Published in final edited form as:
Nature. 2018 November ; 563(7732): 501–507. doi:10.1038/s41586-018-0692-z.

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

A full list of authors and affiliations appears at the end of the article.

Correspondence to B.J.M: bnmtthws@gmail.com.
*These authors contributed equally to this work.

Author Contributions B.J.M. and L.B.V. conceived the study, coordinated data collection and analysis, designed the figures, and wrote the paper with input from all authors. B.J.M. developed and distributed animals and/or DNA of the LVP_AGWG strain. Pacific Biosciences sample preparation and sequencing: (P.P., M.L.S., J.M.); Assembly: (S.B.K., R.H., J.K., S.K., A.M.P.); Bionano optical mapping: (A.R.H., S.C., J.L., H.C.); Hi-C sample preparation, scaffolding, and deduplication: (O.D., S.S.B. A.D.O., A.P.A., E.L.A); Fig. 1c (B.R.E., A.G.-S., J.R.P); Fig. 1d (J.S.J.); Fig. 1f (L.Z.); Fig. 1g (E.C., V.S.J., V.K.K., M.R.M., T.D.M., B.J.M.);Fig. 1h (I.A., O.S.A., J.E.C., A.M.W. B.J.W., R.G.G.K., S.N.M., B.J.M).; Fig. 2 (C.S.M., H.M.R., Z.Z., N.H.R., B.J.M.); Fig. 3a-d (Z.T., M.V.S., I.V.S., A.S., Y.W., J.T., A.C.D., A.R.H., B.J.M.); Fig. 4 (G.D.W., B.J.M., A.R.H., S.B.K., A.M.P., S.K.); Fig. 5a-d (A.F., I.F., T.F., G.R., L.L.); Fig. 5e-g (C.L.C, K.S.-R., W.C.B.IV, B.J.M.); Extended Data Fig. 1a (B.J.M.); Extended Data Fig. 1b (J.S.J.); Extended Data Fig. 1c-d (O.D., S.S.B. A.D.O., A.P.A., E.L.A); Extended Data Fig 1e (S.B.K., J.K., O.D., E.L.A., S.K., A.M.P., B.J.M.); Extended Data Fig. 2a (A.R.H., B.J.M.); Extended Data Fig. 2b (E.C., V.S.J., V.K.K., M.R.M., T.D.M., B.J.M.); Extended Data Fig. 2c-d (M.H., B.J.M.); Extended Data Fig. 2e (A.S., I.V.S., M.V.S.); Extended Data Fig. 2f (C.A.B.-S., S.S., C.A.H.); Extended Data Fig. 3 (C.S.M., H.M.R., Z.Z., N.H.R., B.J.M.)Extended Data Fig. 4–7 (C.S.M., H.M.R., Z.Z., N.H.R., B.J.M.); Extended Data Fig. 8a (S.N.R., D.E.N.); Extended Data Fig. 8b-c (W.J.G., R.S.M., O.D., E.L.A. B.J.M.); Extended Data Fig. 8d (W.J.G., R.S.M.); Extended Data Fig. 9a-b (J.E.C., A.M.W., B.J.W., R.G.G.K., S.N.M); Extended Data Fig 9c-d (B.R.E., A.G.-S., J.R.P); Extended Data Fig. 10a-b (A.F., I.F., T.F, G.R., L.L.); Extended Data Fig. 10c-d (G.J.L., A.K.J., V.R., S.D.B., F.A.P, D.B.S.); Supplementary Data 1 (A.R.H.); Supplementary Data 2–3 (L.Z.); Supplementary Data 4–9 (I.A., O.S.A., J.E.C., A.M.W. B.J.W., R.G.G.K., S.N.M., B.J.M.); Supplementary Data 10–11 (E.C., V.S.J., V.K.K., M.R.M., T.D.M); Supplementary Data 12 (A.S., I.V.S., M.V.S.); Supplementary Data 13 (S.R., A.S.R.); Supplementary Data 14–16 (C.A.B.-S., S.S., C.A.H.); Supplementary Data 17–20 (C.S.M., H.M.R., Z.Z., N.H.R., B.J.M.); Supplementary Data 21 (S.N.R., D.E.N.); Supplementary Data 22 (W.J.G., R.S.M.); Supplementary Data 23 (G.D.W., B.J.M.); Supplementary Data 24 (G.J.L., A.K.J., V.R., S.D.B., F.A.P, D.B.S.).

The authors declare competing financial interests. P.P., M.L.S., J.M., S.B.K., R.H., J.K. are employees of Pacific Biosciences, a company developing single-molecule sequencing technologies. J.L., S.C., H.C., A.R.H. are employees of Bionano Genomics, Inc. and own company stock options. O.D., S.S.B., A.D.O., A.P.A., E.L.A. are inventors on U.S. provisional patent application 62/347,605, filed 8 June 2016, by the Baylor College of Medicine and the Broad Institute.

Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to B.J.M. (bnmtthws@gmail.com).

Online Content Methods and Extended Data Fig. 1–10 are available in the online version of the paper; references unique to these sections appear only in the online paper. Supplementary Data 1–24 contain raw data supporting the conclusions in the paper.

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[13] Extended Data Fig. 2c-d and Supplementary Data 4, 5, 7, 9); PRJNA209388 (RNA-seq reads for developmental time points[57] 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, 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 final genome assembly and annotation are available from the NCBI Assembly Resource under accession GCF_002204515.2.

Code Availability Statement. The overview of the Hi-C workflow, as well as modifications to 3D-DNA associated with AaegL5, is shared on GitHub at https://github.com/theaidenlab/AGWG-merge. The source code and executable version of Juicebox Assembly Tools are available at http://aidenlab.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.

Abstract

Female *Aedes aegypti* mosquitoes infect >400 million people each year with dangerous viral pathogens including dengue, yellow fever, Zika, and chikungunya. Progress in understanding mosquito biology and developing tools to fight them has been slowed by the lack of a high-quality genome assembly. Here we combine diverse technologies to produce the dramatically 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 elusive sex-determining M locus, and revealed copy-number variation among glutathione S-transferase genes important for insecticide resistance. Using high-resolution quantitative trait locus (QTL) 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 unique aspects of mosquito biology and to develop control strategies to reduce their capacity to spread pathogens[1]. The *Ae. aegypti* genome is large (~1.3 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[3] (Fig. 1a). A more recent assembly, AaegL4[5], 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, or half of the assembly 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 scaffolding to produce a new reference genome (AaegL5) that is highly contiguous, with a decrease of 93% in the number of contigs, and anchored end-to-end to the three *Ae. aegypti* chromosomes (Fig. 1, Extended Data Fig. 1–2). Using optical mapping and linked-read sequencing, we validated local structure and predicted structural variants between haplotypes. We generated an improved geneset annotation (AaegL5.0), as assessed by a mean increase in RNA-Seq read alignment of 12%, connections between many gene models previously split across multiple contigs, and a roughly two-fold increase in the enrichment of 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 (Fig. 2–5, Extended Data Fig. 3–10, Supplementary Data 13–24, Supplementary Methods and Discussion).

This project used the Liverpool *Aedes* Genome Working Group (LVP_AGWG) strain, related to the AaegL3 Liverpool ib12 (LVP_ib12) assembly strain[2] (Fig. 1c, Extended Data Fig. 1a). Using flow cytometry, we estimated the genome size of LVP_AGWG as approximately 1.22 Gb (Fig. 1d, Extended Data Fig. 1b). To generate our primary assembly, we produced 166 Gb of Pacific Biosciences data (~130X coverage for a 1.28 Gb genome) and assembled with FALCON-Unzip[4]. This resulted in a total assembly length of 2.05 Gb (contig N50: 0.96 Mb and NG50: 1.92 Mb, or 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; N50 193 kb) representing alternative haplotypes present in the ~80 male siblings pooled for sequencing (Fig. 1e, Table 1, Extended Data Fig. 1e). The primary assembly was longer than expected for a haploid *Ae. aegypti* genome predicted by flow cytometry and prior assemblies, consistent with remaining alternative haplotypes too divergent to be automatically identified as primary/alternative haplotig pairs.

To generate a linear chromosome-scale reference genome assembly, we combined the primary contigs and haplotigs generated by FALCON-Unzip to create an assembly comprising 7,790 contigs. We used Hi-C to order and orient these contigs, correct misjoins, and merge overlaps (Extended Data Fig. 1c-e). We set aside 359 contigs shorter than 20 kb and used Hi-C data to identify 258 misjoins, 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 corresponding 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, Table 1). We used Hi-C contact maps to estimate centromere position with a resolution of ~5 Mb: Chr1 ~150–154 Mb, Chr2 ~227–232 Mb, Chr3 ~196–201Mb. There are 229 remaining gaps in the primary assembly, including 173 on the three primary chromosomal scaffolds (Extended Data Fig. 2a, Supplementary Data 1). Analysis of near-universal single-copy orthologues via BUSCO[5] revealed a slight increase in complete single-copy orthologues and a reduction in fragmented and missing genes compared to previous assemblies (see Supplementary Methods and Discussion). AaegL5 is dramatically more contiguous than AaegL3 and AaegL4 assemblies (Fig. 1a, b, e, Table 1)[2,6]. Using the TEfam, Repbase, and *de novo* identified repeat databases, we found that 65% of AaegL5 was composed of transposable elements (TEs) and other repetitive sequence (Fig. 1f, Supplementary Data 2–3).

Complete and correct gene models are essential for studying all aspects of mosquito biology. We used the NCBI RefSeq annotation pipeline to produce annotation version 101 (AaegL5.0; Extended Data Fig. 2b) followed by manual curation of key gene families. AaegL5.0 formed the basis for a comprehensive quantification of transcript abundance in 253 sex-, tissue-, and developmental stage-specific RNA-Seq libraries (Supplementary Data 4–8). The AaegL5.0 geneset is considerably more complete and correct than previous versions. Many more genes have high protein coverage when compared to *Drosophila melanogaster* orthologues (915 more genes with >80% coverage, a 12.5% increase over AaegL3.4; Fig. 1g) and >12% more RNA-Seq reads map to the AaegL5.0 geneset annotation than AaegL3.4 (Fig. 1h, Supplementary Data 9). 1,463 genes previously annotated separately as paralogues were collapsed into single gene models and 481 previously fragmented gene models were completed (Supplementary Data 10–11). For example, *sex peptide receptor* (*SPR*), is represented by a 6-exon gene model in AaegL5.0 compared to two partial gene fragments on separate scaffolds in AaegL3.4 (Extended Data Fig. 2c). Genome-wide, we mapped a 1.8-fold higher number of ATAC-Seq reads, known to co-localise with promoters and other cis-regulatory elements[7], to predicted transcription start sites in AaegL5.0 as compared to AaegL3.4, consistent with more complete gene models in AaegL5.0 (Extended Data Fig. 2d).

We next validated the base-level and structural accuracy of the AaegL5 assembly. We estimate the lower bound of base-level accuracy of the assembly to be QV = 34.75 (meaning that 99.9665% of bases are correct, see Supplementary Methods and Discussion). To develop a fine-scale physical genome map based on AaegL5, we compared the assembly coordinates of 500 bacterial artificial chromosome (BAC) clones containing *Ae. aegypti* genomic DNA with physical mapping by fluorescence *in situ* hybridization (FISH) (Extended Data Fig. 2e, Supplementary Data 12). After removing repetitive BAC-end sequences and those with ambiguous FISH signals, 377/387 (97.4%) of probes showed concordance between physical mapping and BAC-end alignment. The 10 remaining discordant signals were not supported by Bionano or 10X analysis, and so likely do not reflect misassemblies in AaegL5. The genome coverage of this physical map is 93.5%, compared to 45% of AaegL3[8], and is among the most complete genome map across mosquito species[9,10].

Curation of multi-gene families

Large multi-gene families are notoriously difficult to assemble and correctly annotate because recently duplicated genes typically share high sequence similarity or can be misclassified as alleles of a single gene. We curated genes in large multi-gene families encoding proteases, G protein-coupled receptors, and chemosensory receptors using the improved AaegL5 genome and AaegL5.0 annotation. Serine proteases mediate immune responses[11] and metalloproteases have been linked to vector competence and mosquito-*Plasmodium* interactions[12]. Gene models for over 50% of the 404 annotated serine proteases/metalloproteases in AaegL3.4 were improved in AaegL5.0, and we found 49 previously unannotated protease genes (Supplementary Data 13). G protein-coupled receptors are membrane proteins that respond to diverse external and internal sensory stimuli. We provide major corrections to gene models encoding 10 visual opsins and 17 dopamine and serotonin receptors (Extended Data Fig. 2f, Supplementary Data 14–16). Insect chemosensory receptors are ligand-gated ion channels in three multi-gene families: odorant receptors (*ORs*), gustatory receptors (*GRs*), and ionotropic receptors (*IRs*). These collectively allow insects to sense a vast array of chemical cues, including carbon dioxide and human body odour that activate and attract female mosquitoes. We identified 117 *ORs*, 72 *GRs* (encoding 107 transcripts), and 135 *IRs* in the AaegL5 assembly (Fig. 2a-b, Extended Data Fig. 3, Supplementary Data 17–20), inferred new phylogenetic trees for each family to investigate the relationship of these receptors in *Ae. aegypti, Anopheles gambiae* malaria mosquitoes, and *D. melanogaster* (Extended Data Fig. 4–6), and revised expression estimates for adult male and female neural tissues using deep RNA-Seq[13] (Extended Data Fig. 5). Our annotation identified 54 new *IR* genes (Fig. 2b, Extended Data Fig. 3, Supplementary Data 17) nearly doubling the known members of this family in *Ae. aegypti.* We additionally reannotated *IRs* in *An. gambiae* and found 64 new genes. In *Ae. aegypti,* chemoreceptors are extensively clustered in tandem arrays (Fig. 2a, Extended Data Fig. 3), in particular on chromosome 3p, where over a third of all chemoreceptor genes (n=111) are found within a 109 Mb stretch. Although 71 *GR* genes are scattered across chromosomes 2 and 3, only *AaegGr2*, a subunit of the carbon dioxide receptor, is found on chromosome 1. Characterization of the full chemosensory receptor repertoire will enable the development of novel strategies to disrupt mosquito biting behaviour.

Structure of the sex-determining M locus

Sex determination in *Aedes* and *Culex* mosquitoes is governed by a dominant male-determining factor (M factor) at a male-determining locus (M locus) on chromosome 1[14–16]. This chromosome is homomorphic between the sexes except for the M/m karyotype, meaning that males are M/m and females are m/m. Despite the recent discovery of the M factor *Nix* in *Ae. aegypti*[17], which was entirely missing in the previous *Ae. aegypti* genome assemblies[2,6], the full molecular properties of the M locus remain unknown. We aligned AaegL5 (from M/m males) and AaegL4 (from m/m females), and identified a region that contained *Nix* in AaegL5 where the assemblies diverged and that 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 a ~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 previously 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 (LTR) 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-localise 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 method to quantify regions of the genome with strictly male-specific signal[20]. A sex-differentiated region in the LVP_AGWG strain extends to a ~100 Mb region surrounding the ~1.5 Mb M locus. This is consistent with the recent analysis of male-female F_{ST} 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 exciting opportunities to study the evolution and maintenance of homomorphic sex-determining chromosomes. The sex-determining 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 capacity to vector pathogens[24]. We produced ‘read cloud’ Illumina sequencing libraries of linked-reads with long-range (~80 kb) phasing information from one male and one female mosquito using the 10X Genomics Chromium platform to investigate structural variants (SVs), including insertions, deletions, translocations, and inversions, in individual mosquitoes. We observed abundant small-scale insertions/deletions (indels; 26 insertions and 81 deletions called, median 42.9 kb) and inversions/translocations (29 called) in these two individuals (Extended Data Fig. 8a, Supplementary Data 21). Eight of the inversions/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, while 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* (*lab*) and *proboscipedia* (*pb*) placing *lab* on a separate chromosome (Extended Data Fig. 8b, Supplementary Data 22). We confirmed this in AaegL4, which was generated with Hi-C contact maps from a different *Ae. aegypti* strain[5], 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 *lab* and *pb* appear to be close to telomeres.

Glutathione S-transferases (GSTs) are a large multi-gene family involved in detoxification of compounds including insecticides. Increased GST activity has been associated with resistance to multiple classes of insecticide, 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 genes on chromosome 2 located centrally in the cluster (*GSTe2*, *GSTe5*, *GSTe7*) are duplicated four times in AaegL5 relative to AaegL3 (Fig. 4a-b, 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-number across this gene cluster (Fig. 4d). GSTe2 is a highly efficient metaboliser of DDT[30], and it is interesting to note that the cDNA from three GST genes in the quadruplication was detected at higher levels in DDT-resistant *Ae. aegypti* mosquitoes from southeast Asia[31].

Genome-wide genetic variation

Measurement of genetic variation within and between populations is key to 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[33]. 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 (LD) in geographically diverse populations of *Ae. aegypti*, we first mapped Affymetrix SNP chip markers designed using AaegL3[34] to positions on AaegL5. We genotyped 28 individuals from two populations from Amacuzac, Mexico and Lopé National Park, Gabon and calculated pairwise LD of SNPs from 1 kb bins both genome-wide and within each chromosome (Extended Data Fig. 9c-d). The maximum LD in the Mexican population is approximately twice that of the Gabon population, likely reflecting a recent bottleneck associated with spread of this species out of Africa.

Dengue competence and pyrethroid resistance

To illustrate the value of AaegL5 for QTL mapping, we employed restriction site-associated DNA (RAD) markers to locate QTLs underlying dengue virus (DENV) vector competence. We identified and genotyped RAD markers in the F_2 progeny of a laboratory cross between wild *Ae. aegypti* founders from Thailand (Extended Data Fig. 10a)[35]. 197 F_2 females in this mapping population 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 underlying the likelihood of DENV dissemination from the midgut (i.e., systemic infection), an important component of DENV vector competence[36]. 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 problem[37]. Understanding mechanisms underlying 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 (cysLGIC) superfamily in AaegL5. We found 22 subunit-encoding cysLGICs (Extended Data Fig. 10d, Supplementary Data 24), of which 14 encode nicotinic acetylcholine receptor (nAChR) subunits. nAChRs consist of a core group of subunit-encoding genes (α1 to α8 and β1) that are highly conserved between insect species, and at least one divergent subunit[38]. While *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-LGIC-targeting compounds have potential as mosquito larvicides. The improved annotation presented here provides an invaluable 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_{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 cysLGICs will catalyse the search for new resistance-breaking insecticides. A doubling in the known number of chemosensory *IRs* 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 populations[39], 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 generations[40], 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[1].

METHODS

Ethics information

The participation of humans in blood-feeding mosquitoes during routine colony maintenance at The Rockefeller University was approved and monitored by The Rockefeller University Institutional Review Board (IRB protocol LVO-0652). All human subjects gave their written 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 10X Genomics linked-reads (Extended Data Fig. 1a). The same single male was crossed to a single female in two additional generations to generate high-molecular weight (HMW) genomic DNA for Pacific Biosciences long-read sequencing and to establish a colony (LVP_AGWG). Rearing was performed as previously described[13] and all animals were offered a human arm as a blood source.

SNP analysis of mosquito strains.

Data were generated as described[34], and PCA using LEA 2.0 available for R v3.4.0[41,42]. The following strains were used: *Ae. aegypti* LVP_AGWG (Samples from the laboratory strain used for the AaegL5 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 David Severson, University of Notre Dame), *Ae. aegypti* LVP_MR4 (laboratory strain labelled LVP_ib12 obtained 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 Basile Kamgang), *Ae. aegypti* Rockefeller (laboratory strain provided in 2016 by George Dimopoulos, Johns Hopkins Bloomberg School of Public Health), *Ae. aegypti* Key West, Florida (field specimens collected in 2016 and provided by Walter Tabachnick). Strains used for the LD data presented in Extended Data Fig. 9c-d were: *Ae. aegypti* from Amacuzac, Morelos, Mexico (field specimens collected in 2016 and provided by Cassandra Gonzalez Acosta) and *Ae. aegypti* from La Lope National park forest, Gabon (field specimens collected and provided by Siyang 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 μL/mg, not 50 μL/mg, and samples were stained in the cold and dark for 24 hr 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 ribonuclease A in 1 litre of ddH_2O, adjusted to pH 7.2 with HCl and filtered through a 0.22 μm filter)[44] and grinding with 15 strokes of the A pestle at a rate of 3 strokes/2 sec. The resultant ground mixture was filtered through a 60 μm nylon filter (Spectrum Labs, CA). Samples were stained with 25 μg of propidium iodide and held in the cold (4^oC) and dark for 24 hr 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 2000 nuclei were scored under each 2C peak and all scored peaks had a CV of 2.5 or less[43,44]. Average channel numbers for sample and standard 2C peaks were scored using CytExpert software version 1.2.8.0 supplied with the CytoFlex flow cytometer. Significant differences among strains were determined using Proc GLM in SAS with both a Tukey and a Sheffé option. Significance levels were the same with either option. Genome size was determined as the ratio of the mean channel number of the 2C sample peak divided by the mean channel number of the 2C *D. virilis* standard peak times 328 Mb, where 328 Mb is the amount of DNA in a gamete of the standard. The following species/strains were used: *Ae. mascarensis* (collected by A. Bheecarry on Mauritius in December 2014. Colonized and maintained by J.R.P.), *Ae. aegypti* Ho Chi Minh City F13 (provided by Duane J. Gubler, Duke-National University of Singapore as F1 eggs from females collected in Ho Chi Minh City in Vietnam, between August and September 2013. Colonized and maintained for 13 generations by A.G.-S.), *Ae. aegypti* Rockefeller (laboratory strain provided by Dave Severson, Notre Dame), *Ae. aegypti* LVP_AGWG (reared as described in Extended Data Fig. 1a from a strain labelled LVP_ib12 maintained by M.V.S. at Virginia Tech), *Ae. aegypti* New Orleans F8 (collected by D. Wesson in New Orleans 2014, colonized and maintained by J.R.P. through 8 generations of single pair mating), *Ae. aegypti* Uganda 49-ib-G5 (derived by C.S.M. through 5 generations of full-sibling mating of the U49 colony established from eggs collected by John-Paul Mutebi in Entebbe, Uganda in March 2015).

Pacific Biosciences library construction, sequencing, and assembly.

HMW DNA extraction for Pacific Biosciences sequencing—HMW DNA extraction for Pacific Biosciences sequencing was performed using the Qiagen MagAttract Kit (#67563) following the manufacturer's protocol with approximately 80 male sibling pupae in batches of 25 mg.

SMRTbell Library Construction and Sequencing—Three libraries were constructed using the SMRTbell Template Prep Kit 1.0 (Pacific Biosciences). Briefly, genomic DNA (gDNA) was mechanically sheared to 60 kb using the Megaruptor system (Diagenode) followed by DNA damage repair and DNA end repair. Universal blunt hairpin adapters were then ligated onto the gDNA molecules after which non-SMRTbell molecules were removed with exonuclease. Pulse field gels were run to assess the quality of the SMRTbell libraries. Two libraries were size selected using SageELF (Sage Science) at 30 kb and 20 kb, the third library was size selected at 20 kb using BluePippin (Sage Science). Prior to sequencing, another DNA damage repair step was performed and quality was assessed with pulse field gel electrophoresis. A total of 177 SMRT cells were run on the RS II using P6-C4 chemistry and 6 hr movies.

Contig Assembly and Polishing—A diploid contig assembly was carried out using FALCON v.0.4.0 followed by the FALCON-Unzip module (revision 74eefabdcc4849a8cef24d1a1bbb27d953247bd7)[4]. The resulting assembly contains primary contigs, a partially-phased haploid representation of the genome, and haplotigs, which represent phased alternative alleles for a subset of the genome. Two rounds of contig polishing were performed. For the first round, as part of the FALCON-Unzip pipeline, primary contigs and secondary haplotigs were polished using haplotype-phased reads and the Quiver consensus caller[45]. For the second round of polishing we used the "resequencing" pipeline in SMRT Link v.3.1, with primary contigs and haplotigs concatenated into a single reference. Resequencing maps all raw reads to the combined assembly reference with BLASR (v. 3.1.0)[46], followed by consensus calling with arrow.

Hi-C sample preparation and analysis.

Library Preparation—Briefly, insect tissue was crosslinked and homogenized. The nuclei were then extracted and permeabilised, and libraries were prepared using a modified version of the *in situ* Hi-C protocol that we optimized for insect tissue[47]. Separate libraries were prepared for samples derived from three individual male pupae. The resulting libraries were sequenced to yield 118M, 249M and 114M reads (coverage: 120X), and processed using Juicer[48].

Hi-C Approach—Using the results of FALCON-Unzip as input, we used Hi-C to correct misjoins, to order and orient contigs, and to merge overlaps (Extended Data Fig. 1c-e). The Hi-C based assembly procedure we employed is described in detail in the Supplementary Methods and Discussion. Notably, both primary contigs and haplotigs were used as input. This was essential because Hi-C data identified genomic loci where the corresponding sequence was absent in the primary FALCON-Unzip contigs, and present only in the haplotigs; the loci would have led to gaps, instead of contiguous sequence, if the haplotigs were excluded from the Hi-C assembly process (Extended Data Fig. 1e).

Hi-C Scaffolding—We set aside 359 FALCON-Unzip contigs shorter than 20 kb, because such contigs are more difficult to accurately assemble using Hi-C. To generate chromosome-length scaffolds, we used the Hi-C maps and the remaining contigs as inputs to the previously described algorithms[6]. Note that both primary contigs and haplotigs were used as input. We performed quality control, manual polishing, and validation of the scaffolding results using Assembly Tools[49]. This produced 3 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 Assembly Tools[49] 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 (AaegL3)[2], include high heterozygosity levels in the species and incomplete inbreeding in the samples we sequenced. The merge algorithm described in Dudchenko et al.[6] detects and merges draft contigs that overlap one another due to undercollapsed heterozygosity. Since undercollapsed heterozygosity does not affect most loci in a typical draft assembly, the default parameters are relatively stringent. We adopted more permissive parameters for AaegL5 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 Hi-C maps and LASTZ alignments (Extended Data Fig. 1c and Supplementary Methods and Discussion). We then reran the merge step, using the whitelist as an additional input. Finally, we performed quality control of the results using Assembly Tools[49], which confirmed the absence of the undercollapsed heterozygosity that we had previously observed. The resulting assembly contained 3 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). Importantly, 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. The assembly contains three chromosome-length scaffolds and 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.

Final gap-filling and polishing.

Scaffolded Assembly Polishing—Following scaffolding and de-duplication, we performed a final round of arrow polishing. PBJelly[50] from PBSuite version 15.8.24 was used for gapfilling of the de-duplicated HiC assembly (see Protocol.xml in Supplementary Methods and 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/skingan/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 (QV) was estimated using independent Illumina sequencing data from a single individual male pupa (library H2NJHADXY_1/2). Reads were aligned with bwa mem 0.7.12-r1039[51]. FreeBayes v1.1.0–50-g61527c5-dirty[52] was used to call SNPs and short indels with the parameters -C 2 −0 -O -q 20 -z 0.10 -E 0 -X -u -p 2 -F 0.6. Any SNP and short indels showing heterozygosity (e.g. 0/1 genotype) was excluded. The QV was estimated at 34.75 using the PHRED formula with SNPs as the numerator (597,798) and number of bases with at least 3-fold coverage 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, RefSeq accession number NC_035159.1.

Bionano optical mapping.

High-molecular weight DNA extraction—High-molecular weight (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. 47 mg of frozen male pupae were fixed in 2% v/v formaldehyde in kit Homogenization Buffer (Bionano #20278), for 2 min on ice. The pupae were roughly homogenized by blending for 2 sec, 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 sec. Homogenate was filtered with a 100 μm nylon filter, fixed with ethanol for 30 min on ice, spun down, and washed with more Homogenization Buffer (to remove residual formaldehyde). The final pellet was resuspended in Homogenization Buffer.A single agarose plug was made using the resuspended cells, using the CHEF Mammalian Genomic DNA Plug Kit (BioRad #170–3591), following the manufacturer’s instructions. The plug was incubated with Lysis Buffer (Bionano #20270) and Puregene Proteinase K (Qiagen #1588920) overnight at 50°C, then again the following morning for 2 hr (using new buffer and Proteinase K). The plug was washed, melted, and solubilized with GELase (Epicentre #G09200). The purified DNA was subjected to four hr of drop dialysis (Millipore, #VCWP04700) and quantified using the Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen/ Molecular Probes #P11496).

DNA labelling—DNA was labelled according to commercial protocols using the DNA Labelling Kit –NLRS (RE-012–10, Bionano Genomics, Inc). Specifically, 300 ng of purified genomic DNA was nicked with 7 U nicking endonuclease Nt.BspQI (New England BioLabs, NEB) at 37°C for 2 hr in NEBuffer3. The nicked DNA was labelled with a fluorescent-dUTP nucleotide analogue using Taq polymerase (NEB) for 1 hr at 72°C. After labelling, the nicks were ligated with Taq ligase (NEB) in the presence of dNTPs. The backbone of fluorescently labelled DNA was counterstained with YOYO-1 (Invitrogen).

Data collection—The DNA was loaded onto the nanochannel array of Bionano Genomics IrysChip by electrophoresis of DNA. Linearized DNA molecules were then imaged automatically followed by repeated cycles of DNA loading using the Bionano Genomics Irys system. The DNA molecules backbones (YOYO-1 stained) and locations of fluorescent labels along each molecule were detected using the in-house software package, IrysView. The set of label 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 Gb and 259 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_irys.xml) and pre-release version of Bionano Solve3.1 (Pipeline version 6703 and RefAligner version 6851). Based on Overlap-Layout-Consensus paradigm, pairwise comparison of all DNA molecules was performed to create an overlap graph, which was then used to create the initial consensus genome maps. By realigning molecules to the genome maps (RefineB P-value = $10e^{-11}$) 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 extension step, the software aligned molecules to genome maps (Extension P-value = $10e^{-11}$), and extended the maps based on the molecules aligning past the map ends. Overlapping genome maps were then merged using a Merge P-value cutoff of $10e^{-15}$. These extension and merge steps were repeated five times before a final refinement was applied to “finish” all genome maps (Refine Final P-value = $10e^{-11}$). Two genome map *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 cutoff of $10e^{-12}$.

Transposable element identification.

Identification of known transposon elements (TE)—We first identified known TEs using RepeatMasker (version 3.2.6)[53] against the mosquito TEfam (https://tefam.biochem.vt.edu/tefam/, data downloaded on July 2017), a manually curated mosquito TE database. We then ran RepeatMasker using the TEfam database and Repbase TE library (version 10.05). RepeatMasker was set to default parameters with the –s (slow search) and NCBI/RMblast program (2.2.28).

***De novo* repeat family identification—**We searched for repeat families and consensus sequences using the *de novo* repeat prediction tool RepeatModeler (version 1.0.8)[54] 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 Finder (TRF)[55]. 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 TE structures in AaegL5 can be found at: https://github.com/VosshallLab/AGWG-AaegL5

Generation of RefSeq geneset annotation.

The AaegL5 assembly was deposited at NCBI in June 2017 and annotated using the NCBI RefSeq Eukaryotic gene annotation pipeline[56]. Evidence to support the gene predictions came from over 9 billion Illumina RNA-seq reads, 67k Pacific Biosciences IsoSeq reads, 300k ESTs, and well-supported proteins from *D. melanogaster* 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 ligand-gated ion channels) are in Supplementary Methods and 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 [13,57] and unpublished RNA-Seq reads from tissue-specific libraries produced by Verily Life Sciences were mapped to the RefSeq assembly GCF_002204515.2_AaegL5.0 with STAR aligner (v2.5.3a)[58] using the 2-pass approach. Reads were first aligned in the absence of gene annotations using the following parameters: --outFilterType BySJout; --alignIntronMax 1000000; --alignMatesGapMax 1000000; --outFilterMismatchNmax 999; --outFilterMismatchNoverReadLmax 0.04; --clip3pNbases 1; --outSAMstrandField intronMotif; --outSAMattrIHstart 0; --outFilterMultimapNmax 20; --outSAMattributes NH HI AS NM MD; --outSAMattrRGline; --outSAMtype BAM SortedByCoordinate. Splice junctions identified during the 1st pass mapping of individual libraries were combined and supplied to STAR using the –sjdbFileChrStartEnd 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 featureCounts[59] with default parameters. Count data were transformed to TPM (transcripts per million) values using a custom Perl script. Details on libraries, alignment statistics, and gene expression estimates (expressed in transcripts per million; TPM) are provided as Supplementary Data 4–8.

Identification of ‘collapsed’ and ‘merged’ gene models from AaegL3.5 to AaegL5.0.

VectorBase annotation AaegL3.5 was compared to NCBI *Aedes aegypti* annotation release 101 on AaegL5.0 using custom code developed at NCBI as part of NCBI’s eukaryotic genome annotation pipeline. First, assembly-assembly alignments were generated for AaegL3 (GCA_000004015.3) x AaegL5.0 (GCF_002204515.2) as part of NCBI’s Remap coordinate remapping service, as described at https://www.ncbi.nlm.nih.gov/genome/tools/remap/docs/alignments. The alignments are publicly available in NCBI’s Genome Data Viewer (https://www.ncbi.nlm.nih.gov/genome/gdv/), the Remap interface, and by FTP in either ASN.1 or GFF3 format (ftp://ftp.ncbi.nlm.nih.gov/pub/remap/Aedes_aegypti/2.1/). Alignments are categorized as either ‘first pass’ (aka reciprocity=3) or ‘second pass’ (aka reciprocity=1 or 2). First pass alignments are reciprocal best alignments, and are used to identify regions on the two assemblies that can be considered equivalent. Second pass alignments are cases where two regions of one assembly have their best alignment to the same region on the other assembly. These are interpreted to represent regions where two paralogous regions in AaegL3 have been collapsed into a single region in AaegL5, or vice versa.

For comparing the two annotations, both annotations were converted to ASN.1 format and compared using an internal NCBI program that identifies regions of overlap between gene, mRNA, and CDS features projected through the assembly-assembly alignments. The comparison was performed twice, first using only the first pass alignments, and again using only the second pass alignments corresponding to regions where duplication in the AaegL3 assembly had been collapsed. Gene features were compared, requiring at least some overlapping CDS in both the old and new annotation to avoid noise from overlapping genes and comparisons between coding vs. non-coding genes. AaegL5.0 genes that matched to two or more VectorBase AaegL3.5 genes were identified. Matches were further classified as collapsed paralogs if one or more of the matches was through the second pass alignments, or as improvements due to increased contiguity or annotation refinement if the matches were through first pass alignments (e.g. two AaegL3.5 genes represent the 5’ and 3’ ends of a single gene on AaegL5.0, such as SPR). Detailed lists of merged genes are in Supplementary Data 10–11.

Comparison of alignment to AaegL3.4 and AaegL5.0.

The sequences comprising transcripts from the AaegL5.0 geneset annotation was extracted from coordinates provided in GCF_002204515.2_AaegL5.0_genomic.gtf. Sequences corresponding to AaegL3.4 geneset annotations were downloaded from Vectorbase (https://www.vectorbase.org/download/aedes-aegypti-liverpooltranscriptsaaegl34fagz. Salmon (v0.8.2)[60] indices were generated with default parameters, and all libraries described in Supplementary Data 4 were mapped to both AaegL3.4 and AaegL5 sequences using ‘quant’ mode with default parameters. Mapping results are presented as Supplementary Data 9 and Fig. 1h.

ATAC-Seq.

The previously described ATAC-Seq protocol was adapted for *Ae. aegypti* brains[61]. Individual brains from LVP_MR4 non-blood-fed females (Extended Data Fig. 2c-d) or females 48 h or 96 hr after taking a human blood-meal (data not shown) were dissected in 1X PBS, immediately placed in 100 μL ice-cold ATAC lysis buffer (10 mM Tris-Hcl, pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% IGEPAL CA-630), and homogenized in a 1.5 ml Eppendorf tube using 50 strokes of a Wheaton 1 ml PTFE tapered tissue grinder. Animals at 96 hr after the blood-meal were deprived access to a water oviposition site and were considered gravid at the time of dissection. Lysed brains were centrifuged at 400 g for 20 min at 4^{o}C and the supernatant was discarded. Nuclei were resuspended in 52.5 μL 1X Tagmentation buffer (provided in the Illumina Nextera DNA Library Prep Kit) and 5 μL were removed to count nuclei on a hemocytometer. 50,000 nuclei were used for each transposition reaction. The concentration of nuclei in Tagmentation buffer was adjusted to 50,000 nuclei in 47.5 μL Tagmentation buffer and 2.5 μL Tn5 enzyme was added (provided in the Illumina Nextera DNA Library Prep Kit). The remainder of the ATAC-Seq protocol was performed as described[61]. The final library was purified and size-selected using double-sided AMPure XP beads (0.6×, 0.7×). The library was checked on an Agilent Bioanalyzer 2100 and quantified using the Qubit dsDNA HS Assay Kit. Resulting libraries were sequenced 75 bp paired-end on an Illumina NextSeq500 platform at an average read depth of 30.5 million reads per sample. Raw fastq reads were checked for nucleotide distribution and read quality using FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) and mapped to the AaegL5 and AaegL3 versions of the *Ae. aegypti* genome using Bowtie 2.2.9[62]. Aligned reads were processed using Samtools 1.3.1[63] and Picard 2.6.0 (http://broadinstitute.github.io/picard/index.html), and only uniquely mapped and non-redundant reads were used for downstream analyses. To compare the annotation and assembly of the *sex peptide receptor* (*SPR*) gene in AaegL3 and AaegL5, we used NCBI BLAST[64] to identify AAEL007405 and AAEL010313 as gene fragments in AaegL3.4 annotation that map to *SPR* in the AaegL5.0 genome (BLAST e-values for both queries mapping to *SPR* were 0.0). Next, we used GMAP[65] to align AAEL007405 and AAEL010313 fasta sequences to AaegL5. The resulting .gff3 annotation file was utilized by Gviz[66] to plot RNA-Seq reads and sashimi plot as well as ATAC-Seq reads in the region containing *SPR*. Transcription start site analysis was performed using HOMER 4.9[67]. Briefly, databases containing 2 kb windows flanking transcription start sites genome-wide were generated using the ‘parseGTF.pl’ HOMER script from AaegL3.4 and AaegL5.0 gff3 annotation files. Duplicate transcription start sites and transcription start sites that were within 20 bp from each other were merged using the ‘mergePeaks’ HOMER script. Coverage of ATAC-Seq fragments in predicted transcription start site regions was calculated with the ‘annotatePeaks.pl’ script. Fold change in predicted transcription site regions was calculated by dividing the ATAC fragments per base pair per predicted transcription start site in the AaegL5.0 genome by ATAC fragments per base pair per predicted transcription start site in the AaegL3.4 genome at the 0 base pair point in each predicted transcription start site. Coverage histograms were plotted using ggplot2 2.2.1 in RStudio 1.1.383, R 3.4.2[42].

M locus analysis.

Aligning chromosome assemblies and Bionano scaffolds—The boundaries of the M locus were identified by comparing the current AaegL5 assembly and the AaegL4 assembly[6] using a program called LAST[68] (data not shown). To overcome the challenges of repetitive hits, both AaegL5 and AaegL4 assemblies were twice repeat-masked[53] against a combined repeat library of TEfam annotated TEs (https://tefam.biochem.vt.edu/tefam/)[2] and a RepeatModeler output[54] from the *Anopheles* 16 genomes project[69]. The masked sequences were then compared using BLASTN[64] and we then set a filter for downstream analysis to include only alignment with ≥98% identity over 1000 bp. After the identification of the approximate boundaries of the M locus (and m locus), which contains two male-specific genes, *myo-sex*[18] and *Nix*[17], 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[64] that spans the entire M locus and extends beyond its two borders.

Chromosome quotient (CQ) analysis—CQ[20] was calculated for each 1000 bp window across all AaegL5 chromosomes. To calculate 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 LVP_AGWG male to 10 virgin females. Eggs from this cross were hatched, and virgin male and female adults collected within 12 hr of eclosion to verify their non-mated status. We generated genomic DNA from 5 males and 5 females from each of these crosses. Sheared genomic DNA was used to generate libraries for Illumina sequencing with the Illumina Tru-Seq 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_i of a 1000 bp window, $CQ_{(Si)}=F_{(Si)}/M_{(Si)}$, where $F_{(Si)}$ is the number of female Illumina reads aligned to S_i, and $M_{(Si)}$ is the number of male Illumina reads aligned to S_i. 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 1000 bp window had at least 20 fold more hits to the male Illumina data than to the female Illumina data. Not every 1000 bp window produces a CQ value because many were completely masked by RepeatMasker[53]. To ensure that each CQ value represents a meaningful data point obtained with sufficient alignments, 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 1000 bp window (Fig. 3d). Under these conditions, there is not a single 1000 bp fragment on chromosomes 2 and 3 that showed CQ=0.05 or lower.

Chromosome fluorescent *in situ* hybridization.—Slides of mitotic chromosomes were prepared from imaginal discs of 4th instar larvae following published protocols[3,70,71]. BAC clones were obtained from the University of Liverpool[19] or from a previously described BAC library[72]. 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^{o}C. DNA from the BACs was extracted using Sigma PhasePrep TM BAC DNA Kit (Sigma-Aldrich, catalogue #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 1000× 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[17].

Identification and analysis of *Ae. aegypti* GST and P450 genes and validation of the repeat structure of the GSTe cluster.

Genes were initially extracted from the AaegL5.0 genome annotation (NCBI Release 101) by text search and filtered to remove "off target" matches (e.g. "cytochrome P450 reductase"), then predicted protein sequences of a small number of representative transcripts used to search the protein set using BLASTp, to identify by sequence similarity sequences not captured by the text search (resulting in two additional P450s, no GSTs). For each gene family, predicted protein sequences were used to search the proteins of the AaegL3.4 geneset using BLASTp. All best matches, and additional matches with amino acid identity >90% were tabulated for each gene family (Supplementary Data 23) to identify both closely related paralogues and alleles annotated as paralogues in AaegL3.4. Based on 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 (CYP9J20 and CYP9J21 for XP_001649103.2; CYP6P12 and CYP6BZ1 for XP_021694388.1). Compared to AaegL3.4, 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 4 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 due to frameshifts, but these could be due to uncorrected errors in the assembly. Error correction of all duplicated units was not possible due 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 AaegL3 reference genome using bowtie2 v2.2.4[73], with '--very-fast-local' alignment parameters, an expected fragment size between 0 and 1500 bp and relative orientation "forward-reverse" ("-I 0 -X 1500 -fr"). Aligned reads with a mapping quality less than 10 were removed using Samtools[63]. 'featureCounts', part of the 'Subread' v1.5.0-p2 package[74], was used to assign read pairs or reads ('tags') aligned to either DNA strand ("-s 0") and overlapping a gene's coding regions by at least 100 bp ("-t CDS --minOverlap 100") to genes as an estimate of representation in the genome. Gene-wise tag counts were normalised by calculating the FPKM (fragments per kilobase of gene length per million mapped reads), using the following formula:

(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 GST epsilon 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 (see below). Median FPKM values across all genes ranged from 47.68 to 48.46 and gene-wise FPKM normalised 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 AaegL5 chromosome 2 (351,597,324 – 351,719,186 bp) and performed alignment of Pacific Biosciences reads using Gepard v1.4.0[75]. To validate this repeat structure, we aligned two *de novo* optical maps created by Bionano using linearized DNA labelled with Nt.BspQI or Nb.BssSI. Single molecules from both maps span the entire region and the predicted restriction pattern provides support for the repeat structure as presented in AaegL5 (Fig. 4c).

QTL mapping of dengue virus vector competence.

In theory, a good-quality genome assembly is not necessary for QTL mapping procedures 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 exemplified 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 AaegL3 or AaegL5 genome assemblies. Although the initial number of markers was 616 in both cases, the final linkage map was 3.3-fold denser with AaegL5 than with AaegL3, as shown in Extended Data Fig. 10b. The difference in marker density between the two linkage maps is due to many markers being filtered out from the AaegL3 data. Because the AaegL3 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 AaegL4, which de-duplicated and scaffolded AaegL3 onto chromosomes[6], would likely 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 F_2 intercross design, the fully sex-linked markers lacked the F_0 paternal genotype in F_2 females and segregated in the same manner as a backcross design. No linkage analysis method is readily available to deal with a chromosome that behaves like a mixture of intercross and backcross designs. Therefore, AaegL3-guided linkage analysis and QTL mapping for chromosome 1 were restricted to the fully sex-linked region based on a backcross design. In contrast, AaegL5-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 under-estimation 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, AaegL5 allowed a dramatically improved QTL mapping resolution over AaegL3. For instance, we mapped the same QTL underlying systemic DENV dissemination at the extremity of chromosome 2 with both AaegL3 and AaegL5. The 1.5 LOD support interval was much larger for the AaegL3-guided linkage map (0–50 cM, 74% of the linkage group) than for the AaegL5-guided linkage map (0–17 cM, 9% of the linkage group). We now present this analysis in Extended Data Fig. 10b.

Mosquito crosses—A large F_2 intercross was created from a single mating pair of field-collected F_0 founders. Wild mosquito eggs were collected in Kamphaeng Phet Province, Thailand in February 2011 as previously described[35]. Briefly, F_0 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 hr light-dark cycle) with access to 10% sucrose. The F_0 male and female initiating the cross were chosen from different collection sites to avoid creating a parental pair with siblings from the same wild mother[76,77]. Their F_1 offspring were allowed to mass-mate and collectively oviposit to produce the F_2 progeny (Extended Data Fig. 10a). A total of 197 females of the F_2 progeny were used as a mapping population to generate a linkage map and detect QTLs underlying vector competence for dengue virus (DENV).

Vector competence—Four low-passage DENV isolates were used to orally challenge the F_2 females as previously described[35]. Briefly, four random groups of females from the F_2 progeny were experimentally exposed to two virus isolates belonging to dengue serotype 1 (KDH0026A and KDH0030A) and two virus isolates belonging to dengue serotype 3 (KDH0010A and KDH0014A), respectively. All four virus isolates were derived from human serum specimens collected in 2010 from clinically ill dengue patients at the Kamphaeng Phet Provincial Hospital[35]. 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 deposited into GenBank (accession numbers HG316481–HG316484). Phylogenetic analysis assigned the viruses to known viral lineages that were circulating in Southeast Asia in the previous years[35]. Each isolate was amplified twice in C6/36 (*Ae. albopictus*) cells prior to vector competence assays. Four- to seven-day-old F_2 females were starved for 24 hr and offered an infectious blood-meal for 30 min. Viral titres in the blood meals ranged from 2.0×10^4 to 2.5×10^5 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/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 described[35,78].

Genotyping—Mosquito genomic DNA was extracted using the NucleoSpin 96 Tissue Core Kit (Macherey-Nagel). For the F_0 male, it was necessary to perform whole-genome amplification using the Repli-g Mini kit (Qiagen) to obtain a sufficient amount of DNA. F_0 parents and females of the F_2 progeny were genotyped using a modified version of the original double-digest restriction-site associated DNA (RAD) sequencing protocol[79], as previously described[80]. 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 pipeline[80] was used to process the raw sequence reads. High-quality reads (Phred scores >25) trimmed to the 140-bp length were aligned to the AaegL5 reference genome (July 2017) using Bowtie v0.12.7[62]. Parameters for the ungapped 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 v1.19[81,82]. Downstream analyses only used high-quality genotypes at informative markers that were homozygous for alternative alleles in the F_0 parents (e.g., AA in the F_0 male and BB in the F_0 female), had a sequencing depth ≥10×, and were present in ≥60% of the mapping population.

Linkage map—A comprehensive linkage map based on recombination fractions among RAD markers in the F_2 generation was constructed using the R package OneMap v2.0–3[83]. Every informative autosomal marker is expected to segregate in the F_2 mapping population at a frequency of 25% for homozygous (AA and 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 χ^2 test. Due 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[21], strong deviations from the expected Mendelian segregation ratios were observed for a large proportion of markers assigned to chromosome 1 in the female 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. Due 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 *rf.2pts* function with default parameters for all three chromosomes. The *rf.2pts* function that implements the expectation-maximization (EM) algorithm was used to estimate haplotype frequencies and recombination rates between markers[11] under the assumption of autosomal Hardy-Weinberg proportions. Due to this analytical assumption, the estimates of centiMorgans (cM) distances could be over- or under-estimated for markers on chromosome 1. Markers linked with a logarithm-of-odds (LOD) score ≥11 were assigned to the same linkage group. Linkage groups were assigned to the three distinct *Ae. aegypti* chromosomes based on the physical coordinates of the AaegL5 assembly. Recombination fractions were converted into genetic distances in cM using the Kosambi mapping function[84]. Linkage maps were exported in the R/qtl environment[85] where they were corrected for tight double crossing-overs with the *calc.errorlod* function based on a LOD cutoff threshold of 4. Duplicate markers with identical genotypes were removed with the *findDupMarkers* function. To remove markers located in highly repetitive sequences, RAD sequences were blasted against the AaegL5 assembly using BLASTn v2.6.0. Markers with >1 blast hit on chromosomes over their 140-bp length and 100% identity were excluded from linkage analysis. Reported RAD markers were distributed as follows: Chr 1, n=76; Chr 2, n=80; Chr 3, n=99.

QTL mapping—The newly developed linkage map was used to detect and locate QTL underlying the DENV vector competence indices described above. Midgut infection was analysed in all F_2 females whereas viral dissemination was analysed only in midgut-infected females. The four different DENV isolates were included as a covariate to detect QTL x isolate interactions. Single QTL detection was performed in the R/qtl environment[85] using the EM algorithm of the *scanone* function using a binary trait model. Genome-wide statistical significance was determined by empirical permutation tests, with 1,000 genotype-phenotype permutations of the entire data set.

AaegL5 vs. AaegL3 comparison—To assess the improvement brought by AaegL5 to perform QTL mapping, a linkage map was built by aligning RAD markers to the AaegL3 assembly. The AaegL3-guided linkage map was built by assigning markers to chromosomes and by ordering them within each linkage group only based on their recombination fractions. Markers were initially filtered based on their segregation ratios as described above and assigned to the same linkage group based on a LOD score ≥14 threshold. Linkage groups were assigned to the three *Ae. aegypti* chromosomes using supercontigs that were previously mapped to the chromosomes[22]. For each linkage group, a backbone was created with a small subset of informative markers (n=6) using the *rf.2pts* two-point algorithm of the OneMap package. The remaining markers were positioned one at a time using the OneMap *order.seq* multi-point method, which compares all maps including the new marker at all possible positions keeping the original linkage map unchanged. This procedure produces both a “safe” and a “forced” marker order. The “forced” marker map indicates the most likely position for each marker, whereas the “safe” marker map only displays the unambiguously positioned markers. The AaegL3-guided QTL mapping was performed with the “safe” marker map. Strong bias in Mendelian segregation ratios of markers anchored to chromosome 1 impeded their ordering. Fully sex-linked markers lacked the F_0 paternal (AA) genotype in F_2 females, and segregated analogously to a backcross design in which F_1 AB heterozygotes are backcrossed to F_0 BB homozygotes. No linkage analysis method is readily available to deal with a chromosome that behaves like a mixture of intercross and backcross designs. Therefore, AaegL3-guided linkage analysis and QTL mapping for chromosome 1 were restricted to the fully sex-linked region based on a backcross design. A new OneMap input file only including markers lacking the F_0 paternal (AA) genotype was made by setting the population type to “backcross” instead of “F2 intercross”. Markers were ordered using the *order.seq* function of the OneMap package as described above. A table summarizing this comparison is available as Extended Data Fig. 10b.

Mapping insecticide resistance and VGSC.

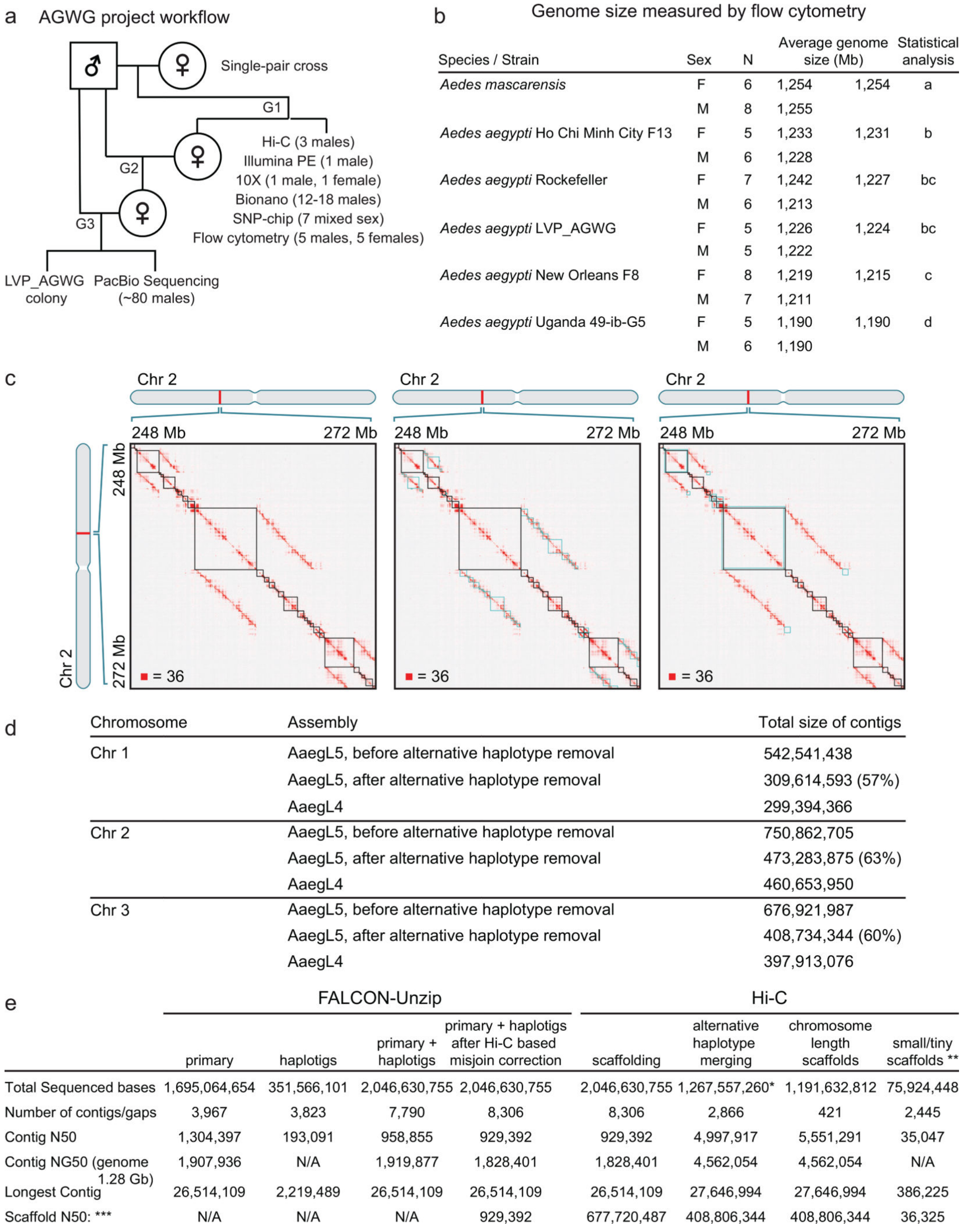
The mosquito population Viva Caucel from Yucatán State in Southern Mexico (Longitude −89.71827, Latitude 20.99827), was collected in 2011 by Universidad Autónoma de Yucatán. We identified up to 25 larval breeding sites from 3–4 city blocks and collected ~1000 larvae. Larvae were allowed to eclose, and twice a day we aspirated the adults from the cartons, discarding anything other than *Ae. aegypti*. 300–400 *Ae. aegypti* were released into a 2-foot cubic cage where they were allowed to mate for up to 5 days with *ad libitum* access to sucrose, after which they were blood fed to collect eggs for the next generation. 390 adult mosquitoes were then phenotyped for deltamethrin resistance. We exposed groups of 50 mosquitoes (3–4 days old) to 3 μg of deltamethrin-coated bottles for 1 hr. After this time, active mosquitoes were transferred to cardboard cups and placed into an incubator (28°C and 70% humidity) for 4 hr; these mosquitoes were referred as the resistant group. Knockdown mosquitoes were transferred to a second cardboard cup. After 4 hr, newly recovered mosquitoes were aspirated, frozen, and labelled as recovered; these were excluded from the current study. The mosquitoes that were knocked down and remained inactive at 4 hr post-treatment were scored as susceptible. DNA was isolated from individual mosquitoes by the salt extraction method[86] and resuspended in 150 μL of TE buffer (10 mM Tris-HCl, 1 mM EDTA pH 8.0). We constructed a total of four gDNA libraries. Two groups were pooled from DNA of 25 individual females that survived 1 hr of deltamethrin exposure (resistant replicates 1 and 2). The second set of two libraries was obtained by pooling DNA from 25 females that were knocked down and inactive at 4 hr 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 Inc.) 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 300–500 bp (Covaris Ltd., Brighton, U.K.). We prepared one library for each of the four DNA pools following the Low Sample (LS) protocol from the Illumina TrueSeqDNA PCR-Free Sample preparation guide (Illumina, San Diego CA). Because 65% of the *Ae. aegypti* 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 UTRs) in the AaegL1.3 genebuild using previously specified exonic coordinates[87]. In total, 26.7 Mb of the genome (2%) was targeted for enrichment. TruSeq libraries were hybridized to the probes using the xGen®Lock®Down 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 gDNA library were aligned to a custom reference physical map generated from the assembly AaegL5. Nucleotide counts were loaded into a contingency table with 4 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. Three (2 × c) contingency tables were subjected to χ^2 analyses (c-1 degrees of freedom) to determine if there are significant ($p \leq 0.05$) differences between 1) Alive replicates, 2) Dead replicates, and 3) Alive vs. Dead. If analysis 1) or 2) was significant, then that SNP locus was discarded. Otherwise the third contingency table consisted of 2 rows corresponding to Alive (sum of Reps 1 and 2), Dead (Reps 1 and 2 summed), and c columns. The χ^2 value from the (2 × c) contingency χ^2 analysis with (c-1) degrees of freedom was loaded into Excel to calculate the one-tailed probability of the χ^2 distribution probability (p). This value was transformed with $-\log_{10}(p)$. The experiment–wise error rate was then calculated following the method of Benjamini and Hochberg[88] to lower the number of Type I errors (false positives).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Extended Data



Genome size measured by flow cytometry

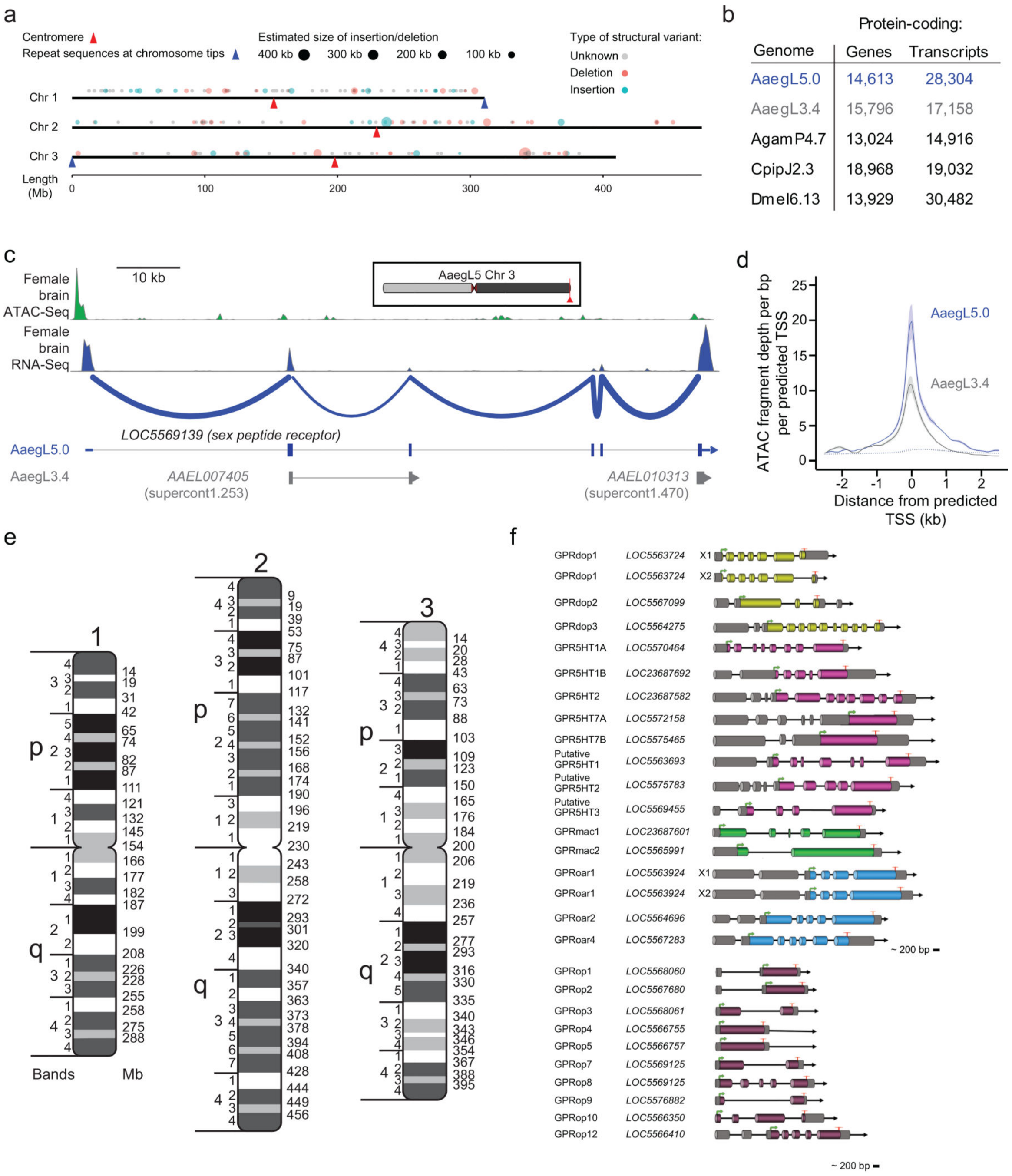
Species / Strain	Sex	N	Average genome size (Mb)		Statistical analysis
Aedes mascarensis	F	6	1,254	1,254	a
	M	8	1,255		
Aedes aegypti Ho Chi Minh City F13	F	5	1,233	1,231	b
	M	6	1,228		
Aedes aegypti Rockefeller	F	7	1,242	1,227	bc
	M	6	1,213		
Aedes aegypti LVP_AGWG	F	5	1,226	1,224	bc
	M	5	1,222		
Aedes 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		

Chromosome	Assembly	Total size of contigs
Chr 1	AaegL5, before alternative haplotype removal	542,541,438
	AaegL5, after alternative haplotype removal	309,614,593 (57%)
	AaegL4	299,394,366
Chr 2	AaegL5, before alternative haplotype removal	750,862,705
	AaegL5, after alternative haplotype removal	473,283,875 (63%)
	AaegL4	460,653,950
Chr 3	AaegL5, before alternative haplotype removal	676,921,987
	AaegL5, after alternative haplotype removal	408,734,344 (60%)
	AaegL4	397,913,076

	FALCON-Unzip				Hi-C			
	primary	haplotigs	primary + haplotigs	primary + haplotigs after Hi-C based misjoin correction	scaffolding	alternative haplotype merging	chromosome length scaffolds	small/tiny scaffolds **
Total Sequenced bases	1,695,064,654	351,566,101	2,046,630,755	2,046,630,755	2,046,630,755	1,267,557,260*	1,191,632,812	75,924,448
Number of contigs/gaps	3,967	3,823	7,790	8,306	8,306	2,866	421	2,445
Contig N50	1,304,397	193,091	958,855	929,392	929,392	4,997,917	5,551,291	35,047
Contig NG50 (genome 1.28 Gb)	1,907,936	N/A	1,919,877	1,828,401	1,828,401	4,562,054	4,562,054	N/A
Longest Contig	26,514,109	2,219,489	26,514,109	26,514,109	26,514,109	27,646,994	27,646,994	386,225
Scaffold N50: ***	N/A	N/A	N/A	929,392	677,720,487	408,806,344	408,806,344	36,325

Extended Data Figure 1 |. Project flowchart, measured genome size, and assembly process.

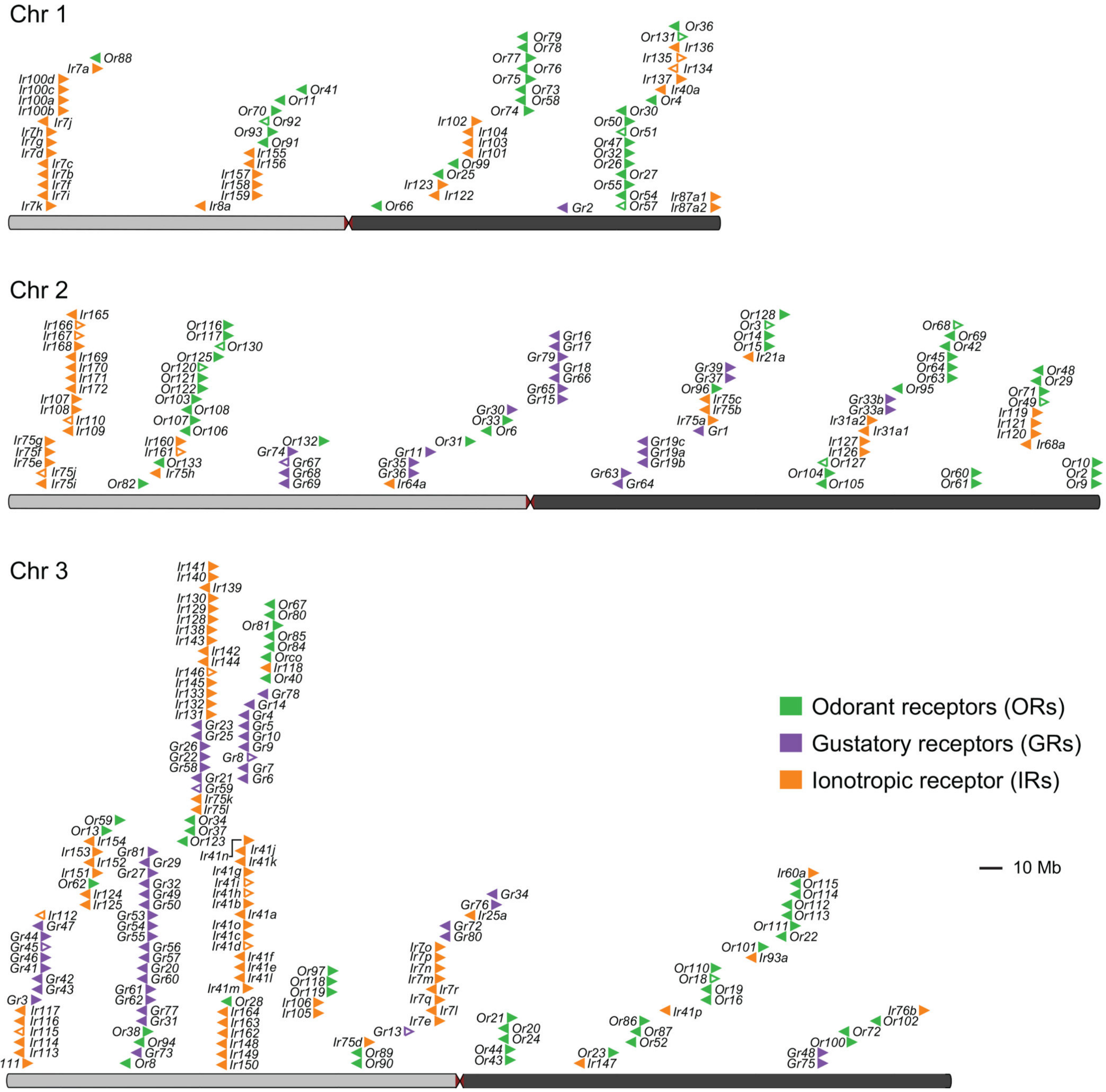
a, Flowchart of LVP_AGWG strain inbreeding, data collection, and experimental design of the AaegL5 assembly process. **b**, Estimated average 1C genome size for each strain for 5 *Ae. aegypti* strains and *Ae. mascarensis*, the sister taxon of *Ae. aegypti* whose genome size has not previously been measured. There were no significant differences between the sexes within and between the species/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 ($p < 0.01$). **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 x 24 Mb fragment of a contact map generated by aligning a Hi-C data set 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 pair 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. Thus, 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 due 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 (resp., below) the diagonal so that the overlapping sequences are horizontally (resp., 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 prior to 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. **See (ref. [6]) for definition of scaffold groups. ***Gaps between contigs were set to 500 bp for calculating scaffold statistics.



Extended Data Figure 2 |. Chromosome map and structural variants in the *Ae. aegypti* AaegL5 genome assembly.

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 size or type of SV 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 triangle) and telomere-associated repeat sequences (blue triangle) are indicated. Raw data available as

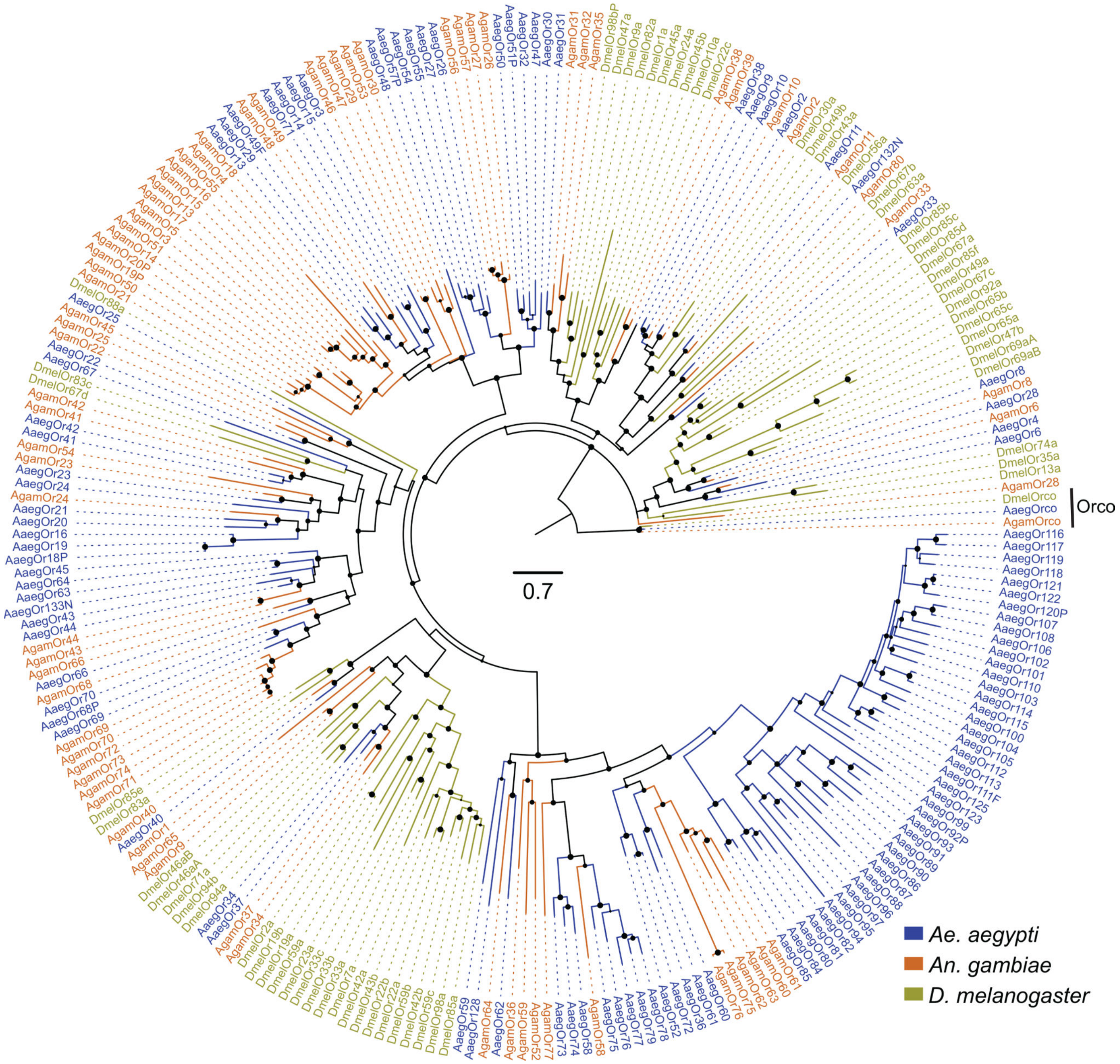
Supplementary Data 1. **b,** Comparison of protein-coding genes and transcripts in AaegL5.0 (NCBI RefSeq Release 101) and geneset annotations from *Anopheles gambiae* (*Agam*), *Culex pipiens* (*Cpip*), and *Drosophila melanogaster* (*Dmel*). **c,** *Sex peptide receptor* (*SPR*) structure in AaegL3.4 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 ± 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 standard deviation. **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); 5’ non-coding exons (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 Figure 3 |. Chromosomal arrangement of chemosensory receptor genes.

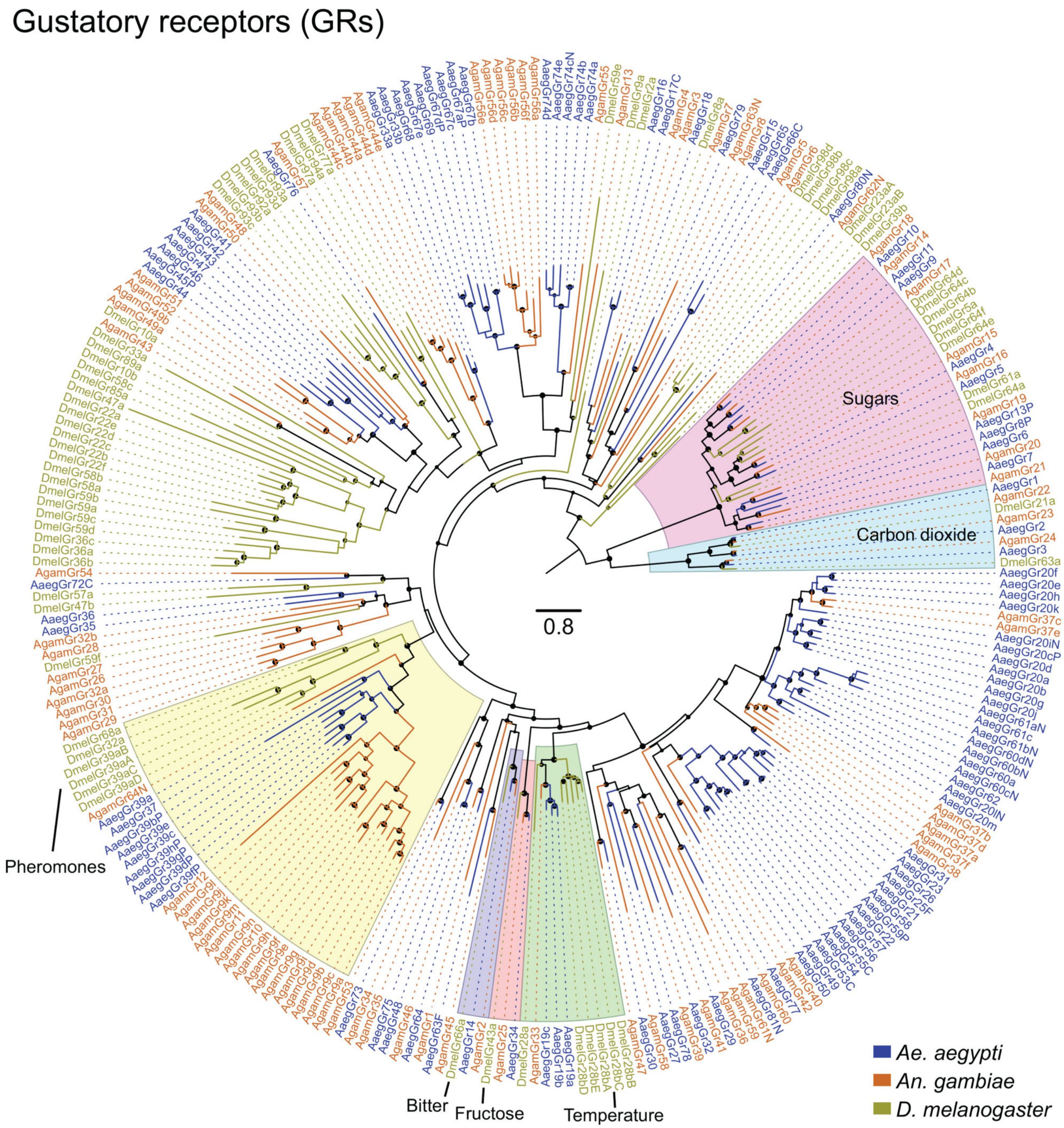
The location of predicted chemoreceptors (*ORs*, *GRs*, and *IRs*) across all three chromosomes in AaegL5. The blunt end of each arrowhead plotted above each chromosome marks gene position and arrowhead indicates orientation. Filled and open arrowheads represent intact genes and pseudogenes, respectively (Supplementary Data 17–20). This figure is identical to Fig. 2a, but here includes gene names.

Odorant receptors (ORs)



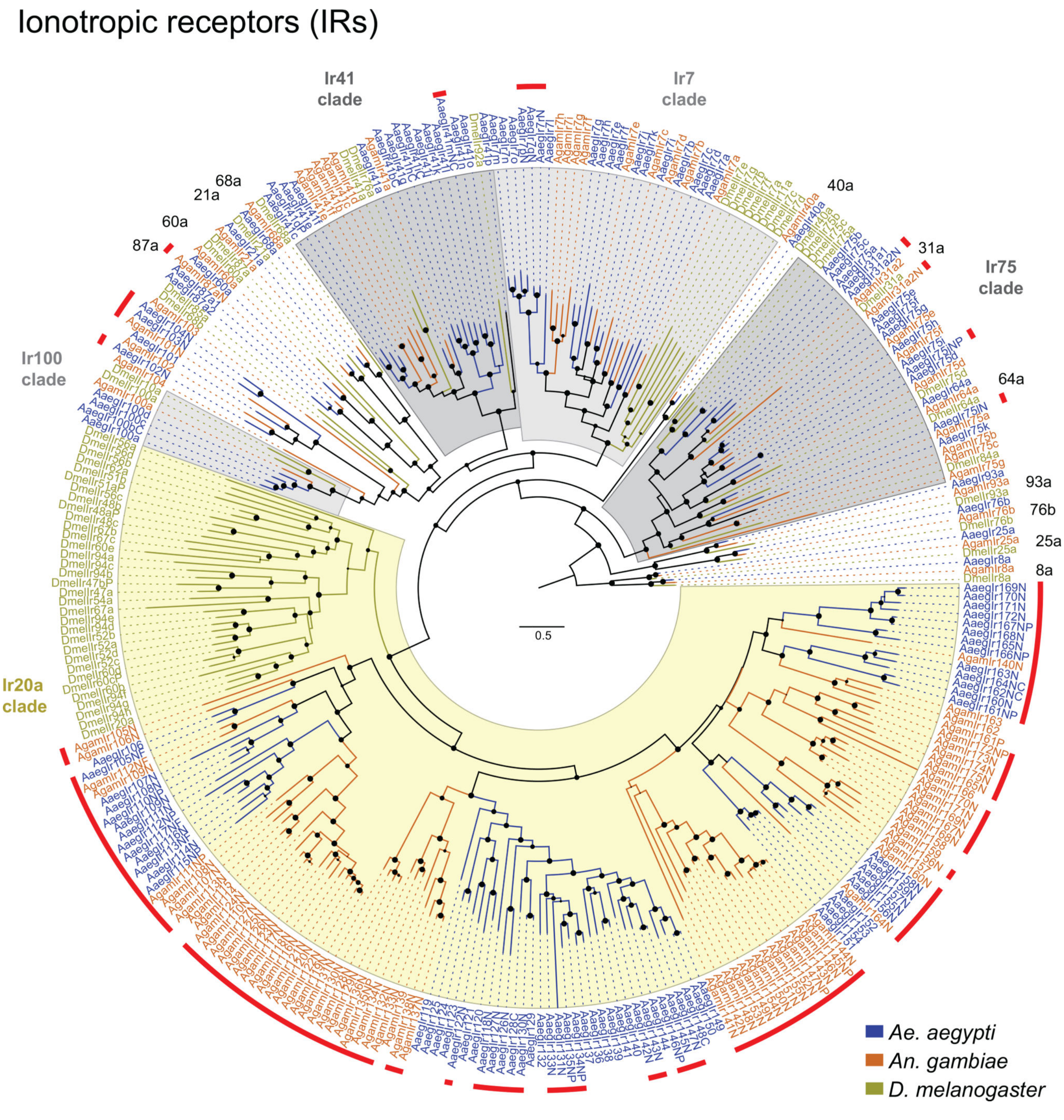
Extended Data Figure 4 |. Phylogenetic trees of odorant receptor (*OR*) gene families from *Ae. aegypti*, *An. gambiae*, and *D. melanogaster*.

Maximum likelihood OR tree was rooted with Orco proteins, which are both highly conserved and basal within the *OR* family[89]. Support levels for nodes are indicated by the size of black circles – reflecting approximate Log Ratio Tests (aLRT values ranging from 0–1 from PhyML v3.0 run with default parameters[90]). Suffixes after protein names are C – minor assembly correction, F – major assembly modification, N – new model, and P – pseudogene. Scale bar: amino acid substitutions per site.



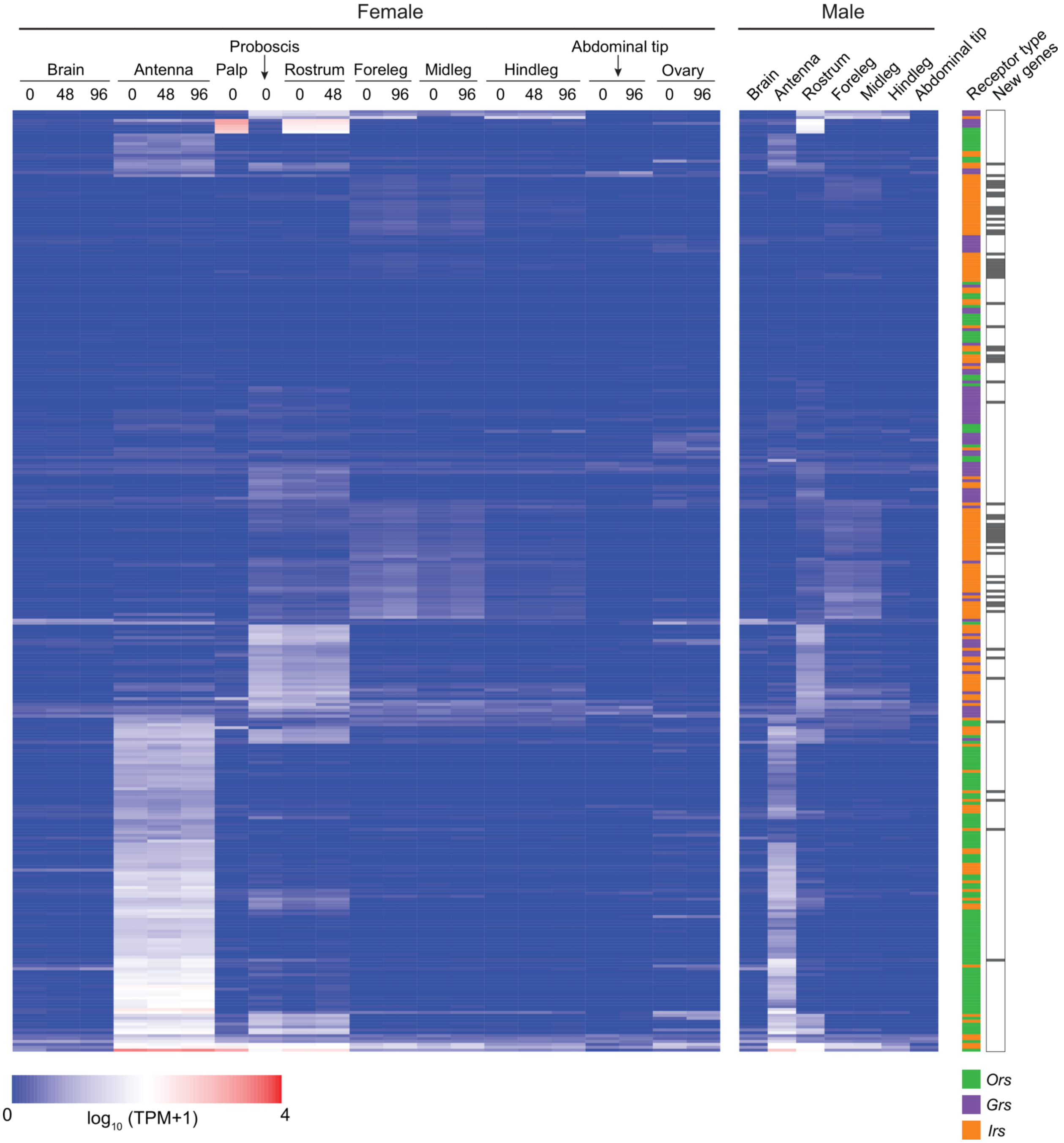
Extended Data Figure 5 |. Phylogenetic trees of the gustatory receptor (*GR*) gene families from *Ae. aegypti*, *An. gambiae*, and *D. melanogaster*.

Maximum likelihood *GR* 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 *GR* family[89]. Subfamilies and lineages closely related to *D. melanogaster GRs* of known function are highlighted. Support levels for nodes are indicated by the size of black circles – reflecting approximate Log Ratio Tests (aLRT values ranging from 0–1 from PhyML v3.0 run with default parameters[90]). Suffixes after protein names are C – minor assembly correction, F – major assembly modification, N – new model, and P – pseudogene. Scale bar: amino acid substitutions per site.



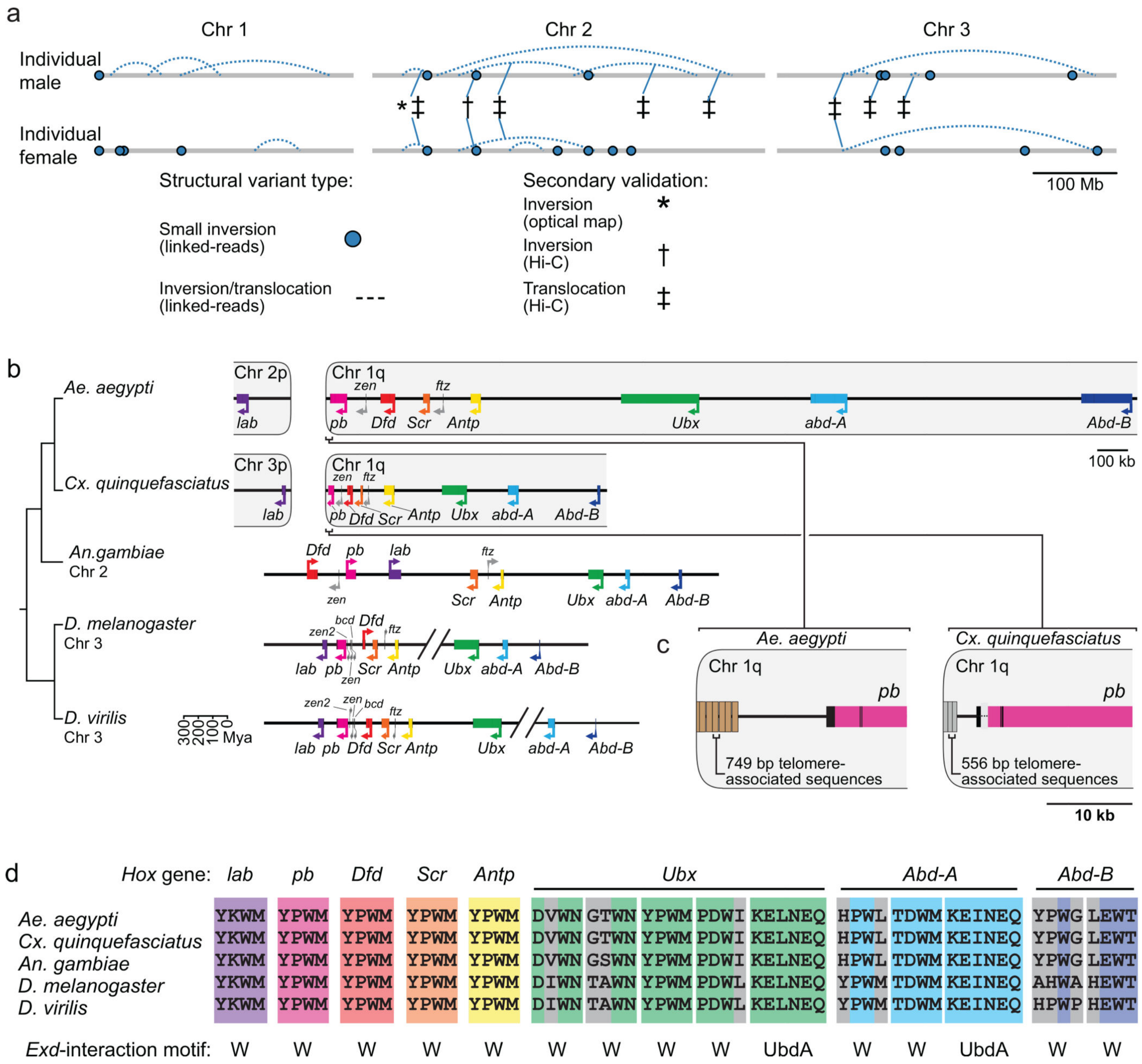
Extended Data Figure 6 |. Phylogenetic trees of the ionotropic receptor (*IR*) gene families from *Ae. aegypti*, *An. gambiae*, and *D. melanogaster*.

Maximum likelihood phylogenetic tree of IR 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 IRs are highlighted with red lines. Suffixes after protein names: C – minor assembly correction, F – major assembly modification, N – new model, and P – pseudogene. Scale bar: amino acid substitutions per site. Filled circles on nodes indicate support levels from approximate likelihood ratio tests from 0–1.



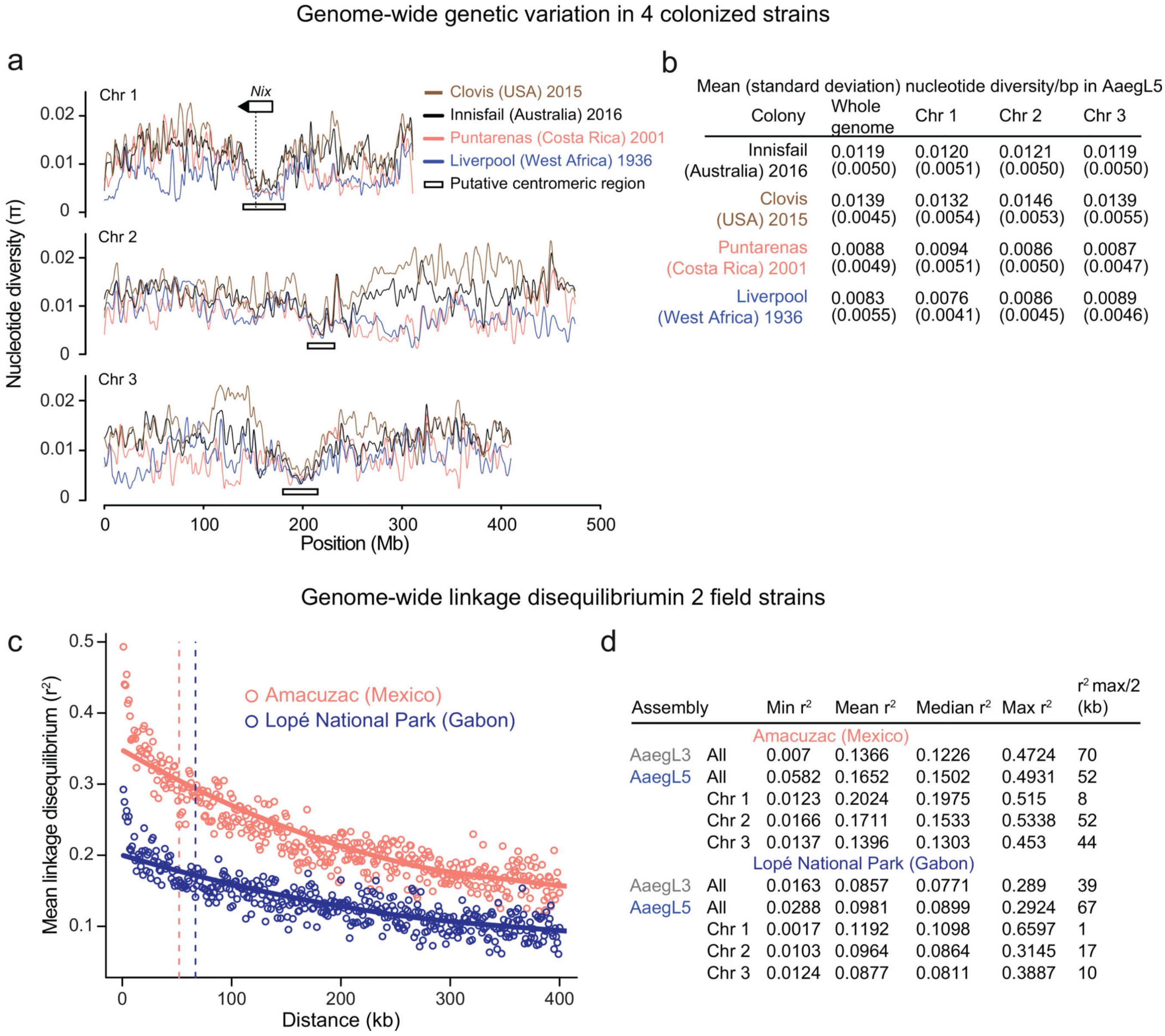
Extended Data Figure 7 |. Chemosensory receptor expression in adult *Ae. aegypti* tissues.

Previously published RNA-Seq data[13] 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 in 3 stages of the gonotrophic cycle (0, 48, or 96 hr after taking a blood-meal, where 0 hr indicates not blood-fed, 48 hr indicates 48 hr after the blood-meal, and 96 hr indicates gravid). New genes are indicated by black bars at right.


a
Chr 1
Chr 2
Chr 3
Individual male
Individual female
*‡ † ‡ ‡ ‡
‡ ‡ ‡
100 Mb
Structural variant type:
Small inversion (linked-reads)
Inversion/translocation (linked-reads)
Secondary validation:
Inversion (optical map) *
Inversion (Hi-C) †
Translocation (Hi-C) ‡
b
Ae. aegypti
Chr 2p
lab
Chr 1q
zen ftz
pb Dfd Scr Antp Ubx abd-A Abd-B
100 kb
Cx. quinquefasciatus
Chr 3p
lab
Chr 1q
zen ftz
pb Dfd Scr Antp Ubx abd-A Abd-B
An.gambiae
Chr 2
Dfd pb lab ftz
zen Scr Antp Ubx abd-A Abd-B
D. melanogaster
Chr 3
bcd Dfd zen2 ftz
lab pb zen Scr Antp Ubx abd-A Abd-B
D. virilis
Chr 3
zen2 zen bcd ftz
lab pb Dfd Scr Antp Ubx abd-A Abd-B
300 200 100 0 Mya
c
Ae. aegypti
Chr 1q
pb
749 bp telomere-associated sequences
Cx. quinquefasciatus
Chr 1q
pb
556 bp telomere-associated sequences
10 kb
d
Hox gene: lab pb Dfd Scr Antp Ubx Abd-A Abd-B
Ae. aegypti YKWM YPWM YPWM YPWM YPWM DVWN GTWN YPWM PDWI KELNEQ HPWL TDWM KEINEQ YPWG LEWT
Cx. quinquefasciatus YKWM YPWM YPWM YPWM YPWM DVWN GTWN YPWM PDWI KELNEQ HPWL TDWM KEINEQ YPWG LEWT
An. gambiae YKWM YPWM YPWM YPWM YPWM DVWN GSWN YPWM PDWI KELNEQ HPWL TDWM KEINEQ YPWG LEWT
D. melanogaster YKWM YPWM YPWM YPWM YPWM DIWN TAWN YPWM PDWL KELNEQ YPWM TDWM KEINEQ AHWA HEWT
D. virilis YKWM YPWM YPWM YPWM YPWM DIWN TAWN YPWM PDWL KELNEQ YPWM TDWM KEINEQ HPWP HEWT
Exd-interaction motif: W W W W W W W W W UbdA W W UbdA W W

Extended Data Figure 8 |. Structural variation 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 5 species (Supplementary Data 22). Due to chromosome arm exchange, Chr 3p in *Cx. quinquefasciatus* is the homologue of Chr 2p in *Ae. aegypti*[6]. **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)[91] 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.



Mean (standard deviation) nucleotide diversity/bp in AaegL5

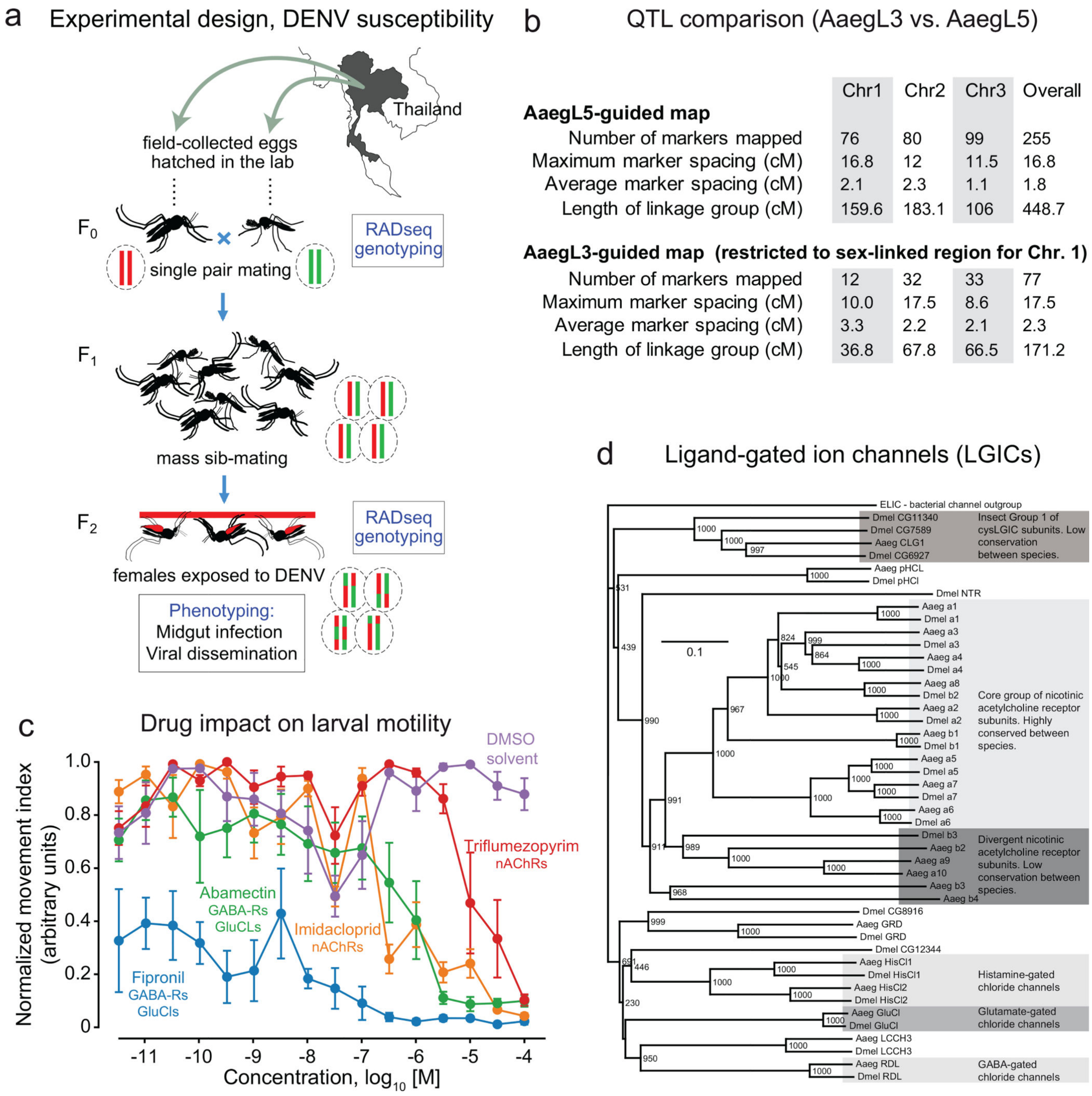
Colony	Whole genome	Chr 1	Chr 2	Chr 3
Innisfail (Australia) 2016	0.0119 (0.0050)	0.0120 (0.0051)	0.0121 (0.0050)	0.0119 (0.0050)
Clovis (USA) 2015	0.0139 (0.0045)	0.0132 (0.0054)	0.0146 (0.0053)	0.0139 (0.0055)
Puntarenas (Costa Rica) 2001	0.0088 (0.0049)	0.0094 (0.0051)	0.0086 (0.0050)	0.0087 (0.0047)
Liverpool (West Africa) 1936	0.0083 (0.0055)	0.0076 (0.0041)	0.0086 (0.0045)	0.0089 (0.0046)

Assembly		Min r^2	Mean r^2	Median r^2	Max r^2	r^2 max/2 (kb)
Amacuzac (Mexico)						
AaegL3	All	0.007	0.1366	0.1226	0.4724	70
AaegL5	All	0.0582	0.1652	0.1502	0.4931	52
	Chr 1	0.0123	0.2024	0.1975	0.515	8
	Chr 2	0.0166	0.1711	0.1533	0.5338	52
	Chr 3	0.0137	0.1396	0.1303	0.453	44
Lopé National Park (Gabon)						
AaegL3	All	0.0163	0.0857	0.0771	0.289	39
AaegL5	All	0.0288	0.0981	0.0899	0.2924	67
	Chr 1	0.0017	0.1192	0.1098	0.6597	1
	Chr 2	0.0103	0.0964	0.0864	0.3145	17
	Chr 3	0.0124	0.0877	0.0811	0.3887	10

Extended Data Figure 9 |. Population genomic structure and QTL analysis of *Ae. aegypti* strains

a, Chromosomal patterns of nucleotide diversity (**π**) in four strains of *Ae. aegypti* measured in 100 kb non-overlapping windows and presented as a LOESS-smoothed curve (Extended Data Fig. 9a-b). **b,** Mean nucleotide diversity in four colonized strains of *Ae. aegypti*, with standard deviation indicated in parentheses. Nucleotide diversity (**π**) 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 (LD) between

SNPs located within the same chromosome estimated from 28 wild-caught individuals from the indicated populations. Each point represents the mean LD for that set of binned SNP-pairs. Solid lines are LOESS-smoothed curves, and dashed lines correspond to $r^2_{max}/2$ (Extended Data Fig. 9a-b). Inclusion of additional individuals available from the Amacuzac population (up to 137) had a minimal effect on the LD estimations ($\Delta R^2 < 0.017$; data not shown). **d,** Table of 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.



	Chr1	Chr2	Chr3	Overall
AaegL5-guided map				
Number of markers mapped	76	80	99	255
Maximum marker spacing (cM)	16.8	12	11.5	16.8
Average marker spacing (cM)	2.1	2.3	1.1	1.8
Length of linkage group (cM)	159.6	183.1	106	448.7
AaegL3-guided map (restricted to sex-linked region for Chr. 1)				
Number of markers mapped	12	32	33	77
Maximum marker spacing (cM)	10.0	17.5	8.6	17.5
Average marker spacing (cM)	3.3	2.2	2.1	2.3
Length of linkage group (cM)	36.8	67.8	66.5	171.2

Extended Data Figure 10 |. QTL analysis of Dengue competence in *Ae. aegypti* and cys-loop ligand-gated ion channels

a, Schematic representation of the experimental workflow for testing Dengue viral 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 impact 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), Dß1 (CAA27641), Dß2 (CAA39211), Dß3 (CAC48166), GluCl (AAG40735), GRD (Q24352), HisCl1 (AAL74413), HisCl2 (AAL74414), LCCH3 (AAB27090), Ntr (NP_651958), pHCl (NP_001034025), RDL (AAA28556). 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.

Authors

Benjamin J. Matthews[1,2,3,*], Olga Dudchenko[4,5,6,7,*], Sarah B. Kingan[8,*], Sergey Koren[9], Igor Antoshechkin[10], Jacob E. Crawford[11], William J. Glassford[12], Margaret Herre[1,3], Seth N. Redmond[13,14], Noah H. Rose[15], Gareth D. Weedall[16,17], Yang Wu[18,19], Sanjit S. Batra[4,5,6], Carlos A. Brito-Sierra[20,21], Steven D. Buckingham[22], Corey L. Campbell[23], Saki Chan[24], Eric Cox[25], Benjamin R. Evans[26], Thanyalak Fansiri[27], Igor Filipović[28], Albin Fontaine[29,30,31,32], Andrea Gloria-Soria[26,33], Richard Hall[8], Vinita S. Joardar[25], Andrew K. Jones[34], Raissa G.G. Kay[35], Vamsi K. Kodali[25], Joyce Lee[24], Gareth J. Lycett[16], Sara N. Mitchell[11], Jill Muehling[8], Michael R. Murphy[25], Arina D. Omer[4,5,6], Frederick A. Partridge[22], Paul Peluso[8], Aviva Presser Aiden[4,5,36,37], Vidya Ramasamy[34], Gordana Rašić[28], Sourav Roy[38], Karla Saavedra-Rodriguez[23], Shruti Sharan[20,21], Atashi Sharma[39], Melissa Laird Smith[8], Joe Turner[40], Allison M. Weakley[11], Zhilei Zhao[15], Omar S. Akbari[41,42], William C. Black IV[23], Han Cao[24], Alistair C. Darby[40], Catherine A. Hill[20,21], J. Spencer Johnston[43], Terence D. Murphy[25], Alexander S. Raikhel[38], David B. Sattelle[22], Igor V. Sharakhov[39,44], Bradley J. White[11], Li Zhao[45], Erez Lieberman Aiden[4,5,6,7,13], Richard S. Mann[12], Louis Lambrechts[29,31], Jeffrey R. Powell[26], Maria V. Sharakhova[39,43], Zhijian Tu[19], Hugh M. Robertson[46], Carolyn S. McBride[15,47], Alex R. Hastie[24], Jonas Korlach[8], Daniel E. Neafsey[13,14], Adam M. Phillippy[9], and Leslie B. Vosshall[1,2,3]

Affiliations

[1]Laboratory of Neurogenetics and Behavior, The Rockefeller University, New York, New York, USA. [2]Howard Hughes Medical Institute, New York, New York, USA. [3]Kavli Neural Systems Institute, New York, New York, USA. [4]The Center for Genome Architecture, Baylor College of Medicine, Houston, Texas, USA. [5]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, USA. [6]Departments of Computer Science and Computational and Applied Mathematics, Rice University, Houston, Texas, USA. [7]Center for Theoretical and Biological Physics, Rice University, Houston, Texas, USA. [8]Pacific Biosciences, Menlo Park, California, USA. [9]National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, USA. [10]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California, USA. [11]Verily Life Sciences, South San Francisco, California, USA. [12]Mortimer B. Zuckerman Mind Brain Behavior Institute, Department of Biochemistry and Molecular

Biophysics, Columbia University, New York, New York, USA. [13]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [14]Department of Immunology and Infectious Disease, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA. [15]Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey, USA. [16]Vector Biology Department, Liverpool School of Tropical Medicine, Liverpool, United Kingdom. [17]Liverpool John Moores University, Liverpool, United Kingdom. [18]Department of Pathogen Biology, School of Public Health, Southern Medical University, Guangzhou, China. [19]Department of Biochemistry, Fralin Life Science Institute, Virginia Tech, Blacksburg, Virginia, USA. [20]Department of Entomology, Purdue University, West Lafayette, Indiana, USA. [21]Purdue Institute for Inflammation, Immunology and Infectious Disease, Purdue University, West Lafayette, Indiana, USA. [22]Centre for Respiratory Biology, UCL Respiratory, University College London, London, United Kingdom. [23]Department of Microbiology, Immunology and Pathology, Colorado State University, Fort Collins, Colorado, USA. [24]Bionano Genomics, San Diego, California, USA. [25]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA. [26]Department of Ecology and Evolutionary Biology, Yale University, New Haven, Connecticut, USA. [27]Vector Biology and Control Section, Department of Entomology, Armed Forces Research Institute of Medical Sciences (AFRIMS), Bangkok, Thailand. [28]Mosquito Control Laboratory, QIMR Berghofer Medical Research Institute, Brisbane, Australia. [29]Insect-Virus Interactions Group, Department of Genomes and Genetics, Institut Pasteur, Paris, France. [30]Unité de Parasitologie et Entomologie, Département des Maladies Infectieuses, Institut de Recherche Biomédicale des Armées, Marseille, France. [31]Centre National de la Recherche Scientifique, Unité Mixte de Recherche 2000, Paris, France. [32]Aix Marseille Université, IRD, AP-HM, SSA, UMR Vecteurs – Infections Tropicales et Méditerranéennes (VITROME), IHU - Méditerranée Infection, Marseille, France [33]The Connecticut Agricultural Experiment Station, New Haven, CT 06504 USA [34]Faculty of Health and Life Sciences, Department of Biological and Medical Sciences, Oxford Brookes University, Oxford, United Kingdom. [35]Department of Entomology, University of California Riverside, Riverside, California, USA. [36]Department of Bioengineering, Rice University, Houston, Texas, USA. [37]Department of Pediatrics, Texas Children's Hospital, Houston, Texas, USA. [38]Department of Entomology, Center for Disease Vector Research and Institute for Integrative Genome Biology, University of California, Riverside, California, USA. [39]Department of Entomology, Fralin Life Science Institute, Virginia Tech, Blacksburg, Virginia, USA. [40]Institute of Integrative Biology, University of Liverpool, Liverpool, United Kingdom. [41]Division of Biological Sciences, University of California, San Diego, La Jolla, California, USA. [42]Tata Institute for Genetics and Society, University of California, San Diego, La Jolla, California, USA [43]Department of Entomology, Texas A&M University, College Station, Texas, USA. [44]Laboratory of Ecology, Genetics, and Environmental Protection, Tomsk State University, Tomsk, Russia. [45]Laboratory of Evolutionary Genetics and Genomics, The Rockefeller University, New York, New York, USA. [46]Department of Entomology, University of Illinois at

Urbana-Champaign, Urbana, Illinois, USA. [47]Princeton Neuroscience Institute, Princeton University, Princeton, New Jersey, USA.

Acknowledgements

We thank Raul Andino; Scott Emrich and Daniel Lawson (Vectorbase), Anthony A. James, Mark Kunitomi, Chad Nusbaum, Dave Severson, Noah Whiteman; Todd Dickinson, Margot Hartley, and Brandon Rice (Dovetail Genomics) for early participation in the AGWG; Cori Bargmann, David Botstein, Erich Jarvis, and Eric Lander for encouragement and facilitation. Nikka Keivanfar, David Jaffe, and Deanna M. Church (10X Genomics) prepared DNA for SV analysis. We thank Amy Harmon of the New York Times and acknowledge generous pro bono data and analysis from our corporate collaborators.

Funding: This research was supported in part by federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Grant Number U19AI110818 to the Broad Institute (S.N.R. and D.E.N.); USDA 2017–05741 (E.L.A.); NSF PHY-1427654 Center for Theoretical Biological Physics (E.L.A.); NIH Intramural Research Program, National Library of Medicine, and National Human Genome Research Institute (A.M.P and S.K.), and the following extramural NIH grants: RO1AI101112 (J.R.P), R35GM118336 (R.S.M., W.J.G.), R21AI121853 (M.V.S., I.V.S., A.S.), RO1AI123338 (Z.T.), T32GM007739 (M.H.), NIH/NCATS UL1TR000043 (Rockefeller University), DP2OD008540 (E.L.A), U01HL130010 (E.L.A.), UM1HG009375 (E.L.A), 5K22AI113060 (O.S.A), 1R21AI123937 (O.S.A); Defence Advanced Research Project Agency: HR0011–17-2– 0047 (O.S.A). Other support: Jane Coffin Childs Memorial Fund (B.J.M), Center for Theoretical Biological Physics postdoctoral fellowship (O.D.), Robertson Foundation (L.Z.), and McNair & Welch (Q-1866) Foundations (E.L.A.), French Government's Investissement d'Avenir program, Laboratoire d'Excellence Integrative Biology of Emerging Infectious Diseases (grant ANR-10-LABX-62-IBEID)(L.L.), Agence Nationale de la Recherche grant ANR-17-ERC2–0016-01 (L.L.), European Union's Horizon 2020 research and innovation program under ZikaPLAN grant agreement No. 734584 (L.L.). A.M.W., B.J.W., J.E.C., and S.N.M. were supported by Verily Life Sciences. L.B.V. is an investigator of the Howard Hughes Medical Institute.

References

1. Bhatt S et al. The global distribution and burden of dengue. Nature 496, 504–507, (2013). [PubMed: 23563266]
2. Nene V et al. Genome sequence of *Aedes aegypti*, a major arbovirus vector. Science 316, 1718–1723, (2007). [PubMed: 17510324]
3. Timoshevskiy VA et al. An integrated linkage, chromosome, and genome map for the yellow fever mosquito *Aedes aegypti*. PLoS Negl Trop Dis 7, e2052, (2013). [PubMed: 23459230]
4. Chin CS et al. Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods 13, 1050–1054, (2016). [PubMed: 27749838]
5. Waterhouse RM et al. BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol, doi: 10.1093/molbev/msx1319, (2017).
6. Dudchenko O et al. *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science 356, 92–95, (2017). [PubMed: 28336562]
7. Denny SK et al. Nfib promotes metastasis through a widespread increase in chromatin accessibility. Cell 166, 328–342, (2016). [PubMed: 27374332]
8. Timoshevskiy VA et al. Genomic composition and evolution of *Aedes aegypti* chromosomes revealed by the analysis of physically mapped supercontigs. BMC Biol 12, 27, (2014). [PubMed: 24731704]
9. George P, Sharakhova MV & Sharakhov IV High-resolution cytogenetic map for the African malaria vector *Anopheles gambiae*. Insect Mol Biol 19, 675–682, (2010). [PubMed: 20609021]
10. Artemov GN et al. The physical genome mapping of *Anopheles albimanus* corrected scaffold misassemblies and identified interarm rearrangements in genus *Anopheles*. G3 (Bethesda) 7, 155–164, (2017). [PubMed: 27821634]
11. Gorman MJ & Paskewitz SM Serine proteases as mediators of mosquito immune responses. Insect Biochem Mol Biol 31, 257–262, (2001). [PubMed: 11167095]

12. Goulielmaki E, Siden-Kiamos I & Loukeris TG Functional characterization of *Anopheles matrix metalloprotease 1* reveals its agonistic role during sporogonic development of malaria parasites. Infect Immun 82, 4865–4877, (2014). [PubMed: 25183733]

13. Matthews BJ, McBride CS, DeGennaro M, Despo O & Vosshall LB The neurotranscriptome of the *Aedes aegypti* mosquito. BMC Genomics 17, 32, (2016). [PubMed: 26738925]

14. Gilchrist BM & Haldane JBS Sex linkage and sex determination in a mosquito, *Culex molestus*. Hereditas 33, 175–190, (1947).

15. McClelland GAH Sex-linkage in Aedes aegypti. Trans. Roy. Soc. Tropical Med. Hyg 56, (1962).

16. Newton ME, Wood RJ & Southern DI Cytological mapping of the M and D loci in the mosquito, Aedes aegypti (L.). Genetica 48, 137–143, (1978).

17. Hall AB et al. Sex determination. A male-determining factor in the mosquito *Aedes aegypti*. Science 348, 1268–1270, (2015). [PubMed: 25999371]

18. Hall AB et al. Insights into the preservation of the homomorphic sex-determining chromosome of *Aedes aegypti* from the discovery of a male-biased gene tightly linked to the M-locus. Genome Biol Evol 6, 179–191, (2014). [PubMed: 24398378]

19. Turner J et al. The sequence of a male-specific genome region containing the sex determination switch in Aedes aegypti. bioRxiv, doi: 10.1101/122804, (2017).

20. Hall AB et al. Six novel Y chromosome genes in *Anopheles* mosquitoes discovered by independently sequencing males and females. BMC Genomics 14, 273, (2013). [PubMed: 23617698]

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). [PubMed: 28945882]

22. Juneja P et al. Assembly of the genome 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). [PubMed: 24498447]

23. Charlesworth D, Charlesworth B & Marais G Steps in the evolution of heteromorphic sex chromosomes. Heredity (Edinb) 95, 118–128, (2005). [PubMed: 15931241]

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

25. Lewis EB A gene complex controlling segmentation in *Drosophila*. Nature 276, 565–570, (1978). [PubMed: 103000]

26. Duboule D The rise and fall of *Hox* gene clusters. Development 134, 2549–2560, (2007). [PubMed: 17553908]

27. Negre B, Ranz JM, Casals F, Caceres M & Ruiz A A new split of the *Hox* gene complex in *Drosophila*: relocation and evolution of the gene *labial*. Mol Biol Evol 20, 2042–2054, (2003). [PubMed: 12949134]

28. Enayati AA, Ranson H & Hemingway J Insect glutathione transferases and insecticide resistance. Insect Mol Biol 14, 3–8, (2005). [PubMed: 15663770]

29. Bass C & Field LM Gene amplification and insecticide resistance. Pest Manag Sci 67, 886–890, (2011). [PubMed: 21538802]

30. Ortelli F, Rossiter LC, Vontas J, Ranson H & Hemingway J Heterologous expression of four glutathione transferase genes genetically linked to a major insecticide-resistance locus from the malaria vector *Anopheles gambiae*. Biochem J 373, 957–963, (2003). [PubMed: 12718742]

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). [PubMed: 21195177]

32. Anopheles gambiae 1000 Genomes Consortium. Genetic diversity of the African malaria vector *Anopheles gambiae*. Nature 552, 96–100, (2017). [PubMed: 29186111]

33. Begun DJ & Aquadro CF Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. Nature 356, 519–520, (1992). [PubMed: 1560824]

34. Evans BR 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). [PubMed: 25721127]

35. Fansiri T et al. Genetic mapping of specific interactions between Aedes aegypti mosquitoes and dengue viruses. PLoS Genet 9, e1003621, (2013). [PubMed: 23935524]

36. Black W. C. t. et al. Flavivirus susceptibility in *Aedes aegypti*. Arch Med Res 33, 379–388, (2002). [PubMed: 12234528]

37. Moyes CL et al. Contemporary status of insecticide resistance in the major Aedes vectors of arboviruses infecting humans. PLoS Negl Trop Dis 11, e0005625, (2017). [PubMed: 28727779]

38. Jones AK & Sattelle DB Diversity of insect nicotinic acetylcholine receptor subunits. Adv Exp Med Biol 683, 25–43, (2010). [PubMed: 20737786]

39. Alphey L Genetic control of mosquitoes. Annu Rev Entomol 59, 205–224, (2014). [PubMed: 24160434]

40. Adelman ZN & Tu Z Control of mosquito-borne infectious disease: Sex and gene drive. Trends Parasitol 32, 219–229, (2016). [PubMed: 26897660]

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. R Foundation for Statistical Computing, Vienna, Austria https://www.R-project.org/ (2017).

43. Hare EE & Johnston JS Genome size determination using flow cytometry of propidium iodide-stained nuclei. Methods Mol Biol 772, 3–12, (2011). [PubMed: 22065429]

44. Galbraith DW et al. Rapid flow cytometric analysis of the cell cycle in intact plant tissues. Science 220, 1049–1051, (1983). [PubMed: 17754551]

45. Chin CS et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods 10, 563–569, (2013). [PubMed: 23644548]

46. Chaisson MJ & Tesler G Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics 13, 238, (2012). [PubMed: 22988817]

47. Rao SS et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159, 1665–1680, (2014). [PubMed: 25497547]

48. Durand NC et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst 3, 95–98, (2016). [PubMed: 27467249]

49. Dudchenko O et al. The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. bioRxiv, https://www.biorxiv.org/content/early/2018/2001/2028/254797, (2018).

50. English AC et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS One 7, e47768, (2012). [PubMed: 23185243]

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

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

53. Smit AFA, Hubley R & Green P RepeatMasker Open-4.0 2013-2015 http://www.repeatmasker.org.

54. Smit AFA & Hubley R RepeatModeler Open-1.0. 2008-2015 http://www.repeatmasker.org.

55. Benson G Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27, 573–580, (1999). [PubMed: 9862982]

56. Thibaud-Nissen F, Souvorov A, Murphy T, DiCuccio M & Kitts P Eukaryotic Genome Annotation Pipeline. The NCBI Handbook [Internet]. 2nd edition. http://www.ncbi.nlm.nih.gov/books/NBK169439/, (2013).

57. Akbari OS et al. The developmental transcriptome of the mosquito Aedes aegypti, an invasive species and major arbovirus vector. G3 (Bethesda) 3, 1493–1509, (2013). [PubMed: 23833213]

58. Dobin A et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21, (2013). [PubMed: 23104886]

59. Liao Y, Smyth GK & Shi W featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30, 923–930, (2014). [PubMed: 24227677]

60. Patro R, Duggal G, Love MI, Irizarry RA & Kingsford C Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods 14, 417–419, (2017). [PubMed: 28263959]

61. Buenrostro JD, Giresi PG, Zaba LC, Chang HY & Greenleaf WJ 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). [PubMed: 24097267]

62. Langmead B, Trapnell C, Pop M & Salzberg SL Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10, R25, (2009). [PubMed: 19261174]

63. Li H et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079, (2009). [PubMed: 19505943]

64. Altschul SF, Gish W, Miller W, Myers EW & Lipman DJ Basic local alignment search tool. J Mol Biol 215, 403–410, (1990). [PubMed: 2231712]

65. Wu TD & Watanabe CK GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21, 1859–1875, (2005). [PubMed: 15728110]

66. Hahne F & Ivanek R Visualizing Genomic Data Using Gviz and Bioconductor. Methods Mol Biol 1418, 335–351, (2016). [PubMed: 27008022]

67. 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). [PubMed: 20513432]

68. Kielbasa SM, Wan R, Sato K, Horton P & Frith MC Adaptive seeds tame genomic sequence comparison. Genome Res 21, 487–493, (2011). [PubMed: 21209072]

69. Neafsey DE et al. Mosquito genomics. Highly evolvable malaria vectors: the genomes of 16 *Anopheles* mosquitoes. Science 347, 1258522, (2015). [PubMed: 25554792]

70. Timoshevskiy VA, Sharma A, Sharakhov IV & Sharakhova MV Fluorescent in situ hybridization on mitotic chromosomes of mosquitoes. J Vis Exp, e4215, (2012). [PubMed: 23007640]

71. Sharakhova MV et al. Imaginal discs--a new source of chromosomes for genome mapping of the yellow fever mosquito *Aedes aegypti*. PLoS Negl Trop Dis 5, e1335, (2011). [PubMed: 21991400]

72. Jimenez LV, Kang BK, deBruyn B, Lovin DD & Severson DW Characterization of an Aedes aegypti bacterial artificial chromosome (BAC) library and chromosomal assignment of BAC clones for physical mapping quantitative trait loci that influence *Plasmodium* susceptibility. Insect Mol Biol 13, 37–44, (2004). [PubMed: 14728665]

73. Langmead B & Salzberg SL Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357–359, (2012). [PubMed: 22388286]

74. Liao Y, Smyth GK & Shi W The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Res 41, e108, (2013). [PubMed: 23558742]

75. Krumsiek J, Arnold R & Rattei T Gepard: a rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics 23, 1026–1028, (2007). [PubMed: 17309896]

76. Apostol BL, Black W. C. t., Reiter P & Miller BR Use of randomly amplified polymorphic DNA amplified by polymerase chain reaction markers to estimate the number of *Aedes aegypti* families at oviposition sites in San Juan, Puerto Rico. Am J Trop Med Hyg 51, 89–97, (1994). [PubMed: 8059920]

77. Rasic G et al. The queenslandensis and the type form of the dengue fever mosquito (Aedes aegypti L.) are genomically indistinguishable. PLoS Negl Trop Dis 10, e0005096, (2016). [PubMed: 27806047]

78. Thomas SJ et al. Dengue plaque reduction neutralization test (PRNT) in primary and secondary dengue virus infections: How alterations in assay conditions impact performance. Am J Trop Med Hyg 81, 825–833, (2009). [PubMed: 19861618]

79. Peterson BK, Weber JN, Kay EH, Fisher HS & Hoekstra HE Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS One 7, e37135, (2012). [PubMed: 22675423]

80. Rasic G, Filipovic I, Weeks AR & Hoffmann AA Genome-wide SNPs lead to strong signals of geographic structure and relatedness patterns in the major arbovirus vector, *Aedes aegypti*. BMC Genomics 15, 275, (2014). [PubMed: 24726019]

81. Catchen JM, Amores A, Hohenlohe P, Cresko W & Postlethwait JH Stacks: building and genotyping loci de novo from short-read sequences. G3 (Bethesda) 1, 171–182, (2011). [PubMed: 22384329]

82. Catchen J, Hohenlohe PA, Bassham S, Amores A & Cresko WA Stacks: an analysis tool set for population genomics. Mol Ecol 22, 3124–3140, (2013). [PubMed: 23701397]

83. Margarido GR, Souza AP & Garcia AA OneMap: software for genetic mapping in outcrossing species. Hereditas 144, 78–79, (2007). [PubMed: 17663699]

84. Kosambi DD The Estimation of Map Distances from Recombination Values, Ramaswamy R *(ed)* (Springer, 2016).

85. Broman KW, Wu H, Sen S & Churchill GA R/qtl: QTL mapping in experimental crosses. Bioinformatics 19, 889–890, (2003). [PubMed: 12724300]

86. Black WC & DuTeau NM RAPD-PCR and SSCP analysis for insect population genetic studies. *In:* Crampton JM, Beard CB, Louis C *(eds)* The Molecular Biology of Insect Disease Vectors. Springer, Dordrecht (1997).

87. Juneja P et al. Exome and transcriptome sequencing of *Aedes aegypti* identifies a locus that confers resistance to *Brugia malayi* and alters the immune response. PLoS Pathog 11, e1004765, (2015). [PubMed: 25815506]

88. Benjamini Y & Hochberg Y Controlling the false discovery rate: A practical and powerful approach to multiple testing. J Roy Stat Soc B Met 57, 289–300, (1995).

89. Robertson HM The insect chemoreceptor superfamily is ancient in animals. Chem Senses 40, 609–614, (2015). [PubMed: 26354932]

90. Guindon S et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59, 307–321, (2010). [PubMed: 20525638]

91. Merabet S & Mann RS To be specific or not: The critical relationship between HOS and TALE proteins. Trends Genet 32, 334–347, (2016). [PubMed: 27066866]

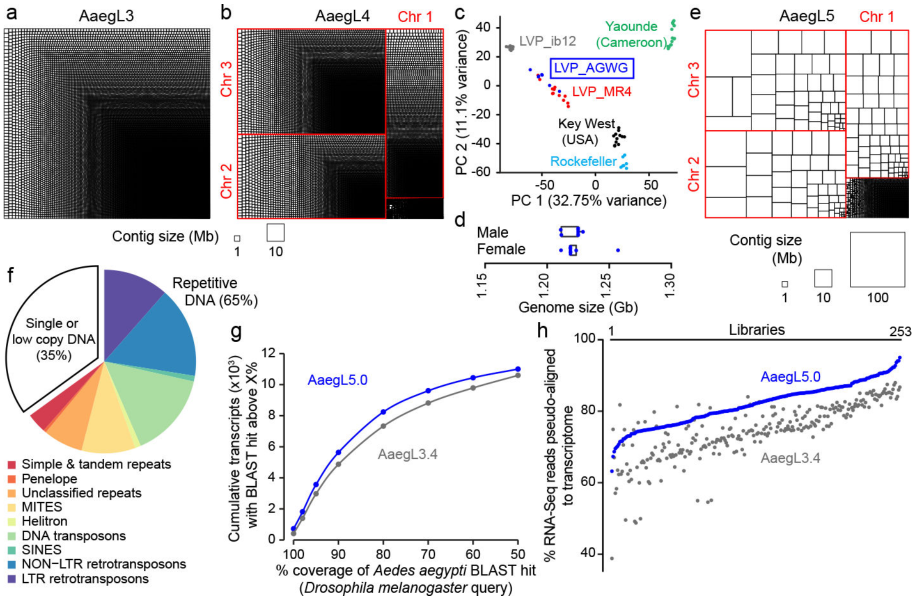


Figure 1 |. AaegL5 assembly statistics, annotation, and chromatin accessibility analysis.

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. **d,** Flow cytometry analysis of LVP_AGWG genome size. Box plot: median= blue line, boxes 1st/3rd quartile, whisker 1.5X interquartile interval (Extended Data Fig. 1b). **e,** Treemap of AaegL5 contigs scaled by length. **f,** Genome composition (Supplementary Data 2–3). **g,** AaegL3.4 vs. AaegL5.0 geneset alignment BLASTp coverage with *D. melanogaster* protein queries. **h,** Alignment of 253 RNA-Seq libraries to AaegL3.4 and AaegL5.0 geneset annotations (Supplementary Data 4–9).

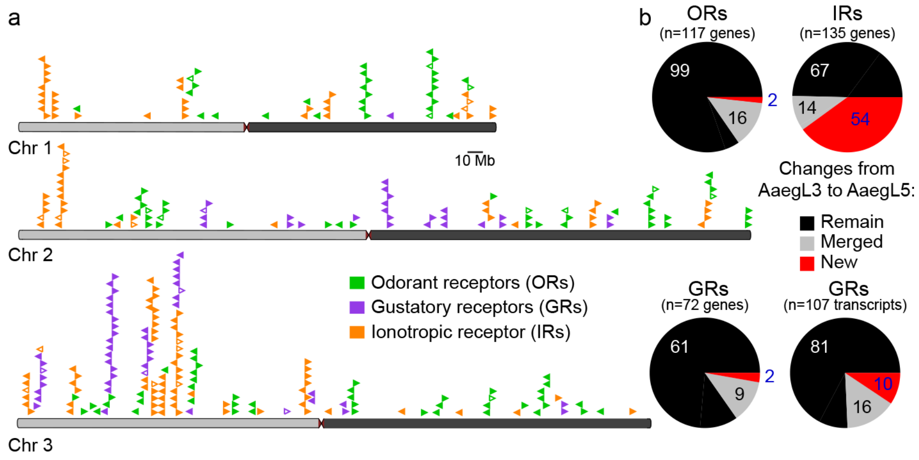


Figure 2 |. Chromosomal arrangement and increased number of chemosensory receptor genes.

a, Location of predicted chemoreceptors (*ORs*, *GRs*, and *IRs*) by chromosome in AaegL5. Blunt end of arrowhead marks gene position and arrow indicates orientation. Filled and open arrowheads represent intact genes and pseudogenes, respectively (Supplementary Data 17-20 and Extended Data Fig. 3).**b,** AaegL5 vs. AaegL3 chemosensory receptor annotation.

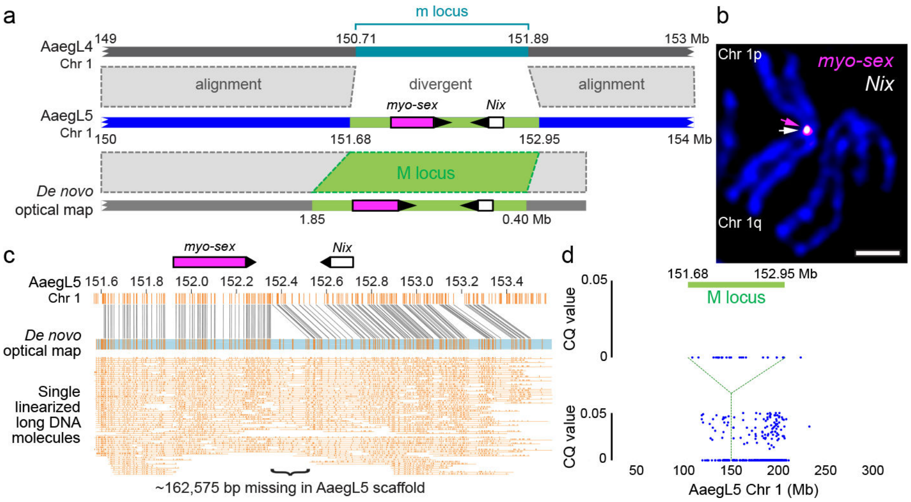


Figure 3 |. Application of AaegL5 to resolve the sex-determining locus and the HOX gene cluster.

a, M locus structure indicating high alignment identity (grey dashed boxes) and boundaries of *myo-sex* and *Nix* gene models (magenta and white boxes, arrowhead represents 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 edges for clarity. **d,** Chromosome-quotient (CQ) analysis of genomic DNA from pure male and female libraries aligned to AaegL5 chromosome 1. Each dot represents the CQ value of a non-repeat-masked 1 kb window with >20 reads aligned from male libraries.

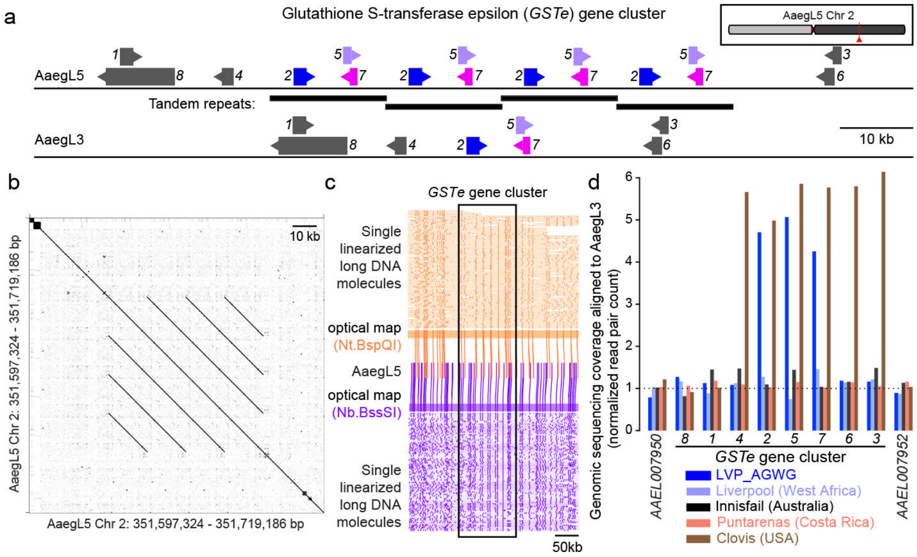


Figure 4 |. Copy-number variation in the glutathione S-transferase epsilon gene cluster.

a, Glutathione S-transferase epsilon (*GSTe*) gene cluster structure in AaegL5 vs. AaegL3 (Supplementary Data 23). Arrowheads indicate gene orientation. **b,** Dot-plot alignment of AaegL5 *GSTe* region to itself. **c,** Optical mapping of DNA labelled with indicated enzymes. DNA molecules are cropped at edges for clarity. **d,** Genomic sequencing coverage of AaegL3 *GSTe* genes (DNA read pairs mapped to each gene, normalized by gene length in kb) from one LVP_AGWG male and pooled mosquitoes from 4 other laboratory strains.

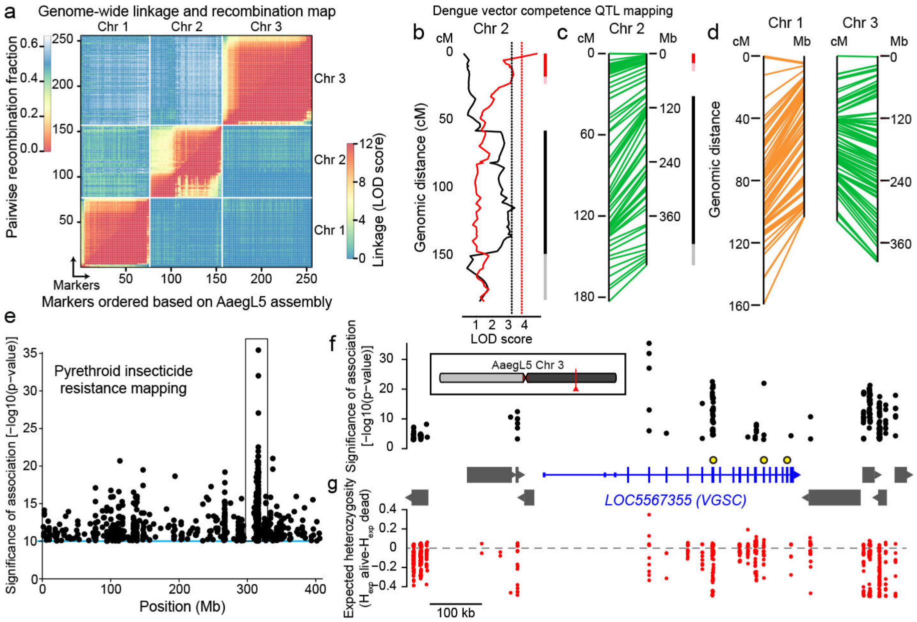


Figure 5 |. Deploying the AaegL5 genome for applied population genetics.

a, Heat map of linkage based on pairwise recombination fractions for 255 RAD markers ordered by AaegL5 physical coordinates. **b,** Significant chromosome 2 QTL underlying systemic DENV dissemination in midgut-infected mosquitoes (Extended Data Fig. 10a). Curves represent log of the odds ratio (LOD) scores obtained by interval mapping. Dotted vertical lines indicate genome-wide statistical significance thresholds (α=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**). **e,** Chromosome 3 SNPs significantly correlated with deltamethrin survival. **f-g,** Zoomed in and inverted view of box in (**e**) centred on new gene model of voltage-gated sodium channel (*VGSC*, transcript variant X3, chromosomal position indicated in red). (**f**). Non-coding genes are omitted for clarity, and other genes indicated with grey boxes. *VGSC* exons are represented by tall boxes and UTRs by short boxes. Arrowheads indicate gene orientation. Non-synonymous *VGSC* SNPs marked with larger black and yellow circles: (V1016I = 315,983,763, F1534C= 315,939,224, V410L = 316,080,722). Difference in expected heterozygosity (H_{exp} alive – H_{exp} dead) for all SNPs (**g**).

Table 1 |

Comparison of assembly statistics

Genome:	AaegL3	AaegL4	AaegL5	AaegL5 (NCBI)
			FALCON-Unzip	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
Contig number	36,205	37,224	3,967	2,539
Contig N50 (bp)	82,618	84,074	1,304,397	11,758,062
Contig NG50 (bp)	85,043	81,911	1,907,936	11,758,062
Scaffold number	4,757	6,206	(N/A)	2,310
Scaffold N50 (bp)	1,547,048	404,248,146 *	(N/A)	409,777,670 *
GC content (%)	38.27	38.28	38.16	38.18
Alternative	(N/A)	(N/A)	351,566,101	591,941,260
haplotypes (bp) Alternative	(N/A)	(N/A)	3,823	4,224
haplotypes (contigs)				

*Scaffold N50 is the length of chromosome 3, N/A: not applicable