Improved reference genome of *Aedes aegypti* informs arbovirus vector control

Benjamin J. Matthews1,2,3,4,6, Olga Dudchenko4,5,6,7,49, Sarah B. Kingan8,49, Sergey Koren9, Igor Antoshechkin10, Jacob E. Crawford11, William J. Glassford12, Margaret Herre1,3, Seth N. Redmond13,14, Noah H. Rose15,16, Gareth D. Weedall17,18, Yang Wu19,20,21, Sanjii S. Batra23,24, Carlos A. Brito-Sierra21,22, Steven D. Buckingham24, Corey L. Campbell25, Saki Chan26, Eric Cox27, Benjamin R. Evans28, Thanyalak Fansiri29, Igor Filipovic30, Albin Fontaine31,32,33,34, Andrea Gloria-Soria26,35, Richard Hall36, Vicente Hernández37, Andrea H. Huerta31,32,33,34, rattananukul k. Kiatkamjorn38, Vamsi K. Kodali27, Joyce Lee39, Gareth J. Lycett32, Sara N. Mitchell31, Jill Muehling32, Michael R. Murphy27, Arina D. Orme4,5,6, Frederick A. Partridge48, Paul Peulos30, Aviva Presser Aiden4,5,8,19,20,21, Vidya Ramasamy36, Goran da Rivi30, Sourav Roy40, Karla Saavedra-Rodriguez25, Shruti Sharan22,23, Atashi Sharma24,41, Melissa Laird Smith24, Joe Turner24, Allison M. Weakley31, Zhlei Zhao31,15,16, Omar S. Akbari3,14,44, William C. Black IV25,28, Han Cao26, Alistair C. Darby32, Catherine A. Hill22,23, T. Spencer Johnston45, Terence D. Murphy22, Alexander S. Raikhel10, David B. Sattelle24, Igor V. Sarakhkov32,41,46, Bradley J. White11, Li Zhao47, Erez Lieberman Aiden4,5,6,7,13, Richard S. Mann22, Louis Lambrechts13,11, Jeffrey R. Powell38, Maria V. Sharakhova21,41,46, Zhijian Tu20,21, Hugh M. Robertson48, Carolyn S. McBride15,16, Alex R. Hastie26, Jonas Krollach8, Daniel E. Neafsey13,14, Adam M. Phillippy9 & Leslie B. Voshall1,2,3

Female *Aedes aegypti* mosquitoes infect more than 400 million people each year with dangerous viral pathogens including dengue, yellow fever, Zika and chikungunya. Progress in understanding the biology of mosquitoes and developing the tools to fight them has been slowed by the lack of a high–quality genome assembly. Here we combine diverse technologies to produce the markedly improved, fully re-annotated AaegL5 genome assembly, and demonstrate how it accelerates mosquito science. We anchored physical and cytogenetic maps, doubled the number of known chemosensory ionotropic receptors that guide mosquitoes to human hosts and egg-laying sites, provided further insight into the size and composition of the sex-determining M locus, and revealed copy-number variation among glutathione S-transferase genes that are important for insecticide resistance. Using high-resolution quantitative trait locus and population genomic analyses, we mapped new candidates for dengue vector competence and insecticide resistance. AaegL5 will catalyse new biological insights and intervention strategies to fight this deadly disease vector.

An accurate and complete genome assembly is required to understand the unique aspects of mosquito biology and to develop control strategies to reduce their capacity to spread pathogens. The *Ae. aegypti* genome is large (approximately 1.25 Gb) and highly repetitive, and a 2007 genome project (AaegL3)2 was unable to produce a contiguous genome fully anchored to a physical chromosome map (Fig. 1a). A more recent assembly, AaegL4, produced chromosome-length scaffolds that made it possible to detect larger-scale syntenic genomic regions in other species but suffered from short contigs (contig N50, 84 kb, meaning that half of the assembly is found on contigs >84 kb) and a correspondingly large number of gaps (31,018; Fig. 1b). Taking advantage of rapid advances in sequencing and assembly technology in the last decade, we used long-read Pacific Biosciences sequencing and Hi-C (a high-throughput sequencing method based on chromosome conformation capture) scaffolding to produce a new reference genome (AaegL5) that is highly contiguous, with a decrease of

---

1 Laboratory of Neurogenetics and Behavior, The Rockefeller University, New York, NY, USA. 2Howard Hughes Medical Institute, New York, NY, USA. 3Kavli Neural Systems Institute, New York, NY, USA. 4The Center for Genome Architecture, Baylor College of Medicine, Houston, TX, USA. 5Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA. 6Pacific Biosciences, Menlo Park, CA, USA. 7National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. 8Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, USA. 9Department of Developmental Biology, California Institute of Technology, Pasadena, CA, USA. 10Center for Theoretical and Biological Physics, Rice University, Houston, TX, USA. 11Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, USA. 12Department of Computer Science, Rice University, Houston, TX, USA. 13Center for Theoretical and Biological Physics, Rice University, Houston, TX, USA. 14Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA. 15Department of Biology, Rice University, Houston, TX, USA. 16Department of Computer Science, Rice University, Houston, TX, USA. 17Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ, USA. 18Division of Molecular Biology, Department of Pathogen Biology, School of Public Health, Southern Medical University, Guangzhou, China. 19Department of Biochemistry, Virginia Tech, Blacksburg, VA, USA. 20Flinn Life Science Institute, Virginia Tech, Blacksburg, VA, USA. 21Department of Entomology, Purdue University, West Lafayette, IN, USA. 22Purdue Institute for Inflammation, Immunology and Infectious Disease, Purdue University, West Lafayette, IN, USA. 23Centre for Respiratory Biology, UCL Respiratory, University College London, London, UK. 24Department of Microbiology, Immunology and Pathology, Colorado State University, Fort Collins, CO, USA. 25Bionano Genomics, San Diego, CA, USA. 26National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA. 27Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT, USA. 28Division of Biological and Control Section, Department of Entomology, Armed Forces Research Institute of Medical Sciences (AFRIMS), Bangkok, Thailand. 29Mosquito Control Laboratory, QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia. 30Insect-Virus Interactions Group, Department of Genomes and Genetics, Institut Pasteur, Paris, France. 31Unité de Parasitologie et Entomologie, Département des Maladies Infectieuses, Institut de Recherche Biomédicale des Armées, Marseille, France. 32Centre National de la Recherche Scientifique, Unité Mixte de Recherche 2000, Paris, France. 33Unité de Parasitologie et Entomologie, Département des Maladies Infectieuses, Institut de Recherche Biomédicale des Armées, Marseille, France. 34Aix Marseille Université, IRD, AP-HM, SSA, UMR Vecteurs – Infections Tropicales et Méditerranéennes (VITROME), IHU - Méditerranée Infection, Marseille, France. 35The Connecticut Agricultural Experiment Station, New Haven, CT, USA. 36Department of Biological and Medical Sciences, Faculty of Health and Life Sciences, Oxford Brookes University, Oxford, UK. 37Department of Entomology, University of California Riverside, Riverside, CA, USA. 38Department of Bioengineering, Rice University, Houston, TX, USA. 39Department of Pediatrics, Texas Children's Hospital, Houston, TX, USA. 40Department of Entomology, Center for Disease Vector Research and Institute for Integrative Genome Biology, University of California, Riverside, CA, USA. 41Department of Pathogen Biology, School of Public Health, Southern Medical University, Guangzhou, China. 42Department of Biochemistry, Virginia Tech, Blacksburg, VA, USA. 43Institute of Integrative Biology, University of Liverpool, Liverpool, UK. 44Department of Biological Sciences, University of California, San Diego, La Jolla, CA, USA. 45Department of Entomology, Texas A&M University, College Station, TX, USA. 46Laboratory of Genetics, Environnemental Protection, Tomsk State University, Tomsk, Russia. 47Laboratory of Evolutionary Genetics and Genomics, The Rockefeller University, New York, NY, USA. 48Department of Entomology, University of Illinois at Urbana-Champaign, Urbana, IL, USA. 49These authors contributed equally: Benjamin J. Matthews, Olga Dudchenko, Sarah B. Kingan. *e-mail: bmmtthws@gmail.com

© 2018 Springer Nature Limited. All rights reserved.
Fig. 1 | AaegL5 assembly statistics and annotation. a, b, Treemap of AaegL3 (a) and AaegL4 (b) contigs scaled by length. c, Principal component analysis of allelic variation of the indicated strains at 11,229 SNP loci. n = 7 per genotype d, Flow cytometry analysis of LVP_AGWG genome size. n = 5 per sex. Box plot: median is indicated by the blue line; boxes show first to third quartiles, whiskers are the 1.5× interquartile interval (Extended Data Fig. 1b). e, Treemap of AaegL5 contigs scaled by length. f, Genome composition (Supplementary Data 2, 3). g, Gene set alignment BLASTp coverage is compared between AaegL3.4 and AaegL5.0, with D. melanogaster protein queries. h, Alignment of 253 RNA-seq libraries to AaegL3.4 and AaegL5.0 gene set annotations (Supplementary Data 4–9). LTR, long terminal repeat retrotransposon; MITES, miniature inverted-repeat transposable elements; SINES, short interspersed nuclear element.

93% in the number of contigs, and anchored end-to-end to the three Ae. aegypti chromosomes (Fig. 1 and Extended Data Figs. 1, 2). Using optical mapping and linked-read sequencing, we validated the local structure and predicted structural variants between haplotypes. We generated an improved gene set annotation (AaegL5.0), as assessed by a mean increase in RNA-sequencing (RNA-seq) read alignment of 12%, connections between many gene models that were previously split across multiple contigs, and a roughly twofold increase in the enrichment of assay for transposase-accessible chromatin using sequencing (ATAC-seq) alignments near predicted transcription start sites. We demonstrate the utility of AaegL5 and the AaegL5.0 annotation by investigating a number of scientific questions that could not be addressed with the previous genome annotations.

This project used the Liverpool Aedes Genome Working Group (LVP_AGWG) strain, related to the AaegL3 Liverpool ib12 (LVP_ib12) assembly strain (Fig. 1c and Extended Data Fig. 1a). Using flow cytometry, we estimated that the genome size of LVP_AGWG is approximately 1.22 Gb (Fig. 1d and Extended Data Fig. 1b). To generate our primary assembly, we produced 166 Gb of Pacific Biosciences data (around 130× coverage for a 1.28-Gb genome) and assembled the genome using FALCON-Unzip. This resulted in a total assembly length of 2.05 Gb (contig N50, 0.96 Mb; and NG50, 1.92 Mb, meaning half of the expected genome size found on contigs >1.92 Mb). FALCON-Unzip annotated the resulting contigs as either primary (3,967 contigs; N50, 1.30 Mb; NG50, 1.91 Mb) or haplotigs (3,823 contigs; NG50, 193 kb), representing alternative haplotypes present in the approximately 80 male siblings pooled for sequencing (Table 1 and Extended Data Fig. 1e). The primary assembly was longer than expected for a haploid Ae. aegypti genome, as predicted by flow cytometry and prior assemblies, which was consistent with remaining alternative haplotypes that were too divergent to be automatically identified as primary and associated alternative haplotype pairs.

To generate a linear chromosome-scale reference genome assembly, we combined the primary contigs and haplotigs that were generated by FALCON-Unzip to create an assembly comprising 7,790 contigs. We used Hi-C to order and orient these contigs, correct misjoined sections and merge overlaps (Extended Data Fig. 1c–e). We set aside 359 contigs that were shorter than 20 kb and used the Hi-C data to identify 258 misjoined sections, resulting in 8,306 ordered and oriented contigs. This procedure revealed extensive sequence overlap among the contigs, consistent with the assembly of numerous alternative haplotypes. We developed a procedure to merge these alternative haplotypes, removing 5,440 gaps and boosting the contiguity (N50, 5.0 Mb; NG50, 4.6 Mb). This procedure placed 94% of sequenced (non-duplicated) bases onto three chromosome-length scaffolds that correspond to the three Ae. aegypti chromosomes. After scaffolding, we performed gap-filling and polishing using Pacific Biosciences reads. This removed 270 gaps and further increased the contiguity (N50, 11.8 Mb; NG50, 11.8 Mb), resulting in a final 1.279-Gb AaegL5 assembly and a complete mitochondrial genome (Fig. 1e and Table 1). We used Hi-C contact maps to estimate the position of the centromere with a resolution of around 5 Mb: chromosome 1, approximately 150–154 Mb; chromosome 2, around 227–232 Mb, chromosome 3, around 196–201 Mb. There are 229 remaining gaps in the primary assembly, including 173 in the three primary chromosome scaffolds (Extended Data Fig. 2a and

**Table 1 | Comparison of assembly statistics**

| Component | AaegL3 | AaegL4 | AaegL5 FALCON-Unzip | AaegL5 (NCBI) FALCON-Unzip | Hi-C | polish |
|------------|--------|--------|----------------------|---------------------------|-----|--------|
| Total length (non-N bp) | 1,310,092,987 | 1,254,548,160 | 1,695,064,654 | 1,278,709,169 | N/A | N/A |
| Contig number | 36,205 | 37,224 | 3,967 | 2,539 | N/A | N/A |
| Contig N50 (bp) | 82,618 | 84,074 | 1,304,397 | 11,758,062 | N/A | N/A |
| Contig NG50 (bp) | 85,904 | 81,911 | 1,907,936 | 11,758,062 | N/A | N/A |
| Scaffold number | 4,757 | 6,206 | N/A | 2,310 | N/A | N/A |
| Scaffold N50 (bp) | 1,547,108 | 404,248,146 | N/A | 409,777,670 | N/A | N/A |
| GC content (%) | 38.27 | 38.28 | 38.16 | 38.18 | N/A | N/A |
| Alternative haplotypes (bp) | N/A | N/A | 351,566,101 | 591,941,260 | N/A | N/A |
| Alternative haplotypes (contigs) | N/A | N/A | 3,823 | 4,224 | N/A | N/A |

N/A, not applicable.

* Scaffold N50 is the length of chromosome 3.
The blunt end of the arrowheads marks gene position and the arrow indicates orientation. Filled and open arrowheads represent intact genes and pseudogenes, respectively (Supplementary Data 17–20 and Extended Data Fig. 3).
Fig. 3 | Application of AaegL5 to resolve the sex-determining locus. 
a. M locus structure indicating high alignment identity (grey-dashed boxes) and boundaries of myo-sex and Nix gene models (magenta and white boxes, arrowheads represent orientation). b. FISH of BAC clones containing myo-sex and Nix. Scale bar, 2 μm. Representative image of 10 samples. c. De novo optical map spanning the M locus and bridging the estimated 163-kb gap in the AaegL5 assembly. DNA molecules are cropped at the edges for clarity. d. Chromosome quotient (CQ) analysis of genomic DNA from male and female libraries aligned to AaegL5 chromosome 1. Each dot represents the CQ value of a repeat-masked 1-kb window with >20 reads aligned from male libraries.

may represent the divergent M/m locus (Fig. 3a). A de novo optical map assembly spanned the putative AaegL5 M locus and extended beyond its two borders. We estimated the size of the M locus at approximately 1.5 Mb, including an approximately 163-kb gap between contigs (Fig. 3a, c). We tentatively identified the female m locus as the region in AaegL4 not shared with the M locus-containing chromosome 1, but note that the complete phased structure of the divergent male M locus and corresponding female m locus remain to be determined. Nix contains a single intron of 100 kb, while myo-sex, a gene encoding a myosin heavy chain protein that has previously been shown to be tightly linked to the M locus 18, is approximately 300 kb in length. More than 73.7% of the M locus is repetitive: long terminal repeat retrotransposons comprise 29.9% of the M locus compared to 11.7% genome-wide. 73.7% of the M locus is repetitive: long terminal repeat retrotransposons comprise 29.9% of the M locus compared to 11.7% genome-wide. Chromosomal FISH with Nix- and myo-sex-containing BAC clones 19 showed that these genes co-localize to the 1p pericentromeric region (1p11) in only one homologous copy of chromosome 1, supporting the placement of the M locus at this position in AaegL5 (Fig. 3b). We investigated the differentiation between the sex chromosomes (Fig. 3d) using a chromosome quotient (CQ) method to quantify regions of the genome with a strictly male-specific signal 20. A sex-differentiated region in the LVP_AGWG strain extends to an approximately 100-Mb region surrounding the approximately 1.5-Mb M locus. This is consistent with the recent analysis of male–female FST in wild population samples and linkage map intercrosses 21 and could be explained by a large region of reduced recombination encompassing the centromere and M locus 22. The availability of a more completely assembled mosquito M locus provides opportunities to study the evolution and maintenance of homomorphic sex-determining chromosomes. The sex-determining

Fig. 4 | Copy-number variation in the glutathione S-transferase epsilon gene cluster. a. Glutathione S-transferase epsilon (GSTε) gene cluster structure in AaegL5 compared to AaegL3 (Supplementary Data 23). Arrowheads indicate gene orientation. b. Dot-plot alignment of AaegL5 GSTε region to itself. c. Optical mapping of DNAlabelled with indicated enzymes. DNA molecules are cropped at the edges for clarity. d. Genomic sequencing coverage of AaegL3 GSTε genes (DNA read pairs mapped to each gene, normalized by gene length in kb) from one LVP_AGWG male and pooled mosquitoes from four other laboratory strains.
chromosome of *Ae. aegypti* may have remained homomorphic at least since the evolutionary divergence between the *Aedes* and *Culex* genera more than 50 million years ago. With the more completely assembled *M* locus, we can investigate how these chromosomes have avoided the proposed eventual progression into heteromorphic sex chromosomes.23

**Structural variation and gene families**

Structural variation is associated with the capacity to vector pathogen-gens.24 We produced ‘read cloud’ Illumina sequencing of linked reads with long-range (around 80 kb) phasing information from one male and one female mosquito using the 10X Genomics Chromium platform to investigate structural variants, including insertions, deletions, translocations and inversions, in individual mosquitoes. We observed abundant small-scale insertions and deletions (indels; 26 insertions and 81 deletions called, median 42.9 kb) and inversions and/or translocations (29 called) in these two individuals (Extended Data Fig. 8a and Supplementary Data 21). Eight of the inversions and translocations coincided with structural variants seen independently by Hi-C or FISH, suggesting that those variants are relatively common within this population and can be detected by different methods. AaegL5 will provide a foundation for the study of structural variants across *Ae. aegypti* populations.

*Hox* genes encode highly conserved transcription factors that specify segment identity along the anterior–posterior body axis of all metazoans.25 In most vertebrates, *Hox* genes are clustered in a co-linear arrangement, although they are often disorganized or split in other animal lineages.26 All expected *Hox* genes are present as a single copy in *Ae. aegypti*, but we identified a split between *labial* and *proboscipedia* placing *labial* on a separate chromosome (Extended Data Fig. 8b and Supplementary Data 22). We confirmed this in AaegL4, which was generated with Hi-C contact maps from a different *Ae. aegypti* strain, and note a similar arrangement in *Culex quinquefasciatus*, suggesting that it occurred before these two species diverged. Although this is not unprecedented,27 a unique feature of this organization is that both *labial* and *proboscipedia* appear to be close to telomeres.

Glutathione S-transferases (GSTs) are a large multi-gene family involved in the detoxification of compounds such as insecticides. Increased GST activity has been associated with resistance to multiple classes of insecticides, including organophosphates, pyrethroids and the organochlorine dichlorodiphenyltrichloroethane (DDT).28 Amplification of detoxification genes is one mechanism by which insects can develop insecticide resistance.29 We found that three insect-specific GST epsilon (GSTe) genes on chromosome 2, located centrally in the cluster (GSTe2, GSTe5 and GSTe7), are duplicated four times in AaegL5 relative to AaegL3 (Fig. 4a, b and Supplementary Data 23). Short Illumina read coverage and optical maps confirmed the copy number and arrangement of these duplications in AaegL5 (Fig. 4c, d), and analysis of whole-genome sequencing data for four additional laboratory colonies showed variable copy numbers across this gene cluster (Fig. 4d). GSTe2 is a highly efficient metaboliser of DDT,30 and it is noteworthy that the cDNA from three GST genes in DDT-resistant *Ae. aegypti* mosquitoes from southeast Asia.31

**Genome-wide genetic variation**

Measurement of genetic variation within and between populations is important for inferring ongoing and historic evolution in a species.32 To understand genomic diversity in *Ae. aegypti*, which spread in the last century from Africa to tropical and subtropical regions around the world, we performed whole-genome resequencing on four laboratory colonies. Chromosomal patterns of nucleotide diversity should correlate with regional differences in meiotic recombination rates.

---

**Fig. 5** | Using the AaegL5 genome for applied population genetics.

a. Heat map of linkage based on pairwise recombination fractions for 255 RAD markers ordered by AaegL5 physical coordinates. b. Significant QTLs on chromosome 1 underlying genome-wide association mapping of DENV variant fitness (odds ratio (LOD) scores obtained by interval mapping). Dotted vertical lines indicate genome-wide statistical significance thresholds ($\alpha = 0.05$). Confidence intervals of significant QTLs: bright colour, 1.5-LOD interval; light colour, 2-LOD interval with generalist effects (black, across DENV serotypes and isolates) and DENV isolate-specific effects (red, indicative of genotype-by-genotype interactions). c, d. Synteny between linkage map (in cM) and physical map (in Mb) for chromosome 2 (c) and chromosomes 1 and 3 (d). The orange color of chromosome 1 denotes uncertainty in the cM estimates because of deviations in Mendelian ratios surrounding the *M* locus. e. Chromosome 3 SNPs significantly correlated with deltamethrin survival. f, g. Magnified and inverted view of box in e, centred in the new gene model of voltage-gated sodium channel (VGSC, transcript variant X3; the chromosomal position is indicated in red). f. Non-coding genes are omitted for clarity, and other genes indicated with grey boxes. VGSC exons are represented by tan boxes and untranslated regions by short boxes. Arrowheads indicate gene orientation. Non-synonymous VGSC SNPs are marked with larger black and yellow circles: *V1016I* = 315,983,763; *F1534C* = 315,939,224; *V410L* = 316,080,722. g. Difference in expected heterozygosity ($H_{exp}$, alive – $H_{exp}$, dead) for all SNPs.
We observed pronounced declines in genetic diversity near the centre of each chromosome (Extended Data Fig. 9a, b), providing independent corroboration of the estimated position of each centromere by Hi-C (Extended Data Fig. 2a).

To investigate linkage disequilibrium in geographically diverse populations of *Ae. aegypti*, we first mapped Affymetrix SNP-Chip markers that were designed using AaegL3,4 to positions on AaegL5. We genotyped 28 individuals from two populations from Amacuzac, Mexico and Lopé National Park, Gabon and calculated the pairwise linkage disequilibrium of single-nucleotide polymorphisms (SNPs) from 1-kb bins both genome-wide and within each chromosome (Extended Data Fig. 9c, d). The maximum linkage disequilibrium in the Mexican population is approximately twice that of the population from Gabon, which probably reflects a recent bottleneck associated with the spread of this species out of Africa.

**Dengue competence and pyrethroid resistance**

To illustrate the value of AaegL5 for mapping quantitative trait loci (QTLs), we used restriction site-associated DNA (RAD) markers to locate QTLs underlying dengue virus (DENV) vector competence. We identified and genotyped RAD markers in the F2 progeny of a laboratory cross between wild *Ae. aegypti* founders from Thailand35 (Extended Data Fig. 10a). For this population, 197 F2 females had previously been scored for DENV vector competence against four different DENV isolates (two isolates from serotype 1 and two from serotype 3)35. The newly developed linkage map included a total of 255 RAD markers (Fig. 5a) with perfect concordance between genetic distances in centiMorgans (cM) and AaegL5 physical coordinates in Mb (Fig. 5a, c, d). We detected two significant QTLs on chromosome 2 that underlie the likelihood of DENV dissemination from the midgut (that is, systemic infection), an important component of DENV vector competence36. One QTL was associated with a generalist effect across DENV serotypes and isolates, whereas the other was associated with an isolate-specific effect (Fig. 5b, c). QTL mapping powered by AaegL5 will make it possible to understand the genetic basis of *Ae. aegypti* vector competence for arboviruses.

Pyrethroid insecticides are used to combat mosquitoes, including *Ae. aegypti*, and emerging resistance to these compounds is a global problem37. Understanding the mechanisms that underlie insecticide targets and resistance in different mosquito populations is critical to combating arboviral pathogens. Many insecticides act on ion channels, and we curated members of the Cys-loop ligand-gated ion channel (Cys-loop LGIC) superfamily in AaegL5. We found 22 subunit-encoding Cys-loop LGICs (Extended Data Fig. 10d and Supplementary Data 24), of which 14 encode nicotinic acetylcholine receptor (nAChR) subunits. nAChRs consist of a core group of subunit-encoding genes (α1–α8 and β1) that are highly conserved between insect species, and at least one divergent subunit38. Whereas *D. melanogaster* possesses only one divergent nAChR subunit, *Ae. aegypti* has five. We found that agricultural and veterinary insecticides impaired the motility of *Ae. aegypti* larvae (Extended Data Fig. 10c), suggesting that these Cys-loop LGIC-targeting compounds have potential as mosquito larvicides. The improved annotation presented here provides a valuable resource for investigating insecticide efficacy.

To demonstrate how a chromosome-scale genome assembly informs genetic mechanisms of insecticide resistance, we performed a genome-wide population genetic screen for SNPs correlating with resistance to deltamethrin in *Ae. aegypti* collected in Yucatán, Mexico, where pyrethroid-resistant and -susceptible populations co-exist (Fig. 5e). We uncovered an association with non-synonymous changes to three amino acid residues of the voltage-gated sodium channel VGSC, a known target of pyrethroids (Fig. 5f). The gene model for VGSC, a complex locus spanning nearly 500 kb in AaegL5, was incomplete and highly fragmented in AaegL3. SNPs in this region have a lower expected heterozygosity (\(H_{\text{exp}}\)) in the resistant compared to the susceptible population, suggesting that they are part of a selective sweep for the resistance phenotype surrounding VGSC (Fig. 5g). Accurately associating SNPs with phenotypes requires a fully assembled genome, and we expect that AaegL5 will be critical to understanding the evolution of insecticide resistance and other important traits.

**Summary**

The high-quality genome assembly and annotation described here will enable major advances in mosquito biology, and has already allowed us to carry out a number of experiments that were previously impossible. The highly contiguous AaegL5 genome permitted high-resolution genome-wide analysis of genetic variation and the mapping of loci for DENV vector competence and insecticide resistance. A new appreciation of copy number variation in insecticide-detoxifying GSTe genes and a more complete accounting of Cys-loop LGICs will catalyse the search for new resistance-breaking insecticides. A doubling in the known number of chemosensory ionotropic receptors provides opportunities to link odorants and tastants on human skin to mosquito attraction, a key first step in the development of novel mosquito repellents. ‘Sterile Insect Technique’ and ‘Incompatible Insect Technique’ show great promise to suppress mosquito populations39, but these population suppression methods require that only males are released. A strategy that connects a gene for male determination to a gene drive construct has been proposed to effectively bias the population towards males over multiple generations38, and improved understanding of M locus evolution and the function of its genetic content should facilitate genetic control of mosquitoes that infect many hundreds of millions of people with arboviruses every year.

**Online content**

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0692-z.

Received: 28 December 2017; Accepted: 5 October 2018; Published online 14 November 2018.
19. Turner, J. et al. The sequence of a male-specific genome region containing the sex determination switch in Aedes aegypti. Parasit. Vectors 11, 549 (2018).

20. Hall, A. B. et al. Six novel Y chromosome genes in Anopheles mosquitoes discovered by independently sequencing males and females. BMC Genomics 14, 273 (2013).

21. Fontaine, A. et al. Extensive genetic differentiation between homomorphic sex chromosomes in the mosquito vector, Aedes aegypti. Genome Biol. Evol. 9, 2322–2335 (2017).

22. Junge, P. et al. Assembly of the disease vector Aedes aegypti onto a genetic linkage map allows mapping of genes affecting disease transmission, PLoS Negl. Trop. Dis. 8, e2652 (2014).

23. Charlesworth, D., Charlesworth, B. & Marais, G. Stasis in the evolution of homothetic sex chromosomes. Heredity 95, 118–128 (2005).

24. Riehle, M. M. et al. The Anopheles gambiae 2La chromosome inversion is associated with susceptibility to Plasmodium falciparum in Africa. eLife 6, e25813 (2017).

25. Lewis, E. A. Gene complex controlling segmentation in Drosophila. Nature 276, 565–570 (1978).

26. Duboule, D. The rise and fall of box genes. Development 134, 2549–2560 (2007).

27. Negre, B., Ranz, J. M., Casals, F., Cáceres, M. & Ruiz, A. A new split of the African malaria vector Anopheles gambiae. Nature 430, 158–162 (2004).

28. Bass, C. & Field, L. M. Gene amplification and insecticide resistance. Insect Mol. Biol. 5, 147–154 (1996).

29. Ortelli, F., Rossiter, L. C., Vontas, J., Ranson, H. & Hemingway, J. Heterologous gene drive. Adv. Exp. Med. Biol. 832, 365–392 (2015).

30. Juneja, P. et al. Assembly of the genome of the disease vector Anopheles gambiae. Nature 505, 179–185 (2014).

31. Lumjuan, N. et al. The role of the Aedes aegypti Epsilon glutathione transferases in conferring resistance to DDT and pyrethroid insecticides. Insect Biochem. Mol. Biol. 41, 203–209 (2011).

32. Anopheles gambiae 1000 Genomes Consortium. Genetic diversity of the African malaria vector Anopheles gambiae. Nature 552, 96–100 (2017).

33. Begun, D. J. & Aquadro, C. F. Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. Nature 356, 519–520 (1992).

34. Evans, B. et al. A multipurpose, high-throughput single-nucleotide polymorphism chip for the dengue and yellow fever mosquito, Aedes aegypti. G3 (Bethesda) 5, 711–718 (2015).

35. Fasini, T. et al. Genetic mapping of specific interactions between D. melanogaster labial and miRNA-128. J. Biol. 12, 1–10 (2013).

36. Black, W. C. V. et al. Flavivirus susceptibility in Aedes aegypti. J. Med. Virol. 34, 489–496 (1991).

37. Moyes, C. L. et al. Contemporary status of insecticide resistance in the major insecticide-resistance locus from the malaria vector Anopheles gambiae. Pest Manag. Sci. 68, 583–588 (2012).

38. Negre, B., Ranz, J. M., Casals, F., Cáceres, M. & Ruiz, A. A new split of the African malaria vector Anopheles gambiae. Nature 430, 158–162 (2004).

39. Lewis, E. B. A gene complex controlling segmentation in Drosophila. Nature 276, 565–570 (1978).

40. Evans, B. et al. A multipurpose, high-throughput single-nucleotide polymorphism chip for the dengue and yellow fever mosquito, Aedes aegypti. G3 (Bethesda) 5, 711–718 (2015).

41. Fasini, T. et al. Genetic mapping of specific interactions between Aedes aegypti mosquito and dengue virus. PLoS Genet. 9, e1003621 (2013).

42. Black, W. C. V. et al. Flavivirus susceptibility in Aedes aegypti. Arch. Med. Res. 33, 379–388 (2002).

43. Moyes, C. L. et al. Contemporary status of insecticide resistance in the major Aedes vectors of arboviruses infecting humans. PLoS Negl. Trop. Dis. 11, e0005625 (2017).

44. Jones, A. & Sattelle, D. B. Diversity of insect nicotinic acetylcholine receptor subunits. Arch. Exp. Med. Biol. 683, 25–43 (2010).

45. Alphey, L. Genetic control of mosquitoes. Annu. Rev. Entomol. 59, 205–224 (2014).

46. Adelman, Z. N. & Tu, Z. Control of mosquito-borne infectious disease: sex and gene drive. Trends Parasitol. 32, 219–226 (2016).

Acknowledgements We thank R. Andino, S. Emrich and D. Lawson (Vectorbase); A. A. James, M. Kountouri, C. Nusbauer, D. Severson, N. Whiteman; T. Dickinson, M. Hartley and B. Rice (Dovetail Genomics) for early participation in the AGWG; C. Bargmann, D. Botstein, E. Jarvis and E. Lander for encouragement and facilitation. N. Keivanfar, D. Jaffe and D. M. Church (10X Genomics) prepared DNA for structural-variant analysis. We thank A. Harmon of the New York Times and acknowledge generous pro claims in published maps and institutional affiliations.

Correspondence and requests for materials should be addressed to B.J.M. Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise. In this case the material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0692-z.

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0692-z.

Reprints and permissions information is available at http://www.nature.com/reprints.

© 2018 Springer Nature Limited. All rights reserved.
METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Ethics information. The participation of one human subject in blood-feeding mosquitoes was approved and monitored by The Rockefeller University Institutional Review Board (IRB protocol LVO-0652). This subject gave their written and informed consent to participate.

Mosquito rearing and DNA preparation. *Ae. aegypti* eggs from a strain labelled 'LVP_ib12' were supplied by M.V.S. from a colony maintained at Virginia Tech. We performed a single pair cross between a male and female individual to generate material for Hi-C, Bionano optical mapping, flow cytometry, SNP-Chip analysis of strain variance, paired-end Illumina sequencing and to establish a colony (LVP_AGWG). Rearing was performed as previously described\(^4,13\) and all animals were offered a human arm as a blood source.

SNP analysis of mosquito strains. Data were generated as described\(^5,14\), and PCA was performed using LEA 2.0 available for R v.3.4.0\(^{41,42}\). The following strains were used: *Ae. aegypti LVP_AGWG* (samples from the laboratory strain used for the AeagL5 genome assembly, reared as described in Extended Data Fig. 1a by a single pair mating in 2016 from a strain labelled LVP_ib12 maintained at Virginia Tech), *Ae. aegypti LVP_ib12* (laboratory strain, LVP_ib12, provided in 2013 by D. Severson, University of Notre Dame), *Ae. aegypti LVP_MR4* (laboratory strain labelled MR4 in 2016 from MR4 at the Centers for Disease Control via BEI Resources catalogue MRA-735), *Ae. aegypti Yaounde*, Cameroon (field specimens collected in 2014 and provided by B. Kamgang), *Ae. aegypti Rockefeller* (laboratory strain provided in 2016 by G. Dimopoulos, Johns Hopkins Bloomberg School of Public Health), *Ae. aegypti Key West*, Florida (field specimens collected in 2016 and provided by W. Tabachnick). Strains used for the linkage disequilibrium data presented in Extended Data Fig. 9c, d were: *Ae. aegypti from Amacuzac, Morelos, Mexico* (field specimens collected in 2016 and provided by C. Gonzalez Acosta) and *Ae. aegypti* from La Lope National park forest, Gabon (field specimens collected and provided by S. Xia).

Flow cytometry. Genome size was estimated by flow cytometry as described\(^43\), except that the propidium iodide was added at a concentration of 25 μg ml\(^{-1}\), not 50 μg ml\(^{-1}\), and samples were stained in the cold and dark for 24 h to allow the stain to fully saturate the sample. In brief, nuclei were isolated by placing a single frozen head of an adult sample along with a single frozen head of an adult *Drosophila virilis* female standard from a strain with 1C = 328 Mb into 1 ml of Galbraith buffer (4.26 g MgCl\(_2\), 8.84 g sodium citrate, 4.2 g 3-[N-morpholino] propane sulfonic acid (MOPS), 1 ml Triton X-100 and 1 mg boiled RNase A in 1 l of ddH\(_2\)O, adjusted to pH 7.2 with HCl and filtered through a 0.22-μm filter\(^43\) and grinding with 15 strokes of the A pestle at a rate of 3 strokes per 2 s. The resultant ground mixture was filtered through a 60-μm nylon filter (Spectrum Labs). Samples were stained with 25 μg of propidium iodide and held in the cold (4 °C) and dark for 24 h at which time the relative red fluorescence of the 2C nuclei of the standard and sample were determined using a Beckman Coulter CytoFlex flow cytometer with excitation at 488 nm. At least 2,000 nuclei were scored under each 2C peak and all scored peaks had a coefficient of variation of 2.5 or less\(^4,43\). Average channel numbers for sample and standard 2C peaks were scored using CytExpert software and all scored peaks had a coefficient of variation of 2.5 or less\(^4,44\). Average channel numbers for sample and standard 2C peaks were scored using CytExpert software and all scored peaks had a coefficient of variation of 2.5 or less\(^4,44\). Average channel numbers for sample and standard 2C peaks were scored using CytExpert software and all scored peaks had a coefficient of variation of 2.5 or less\(^4,44\). Average channel numbers for sample and standard 2C peaks were scored using CytExpert software and all scored peaks had a coefficient of variation of 2.5 or less\(^4,44\). Average channel numbers for sample and standard 2C peaks were scored using CytExpert software and all scored peaks had a coefficient of variation of 2.5 or less\(^4,44\). Average channel numbers for sample and standard 2C peaks were scored using CytExpert software and all scored peaks had a coefficient of variation of 2.5 or less\(^4,44\). Average channel numbers for sample and standard 2C peaks were scored using CytExpert software and all scored peaks had a coefficient of variation of 2.5 or less\(^4,44\). Average channel numbers for sample and standard 2C peaks were scored using CytExpert software and all scored peaks had a coefficient of variation of 2.5 or less\(^4,44\). Average channel numbers for sample and standard 2C peaks were scored using CytExpert software and all scored peaks had a coefficient of variation of 2.5 or less\(^4,44\). Average channel numbers for sample and standard 2C peaks were scored using CytExpert software and all scored peaks had a coefficient of variation of 2.5 or less\(^4,44\). Average channel numbers for sample and standard 2C peaks were scored using CytExpert software and all scored peaks had a coefficient of variation of 2.5 or less\(^4,44\). Average channel numbers for sample and standard 2C peaks were scored using CytExpert software and all scored peaks had a coefficient of variation of 2.5 or less\(^4,44\). Average channel numbers for sample and standard 2C peaks were scored using CytExpert software and all scored peaks had a coefficient of variation of 2.5 or less\(^4,44\). Average channel numbers for sample and standard 2C peaks were scored using CytExpert software and all scored peaks had a coefficient of variation of 2.5 or less\(^4,44\). Average channel numbers for sample and standard 2C peaks were scored using CytExpert software and all scored peaks had a coefficient of variation of 2.5 or less\(^4,44\). Average channel numbers for sample and standard 2C peaks were scored using CytExpert software and all scored peaks had a coefficient of variation of 2.5 or less\(^4,44\).

Hi-C sample preparation and analysis. For technical details, see Methods and Supplementary Discussion. In brief, both primary contigs and haplotigs were used as input. We performed quality control, manual polishing and validation of the scaffolding results using AssemblyTools\(^45\). This produced three chromosome-length scaffolds. Notably, the contig N50 decreased slightly, to 929,392 bp, because of the splitting of misjoined contigs.

Hi-C alternative haplotype merging. Examination of the initial chromosome-length scaffolds using AssemblyTools\(^45\) revealed that extensive undercollapsed heterozygosity was present. In fact, most genomic intervals were repeated, with variations, on two or more unmerged contigs. This suggested that the levels of undercollapsed heterozygosity were unusually high, and that the true genome length was far shorter than either the total length of the Pacific Biosciences contigs (2,047 Mb), or the initial chromosome-length scaffolds (1,973 Mb). Possible factors that could have contributed to the unusually high rate of undercollapsed heterozygosity seen in the FALCON-Unzip Pacific Biosciences contigs relative to prior contig sets for *Ae. aegypti* generated using Sanger sequencing (AeagL3\(^2\)), include high heterozygosity levels in the species and incomplete inbreeding in the samples that we sequenced. The merge algorithm described previously\(^45\) detects and merges draft contigs that overlap one another owing to undercollapsed heterozygosity. Because undercollapsed heterozygosity does not affect most loci in a typical draft assembly, the default parameters are relatively stringent. We adopted more permissive parameters for AeagL5 to accommodate the exceptionally high levels of undercollapsed heterozygosity, but found that the results would occasionally merge contigs that did not overlap. To avoid these false positives, we developed a procedure to manually identify and ‘whitelist’ regions of the genome containing no overlap, based on both primary contigs and haplotigs (Extended Data Fig. 1c, Supplementary Methods and Supplementary Discussion). We then reran the merge step, using the whitelist as an additional input. Finally, we performed quality control of the results using AssemblyTools\(^45\), which confirmed the absence of the undercollapsed heterozygosity.
heterozygosity that we had previously observed. The resulting assembly contained three chromosome-length scaffolds (310 Mb, 473 Mb and 409 Mb), which spanned 94% of the merged sequence length. The assembly also contained 2,364 small scaffolds, which spanned the remaining 6% (Table 1). These lengths were consistent with the results of flow cytometry and the lengths obtained in prior assemblies. Notably, the merging of overlapping contigs using the above procedure frequently eliminated gaps, and thus greatly increased the contig N50, from 929,392 to 4,997,917 bp.

Final gap-filling and polishing. Scaffolded assembly polishing. Following scaffolding and de-duplication, we performed a final round of arrow polishing. PBJelly²⁰ from PBSuite version 15.8.24 was used for gapfilling of the de-duplicated HiC assembly (see ‘Protocol.xml’ in Supplementary Methods and Supplementary Discussion). After PBJelly, the liftover file was used to translate the renamed scaffolds to their original identifiers. For this final polishing step (run with SMRT Link v3.1 resequencing), the reference sequence included the scaffolded, gap-filled reference, as well as all contigs and contig fragments not included in the final scaffolds (https://github.com/skimgan/AaegL5_FinalPolish). This reduces the likelihood that reads map to the wrong haplotype, by providing both haplotypes as targets for read mapping. For submission to NCBI, two scaffolds identified as mitochondrial in origin were removed (see below), and all remaining gaps on scaffolds were standardized to a length of 100 Ns to indicate a gap of unknown size. The assembly quality value was estimated using independent Illumina sequencing data from a single individual male pupa (library H2NJHAXDY_1/2). Reads were aligned with BWA-MEM v0.7.12-r103993. FreeBayes v1.1.0-50-g61527c5-dirty⁴ was used to call SNPs and short indels with the parameters -C 2 -O 0 -Q 20 -o 0.10 -E X/a-p 2 -P 0.01. Any SNP and short indels showing heterozygosity (for example, 0/1 genotypes) were excluded. The quality value was estimated at 34.75 using the PHRED formula with SNPs as the numerator (597,798) and number of bases with at least three-coverage fold as the denominator, including alternate alleles (1,782,885,792).

Identification of mitochondrial contigs. During the submission process for this genome, two contigs were identified as mitochondrial in origin and were removed from the genomic assembly, manually circularized, and submitted separately. The mitochondrial genome is available as GenBank accession number MF194022.1, ReSeq accession number NC_035159.1.

Bionano optical mapping. HMW DNA extraction. HMW DNA extraction was performed using the Bionano Animal Tissue DNA Isolation Kit (RE-013-10), with a few protocol modifications. A single-cell suspension was made as follows. After another 2 min fixation, the tissue was finely homogenized by blending for 2 s, using a rotor–stator tissue homogenizer (TissueRuptor, Qiagen 9001271). After another 2 min fixation, the tissue was finely homogenized by running the rotor–stator for 10 s. Subsequently, the homogenate was filtered with a 100-μm nylon filter, fixed with ethanol for 30 min on ice, spun down, and the DNA was extracted using a QIAGEN Tissue RNeasy kit (Qiagen 74134). The DNA was resuspended in 50 μl EB and incubated for 5 min at 70 °C. After restriction with Nt.BspQI or Nb.BssSI, the DNA was purified using a Sepharose 4B spin column (GE Healthcare). The DNA was resuspended in 5 μl DNAse-free water and used for library preparation.

DNA labeling. DNA was labelled according to commercial protocols using the DNA Labelling Kit NLRs (RE-012-10, Bionano Genomics). Specifically, 300 μg of purified genomic DNA was nicked with 7 U nicking endonuclease Nt.BspQI (New England BioLabs, NEB) at 37 °C for 2 h in NEBuffer3. The nicked DNA was labelled with a fluorescent-DUTP nucleotide analogue using Tag polymerase (NEB) for 1 h at 72 °C. After labelling, the nicks were ligated with T4 ligase (NEB) in the presence of DNA polymerase. The backbone of fluorescently labelled DNA was counter-stained with YOYO-1 (Invitrogen).

Data collection. The DNA was loaded onto the nanochannel array of Bionano Genomics IrisChip by electrophoresis of DNA. Linearized DNA molecules were then imaged automatically followed by repeated cycles of DNA loading using the Bionano Genomics Iris system. The DNA-molecule backbones (YOYO-1 stained) and locations of fluorescent labels along each molecule were detected using the in-house-generated software package, IrisView. The set of labels locations of each DNA molecule defines an individual single-molecule map. After filtering data using normal parameters (molecule reads with length greater than 150 kb, a minimum of 8 labels and standard filters for label and backbone signals), a total of 299,745 GCF_002204515.2 scaffolds with 76,359 Gb of data were collected from Nt.BspQI and Nb.BssSI samples, respectively.

De novo genome map assembly. De novo assembly was performed with non-haplotype aware settings (optArguments_nonhaplotype_noES_iris.xml) and pre-release version of Bionano Solve v3.1 (Pipeline version 6703 and ReAligner version 6851). On the basis of the overlap–layout–Consensus paradigm, pairwise comparisons of all DNA molecules were performed to create an overlap graph, which was then used to create the initial consensus genome maps. By realigning molecules to the genome maps (RealignB P = 10 − 10−11) and by using only the best match molecules, a refinement step was performed to refine the label positions on the genome maps and to remove chimeric joins. Next, during an iterative process, the software creates new genome maps (extension, P = 10 − 10−11), and extended the maps based on the molecules aligning past the map ends. Overlapping genome maps were then merged using a merge P-value cut-off of 10 P = 10 × 10−15. These extension and merge steps were repeated five times before a final refinement was applied to ‘finish’ all genome maps (refine final, P = 10 × 10−11). Two genome maps de novo assemblies, one with nickase Nt.BspQI and the other with nickase Nb.BssSI, were constructed. Alignments between the constructed de novo genome assemblies and the L5 assembly were performed using a dynamic programming approach with a scoring function and a P-value cut-off of P = 10 × 10−12.

Transposable element identification. Identification of transposable elements. We first identified known transposable elements using RepeatMasker (version 3.2.6)⁴³ against the mosquito TEfam (https://tefam.biochem.vt.edu/tefam/; data downloaded July 2017), a manually curated mosquito transposable-elements database. We then ran RepeatMasker using the TEfam database and Repbase transposable-elements library (version 10.05). RepeatMasker was set to default parameters with the -s (slow search) flag and NCBI/RMas blast program (v.2.2.28). De novo repeat family identification. We searched for repeat families and consen sus repeats within the de novo repeat prediction tool RepeatModeler (version 1.0.8)⁴⁵ using default parameters with RECON (version 1.07) and RepeatScout (1.0.5) as core programs. Consensus sequences were generated and classified for each repeat family. Then RepeatMasker was run on the genome sequences, using the RepeatModeler consensus sequence as the library.

Tandem repeats. We also predicted tandem repeats in the whole genome and in the repeatmasked genome using Tandem Repeat Finder²⁸. Long tandem copies were identified using the ‘Match=2, Mismatch=7, Delta=7, PM=80, PI=10, Minscore=50 MaxPeriod=500’ parameters. Simple repeats, satellites and low complexity repeats were found using ‘Match=2, Mismatch=7, Delta=7, PM=80, PI=10, Minscore=50, and MaxPeriod=12’ parameters.

A file representing the coordinates of all identified repeat and transposable-element structures in AaegL5 can be found at https://github.com/VoshallLab/AGWG-AaegL5.

Generation of ReSeq gene set annotation. The AaegL5 assembly was deposited at NCBI in June 2017 and annotated using the NCBI ReSeq Eukaryotic gene annotation pipeline²⁴. Evidence to support the gene predictions came from over 9 billion Illumina RNA-seq reads, 67,000 Pacific Biosciences IsoSeq transcripts, 300,000 expression and EST sequences, and the assembled genome of Aedes aegypti and other insects. Annotation Release 101 was made public in July 2017, and specific gene families were subjected to manual annotation and curation. Detailed descriptions of the manual annotation and curation of multigene families (Hox genes, proteases, opsins and biogenic amine receptors, chemosensory receptors and LGICs) can be found in the Supplementary Methods and Supplementary Discussion. See also https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Aedes_aegypti/101/.

Alignment of RNA-seq data to AaegL5 and quantification of gene expression. Published RNA-seq reads andunaligned RNA-seq reads from tissue-specific libraries produced by Verily Life Sciences were mapped to the ReSeq assembly GCF_002204515.2_AaegL5.0 with STAR aligner (v.2.5.3a)⁵⁸ using the two-pass approach. Reads were first aligned in the absence of gene annotations using the following parameters: −outFilterType BySlopout; −outIntronMax 1000000; −outAlignMatesGapMax 1000000; −outFilterMismatchNmax 999; −outFilterMismatchNovelReads Max 0.04; −clip3pNbases 1; −outSAMstrandField intronMotif; −outSAMattrHstart 0; −outFilterMultimapNmax 20; −outSAMattributes NH AS NM MD; −outSAMattrRGline; −outSAMtype BAM SortedByCoordinate. Splice junctions identified during the first pass mapping of individual libraries were combined and supplied to STAR using the −sdbFileChrStartEnd option for the second pass. Reads mapping to gene models defined by the NCBI annotation pipeline (GCF_002204515.2_AaegL5.0_genomic.gff) were quantified using featureCounts⁵⁹ with default parameters. Count data were transformed to transcripts per million values using a custom Perl script. Details on library, alignment statistics and gene expression estimates (expressed in transcripts per million) are provided as Supplementary Data 4–8.

Identification of ‘collapsed’ and ‘merged’ gene models from AaegL3.5 to AaegL5.0. Vector Base annotation AaegL3.5 was compared to NCBI Ae. aegypti annotation step 101 were aligned molecules to identify AaegL3.5 as NCBI as part of NCBI’s eukaryotic genome annotation pipeline. First, assembly–assembly alignments were generated for AaegL3 (GCA_000004105.3) × AaegL5.0 (GCF_002204515.2) as part of NCBI’s Remap command remerging, copyright © 2018 Springer Nature Limited. All rights reserved.
Alignment of chromosome assemblies and Bionano scaffolds. The boundaries of the M locus were identified by comparing the current AaegL5 assembly and the AaegL4 assembly using a program called LASTN (data not shown). To overcome the challenges of repetitive hits, both AaegL5 and AaegL4 assemblies were twice repeat-masked against a combined repeat library of TFeam-annotatable transposable elements (https://tfam.biochem.vt.edu/tfam/1) and a RepeatModeler output from the Anopheles 16 Genomes project. The masked sequences were then compared using BLASTn, and then we set a filter for downstream analysis to include only alignment with ≥98% identity over 1,000 bp. After the identification of the approximate boundaries of the M locus (and m locus), which contains two male-specific genes, myo-sex and Nix, we zoomed in by performing the same analysis on regions of the M locus and m locus plus 2 Mb flanking regions without repeatmasking. In this and subsequent analyses, only alignment with ≥98% identity over 500 bp were included. Consequently, approximate coordinates of the M locus and m locus were obtained on chromosome 1 of the AaegL5 and AaegL4 assemblies, respectively. Super-scaffold-63 in the Bionano optical map assembly was identified by BLASTN which spans the entire M locus and extends beyond its two borders.

Chromosome quotient analysis. The chromosome quotient (CQ) was calculated for each 1,000-bp window across all AaegL5 chromosomes. To calculate the CQ, Illumina reads were generated from two paired sibling female and male sequencing libraries. To generate libraries for CQ analysis, we performed two separate crosses of a single VLP_AGGWG male to 10 virgin females. Eggs from this cross were hatched, and virgin male and female adults collected within 12 h of eclosion to verify their non-mated status. We generated genomic DNA from five males and five females from each of these crosses. Sheared genomic DNA was used to generate libraries for Illumina sequencing with the Illumina TrueSeq Nano kit and sequencing performed on one lane of 150-bp paired-end sequencing on an Illumina NextSeq 500 in high-output mode.

For a given sequence S of a 1,000-bp window, CQ(S) = Fm/Mf, where Fm is the number of female Illumina reads aligned to S, and Mf is the number of male Illumina reads aligned to S. Normalization was not necessary for these datasets because the mean and median CQs of the autosomes (chromosomes 2 and 3) are all near 1. A CQ value lower than the 0.05 indicates that the sequences within the corresponding 1,000-bp window had at least 20-fold more hits to the male Illumina dataset than the female Illumina dataset. To calculate each CQ value, reads were compressed and only alignment with ≥98% identity over 1,000 bp were included. Because many of the reads were compressed by RepeatMasker, to ensure that each CQ value represents a meaningful data point obtained with sufficient alignment, only sequences with more than 20 male hits were included in the calculation. The CQ values were then plotted against the chromosome location of the 1,000-bp window (Fig. 3d). Under these conditions, there is not a single 1,000-bp fragment on chromosomes 2 and 3 that showed CQ = 0.05 or lower.

Chromosome FISH. Slides of mitotic chromosomes were prepared from imaginal discs of fourth instar larvae following published protocols (14,15). BAC clones were obtained from the University of Liverpool (16) or from a previously described BAC library (17). BACs were plated on agar plates (Thermo Fisher) and a single bacterial colony was used to grow an overnight bacterial culture in LB broth plates (Thermo Fisher) at 37 °C. DNA from the BACs was extracted using Sigma PhasePrep TM BAC DNA Kit (Sigma-Aldrich, NA-0100). BAC DNA for hybridization was labelled by nick translation with Cy3-, Cy5-, dUTP (Enzo Life Sciences) or Fluorescein 12-dUTP (Thermo Fisher). Chromosomes were counterstained with DAPI in Prolong Gold Antifade (Thermo Fisher). Slides were analysed using a Zeiss LSM 880 Laser Scanning Microscope at 1,000× magnification. We note that localization of the M locus to 1p11 is supported by both FISH and genomic analyses, but is contrary to a previously published placement at 1q21 (7).
annotated as paralogues in AaegL3.4. On the basis of a BLASTp search against the AaegL3.4 protein set, the two putative P450 genes not annotated as such in AaegL5.0 (encoding proteins XP_001649103.2 and XP_021694388.1) appear to be incorrect gene models in the AaegL5.0 annotation, which should in fact be two adjacent genes (CYP9J21 and CYP9J22 for XP_001649103.2; CYP9P12 and CYP9P13 for XP_021694388.1). Compared to AaeL3.4, which predicts a single copy each of GSTe2, GSTe5 and GSTe7, the NCBI annotation of AaegL5.0 predicts three copies each of GSTe2 and GSTe5, and four copies of GSTe7, arranged in a repeat structure. BLASTn searches revealed one additional copy each of GSTe2 and GSTe5 in the third duplicated unit. Both contain premature termination codons owing to frameshifts, but these could be owing to uncorrected errors in the assembly. Error correction of all duplicated units was not possible owing to the inability to unequivocally align reads to units not ‘anchored’ to adjacent single-copy sequence.

To validate these tandem duplications, two lanes of Illumina whole-genome sequence data from a single pupa of the LVP_AGWG strain (H2NJHADXY) were aligned to a hard-masked version of the AaeL3.4 reference genome using Bowtie2 v.2.2.4, with ‘–very-fast-local’ alignment parameters, an expected fragment size between 0 and 1,500 bp and relative orientation ‘forward–reverse’ (‘–1 0 -X 1500 –fr’). Aligned reads with a mapping quality less than 10 were removed using Samtools84, ‘featureCounts’, part of the ‘Subread’ v.1.5.0-p2 package84, was used to assign read pairs or reads (‘tags’) aligned to either DNA strand (‘s’-0) and overlapping the coding regions of a gene by at least 100 bp (‘-CDS–minOverlap 100’) to genes as an estimate of representation in the genome. Gene-wise tag counts were normalized by calculating the fragments per kilobase of gene length per million mapped reads (FPKM), using the following equation: (tag count/gene length in kb)/(sum of tag counts for all genes in genome/1,000,000).

Median FPKM for all genes in the genome was calculated (48.22), allowing FPKM of GSTe genes to be expressed relative to this. To examine strain differences in coverage at this cluster, we repeated this analysis for the four laboratory colonies analysed in Extended Data Fig. 9a, b. Median FPKM values across all genes ranged from 47.68 to 48.46 and gene-wise FPKM values normalized relative to these medians are plotted in Fig. 4d.

To visualize the sequence identity of the repeat structure in the GSTe cluster (Fig. 4b), we extracted the region spanning the cluster from AaeL3.4 chromosome 2 (S51,597,324–351,719,186 bp) and performed alignment of Pacific Biosciences reads using minimap2 v.2.1.1 (minimap2 -DP K7 -w1 -r2 -r2o -g1o -mi1.5_gst -fa L_gst.fa)85 and visualized these alignments using D-GENIES v.1.2.0.086 with minimum identity set to 0.15 and ‘Strong Precision’ enabled. To validate this repeat structure, we aligned two de novo optical maps created by Bionano using linearized DNA labelled with Nt. BspQI or Nb.BasSI. Single molecules from both maps span the entire region and the predicted restriction pattern provides support for the repeat structure as presented in AaeL3.4 (Fig. 4c).



## QTL mapping of DENV vector competence

In theory, a good-quality genome assembly is important for QTL mapping and offered by AaeL3.4 genome assembly, because it relies on a linkage map that can be generated de novo from empirical recombination fractions. This typically involves three steps: (i) marker selection based on the Mendelian segregation ratios, (ii) marker assignment to linkage groups and (iii) marker ordering within each linkage group. However, if a high-quality reference genome assembly is available, the physical position of each marker can be determined and this prior information greatly facilitates steps (ii) and (iii), as shown below.

To demonstrate the improvement enabled by our new genome, we generated two linkage maps using the same Illumina sequence data that were aligned either to AaeL3.4 or AaeL5.0 genome assemblies. Although the initial number of markers was 616 in both cases, the final linkage map was 3.3-fold denser with AaeL3.5 than with AaeL3.4, as shown in Extended Data Fig. 10b. The difference in marker density between the two linkage maps is because many markers were filtered out from the AaeL3.4 data. Because the AaeL3.4 assembly is highly fragmented (>4,700 scaffolds), the position of each marker within the linkage groups is primarily determined from the recombination fractions. This ordering step is performed by creating a backbone with a subset of informative markers using a two-point algorithm, followed by the positioning of the remaining markers one at a time using a multi-point method. Only markers that are unambiguously positioned are kept in the final linkage map for QTL mapping. We note that AaeL4.4, which de-duplicated and scaffolded AaeL3.4 onto chromosomes, would probably yield a similar improvement in mapping resolution.

Another complication arises for the chromosome 1 in Ae. aegypti, because recombination is strongly reduced in the region containing the sex-determining M locus. This leads to the severely biased segregation ratios for markers anchored to this linkage group. In our F2 intercross design, the fully sex-linked markers lacked the F0 paternal genotype in F2 females and segregated in the same manner as a backbone linkage analysis. This backbone linkage analysis is readable enough to deal with a chromosome that behaves like a mixture of intercross and backcross designs. Therefore, AaeL3.4-guided linkage analysis and QTL mapping for chromosome 1 were restricted to the fully sex-linked region based on a backcross design. By contrast, AaeL5.5-guided linkage analysis and QTL mapping for chromosome 1 made use of all markers regardless of their segregation ratios, allowing chromosome-wide coverage. As mentioned in the present manuscript, the only caveat is that our analytical procedure assumes autosomal Mendelian proportions, which may have resulted in over- or underestimation of linkage distances between markers on chromosome 1. The linkage map was iteratively refined by checking for misplaced markers based on visual inspection of the LOD/RF matrix.

Ultimately, AaeL5.5 has a markedly improved QTL mapping resolution over AaeL3.4. For instance, we mapped the same QTL underlying systemic DENV dissemination at the extremity of chromosome 2 with both AaeL3.4 and AaeL5.5. The 1.5 LOD support interval was much larger for the AaeL3.4-guided linkage map (0–50 cM, 74% of the linkage group) than for the AaeL5.5-guided linkage map (0–17 cM, 9% of the linkage group). We present this analysis in Extended Data Fig. 10b.

## Mosquito crosses

A large F2 intercross was created from a single mating pair of field-collected F0 founders. Wild mosquito eggs were collected in Kamphaeng Phet Province, Thailand in February 2011 as previously described87. In brief, F0 eggs were allowed to hatch in filtered tap water and the larvae were reared until the pupae emerged in individual vials. Ae. aegypti adults were identified by visual inspection and maintained in an insectary under controlled conditions (28 ± 1°C, 75 ± 5% relative humidity and 12:12-h light:dark cycle) with access to 10% sucrose. The F0 male and female initiating the cross were chosen from different collection sites to avoid creating a parental pair with siblings from the same wild mother77. Their F0 offspring were allowed to mass-mate and collectively oviposit to produce the F2 progeny (Extended Data Fig. 10a). A total of 197 females of the F2 progeny were used to generate a linkage map and detect QTLs underling vector competence for DENV.

## Vector competence

Four low-passage DENV isolates were used to orally challenge the F2 females as previously described55. In brief, four random groups of females from the F2 progeny were experimentally exposed to two virus isolates belonging to DENV serotype 1 (KDH0026A and KDH0030A) and two virus isolates belonging to DENV serotype 3 (KDH0010A and KDH0014A), respectively. All four virus isolates were derived from human serum specimens collected in 2010 from clinically ill patients who were infected with DENV at the Kamphaeng Phet Provincial Hospital88. Because the viruses were isolated in the laboratory cell culture, informed consent of the patients was not necessary for the present study.

Complete viral genome sequences were previously deposited into GenBank (accession numbers HG316481, HG316582, HG316683, and HG316484)89. Phylogenetic analysis assigned the viruses to known viral lineages that were circulating in south-east Asia in the previous years45. Each isolate was amplified twice in C6/36 (Aedes albopictus) cell lines (maintained at AFRIMS in Bangkok, Thailand; used only to grow virus, not explicitly authenticated or checked for mycoplasma contamination) before vector competence assays. Four- to seven-day-old F2 females were orally challenged with four QTL mapping DENV virus isolates. The titre in the blood meals ranged from 2.0 × 105 to 2.5 × 106 plaque-forming units per ml across all isolates. Fully engorged females were incubated under the conditions described above. Vector competence was scored 14 days after the infectious blood meal according to two conventional phenotypes: (i) midgut infection and (ii) viral dissemination from the midgut. These binary phenotypes were scored based on the presence or absence of infectious particles in body and head homogenates, respectively. Infectious viruses were detected by plaque assay performed in LLC-MK2 (rhesus monkey kidney epithelial) cells as previously described90.

## Genotyping

Mosquito gDNA was extracted using the Nuclisens 96 Tissue Core Kit (Macherey-Nagel). For the F3 male, it was necessary to perform whole-genome amplification using the REPLi-g Mini kit (Qiagen) to obtain a sufficient amount of DNA. F3 parents and females of the F3 progeny were genotyped using a modified version of the original double-digest restriction site-associated DNA (RAD) sequencing protocol59, as previously described91. The final libraries were spiked with 15% PhiX and sequenced on an Illumina NextSeq 500 platform using a 150-cycle paired-end chemistry (Illumina). A previously developed bash script pipeline26 was used to process the raw sequence reads. High-quality reads (Phred scores > 25) trimmed to the 140-bp length were aligned to the AaeL5.5 reference genome (July 2017) using Bowtie v.0.12.6.2. Parameters for the untagged alignment included ≤3 mismatches in the seed, suppression of alignments with >1 best reported alignment under a ‘try-hard’ option. Variant and genotype calling was performed from a catalogue of RAD loci created with the ref_map.pl pipeline in Stacks v.1.190.1–82. Downstream analyses only used high-quality genotypes at informative markers that were homozygous for alternative alleles in the F0 parents (for example, AA in the F0 male and BB in the F0 female), had a sequencing depth ≥10× and were present in ≥60% of the mapping population.
BB) genotypes and 50% for heterozygous (AB) genotypes. Autosomal markers that significantly deviated from these Mendelian segregation ratios \( (P < 0.05) \) were filtered out using a \( \chi^2 \) test. Owing to the presence of a dominant male-determining locus on chromosome 1, fully sex-linked markers on chromosome 1 are expected to segregate in \( F_2 \) females with equal frequencies \( (50\%) \) of heterozygous (AB) and \( F_0 \) maternal (BB) genotypes, because the \( F_0 \) paternal (AA) genotype only occurs in \( F_2 \) males. As previously reported\(^4\), strong deviations from the expected Mendelian segregation ratios were observed for a large proportion of markers assigned to chromosome 1 in the \( F_2 \) progeny. Markers on chromosome 1 were included if they had heterozygous (AB) genotype frequencies inside the 40–60% range and \( F_0 \) maternal (BB) genotype frequencies inside the 5%–65% range. These arbitrary boundaries for marker selection were largely permissive for partially or fully sex-linked markers on chromosome 1. Owing to a lack of linkage analysis methods that deal with sex-linked markers when only one sex is genotyped, the recombination fractions between all pairs of selected markers were estimated using the \( r.f.2p \)ts function with default parameters for all three chromosomes. The \( r.f.2p \)ts function, which implements the expectation–maximization (EM) algorithm, was used to estimate haplotype frequencies and recombination rates between markers\(^1\) under the assumption of autosomal Hardy–Weinberg proportions. Owing to this analytical assumption, the estimates of cM distances could be over- or underestimated for markers on chromosome 1. Markers linked with a LOD score \( >11 \) were assigned to the same linkage group. Linkage groups were assigned to the three distinct \( A.e. \ aegyi \) chromosomes based on the physical coordinates of the AaegL5 assembly. Recombination fractions were converted into genetic distances in cM using the Kosambi mapping function\(^6\). Linkage maps were exported in Excel and each of the four DNA pools following the Low Sample protocol from the Illumina TrueSeq DNA PCR-Free Sample preparation guide (Illumina). Because 65% of the \( A.e. \ aegyi \) genome consists of repetitive DNA, we performed an exome-capture hybridization to enrich for coding sequences using custom SeqCap EZ Developer probes (NimbleGen, Roche). Probes covered protein–coding sequences (not including untranslated regions) in the AaegL1.3 genebuild using previously specified exonic coordinates\(^7\). In total, 26.7 Mb of the genome \( (2\%) \) was targeted for enrichment. TruSeq libraries were hybridized to the probes using the \( xGenLockDown \) recommendations (Integrated DNA Technologies). The targeted DNA was eluted and amplified \( (10–15 \) cycles) before being sequenced on one flow cell of a 100-bp HiSeq Rapid duo-paired-end sequencing run (Illumina) performed by the Centers for Disease Control (Atlanta, GA, USA).

The raw sequence files (FASTQ) for each pair-ended dGNA library were aligned to a custom reference physical map generated from the assembly AaegL5. Nucleotide counts were loaded into a contingency table with four rows corresponding to ‘Alive Rep1’, ‘Alive Rep2’, ‘Dead Rep1’ and ‘Dead Rep2’. The numbers of columns \( (c) \) corresponded to the number of alternative nucleotides at a SNP locus. The maximum value for \( c \) is 6, corresponding to A, C, G, T, insert or deletion. The columns \( (c) \) were subjected to yoking the alleles of the \( A.e. \ aegyi \) genome to determine whether there are significant \( (P < 0.05) \) differences between (1) Alive replicates, (2) Dead replicates and (3) Alive versus Dead. If analysis \( (a) \) or \( (b) \) was significant, then that SNP locus was discarded. Otherwise the third contingency table consisted of two rows corresponding to Alive \( (\sum \text{replicates 1 and 2}) \) and Dead \( (\text{replicates 1 and 2 summed}) \), and \( c \) columns. The \( \chi^2 \) value from the \( (2 \times c) \) contingency \( \chi^2 \) analysis with \( (c-1) \) degrees of freedom was loaded into Excel to calculate the one-tailed probability of the \( \chi^2 \) distribution probability \( \left( P \right) \). This value was transformed with \( -\log \left( P \right) \). The experiment-wise error rate was then calculated following the method of Benjamini and Hochberg\(^8\) to lower the number of type I errors (false positives).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Code availability. The overview of the Hi-C workflow, as well as modifications to 3D DNA-association and AaegL5, is shared on GitHub at https://github.com/theaidenlab/AGGW-merge. The source code and executable version of Juicebox Assembly Tools are available at https://aiderenlab.org/assembly. Data files and scripts used for the final polishing of scaffolded, gap-filled assembly are available at https://github.com/skingan/AaegL5_FinalPolish.

Mapping insecticide resistance and VGSC. The mosquito population Viva Cauzel from Yucatán State in Southern Mexico (longitude \(-89.7827\), latitude \(20.99827\)), was collected in 2011 through the Universidad Autónoma de Yucatán. We identified up to 25 larval breeding sites from 3–4 city blocks and collected around 1,000 larvae. Larvae were allowed to eclose, and twice a day we aspirated the adults from the cartons, discarding anything other than \( A.e. \ aegyi \). Then, 300–400 \( A.e. \ aegyi \) were released into a 2-foot (61-cm) cubic cage where they were allowed to mate for up to five days with ad libitum access to sucrose, after which they were blood fed to collect eggs for the next generation. Then, 390 adult mosquitoes were phenotyped for deltamethrin resistance. We exposed groups of 50 mosquitoes \( (3–4 \) days old) to \( 3 \mu g \) of deltamethrin-coated bottles for 1 h. After this time, active mosquitoes were transferred to cardboard cups and placed into an incubator \((28 \) °C and 70% humidity) for 4 h; these mosquitoes were referred to as the resistant group. Immobile mosquitoes were transferred to a second cardboard cup. After 4 h, newly recovered mosquitoes were aspirated, frozen and labelled as recovered; these were excluded from the current study. The mosquitoes that were immobilized and remained inactive at 4 h post-treatment were scored as susceptible. DNA was isolated from individual mosquitoes by the salt extraction method\(^9\) and resuspended in 150 µl of TE buffer \((10 \) mM Tris-\(\text{HCl}, 1 \) mM EDTA pH 8.0). We constructed a total of four dGNA libraries. Two groups were pooled from DNA of 25 individual females that survived 1 h of deltamethrin exposure (resistant replicates 1 and 2). The second set of two libraries was obtained by pooling DNA from 25 females that were immobilized and inactive at 4 h post-treatment (susceptible replicates 1 and 2). Before pooling, DNA from each individual mosquito was quantified using the Quant-IT Pico Green kit (Life Technologies, Thermo Fisher Scientific) and around 40 ng from each individual DNA sample \((25 \) individuals per library) was used for a final DNA pool of 1 µg. Pooled DNA was sheared and fragmented by sonication to obtain fragments between 300 and 500 bp (Covaris). We prepared one library from each DNA pool using the Low Sample protocol from the Illumina TrueSeq DNA PCR-Free Sample preparation guide (Illumina). Because 65% of the \( A.e. \ aegyi \) genome consists of repetitive DNA, we performed an exome-capture hybridization to enrich for coding sequences using custom SeqCap EZ Developer probes (NimbleGen, Roche). Probes covered protein–coding sequences (not including untranslated regions) in the AaegL1.3 genebuild using previously specified exonic coordinates\(^7\). In total, 26.7 Mb of the genome \( (2\%) \) was targeted for enrichment. TruSeq libraries were hybridized to the probes using the \( xGenLockDown \) recommendations (Integrated DNA Technologies). The targeted DNA was eluted and amplified \( (10–15 \) cycles) before being sequenced on one flow cell of a 100-bp HiSeq Rapid duo-paired-end sequencing run (Illumina) performed by the Centers for Disease Control (Atlanta, GA, USA).

The raw sequence files (FASTQ) for each pair-ended dGNA library were aligned to a custom reference physical map generated from the assembly AaegL5. Nucleotide counts were loaded into a contingency table with four rows corresponding to ‘Alive Rep1’, ‘Alive Rep2’, ‘Dead Rep1’ and ‘Dead Rep2’. The numbers of columns \( (c) \) corresponded to the number of alternative nucleotides at a SNP locus. The maximum value for \( c \) is 6, corresponding to A, C, G, T, insert or deletion. The columns \( (c) \) were subjected to yoking the alleles of the \( A.e. \ aegyi \) genome to determine whether there are significant \( (P < 0.05) \) differences between (1) Alive replicates, (2) Dead replicates and (3) Alive versus Dead. If analysis \( (a) \) or \( (b) \) was significant, then that SNP locus was discarded. Otherwise the third contingency table consisted of two rows corresponding to Alive \( (\sum \text{replicates 1 and 2}) \) and Dead \( (\text{replicates 1 and 2 summed}) \), and \( c \) columns. The \( \chi^2 \) value from the \( (2 \times c) \) contingency \( \chi^2 \) analysis with \( (c-1) \) degrees of freedom was loaded into Excel to calculate the one-tailed probability of the \( \chi^2 \) distribution probability \( \left( P \right) \). This value was transformed with \( -\log \left( P \right) \). The experiment-wise error rate was then calculated following the method of Benjamini and Hochberg\(^8\) to lower the number of type I errors (false positives).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Code availability. The overview of the Hi-C workflow, as well as modifications to 3D DNA-association and AaegL5, is shared on GitHub at https://github.com/theaidenlab/AGGW-merge. The source code and executable version of Juicebox Assembly Tools are available at https://aiderenlab.org/assembly. Data files and scripts used for the final polishing of scaffolded, gap-filled assembly are available at https://github.com/skingan/AaegL5_FinalPolish.

Data availability. All raw data have been deposited at NCBI under the following BioProject accession numbers: PRJNA318737 (primary Pacific Biosciences data, Hi-C sequencing primary data and processed contact maps, whole-genome sequencing data from a single male (Fig. 4d) and pools of male and females (Fig. 3d)). Bionano optical mapping data (Figs. 3c, 4c) and 10X linked-read sequences (Extended Data Fig. 8a and Supplementary Data 21)); PRJNA236239 (RNA-seq reads and de novo transcriptome assembly\(^10\)); PRJNA290588 (RNA-seq reads for developmental stage Data 4–6); PRJNA291931 (RNA-seq reads from adult reproductive tissues and developmental time points, Verity Life Sciences (Fig. 1h and Supplementary Data 4, 5, 8, 9)); PRJNA393466 (full-length Pacific Biosciences Iso-Seq transcript sequencing); PRJNA418406 (ATAC-seq data...
from adult female brains at three points in the gonotrophic cycle (Extended Data Fig. 2c, d and data not shown); PRINJA419379 (whole-genome sequencing data from four colonies (Fig. 4d and Extended Data Fig. 9a, b)); PRINJA99617 (restriction-site-associated DNA-sequencing data (Fig. 5a–d)); PRINJA393171 (exon-sequence data (Fig. 5e–g)). Intermediate results related to the AagLS assembly are also available via GitHub (http://github.com/theaidenlab/AGWG-merge) and have been uploaded to GEO (GSE113256). The Hi-C maps are available via http://aid-enlab.org/juicebox. The complete mitochondrial genome is available as Genbank accession MF194022.1, RefSeq accession NC_035159.1. The final genome assembly and annotation are available from the NCBI Assembly Resource under accession GCF_002204515.2.

41. Frichot, E. & François, O. LEA: an R package for landscape and ecological association studies. Methods Ecol. Evol. 6, 925–929 (2015).

42. R Core Team. R: A Language and Environment for Statistical Computing http://www.R-project.org/ (R Foundation for Statistical Computing, Vienna, Austria, 2017).

43. Hare, E. E. & Johnston, J. S. Genome size determination using flow cytometry of propidium iodide-stained nuclei. Methods Mol. Biol. 772, 3–12 (2012).

44. Galbraith, D. W. et al. Rapid flow cytometric analysis of the cell cycle in intact nuclei. Cell 220, 1049–1051 (1983).

45. Chin, C. N. et al. Nonhybrid, finished microbial genome assemblies from long-read sequencing data. Nat. Methods 10, 563–569 (2013).

46. Chandonia, J. M. & Lesk, A. M. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics 13, 238 (2012).

47. Rao, S. S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159, 1665–1680 (2014).

48. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst. 3, 95–98 (2016).

49. Dudchenko, O. et al. The Juicebox Assembly Tools module facilitating de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. Preprint at https://www.biorxiv.org/content/early/2018/01/28/254797 (2018).

50. English, A. C. et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS ONE 7, e47768 (2012).

51. Li, H. Aligning sequencing reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997 (2013).

52. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. Preprint at https://arxiv.org/abs/1207.3907 (2012).

53. Smit, A. F., Hubley, R. & Green, P. RepeatMasker Open version 4.0 http://www.repeatmasker.org (2013–2015).

54. Smit, A. F. A., Hubley, R. & Green, P. RepeatModeler Open version 1.0. http://www.repeatmasker.org (2008–2015).

55. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27, 503–507 (1999).

56. Li, H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094–3100 (2018).

57. Apostoli, B. L., Black, W. C., IV, Reiter, P. & Miller, B. R. Use of randomly amplified polymorphic DNA (RAPD) to examine drug resistance markers of Trypanosoma cruzi. Gene 133, 173–180 (1993).

58. Dupont, J. M., Roffey, T., Pastina, C. & Holcik, M. The CCHC domain: a conserved structural motif in Dicer-like RNase III enzymes. Biochemical and Biophysical Research Communications 352, 16–20 (2007).

59. Ebert, B. A. et al. Stabilization of germline-specific A to U editing by Zc3h12a in C. elegans. Nature 483, 195–198 (2012).

60. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell 38, 576–589 (2010).

61. Kielbasa, S. M., Wan, R., Sato, K., Horton, P. & Frith, M. C. Adaptive seeds tame genomic sequence comparison. Genome Res. 21, 487–493 (2011).

62. Liao, W. et al. Genome-wide discovery of Cbx5-mediated repression sites reveals a role in early development. Development 143, 2845–2855 (2016).

63. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30, 923–930 (2014).

64. Patro, R., Duggal, G., Love, M. I., Inzargar, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods 14, 417–419 (2017).

65. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods 10, 1213–1218 (2013).

66. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10, R25 (2009).

67. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).

68. Attschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. J. Mol. Biol. 215, 403–410 (1990).

69. Neafsey, D. E. et al. Highly evolvable malaria vectors: the genomes of 16 Plasmodium species. Cell 159, 144–155 (2014).

70. Schnell, E., Shaulsky, G., Dietrich, F., Hoffmann, B. & Krex, D. Characterization of a novel immune response regulatory protein (IRRP) in Aedes aegypti. Parasitology 140, 575–583 (2013).

71. Sharakhova, M. V. et al. Imaginal discs—a new source of chromosomes for Drosophila melanogaster. Genome Res. 12, 1071–1076 (2002).

72. Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W. A. & Young, A. E. Stacks: a general software package for handling genome-wide complex genetic data. Mol. Ecol. Res. 14, 7–10 (2014).

73. Merabet, S. & Mann, R. S. To be specific or not: the critical relationship between Hox and TALE proteins. Trends Genet. 32, 334–347 (2016).

74. Cabanettes, F. & Klopp, C. D. GENIES: dot plot large genomes in an interactive, efficient and simple way. Peer J. 6, e4958 (2018).

© 2018 Springer Nature Limited. All rights reserved.
a) AGWG project workflow

b) Genome size measured by flow cytometry

| Species / Strain       | Sex | N  | Average genome size (Mb) | Statistical analysis |
|------------------------|-----|----|--------------------------|----------------------|
| *Aedes mascarenis*     | F   | 6  | 1,254                    | 1,254                | a                   |
|                        | M   | 8  | 1,255                    |                      |                     |
| *Aedes aegypti* Ho Chi Minh City F13 | F   | 5  | 1,233                    | 1,231                | b                   |
|                        | M   | 6  | 1,228                    |                      |                     |
| *Aedes aegypti* Rockefeller | F   | 7  | 1,242                    | 1,227                | bc                  |
|                        | M   | 6  | 1,213                    |                      |                     |
| *Aedes aegypti* LVP_AGWG | F   | 5  | 1,226                    | 1,224                | bc                  |
|                        | M   | 5  | 1,222                    |                      |                     |
| *Aedes aegypti* New Orleans F8 | F   | 8  | 1,219                    | 1,215                | c                   |
|                        | M   | 7  | 1,211                    |                      |                     |
| *Aedes aegypti* Uganda 49-ib-G5 | F   | 5  | 1,190                    | 1,190                | d                   |
|                        | M   | 6  | 1,190                    |                      |                     |

Extended Data Fig. 1 | See next page for caption.
Extended Data Fig. 1 | Project flowchart, measured genome size and assembly process. a, Flowchart of LVP_AGWG strain inbreeding, data collection and experimental design of the AaegL5 assembly process. b, Estimated average 1C genome size for each strain for five Ae. aegypti strains and Ae. mascarensis, the sister taxon of Ae. aegypti, for which the genome size has not previously been measured. There were no significant differences between the sexes within and between the species and strains analysed (P > 0.2). Significant differences between strains were determined using Proc GLM in SAS with both a Tukey and a Scheffé option with the same outcome. Data labelled with different letters are significantly different (P < 0.05). c, Combining Hi-C maps with 2D annotations enabled efficient review of sequences identified as alternative haplotypes by sequence alignment. The figure depicts a roughly 24 Mb × 24 Mb fragment of a contact map generated by aligning a Hi-C dataset to an intermediate genome assembly generated during the process of creating AaegL5. This intermediate assembly was a sequence comprising error-corrected, ordered and oriented FALCON-Unzip contigs. The intensity of each pixel in the contact map correlates with how often pairs of loci co-locate in the nucleus. Maximum intensity is indicated in the lower left of each panel. These maps include reads that do not align uniquely (reads with zero mapping quality); such alignments are randomly assigned to one of the possible genomic locations. Three panels show three types of annotations that are overlaid on top of the contact map. Left, FALCON-Unzip contig boundaries are highlighted as black squares along the diagonal. Notably, large linear features appear above and below the diagonal. These are the result of sequence overlap among contigs, which can indicate the presence of undercollapsed heterozygosity in the contig set. Because reads that do not map uniquely are randomly assigned during the alignment step, Hi-C reads derived from a contig will sometimes be aligned to an overlapping contig. When this happens, the Hi-C read pair may contribute to the formation of a linear feature above and below the diagonal. Therefore, the linear stretches of enriched contact frequency parallel to the diagonal are brought about by the random assignment procedure, and can facilitate the detection of pairs of overlapping contigs. Note that, when the overlap between contigs is owing to undercollapsed heterozygosity, both contigs will exhibit similar long-range contact patterns. This aspect of Hi-C data also provides evidence for the presence of undercollapsed heterozygosity. Centre, LASTZ-alignment-based annotations for fully redundant contigs. The squares shown in blue are obtained by taking diagonal contig boundary annotations (in black) and shifting them up (respectively, left) when drawing above (respectively, below) the diagonal so that the overlapping sequences are horizontally (respectively, vertically) aligned. Note that, as expected, the squares typically span linear, off-diagonal features in the Hi-C data. When one contig is entirely contained in another contig, the redundant contig does not contribute sequence to the merged chromosome-length scaffolds. Right, LASTZ-alignment-based annotations for partially redundant contigs. Again, the squares shown in blue are obtained by taking diagonal contig boundary annotations (in black) and shifting them up and left. The overlaps shown in this panel correspond to contigs that only partially overlap in sequence with other contigs. Consequently, some of their sequence is incorporated in the final fasta. d, Comparison of chromosome lengths between AaegL4 and AaegL5. Numbers are given before post-Hi-C polishing and gap closing. e, Step-wise assembly statistics for Hi-C scaffolding, alternative haplotype removal and annotation. *Removed length, 779,073,495 bp. **The definition of scaffold groups can be found in a previously published study4. ***Gaps between contigs were set to 500 bp for calculating scaffold statistics. N/A, not applicable.
Extended Data Fig. 2 | See next page for caption.
Extended Data Fig. 2 | Remaining assembly gaps, summary of geneset annotation improvement, chromatin accessibility analysis, physical genome map and gene structures of biogenic amine-binding receptors and opsins in AaegL5. 

a. Representation of structural variants identified at assembly gaps by alignment of Bionano optical maps. The estimated size of an insertion (blue) or deletion (red) relative to the reference is represented by the size of the circle. When the size or type of structural variants could not be determined or did not agree between the two optical maps, the location of the assembly gap is plotted in grey. Approximate locations of the centromeres (red triangles) and telomere-associated repeat sequences (blue triangles) are indicated. Raw data are available as Supplementary Data 1.

b. Comparison of protein-coding genes and transcripts in AaegL5.0 (NCBI RefSeq Release 101) and gene set annotations from An. gambiae (Agam), Culex pipiens (Cpip) and D. melanogaster (Dmel).

c. Sex peptide receptor structure in AaegL3.4 and AaegL5.0, and female brain RNA-seq and ATAC-seq reads aligned to AaegL5. Blue lines on the RNA-seq track indicate splice junctions, with the number of reads spanning a junction represented by line thickness. Exons are represented by tall filled boxes and introns by lines. Arrowheads indicate gene orientation.

d. Average read profiles across promoter regions, defined as the transcription start site (TSS) ±2.5 kb. Solid lines represent Tn5-treated native chromatin using the ATAC-seq protocol (n = 4), dotted lines represent Tn5-treated naked genomic DNA (n = 1). Shaded regions represent s.d.

e. A physical genome map was developed by localizing 500 BAC clones to chromosomes using FISH. For the development of a final chromosome map for the AaegL5 assembly, we assigned the coordinates of each outmost BAC clone within a band (Supplementary Data 12) to the boundaries between bands. The final resolution of this map varies on average between 5 and 10 Mb because of the differences in BAC mapping density in different regions of chromosomes.

f. Schematic of predicted gene structures of the Ae. aegypti biogenic amine-binding receptors and opsins. Exons, cylindrical bars; introns, black lines; dopamine receptors, yellow bars; serotonin receptors, magenta bars; muscarinic acetylcholine receptors, green bars; octopamine receptors, blue bars; opsins, dark purple bars; predicted 3' and 5' non-coding sequence (dark shading). The 'unclassified receptor' GPRnna19 is not shown. Details on gene models compared to previous annotations and the predicted amino acid sequences of each gene are available in Supplementary Data 14–16.
Extended Data Fig. 3 Chromosomal arrangement of chemosensory receptor genes. The location of predicted chemoreceptors (odorant receptors (ORs), gustatory receptors (GRs) and ionotropic receptors (IRs)) across all three chromosomes in AaegL5. The blunt end of each arrowhead plotted above each chromosome marks gene position and arrowhead indicates orientation. Filled and open arrowheads represent intact genes and pseudogenes, respectively (Supplementary Data 17–20). This figure is identical to Fig. 2a, but here includes gene names.
Odorant receptors (ORs)

Extended Data Fig. 4 | Phylogenetic tree of odorant receptor gene families from *Ae. aegypti*, *An. gambiae* and *D. melanogaster*. Maximum likelihood odorant receptor tree was rooted with Orco proteins, which are both highly conserved and basal within the odorant receptor family\(^9\). Support levels for nodes are indicated by the size of black circles—reflecting approximate likelihood ratio tests (aLRT values ranging from 0 to 1 from PhyML v.3.0 run with default parameters\(^90\)). Suffixes after protein names are C, minor assembly correction; F, major assembly modification; N, new model; P, pseudogene. Scale bar, amino acid substitutions per site.
Gustatory receptors (GRs)

Extended Data Fig. 5 | Phylogenetic tree of the gustatory receptor gene families from *Ae. aegypti, An. gambiae* and *D. melanogaster*. Maximum likelihood gustatory receptor tree was rooted with the highly conserved and distantly related carbon dioxide and sugar receptor subfamilies, which together form a basal clade within the arthropod gustatory receptor family. Subfamilies and lineages closely related to *D. melanogaster* gustatory receptors of known function are highlighted. Support levels for nodes are indicated by the size of black circles—reflecting approximate likelihood ratio tests (aLRT values ranging from 0 to 1 from PhyML v.3.0 run with default parameters). Suffixes after protein names are C, minor assembly correction; F, major assembly modification; N, new model; P, pseudogene. Scale bar, amino acid substitutions per site.
Extended Data Fig. 6 | Phylogenetic tree of the ionotropic receptor gene families from Ae. aegypti, An. gambiae and D. melanogaster. Maximum likelihood phylogenetic tree of ionotropic receptor protein sequences from the indicated species rooted with highly conserved Ir8a and Ir25a proteins. Conserved proteins with orthologues in all species are named outside the circle, and previously unannotated ionotropic receptors are highlighted with red lines. Support levels for nodes are indicated by the size of black circles—reflecting approximate likelihood ratio tests (aLRT values ranging from - to 1 from PhyML v.3.0 run with default parameters90). Suffixes after protein names are C, minor assembly correction; F, major assembly modification; N, new model; P, pseudogene. Scale bar, amino acid substitutions per site.
Extended Data Fig. 7 | Chemosensory receptor expression in adult *Ae. aegypti* tissues. Previously published RNA-seq data\(^1\) were reanalysed using the new chemoreceptor annotations and genome assembly. Chemoreceptors have been clustered according to Euclidian distance of their expression vectors using the R function hclust. Expression is given for females at three stages of the gonotrophic cycle (0, 48 or 96 h after taking a blood-meal, for which 0 h indicates not blood-fed, 48 h indicates 48 h after the blood-meal, and 96 h indicates gravid). New genes are indicated by black bars on the right.
Extended Data Fig. 8 | Structural variation, the Hox gene cluster and Hox cofactor motifs. a, Linked-read sequencing of two individuals from the LVP_AGWG strain identified putative structural variants in the AaegL5 assembly. b, Comparative genomic arrangement of the Hox cluster (HOXC) in five species (Supplementary Data 22). Note the split of labial (lab) and proboscipedia (pb) between two chromosomes in Ae. aegypti and Cx. quinquefasciatus. Owing to chromosome arm exchange, chromosome 3p in Cx. quinquefasciatus is the homologue of chromosome 2p in Ae. aegypti. c, Repeats in putative telomere-associated sequences downstream of pb in both species. d, Motifs known to mediate protein–protein interactions with the Hox cofactor Extradenticle (Exd) from the five indicated species are aligned using Clustal-Omega. Perfectly aligned residues are coloured according to Hox gene identity, non-conserved residues are grey.
Extended Data Fig. 9 | Population genomic structure and linkage disequilibrium analysis of *Ae. aegypti* strains. 
**a**, Chromosomal patterns of nucleotide diversity ($\pi$) in four strains of *Ae. aegypti* measured in 100-kb non-overlapping windows and presented as a LOESS-smoothed curve. 
**b**, Mean nucleotide diversity in the strains in a, with s.d. indicated in parentheses. Nucleotide diversity ($\pi$) was measured in non-overlapping 100-kb windows. The Liverpool and Costa Rica colonies maintain extensive diversity despite being colonized in the laboratory more than a decade ago, but show reduced genome-wide diversity (on the order of 30–40%) relative to the more recently laboratory colonized Innisfail and Clovis. 
**c**, Pairwise linkage disequilibrium between SNPs located within the same chromosome estimated from 28 wild-caught individuals from the indicated populations. Each point represents the mean linkage disequilibrium for that set of binned SNP pairs. Solid lines are LOESS-smoothed curves, and dashed lines correspond to $r^2_{\text{max}}/2$. Inclusion of additional individuals available from the Amacuzac population (up to 137) had a minimal effect on the linkage disequilibrium estimations ($\Delta R^2 < 0.017$; data not shown). 
**d**, Linkage disequilibrium ($r^2$) values along the *Ae. aegypti* AaegL5 genome assembly based on pairwise SNP comparisons. Data were obtained from the average $r^2$ of SNPs in 1-kb bins.
Extended Data Fig. 10 | QTL analysis of DENV competence in *Ae. aegypti* and Cys-loop LGICs. **a**, Schematic representation of the experimental workflow for testing DENV competence in *Ae. aegypti*, related to Fig. 5b–d. **b**, Comparison of QTL map density constructed against AaegL3 or AaegL5 assemblies. **c**, Concentration–response curves showing the effect on *Ae. aegypti* larval motility of insecticides currently used in veterinary and agricultural applications (mean ± s.e.m., *n* = 7). **d**, Phylogenetic tree of Cys-loop LGIC subunits for *Ae. aegypti* and *D. melanogaster*. The accession numbers of the *D. melanogaster* sequences used in constructing the tree are: Dα1 (CAA30172), Dα2 (CAA36517), Dα3 (CAA75688), Dα4 (CAB77445), Dα5 (AAM13390), Dα6 (AAM13392), Dα7 (AAK67257), Dβ1 (AA27641), Dβ2 (AAK67257), Dβ3 (CAC48166), GluCl (AAG40735), GRD (Q24352), HisCl1 (AAL774413), HisCl2 (AAL774414), LCCH3 (AAB27090), Ntr (NP_651958), pHCl (NP_001034025), RDL (AA28556). For *Ae. aegypti* sequences, see Supplementary Data 24. ELIC (Erwinia ligand-gated ion channel), which is an ancestral Cys-loop LGIC found in bacteria (accession number P0C7B7), was used as an outgroup. Scale bar, amino acid substitutions per site.
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
- Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated
- Clearly defined error bars
- State explicitly what error bars represent (e.g. SD, SE, CI)

Software and code

Policy information about availability of computer code

Data collection

Flow cytometry data were collected and scored using CytExpert software version 1.2.8.0 (supplied with a Beckman Coulter CytoFlex flow cytometer).

Data analysis

Common bioinformatic and statistical analysis software packages were used, including: R, NCBI BLAST, Samtools, Picard, GATK, FALCON, freebayes BLASR, Quiver, arrow, PBjelly, RepeatMasker, RepeatModeler, RepeatScout, Tandem Repeats Finder, Salmon, gmap, HOMER, bowtie, Juicebox, Assembly Tools. Version numbers and specific parameters used during run-time are provided in the methods when appropriate. All custom software related to the Hi-C assembly is open source and available on the Aiden Lab GitHub page, as indicated in the methods.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.
Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data availability statement. All raw data have been deposited at NCBI under the following BioProject Accession numbers: PRJNA318737 (Primary Pacific Biosciences data, Hi-C sequencing primary data and processed contact maps, whole-genome sequencing data from a single male (Fig. 4d), and pools of male and females (Fig. 3d), Bionano optical mapping data (Fig. 3c and Fig. 4c), and 10X linked-read sequences (Extended Data Fig. 8a and Supplementary Data 21)); PRJNA236239 (RNA-seq reads and de novo transcriptome assembly, Extended Data Fig. 2c and Supplementary Data 4, 5, 7, 9); PRJNA209388 (RNA-seq reads for developmental time points, Fig. 1h and Supplementary Data 4–6, 9); PRJNA419241 (RNA-Seq reads from adult reproductive tissues and developmental time points, Verily Life Sciences Fig. 1h and Supplementary Data 4, 5, 8, 9); PRJNA393466 (full-length Pacific Biosciences Iso-Seq transcript sequencing); PRJNA418406 (ATAC-Seq data from adult female brains at three points in the gonotrophic cycle, Extended Data Fig. 2c, d and data not shown); PRJNA419379 (whole-genome sequencing data from colonies Fig. 4d and Extended Data Fig. 9a, b); PRJNA399617 (RAD-Seq data Fig. 5a-d); PRJNA393171 (exome sequencing data Fig. 5e-g). Intermediate results related to the AaegL5 assembly are also available via GitHub (http://github.com/theaidenlab/AGWG-merge) and have been uploaded to GEO (GEO Record: GSE113256). The Hi-C maps are available via http://aidenlab.org/juicebox. The complete mitochondrial genome is available as Genbank accession MF194022.1, RefSeq accession NC_035159.1. The final genome assembly and annotation are available from the NCBI Assembly Resource under accession GCF_002204515.2.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☑ Life sciences  ☐ Behavioural & social sciences  ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| Sample size                          | Sample sizes for genome variability analysis via SNP-chip (Fig. 1c) were determined according to previously published work (Evans et al., 2015 PMID 25721127). Sample sizes for genome size determination (Fig. 1d) were determined according to the standards of the field (see Hare and Johnston, 2011 PMID 22065429). Samples sizes for FISH were determined according to the standards of the field (see Timoshevskiy et al., 2012 PMID 23007640). Sample sizes for dengue virus competence (Fig. 5b-d and Extended Data Fig. 10a), pyrethroid resistance (Fig. 5e-g) and larval motility Ext. Data Figure 10c) were determined by the limited availability of animals, biological or chemical reagents. Bioinformatic analyses were performed with all available data. |
| Data exclusions                      | None |
| Replication                          | Replication does not apply to the primary results of this paper - it was not feasible to independently resequence/reassemble the genome twice within the scope of the funding available to us. |
| Randomization                        | Randomization was not performed in this study. Samples were divided into experimental groups based on species, strain or biological phenotype according to the criteria listed in the methods. |
| Blinding                             | Blinding was not performed for this study. The diversity of sourcing of samples and data precluded centralized collection and blinding of biological material or sequencing data. |

Reporting for specific materials, systems and methods

| Materials & experimental systems     | Methods                  |
|-------------------------------------|--------------------------|
| n/a Involved in the study           | n/a Involved in the study |
| ☑ Unique biological materials      | ☑ ChIP-seq               |
| ☑ Antibodies                       | ☑ Flow cytometry          |
| ☑ Eukaryotic cell lines            | ☑ MRI-based neuroimaging  |
| ☑ Palaeontology                    |                          |
| ☑ Animals and other organisms      |                          |
| ☑ Human research participants      |                          |
Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

| Laboratory animals | Male mosquitoes at pupal stage (6-7 days post-hatching) were used to generate the high molecular weight DNA for the primary assembly, Hi-C data, Illumina sequencing data and Bionano optical mapping data. Male and female pupal or adult mosquitoes of various ages were used for all other data collection. Specific details are provided in the methods. Established laboratory strains used include: Aedes aegypti: LVP_AGWG (Rockefeller University), LVP_ib12 (Virginia Tech and Notre Dame), LVP_MR4 (Centers for Disease Control), Rockefeller (Johns Hopkins), Ho Chi Minh City Vietnam (Yale University), New Orleans USA (Yale University), Uganda (Princeton University), Kamphaeng Phet Province Thailand (Institut Pasteur), Viva Caucel Mexico (Colorado State), Clovis USA (Verily Life Sciences), Innisfail Australia (Verily Life Sciences), Puntarenas Costa Rica (Verily Life Sciences), Liverpool (Verily Life Sciences). Aedes mascarensis: Mauritius (Yale University). |
| Wild animals | Field-collected mosquitoes (Aedes aegypti) were obtained from locations in Australia, Cameroon, Florida, Gabon, Mexico, and Thailand. All appropriate local permits were in place to authorize such collections. Mosquitoes were trapped as adults in the field, or as eggs or larvae reared to adulthood in field laboratories, and euthanized by placement into 100% ethanol to preserve genomic DNA. These animals were shipped as dead samples to the investigators who carried out the analysis. |
| Field-collected samples | Field-collected mosquitoes (Aedes aegypti) were obtained from locations in Australia, Cameroon, Florida, Gabon, Mexico, and Thailand. All appropriate local permits were in place to authorize such collections. Mosquitoes were trapped as adults in the field, or as eggs or larvae reared to adulthood in field laboratories, and euthanized by placement into 100% ethanol to preserve genomic DNA. These animals were shipped as dead samples to the investigators who carried out the analysis. |

Human research participants

Policy information about studies involving human research participants

| Population characteristics | Work with human subjects was covered under Rockefeller IRB protocol LVO-0652 (Laboratory of Leslie Vosshall). Only one subject participated in this study as a source of blood for mosquitoes. |
| Recruitment | One of the authors was the subject for this work, and the subject's participation followed vetting by Rockefeller University administration officials that no coercion by the laboratory head, Leslie Vosshall, to participate in this study had taken place. Written informed consent was obtained prior to enrolling the subject in this study. |