areas that are geographically closer to Equatorial cohort populations. Further exploration of genetic differentiation using isolation by distance (IBD) showed that North Ghana and South Benin are more isolated than expected by distance and more isolated than Kiribina or Folonzo (fig. S3C) and that North Ghana and South Benin appeared as distinct populations before Kiribina (fig. S4). IBD analysis also showed that although the Central, Southeastern, and South Mozambique clusters drove PC1 (Fig. 1C), their differentiation from other cohorts was consistent with the pattern expected under isolation by distance, and thus, we do not consider them potential ecotypes (fig. S3, C and D). Therefore, we further explored South Benin and North Ghana as potential ecotypes.

Compared with all other populations, the North Ghana cohort had reduced levels of diversity (Fig. 1B and fig. S2A), an excess of both long and short ROH (figs. S2B and S5D), and high divergence from all other cohorts (fig. S2F), including from GH-A <400 km away, consistent with a recent population bottleneck. Only a single individual from North Ghana clustered with the GH-A population, but it is unclear whether this was due to very recent migration, perhaps from being transported in a car or other vehicle, or whether this individual is representative of a sympatric Equatorial population in North Ghana, which is not well represented in our sample set. Comparing the two populations from Ghana revealed elevated F_{ST} in the genomic regions associated with the 2R and 3L inversions (fig. S5E), and indeed, North Ghana appears to be an outlier for its 2Ra karyotype in comparison to its nearest neighbors (fig. S6E). We were also unable to karyotype North Ghana for the 3La inversion owing to high divergence on the 3L arm, particularly in the inversion region (fig. S5, E and F). These karyotypic anomalies, together with a potentially sympatric population that clusters with the Equatorial cohort, hint that the North Ghana cohort may indeed be a new *An. funestus* ecotype, but further sampling will be needed to confirm this hypothesis.

The South Benin (BJ) cohort has similar levels of variation as the Equatorial cohort but was clearly differentiated in the PCA (Fig. 1, B and C). To further explore genomic patterns of differentiation, we ran sliding window PCAs that resulted in a view of population structure along the genome [Fig. 1D; North Ghana was excluded from these analyses because of excessive ROH; see (*24*) for interactive plots]. Although BJ follows the Equatorial cohort for most of the genome, it is moderately differentiated genome wide (fig. S2F) and highly differentiated from all other geographic cohorts in two genomic regions (2R, 47 Mbp; 3R, 37 Mbp) (Fig. 1D). Both loci have plenty of haplotype variation and show no evidence of selective sweeps in either BJ or any other cohort; however, both loci show that all BJ haplotypes are highly differentiated from every other haplotype sequenced here (fig. S5, B and C). Within the 2R locus, there is one large gene, "hemicentin-1" (AFUN2_013216), and within the 3R locus, the peak of F_{ST} between BJ and the neighboring GH-A shows that the strongest differentiation between the two populations falls between two genes characterized as "semaphorin-2A–like" (AFUN2_001563, AFUN2_006509) (fig. S5A). In both cases, we found some relatively weak nonsynonymous differentiation between BJ and other populations, so it may be that the divergence is regulatory in nature, or that the reference genome is inadequately representing the sequences of these individuals in these genomic regions (table S3). Populations from Benin are exceptionally resistant to DDT (0% mortality after one hour of exposure) (*25*), but there is currently no evidence in the literature to support a role in DDT resistance for hemicentin-1 or semaphorin-2A, which are secreted or transmembrane immunoglobulins and signaling proteins; instead, semaphorins have been implicated in nerve development, limb development, and olfaction (*26*, *27*). Strong divergence could be caused by differential regulation of genes involved in DDT resistance, assortative mating, or habitat adaptation, but genome accessibility drops off as F_{ST} increases, and the cohort is too diverged from the reference genome in this genomic region to speculate further. BJ has inversion frequencies that are also strongly contrasting with those of its close geographic neighbors: It is fixed standard for all inversions except 3La (Fig. 2 and fig. S6E). This is reminiscent of the Kiribina ecotype, which is (nearly) fixed standard for 2Ra, 3Ra, and 3Rb inversions (*22*, *23*); however, BJ is clearly differentiated from Kiribina (fig. S3B). Furthermore, in this sample set, we found no evidence of a Folonzo or Equatorial-like population in South Benin, so BJ may be the only form present in the region, but this requires further sampling to confirm.

We explored the mitochondrial genome for every individual, including those failing QC (838 modern and 75 historic), and included a selection of previously published mitochondrial genomes across the *Anopheles* genus (fig. S11E and table S1). Previous studies identified two lineages in the mitochondrial tree of the Funestus Subgroup; lineage I, which only contains *An. funestus*, and lineage II, which contains *An. funestus* as well as other species in the subgroup (*28*–*30*). In these studies, the *An. funestus* samples in lineage II all originated from east Africa (Madagascar, Mozambique, Tanzania, and Zambia), and for all but one location, lineage I individuals were found sympatrically. We generated mitochondrial consensus sequences for our samples and built a tree that included samples from these previous studies (*29*, *30*). Roughly half of the South Mozambique cohort belongs to lineage II, and we found a small number (12 total) of individuals from other cohorts (GH-A, CD-H, UG, CD-K, TZ, and ZM) in this lineage. Within lineage I, two clusters have been identified, where cluster A contained individuals from Zambia, DRC, and Tanzania, where cluster B contained only individuals from Zambia (*29*). We observed clear geographic structure between the clusters, consistent with the most pronounced structure in the analysis of the nuclear genome, where the vast majority of samples from the North Ghana, South Benin and Equatorial cohorts belong to cluster B, whereas the vast majority of samples from the Central and Southeastern cohorts belong to cluster A, and the South Mozambique cohort is relatively evenly split between lineage I cluster A and lineage II. However, the separation is not perfect, and we found eight samples from GH-N, GH-A, UG, and KE-N in cluster A and, vice versa, three samples from TZ and MW in cluster B. It is notable that we did not observe the split between lineage I and II samples from South Mozambique in the analysis of the nuclear genomic data, suggesting that the mitochondrial and nuclear genome can have distinct evolutionary histories.

Although further sampling across the continent is needed, these results indicate that there is one genetically connected population of *An. funestus* across the equator, whereas other populations, both geographically proximal and distant to the Equatorial cohort, are much more genetically isolated.

Inversions

Chromosomal inversions can capture coadapted alleles and protect them from recombination (*31*). In the sliding window PCA, we found five large segregating inversions that drive population structure in their respective genomic regions, and after determining inversion breakpoints, we identified these inversions as 2Ra, 2Rh, 3Ra, 3Rb, and 3La (Fig. 1D, fig. S1B, and table S2; supplementary text). We performed in silico karyotyping for all samples and all inversions using a sliding window PCA and aggregated PCA on each inversion region (Fig. 2, fig. S6, and table S1). Inversion frequencies vary greatly across the continent, and previous studies have linked inversions to behavior and adaptation (*12*, *32*, *33*). Although Hardy-Weinberg equilibrium (HWE)

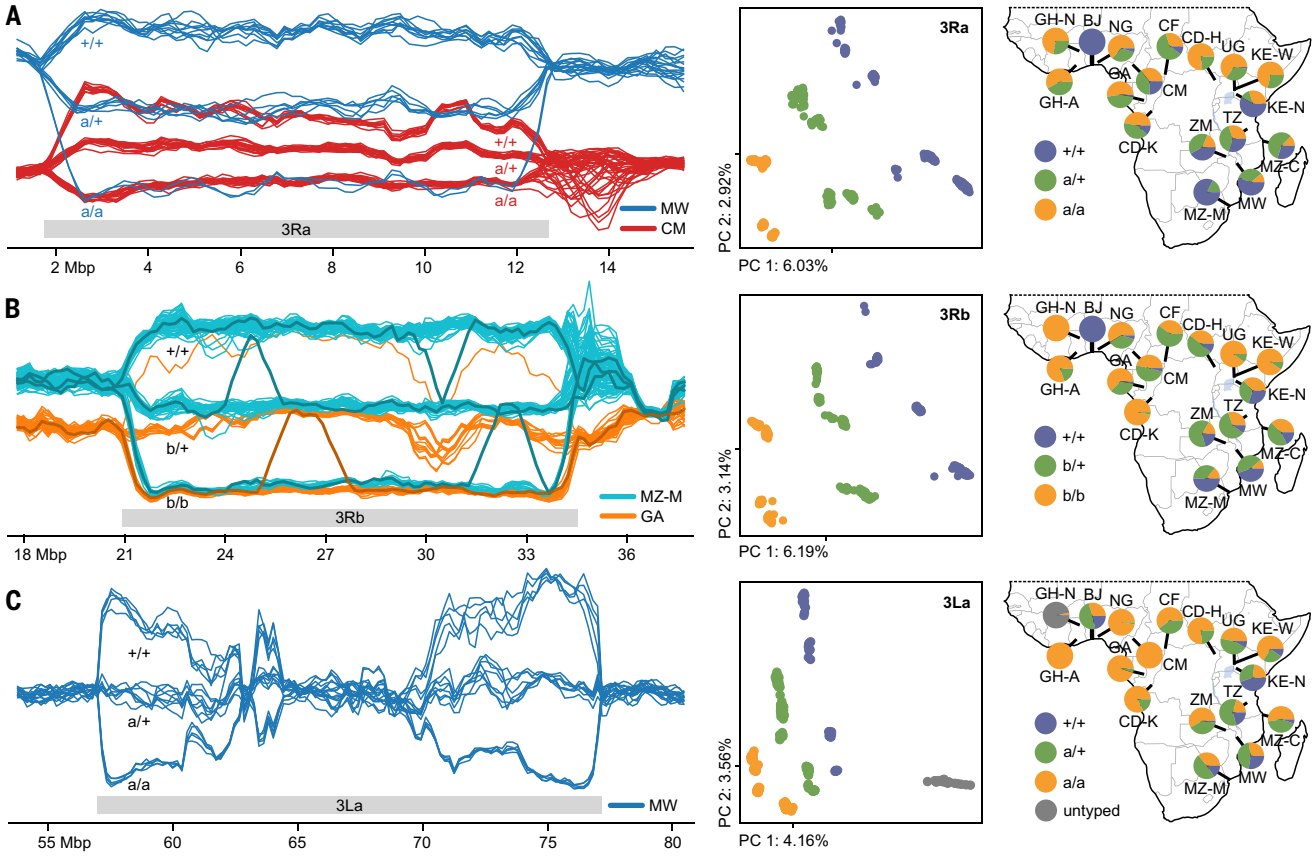


Fig. 2. Segregating inversions on chromosome 3. Sliding window PCAs (left) were computed on all samples except North Ghana. For visualization purposes, only a subset of samples is shown for each sliding window PCA. (**A**) (Left) Sliding window PCA showing CM (red) and MW (blue) samples on the 3Ra inversion region; window centers are plotted in mega–base pairs along the *x* axis, and PC1 values for each individual, along the *y* axis. The different karyotypes are visible as three different horizontal trajectories: from top to bottom, 3R+/+, 3Ra/+, and 3Ra/a. All three karyotypes are present in both cohorts, and because PC1 captures a signal that is a combination of the inversion and geographic population structure, the upper and middle trajectories (3R+/+ and 3Ra/+, respectively) do not overlap for the different cohorts; see fig. S7D. (Middle) Projection along the first two PCs computed on the entire inversion region using all samples; samples are colored by inversion karyotype. (Right) Map of karyotype frequencies per geographic cohort. (**B**) (Left) Sliding window PCA on the 3Rb inversion region displaying GA (orange) and MZ-M (cyan). Several samples change trajectory for part of the inversion region (indicated by thicker, darker lines); these samples appear to be the product of double recombination, and they locally exhibit a different karyotype than they have for the rest of the inversion (see fig. S7E and supplementary materials). The middle and right panels are the same as in (A) but for the 3Rb inversion. (**C**) Sliding window PCA displaying MW on the 3La inversion region, showing a strong separation of karyotypes near the breakpoints, but the signal decays toward the inversion center (fig. S7B). The middle and right panels are the same as in (A) but for the 3La inversion.

of inversion karyotypes is typically satisfied within PCA cohorts, there are a few exceptions where samples collected in different seasons do not satisfy HWE when grouped together. Changes in inversion frequencies during the wet and dry seasons have previously been reported for *An. funestus* (*34*) as well as for the Gambiae Complex (*35*). For the 2Rh inversion, we observed a difference in heterozygosity between the homokaryotypes: 2Rh/h has lower heterozygosity than $2R+^{h}/+^{h}$, hinting that the standard orientation is likely ancestral (fig. S7C). For the other inversions, we did not observe such a clear difference in heterozygosity between the homokaryotypes, possibly because these inversions are older or the signal is confounded by other factors, such as unequal inversion frequencies or geographical spread.

The sliding window PCA enables the investigation of local structure within the inversions. In the genomic regions where large inversions are segregating, PC1 captures a compound signal of inversion karyotype and population structure (Figs. 1D and 2 and figs. S6 and S7). Size, age, recombination rates, and demographic history of the inversion can all affect the relative strength of the inversion and population structure signals as well as the observed amount of genetic variation in the different inversion orientations (*36*). Although for most inversions, the relative strength of these two signals is constant for the entire inversion (e.g., Fig. 2A), for 3La, the inversion signal decays in the middle (Fig. 2C and fig. S7B). This distinct pattern of 3La in comparison to that of the other inversions may have been driven by more frequent double recombination due to its large size or due to its age. Sliding window PCA revealed individuals with inversion karyotypes that are the product of double recombination in 3Rb and 3La (Fig. 2B and fig. S7, E and F). These double recombinants can be seen in the sliding window plots where individuals change karyotype trajectories for a part of the inversion region (Fig. 2B). Events like these move alleles from one karyotype into the other and are important for the spread of variants (*37*).

Consistent with Sharakhov *et al.* (*19*), we observed several overlapping inversions on chromosome arm 2R: 2Ra and 2Rh, which occur in many cohorts, and 2Rt, which occurs only in the CM and CD-H cohorts (Fig. 1D and fig. S6). There are six possible combined karyotypes for 2Rah, all resulting in distinct trajectories in the sliding window PCA (fig. S6, B and D). The relatively rare 2Rt inversion results in two additional states in the region where it overlaps 2Ra (fig. S6, C and D). We believe that the 2Rt inversion only occurs in the heterozygous state in this dataset, firstly, because that results in a consistent interpretation

of the sliding window PCA trajectories of the combined 2Raht karyotypes, and secondly, because the samples carrying 2Rt never interpolate between two groups of homozygotes but drive a PC by themselves. Assuming that this dataset does not contain individuals homozygous for the 2Rt inversion, both CM and CD-H are in HWE.

Signals of selection

To explore signals of selection, we computed H12 (*38*), a statistic that measures the haplotype homozygosity along the genome (Fig. 3A; Materials and methods). H12 quantifies excess homozygosity indicative of hard or soft selective sweeps. We identified four regions that are putatively under very strong selection (H12 > 0.4) in at least two geographic cohorts (Fig. 3A, gray boxes); these regions are also visible as peaks in the sliding window PCA (Fig. 1D, circled regions). All four H12 peaks are centered on known insecticide resistance genes that are important in many insect species (*Gste2*, *Gaba*, *Cyp6p*, and *Cyp9k1*) (*39*).

The H12 peak on chromosome arm 2L—observed in BJ, GH-A, NG, and CD-K—is centered on a cluster of seven glutathione S-transferase (*Gste*) genes. The nonsynonymous L119F mutation in *Gste2* is important in both DDT and pyrethroid resistance in *An. funestus* (*25*). We observed this mutation across the continent, at low frequencies in the east and at particularly high frequencies in the aforementioned cohorts, being nearly fixed in BJ (Fig. 3, A, D, and F, and table S3). Most individuals from BJ, GH-A, and NG with the L119F mutation share the same haplotype in the *Gste* region, whereas in CD-K, the L119F mutation is found on a different haplotypic background (Fig. 3D and fig. S8A), suggesting two distinct selective sweeps rather than the spread of a single resistance mutation.

The H12 peak on chromosome arm 3R in GH-A and NG is centered on the γ-aminobutyric acid receptor subunit β (*Gaba*) gene (Fig. 3A and table S3). In many species, a mutation at one amino acid residue in *Gaba* confers dieldrin resistance, hence it being named the resistance to dieldrin (rdl) mutation (*40*). The A296S rdl mutation in *An. funestus* has previously been detected at high frequency in Central and Western Africa (*41*). The same A296S mutation is found here in 10 geographic cohorts, predominantly on two distinct swept haplotypic backgrounds that occur sympatrically in GH-A and NG (Fig. 3, E and G), but also spread throughout the continent at low frequency.

The other strong H12 peaks occurring in multiple cohorts include two other well-studied insecticide resistance loci, *rp1* and *Cyp9k1*. There are also strong H12 peaks that are only found in a single cohort each (TZ and CM), including a clear sweep on the well-studied knockdown resistance (kdr) mutation at the voltage-gated sodium channel (*vgsc*) locus only in the Tanzanian samples. These samples were collected from a region that has a DDT stockpile that was leaking into the environment for decades (*42*). These loci are all discussed in the supplementary text and figs. S9 and S10.

Some of the strongest selective pressures observed in this species are driven by insecticides. Widespread use of synthetic insecticides began in Africa in the late 1940s with the advent of DDT (*43*, *44*). To explore both when and where the mutations we observe arose, we sequenced 75 historic mosquitoes labeled as *An. funestus* from two museum collections (the Natural History Museum in London and the Research Institute for Development in Montpellier) spanning 1927 to 1973 (fig. S11 and table S4; materials and methods) (*45*). On the basis of mitochondrial and nuclear sequencing data, only 45 of these (spanning 1927 to 1967) were determined to be *An. funestus* s.s. (table S4). These 45 samples originated from nine countries and clustered together with the Equatorial and Southeastern PCA cohorts, which was expected based on their origins (fig. S11C). We integrated the historic data into the sliding window PCA, which showed that most of the genome in historic samples followed their present-day cohorts well, suggesting that population structure has been relatively stable over the past century [fig. S11D; see (*24*) for interactive plots]. Nine historic samples follow the trajectory of South Benin in the divergence peak in chromosome arm 3R, and the alleles observed in the historic samples are highly similar to those found in modern-day South Benin (fig. S12). As previously mentioned, there is exceptionally strong DDT resistance in South Benin (*46*), but it remains to be determined what this allele, which is presently found only in South Benin, was associated with in these historic populations.

Emerging DDT resistance was phenotypically detected in *An. funestus* in the late 1950s (*47*); however, we did not find any evidence of mutations now associated with DDT resistance in our 45 historic individuals, including L119F in *Gste* and kdr in *vgsc* (table S4). We suspect that if the kdr mutation conferred resistance to DDT 60 years ago, then it would have been rapidly reselected when pyrethroid-treated bed nets were distributed at scale. Historic genomes also do not have evidence of insecticide resistance mutations at *rp1* or *Cyp9k1* (table S4), suggesting that selection pressures on these genes are more recent. However, we found that six individuals from the 1960s, the time when dieldrin resistance in *An. funestus* was first reported (*47*, *48*), follow the PC1 trajectory of present-day resistant populations that carry the rdl A296S mutation (Fig. 3B and fig. S11D), and indeed, each of those individuals carry the mutation in a heterozygous or homozygous state. *Gaba* might act as a secondary target for pyrethroids (*49*), which could explain the persistence of rdl in modern populations. Dieldrin, like DDT, is a persistent organic pollutant and has stockpiles across Africa. Perhaps leaking dieldrin stockpiles continue to exert selective pressure on modern populations, or unregulated usage continues in some areas. Undoubtedly, these long-lasting insecticides that are stored in poor conditions in many locations across Africa are complicating resistance management and vector control (*50*).

Today, although we observe that both the L119F mutation in *Gste2* and the rdl mutation in *Gaba* are on haplotypes undergoing selective sweeps in west and central Africa, they are also found scattered across the continent at lower frequencies (Fig. 3, D to G). This may hint that historic selection for these mutations was relaxed during the period of the 1970s to 1990s, when DDT and dieldrin usage declined, but these mutations may have remained at low frequencies, preadapting the species to rapidly evolve resistance to pyrethroids when those came into common usage in the 2000s.

Gene drive

The use of gene drive to eliminate or modify malaria vector species holds great promise as a targeted malaria vector control strategy (*51*). Knowledge of genetic variation and structure throughout the species range is important to both assess the likelihood of geographical spread and to identify candidate targets for CRISPR-Cas9 gene drive, given that polymorphism within target sites affects gene drive efficacy (*16*). Suppression based gene drive targeting the *doublesex* (*dsx*) gene has already been developed and successfully tested in the laboratory and large cage settings for *An. gambiae* (*52*, *53*). Approaches developed for *An. gambiae* should be technically sound for *An. funestus*, but population structure and levels of genetic diversity will differ between these species.

In this work, we evaluated potential target sites within coding sequence for *An. funestus*. Gene drive targets were identified as 20-bp sequences located entirely within coding sequence, containing the protospacer adjacent motif (PAM). In total, we identified 2,349,313 target sites in 11,086 genes based on the reference genome. However, only 30,459 sites in 3927 genes remained after excluding targets with variation in any subset_2 individual. In comparing target site availability in *An. funestus* to *An. gambiae*, we found that *An. gambiae* has more available targets based on the reference genome alone (2,718,188), but the decay in number of targets as we consider variation in cumulatively more individuals is similar to that of *An. funestus* (Fig. 4A; supplementary text). We also specifically explored the *dsx* gene drive target region in *An. funestus*, finding that the reference genomes (AgamP4 and AfunGA1) differ at one site (Fig. 4B) and that this site also harbors

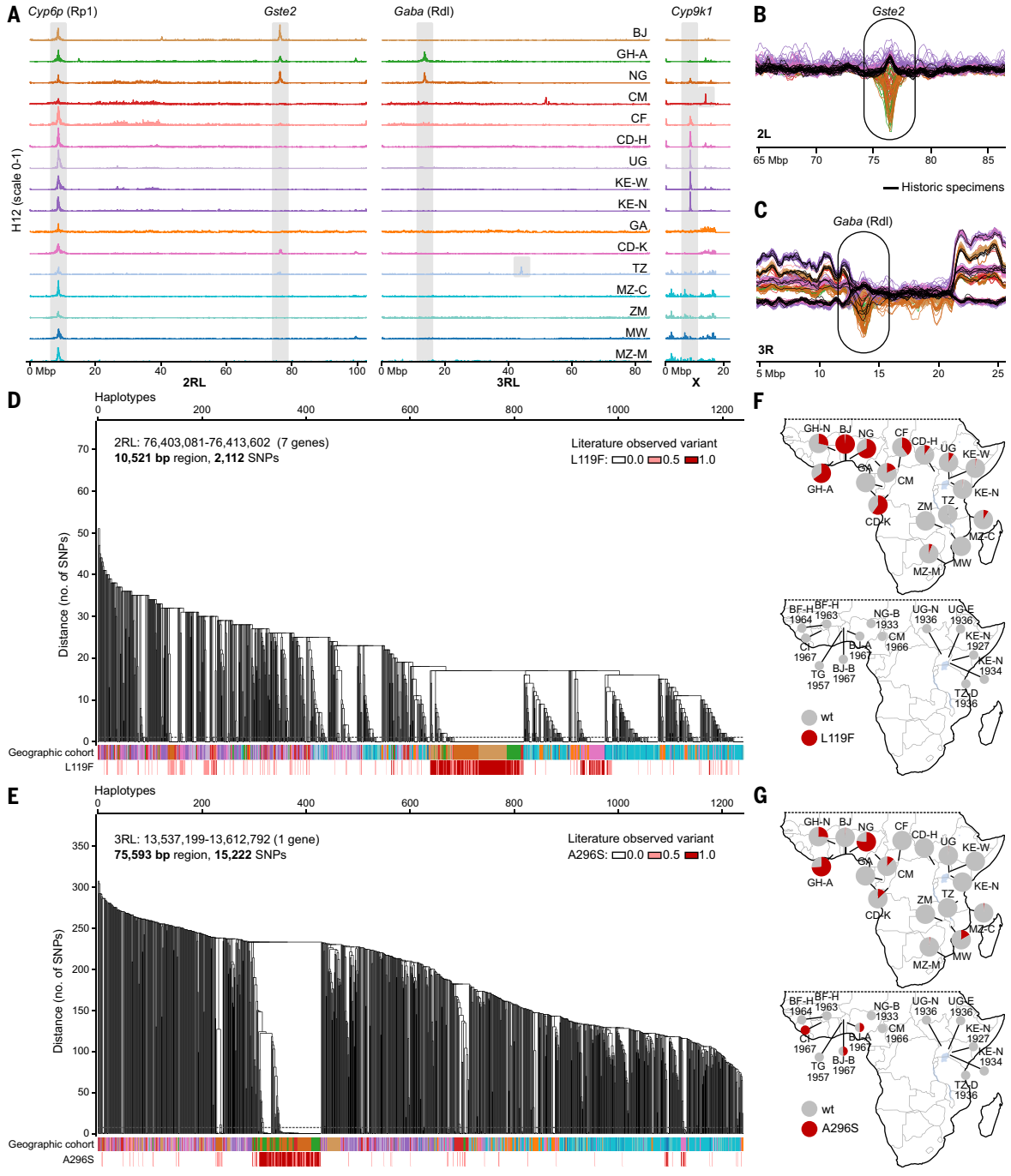


Fig. 3. Selection sweeps and known insecticide resistance variants in *Gste2* and *Gaba*. (**A**) Genome-wide H12 scans for signals of recent selection. H12 values range from 0 to 1, with higher values indicating excessive haplotype sharing, which is a signature of recent selection. North Ghana was excluded owing to elevated runs of homozygosity, which adds noise and makes H12 values unreliable for detecting genuine selection signals. The *y* axis runs from 0 to 1 for each cohort, and the *x* axis shows positions along the genome. Peaks of H12 values ≥0.4 are highlighted with a gray vertical bar; for peaks where only one cohort reaches this threshold, the bar is restricted to this cohort. (**B**) Sliding window PCA computed on modern samples and historic samples combined, excluding North Ghana. For visualization purposes, only present-day Equatorial and South Benin cohorts and the historic Equatorial cohort are shown (for full sliding window PCA, see fig. S11D), and the plot is restricted to a 20-Mbp region around the Gste2 gene. a 20-Mbp region around the *Gste2* gene. Present-day samples show a peak in the *Gste2* region (indicated by black ellipse), which is likely a signal of selection, whereas historic samples (shown in black) do not exhibit this signal. (**C**) Sliding window PCA computed and samples subsetted as in (B), here showing a 20-Mbp region around the *Gaba* gene. Several historic samples follow the peak seen in present-day samples at the *Gaba* locus. (**D**) Haplotype clustering within a region containing seven *Gste* genes. The dendrogram was obtained by hierarchical clustering of phased haplotypes and used to define haplotype clusters as groups of haplotypes with SNP divergence <0.0005 (cutoff indicated as dashed horizontal line on the dendrogram). The first bar below the dendrogram shows the population of origin for each haplotype, the second bar shows the genotype for the known *Gste2* Leu^{119}Phe (L119F) mutation (note that this mutation was filtered out before haplotype phasing, so each haplotype is colored by the genotype of the individual it belongs to; fig. S8A and table S3). (**E**) Haplotype clustering within the *Gaba* gene. Same structure as in (D), with the red bar showing the genotype for the rdl Ala^{296}Ser (A296S) variant (fig. S8B and table S3). (**F**) Maps showcasing the frequency of the L119F mutation in present-day (above) and historic (below) sample sets. (**G**) Maps showcasing the frequency of the rdl A296S mutation in present-day (above) and historic (below) sample sets. Single-letter abbreviations for the amino acid residues referenced throughout this paper are as follows: L, Leu; F, Phe; A, Ala; S, Ser.

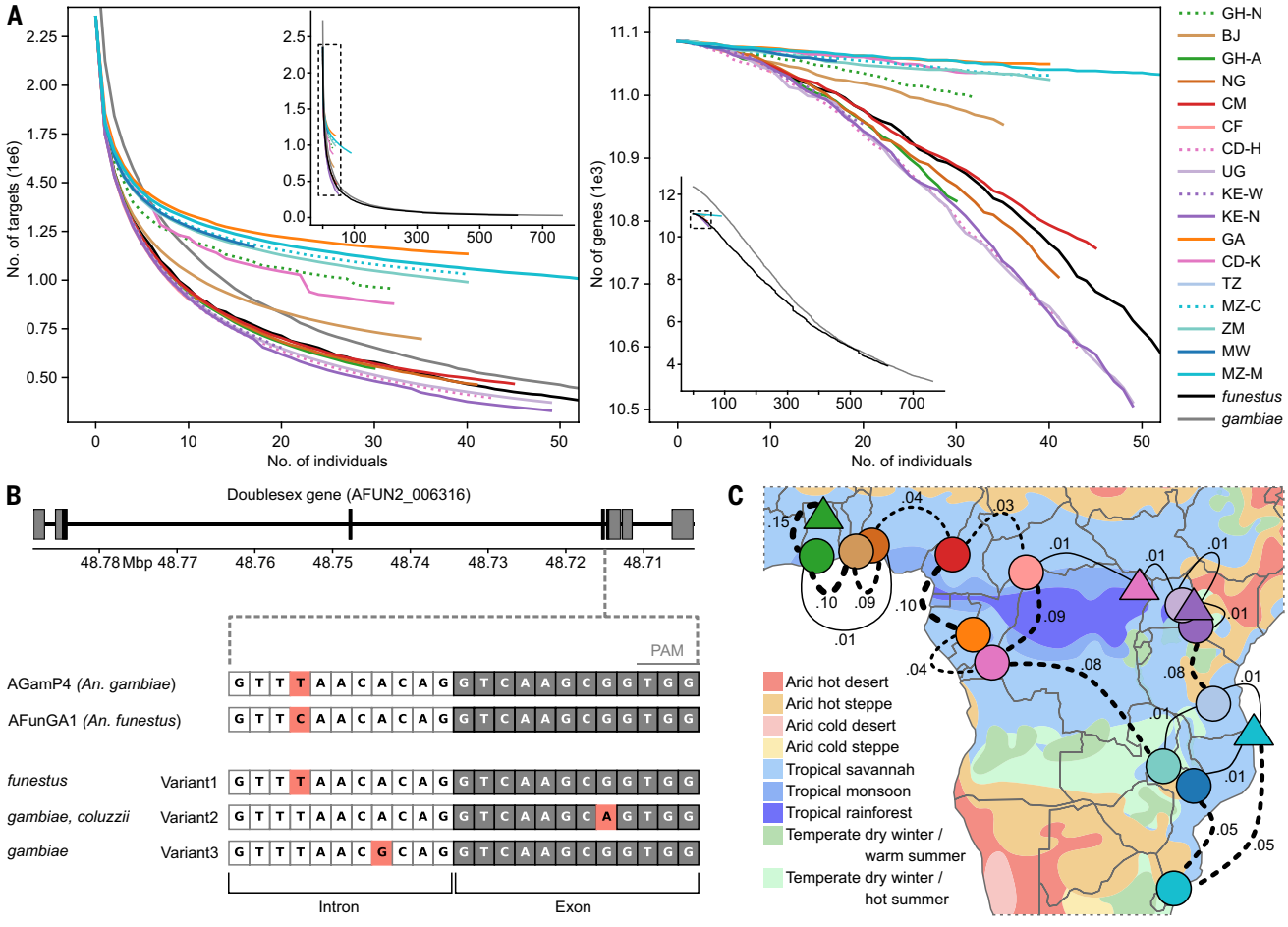


Fig. 4. Challenges for vector control. (**A**) Cumulative number of gene drive targets (left) and number of genes containing gene drive targets (right) per geographic cohort as well as all subset_2 individuals (*funestus*) and *An. gambiae* and *An. coluzzii* individuals from Ag1000G phase 1 (*16*) (*gambiae*). The insets are zoomed-out versions of the main plots, showing the results for the fully explored datasets; the dashed boxes in the insets indicate the zoomed-in areas shown in the main plots. (**B**) Sequence variation in the *dsx* gene drive target on chromosome arm 2R. The *dsx* gene drive target is located at the boundary of intron 4 and exon 5 of the female-specific isoform. At the top, the seven exons of the female-specific isoform are depicted as boxes on the AGamP4 reference genome; the coding sequence is shown in black, and untranslated exonic regions (UTRs) are shown in grey. The target sequence is shown below; nucleotides in white boxes are located in intron 4, and nucleotides in gray boxes are located in exon 5. The last three nucleotides constitute the PAM. The AgamP4 (*An. gambiae*) and AfunGA1 (*An. funestus*) reference genomes differ at the fourth base of the target sequence (highlighted in pink) in the intronic region. Variant 1 was found at very low frequencies in two cohorts from this study, variant 2 was found at low to intermediate frequencies in four cohorts of *An. gambiae* and *An. coluzzii*, and variant 3 was found at very low frequency in one cohort of *An. gambiae* (see supplementary materials). (**C**) Map showing F_{ST} for neighboring populations. F_{ST} was calculated on all accessible sites on the 2L chromosome arm. $F_{ST} > 0.01$ lines are dotted and become thicker as F_{ST} becomes higher. The background of the map is colored by the Köppen-Geiger climate classification, adapted from Beck *et al*. (*56*).

a SNP in two individuals among all our newly sequenced *An. funestus*. The alternative allele found in these two individuals is the *An. gambiae* reference allele. Altogether, this bodes well for using the same population suppression gene drive approach in *An. funestus*, as is underway for *An. gambiae*.

Discussion

Even if the Gambiae Complex disappeared today, malaria will still rage through Africa until *An. funestus* is also effectively targeted. The greater understanding of the high levels of genetic diversity and the complex population structure of *An. funestus* presented in this study will underpin smarter surveillance and targeted vector control (*3*, *54*, *55*). More than 4000 km separates the sampling sites of the Equatorial cohort (from Ghana Ashanti to Western Kenya), yet populations from across that range are genetically connected, whereas much geographically closer populations, such as South Benin and North Ghana, are genetically distinct (Fig. 4C). Some of this structure may originate from geographic discontinuities, such as the Congo basin rainforest and the Rift valley; some may originate from differences in climate and rainfall (*56*); and some may be due to local adaptation that we do not fully understand. Small *et al.* (*23*) predicted that other ecotypes are likely to be found as genomic investigations into the species proceed (*57*, *58*), and we argue that both South Benin and North Ghana may provide two more examples. The existence of these ecotypes near the Equatorial population, which ranges across a vast area, highlights the complexity of population structure in *An. funestus*. This complexity together with the high genetic diversity of the species suggests that a one-size-fits-all approach to *An. funestus* vector control may be ineffective.

The observations in this dataset regarding insecticide resistance further underscore the need to consider locally tailored control strategies. We identified strong selective sweeps centered on known insecticide resistance genes, where the same mutations tend to confer resistance and be under selection in a wide range of insects (*39*) as well as in *An. gambiae* (*59*). However, these shared mutations clearly sometimes occur independently on multiple distinct haplotypic backgrounds, suggesting that *An. funestus* populations may not be mutation limited.

This convergent evolution of resistance mechanisms highlights the adaptability of these vector populations and emphasizes the need for tailored interventions that consider local genetic backgrounds and resistance profiles (*60*). Insecticide resistance in *An. funestus* results from a complex interplay between convergent evolution, sharing of resistance alleles between populations, and changing selective pressures in space and time and calls for the continued monitoring of resistance alleles alongside the development and deployment of newly identified insecticides or alternative control strategies (*61*). Population suppression or modification gene drives are a promising alternative to insecticide-based control, and we identified a comparable number of candidate gene drive targets in *An. funestus* as in *An. gambiae* and *An. coluzzii*. Although extremely low levels of natural variation across the full species range at one of the most promising targets (*dsx*) is encouraging, gene drive release for *An. funestus* will need to account for the greater complexity of the species' population structure observed in this work than was previously appreciated.

A comprehensive understanding of gene flow within and between each of the major malaria vector species is critical for implementing effective malaria control, whether that be through gene drive release or successful insecticide use and resistance management. This work represents the first step in generating a foundational genomic understanding of *An. funestus* akin to what is available for the other major malaria vector species in Africa (*16*, *62*). Population genomic data resources are increasingly critical for informing vector control and are needed to underpin both technical strategies and global health policies. Further studies will need to focus on increased spatiotemporal sampling across the full species range. The complex patterns of population connectivity revealed in this study show that population differentiation occurs at wildly different spatial scales, with some geographically restricted ecotypes but also some extremely widespread forms. This kind of variable structure will impact both insecticide resistance and gene drive, influencing the rate and success of spread. This clearer understanding of structure should be incorporated into models to guide the challenge of gene drive release strategies. This challenge will be difficult owing to the complex social setting. For example, more extensive sampling across West Africa could reveal the range of the new putative ecotypes identified in this work. If they turn out to be geographically restricted and continue to look genetically differentiated, then they may be excellent candidates for the first releases of gene drive, as they would be less likely to spread into neighboring countries than, for example, a gene drive release in a genetic background that was clearly compatible with the Equatorial cohort. This is one theoretical example of the value of increased spatiotemporal sampling. In general, more data are needed to monitor and respond to both insecticide resistance and, when it arises, gene drive resistance. Beyond these important vector control motivations, increased spatiotemporal sampling will also enable the exploration of ecologically notable patterns, such as seasonal changes in inversion frequencies and the range and persistence of ecotypes. To link the variable genotypes observed in this study with functional changes or other adaptation, phenotypic data should routinely be included in collections conducted as part of vector surveillance. Beyond increased spatiotemporal sampling, long-read sequencing of ecotypes and resistance alleles will enhance our understanding of gene flow in this major malaria vector.

Materials and methods

Population sampling

In 2017, we circulated an open call to vector biologists working in Africa to establish a baseline understanding of genomic diversity and population structure in *Anopheles funestus*. We received mosquito carcases or DNA collected from 13 countries between 2014 and 2018 (Fig. 1A and table S1). Specimens were collected indoors and outdoors using a variety of methods (human landing catch, pyrethroid spraying, manual aspiration, CDC light traps, and larval dipping followed by rearing to adulthood) and most were stored on silica gel after collection. Specimens were morphologically identified by collectors as *An. funestus*. The majority of specimens are females and comprise whole mosquitoes, but the dataset includes a small number of males and partial specimens (e.g., head/thorax or abdomen only). No phenotypic characterisation (e.g., bioassay outcomes, inversion karyotypes) was carried out on the specimens.

Additionally, a total of 75 dry pinned or ethanol stored historic specimens were selected from the anopheline collections at the London Natural History Museum (NHMUK) and Institut de Recherche pour le Développement in Montpellier (IRDFR) (table S4). These samples were collected from 10 countries between 1927 and 1973 (fig. S11C). Most samples were labeled as *An. funestus*, however 30 turned out to be *An. leesoni*, *An. rivulorum,* and another unknown Rivulorum Subgroup species, as observed from mitochondrial DNA ML tree clustering with previously published full *Anopheles* mitogenomes (fig. S11E).

Whole-genome sequencing

Samples that arrived at the Wellcome Sanger Institute as mosquitoes were typically non-destructively extracted using Buffer C (*45*) and the lysates were purified using Qiagen MinElute kits (table S1). A small number of samples were extracted with Buffers A or G (*45*). Samples provided by contributors as DNA had been previously extracted in a variety of ways including CTAB and Qiagen DNeasy kits. All DNAs were quantified using Quant-iT picogreen dsDNA assays (ThermoFisher Scientific) following manufacturer protocols. Every DNA extract was subjected to a species-diagnostic polymerase chain reaction (PCR) (*63*) and the majority showed the expected band size for *An. funestus*. A small number of samples were sequenced despite unclear results for the band-size based assay, but typically did not pass QC, as they turned out to be different species based on mitochondrial genome analysis (figs. S1A, S11E). Samples that had at least 70 ng of DNA were submitted for standard library preparation, which included shearing to 450 bp using a Covaris LE220 instrument, purification by SPRIselect Beads on an Agilent Bravo WorkStation, library construction using a custom protocol for the NEBNext Ultra II DNA Library Prep Kit for Illumina on an Agilent Bravo Workstation, tagging with KAPA HiFi HotStart ReadyMix and custom Integrated DNA Technologies (IDT) primers with Illumina UDI 1-96 barcodes, library quantification by quantitative PCR (qPCR), and finally pooling libraries in equimolar amounts before sequencing (*64*). For more than 200 samples with < 30 ng of DNA, libraries were prepared using an established low input method designed for laser capture microdissection that uses the NEBNextUltra II Fragmentase System (*65*). If samples had between 30 and 70 ng of DNA, they went for one of these two library preparation methods (table S1). Each pool, containing on average 30 individual samples, was sequenced on three lanes of an Illumina HiSeq X10 platform using PE150 kits, aiming for 30x coverage per individual.

Historic samples were extracted and sequenced following a previously described method (*45*). Briefly, samples were extracted with the same type of Buffer (mostly C, some A and G) but with shorter incubation times, and double-stranded library preparation was performed using the same NEBNext Ultra II DNA Library Prep Kit for Illumina with modifications to retrieve short and damaged DNA fragments. Libraries were then sequenced on an Illumina HiSeq or NovaSeq instrument using PE75 and PE150 kits.

Sequence data processing and variant calling

Alignment and genotyping was performed with the MalariaGEN alignment and genotyping pipelines with their default parameters (*66*, *67*). For each sample, sequencing reads were aligned to the AfunGA1 reference genome (*18*) using bwa mem v0.7.15 (*68*) and all alignments were post-processed and combined across lanes with samtools v1.4.1 (*69*) and Picard v2.9.2 (*70*). PCR duplicates were marked with the biobambam v2.0.73 (*71*) bammarkduplicates command. Reads were realigned

around indels with GATK v3.7.0 (*72*) IndelRealigner, with target intervals generated for each sample separately. Variant calling was performed using GATK v3.7-0 (*72*) UnifiedGenotyper with all possible substitutions across all non-N sites of the reference genome marked as–alleles. Resulting VCF files were converted to a zarr format using the scikit-allel v1.2.1 (*73*) vcf_to_zarr function. Links to raw sequenced library fasta files, mapped BAM files, and both variant call files (VCF, zarr) per sample are listed in table S1.

Historic samples were preprocessed with the ancient DNA pipeline EAGER v2.3.5 (*74*), which performs adapter trimming, merging of paired-end reads if there is a ≥11 bp overlap, alignment with bwa mem v0.7.17 (*68*), removal of PCR duplicates and unmapped reads. Variant calls are generated as above. All libraries show characteristics of ancient DNA, such as C>T 5′ and G>A 3′ substitutions and short reads without prior shearing, highlighting we are indeed generating data from old mosquito DNA (table S4).

Sample and population QC

All 838 modern sequenced samples underwent QC, which assessed median coverage, fraction of genome covered, divergence from the reference genome, estimated contamination percentage, sex determination, and replication likelihood. Any sample that failed to meet our specified parameter thresholds was excluded from further analysis. Sample failures appeared to be random, irrespective of DNA extraction type or library preparation method (table S1). The thresholds we used are: a minimum of 10x median coverage, at least 85% of the reference genome sites covered by at least 1 read, at most 4% divergence from the reference genome (computed as the fraction of non-reference alleles), a maximum of 4.5% estimated proportion of sites affected by cross-contamination, a ratio of modal coverage on the X chromosome to that on the 3RL chromosome of between 0.4 and 0.6 (male) or between 0.8 and 1.2 (female), and a genetic distance of at least 0.006 between all pairs of individuals, to exclude excessively similar samples (fig. S1A; supplementary text). In total, 665 individuals (79%) passed the QC steps outlined above. Out of the 173 samples that were removed, 36 fell outside the Funestus Subgroup mitochondrial haplogroup (fig. S11E).

After this per-individual QC, we performed QC on the dataset as a whole, to check for outliers and potential sample swaps. Dataset QC was performed on all samples passing the first QC step, using PCA of the 2L chromosomal arm, which in *An. funestus* does not contain any large, common chromosomal inversions (*19*). The identification of outliers within the dataset was based on their isolation from other samples in the first 11 principal components. Identified outliers were removed and the PCA was recomputed on the remaining samples, until no outliers were identified. We excluded nine samples in two rounds of PCA, resulting in a total of 656 samples remaining for further analysis (table S1).

Historic samples

Historic samples have varied, but generally low coverage and hence we relaxed the QC thresholds and only applied the divergence filter; we excluded 30 individuals with > 4% divergence from the reference genome and retained 45 (table S4). As a dataset QC substitute, we performed a PCA with historic and modern samples (fig. S11C); all individuals fell within the expected clusters.

Public Datasets

For analyses including publicly available samples from the Funestus Subgroup (*30*) and the Folonzo and Kiribina ecotypes (*23*), we aligned all samples from these studies to the AfunGA1 reference genome and performed variant calling and QC as described above. We compared gene drive target availability in *An. funestus* with 762 Gambiae Complex mosquitoes (*16*) and assessed variation in the *dsx* target in 3,081 individuals from the Gambiae Complex (*75*).

Site filters

Following QC, we implemented a site filtering procedure to address the inherent variation along the genome in our ability to confidently call genotypes. We computed various site statistics from the data of all females passing sample QC. We generated two distinct site filters, namely the static-cutoff (sc) and decision-tree (dt) filters (fig. S1, C and D, and table S2; Supplementary Text).

For most of our analyses we used the sc filter, which retains sites with mean genotype quality (GQ) of at least 80, mean mapping quality (MQ) of at least 50, and median genomic coverage of samples with data at the focal site between 30x and 40x (fig. S1C). Depending on the analysis, further filtering based on the fraction of samples missing data was completed. These thresholds were chosen by considering the observed distributions of these statistics (fig. S1C).

For all haplotype-based analyses, we used the dt filter because we completed haplotype phasing using this filter. The decision tree was trained on 15 lab colony crosses and approximately 2,000 wild-caught *An. coluzzii* and *An. gambiae* mosquitoes (*76*). Mendelian errors were computed for the colony crosses and split into training and validation data. The decision tree inputs were 15 site summary statistics computed on the wild-caught mosquitoes and the optimal tree was selected using the validation data. Because the decision tree takes only site summary statistics as input, the trained model can be applied to a different population, a different species, or a different reference genome. We applied the trained model to the site summary statistics computed from all female samples passing QC to generate the dt site filter for *An. funestus*. It resulted in fewer variant sites than the sc filter, though overall concordance was high (fig. S1D and table S2).

Variant annotation

We extracted features from the AfunGA1 reference genome, using annotations from Vectorbase gff3 version 61 (*77*). Within each canonical coding sequence, we assess the effect of all possible SNPs, classify them as 'synonymous' or 'non-synonymous' mutations and record the amino acid changes for the latter category, using the CodonTable module in biopython (*78*). The AfunGA1 annotation gff3 file has not been manually curated, and we have noted some inaccuracies (supplementary text).

Haplotype phasing

We performed haplotype phasing on genomic sites that met three criteria: (i) they passed our dt site filtering process, (ii) were biallelic in our dataset, and (iii) contained the reference allele. To manage this large dataset, we first created interval tables for each chromosome, which defined intervals of 200,000 SNPs, with a 40,000 SNP overlap between adjacent intervals. We also incorporated a genetic map into the phasing process, which detailed recombination rates (2.0 cM/Mb for euchromatic and 0.5 cM/Mb for heterochromatic regions) inferred from average values in *An. gambiae (16)*. The haplotype phasing process was carried out using a specialized pipeline developed by the Broad Institute's Data Engineering team (*79*), with computational tasks conducted on the Terra platform (*80*). We applied two phasing methods: read-backed phasing conducted with WhatsHap v1.0 (*81*) and statistical phasing carried out using SHAPEIT4 v4.2.1 (*82*), successfully phasing all 656 individuals that passed our sample and dataset QC.

CNV calling

In the 2RL:8.2-9.8 Mbp region surrounding the *rp1* locus, we conducted copy number variation (CNV) analysis following the method described in Lucas *et al.* (*83*). Briefly, for each individual, we record read counts in 300 bp non-overlapping windows and normalize by the per-individual mean number of reads in genome-wide autosomal 300 bp windows, stratified by the GC content. These normalized coverage values were used as the observations in a Gaussian Hidden Markov Model (HMM) with the copy number states as hidden variables. CNVs were defined as having at least five consecutive 300 bp windows with elevated

HMM-predicted copy number states. Using unique patterns of discordant read pairs and split reads at the CNV breakpoints, we manually characterized nine CNV alleles. We say an identified CNV allele is present in an individual if we find at least two supporting diagnostic reads.

Cohort definitions

For the 656 samples that passed all QC filtering we defined two sets of cohorts: *geographic cohorts* based on collection metadata, and *PCA cohorts* based on observed structure in PCA projections (Fig. 1 and table S1). Geographic cohorts were defined following the same strategy used in MalariaGEN Vector Observatory (*84*), where we use the coarsest administrative subdivision within each country to assign samples to cohorts based on their collection location. We did not incorporate a time component in the cohort division, because its influence on the structure of this sampleset is vastly outweighed by the location component and all samples were collected within four years of each other. The PCA cohorts were defined based on the structure observed in the PC projection of chromosome arm 2L (Fig. 1A). Each PCA cohort contains one or multiple geographical cohorts in their entirety, with the exception of GH-N. In this case, nearly all (35/36) samples formed a separate cluster on the PCA plot and were assigned to the PCA cohort North Ghana, while one sample clustered with the Equatorial PCA cohort and was not assigned to any PCA cohort. PCA cohorts were used in analyses where the advantage of a larger sample size outweighs the disadvantage of reduced homogeneity within a cohort, e.g., analyses where cohorts are further divided by inversion karyotype.

Sample subsets

As a default, we perform analyses on all individuals that passed QC and refer to them as subset_1; if specific cohorts are excluded from subset_1, this is stated explicitly. We observed an excess of heterozygous calls in samples with median coverage below 20x and suspect this is a technical artefact (Fig. 1B and fig. S2A). Analyses that estimate genetic diversity, e.g., nucleotide diversity, are sensitive to samples with outlier heterozygosity, so these analyses were performed only on individuals with median coverage ≥20x (619 samples), referred to as subset_2. Some analyses (e.g., Tajima's *D*) are additionally affected by cohort sizes. For those analyses, we select 30 samples from each geographic cohort with median coverage ≥20x and the lowest contamination estimates. These 390 samples are referred to as subset_3. Four geographic cohorts (CF, KE-W, TZ, MW), have fewer than 30 samples and are thus not represented in subset_3.

Population genetic and selection analysis

Most analyses were performed within the MalariaGEN computational environment using scikit-allel v1.3.5-8 (*73*) functions. Analyses were performed using the sc filter, unless stated otherwise. Maps were generated with SciTools/cartopy v0.22.0 (*85*), and other plots were generated with matplotlib (*86*).

PCA

PCAs were computed using the pca function on a random sample of typically 200,000 biallelic sites within the specified region, with a minor allele frequency ≥0.01 and less than 5% of samples with missing genotype calls.

The sliding window PCA [inspired by (*87*)] uses the pca function within windows of 5,000 variants, moving with 1,000 variant steps. We selected nearly-biallelic variants (defined as having second minor allele frequency ≤0.001) with a minor allele frequency ≥0.02, allowing for at most 10% of samples with missing data, and downsampled to an average of 5,000 variants per Mbp. This subsetting of variants was necessary to include sufficient sites along the genome. Continuity between windows is ensured by minimizing the absolute distance of PC1 values between the focal and preceding window, flipping the PC1 axis if required.

For PCAs including historic samples, we excluded C>T and G>A substitutions from our variant selection to avoid confounding signals from DNA deamination. This correction was not implemented for the sliding window PCA, because it would substantially increase the computation time.

ADMIXTURE

We ran ADMIXTURE v1.3.0 (*88*) on 50,000 accessible (sc-filter) biallelic variants on the 2L chromosome arm with minor allele frequency ≥0.01, at most 0.1% missing genotype calls and including the reference allele. We selected three different sets of variants and ran ADMIXTURE with three different seeds for each set, with K ranging from 2 to 10. The violin plot was produced with seaborn 0.13.2 (*89*).

Nucleotide diversity, Watterson θ estimator, and Tajima's *D*

We estimated nucleotide diversity (π) as the mean pairwise differences for subset_2 individuals in each geographic cohort using the mean_pairwise_difference function. We split the genome into non-overlapping windows of 20,000 accessible sites with genotype calls for ≥90% of samples in each cohort and averaged the values within each window. Tajima's *D* is affected by sample size and smaller cohorts tend to have higher (less negative) Tajima's *D* values (data not shown). We partitioned the genome into windows as above, but using individuals from subset_3, ran the tajima_D function and averaged the values within each window. Watterson's theta can also be affected by sample size, so we counted the number of segregating sites amongst individuals from subset_3 using the count_segregating function in 100 kbp non-overlapping windows and divided by the number of sites and the 60th harmonic number (in subset_3 all cohorts contain 30 individuals, so 60 haplotypes) to obtain the Watterson θestimator per site.

F_{ST}

We followed Hudson's method (*90*) to estimate pairwise F_{ST} values. We computed F_{ST} using the hudson_fst function, which gives the numerator (number of differences between cohorts minus number of differences within cohorts) and denominator (number of differences between cohorts) for each site (for non-variable sites both are equal to 0 and hence the F_{ST} value for these sites is undefined). As per recommendation (*91*), we reported the "ratio of averages" per chromosomal arm (mean of numerators divided by the mean of denominators).

Isolation by distance

Genetic distance between geographic cohorts (where individuals of each ecotype from Burkina Faso (*23*) were grouped together into one geographic cohort) was computed as F_{ST} / (1 - F_{ST}) on chromosome arm 2L, including subset_2 individuals. Geographic distance between cohorts was computed as the great circle distance (in km) between the mean latitude and longitude of the individuals in the cohort and transformed by the natural logarithm. These transformations ensure an expected linear relationship between genetic and geographic distance under the island model in two dimensions (*92*). Linear regression between genetic and geographic distances was performed using the LinearRegression() function from scikit-learn v1.6.1 (*93*), which also reports the determinant of correlation R^2.

Heterozygous sites and ROH

We used the count_het function to count the number of heterozygous sites for every individual. We computed ROH only for females (21 males excluded) on genomic sites that passed the dt-filter using the roh_mhmm function with min_roh=100000 and default parameters otherwise. This function utilizes a multimodal hidden Markov model (MHMM) to estimate the positions and lengths of ROH.

Doubletons

We identified doubletons as alleles at accessible sites occurring exactly twice in subset_3 using the count_alleles function. We then identified

the number of doubletons shared within and between all geographic cohorts. Some of these doubletons (around 16% in *An. gambiae*) will be caused by recurrent mutations and thus not have shared ancestry (*94*).

In silico inversion karyotyping

Initial karyotyping was performed by identifying two horizontal thresholds in the sliding window PCA, partitioning the samples into three groups. To account for the effect of geographic structure, we used two different sets of thresholds. We determined the standard orientation by incorporating AfunGA1 as a fully homozygous sample, representing the 2R+^{t}+^{a}h, 3Rab, 3La orientation. We checked the results by computing PCA on the entire inversion region and assess whether samples clustered by karyotype; by this procedure we assigned karyotypes for GH-N, which is not included in the sliding window PCA, for all inversions except 3La.

Mitochondrial tree

We extracted reads mapping to scaffold_MT of AfunGA1 into separate BAM files and used bcftools 1.10.2 (*69*) mpileup to create consensus mitochondrial sequences for each sample. Together with publicly available Funestus Subgroup and sister species mitochondrial sequences (*29*, *30*) (table S1), we created a maximum likelihood tree using MAFFT v7.520 (*95*) and FastTree v2.1.11 (*96*), and visualized it using TreeViewer v2.2.0 (*97*).

Garud's H12 scan, haplotype trees, and variants potentially under selection

We conducted Garud's H12 scans (*38*) using the moving_garud_h function with cohort-specific window sizes (table S2). H12 was performed on sites passing the dt-filter. H12 scans were not presented for GH-N due to a high noise-to-signal ratio caused by a high number of ROH. For regions with H12 ≥ 0.4, we defined a peak as a 0.2 Mbp region centered on the highest value. Next, we calculated SNP allele frequencies for all non-synonymous variants (without site filter) for each gene in this region with minor allele frequency ≥ 0.055 using the snp_allele_frequencies function (table S3). Within the peaks, we identified genes or gene families likely to be under selection and constructed haplotype trees using the plot_haplotype_clustering function.

For historic samples, the genotype at potential insecticide resistance mutations was confirmed in IGV v.2.17.1 (*98*), as DNA deamination can be erroneously called as variants, especially at low coverage.

Gene drive targets

We identified gene drive targets in the reference genome as 20 bp sequences entirely within coding sequence, (using Vectorbase gff version 65 (*77*)) containing the protospacer adjacent motif (ending in -GG on the + strand or starting with CC- on the - strand). To account for resistance due to natural variation at the target site, we eliminated all gene drive targets that contained any variants in the group of samples under consideration (from subset_2).

Doublesex

The doublesex (*dsx*) target site is found at 2R:48,714,637-659 in AgamP4 and at 2RL:15,613,532-554 in AfunGA1. We searched for any variants at these sites in 3,081 Gambiae Complex individuals (*75*) and *An. funestus* subset_1 individuals.

REFERENCES AND NOTES

1. M. E. Sinka *et al.*, The dominant Anopheles vectors of human malaria in Africa, Europe and the Middle East: Occurrence data, distribution maps and bionomic précis. *Parasit. Vectors* **3**, 117 (2010). doi: 10.1186/1756-3305-3-117; pmid: 21129198
2. R. S. McCann *et al.*, Reemergence of Anopheles funestus as a vector of Plasmodium falciparum in western Kenya after long-term implementation of insecticide-treated bed nets. *Am. J. Trop. Med. Hyg.* **90**, 597–604 (2014). doi: 10.4269/ajtmh.13-0614; pmid: 24470562
3. W. Ntabaliba *et al.*, Life expectancy of Anopheles funestus is double that of Anopheles arabiensis in southeast Tanzania based on mark-release-recapture method. *Sci. Rep.* **13**, 15775 (2023). doi: 10.1038/s41598-023-42761-3; pmid: 37737323
4. E. O. Ogola *et al.*, Insights into malaria transmission among Anopheles funestus mosquitoes, Kenya. *Parasit. Vectors* **11**, 577 (2018). doi: 10.1186/s13071-018-3171-3; pmid: 30400976
5. I. Dia, M. Wamdaogo, D. Ayal, "Advances and Perspectives in the Study of the Malaria Mosquito Anopheles funestus" in *Anopheles Mosquitoes - New Insights into Malaria Vectors*, S. Manguin, Ed. (InTech, 2013). doi: 10.5772/55389
6. R. Djouaka *et al.*, Mapping the distribution of *Anopheles funestus* across Benin highlights a sharp contrast of susceptibility to insecticides and infection rate to *Plasmodium* between southern and northern populations. *Wellcome Open Res.* **1**, 28 (2016). doi: 10.12688/wellcomeopenres.10213.1; pmid: 28191507
7. B. J. Msugupakulya *et al.*, Changes in contributions of different Anopheles vector species to malaria transmission in east and southern Africa from 2000 to 2022. *Parasit. Vectors* **16**, 408 (2023). doi: 10.1186/s13071-023-06019-1; pmid: 37936155
8. N. Moiroux *et al.*, Changes in Anopheles funestus biting behavior following universal coverage of long-lasting insecticidal nets in Benin. *J. Infect. Dis.* **206**, 1622–1629 (2012). doi: 10.1093/infdis/jis565; pmid: 22966127
9. S. Sougoufara *et al.*, Biting by Anopheles funestus in broad daylight after use of long-lasting insecticidal nets: A new challenge to malaria elimination. *Malar. J.* **13**, 125 (2014). doi: 10.1186/1475-2875-13-125; pmid: 24678587
10. S. Omondi *et al.*, Late morning biting behaviour of Anopheles funestus is a risk factor for transmission in schools in Siaya, western Kenya. *Malar. J.* **22**, 366 (2023). doi: 10.1186/s12936-023-04806-w; pmid: 38037026
11. I. H. Nambunga *et al.*, Aquatic habitats of the malaria vector Anopheles funestus in rural south-eastern Tanzania. *Malar. J.* **19**, 219 (2020). doi: 10.1186/s12936-020-03295-5; pmid: 32576200
12. D. Ayala, A. Ullastres, J. González, Adaptation through chromosomal inversions in Anopheles. *Front. Genet.* **5**, 129 (2014). doi: 10.3389/fgene.2014.00129; pmid: 24904633
13. D. W. Wangrawa, J. O. Odero, F. Baldini, F. Okumu, A. Badolo, Distribution and insecticide resistance profile of the major malaria vector *Anopheles funestus* group across the African continent. *Med. Vet. Entomol.* **38**, 119–137 (2024). doi: 10.1111/mve.12706; pmid: 38303659
14. P. G. Pinda *et al.*, Comparative assessment of insecticide resistance phenotypes in two major malaria vectors, Anopheles funestus and Anopheles arabiensis in south-eastern Tanzania. *Malar. J.* **19**, 408 (2020). doi: 10.1186/s12936-020-03483-3; pmid: 33176805
15. G. D. Weedall *et al.*, An Africa-wide genomic evolution of insecticide resistance in the malaria vector Anopheles funestus involves selective sweeps, copy number variations, gene conversion and transposons. *PLOS Genet.* **16**, e1008822 (2020). doi: 10.1371/journal.pgen.1008822; pmid: 32497040
16. The Anopheles gambiae 1000 Genomes Consortium, Genetic diversity of the African malaria vector Anopheles gambiae. *Nature* **552**, 96–100 (2017).
17. E. Bier, Gene drives gaining speed. *Nat. Rev. Genet.* **23**, 5–22 (2022). doi: 10.1038/s41576-021-00386-0; pmid: 34363067
18. D. Ayala *et al.*, DNA Pipelines collective, S. McCarthy, D. Neafsey, A. Makunin, M. Lawniczak, The genome sequence of the malaria mosquito, Anopheles funestus, Giles, 1900. *Wellcome Open Research*, doi: (2022). doi: 10.12688/wellcomeopenres.18445.1
19. I. Sharakhov *et al.*, A microsatellite map of the African human malaria vector Anopheles funestus. *J. Hered.* **95**, 29–34 (2004). doi: 10.1093/jhered/esh011; pmid: 14757727
20. R. C. Griffiths, S. Tavaré, The age of a mutation in a general coalescent tree. *Commun. Stat. Stoch. Models* **14**, 273–295 (1998). doi: 10.1080/15326349808807471
21. A. V. Stronen, A. J. Norman, E. Vander Wal, P. C. Paquet, The relevance of genetic structure in ecotype designation and conservation management. *Evol. Appl.* **15**, 185–202 (2022). doi: 10.1111/eva.13339; pmid: 35233242
22. C. Costantini, N. Sagnon, E. Ilboudo-Sanogo, M. Coluzzi, D. Boccolini, Chromosomal and bionomic heterogeneities suggest incipient speciation in Anopheles funestus from Burkina Faso. *Parassitologia* **41**, 595–611 (1999). pmid: 10870569
23. S. T. Small *et al.*, Standing genetic variation and chromosome differences drove rapid ecotype formation in a major malaria mosquito. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2219835120 (2023). doi: 10.1073/pnas.2219835120; pmid: 36881629
24. M. Boddé, P. Korlevic, M. Lawniczak, Interactive sliding window PCAs for An. Funestus, Zenodo (2024); https://doi.org/10.5281/ZENODO.13993020.
25. J. M. Riveron *et al.*, Genome-Wide Transcription and Functional Analyses Reveal Heterogeneous Molecular Mechanisms Driving Pyrethroids Resistance in the Major Malaria Vector *Anopheles funestus* Across Africa. *G3* **7**, 1819–1832 (2017). doi: 10.1534/g3.117.040147; pmid: 28428243
26. K. R. Maynard, S. S. McCarthy, E. Sheldon, H. W. Horch, Developmental and adult expression of semaphorin 2a in the cricket Gryllus bimaculatus. *J. Comp. Neurol.* **503**, 169–181 (2007). doi: 10.1002/cne.21392; pmid: 17480023
27. J. Bonner, T. P. O'Connor, Semaphorin function in the developing invertebrate peripheral nervous system. *Biochem. Cell Biol.* **78**, 603–611 (2000). doi: 10.1139/o00-076; pmid: 11103951
28. A. P. Michel *et al.*, Rangewide population genetic structure of the African malaria vector Anopheles funestus. *Mol. Ecol.* **14**, 4235–4248 (2005). doi: 10.1111/j.1365-294X.2005.02754.x; pmid: 16313589
29. C. M. Jones *et al.*, Complete Anopheles funestus mitogenomes reveal an ancient history of mitochondrial lineages and their distribution in southern and central Africa. *Sci. Rep.* **8**, 9054 (2018). doi: 10.1038/s41598-018-27092-y; pmid: 29899497

30. S. T. Small *et al.*, Radiation with reticulation marks the origin of a major malaria vector. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 31583–31590 (2020). doi: 10.1073/pnas.2018142117; pmid: 33262284
31. M. Wellenreuther, L. Bernatchez, Eco-Evolutionary Genomics of Chromosomal Inversions. *Trends Ecol. Evol.* **33**, 427–440 (2018). doi: 10.1016/j.tree.2018.04.002; pmid: 29731154
32. D. Ayala, R. F. Guerrero, M. Kirkpatrick, Reproductive isolation and local adaptation quantified for a chromosome inversion in a malaria mosquito. *Evolution* **67**, 946–958 (2013). doi: 10.1111/j.1558-5646.2012.01836.x; pmid: 23550747
33. W. M. Guelbeogo, N. Sagnon, F. Liu, N. J. Besansky, C. Costantini, Behavioural divergence of sympatric Anopheles funestus populations in Burkina Faso. *Malar. J.* **13**, 65 (2014). doi: 10.1186/1475-2875-13-65; pmid: 24559382
34. W. M. Guelbeogo *et al.*, Seasonal distribution of Anopheles funestus chromosomal forms from Burkina Faso. *Malar. J.* **8**, 239 (2009). doi: 10.1186/1475-2875-8-239; pmid: 19857258
35. C. Sangbakembi-Ngounou *et al.*, Temporal and biting dynamics of the chromosomal inversion 2La in the malaria vectors Anopheles gambiae and Anopheles coluzzii in Bangui, Central African Republic. *Front. Ecol. Evol.* **10**, 986925 (2022). doi: 10.3389/fevo.2022.986925
36. R. F. Guerrero, F. Rousset, M. Kirkpatrick, Coalescent patterns for chromosomal inversions in divergent populations. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **367**, 430–438 (2012). doi: 10.1098/rstb.2011.0246; pmid: 22201172
37. X. Grau-Bové *et al.*, Evolution of the Insecticide Target Rdl in African Anopheles Is Driven by Interspecific and Interkaryotypic Introgression. *Mol. Biol. Evol.* **37**, 2900–2917 (2020). pmid: 32449755
38. N. R. Garud, P. W. Messer, E. O. Buzbas, D. A. Petrov, Recent selective sweeps in North American Drosophila melanogaster show signatures of soft sweeps. *PLOS Genet.* **11**, e1005004 (2015). doi: 10.1371/journal.pgen.1005004; pmid: 25706129
39. R. H. Ffrench-Constant, The molecular genetics of insecticide resistance. *Genetics* **194**, 807–815 (2013). doi: 10.1534/genetics.112.141895; pmid: 23908373
40. J. Taylor-Wells, A. K. Jones, "Variations in the Insect GABA Receptor, RDL, and Their Impact on Receptor Pharmacology" in *Advances in Agrochemicals: Ion Channels and G Protein-Coupled Receptors (GPCRs) as Targets for Pest Control*, **vol. 1265** of *ACS Symposium Series* (American Chemical Society, 2017), pp. 1–21.
41. C. S. Wondji *et al.*, Identification and distribution of a GABA receptor mutation conferring dieldrin resistance in the malaria vector Anopheles funestus in Africa. *Insect Biochem. Mol. Biol.* **41**, 484–491 (2011). doi: 10.1016/j.ibmb.2011.03.012; pmid: 21501685
42. J. O. Odero *et al.*, Discovery of knock-down resistance in the major African malaria vector Anopheles funestus. *Mol. Ecol.* **33**, e17542 (2024). doi: 10.1111/mec.17542; pmid: 39374937
43. D. B. Wilson, A. G. Robertson, DDT air-spray in malaria control in East Africa. *Trans. R. Soc. Trop. Med. Hyg.* **40**, 823–850 (1947). doi: 10.1016/0035-9203(47)90040-0; pmid: 20262306
44. R. C. M. Thomson, The effects of house spraying with pyrethrum and with DDT on Anopheles gambiae and A. melas in West Africa. *Bull. Entomol. Res.* **38**, 449–464 (1947). doi: 10.1017/S0007485300022276; pmid: 18918650
45. P. Korlević *et al.*, A Minimally Morphologically Destructive Approach for DNA Retrieval and Whole-Genome Shotgun Sequencing of Pinned Historic Dipteran Vector Species. *Genome Biol. Evol.* **13**, evab226 (2021). doi: 10.1093/gbe/evab226; pmid: 34599327
46. R. Djouaka *et al.*, Multiple insecticide resistance in an infected population of the malaria vector Anopheles funestus in Benin. *Parasit. Vectors* **9**, 453 (2016). doi: 10.1186/s13071-016-1723-y; pmid: 27531125
47. V. Ramakrishna, R. Elliott, Normal resistance-level of Anopheles funestus Giles to insecticides. *Nature* **179**, 1140–1141 (1957). doi: 10.1038/1791140b0; pmid: 13430815
48. J. Hamon, S. Sales, P. Venard, J. Coz, J. Brengues, Others, Dieldrin-resistant Anopheles funestus in south-west Upper Volta. *Med. Trop.* **28**, 222–226 (1968).
49. J. Taylor-Wells, B. D. Brooke, I. Bermudez, A. K. Jones, The neonicotinoid imidacloprid, and the pyrethroid deltamethrin, are antagonists of the insect Rdl GABA receptor. *J. Neurochem.* **135**, 705–713 (2015). doi: 10.1111/jnc.13290; pmid: 26296809
50. O. E. Akinrinade, F. O. Agunbiade, R. Alani, O. O. Ayejuyo, Implementation of the Stockholm Convention on persistent organic pollutants (POPs) in Africa – progress, challenges, and recommendations after 20 years. *Environ. Sci. Adv.* **3**, 623–634 (2024). doi: 10.1039/D3VA00347G
51. T. Nolan, Control of malaria-transmitting mosquitoes using gene drives. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **376**, 20190803 (2021). doi: 10.1098/rstb.2019.0803; pmid: 33357060
52. K. Kyrou *et al.*, A CRISPR-Cas9 gene drive targeting doublesex causes complete population suppression in caged Anopheles gambiae mosquitoes. *Nat. Biotechnol.* **36**, 1062–1066 (2018). doi: 10.1038/nbt.4245; pmid: 30247490
53. A. Hammond *et al.*, Gene-drive suppression of mosquito populations in large cages as a bridge between lab and field. *Nat. Commun.* **12**, 4589 (2021). doi: 10.1038/s41467-021-24790-6; pmid: 34321476
54. A. M. Adja *et al.*, Contribution of Anopheles funestus, An. gambiae and An. nili (Diptera: Culicidae) to the perennial malaria transmission in the southern and western forest areas of Côte d'Ivoire. *Ann. Trop. Med. Parasitol.* **105**, 13–24 (2011). doi: 10.1179/136485910X12851868780388; pmid: 21294945
55. N. F. Kahamba *et al.*, Using ecological observations to improve malaria control in areas where Anopheles funestus is the dominant vector. *Malar. J.* **21**, 158 (2022). doi: 10.1186/s12936-022-04198-3; pmid: 35655190
56. H. E. Beck *et al.*, Present and future Köppen-Geiger climate classification maps at 1-km resolution. *Sci. Data* **5**, 180214 (2018). doi: 10.1038/sdata.2018.214; pmid: 30375988
57. L. Lochouarn, I. Dia, D. Boccolini, M. Coluzzi, D. Fontenille, Bionomical and cytogenetic heterogeneities of Anopheles funestus in Senegal. *Trans. R. Soc. Trop. Med. Hyg.* **92**, 607–612 (1998). doi: 10.1016/S0035-9203(98)90782-9; pmid: 10326101
58. E. O. Ogola, J. O. Odero, J. M. Mwangangi, D. K. Masiga, D. P. Tchouassi, Population genetics of Anopheles funestus, the African malaria vector, Kenya. *Parasit. Vectors* **12**, 15 (2019). doi: 10.1186/s13071-018-3252-3; pmid: 30621756
59. C. S. Clarkson *et al.*, The genetic architecture of target-site resistance to pyrethroid insecticides in the African malaria vectors Anopheles gambiae and Anopheles coluzzii. *Mol. Ecol.* **30**, 5303–5317 (2021). doi: 10.1111/mec.15845; pmid: 33590926
60. H. Ranson, N. Lissenden, Insecticide resistance in African anopheles mosquitoes: A worsening situation that needs urgent action to maintain malaria control. *Trends Parasitol.* **32**, 187–196 (2016). doi: 10.1016/j.pt.2015.11.010; pmid: 26826784
61. J. Hemingway *et al.*, Averting a malaria disaster: Will insecticide resistance derail malaria control? *Lancet* **387**, 1785–1788 (2016). doi: 10.1016/S0140-6736(15)00417-1; pmid: 26880124
62. Anopheles gambiae 1000 Genomes Consortium, Genome variation and population structure among 1142 mosquitoes of the African malaria vector species *Anopheles gambiae* and *Anopheles coluzzii*. *Genome Res.* **30**, 1533–1546 (2020). doi: 10.1101/gr.262790.120; pmid: 32989001
63. L. L. Koekemoer, L. Kamau, R. H. Hunt, M. Coetzee, A cocktail polymerase chain reaction assay to identify members of the Anopheles funestus (Diptera: Culicidae) group. *Am. J. Trop. Med. Hyg.* **66**, 804–811 (2002). doi: 10.4269/ajtmh.2002.66.804; pmid: 12224596
64. I. F. Bronner, M. A. Quail, Best Practices for Illumina Library Preparation. *Curr. Protoc. Hum. Genet.* **102**, e86 (2019). doi: 10.1002/cphg.86; pmid: 31216112
65. P. Ellis *et al.*, Reliable detection of somatic mutations in solid tissues by laser- capture microdissection and low-input DNA sequencing. *Nat. Protoc.* **16**, 841–871 (2021). doi: 10.1038/s41596-020-00437-6; pmid: 33318691
66. A. Miles, J. Stalker, G. Grant, Mosquito short read alignment pipeline specification (Version 1.2.1) (2020). https://github.com/malariagen/pipelines/blob/c1531ef6120021106cb0159150a297a5d8473e07/docs/specs/short-read-alignment-vector.md.
67. A. Miles, J. Stalker, Mosquito SNP genotyping pipeline specification (Version 1.4.1) (2020). https://github.com/malariagen/pipelines/blob/a910bde50e824623a709d86465658d9db44e1ccd/docs/specs/snp-genotyping-vector.md.
68. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, *arXiv [q-bio.GN]* (2013). https://arxiv.org/abs/1303.3997.
69. P. Danecek *et al.*, Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021). doi: 10.1093/gigascience/giab008; pmid: 33590861
70. Picard, https://broadinstitute.github.io/picard/.
71. G. Tischler, S. Leonard, biobambam: Tools for read pair collation based algorithms on BAM files. *Source Code Biol. Med.* **9**, 13 (2014). doi: 10.1186/1751-0473-9-13
72. G. van der Auwera, B. D. O'Connor, *Genomics in the Cloud* (O'Reilly Media, 2020).
73. A. Miles, pyup.io bot, M. F. Rodrigues, P. Ralph, J. Kelleher, M. Schelker, R. Pisupati, S. Rae, Tim, scikit-allel, zenodo (2024). https://doi.org/10.5281/zenodo.597309.
74. J. A. Fellows Yates *et al.*, Reproducible, portable, and efficient ancient genome reconstruction with nf-core/eager. *PeerJ* **9**, e10947 (2021). doi: 10.7717/peerj.10947; pmid: 33777521
75. The Anopheles gambiae 1000 Genomes Consortium, Ag1000G phase 3 SNP data release, *MalariaGEN.* (2021). https://www.malariagen.net/data_package/ag1000g-phase3-snp.
76. G. E. N. Malaria, Site filtering, *Ag3 sequencing and variant calling methods*. https://malariagen.github.io/vector-data/ag3/methods.html.
77. G. I. Giraldo-Calderón *et al.*, VectorBase.org updates: Bioinformatic resources for invertebrate vectors of human pathogens and related organisms. *Curr. Opin. Insect Sci.* **50**, 100860 (2022). doi: 10.1016/j.cois.2021.11.008; pmid: 34864248
78. P. J. A. Cock *et al.*, Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009). doi: 10.1093/bioinformatics/btp163; pmid: 19304878
79. A. Miles, J. Brenas, Mosquito phasing pipeline specification (Version 1.0.0) (2021). https://github.com/malariagen/pipelines/blob/VectorCNV_v1.0.0/docs/specs/phasing-vector.md.
80. T. Garfinkel, B. Pfaff, J. Chow, M. Rosenblum, D. Boneh, "Terra: a virtual machine-based platform for trusted computing" in *Proceedings of the Nineteenth ACM Symposium on Operating Systems Principles* (Association for Computing Machinery, 2003), pp. 193–206. doi: 10.1145/945445.945464
81. M. Martin *et al.*, WhatsHap: fast and accurate read-based phasing. bioRxiv 085050 [Preprint] (2016); https://doi.org/10.1101/085050.
82. O. Delaneau, J.-F. Zagury, M. R. Robinson, J. L. Marchini, E. T. Dermitzakis, Accurate, scalable and integrative haplotype estimation. *Nat. Commun.* **10**, 5436 (2019). doi: 10.1038/s41467-019-13225-y; pmid: 31780650

83. E. R. Lucas *et al.*, Whole-genome sequencing reveals high complexity of copy number variation at insecticide resistance loci in malaria mosquitoes. *Genome Res.* **29**, 1250–1261 (2019). doi: 10.1101/gr.245795.118; pmid: 31345938
84. C. Clarkson, Grouping samples into cohorts (2022); https://anopheles-genomic-surveillance.github.io/workshop-1/module-4-vgsc-snps.html.
85. P. Elson *et al.*, SciTools/cartopy, Zenodo (2024); https://doi.org/10.5281/zenodo.1182735.
86. J. D. Hunter, Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9**, 90–95 (2007). doi: 10.1109/MCSE.2007.55
87. H. Li, P. Ralph, Local PCA Shows How the Effect of Population Structure Differs Along the Genome. *Genetics* **211**, 289–304 (2019). doi: 10.1534/genetics.118.301747; pmid: 30459280
88. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009). doi: 10.1101/gr.094052.109; pmid: 19648217
89. M. Waskom, seaborn: Statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021). doi: 10.21105/joss.03021
90. R. R. Hudson, M. Slatkin, W. P. Maddison, Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**, 583–589 (1992). doi: 10.1093/genetics/132.2.583; pmid: 1427045
91. G. Bhatia, N. Patterson, S. Sankararaman, A. L. Price, Estimating and interpreting FST: The impact of rare variants. *Genome Res.* **23**, 1514–1521 (2013). doi: 10.1101/gr.154831.113; pmid: 23861382
92. F. Rousset, Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics* **145**, 1219–1228 (1997). doi: 10.1093/genetics/145.4.1219; pmid: 9093870
93. F. Pedregosa *et al.*, Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
94. J. J. Reynolds, V. Koufopanou, A. Burt, Identifying single origin rare variants in population genomic data, bioRxiv (2025). https://doi.org/10.1101/2025.02.13.638116.
95. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013). doi: 10.1093/molbev/mst010; pmid: 23329690
96. M. N. Price, P. S. Dehal, A. P. Arkin, FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26**, 1641–1650 (2009). doi: 10.1093/molbev/msp077; pmid: 19377059
97. G. Bianchini, P. Sánchez-Baracaldo, TreeViewer: Flexible, modular software to visualise and manipulate phylogenetic trees. *Ecol. Evol.* **14**, e10873 (2024). doi: 10.1002/ece3.10873; pmid: 38314311
98. J. T. Robinson *et al.*, Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011). doi: 10.1038/nbt.1754; pmid: 21221095
99. An. funestus phase 1.0 data release. [Preprint] (2025). https://malariagen.github.io/vector-data/af1/af1.0.html.
100. M. Boddé, A. Makunin, P. Korlević, J. Nwezeobi, M. Lawniczak, analysis code for Genomic diversity of An. funestus. Zenodo (2025). https://doi.org/10.5281/ZENODO.15773526.

ACKNOWLEDGMENTS

We thank K. von Wyschetzki and S. Goncalves for assistance in sequencing submissions and Sanger's Scientific Operation Teams for carrying out all library generation and sequencing on data presented here. We also express our gratitude to the entomologists who, decades ago, collected and identified the *Anopheles* specimens used in this study: Adam, Brunhes, Buxton, Garnham, Hamon, Leeson, Lewis, Pichon, Rickenbach, Stymes, and all supporting fieldwork and collection assistants whose names did not fit on the specimen labels. **Funding:** Wellcome Trust Quinquennial awards 206194 and 220540/Z/20/A (M.K.N.L., J.N., A.Ma., P.K., and D.K.); Wellcome 4 year PhD studentship RG92770 (M.B.); Bill & Melinda Gates Foundation INV-001927 (A.Mi., L.H.); Wellcome award WT207492 (R.D.); National Research Foundation of South Africa SRUG2203311457 (L.K.); Gates Foundation Grand Challenges grant no. INV024969 and Investment Grant OPP1210769 (B.P., E.O.); Wellcome Trust International Intermediate fellowship 222005/Z/20/Z (D.P.T.); Wellcome Trust Senior Research fellowship 217188/Z/19/Z (C.S.W.); ANR, France ANR-18-CE35-0002-01–WILDING (D.A.); Bill and Melinda Gates Foundation and Obra Social "la Caixa" Partnership for the Elimination of Malaria in Southern Mozambique INV-008483 (K.P., M.M.); UK Medical Research Council MR/T001070/1 (E.R.L., S.J.N., awarded to M. Donnelly); US National Institute of Health R01AI116811 (E.R.L., S.J.N., awarded to M. Donnelly). **Author contributions:** Conceived and coordinated the study: M.K.N.L.; Provided samples and secured funding to do so: O.A.-E., M.G., E.K., M.M., C.S., S.D., L.L.K., E.M., E.O., F.O., K.P., D.P.T., C.S.W., D.A.; Carried out DNA extractions: P.K., K.L.; Secured funding to support sequencing and analysis: D.K., M.K.N.L.; Data curation lead: J.N.; Data curation contributions: L.H., A.Mi.; Data analysis lead: M.B.; Data analysis major contributions: J.N., P.K.; Data analysis minor contributions: A.Ma., S.B., E.L., I.L., S.N., J.O.O., R.D.; Writing: M.B., M.K.N.L.; Writing major contributions: P.K., J.N.; Writing minor contributions: all authors; Supervision of lead authors: D.A., R.D., A.Mi., M.K.N.L. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** ENA accession IDs for each sample can be found in table S1 (modern) and table S4 (historic). All metadata associated with each sample is in the same tables. These data are released as part of the ENA BioProject PRJEB2141, the Vector Observatory Project. Processed data can be accessed through the malariagen_data python package at https://malariagen.github.io/vector-data/af1/af1.0.html (*99*). Analysis code is available at github.com/malariagen/funestus_popgen_ms (*100*).

This research was funded in whole or in part by Wellcome (Quinquennial awards 206194 and 220540/Z/20/A, 4 year PhD studentship RG92770, award WT207492, International Intermediate fellowship 222005/Z/20/Z, and Senior Research fellowship 217188/Z/19/Z), a cOAlition S organization. The author will make the Author Accepted Manuscript (AAM) version available under a CC BY public copyright license.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adu3596
Supplementary Text; Figs. S1 to S12; Tables S1 to S4; References (*101*–*156*)

Submitted 6 November 2024; accepted 3 July 2025

10.1126/science.adu3596