HHS Public Access

Author manuscript

Science. Author manuscript; available in PMC 2026 March 02.

Published in final edited form as:

Science. 2025 September 18; 389(6766): eads3732. doi:10.1126/science.ads3732.

1206 genomes reveal origin and movement of *Aedes aegypti* driving increased dengue risk

Jacob E. Crawford[1,*,†], **Dario Balcazar**[2,3,4], **Seth Redmond**[5], **Noah H. Rose**[6,‡], **Henry A. Youd**[7], **Eric R. Lucas**[7], **Rusdiyah Sudirman Made Ali**[8], **Ashwaq Alnazawi**[9], **Athanase Badolo**[10], **Chun-Hong Chen**[11], **Luciano V. Cosme**[12], **Jennifer A. Henke**[13], **Kim Y. Hung**[13], **Susanne Kluh**[14], **Wei-Liang Liu**[11], **Kevin Maringer**[15], **Ademir Martins**[16], **María Victoria Micieli**[3], **Evlyn Pless**[2,§], **Aboubacar Sombié**[10], **Sinnathamby N. Surendran**[17], **Isra Wahid**[18], **Peter A. Armbruster**[19], **David Weetman**[7], **Carolyn S. McBride**[6], **Andrea Gloria-Soria**[4], **Jeffrey R. Powell**[2], **Bradley J. White**[1,†]

[1]Verily Life Sciences, South San Francisco, CA, USA

[2]Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT, USA

[3]Centro de Estudios Parasitológicos y de Vectores (CEPAVE) CONICET-Universidad Nacional de la Plata, La Plata, Argentina

[4]Department of Entomology, Center for Vector Biology & Zoonotic Diseases, The Connecticut Agricultural Experiment Station, New Haven, CT, USA

[5]Department of Epidemiology of Microbial Diseases, Yale School of Public Health, New Haven, CT, USA

[6]Department of Ecology and Evolutionary Biology and Princeton Neuroscience Institute, Princeton University, Princeton, NJ, USA

[7]Department of Vector Biology, Liverpool School of Tropical Medicine, Pembroke Place, Liverpool, UK

[8]National Research and Innovation Agency (BRIN), Jakarta Pusat, Indonesia

[9]Department of Epidemiology, Faculty of Public Health and Tropical Medicine, Jazan University, Jazan, Saudi Arabia

[*]**Corresponding author.** jacobcrawford@google.com.
[†]Present address: Google LLC, Mountain View, CA, USA.
[‡]Present address: Department of Ecology, Behavior, and Evolution, University of California, San Diego, La Jolla, CA, USA.
[§]Present address: Delfi Diagnostics, Baltimore, MD, USA.

Author contributions: Conceptualization: J.E.C., N.H.R., S.R., C.S.M., A.G.-S., J.R.P., B.J.W.; Formal analysis: J.E.C., D.B., S.R., N.H.R., H.Y., E.R.L., P.A.A.; Methodology: J.E.C., D.B., S.R., N.H.R., A.G.-S.; Project administration: J.E.C.; Resources: J.E.C., N.H.R., R.S.M.A., A.A., A.M., A.B., C.-H.C., L.V.C., J.A.H., K.Y.H., S.K., W.-L.L., K.M., M.V.M., E.P., A.S., S.N.S., I.W., P.A.A., A.G.-S., J.R.P.; Supervision: D.W., C.S.M., A.G.-S., J.R.P., B.J.W.; Visualization: J.E.C., D.B., S.R., N.H.R.; Writing – original draft: J.E.C., D.B., S.R., N.H.R., C.S.M., A.G.-S.; Writing – review & editing: All co-authors.

Competing interests: J.E.C. and B.J.W. report employment and equity ownership at Verily Life Sciences LLC, a for-profit company developing new technologies for mosquito control.

[10]Laboratory of Fundamental and Applied Entomology, Université Joseph Ki-Zerbo, Ouagadougou, Burkina Faso

[11]National Health Research Institutes, National Mosquito-Borne Disease Control Research Center and National Institute of Infectious Diseases and Vaccinology, Zhunan, Taiwan

[12]Department of Entomology, University of California, Riverside, Riverside, CA, USA

[13]Coachella Valley Mosquito and Vector Control District, Indio, CA, USA

[14]Greater Los Angeles County Vector Control District, Santa Fe Springs, CA, USA

[15]The Pirbright Institute, Pirbright, UK

[16]Laboratório de Fisiologia e Controle de Artrópodes Vetores, Instituto Oswaldo Cruz, Fiocruz, Rio de Janeiro, RJ, Brazil

[17]Department of Zoology, University of Jaffna, Jaffna, Sri Lanka

[18]Hasanuddin University Medical Research Centre (HUMRC), Center for Zoonotic and Emerging Diseases, Makassar, Indonesia

[19]Department of Biology, Georgetown University, Washington, DC, USA

Abstract

The emergence and global expansion of *Aedes aegypti* puts more than half of all humans at risk of arbovirus infection, but the origin of this mosquito and the impact of contemporary gene flow on arbovirus control are unclear. We sequenced 1206 genomes from 73 globally distributed locations. After evolving a preference for humans in Sahelian West Africa, the invasive subspecies *Ae. aegypti aegypti* (*Aaa*) emerged in the Americas after the Atlantic slave trade era and expanded globally. Recent back-to-Africa *Aaa* migration introduced insecticide resistance and anthropophily into regions with recent dengue outbreaks, raising concern that *Aaa* movement could increase arbovirus risk in urban Africa. These data underscore developing complexity in the fight against dengue, Zika, and chikungunya and provide a platform to further study this important mosquito vector.

Abstract

INTRODUCTION: Approximately 4 billion people are at risk of contracting the mosquito-transmitted disease dengue each year in tropical and subtropical regions and increasingly in temperate regions. An effective vaccine for dengue is not widely available, so controlling the primary mosquito vector *Aedes aegypti* is key to limiting the impact of dengue and other viral diseases transmitted by this mosquito, including Zika, chikungunya, and yellow fever. The ancestral subspecies of *Ae. aegypti* is a forest-dwelling ecological generalist, *Ae. aegypti formosus* (*Aaf*), that feeds on a variety of hosts in Africa, but viral transmission is driven by a globally invasive, human-preferring subspecies, *Ae. aegypti aegypti* (*Aaa*), that emerged more recently. Historical disease records suggest that *Ae. aegypti* emigrated from Africa to the Americas on ships during the Atlantic slave trade (AST), but the role that the AST played in the origin of invasive *Aaa* is not clear. Previous global population genetic surveys revealed populations both inside and

outside Africa distributed along an ancestry gradient between *Aaf* and *Aaa* that provided clues into the early stages of the domestication process leading to *Aaa*.

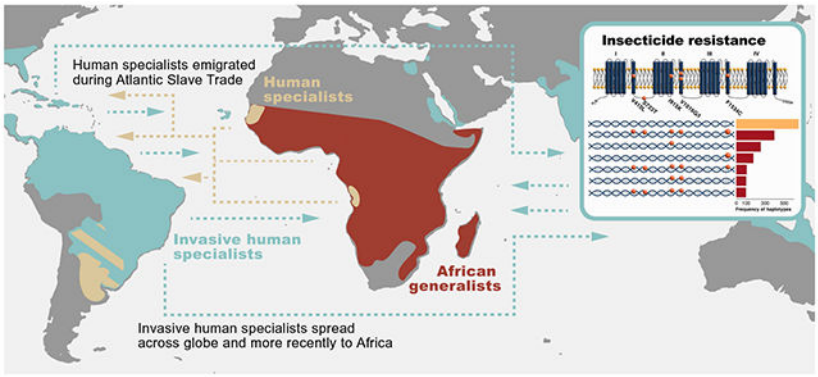
RATIONALE: Advances in whole-genome sequencing (WGS) and methods for inferring evolutionary patterns from WGS data have advanced the ability to resolve complex signals in population genetics data. However, WGS analysis of *Ae. aegypti* has been limited in large part by its genome that is both larger and more repetitive than most other insect pests. A WGS dataset including populations from both the ancestral subspecies *Aaf* and the invasive human-specialist *Aaa* would allow high-resolution inference of genetic signals that could produce a clearer understanding of historical shifts as well as contemporary movement and molecular changes that could impact public health.

RESULTS: In this work, we present the Aaeg1200 WGS dataset, including 1206 whole genomes from 73 locations throughout the distributions of *Aaf* and *Aaa*. We aligned sequencing reads to a chromosome-level genome reference sequence for analysis and identified sites of more than 141 million single-nucleotide polymorphisms (SNPs) , 53.3% of which are found exclusively in *Aaf*. To determine whether *Aaf*-like *Ae. aegypti* from Argentina is a recent introduction or relict from the AST, we fitted coalescent models to phased haplotypes and SNP frequency data and showed that Argentine *Ae. aegypti* split from African populations ~320 years ago around the time of the AST and that invasive *Aaa* split from Argentina ~100 years later. Representative invasive *Aaa* populations show signs of natural selection at several regions, suggesting a role for adaptation to new pathogens and feeding habits that may have enabled further expansion into new environments. On the basis of phased haplotype data, we show evidence for recent secondary contact among the subspecies that has, in some cases, resulted in extensive sharing of insecticide-resistance mutations and introduction of these mutations into Africa.

CONCLUSION: *Ae. aegypti* is a major threat to public health and also a useful model for the study of domestication, the evolution of human blood feeding, and neo-sex chromosome evolution, as well as additional basic and applied research questions. The Aaeg1200 WGS dataset was assembled to help address these and other questions to enable improved understanding of the molecular changes and evolutionary processes underlying this disease vector. Chemical insecticides are an essential tool in the fight against *Ae. aegypti*, and datasets such as Aaeg1200 will help understand and manage the spread of resistance and enable new tools to be developed in the fight to reduce the burden of dengue and other mosquito-borne viruses.

Graphical Abstract

Invasive *Aedes aegypti* global migration spreads insecticide resistance and increases arboviral risk. DNA sequence analysis revealed that the origin of invasive *Ae. aegypti aegypti* occurred after emigration to the Americas and the Caribbean, where the subspecies adapted to new environments. Mutations conferring resistance to common insecticides have emerged independently in multiple locations and spread through recent transcontinental migration, increasing arboviral risk in Africa, among other locations.



The mosquito *Aedes aegypti* is the primary vector of dengue, chikungunya, Zika, and yellow fever viruses in the tropics and subtropics. *Ae. aegypti* occurs globally, and its distribution continues to expand to temperate zones with a warming climate and increased urbanization (1), putting nearly 4 billion people at risk of contracting dengue each year (2). Over the past two decades, the World Health Organization (WHO) has recorded a 10-fold increase in global cases of dengue, with more than 12 million cases and 7700 deaths reported in the Americas and Caribbean (2, 3). Model estimates suggest as many as 390 million infections per year (4). Effective vaccines against these viruses are not widely available, so mosquito control is critical to reduce the impact of these diseases on human populations. Improved understanding of the genetic factors and evolutionary processes underlying the global spread of *Ae. aegypti* will improve the efficacy of existing control tools and facilitate development of new ones.

Ae. aegypti varies in ecological niche and preference for bloodmeal hosts. It is hypothesized that after originating in the Southwest Indian Ocean region, *Ae. aegypti* invaded continental Africa from the southeast approximately 85,000 years ago (5). It then spread across the continent, where it has inhabited forested areas of Africa, breeding in tree holes and other natural containers and feeding opportunistically on a variety of hosts for bloodmeals, including reptiles and a variety of mammals (6–12). In contrast to the predominantly forest-dwelling generalist form of *Ae. aegypti* in Africa, outside of Africa *Ae. aegypti* lives in close association with humans, living in and around human dwellings and preferring human bloodmeals (7, 9, 13). Two subspecies were defined (14) on the basis of these behavioral differences and phenotypic features, with the generalist named *Ae. aegypti formosus* (*Aaf*) and the human specialist named *Ae. aegypti aegypti* (*Aaa*). The two subspecies are genetically differentiated, although they remain reproductively compatible, and evidence points to *Aaf* as the ancestral subspecies (15–20).

How did a forest-dwelling, generalist mosquito species evolve into such a successful, invasive human disease vector? Initial genetic surveys suggested the possibility of multiple “domestication” events on continental Africa, with mosquitoes moving into human habitats independently in different locations across the continent (15). More detailed genetic data clearly show, however, that human-biting populations descend from a single ancestral lineage, which is consistent with a single origin of invasive *Aaa* occurring either in Africa or after the species expanded to other continents (16, 20–23). Expansion of the Sahara

desert ~5000 years ago forced *Ae. aegypti* populations to adapt by breeding in human water storage containers during extended dry periods (8, 24–27). Indeed, systematic host-preference testing of African populations revealed human-preferring populations in Senegal (13), and subsequent genomic analysis dated the split of these human-specialist populations from forest-dwelling generalists to the expansion of the Sahara ~5000 years ago (28).

Globally invasive *Aaa* (human specialists) are genetically distinct from African human-specialist populations (hereafter proto-*Aaa*), however, suggesting that *Aaa* emerged from a second evolutionary shift involving a strong bottleneck and possibly additional canalization of a human-specialist lifestyle (5, 20, 21). Demographic model fitting to exome sequencing data suggests the split between a Senegalese proto-*Aaa* population and an *Aaa* population from Mexico occurred ~300 to 400 years ago (21), which is consistent with historical records suggesting that *Aa. aegypti* spread from Africa on ships involved in the Atlantic slave trade [AST (26, 29)], but it remains unclear whether invasive *Aaa* emerged in Africa or after migrating to the Americas. Recently, *Ae. aegypti* populations from Argentina were shown to carry *Aaf*-like ancestry (20, 30, 31), raising the possibility that proto-*Aaa* populations made the transatlantic migration, and thus invasive *Aaa* emerged in the Americas. Genomic analysis of populations sampled across the globe will help understand the origin of invasive *Aaa* and contemporary movement of genetic elements relevant to public health.

Aaeg1200 genome sequencing project

Population genetics studies on *Ae. aegypti* have progressed with available methodologies, beginning with allozymes in the 1970s and advancing to determination of nucleotide variation from DNA chips and then to the use of limited sequencing (15–17, 20, 21). Until now, few whole-genome sequences, which provide the most detailed information, have been produced. We sequenced *Ae. aegypti* genomes from around the globe to further disentangle the complex historical shifts and genetic changes underlying the explosive global invasion of this important vector species. Following a similar effort in *Anopheles gambiae* (32), we named this endeavor the “Aaeg1200 genomes project.” Scientific resources for studying *Ae. aegypti* have lagged behind *Anopheles* malaria vectors, but the Aaeg1200 genome project has benefited from and builds upon the recent improved genome reference sequence available (33). These resources provide a platform to continue studying this important mosquito disease vector. We anticipate that these data, combined with valuable field and laboratory research, will advance understanding of the biology of this devastating disease vector and contribute to the fight against *Ae. aegypti* and the viruses it transmits to millions of humans around the globe.

We generated 31.4 terabases of DNA sequence with the Illumina HiSeq platform from 1304 specimens representing 73 populations from throughout the global distribution of *Ae. aegypti* (Fig. 1A and table S1). Samples were collected at all life stages and were 50.4% males. Special emphasis was given to sampling *Aaf* inside the native sub-Saharan African range of this species, with 510 samples representing 32 populations of this subspecies. A subset of our full dataset, including many of the *Aaf* populations, has been described previously (13). Short-read data were aligned to the updated versions of the AaegL5 (33)

genome assembly, deriving from the Liverpool Colony originally collected from West Africa (34) but genetically close to Asian *Aaa* populations (35), resulting in an average read depth at accessible, robust sites of 11.24× across the genome (fig. S1). After quality control and removal of closely related individuals, the panel included 1206 individuals. We identified 141.42 million high-confidence single-nucleotide polymorphisms (SNPs) across the panel.

Global patterns and differentiation

To understand how these populations are related to each other, we estimated genetic ancestry and genetic exchange among populations. Consistent with previous findings (13, 15, 20), the largest signal of genetic differentiation is between *Aaf* and *Aaa*, with further large subdivisions aligning with geographical barriers and distance (Fig. 1 and figs. S2 and S3). *Aaf* populations fall into two clusters on either side of the East African Rift Valley, and *Aaa* populations cluster by region and continent. The ΔK method (36) (where K is the number of clusters) supports a model with five genetic clusters (max $\Delta K = 6.27$), but model likelihoods increase continuously as more genetic clusters are added. The most complex model we evaluated (14 genetic clusters) was the best fit to the data, which is consistent with differentiation among regional and local populations in the data. Several populations emerge as genetically intermediate between the large genetic clusters (Fig. 1, B to D), including Ngoye (Senegal), Dakar/Thies (Senegal), Luanda (Angola), Rabai (2009, Kenya), and El Dorado (Argentina), which could be explained by a serial founder scenario describing the origins of proto-*Aaa* and *Aaa* or a model with a simple split between *Aaa* and *Aaf* followed by admixture among the subspecies. Admixture models with eight or more genetic components assign individuals from Argentina almost entirely to one genetic component (the proto-*Aaa* component), with Ngoye, Dakar, Luanda, and others from West Africa sharing the proto-*Aaa* component with additional ancestry from the West African *Aaf* component as well as some non-African *Aaa* ancestry (Fig. 1B). Consistent with previous descriptions (5, 9, 16, 37), Rabai (Kenya) is identified as an admixture of East African *Aaf* and an Asian *Aaa* cluster (Fig. 1, B to D).

Genetic diversity varies globally

It is known that *Ae. aegypti* has expanded in distribution and experienced a history of demographic shifts, but the effects of this history on genomic diversity are not well understood. Consistent with *Aaf* being the ancestral subspecies, genetic diversity is concentrated within *Aaf*, with the majority of variants (53.3%) segregating exclusively in *Aaf*, 14.7% found only in *Aaa*, and close to one-third (32.0%) shared between the two subspecies (Fig. 2B). Mean genetic diversity (π) within each population varies by nearly threefold among populations across the globe, from 1.11% in Cali, Colombia, to 2.87% in Skukuza, South Africa (Fig. 2). Nucleotide diversity varies along each chromosome with diversity significantly lower (Wilcoxon rank sum test; $P < 2.2 \times 10^{-16}$) around the centromeres and increasing with distance from centromere until it reaches a plateau at the center of each chromosomal arm (Fig. 2 and figs. S4 to S6), a pattern characteristic of the interaction between natural selection and recombination along chromosomal arms (38). Levels of nucleotide diversity are lower in highly recombining chromosomal arm regions in *Aaa* populations, making the distinction with the centromeric regions less pronounced.

Large demographic shifts can distort frequencies of genetic variation in a population and lead to increased mutational load that can have deleterious effects in extreme cases. We summarized derived SNP frequencies as the site frequency spectrum (SFS) in each population and revealed strong deficits of rare SNPs in many *Aaa* populations compared with *Aaf* populations and a hypothetical population with a constant population size (fig. S7). In some extreme cases, such as Jaffna (Sri Lanka), there are fewer rare variants than intermediate variants. To determine whether distorted frequency spectra could be explained by higher concentrations of chromosomal segments that are identical by descent (IBD), we identified IBD tracts within individuals and found that the mean length of IBD tracts is 43% longer in *Aaa* populations compared with *Aaf* populations (51.4 kb in *Aaf* versus 73.6 kb in *Aaa*). *Aaa* populations carry a higher number of IBD tracts that cover more of the genome compared with *Aaf* (fig. S8). We also measured correlations among SNP frequencies as a function of distance [i.e., linkage disequilibrium (LD)] and found evidence of disequilibrium at more than two times greater distances in *Aaa* populations [mean LD decay distance in *Aaa* = 14,571 base pairs (bp)] than in *Aaf* populations (mean LD decay distance in *Aaf* = 5669 bp; fig. S9). These results show that the population histories of *Aaa* populations have impacted patterns of inheritance and diversity, potentially exposing recessive deleterious alleles in IBD tracts.

Origin of *Aedes aegypti* invasive form

Genetic studies suggest that all invasive populations of *Aaa* are monophyletic with a single origin (5, 13, 15, 20, 21), but when and where invasive *Aaa* first evolved remains a key unanswered question. Tracing the origin of *Aaa* is complicated because *Aaa* was partially or completely eradicated from large regions of the Americas, the Caribbean, and Europe (39)—the most likely entry points when this species emigrated from Africa. Thus, *Ae. aegypti* populations currently found in these regions result from a massive reinvasion after eradication efforts ceased in the 1960s (39). With the exception of some *Aaa* populations from the United States, certain Caribbean islands, and possibly Argentina that were not eradicated, these groups are unlikely to represent direct descendants of pre-eradication populations (39, 40).

To understand where and when invasive *Aaa* emerged, we first need to identify their closest ancestral population(s). Outside Africa, most *Aaa* populations trace back to a single ancestral lineage except for Argentinian populations (20, 30, 31). Our data also point to Argentinian populations of *Ae. aegypti* as outliers from other populations outside of Africa with higher genetic diversity and proto-*Aaa* ancestry, suggesting that the origin of *Aaa* could be described by an *Aaf*-to-proto-*Aaa*-to-*Aaa* stepping stone model with American proto-*Aaa* as a middle step. However, populations of proto-*Aaa* outside of Africa could represent either a relict population that has survived since an introduction during the AST or a more recent introduction from Africa (20, *26*). We conducted cross-coalescent analyses to estimate the split time between El Dorado (Argentinian proto-*Aaa*) and Ngoye (Senegalese proto-*Aaa*) and found that the timing of the split (Fig. 3A) is most consistent with Argentinian populations representing relict populations of proto-*Aaa* introduced around the time of the AST. Angola was a frequent last port for embarkations during the AST (29), so we also compared El Dorado with Luanda (Angola), another African population with

proto-*Aaa* ancestry (22, 26, 41), and found very similar patterns (fig. S10), confirming that the split timing between American and African proto-*Aaa* populations is not sensitive to the choice of African proto-*Aaa* population.

The similarity among cumulative migration curves points to close ties among the three proto-*Aaa* populations in our dataset before and during the AST. To test whether one of the proto-*Aaa* populations is more closely related to invasive *Aaa* populations than others, but using an alternative method, we calculated phylogenetic trees in 10-kb windows along the genome using a set of representative *Aaf*, proto-*Aaa*, and *Aaa* populations and found large proportions of the trees with support for each of the three proto-*Aaa* populations. Of the trees where *Aaa* populations were monophyletic, 0.41 [95% confidence interval (CI): 0.40 to 0.42] trees include Luanda closest to *Aaa*, 0.32 (95% CI: 0.31 to 0.33) trees include El Dorado closest, and 0.27 (95% CI: 0.26 to 0.28) trees place Ngoye closest. Cross-coalescent analysis and subtree weightings agree that El Dorado is not a recent introduction but instead appears to be a relict of a highly connected metapopulation of proto-*Aaa* that was likely spread around the Atlantic basin around the time of the AST.

If proto-*Aaa* populations were introduced during the AST and invasive *Aaa* evolved from proto-*Aaa* in the Americas or the Caribbean, then we would expect the split between proto-*Aaa* and invasive *Aaa* populations to come after the split among proto-*Aaa* populations. We estimated the timing of the split between El Dorado proto-*Aaa* and several representative *Aaa* populations using Multiple Sequentially Markovian Coalescent–Isolation-Migration (MSMC-IM) and discovered support for multiple histories (Fig. 2B). The *Aaa* population from Cebu City (Philippines) showed a sharp increase in cumulative migration from ~0.1 to 1, going backward in time around the period of the AST, implying complete population merging with El Dorado at that time. Passo Fundo, Brazil, by contrast, shared migrants with El Dorado more recently than did Cebu City, but approximately one-fourth of the Passo Fundo genome coalesced further back in time on the order of thousands of years, not hundreds. The remaining *Aaa* populations showed similar step functions in their cumulative migration curves, in which a substantial proportion (~0.4 to 0.6) of the genome coalesced with El Dorado around the time of the AST, but the remaining genomic haplotypes remained isolated until around the same time that Passo Fundo showed an ancient increase in migration. We repeated the analyses with bootstrap-resampled genomes and found that some populations show similar curves for both the empirical and bootstrap replicates (e.g., Arizona, Miami, and Pazar; fig. S11), but many *Aaa* populations showed both the step function shape as well as recent population merging with El Dorado, similar to what was observed for Cebu City, for which the curve reaches one around the time of the AST (fig. S11).

We also calculated cumulative migration curves, comparing Arizona and Trinidad with Luanda, another proto-*Aaa* population, and recovered both shapes in the bootstrap replicate curves (fig. S12). These results suggest that genomes from *Aaa* and proto-*Aaa* populations comprise a mixture of haplotypes with different histories: haplotypes that coalesce between *Aaa* and proto-*Aaa* around the time of the AST and haplotypes that coalesce further back in time, around the time when proto-*Aaa* split from *Aaf* ~5000 years ago (28). The presence of haplotypes in proto-*Aaa* and *Aaa* genomes with such distinct histories could be explained by

either a history of substructure within African proto-*Aaa* with variable levels of admixture exchange with *Aaf* or introduction of *Aaf* haplotypes through more recent admixture with proto-*Aaa*. The similarity between cumulative migration between Arizona and both El Dorado and Luanda suggests that recent admixture among proto-*Aaa* and *Aaf* cannot explain the entire signal because El Dorado is geographically isolated from *Aaf.*

Using the changes in effective population size and migration suggested by the MSMC-IM results, we defined two competing demographic models to explicitly test, using an independent method, whether the data support a stepping stone model with proto-*Aaa* making the transatlantic trip or a model with both *Aaa* and proto-*Aaa* making transatlantic trips. We fitted the four-population models to two-dimensional (2D) SFS data (fig. S13) from representative populations [*Aaf*: PKT; proto-*Aaa*: NGO; proto-*Aaa*: ELD; *Aaa* (invasive): ARI] using fastsimcoal2 (42) and found that the one-emigration stepping stone model resulted in a better fit to the data with a substantially lower Akaike information criterion value (Fig. 3; ΔAIC = 20,187.86). The best-fitting model included a split between Argentinian proto-*Aaa* and North American *Aaa* ~220 (95% CI: 173 to 328; assuming 15 generations per year) years ago and a split between Argentinian proto-*Aaa* and Senegalese proto-*Aaa* ~100 years earlier (95% CI: 275 to 438; fig. S14). Our best-fit model includes an estimate of ~2300 (95% CI: 2148 to 2576) years ago for the split between *Aaf* and proto-*Aaa*, substantially more recent than a previous study based on a similar approach that reached an estimate of ~11,000 years ago (21). An MSMC-IM–based study allowing greater flexibility for variation along the genome timed this split to be ~5000 years ago (28). Together with the historical absence of *Aaa* populations within Africa until recently (37, 43), the most parsimonious scenario is that proto-*Aaa* emerged in West Africa with varying degrees of isolation from *Aaf*, slave trade vessels transported populations of proto-*Aaa* to the Americas during the AST, and *Aaa* evolved from proto-*Aaa* in the Americas and subsequently spread around the globe.

Global invasion of *Aaa*

Historical records and genetic data suggest that Asia, the Pacific region, and the Middle East were colonized by *Aaa* in the 1800s, well after *Aaa* establishment in the Americas (20, 44), but the source of these populations is not well understood. Our genetic structure analysis, principal components analysis (PCA), and genetic distance tree all show that populations from the Middle East, Pacific, and Asia have genetic affinities with populations from the Caribbean (Fig. 1), which is potentially consistent with movement from the Caribbean to these regions as the source of the founding populations. Alternatively, recent Asian migration into the Caribbean could explain this pattern. Cross-coalescent analysis of migration, however, comparing the population from Bangkok with representative *Aaa* American populations, shows that Bangkok split from all the representative populations around the same time. However, the population from Puerto Rico exchanged migrants with Bangkok more recently than did the other populations, likely explaining the higher level of shared ancestry with populations from the Caribbean (Fig. 2 and fig. S15). This hypothesis is consistent with the presence of *Ae. aegypti* in the Mediterranean Basin in the 1800s (41, 45), near ports involved in trade with the Americas that frequently stopped at Caribbean

islands (especially Santo Domingo) before returning to Europe, providing a potential route for *Aaa* introduction to Asia through the Suez Canal.

Invasive *Aaa* adaptation

Ae. aegypti has undergone a series of phenotypic shifts in recent evolutionary history, beginning with the shift from forest-dwelling host generalist phenotypes to preferring human hosts and living around human settlements, and later further adapting to modern human landscapes and interventions such as insecticides (46–48). Rose *et al.* (13) identified several genomic regions that are highly differentiated between generalist and human-preferring populations in West Africa. To discover genomic regions targeted by natural selection during the more recent invasion of the Americas and the Caribbean, we scanned the genome for regions that are highly differentiated in North American populations (Miami, New Orleans, and Mexico) when compared with a proto-*Aaa* population (El Dorado), after polarizing with a West African *Aaf* population (PK10) to isolate signals specific to North American populations. North American populations were chosen because they likely represent the closest proxy to early New World populations introduced during the AST, because most of the original founding South American populations were eradicated in the 1950s (39).

We screened for signatures of natural selection using the population branch statistic (49) (PBS) and found putative evidence for adaptation to novel pathogens and feeding environments (fig. S16), which is consistent with niche expansion after the emergence of invasive *Aaa*. The strongest window of differentiation shared among several North American populations (PBS = 0.5713) was located on the q arm of chromosome 2 and centered at 401.75 Mb (fig. S17). The window of increased differentiation is ~2 Mb wide and centered over a region with relatively few annotated genes (fig. S17). The gene *AAEL025736*, at the center of the peak, comprises 18 exons spanning ~477 kb and is annotated as a Toll-like receptor Tollo, a transmembrane receptor involved in antimicrobial and antiviral immune responses (50). The second-strongest window of differentiation is located on chromosome 1 centered at 79.75 Mb (fig. S18). The width of the peak is somewhat smaller compared with the first peak, but it covers a more gene-dense area (fig. S18). The gene at the center of the peak, *AAEL017005*, is the RYamide receptor that pairs with RYamide neuropeptides and is thought to play a role in regulating feeding and digestion as well as male reproduction (51).

Contemporary admixture between *Aaa* and *Aaf*

After historical shipping and trade helped spread *Aaa* throughout the tropical world, modern shipping, trade, and travel have increased global connections with potentially important implications for mosquito populations and viral transmission risk. Our admixture analysis suggests that *Aaa* populations from Southern Brazil are admixed with proto-*Aaa* ancestry, likely from nearby Argentina (Fig. 1B). After the collapse of eradication efforts in South America in the 1960s, *Aaa* populations likely reinvaded from the North (52) and reestablished throughout eradicated areas, potentially making secondary contact with relict proto-*Aaa* populations upon expansion. We analyzed admixture tracts along genomes from Passo Fundo from Southern Brazil to estimate the timing of admixture between *Aaa* and proto-*Aaa* from El Dorado (fig. S19). We modeled Passo Fundo as a mixture of El Dorado

and the closest *Aaa* population, Patos de Minas, and estimated the time of contact to be ~18 years ago (assuming 15 generations per year, 95% CI: 14 to 22). Considering that our Passo Fundo samples were collected in 2017, the admixture we detected occurred in the late 1990s. It is possible that during this time period, *Aaa* completed reinvasion from the north into Southern Brazil. Alternatively, the establishment of Mercosur (Mercado Común del Sur), an economic and political bloc consisting of Argentina, Brazil, Paraguay, Uruguay, and Bolivia, increased trade in this area since 1991, resulting in increased migration and secondary contact between proto-*Aaa* and *Aaa* (31).

We also searched for contemporary admixture between *Aaa* and *Aaf* populations more broadly, using Patterson's F_3. Admixture signals were identified predominantly in coastal populations on both coasts of Africa, with particularly strong signals detected in coastal Senegal, Angola, and Kenya, whereas Ouagadougou, Burkina Faso, was the only inland population with clear evidence of admixture (Fig. 4). These results were robust to selection of alternative *Aaa*/*Aaf* ancestral populations (fig. S20). We applied Patterson's F4 to pairs of *Aaa*/*Aaf* and pairs of proto-*Aaa*/*Aaf* to rule out the possibility that the admixture signals derived from admixture between proto-*Aaa* African populations and unadmixed *Aaf* African populations. In all cases, the magnitude of F4 was greatest within *Aaa*/*Aaf* admixture pairs (fig. S21) and could not derive from admixture within sub-Saharan Africa. To understand the timing of admixture, we fit a two-pulse admixture model to identify admixture tracts in samples from Rabai, Kenya, and found evidence for first contact to be ~52 years (assuming 15 generations per year, 95% CI 39 to 71 years) before sample collection and a secondary wave of admixture ~10 years before sample collection (95% CI 7 to 18 years). The first report of *Aaa* along the coast of East Africa was in 1952, which is consistent with the presence of an imported domesticated *Aaa* population (23). Together, our admixture results indicate that contemporary secondary contact between global *Aaa* and African *Aaf* is extensive in African coastal regions. Dengue outbreaks have become more common in Africa and are geographically correlated with the signatures of *Aaa*/*Aaf* admixture (Fig. 4). Although most of sub-Saharan Africa does not report extensive dengue outbreaks, four (Angola, Burkina Faso, Kenya, and Senegal) out of seven locations that show signals of *Aaf/Aaa* admixture have reported regular or major dengue outbreaks in recent years (53).

Knockdown resistance mutation sharing

Pyrethroid chemical insecticides are a critical component of modern mosquito control, but the spread of insecticide-resistance mutations is threatening the utility of this tool globally. We find multiple SNPs that have demonstrated roles in resistance pyrethroids and DDT (54, 55) at high frequency in our dataset (table S2). Knockdown resistance (KDR) SNPs are common in the Americas, with high F1534C (F1534→C; F, phenylalanine; C, cysteine) prevalence across all regions (USA 76%; Mexico 97%; Brazil 64%), and >90% frequency across the Caribbean (table S2). Two mutations at V1016 (V, valine) are mutually exclusive, with V1016I (I, isoleucine) predominating in the Americas, often in linkage with F1534C and V1016G (G, glycine) at high frequency across Asia, which is consistent with earlier reports (56).

To infer the origins and spread of KDR-related pyrethroid resistance in *Ae. aegypti*, we derived a haplotype network from all (n = 15) non-synonymous SNPs across the locus (Fig. 4 and fig. S22). Whereas the susceptible (wild-type H001) haplotype represents 31% of the dataset, just eight mutated haplotypes (H002 to H009) make up the next 58% of the total, each of which contains at least one previously characterized resistance-associated SNP (fig. S23). In all regions outside sub-Saharan Africa, susceptible haplotypes are in the minority; the average North American population exhibits 93% resistant haplotypes (72 to 100%), and the mean frequency of resistant haplotypes is similarly high across Asia, albeit across a much wider range (mean 76%; range 8 to 100%) (table S2). The haplotype network shows at least five independent origins of resistance haplotypes with further branching and expansions, with one branch (starting with H003) including most haplotypes from Asia and the Americas. One haplotype node common in South America and another common in Asia are both preceded in the network with a haplotype node with no known resistance mutations, but the second node includes at least one resistance mutation, raising the possibility that the first nodes may carry mutations with no currently known role in resistance. Stark haplotype frequency differences between Asia and the Americas and the topology of the network support independent origins of pyrethroid resistance in each continent, in agreement with previous reports (56).

Within Africa, many of the populations with strong signals of *Aaa*/*Aaf* introgression also exhibit a high prevalence of KDR SNPs (Fig. 4); F1534C and V1016I are both at high frequency in Luanda, Angola (F1534C: 94%; V1061I: 44%), and Ouagadougou, Burkina Faso (65%; 33%). These mutations are rare within Africa but more common in the Americas, which is consistent with admixture-aided introduction and spread of the alleles. Haplotype dendrograms confirm clear resistance-haplotype sharing among West African populations and non-African populations (figs. S24 and S25). Multiple haplotype clades include large polytomies spanning continents underscoring the rapid spread of haplotypes carrying multiple known resistance mutations. However, this pattern is not universal, with admixed coastal populations in Senegal showing a complete absence of the three major KDR alleles, despite strong evidence of admixture.

The prevalence of resistance haplotypes within sub-Saharan Africa is consistently lower than in the rest of the world, with mean prevalence of resistant haplotypes in West and East Africa being 23 and 29%, respectively, and many populations carrying an entirely susceptible KDR profile. Resistant haplotype H009, which contains only the I915K (K, lysine) SNP, is found only in East Africa and may represent a reversion of the locus.

Conclusions

The evolution of preference for human hosts and adaptations to human habitats set the stage for the explosive expansion of *Aaa* and increased risk of arboviral diseases around the globe. Building on prior knowledge, our results help define four historical eras in the evolution of invasive *Aaa* (Fig. 5). Similarities among cumulative migration curves and subtree weights suggest that the first era, the period leading up to and during the AST, was characterized by partial separation between proto-*Aaa* and *Aaf*, followed by large-scale movement and shuffling of proto-*Aaa* populations around the Atlantic coasts of Africa and the Americas,

introducing populations to an open, human-centered niche in the Americas. Cross-coalescent analyses and 2D SFS-based model fitting suggest that ~100 years after proto-*Aaa* emigrated from Africa to the Americas and the Caribbean during the AST, canonical (invasive) *Aaa* emerged in the Americas. Scans for signals of differentiation specific to invasive *Aaa* point to adaptation to new immune challenges and feeding habits that may have enabled invasive *Aaa* to expand into new environments, leading to rapid expansion throughout the continent and eventually the rest of the tropical world in the second era. The third era involved large-scale elimination of *Ae. aegypti* from many regions of the Americas, the Caribbean, and Europe (39).

Our analysis indicates that the global spread of *Aaa* continues today, in the fourth era, with important implications for control of *Aa. aegypti* and dengue transmission, especially in Africa. Back-to-Africa migration of *Aaa* was first recognized in Tanzania in the 1950s (23), and our data suggest that such movement of the domestic, human-preferring *Aaa* into Africa is more widespread and common than previously understood. We also show insecticide-resistant haplotype sharing among populations in the Americas and populations in Africa, pointing to the importation of genetic variants that could hamper mosquito control efforts. To assist in control efforts, diagnostic assays could be developed from the SNPs discovered in the Aaeg1200 dataset for monitoring both *Aaa* ancestry and insecticide resistance in urban African populations. Our results underscore how the combined effects of increased urbanization, global connectivity, and global warming enable the emergence and spread of invasive species, and in this case the highly domesticated, invasive *Aaa* may already be driving increased dengue transmission (57, 58).

Materials and methods

Sample collection

Mosquito samples were provided by a number of contributors either as carcasses or DNA extractions. Samples were collected from 2009 to 2019. See table S1 for sample sizes, location, and other collection information.

Whole-genome resequencing and processing

We sequenced whole genomes using the Illumina HiSeq X and 4000 platforms. When adult or larval carcasses were available, we extracted DNA using the protocol described previously in Crawford *et al.* (59). Briefly, individual carcasses were homogenized in lysis buffer using steel grinding balls on a SPEX grinder. Lysate was transferred to a new plate as input for DNA extraction on Chemagic360 according to the manufacturer's protocol. DNA sequencing library preparation was performed using a PCR-free protocol also described in detail in Crawford *et al.* (59). Briefly, at least 150 ng of gDNA from each sample were fragmented to target size between 350 and 400 bp, double-side size selected to narrow the size range to 350 to 400 bp, end repaired, A-tailed, and adapter-ligated. Purified libraries were pooled and sequenced on the Illumina HiSeq 4000 and HiSeqX platforms to generate 2×151 bp reads.

DNA sequence reads for samples published in Rose *et al.* (13) have been deposited to NCBI SRA under accession number PRJNA602495. The remaining samples have been deposited under accession number PRJNA1185803.

Short-read mapping and genome reference updating

We mapped Illumina paired-end reads to the *Ae. aegypti* L5 genome reference [NCBI WGS Project NIGP01; (33)] using BWA *mem* (v0.7.16) (60) with default parameters, then sorted using samtools *sort* (1.1) (61), and duplicate reads were marked using Picard's *markdup* function (2.1.0, http://broadinstitute.github.io/picard/).

We calculated mapping quality metrics using the Picard (62) CollectAlignmentSummaryMetrics tool and observed a trend where samples with higher mismatch rates between reads and the L5 reference sequence mapped at lower rates, especially samples from Africa (fig. S26). We hypothesized that the relationship between mismatch rates and read mapping rates could bias variant recovery against high diversity populations (i.e., *Aaf* populations), potentially impacting downstream analyses. To minimize the effects of SNP variants and genetic distance on short read mapping, we conducted an iterative process to update the reference sequences. This routine was described previously in Rose *et al.* (13), and the resulting reference sequence was used for mapping *Aaf* populations for this study as well. We analyzed various aspects of read mapping for each iteration and found that the number of reads mapping, the number of sites covered, and the number of heterozygous sites discovered all increased, while mismatches between reads and the reference decreased, as shown in fig. S5 in Rose *et al.* (13). We made an analogous updated reference specifically for *Aaa* populations as well using the same method. Briefly, a random selection of 100 *Aaa* males were aligned to the L5 *Ae. aegypti* reference, and then the reference was updated with the consensus allele from the pileup. This process was repeated three times and then used this updated reference for remapping all *Aaa* samples.

After re-mapping all short reads to updated reference sequences, we conducted indel realignment and read clipping. We used GATK (v3.8-1-0-gf15c1c3ef) (63) for indel realignment on each population individually with default settings. After indel realignment, we clipped overlapping read pairs using the clipOverlap function of the bam program within the bamUtil analysis kit (v. 1.0.14) (64). We calculated read depths with BAM files after read-clipping at population-specific robust sites using ANGSD (-minMapQ 20 -minQ 10 -remove_bads 1 -uniqueOnly 0 -doCounts 1 -doDepth 1 -maxDepth 1000).

Outgroup sequences

We used publicly available sequence data for *Aedes bromeliae* to assemble an outgroup sequence. We downloaded fastq files (accession numbers SRX2323182, SRX2323157, and SRX2323117) from NCBI SRA for an *Ae. bromeliae* individual. These short reads were aligned to the updated *Aaf* reference sequence using the same pipeline described above including read clipping, but no indel realignment. We created a fasta sequence from the short data using ANGSD (-minMapQ 20 -minQ 10 -only_ proper_pairs 1 -remove_bads 1 -uniqueOnly 0 -doFasta 3 -doCounts 1 -setMaxDepth 150 -setMinDepth 8).

We sequenced four *Aedes mascarensis* individuals from a laboratory colony using the same protocols as for the full *Ae. aegypti* panel.

Close relative screen

To identify close relatives within each population sample, we calculated relatedness among all individuals using NgsRelate (65). GLF input files were prepared using ANGSD (66) at 60 million randomly selected sites from the robust sites set described below and GATK genotype likelihood model (-GL 2), as recommended by the authors of NgsRelate. We further limited the analysis to sites with read depth greater than seven reads and less than 46 reads, and applied the SNP-pval flag with a value of 1×10^{-6}. We ran NgsRelate with a minimum allele frequency of 0.05 and plotted the KING-robust and R1 statistics to evaluate relatedness based on theoretical thresholds for each level of relatedness (67).

To confirm the KING-robust and R1 statistics behaved as expected with our samples, we calculated the physical distance between samples, all *Aaf*, collected for Rose *et al.* (13) based on the location of ovicups used for collections. We binned sample pairs into four categories: (i) cupmates collected from the same ovicup, (ii) pairs of samples from different cups but <50 m apart, (iii) pairs of samples from cups separated by 50 to 100 m, and (iv) pairs of samples from cups >100 m apart. We plotted KING-robust on R1 for each category separately (fig. S27). As expected, close relatives (KING-robust > 0.25 and R1 > 0.5) were predominantly found in the cupmate and <50-m categories with only two exceptions in each of the other categories.

We then plotted the same statistics for all populations (fig. S28) with *Aaa* and *Aaf* populations separated by color. In contrast to *Aaf* populations where most pairs had KING-robust and R1 values consistent with low relatedness (lower left quadrant), *Aaa* populations followed a different pattern with elevated R1 values overall and a near continuous distribution extending into the high-relatedness quadrant (fig. S28). Although the relatedness statistics followed the expected distribution in *Aaf* populations, the statistics appeared elevated in *Aaa* populations. The authors of NgsRelate conducted simulations to show that these statistics can be robust to non-equilibrium demographics, but we suspect that the demographic histories of the *Aaa* populations likely deviate substantially from the relatedness models underlying KING-robust and R1. For example, the standard thresholds would suggest that over half of the samples in the *Aaa* population from Clovis, USA are siblings. However, collection information would make that scenario extremely unlikely. Given that many *Aaa* populations have extreme demographic shifts in their recent history, we could not identify a sensible *Aaa*-specific set of thresholds that would be appropriate while ensuring that we were not inadvertently biasing patterns of genetic diversity in these populations. As such, we used the standard thresholds (KING-robust < 0.25 and R1 < 0.5) to identify unrelated pairs and removed one individual of every pair that fell outside those bounds for all *Aaf* populations but not *Aaa* populations. Relatedness filtering resulted in the removal of 98 individuals from *Aaf* populations leaving 412 *Aaf* and 794 *Aaa* individuals used in all downstream analyses. Sample IDs for unrelated males and females are available in a Zenodo record (68).

Robust site identification

For each analysis, we identified genomic locations considered robust for analysis by applying a series of filters and only keeping sites that passed all filters. First, we generated a pileup VCF annotated with a series of quality statistics and tests using samtools mpileup (1.6-2-gf068ac2) and the following flags: '-q 10 -Q 20 -I -u -t SP,DP'. Putative variant sites were called from this VCF using BCFtools (1.6, https://samtools.github.io/bcftools/) with '-f GQ -c' flags, which were then filtered using SNPcleaner.pl from ngsQC (https://github.com/tplinderoth/ngsQC) with the following settings '-k MIN_IND -u 2 -D MAXD -a 0 -f 1e-4 -H 1e-6 -L 1e-6 -b 1e-10'. For population samples with more than 12 individuals, MIN_IND, or the minimum number of individuals with data at a given site, was set to 12. For population samples with less than 12 individuals, MIN_IND was set to the number of individuals in the sample. The total maximum depth across all individuals in a population sample, MAXD, was set to 45 X MIN_IND. In all cases, the autosomal genome was divided into 10-Mb chunks and analyzed in parallel using GNU parallel (69). This pipeline produced a set of sites, including both putative variant and invariant sites, for each population separately that have passed all of the filters that could be considered robust for analysis of that population. To obtain a global set of robust sites for analysis of the full panel, we found the intersection of robust sites sets across all *Aaa* populations and *Aaf* populations separately.

To further filter sites that are potentially difficult to map in the genome, we calculated the total depth of reads with map quality >= 20 for all African and non-African samples on their respective updated references at sites identified in the robust set above. Supplementary fig. S29 shows density curves for each set up to 20,000 reads. For individuals mapped to the African reference, 1,180,248,427 (98.76%) sites had at least one read. For the non-African reference, 1,169,822,108 (97.89%) sites were covered by at least one read. While both sets show clear local maxima at 6410 and 9701 for African and non-African, respectively, they differ in the density of lower-depth sites, with a sharp increase in lower-depth sites in the African set. Regional genomic biases in accessibility impacting read mapping rates likely contribute substantially to variation in read depth across the genome. We excluded sites in the African set with a total read depth greater than 50% above the 6410 maximum, or 9615 reads. Using the same logic, we excluded sites with a total read depth greater than 14,552, for the non-African set. High read-depth sites were excluded from the *Aaa* and *Aaf* robust sites sets from above and then the filtered sets were merged to find the intersection to arrive at a final set of 1,138,636,693 sites for analysis.

Variant calling

We called SNPs using the likelihood ratio test in ANGSD and a *P*-value threshold of 1×10^{-6} in the *Aaf* ($n = 425$) and *Aaa* ($n = 778$) sets separately. We identified 120.68 million variant sites in *Aaf* and 66.02 million variant sites in *Aaa*. We filtered all variants to include only sites with data present for at least 80% of individuals and the minor allele was found on at least 10 chromosomes. We then found the union of the filtered *Aaf* and *Aaa* variant sets for each chromosome resulting in a total of 141.42 million unique variants.

In addition to the full set, we generated three linkage disequilibrium-pruned variant sets to facilitate genome-level analysis of independent loci. First, we used PLINK (70) to calculate

linkage disequilibrium (LD) in 100 variant windows with a step size of and an LD cutoff of 0.1 resulting in a set of 7,161,365 from the Aaa variant set. From this LD-pruned set, we randomly selected 1 million sites (1×10^6 set). From the 1×10^6 set, we randomly selected a smaller set of 100,000 sites. For each LD-pruned SNP set, variable sites were distributed across all three chromosomes such that each chromosome was represented in proportion to its length. For analysis, we filtered the gVCFs to generate a set of variant VCFs with each of these LD-pruned variant sets.

Haplotype phasing

We phased genomes as described in Rose *et al.* (28), first pre-phasing variants spanned by single sequencing fragments using HAPCUT2 (71), then statistically phasing chromosomes using SHAPEIT4 (version 4.2) (72). In order to manage memory usage, we phased different global regions (North America, South America, African samples not previously analyzed in Rose *et al.* (28), and Asia) in separate SHAPEIT4 runs. We then used bcftools merge to combine phased regional genome sets. For each SHAPEIT4 analysis, we included the Rose *et al.* (28) set of phased genomes as a reference panel.

Genetic differentiation and structure

We calculated genetic differentiation using multiple approaches. First, we used ANGSD to prepare a beagle-format genotype likelihood file at the unlinked 1×10^6 SNP set described above and used NGSadmix (73) to estimate admixture proportions assuming (K) 2 to 14 population components. We applied a minimum minor allele frequency threshold of 0.05, and missingness tolerance of 0.05, and a minimum number of individuals with a data filter of 1100 resulting in a filtered set of 292,710 SNPs for analysis. To compare among models with different numbers of population components, we ran each K level 20 times with different starting seed values and submitted output of all replicates to the CLUMPAK server for analysis using the best K method (https://clumpak.tau.ac.il/bestK.html). The output of both the Evanno (36) best-K and model probability analyses are shown in fig. S30 and S31, respectively.

Second, we used the same unlinked 1×10^6 SNP set for PCA. We calculated the covariance matrix using PCangsd (version 0.982, https://github.com/Rosemeis/pcangsd) (74) and -minMaf threshold of 0.01. Eigenvectors were calculated using the eigen function in R (75). For Fig. 1, we found the largest admixture component from the admixture analysis with $K = 8$ and set the color of the individual in the PCA plot to match the color in the admixture plot. We also plotted subsets of the full panel by first subsetting the covariance matrix and then recalculating the eigenvectors. Analysis of only non-African *Aaa* populations is shown in fig. S32 and proto-*Aaa* and *Aaf* populations are analyzed and shown in fig. S33. To reveal additional substructure, we plotted principal components 3 and 4 (fig. S34) and see *Aaa* distributed across PC3 and *Aaf* distribute across PC4. We also calculated covariance matrices and eigen vectors for each of the three chromosomes independently and see that they each recapitulate the high level genome-level result (fig. S35).

Third, we calculated genetic distance as D_{xy}. To calculate D_{xy} while avoiding any biases associated with hard SNP and genotype calling, we used realSFS (76) and ANGSD to

calculate 2D SFS pairwise between all individuals and used the entries of the 2D SFS to quantify the number of differences and the total number of sites. The 2D SFS was calculated at all robust sites after removing all sites inside and within 100 bp from boundaries of annotated exons. For computational efficiency and to avoid the sex-differentiated region on chromosome 1, we limited this analysis to 84 1-Mb windows on chromosomes 2 and 3 excluding centromeric regions as defined for the AaegL5 assembly. We used this same protocol for the two outgroups, *Ae. bromelia* and *Ae. mascarensis*, and included them directly in the overall distance matrix. The neighbor-joining tree was calculated and plotted using the *ape* package (77) in R (75).

Site-frequency spectra

To avoid any biases associated with hard SNP calling, we calculated unfolded SFS directly from sequencing read data using ANGSD and realSFS for each population separately. First, we calculated site allele frequency (.saf) files at global robust sites on all three chromosomes (unplaced scaffolds excluded) using *doSaf* in ANGSD with *Ae. bromeliae* specified as the ancestral sequence. Second, we used realSFS to estimate the genome-wide SFS for each population. For plotting, allele counts were converted to proportions to facilitate comparison among populations.

Nucleotide diversity

We calculated nucleotide diversity (π) by calculating folded site allele frequency files for each population. We then used realSFS to estimate folded site frequency spectra, calculate theta statistics for each population using the thetaStat program (78) distributed with ANGSD, and conducted sliding window analysis with 50-kb non-overlapping windows. We summarized π using smoothed-loess functions in R (75) and plotted along chromosomes 1 to 3 (figs. S4 and S5). We also calculated density curves for each population that were based on all 50-kb windows (fig. S36) and summarized the distributions by calculating the mean and standard deviations across all windows for each population and plotted on the global map (Fig. 2A) using the R package *ggmap* (79).

We tested for differences between centromeric regions and chromosomal arm regions by assigning 50-kb windows to centromeric or arm groups and comparing the distributions using the Wilcox rank sum (wilcox.test function) test in R. Centromeric regions were defined as Mb 145 to 175 for chromosome 1, Mb 205 to 235 for chromosome 2, and Mb 180 to 210 for chromosome 3. For the arm regions, Mb 50 to 75 was included for all three chromosomes as well as Mb 250 to 275 for chromosome 1, Mb 350 to 375 for chromosome 2, and Mb 300 to 325 for chromosome 3.

Linkage disequilibrium

We calculated r^2 using ngsLD (80). For efficiency, we extracted 1-Mb regions from each of the six chromosomal arms. On the p arms, the regions started at 110 Mb on all three chromosomes. On the q arms, the region started at 270 Mb on chromosome one, and 320 Mb on chromosomes 2 and 3. These regions were extracted from the full genome beagle and then each population was extracted to make population specific input files. ngsLD was run with the following flags–probs–max_kb_dist 50–min_maf 0.05–ignore_miss_data–

rnd_sample 0.1–extend_out to randomly select 10% of sites, apply a minimum allele frequency filter and limit comparisons to SNPs within 50 kb of each other. Decay curves were fitted using the fit_LDdecay.R script included with ngsLD and plotted (fig. S9).

IBD tracts

We identified IBD tracts within individuals using ngsF-HMM (81), a two-state Hidden Markov Model that uses a probabilistic framework based on genotype likelihoods. We called IBD tracts in four representative *Aaf* populations (PK10, La Lope, Bantata, and Arabuko), one proto-*Aaa* population (Ngoye), and 10 representative *Aaa* populations (Manaus, Jacobina, Passo Fundo, Bangkok, Jeddah, Pazar, Jaffna, Amacuzac, Clovis). For each population, we assembled Beagle-style genotype likelihood files including the 1×10^6 SNP sites and ran ngsF-HMM with the–lkl flag bug otherwise default settings. IBD tracts were extracted from individual .ibd output files, summarized and plotted using custom R scripts.

Cross-coalescent analysis

All cross-coalescent analyses were performed following the methodology outlined in Rose *et al.* (28). In summary, for the cross-coalescent analyses, we first utilized the bam files corresponding to each sample to generate a genome-specific mask for every chromosome and sample. This was achieved using the bamCaller.py script from the MSMC tools package (https://github.com/stschiff/msmc-tools). Subsequently, we obtained the final genotype for each sample by merging its phased genotype from the *Ae. aegypti* phased reference panel (as inferred in Rose *et al.* (13, 28) with its original genotype obtained using samtools 1.9 (mpileup -B -q20 -Q20 -C50). Next, we filtered the SNPs using both the specific mask and the general mask (13), retaining only the best SNPs that met the following criteria: not in outlier regions identified as under natural selection between human specialists and generalist in Rose *et al.* 2020, in callable sites identified in Rose *et al.* (13), and not in the sex locus or on repetitive and centromeric regions. Finally, we inferred the cross-coalescent events by conducting MSMC2 (82) analysis followed by MSMC-IM (83) analysis, to fit the isolation-with-migration model to the inferred cross-coalescent estimations. For each population, we used one male and one female representative. These analyses were performed with default parameters. Bootstrapped MSMC2 and MSMC-IM analyses were conducted using 10 replicates of three 400-Mb chromosomes constructed with resampled blocks of 20 Mb. For plotting, we used the same approach as Rose *et al.* (28) including a scaling factor corresponding to a mutation rate (μ) of 4.85×10^{-9} and generation time (g) of 0.067 years, and providing the same visual reference points of the AST and the African Humid Period. Effective population size and cumulative migration curves for bootstrap replicates are shown in fig. S37.

Ancestry tract analysis

We used AncestryHMM (84) to detect evidence of admixture over the genome of two different populations: Passo Fundo from Brazil and outdoor-Rabai collected in 2017 from Kenya.

For Passo Fundo, El Dorado (Argentina) population was selected as a representative of the proto-*Aaa* ancestor and Patos de Minas (Brazil) as a proxy of *Aaa* as the closest, non-admixed Brazilian population to Passo Fundo.

For Rabai 2017 analysis, African mosquitoes from different populations but with a clean Aaf ancestry (see NGSadmix results: three from Arabuko, four from Kwale, three from Shimba Hills, and two from Skukusa) were used as representatives of *Aaf*, while individuals from Jeddah were used as the proxy for the *Aaa* donor population.

AncestryHMM was applied following the detailed instructions on (28) (28) (28), with minor modifications: after filtering in the best sites as for cross-coalescent analysis, the input data was obtained by considering a recombination rate of 1.7×10^{-9} M/b. The admixture timing was estimated for each comparison. An analysis of 100 bootstrap samples with blocks of 2000 informative sites was used to infer the 95% CI for the admixture timing.

Demographic model fitting with FastSimCoal2

We assembled 2D site-frequency-spectra (2D SFS) by generating .saf files with ANGSD (-minMapQ 20 -minQ 10 -remove_bads 1 -uniqueOnly 0 -doSaf 1 -domaf 1 -GL 1 -doMajorMinor 5 -minInd 10 -setMinDepthInd 2) and assembling pairwise 2D SFS using realSFS (76) (76) (76). Saf files were calculated at robust sites after removing positions inside or within 100 bp of exon coding positions as defined by GFF file (VectorBase-66_AaegyptiLVP_AGWG.gff) associated with the L5 assembly on Vectorbase.org. To reduce linkage disequilibrium among sites in the data, we randomly sampled 0.01 of the robust sites. Additionally, to avoid effects from the sex-differentiated region on chromosome 1, only sites on arms of chromosome 2 (10 Mb to 175 Mb and 275 to 450 Mb) and chromosome 3 (10 Mb to 150 Mb and 250 Mb and 400 Mb). We included four populations representing West African *Aaf* (PK10 from Senegal, PKT), African proto-*Aaa* (Ngoye, Senegal; NGO), South American proto-*Aaa* (El Dorado, Argentina; ELD), and *Aaa* (Maricopa County, Arizona, USA; ARI). To ease comparison among populations, we subsampled individuals from each population to include only 13 individuals per population to match Luanda, which only includes 13 individuals.

We defined two demographic models. The first model (one transatlantic emigration) describes four populations that merge sequentially going back in time in a stepping stone style tree (fig. S14). We allowed each population to change size once, except for PKT, which was allowed to change size twice. We also allowed migration between NGO and PKT as they are geographically proximate. The second model includes the same four populations (two transatlantic emigrations) but with two transatlantic emigrations where population 4 (ARI) merges into population 2 (NGO) and then population 3 (ELD) merges independently into population 2 (NGO), going backward in time. Like the one-emigration model, each population was allowed to change population size, and PKT and NGO were allowed to exchange migrants. The only difference between the models is whether ARI merges with ELD or with NGO, going backward in time. Both models included 18 free parameters. Model and parameter definition file contents are available from Zenodo (68).

We fit these models to 2D SFS using FastSimCoal2 (42) (ver 2.8.0.0 - 22.09.23) with default parameters except the following flags: -n 50000 -M -L 40 -c8. For mutation rate, we used the rate (4.85×10^{-9}) obtained in Rose *et al.* (2020) using a parameter maximization routine to find the value that best explains the data. In line with Crawford *et al.* (2017), we used the same value for recombination rate. To ensure we find the most likely model, we ran the model fitting 100 times for each model. The best-fitting model is shown in fig. S14. To obtain 95% CIs for each parameter value for both models, we assembled 100 bootstrapped genomes by dividing the regions included in the 2D SFS into 1-Mb windows and sampling with replacement. We specified bounded search ranges for each parameter in the .est files and fit each model to the bootstrapped genomes using the same approach as the initial optimization. To compare models, we calculated the Akaike Information Criterion (AIC) for each model as AIC = $2d - 2LL$, where d is the number of free parameters in the model and LL is the log likelihood of the model.

Subtree weighting analysis

To quantify support for various phylogenetic relationships along the genome, we used a method similar to *Twisst* (85) where we built neighbor-joining trees based on genetic distance among a representative set of individuals within 10-kb windows. First, we called consensus fasta sequences for eight individuals using ANGSD. We used the following commands for each individual: '-doFasta 3 -explode 0 -doCounts 1 -setMaxDepth 50 -setMinDepth 8 -minMapQ 20 -minQ 10 -only_proper_pairs 1 -remove_bads 1 -uniqueOnly 0'. We included one individual each from PK10 (Senegal), Ouahigouya (Burkina Faso), Luanda (Angola), Ngoye (Senegal), El Dorado (Argentina), New Orleans (USA), Playa Puerto Rico (USA), and Cebu City (Philippines). We then subdivided the genome into 10-kb windows, calculated the number differences among all pairs, and built neighbor-joining trees using the *ape* package (77) in R (75). We excluded windows where one or more comparisons in the matrix had less than 2000 sites with data. We tested for monophyly among certain groups using the *is.monophyletic* function in *ape*. To avoid biases from back-to-Africa *Aaa* admixture in proto-*Aaa* populations, we excluded any trees where *Aaa* and proto-*Aaa* were not monophyletic with respect to *Aaf.* We conducted bootstrap analyses by resampling 10-kb windows with replacement (1000 replicates) and repeating the analysis for each bootstrap replicate.

F_3 and F_4 admixture inference

We next examined contemporary patterns of long distance vector migration by identifying recent admixture events or secondary contact.

Putatively unadmixed parent populations from each subspecies were selected based on prior studies (13). *Aaf* populations included: Franceville, Gabon; Bantata and Kedogou, Senegal. *Aaa* populations included: New Orleans, USA and Ho Chi Minh City, Vietnam.

To preserve admixture signals related to LD while avoiding misleading genetic signals, repetitive regions and those of consistently low or high coverage were removed from the full SNP set followed by removal of SNPs with a minor allele frequency below 0.01 and random subsampling to 1×10^7 SNPs genome-wide.

Admixture was examined using Patterson's F_3 applied genome-wide with block jackknifing for each of our sampling sites, using unadmixed *Aaa* and *Aaf* parents. Directionality of the admixture was established with the Patterson's F_4 test, using Skukusa as an outgroup, due to its basal position in the phylogeny (Fig. 1). Both tests were implemented through scikit allel (https://github.com/cggh/scikit-allel). To avoid false positives from cryptic population structure we repeated all F_3 and F_4 tests with alternate parent populations.

To distinguish between scenarios of *Aaa*/*Aaf* admixture and *Aaf*/proto-*Aaa* admixture, Patterson's F_4 test was also applied to pure *Aaf* populations with the putative proto-*Aaa* population from El Dorado, Argentina.

KDR genotyping and haplotype network reconstruction

Initial genotyping on the assembled chromosome of AaegL5 (33) did not result in a call for the well characterized F1534C variant. In the likelihood that large and highly structured haplotypes in this region were assembled as alternative haplotypes, we aligned all alternative haplotypes to their most probable analog in the core genome using minimap2 (86), allowing up to 1% sequence divergence and alternative matches within 95% of the primary alignment [-ASM10 -p0.95]. Alternative haplotype NW_018735214.1 aligned across the 5′ end of KDR encompassing many of the known resistance loci. Variants called across this alternative haplotype were spliced into the original locus allowing the recovery of known alleles for F1534C, V1016G and V1016I, and others. Population level KDR mutation frequencies are listed in table S3.

Reads were realigned to this reconstructed region using BWA (60) and individual read-based phasing was performed using WhatsHap (87); the resulting phase blocks were phased across the whole dataset using SHAPEIT4 (72). Haplotype networks for this region were reconstructed across all non-synonymous coding SNPs using GeneHapR (88).

For an alternative view of KDR haplotype relationships, we constructed two hierarchical dendrograms and plotted with geographical locations and resistance mutation information below. The first dendrogram (fig. S24) is based on all callable SNPs (i.e., synonymous and non-synonymous) across the KDR locus, and the second dendrogram (fig. S25) is based on only non-synonymous SNPs as in the haplotype network described above. Haplotype distances were calculated via hamming distance and hierarchically clustered using the *hclust* function in R. SNP coordinates of resistance-associated loci within the AaegL5 reference were determined according to Mack *et al.* (89).

Scans for positive selection

We scanned the genome for highly differentiated regions using the PBS (49), which is a polarized measure of genetic differentiation intended to localize differentiation to one branch on a three-branch population tree. To minimize any impact of admixture on signals of differentiation, we used admixture-corrected allele frequencies obtained using NGSadmix (73). We ran NGSadmix on the full set of genotype likelihoods at SNPs described above with the minimum number of individuals with data (-minInd) set to 500, no minimum allele frequency filter (n = 67,182,646 SNPs after filtering), and the Q matrix with admixture proportions at K = 8 obtained from genetic structure analysis described above.

By supplying an existing Q matrix, NGSadmix assumes admixture proportions for each individual according to this matrix, and then estimates allele frequencies at each SNP under the assumed admixture model. We calculated PBS at SNPs with a minor allele frequency of at least 0.05 in the target population using sample sizes of 155, 18, and 231 for the target, contrast, and outgroup populations, respectively. We counted the number of individuals in each of the admixture components with admixture proportions of at least 0.5 to define sample sizes used in PBS calculations. We calculated PBS for individual SNPs as well as in 20-SNP windows. After removing windows greater than 10,000 bp in size, we ranked windows based on PBS scores, then ranked the top 1000 windows based on the number of individual SNPs in the top 0.01% of the SNP-wise distribution genome-wide. We obtained gene names and exon boundaries from the AaegL5 AGWG GFF hosted on Vectorbase.org (version 66, VectorBase-66_AaegyptiLVP_AGWG.gff). We plotted individual window values as well as the rolling mean of 30 windows. Additional regions of differentiation were identified and are listed in table S4.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We thank H. Bossin, P. Howell, and J. Mains, as well as the teams at Puerto Rico Vector Control Unit, US Virgin Islands Department of Health, and Consolidated Mosquito Abatement District for help with field collections. We also thank B. Shelden, A. Weakley, N. Barbaro, E. Rumsey, and J. Sullivan for helping prepare and sequence mosquito samples. We thank three reviewers for constructive feedback that helped substantially improve this manuscript.

Funding:

This work was supported by the Helen Hay Whitney Foundation Postdoctoral Fellowship (NHR); NIH NIAID award K22AI166268 (N.H.R.); the Robertson Investigator Award from the New York Stem Cell Foundation (C.S.M.); NIH grant R00DC012069 (C.S.M.); NIH grant R01AI155562 (J.R.P.); Biotechnology and Biological Sciences Research Council grant BBS/E/PI/230002C (K.M.); National Scientific and Technical Research Council grant CONICET PIP N 584 (M.V.M.); NIH NIAID award R01AI132409 (P.A.A.); National Institute for Food and Agriculture Multistate Research Project NE1943 (A.G.-S.); and UK Research and Innovation (UKRI) Medical Research Council (MRC) Doctoral Training Partnership grant MR/N013514/1 (H.A.Y.).

Data and materials availability:

DNA sequence reads for samples published in Rose *et al.* 2023 (28) have been deposited to NCBI SRA under accession no. PRJNA602495. The remaining samples have been deposited under accession no. PRJNA1185803. Variant data will be available on the Vectorbase.org platform. Custom scripts and supporting files will be available at Zenodo (68).

REFERENCES AND NOTES

1. Kraemer MUG et al. , Past and future spread of the arbovirus vectors *Aedes aegypti* and *Aedes albopictus*. Nat. Microbiol 4, 854–863 (2019). doi: 10.1038/s41564-019-0376-y [PubMed: 30833735]
2. World Health Organization, “Dengue – Global situation” (2023); https://www.who.int/emergencies/disease-outbreak-news/item/2023-DON498.

3. Pan American Health Organization, “Report on the epidemiological situation of dengue in the Americas” (PAHO, 2024); https://www.paho.org/sites/default/files/2024-12/2024-cde-dengue-sitrep-americas-epi-week-47-12-dec.pdf.
4. Bhatt S et al. , The global distribution and burden of dengue. Nature 496, 504–507 (2013). doi: 10.1038/nature12060; [PubMed: 23563266]
5. Soghigian J et al. , Genetic evidence for the origin of *Aedes aegypti*, the yellow fever mosquito, in the southwestern Indian Ocean. Mol. Ecol 29, 3593–3606 (2020). doi: 10.1111/mec.15590; [PubMed: 33463828]
6. Lounibos LP, Habitat segregation among African treehole mosquitoes. Ecol. Entomol 6, 129–154 (1981). doi: 10.1111/j.1365-2311.1981.tb00601.x
7. Gouck HK, Host preferences of various strains of Aedes aegypti and A. simpsoni as determined by an olfactometer. Bull. World Health Organ 47, 680–683 (1972). [PubMed: 4540689]
8. Petersen JL, “Behavior differences in two subspecies of *Aedes aegypti* (L.) (Diptera: Culicidae) in East Africa,” thesis, University Notre Dame, Notre Dame, IN (1977).
9. McBride CS et al. , Evolution of mosquito preference for humans linked to an odorant receptor. Nature 515, 222–227 (2014). doi: 10.1038/nature13964; [PubMed: 25391959]
10. Xia S et al. , Larval sites of the mosquito *Aedes aegypti formosus* in forest and domestic habitats in Africa and the potential association with oviposition evolution. Ecol. Evol 11, 16327–16343 (2021). doi: 10.1002/ece3.8332; [PubMed: 34824830]
11. Diouf B et al. , Resting Behavior of Blood-Fed Females and Host Feeding Preferences of Aedes aegypti (Diptera: Culicidae) Morphological Forms in Senegal. J. Med. Entomol 58, 2467–2473 (2021). doi: 10.1093/jme/tjab111; [PubMed: 34165556]
12. McClelland GA, Weitz B, Serological identification of the natural hosts of aedes aegypti (l.) and some other mosquitoes (diptera, culicidae) caught resting in vegetation in kenya and uganda. Ann. Trop. Med. Parasitol 57, 214–224 (1963). doi: 10.1080/00034983.1963.11686176; [PubMed: 14042652]
13. Rose NH et al. , Climate and urbanization drive mosquito preference for humans. Curr. Biol 30, 3570–3579.e6 (2020). doi: 10.1016/j.cub.2020.06.092; [PubMed: 32707056]
14. Mattingly PF, Genetical aspects of the *Aedes aegypti* problem. I. Taxonom: And bionomics. Ann. Trop. Med. Parasitol 51, 392–408 (1957). doi: 10.1080/00034983.1957.11685829; [PubMed: 13498658]
15. Brown JE et al. , Worldwide patterns of genetic differentiation imply multiple ‘domestications’ of *Aedes aegypti*, a major vector of human diseases. Proc. Biol. Sci 278, 2446–2454 (2011). doi: 10.1098/rspb.2010.2469; [PubMed: 21227970]
16. Brown JE et al. , Human impacts have shaped historical and recent evolution in *Aedes aegypti*, the dengue and yellow fever mosquito. Evolution 68, 514–525 (2014). doi: 10.1111/evo.12281; [PubMed: 24111703]
17. Tabachnick WJ, Powell JR, A world-wide survey of genetic variation in the yellow fever mosquito, *Aedes aegypti*. Genet. Res 34, 215–229 (1979). doi: 10.1017/S0016672300019467; [PubMed: 544311]
18. Paupy C et al. , Gene flow between domestic and sylvan populations of Aedes aegypti (Diptera: Culicidae) in North Cameroon. J. Med. Entomol 45, 391–400 (2008). doi: 10.1093/jmedent/45.3.391; [PubMed: 18533431]
19. Sylla M, Bosio C, Urdaneta-Marquez L, Ndiaye M, Black WC 4th, Gene flow, subspecies composition, and dengue virus-2 susceptibility among *Aedes aegypti* collections in Senegal. PLOS Negl. Trop. Dis 3, e408 (2009). doi: 10.1371/journal.pntd.0000408; [PubMed: 19365540]
20. Gloria-Soria A et al. , Global genetic diversity of *Aedes aegypti*. Mol. Ecol 25, 5377–5395 (2016). doi: 10.1111/mec.13866; [PubMed: 27671732]
21. Crawford JE et al. , Population genomics reveals that an anthropophilic population of Aedes aegypti mosquitoes in West Africa recently gave rise to American and Asian populations of this major disease vector. BMC Biol. 15, 16 (2017). doi: 10.1186/s12915-017-0351-0; [PubMed: 28241828]

22. Kotsakiozi P et al. , Population structure of a vector of human diseases: *Aedes aegypti* in its ancestral range, Africa. Ecol. Evol 8, 7835–7848 (2018). doi: 10.1002/ece3.4278; [PubMed: 30250667]
23. Lumsden WHR, An epidemic of virus disease in Southern Province, Tanganyika Territory, in 1952-53. II. General description and epidemiology. Trans. R. Soc. Trop. Med. Hyg 49, 33–57 (1955). doi: 10.1016/0035-9203(55)90081-X; [PubMed: 14373835]
24. Gillett JD, The Inherited Basis of Variation in the Hatching-response of *Aëdes* Eggs (Diptera: Culicidae). Bull. Entomol. Res 46, 255–265 (1955). doi: 10.1017/S0007485300030893
25. Leahy SMG, VandeHey RC, Booth KS, Differential response to oviposition site by feral and domestic populations of *Aedes aegypti* (L.) (Diptera: Culicidae). Bull. Entomol. Res 68, 455–463 (1978). doi: 10.1017/S0007485300009433
26. Powell JR, Gloria-Soria A, Kotsakiozi P, Recent history of *Aedes aegypti*: Vector genomics and epidemiology records. Bioscience 68, 854–860 (2018). doi: 10.1093/biosci/biy119; [PubMed: 30464351]
27. Metz HC et al. , Evolution of a Mosquito's Hatching Behavior to Match Its Human-Provided Habitat. Am. Nat 201, 200–214 (2023). doi: 10.1086/722481; [PubMed: 36724468]
28. Rose NH et al. , Dating the origin and spread of specialization on human hosts in *Aedes aegypti* mosquitoes. eLife 12, e83524 (2023). doi: 10.7554/eLife.83524; [PubMed: 36897062]
29. Eltis D, Richardson D, Atlas of the Transatlantic Slave Trade (Yale Univ. Press, 1995).
30. Gómez-Palacio A et al. , Robustness in population-structure and demographic-inference results derived from the Aedes aegypti genotyping chip and whole-genome sequencing data. G3 14, jkae082 (2024). doi: 10.1093/g3journal/jkae082; [PubMed: 38626295]
31. Dueñas JCR, Llinás GA, Panzetta-Dutari GM, Gardenal CN, Two different routes of colonization of Aedes aegypti in Argentina from neighboring countries. J. Med. Entomol 46, 1344–1354 (2009). doi: 10.1603/033.046.0613; [PubMed: 19960679]
32. The Anopheles gambiae 1000 Genomes Consortium, Genetic diversity of the African malaria vector *Anopheles gambiae*. Nature 552, 96–100 (2017). doi: 10.1038/nature24995; [PubMed: 29186111]
33. Matthews BJ et al. , Improved reference genome of *Aedes aegypti* informs arbovirus vector control. Nature 563, 501–507 (2018). doi: 10.1038/s41586-018-0692-z; [PubMed: 30429615]
34. Macdonald WW, The Genetic Basis of Susceptibility to Infection with Semi-Periodic *Brugia Malayi* in *aëdes Aegypti*. Ann. Trop. Med. Parasitol 56, 373–382 (1962). doi: 10.1080/00034983.1962.11686135
35. Gloria-Soria A, Soghigian J, Kellner D, Powell JR, Genetic diversity of laboratory strains and implications for research: The case of Aedes aegypti. PLOS Negl. Trop. Dis 13, e0007930 (2019). doi: 10.1371/journal.pntd.0007930; [PubMed: 31815934]
36. Evanno G, Regnaut S, Goudet J, Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. Mol. Ecol 14, 2611–2620 (2005). doi: 10.1111/j.1365-294X.2005.02553.x; [PubMed: 15969739]
37. Tabachnick WJ, Munstermann LE, Powell JR, Genetic Distinctness of Sympatric Forms of Aedes aegypti in East Africa. Evolution 33, 287–295 (1979). doi: 10.2307/2407619; [PubMed: 28568173]
38. Begun DJ, Aquadro CF, Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. Nature 356, 519–520 (1992). doi: 10.1038/356519a0; [PubMed: 1560824]
39. Severo OP, "Eradication of the Aedes Aegypti Mosquito from the Americas" in "Yellow fever, a symposium in commemoration of Carlos Juan Finlay, 1955" (Thomas Jefferson University, Paper 6, 1955); https://jdc.jefferson.edu/yellow_fever_symposium/6.
40. Popkin BM, Agricultural policies, food and public health. EMBO Rep. 12, 11–18 (2011). doi: 10.1038/embor.2010.200; [PubMed: 21151043]
41. Kotsakiozi P, Gloria-Soria A, Schaffner F, Robert V, Powell JR, Aedes aegypti in the Black Sea: Recent introduction or ancient remnant? Parasit. Vectors 11, 396 (2018). doi: 10.1186/s13071-018-2933-2; [PubMed: 29980229]

42. Excoffier L et al. , fastsimcoal2: Demographic inference under complex evolutionary scenarios. Bioinformatics 37, 4882–4885 (2021). doi: 10.1093/bioinformatics/btab468; [PubMed: 34164653]
43. Trpis M, Hausermann W, Genetics of house-entering behaviour in East African populations of *Aedes aegypti* (L.) (Diptera: Culicidae) and its relevance to speciation. Bull. Entomol. Res 68, 521–532 (1978). doi: 10.1017/S0007485300009494
44. Tabachnick WJ, Evolutionary Genetics and Arthropod-borne Disease: The Yellow Fever Mosquito. Am. Entomol 37, 14–26 (1991). doi: 10.1093/ae/37.1.14
45. Schaffner F, Mathis A, Dengue and dengue vectors in the WHO European region: Past, present, and scenarios for the future. Lancet Infect. Dis 14, 1271–1280 (2014). doi: 10.1016/S1473-3099(14)70834-5; [PubMed: 25172160]
46. Barrera R et al. , Unusual productivity of Aedes aegypti in septic tanks and its implications for dengue control. Med. Vet. Entomol 22, 62–69 (2008). doi: 10.1111/j.1365-2915.2008.00720.x; [PubMed: 18380655]
47. Lima A, Lovin DD, Hickner PV, Severson DW, Evidence for an Overwintering Population of Aedes aegypti in Capitol Hill Neighborhood, Washington, DC. Am. J. Trop. Med. Hyg 94, 231–235 (2016). doi: 10.4269/ajtmh.15-0351; [PubMed: 26526922]
48. Arsenault-Benoit A, Greene A, Fritz ML, Paved paradise: Belowground parking structures sustain urban mosquito populations in washington, DC. J. Am. Mosq. Control Assoc 37, 291–295 (2021). doi: 10.2987/21-7023; [PubMed: 34817615]
49. Yi X et al. , Sequencing of 50 human exomes reveals adaptation to high altitude. Science 329, 75–78 (2010). doi: 10.1126/science.1190371; [PubMed: 20595611]
50. Ramirez JL, Dimopoulos G, The Toll immune signaling pathway control conserved anti-dengue defenses across diverse *Ae. aegypti* strains and against multiple dengue virus serotypes. Dev. Comp. Immunol 34, 625–629 (2010). doi: 10.1016/j.dci.2010.01.006; [PubMed: 20079370]
51. Luong T, Characterization of RYamide neuropeptides and their receptor in the disease vector mosquito, Aedes aegypti. thesis, York University, Toronto, Ontario (2023).
52. Kotsakiozi P et al. , Tracking the return of Aedes aegypti to Brazil, the major vector of the dengue, chikungunya and Zika viruses. PLOS Negl. Trop. Dis 11, e0005653 (2017). doi: 10.1371/journal.pntd.0005653; [PubMed: 28742801]
53. Gainor EM, Harris E, LaBeaud AD, Uncovering the burden of dengue in africa: Considerations on magnitude, misdiagnosis, and ancestry. Viruses 14, 233 (2022). doi: 10.3390/v14020233; [PubMed: 35215827]
54. Chen M, Du Y, Nomura Y, Zhorov BS, Dong K, Chronology of sodium channel mutations associated with pyrethroid resistance in Aedes aegypti. Arch. Insect Biochem. Physiol 104, e21686 (2020). doi: 10.1002/arch.21686; [PubMed: 32378259]
55. Moyes CL et al. , Contemporary status of insecticide resistance in the major *Aedes* vectors of arboviruses infecting humans. PLOS Negl. Trop. Dis 11, e0005625 (2017). doi: 10.1371/journal.pntd.0005625; [PubMed: 28727779]
56. Cosme LV, Gloria-Soria A, Caccone A, Powell JR, Martins AJ, Evolution of kdr haplotypes in worldwide populations of Aedes aegypti: Independent origins of the F1534C kdr mutation. PLOS Negl. Trop. Dis 14, e0008219 (2020). doi: 10.1371/journal.pntd.0008219; [PubMed: 32298261]
57. Mboera LEG et al. , The risk of dengue virus transmission in Dar es Salaam, Tanzania during an epidemic period of 2014. PLOS Negl. Trop. Dis 10, e0004313 (2016). doi: 10.1371/journal.pntd.0004313; [PubMed: 26812489]
58. Badolo A et al. , First comprehensive analysis of Aedes aegypti bionomics during an arbovirus outbreak in west Africa: Dengue in Ouagadougou, Burkina Faso, 2016-2017. PLOS Negl. Trop. Dis 16, e0010059 (2022). doi: 10.1371/journal.pntd.0010059; [PubMed: 35793379]
59. Crawford JE et al. , Efficient production of male Wolbachia-infected Aedes aegypti mosquitoes enables large-scale suppression of wild populations. Nat. Biotechnol 38, 482–492 (2020). doi: 10.1038/s41587-020-0471-x; [PubMed: 32265562]
60. Li H, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997 [q-bio.GN] (2013).

61. Li H, A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27, 2987–2993 (2011). doi: 10.1093/bioinformatics/btr509; [PubMed: 21903627]
62. Picard (2025); http://broadinstitute.github.io/picard.
63. DePristo MA et al. , A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet 43, 491–498 (2011). doi: 10.1038/ng.806; [PubMed: 21478889]
64. Jun G, Wing MK, Abecasis GR, Kang HM, An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. Genome Res. 25, 918–925 (2015). doi: 10.1101/gr.176552.114; [PubMed: 25883319]
65. Korneliussen TS, Moltke I, NgsRelate: A software tool for estimating pairwise relatedness from next-generation sequencing data. Bioinformatics 31, 4009–4011 (2015). doi: 10.1093/bioinformatics/btv509; [PubMed: 26323718]
66. Korneliussen TS, Albrechtsen A, Nielsen R, ANGSD: Analysis of next generation sequencing data. BMC Bioinformatics 15, 356 (2014). doi: 10.1186/s12859-014-0356-4; [PubMed: 25420514]
67. Waples RK, Albrechtsen A, Moltke I, Allele frequency-free inference of close familial relationships from genotypes or low-depth sequencing data. Mol. Ecol 28, 35–48 (2019). doi: 10.1111/mec.14954; [PubMed: 30462358]
68. Crawford J, Rose N, Redmond S, Scripts and supporting files for Aaeg1200 Aedes aegypti genome sequencing dataset analysis, version v1, Zenodo (2025); 10.5281/zenodo.14503663.
69. Tange O, Parallel GNU, The Command-Line Power Tool. USENIX Mag. 36, 42–47 (2011).
70. Purcell S et al. , PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet 81, 559–575 (2007). doi: 10.1086/519795; [PubMed: 17701901]
71. Edge P, Bafna V, Bansal V, HapCUT2: Robust and accurate haplotype assembly for diverse sequencing technologies. Genome Res. 27, 801–812 (2017). doi: 10.1101/gr.213462.116; [PubMed: 27940952]
72. Delaneau O, Zagury J-F, Robinson MR, Marchini JL, Dermitzakis ET, Accurate, scalable and integrative haplotype estimation. Nat. Commun 10, 5436 (2019). doi: 10.1038/s41467-019-13225-y; [PubMed: 31780650]
73. Skotte L, Korneliussen TS, Albrechtsen A, Estimating individual admixture proportions from next generation sequencing data. Genetics 195, 693–702 (2013). doi: 10.1534/genetics.113.154138; [PubMed: 24026093]
74. Meisner J, Albrechtsen A, Inferring Population Structure and Admixture Proportions in Low-Depth NGS Data. Genetics 210, 719–731 (2018). doi: 10.1534/genetics.118.301336; [PubMed: 30131346]
75. R Core Team, R : A language and environment for statistical computing (R Foundation for Statistical Computing, 2024); https://www.R-project.org/.
76. Nielsen R, Korneliussen T, Albrechtsen A, Li Y, Wang J, SNP calling, genotype calling, and sample allele frequency estimation from New-Generation Sequencing data. PLOS ONE 7, e37558 (2012). doi: 10.1371/journal.pone.0037558; [PubMed: 22911679]
77. Paradis E, Schliep K, ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35, 526–528 (2019). doi: 10.1093/bioinformatics/bty633; [PubMed: 30016406]
78. Korneliussen TS, Moltke I, Albrechtsen A, Nielsen R, Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. BMC Bioinformatics 14, 289 (2013). doi: 10.1186/1471-2105-14-289; [PubMed: 24088262]
79. Kahle D, Wickham H, ggmap: Spatial Visualization with ggplot2. R J. 5, 144–161 (2013). doi: 10.32614/RJ-2013-014
80. Fox EA, Wright AE, Fumagalli M, Vieira FG, ngsLD: Evaluating linkage disequilibrium using genotype likelihoods. Bioinformatics 35, 3855–3856 (2019). doi: 10.1093/bioinformatics/btz200; [PubMed: 30903149]
81. Vieira FG, Albrechtsen A, Nielsen R, Estimating IBD tracts from low coverage NGS data. Bioinformatics 32, 2096–2102 (2016). doi: 10.1093/bioinformatics/btw212; [PubMed: 27153648]
82. Schiffels S, Wang K, MSMC and MSMC2: The multiple sequentially markovian coalescent. Methods Mol. Biol 2090, 147–166 (2020). doi: 10.1007/978-1-0716-0199-0_7;

83. Wang K, Mathieson I, O'Connell J, Schiffels S, Tracking human population structure through time from whole genome sequences. PLOS Genet. 16, e1008552 (2020). doi: 10.1371/journal.pgen.1008552; [PubMed: 32150539]
84. Corbett-Detig R, Nielsen R, A hidden markov model approach for simultaneously estimating local ancestry and admixture time using next generation sequence data in samples of arbitrary ploidy. PLOS Genet. 13, e1006529 (2017). doi: 10.1371/journal.pgen.1006529; [PubMed: 28045893]
85. Martin SH, Van Belleghem SM, Exploring evolutionary relationships across the genome using topology weighting. Genetics 206, 429–438 (2017). doi: 10.1534/genetics.116.194720; [PubMed: 28341652]
86. Li H, Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094–3100 (2018). doi: 10.1093/bioinformatics/bty191; [PubMed: 29750242]
87. Martin M et al. , WhatsHap: Fast and accurate read-based phasing. BioRxiv 085050 [Preprint] (2016). 10.1101/085050.
88. Zhang R, Jia G, Diao X, geneHapR: An R package for gene haplotypic statistics and visualization. BMC Bioinformatics 24, 199 (2023). doi: 10.1186/s12859-023-05318-9; [PubMed: 37189023]
89. Mack LK et al. , Frequency of sodium channel genotypes and association with pyrethrum knockdown time in populations of Californian *Aedes aegypti*. Parasit. Vectors 14, 141 (2021). doi: 10.1186/s13071-021-04627-3; [PubMed: 33676552]
90. SlaveVoyages.org, The Trans-Atlantic Slave Trade Database (2023); https://www.slavevoyages.org/voyage/trans-atlantic#voyages.
91. Beserra EB, de Castro FP, dos Santos JW, Santos T. da S., Fernandes CRM, Biology and thermal exigency of Aedes aegypti (L.) (Diptera: Culicidae) from four bioclimatic localities of Paraíba. Neotrop. Entomol 35, 853–860 (2006). doi: 10.1590/S1519-566X2006000600021; [PubMed: 17273720]

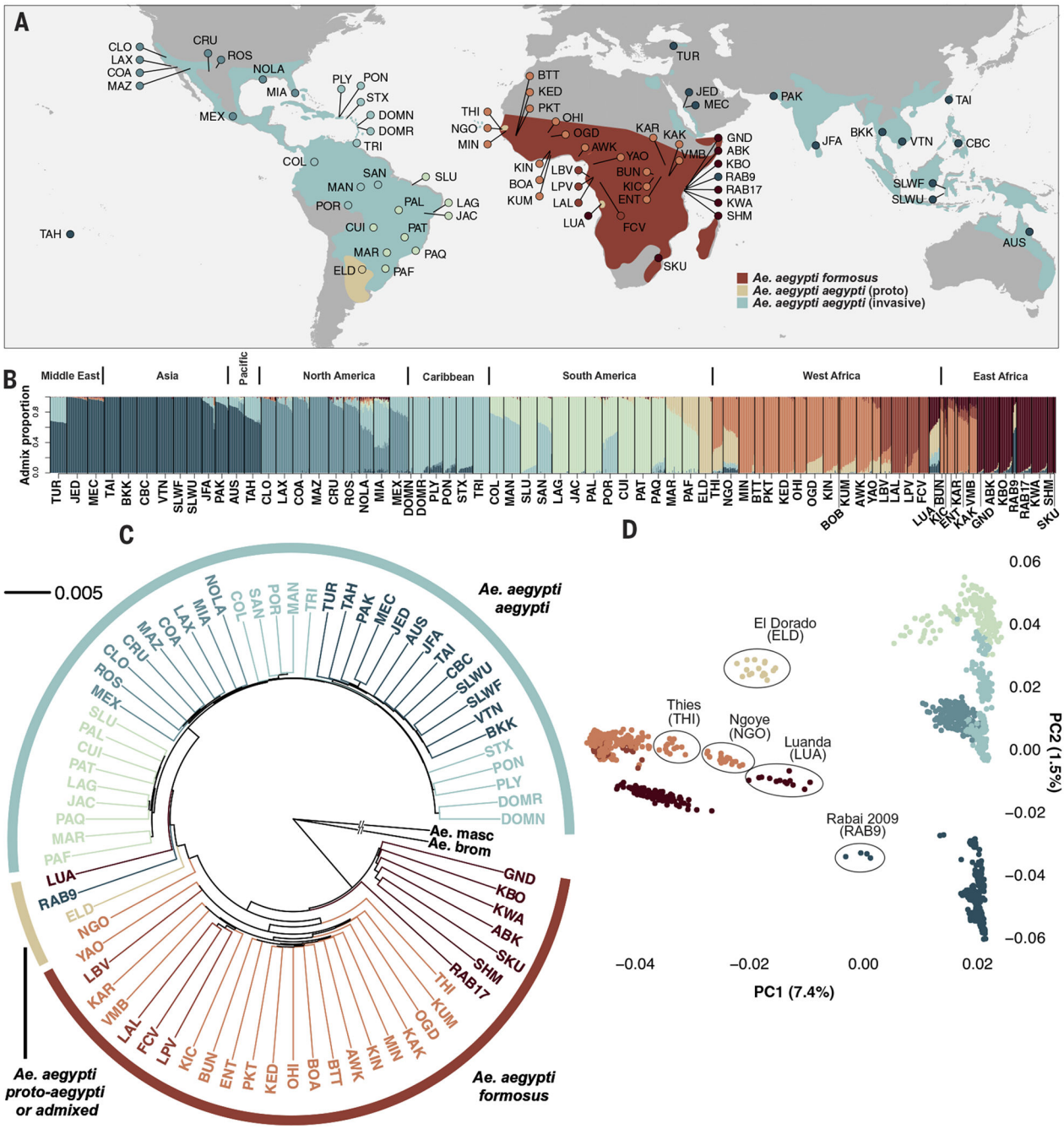


Fig. 1. Global genetic differentiation of *Aedes aegypti*.

(**A**) Global map showing locations of population samples included in Aaeg1200 sequencing. Details for individual population samples and full names can be found in table S1. (**B**) Bar plot showing admixture proportions assuming eight ancestral populations ($K = 8$). Each vertical bar shows admixture proportions for an individual mosquito, with each color corresponding to an ancestral population. (**C**) Neighbor-joining tree calculated on the basis of pairwise genetic distance (D_{xy}) among three representative individuals from each population and averaged at the population level. Outgroup branches are broken for ease of

visualization. Branch and label colors correspond to admixture plot colors and outer band indicates subspecies designation with colors corresponding to Fig. 1B. (**D**) PCA of genetic variation in all 1206 individuals. The first two principal component axes are shown with the percent of genetic variation explained shown in parentheses. Each dot represents one individual mosquito, with color indicating the largest admixture component from (B).

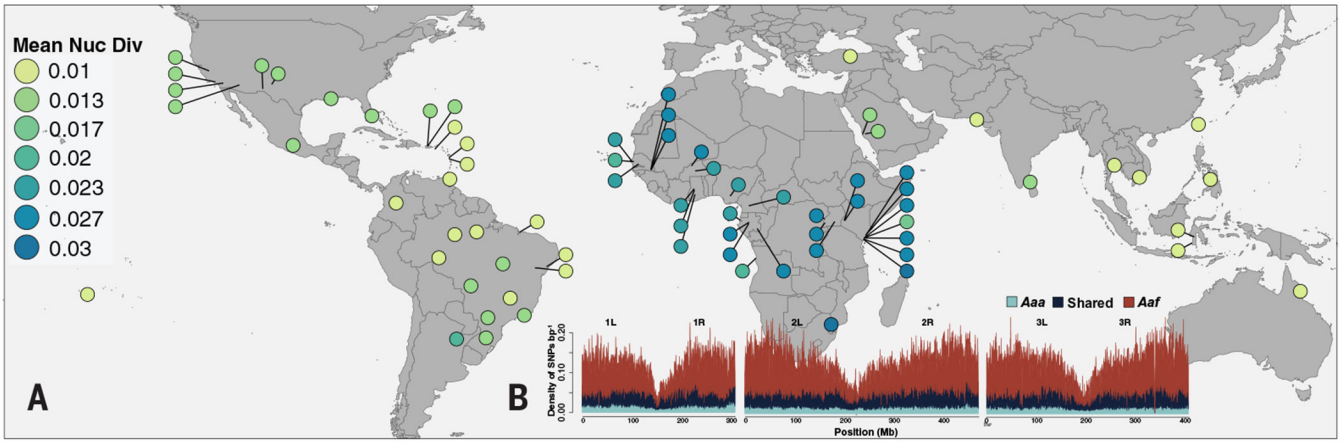
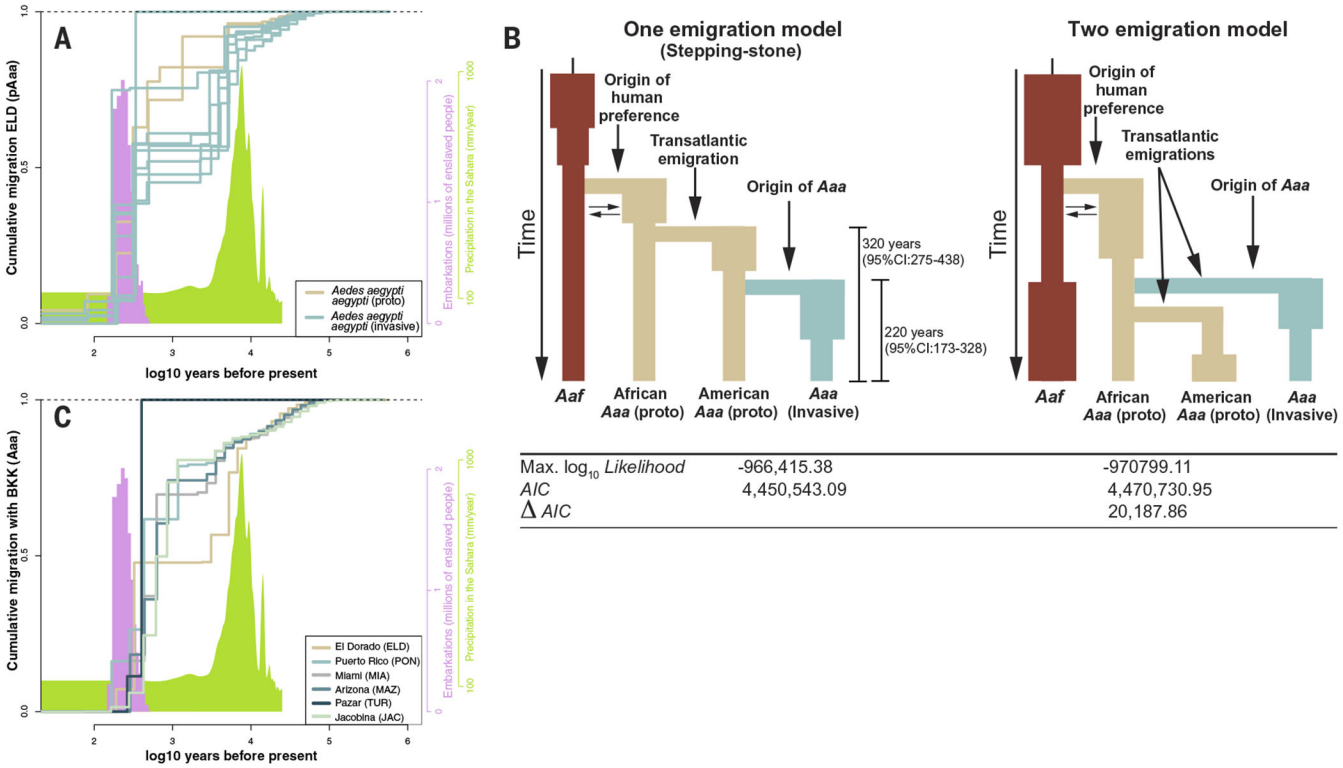


Fig. 2. Global genetic variation of *Aedes aegypti*.

(**A**) Genetic diversity (**π**) averaged across 50-kb windows for each population shown, with colors indicating value according to scale. Mean Nuc Div, mean nucleotide diversity. (**B**) High-confidence SNPs were classified as either exclusive to *Aaf*, exclusive to *Aaa*, or shared across the subspecies, counted within 250-kb windows, and plotted for each chromosomal arm.



Max. $\log_{10}$ *Likelihood*	-966,415.38	-970799.11
AIC	4,450,543.09	4,470,730.95
Δ *AIC*		20,187.86

Fig. 3. Population histories of *Aaf* and *Aaa*.

(**A**) Cumulative migration plotted as a function of time on $\log_{10}$ scale (assuming 15 generations per year) between representative populations and South American proto-*Aaa* (El Dorado and Argentina) estimated by using MSMC-IM. Cumulative migration is expected to plateau at one, going back in time. The pink histogram shows the estimated number of slave vessel embarkations scaled by millions of enslaved people (90) on the alternative, pink *y* axis at right. The green histogram shows precipitation levels in the Sahara (91) according to the green scale at right. Both histograms are shown for approximate temporal reference. (**B**) Schematic (not to scale) showing four-population demographic models describing evolution from *Aaf* to invasive *Aaa* that were fitted to 2D site-frequency spectra data. The one-emigration model describes a scenario in which invasive *Aaa* split from American proto-*Aaa* (ELD), and the two-emigration model describes the scenario in which invasive *Aaa* and American proto-*Aaa* both split separately from West African proto-*Aaa* and emigrated to the Americas. The $\log_{10}$ likelihood and AIC values for both best-fit models are below. Full description of models and parameters is found in fig. S14. (**C**) Same as shown in (A) but plotting cumulative migration between representative populations and Bangkok.

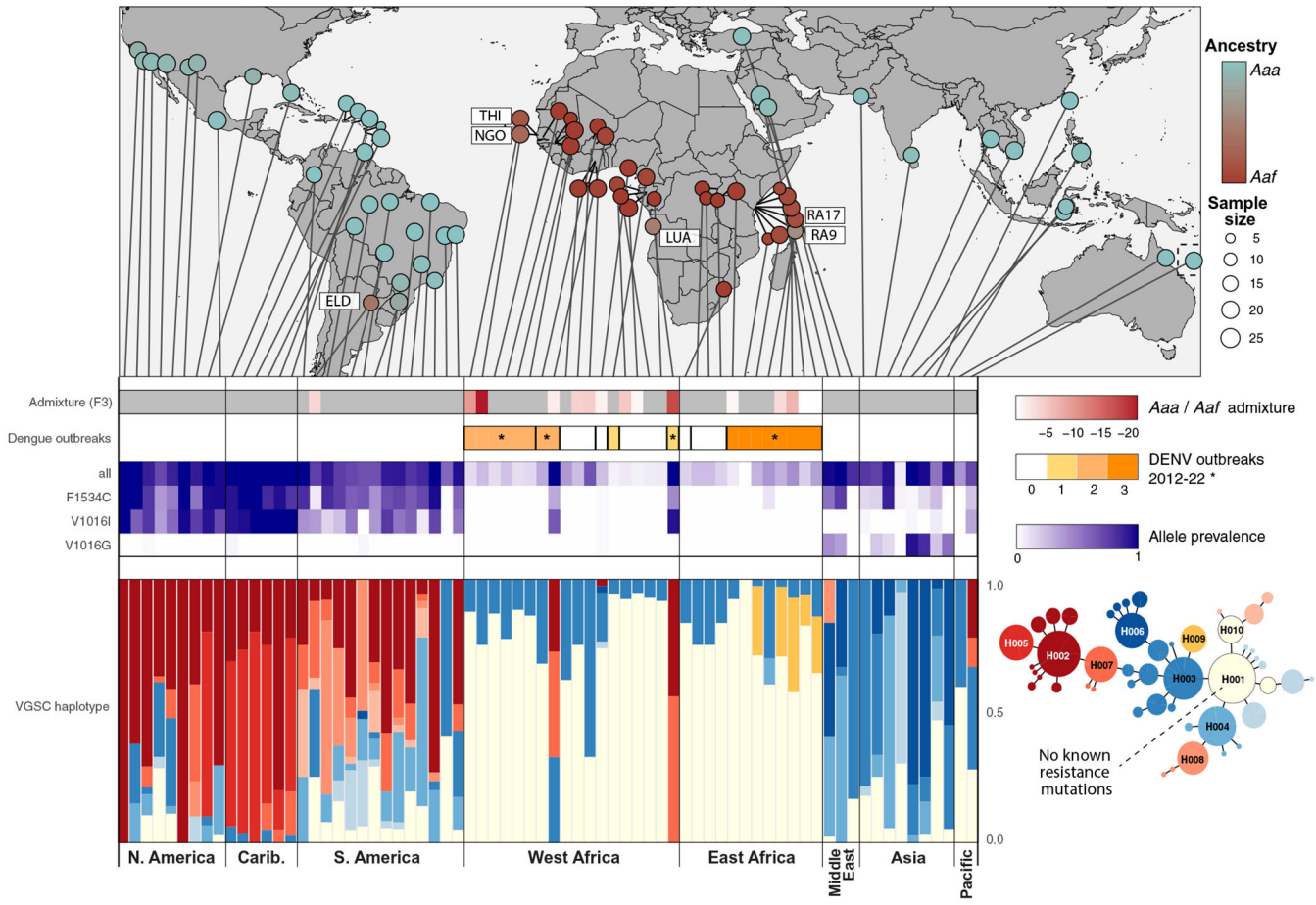


Fig. 4. *Aaa/Aaf* admixture spreads insecticide resistance and dengue risk.

The first (top) panel shows a global map presenting evidence of *Aaa/Aaf* admixture proportions from NGSadmix analysis, with $K = 2$. Lollipop markers used to offset populations when actual locations were too tightly clustered for visualization, with notable populations given labels as in Fig. 1. Gray lines indicate the position of each population in bar plots below. The second panel shows results of the F_3 test for admixture according to Z-score scale at the right, with darker red corresponding to evidence of stronger admixture. The third panel shows the number of dengue virus (DENV) outbreaks in African countries according to Gainor, Harris, and LaBeaud (53), and asterisks show countries with at least one admixed location and one dengue outbreak. The fourth panel shows population-level prevalence of KDR mutations as labeled on the left with colors indicating prevalence according to the scale on right. The haplotype network (bottom right) shows the global frequency and relationships among the top 10 KDR haplotypes (labeled H001 to H010); haplotypes at lower frequencies are clustered with their nearest-related larger haplotype. Haplotypes colored in cream lack known resistance mutations, and resistant haplotypes are colored according to the continent with the highest frequency (Asia: blue; Americas: red; Africa: yellow). The bottom panel bar plot shows population-level frequency of each of the top eight resistant KDR haplotypes, with colors indicating related haplotypes in the network on the right. See fig. S22 for a detailed KDR haplotype network.

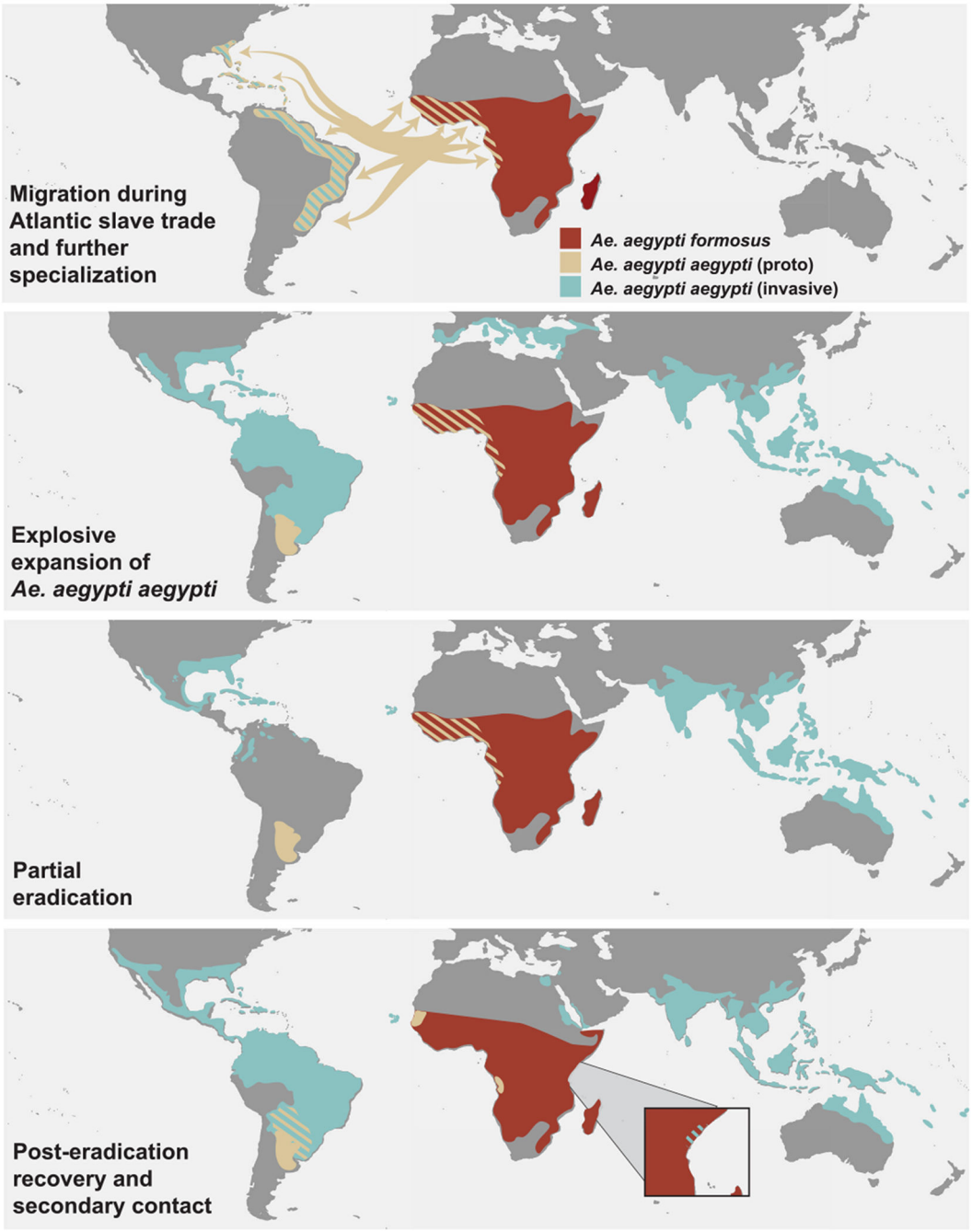


Fig. 5. Four eras in the evolution of invasive *Aedes aegypti aegypti*.

Each subspecies and subgroup is shown in different colors according to legend. Distributions are inferred from historical records and interpreted from our results, including the region in Argentina where proto-*Aaa* is thought to have escaped eradication. Each phase is described in the bottom left corner. Arrows in the first phase indicate movement of proto-*Aaa* during the Atlantic slave trade. Striped areas indicate more than one form or subspecies found in

that region at that time, or areas where ancestry from both groups are found in current samples consistent with recent or historical migration.