Genome sequence of the Asian Tiger mosquito, Aedes albopictus, reveals insights into its biology, genetics, and evolution

The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>As Published</td>
<td><a href="http://dx.doi.org/10.1073/pnas.1516410112">http://dx.doi.org/10.1073/pnas.1516410112</a></td>
</tr>
<tr>
<td>Publisher</td>
<td>National Academy of Sciences (U.S.)</td>
</tr>
<tr>
<td>Version</td>
<td>Final published version</td>
</tr>
<tr>
<td>Accessed</td>
<td>Thu Jan 03 14:39:28 EST 2019</td>
</tr>
<tr>
<td>Citable Link</td>
<td><a href="http://hdl.handle.net/1721.1/105183">http://hdl.handle.net/1721.1/105183</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>Article is made available in accordance with the publisher's policy and may be subject to US copyright law. Please refer to the publisher's site for terms of use.</td>
</tr>
<tr>
<td>Detailed Terms</td>
<td></td>
</tr>
</tbody>
</table>
Genome sequence of the Asian Tiger mosquito, *Aedes albopictus*, reveals insights into its biology, genetics, and evolution

Xiao-Guang Chen\(^{a,1}\), Xuanting Jiang\(^b\), Jinhao Gu\(^a\), Meng Xu\(^b\), Yang Wu\(^a\), Yuhua Deng\(^a\), Chi Zhang\(^b\), Mariangela Bonizzoni\(^c,d\), Wannes Dermauw\(^e\), John Vontas\(^f,g\), Peter Armbruster\(^h\), Xin Huang\(^b\), Yulan Yang\(^b\), Hao Zhang\(^b\), Weiming He\(^b\), Hongjian Peng\(^b\), Yongfeng Liu\(^b\), Kun Wu\(^b\), Jiahua Chen\(^b\), Manolis Lirakis\(^a\), Pantelis Topalis\(^h\), Thomas Van Leeuwen\(^d,j\), Andrew Brantley Hall\(^a,i\), Xiaofang Ji\(^a\), Chevon Thorpe\(^e\), Rachel Lockridge Mueller\(^b\), Cheng Sun\(^f\), Robert Michael Waterhouse\(^o-p,q\), Guiyun Yan\(^a\), Zhijian Jake Tu\(^k\), Xiaodong Fang\(^b,1\), and Anthony A. James\(^b,3\)

\(^{a}\)Department of Pathogen Biology, School of Public Health and Tropical Medicine, Southern Medical University, Guangzhou 510515, China; \(^{b}\)Beijing Genomics Institute-Shenzhen, Shenzhen 518083, China; \(^{c}\)Program in Public Health, University of California, Irvine, CA 92697; \(^{d}\)Department of Biology and Biotechnology, University of Pavia, 27100 Pavia, Italy; \(^{e}\)Laboratory of Agrozoology, Department of Crop Protection, Faculty of Bioscience Engineering, Ghent University, B-9000 Ghent, Belgium; \(^{f}\)Institute of Molecular Biology and Biotechnology, Foundation for Research and Technology-Hellas, 73100 Heraklion, Greece; \(^{g}\)Faculty of Crop Science, Pesticide Science Lab, Agricultural University of Athens, 11855 Athens, Greece; \(^{h}\)Department of Biology, Georgetown University, Washington, DC 20057; \(^{i}\)Department of Biology, University of Crete, Heraklion, GR-74100, Crete, Greece; \(^{j}\)Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, 1090 GE Amsterdam, The Netherlands; \(^{k}\)Interdisciplinary PhD Program in Genetics, Bioinformatics, and Computational Biology, Virginia Tech University, Blacksburg, VA 24061; \(^{l}\)Department of Biochemistry, Fralin Life Science Institute, Virginia Tech University, Blacksburg, VA 24061; \(^{m}\)Cellular and Molecular Physiology, Edward Via College of Osteopathic Medicine, Blacksburg, VA 24060; \(^{n}\)Department of Biology, Colorado State University, Fort Collins, CO 80523; \(^{o}\)Department of Genetic Medicine and Development, University of Geneva Medical School, 1211 Geneva, Switzerland; \(^{p}\)Swiss Institute of Bioinformatics, 1211 Geneva, Switzerland; \(^{q}\)Contemporary Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139; \(^{r}\)The Broad Institute of MIT and Harvard, Cambridge, MA 02142; and \(^{s}\)Departments of Microbiology & Molecular Genetics and Molecular Biology & Biochemistry, University of California, Irvine, CA 92697

Edited by David L. Denlinger, Ohio State University, Columbus, OH, and approved September 22, 2015 (received for review August 20, 2015)

The Asian tiger mosquito, *Aedes albopictus*, is a highly successful invasive species that transmits a number of human viral diseases, including dengue and Chikungunya fevers. This species has a large genome with significant population-based size variation. The complete genome sequence was determined for the Foshan strain, an established laboratory colony derived from wild mosquitoes from southeastern China, a region within the historical range of the origin of the species. The genome comprises 1,967 Mb, the largest mosquito genome sequenced to date, and its size results principally from an abundance of repetitive DNA classes. In addition, expansions of the numbers of members in gene families involved in insecticide-resistance mechanisms, diapause, sex determination, immunity, and olfaction also contribute to the larger size. Portions of integrated flavivirus-like genomes support a shared evolutionary history of association of these viruses with their vector. The large genome repertoire may contribute to the adaptability and success of *Ae. albopictus* as an invasive species.

mosquito genome \(\rightarrow\) transposons \(\rightarrow\) flavivirus \(\rightarrow\) diapause \(\rightarrow\) insecticide resistance

The Asian tiger mosquito, *Aedes albopictus*, is an aggressive daytime-biting insect that is an increasing public health threat throughout the world (1). Its impact on human health results from its rapid and aggressive spread from its native home range, along with its ecological adaptability in different traits, including feeding behavior, diapause, and vector competence (2). This species is indigenous to East Asia and islands of the western Pacific and Indian Ocean, but has spread in the past 40 y to every continent except Antarctica (1). This widespread geographic distribution includes tropical and temperate zones, which is unusual for mosquitoes. *Ae. albopictus* is a competent vector for at least 26 arboviruses, and is important in the transmission of those that cause dengue and Chikungunya fevers (2, 3). It also is implicated as a vector of filarial nematodes of veterinary and zoonotic significance (4, 5). Although this species is considered a less efficient dengue vector than *Aedes aegypti* (2), it is the sole vector of recent outbreaks in southern China, Hawaii, and Gabon, and the first local (autochthonous) transmission in Europe (1, 2). *Ae. albopictus* vector competence for viruses is dynamic (3); for example, recent Chikungunya fever outbreaks in Réunion (Island), Mauritius, Madagascar, and Mayotte (2005–2007); Central Africa (2006–2007); and Italy (2007) were caused by viruses carrying at least one mutation that improved their transmission efficiency by the mosquito, making it the primary vector (2). The genome sequence of *Ae. albopictus* provides the basis for probing and understanding the mechanisms underlying its fast expansion and the development of strategies for controlling it and the pathogens it transmits.

Significance

*Aedes albopictus* is a highly adaptive species that thrives worldwide in tropical and temperate zones. From its origin in Asia, it has established itself on every continent except Antarctica. This expansion, coupled with its ability to vector the epidemic human diseases dengue and Chikungunya fevers, make it a significant global public health threat. A complete genome sequence and transcriptome data were obtained for the *Ae. albopictus* Foshan strain, a colony derived from mosquitoes from its historical origin. The large genome (1,967 Mb) comprises an abundance of repetitive DNA classes and expansions of the numbers of gene family members involved in insecticide resistance, diapause, sex determination, immunity, and olfaction. This large genome repertoire and plasticity may contribute to its success as an invasive species.


The authors declare no conflict of interest.

This article is a PNAS Direct Submission. Freely available online through the PNAS open access option.

Data deposition: The data reported in this paper have been deposited in the GenBank database (accession nos. SRA245721 and SRA215477, National Center for Biotechnology Information (NCBI); ID code JXUM0000000 [genome assembly]), and NCBI Transcriptome Shotgun Assembly database, www.ncbi.nlm.nih.gov/assembly/TSA.html (ID code GCLM00000000).\(^{1}\)

\(^{1}\)To whom correspondence may be addressed. Email: 18664867266@qq.com, Fangxu@gnomics.cn, or aajames@uci.edu.

This article contains supporting information online at [www.pnas.org/lookup/suppl/doi:10.1073/pnas.1516410112/-/DCSupplemental].

www.pnas.org/cgi/doi/10.1073/pnas.1516410112

PNAS Early Edition | 1 of 9
Results and Discussion
Genome Properties and Evolution.

Genome sequencing and assembly. A total of 23 libraries with insert sizes ranging from 170 bp to 20 kb in length were sequenced to yield a total of 689.59 Gb after filtering low-quality reads. An assembly generated a total of 1,967 Mb of sequence with scaffold and contig N50s of 195.54 kb and 17.28 kb, respectively, and an estimated whole-genome coverage of ~350 fold (SI Appendix, Figs. S1.1–S1.4 and Tables S1.1–S1.6).

Genome size variation. The *Ae. albopictus* genome is the largest of any mosquito species sequenced to date, which vary from 174 Mb for *Anopheles darlingi* to 540 Mb and 1,376 Mb for *Culex quinquefasciatus* and *Ae. aegypti*, respectively (6–9). Genome size variation also is observed among different populations of *Ae. albopictus* (10). An analysis of 47 geographic isolates from 18 countries showed a 2.5-fold variation in haploid genome weights (i.e., C-value) ranging from 0.62 pg in a population from Koh Samui (Thailand) to 1.66 pg in those recovered from Houston, TX (11). Inter- and intraspecific variation in genome size among mosquitoes appears to be caused mainly by changes in the amounts and organization of repetitive DNA. Increases in abundance of all classes of repetitive DNA sequences are correlated linearly with total genome size (12) (SI Appendix, Fig. S1.5 and Table S1.7).

Gene predictions. A total of 17,539 protein-encoding gene models were predicted de novo and supported by evidence-based searches using reference gene sets from mosquito (*Ae. aegypti*, *An. gambiae*, and *Cx. quinquefasciatus*) and fruit fly (*Drosophila melanogaster*) genome annotations, and this number is larger than those species with the exception of *Cx. quinquefasciatus* (SI Appendix, Table S1.9). These predictions were supported by RNA sequencing (RNA-seq)-based transcriptome data from multiple developmental stages (SI Appendix, Table S1.8). Approximately 93.6% of the predicted proteins matched entries in the SWISS-PROT, InterPro, or TrEMBL databases (SI Appendix, Tables S1.10–S1.12). Noncoding RNAs discovered in RNA-seq analyses include as many as 57 previously undescribed miRNAs putatively unique to *Ae. albopictus* (SI Appendix, Tables S1.13 and S4.4).

Evolution. Phylogenetic relationships based on 2,096 single-copy orthologous genes from five mosquito and one fruit fly species are consistent with previous reports that place *Ae. albopictus* within the Culicinae (www.fossilrecord.net) and estimate a divergence time of ~71 Mya from *Ae. aegypti* (Fig. 1), longer than a previous

![Fig. 1. Comparative analyses of the *Ae. albopictus* genome. (A) Phylogeny and divergence time estimation by molecular clock analysis. The mosquitoes *An. gambiae*, *An. darlingi*, *Cq. quinquefasciatus*, *Ae. albopictus*, and *Ae. aegypti* are estimated to have last shared a common ancestor with the fruit fly *D. melanogaster* ~260 Mya. The anopheline and culicine branches are estimated to have diverged 217 Mya (SE = 180.8–256.9 Mya). *Ae. albopictus* and *Ae. aegypti* are estimated to have last shared a common ancestor 71.4 Mya (SE = 44.3–107.5 Mya). (B) Comparisons of repeat content among the six species. Families of repetitive elements are represented by colors: green, Gypsy; blue, Pao, light blue, Copia, red, RTE-BovB; orange, LOA; purple, R1; white, other. The horizontal extent of each bar represents the relative length of repeat within each family. Numbers to the right represent the total length of repeats in the respective genomes. (C) Copy number of TE insertions relative to the estimated time of insertion. Shown here are type I LINE (LINE/I) transposons. AEDALB (red), *Ae. albopictus*; Aedaeg (blue), *Ae. aegypti*. (D) Statistics of TE representation in the genomes of six species. 1. Length in base pairs; 2, percentage of genome represented. DNA, DNA transposon; SINE, short interspersed nuclear element.](https://www.pnas.org/content/112/1/242030112)
estimate of ∼60 Mya (13). Mosquito/fruit fly divergence is estimated to have occurred ∼260 Mya. A total of 86 expansion (773 genes) and 26 contraction (108 genes) Ae. albopictus gene families were identified, and function enrichment of the former was determined (Fig. 2 and SI Appendix, Fig. S1.6 and Tables S1.14 and S1.15). Furthermore, 239 of the 2,086 orthologs (∼11%) show evidence of positive selection in Ae. albopictus with 32 Gene Ontology (GO) classes exhibiting significant enrichment (P < 0.05; SI Appendix, Table S1.16).

Properties of Specific Gene Categories. Repetitive DNA. The Ae. albopictus genome harbors all major groups of transposable elements (TEs) (Fig. 1). Repetitive sequences represent ∼68% of the genome, the most of all sequenced mosquito species (9). This high repeat content is consistent with the large genome size, and the total length of these DNAs is ∼40% more than that of Ae. aegypti, a member of the same subgenus, Stegomyia, and the only other mosquito with a sequenced genome larger than 1 Gb (5) (SI Appendix, Table S2). Non-LTR (long terminal repeats) retrotransposons or long interspersed nuclear elements (LINEs) showed the highest genome abundance in both species (Fig. 1). The LINE family, RTE-Bov, represents ∼15.7% (308 Mb) of the entire Ae. albopictus genome. Interestingly, a single Ae. albopictus LINE element, Duo (SI Appendix, Fig. S2.1), and its Ae. aegypti homolog, TF000022, occupy ∼4.1% (∼82 Mb) and ∼3.17% (∼44 Mb) of their respective genomes. The shared element and its abundance support the conclusion that it was present in the ancestral lineage of the two species. In contrast, >20% of the Ae. albopictus genome is occupied by interspersed repeats that have no similarity (i.e., e-value ≤ 1e-5) to Ae. aegypti sequences, and this provides support for the hypothesis that there was a rapid expansion of repeat DNA after divergence of the species.

The relative times of insertion of LINE and LTR retrotransposons were estimated by comparing sequence similarities among the best matching TE pairs within clusters (SI Appendix, Fig. S2.2). This analysis determined that the highest number of insertions in Ae. albopictus occurred within the last 10 My. Similar recent activity maxima were not observed or were at lower levels in Ae. aegypti TEs of the same clade. Thus, recent transposition of LTR and LINE retrotransposons contributes to the expansion of the Ae. albopictus genome.

Varied deletion rates also drive genome size differences (14, 15). Deletion rate analysis using the “dead-on-arrival” (i.e., neutral or evolving) non-LTR retrotransposon sequences from Ae. albopictus, Ae. aegypti, and Cx. quinquefasciatus reveals that there are more deletions than insertions, a result consistent with what is seen in other similarly analyzed organisms (15) (Table 1). Ae. albopictus has a slightly lower DNA loss rate than Ae. aegypti and Cx. quinquefasciatus, and this also may contribute to its large genome size. Flavivirus-like sequences in the Ae. albopictus genome. Sequences with similarity to flaviruses are detected in the genome of Ae. albopictus (16–18). Integrations from nonretroviral RNA viruses are referred to as nonretroviral integrated RNA viruses (NIRVs) (19, 20); the first integrations from flaviruses in the Ae. albopictus genome were referred to as Cell Silent Agents (16). NIRV representation in the genomes of the Ae. albopictus Foshan strain, Ae. aegypti (Assembly AaegL3), An. gambiae (AgamP4 assembly), and Cx. quinquefasciatus (CipiP2 assembly) were queried bioinformatically by using 261 sequences of previously characterized NIRVs along with the complete or portions of the genomes of representative insect-specific flaviruses (ISFs), mosquito-borne viruses (MBVs), and tick-borne viruses (TBVs) and flaviviruses with no known vector (NBVs; Dataset S2). No matches were returned for An. gambiae or Cx. quinquefasciatus, whereas thousands with e-values <10^-4 were detected in Ae. albopictus and Ae. aegypti (Datasets S3 and S4). Ae. albopictus has more variability than Ae. aegypti among viral types, including those with similarities to dengue viruses, and integrations were longer in length. Analyses and functional annotation of the sequences corresponding to basic local alignment search tool (BLAST) hits in Ae. albopictus revealed 24 sequences spanning partial or complete flaviviral ORFs, primarily NS1 and NS5, across 10 scaffolds (Dataset S5). NIRVs were embedded in regions rich with LTR retrotransposon sequences, primarily Ty3-copia and Ty3-gypsy (21, 22). No nucleotide repeats (direct or inverted) were observed at integration sites, supporting the conclusion that flaviviral integrations were derived from ectopic recombination with retrotransposons rather than being catalyzed by classical transposition activity (22, 23).

The larger number of NIRVs identified in the Foshan strain with respect to previous reports may result from the fact that past characterizations were based on gene-amplification analyses with flavivirus-specific primers (16, 18, 24). Alternatively, the larger number may indicate that these are ancient integrations and that migration out of its native range results in integration loss associated with founder effects. The presence of a variable number of integrations across geographic populations also may contribute to the observed variation in genome size of Ae. albopictus populations (10). The current variability in the NIRV integration sites and sequences support the conclusion that different regions of different length of the flavivirus genome can integrate. NIRVs phylogenetic relationship with respect to previously characterized NIRVs, ISFs and MBVs, TBVs, and NBVs indicate that flaviviral

Fig. 2. Expansion and contraction of gene families among mosquito species. The numbers designate the number of gene families that have expanded (green) and contracted (red) since the split from the last common ancestor. The most recent common ancestor (MRCA) has 7,590 gene families. Aedae, Ae. aegypti; Aedal, Ae. albopictus; Anoda, An. darlingi; Anoga, An. gambiae; Cxqu, Cx. quinquefasciatus; Drome, D. melanogaster.
integrations may occur in germ-line cells and may be inherited in *Ae. albopictus* as mosquitoes of the Foshan strain have been excluded from contact with wild-caught mosquitoes and viruses for more than 30 y (*SI Appendix*, Figs. S3.1–S3.4). At the same time, these data support the hypothesis that integrations of flaviviral sequences may be an ongoing regional process because NIRVs reported recently for *Ae. albopictus* collected in northern Italy (17) formed a separate cluster from any of the identified NIRVs (*SI Appendix*, Fig. S3.1), and sequences with an intact ORF and a high level of identity to circulating viruses were detected along with sequences harboring extensive rearrangements. Whether these sequences affect the replication and/or dissemination of mosquito-infecting arboviruses and contribute to vector competence is unknown.

**Diapause-related genes.** Diapause in insects is characterized by no or low growth, low metabolic activity, and an increased ability to survive environmental temperature and humidity extremes, and may be specific to one developmental stage. In mosquitoes, diapause may occur at the embryonic, larval, or adult stage (25), and it is observed in *Ae. albopictus* at low frequencies in populations derived from subtropical habitats (26). The Foshan strain was not tested for a diapause response. However, our analysis addresses regions of the genome that have been demonstrated to be expressed differentially as part of the diapause program in temperate, fully diapause-capable populations based on extensive previously published RNA-seq studies (27–30). A total of 71 genes with a putative diapause function were annotated based on these previous studies (Dataset S9). Of these, 14 are duplicated in the *Ae. albopictus* genome, including several with known diapause-related functions such as lipid metabolism, elongation of long-chain hydrocarbons, and hormone signaling.

Approximately 211 *Ae. albopictus* genes in expansion families are represented in transcriptomes of mosquitoes in diapause and nondiapause conditions in preadult and adult stages (27–30) (Dataset S10). Of these, 140 (66%) are expressed differentially during at least one of the life-cycle stages examined under diapause vs. nondiapause conditions, which is a greater percentage than the overall proportion of differentially expressed gene models in the transcriptome as a whole (*P* = 0.022, *SI Appendix*, Table S4.3). Furthermore, 96 of the 140 differentially expressed genes represent superfamilies of stress response, lipid metabolism, gene expression regulation, serine protease-related, and other genes (*SI Appendix*, Table S4.1). The proportion of genes within each category that show contrasting patterns of diapause-associated differential transcript accumulation (e.g., higher under diapause conditions at the preadult stage, no differential accumulation or lower under diapause conditions at the adult stage) was determined across preadult vs. adult stages of the life cycle. The superfamily categories of lipid metabolism ( *P* = 0.043), gene expression regulation ( *P* = 0.012), serine protease-related ( *P* = 0.003), and stress response ( *P* = 0.043) were enriched significantly for genes with contrasting patterns of diapause-associated differential transcript accumulation across the life cycle relative to the transcriptome database as a whole. These results are consistent with the hypothesis that gene family expansion can give rise to flexible gene expression across the life cycle and thereby contribute to the tolerance of environmental heterogeneity. Lipid metabolism also was implicated previously as an important transcriptional component of the diapause program in *Ae. albopictus*, and gene expression regulation is implicated as an important component of diapause-based extensive differential transcript accumulation (>5,000 genes represented) under diapause and nondiapause conditions (27–30). The role of contrasting differential transcript accumulation for serine protease genes across the life cycle remains unclear.

**Detoxification (cytochrome-oxidase P450, carboxy/cholinesterase, glutathione S-transferases) and ABC transporter gene families.** The *Ae. albopictus* genome contains 186 full-length cytochrome-oxidase P450 (CYP) genes, compared with 168, 104, and 87 in *Ae. aegypti*, *An. gambiae*, and *D. melanogaster*, respectively (Table 2, *SI Appendix*, Figs. S5.1–S5.6 and Table S5.3, and Datasets S12 and S13), and 196 are reported in *Cq. quinquefasciatus* (31). Approximately 24 CYP pseudogenes also were found. Approximately 20% of the *Ae. albopictus* genes are clustered on three scaffolds (scaffolds 64, 501, and 4011). All orthologs of the CYP9J family, the main pyrethroid metabolizers in *Ae. aegypti* (32), were identified in *Ae. albopictus*.

Most insects have one or two genes in the CYP4G family (33), and we identified three each in *Ae. albopictus*—AalbCYP042, AalbCYP052, and AalbCYP125—and *Ae. aegypti* (*SI Appendix*, Figs. S5.2 and S5.3). Cyp4g1 in *D. melanogaster* is the most highly expressed of all fruit fly CYP genes, and encodes an insect-specific P450 oxidative decarboxylase with a role in cuticular hydrocarbon biosynthesis (33). Phylogenetic analysis shows that two *Ae. albopictus* CYP4Gs (042 and 052) cluster together and do not cluster on a 1:1 basis with the *Ae. aegypti* CYP4Gs, supporting the conclusion

<table>
<thead>
<tr>
<th>Gene type</th>
<th>D. melanogaster</th>
<th>An. gambiae</th>
<th>Ae. aegypti</th>
<th>Cx. quinquefasciatus</th>
<th>Ae. albopictus</th>
<th><em>D. melanogaster</em></th>
<th>An. gambiae</th>
<th>Ae. aegypti</th>
<th>Cx. quinquefasciatus</th>
<th>Ae. albopictus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cytochrome P450s†</td>
<td>87</td>
<td>104</td>
<td>168</td>
<td>196</td>
<td>186 (210)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glutathione-S-transferases ³�†</td>
<td>37</td>
<td>28</td>
<td>26</td>
<td>35</td>
<td>32 (37)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCEs³</td>
<td>34</td>
<td>46</td>
<td>59</td>
<td>71</td>
<td>64 (71)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABC transporters³</td>
<td>56</td>
<td>52</td>
<td>58</td>
<td>—</td>
<td>71</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
that gene duplication likely occurred after divergence of the two lineages. The abundant expression of AalbCYP042 and AalbCYP052 during egg formation and diapause (27) is consistent with a potential role in promoting mosquito survival during unfavorable environmental conditions and may contribute to the invasion success of the species. An ortholog of D. melanogaster Cyp4gl5 in the wild silk moth Antheraea yamamai (CYP4G25) is also expressed highly during diapause of the pharate first-instar larvae (34).

Sixty-four full-length carboxy/cholinesterase (CCE) genes were identified in Ae. albopticus, a number similar to that found in Ae. aegypti and Cx. quinquefasciatus, but higher than the numbers found in D. melanogaster and An. gambiae (Table 2, SI Appendix, Fig. S5.7 and Table S5.6, and Datasets S14 and S15). A CCE gene, CCCEa3A, implicated in temephos resistance in Ae. aegypti (35) and Ae. albopticus (36), is present in Ae. albopticus as two tandemly duplicated genes (AalbCCE013 and AalbCCE014). Acetylcholinesterases are major insecticide targets, and, in contrast to other insects that only have one or two such genes, three (AalbCCE031 and AalbCCE100, orthologs of Ace1, and AalbCCE101, an ortholog of Ace2) are annotated in the Ae. albopticus genome. Furthermore, a notable expansion to 18 and 9 genes was found for the subfamily of cricket co-orthologs and juvenile hormone esterases, respectively (37, 38). In D. melanogaster the cricket gene is located at a locus essential for mediating the response of adult tissues to juvenile hormone (37, 38) and allelic variants in the gene contribute to altitudinal variation in development time (39). Among the mosquito species, Ae. albopticus had the highest number of cricket orthologs, with five cases in which Ae. albopticus has two or three copies compared with one in Ae. aegypti (SI Appendix, Fig. S5.7).

The numbers of glutactins in Ae. albopticus and Ae. aegypti are nearly double those found in An. gambiae and D. melanogaster. The function of glutactins is not well understood, but a role in the formation of the eggshell matrix was proposed (7, 40).

Ae. albopticus has 32 full-length cytosolic glutathione S-transferase (GST) genes (Table 2, SI Appendix, Fig. S5.8 and Table S5.9, and Datasets S16 and S17), more than Ae. aegypti and An. gambiae, but fewer than D. melanogaster and Cx. quinquefasciatus. This expansion results mainly from the higher number of delta- and epsilon-class GSTs, the majority of which are associated with insecticide resistance (41). Finally, we annotated 71 putative ABC genes in Ae. albopticus, more copies than are found in Ae. aegypti, D. melanogaster and most other insect species (Table 2, SI Appendix, Figs. S5.9–S5.16 and Table S5.12, and Datasets S18 and S19). Orthologs of ABC proteins conserved widely in metazoans were identified, with five cases of duplicated ABCC transporter genes, a family known for its role in multidrug resistance in humans. Similarly, six duplications of Ae. aegypti ABCG genes are found in Ae. albopticus. Human ABCG transporters are involved in lipid transport, and the duplication in Ae. albopticus of genes encoding these proteins may be related to the complex regulation of increased lipid content in diapausing vs. nondiapausing pharate larvae. These combined findings provide genomic support for the potential of a robust response of Ae. albopticus to environmental stresses and insecticides.

**Odorant-binding and odorant receptor proteins.** A total of 86 odorant-binding proteins (OBPs) and 158 odorant receptor (OR) genes are predicted in the Ae. albopticus genome (Table 3 and SI Appendix, Table S6.1). All the OBPs are members of the pheromone-binding protein (PBP)/GOBP family, 47 of which are PBPs with putative functions associated with communication (42–45). Orthologs of 156 of the OR genes could be found in Ae. aegypti. Comparisons of the Ae. albopticus repertoire with An. gambiae, Ae. aegypti, Cx. quinquefasciatus, and D. melanogaster confirmed previous reports of smaller numbers of genes encoding OBPs and ORs in the fruit fly than the mosquito species (46–56) (Table 3 and SI Appendix, Fig. S6.2). Both Ae. albopticus and Ae. aegypti have more of both classes of genes than An. gambiae and Cx. quinquefasciatus, and 43 OBPs and two OR putative novel genes (i.e., no orthologs identified in other species) contribute to these differences (SI Appendix, Table S6.1). Most of the putative OBPs genes encode a predicted N-terminal signal peptide, a feature characteristic of their respective proteins (52, 56, 57), and had molecular weights ranging from 14 to 41 kDa. Conserved domain database (CDD) predictions showed that they belong to the PBP/GOBP family, and amino acid alignments confirmed the conservation of six characteristic cysteines (SI Appendix, Fig. S6.3). The putative OR genes encode seven transmembrane domain proteins characteristic of this family.

The expression profiles of OBPs and ORs in Ae. albopticus and other mosquitoes in which data are available show an increasing complexity in the number of transcriptionally active genes as the insects progress through development (47, 56, 57) (Fig. 3 and SI Appendix, Fig. S6.4 and Table S6.2). Furthermore, although their mRNAs are present at relatively low abundance, these genes exhibit distinct temporal- and tissue-specific expression. The increasing transcriptional activity may contribute to the ability of this group of insects to navigate increasingly complex environments as they transition from food location in aqueous larval habitats to the mate-detection, host-seeking, feeding-preference and oviposition site-identification abilities of the adults.

**Sex-biased gene expression.** Sex-biased gene expression is responsible for the extensive phenotypic and behavioral differences exhibited by male and female mosquitoes (58). RNA-seq analysis of separate samples derived from Ae. albopticus adult females and males identified 8,559 and 4,140 genes, respectively, with sex-biased expression profiles, and the total represents ∼50% of all annotated genes in the genome (Dataset S24). A total of 246 and 268 genes in females and males, respectively, exhibited sex-specific expression (Datasets S25 and S26). Genes with sex-biased expression are enriched significantly (P < 0.01) in GO terms for 26 biological process, 11 cellular compartment, and 34 molecular function categories, with the highest representations in RNA metabolic processes, nucleus, and ion binding (Datasets S27–S30). Further studies are needed to link many of these genes with specific roles in sex-specific biology.

**Sex-determination genes.** Aedes mosquitoes, including Ae. albopticus, have a homomorphic sex-determining chromosome with a small male-specific region called the M-locus containing DNA functioning as a dominant male-determining factor (M factor) (59, 60). Importantly, an Ae. aegypti gene, Nix, encoding the phenotypic properties of the M factor (59), has an ortholog, KP765684, in Ae. albopticus. Orthologs of other genes likely involved in the sex-determination pathway also were found. Several orthologs of transformer2 (tra2) and those of doublesex and fruitless, the terminal regulatory genes in the sex-determination pathway, were identified (SI Appendix, Table S7.4). No ortholog of transformer was found, most likely because it evolves rapidly resulting in sequence divergence (61–63).

### Table 3. Numbers of annotated genes encoding OBP and OR

<table>
<thead>
<tr>
<th>Protein</th>
<th>Ae. albopticus</th>
<th>Ae. aegypti</th>
<th>An. gambiae</th>
<th>Cx. quinquefasciatus</th>
<th>D. melanogaster</th>
</tr>
</thead>
<tbody>
<tr>
<td>OBP</td>
<td>86</td>
<td>64</td>
<td>58</td>
<td>50</td>
<td>51</td>
</tr>
<tr>
<td>OR</td>
<td>158</td>
<td>110</td>
<td>80</td>
<td>88</td>
<td>61</td>
</tr>
</tbody>
</table>
Comparative analysis with curated sets of cages and allowed access to a cotton. The raw sequence data were filtered was obtained from the ∼μ Ae. albopictus data reveal increase | Ae. albopictus and low-quality bases. The assembler SOAPdenovo (version Ae. albopictus The Foshan strain of Ae. aegypti pupa of a ninth-generation isofemale line was sub-

Immune-related genes. Comparative analysis with curated sets of immune-related genes (64, 65) identified 554, 476, 536, 400, and 345 immunity genes (includes candidate pseudogenes) in Ae. albopictus, Ae. aegypti, An. gambiae, Cx. quinquefasciatus, and D. melanogaster, respectively (SI Appendix, Table S8.1). Expansions of several gene families (SPZ, BGBP, SRRP, GALE, TOLL, SCR, TOLLPATH, SOD, APHAG, PPO, and CLIP) account for the large Ae. albopictus immunity-related gene repertory. Analyses of the transcriptome data reveal increased abundances of the immune-related gene products in the postembryonic insects (SI Appendix, Fig. S8.6). A total of 468 immune-related transcripts [representing ~88% (486 of 554) of the total predicted genes] were found in adult mosquitoes, including 166 related to immune recognition, 106 involved in gene modulation, 100 in signal transduction, and 96 in effector molecule (SI Appendix, Fig. S8.7). The top three most abundant transcripts represent effector AMPs, recognition LRRs, and modulation CLIPs (56, 53, and 51, respectively).

Summary and Conclusions
It was known previously that Ae. albopictus had a large genome, and this is confirmed in the present report. This large size is evident in the greater number of repetitive DNA elements, expansion in all protein-encoding gene categories examined, and the amount of the genome represented by insertions of DNA copies of RNA viruses. The genome size may account in part for why this mosquito is successful as an invasive species. The large repertory of noncoding and coding DNA may provide the genetic substrates from which adaptation emerges following selection in novel environments. A draft sequence was published recently of an Ae. albopictus strain, Fellini, that recently invaded Italy (66). Detailed molecular comparisons of that strain with the ancestral one presented here are expected to highlight aspects of genome evolution as this species adapts to more temperate zones.

Materials and Methods
Mosquitoes. The Foshan strain of Ae. albopictus was obtained from the Center for Disease Control and Prevention of Guangdong Province, China, where it has been in culture since 1981. Mosquitoes were reared at 28 °C and 70–80% relative humidity with 14/10 h light/dark cycles. Larvae were reared in pans and fed on finely ground fish food mixed at a 1:1 ratio with yeast powder. Adults were kept in 30-cm³ cages and allowed access to a cotton wick soaked in 0.2 g/ml sucrose as a carbohydrate source. Adult females were allowed to feed on anesthetized mice 3–4 d after eclosion.

DNA Sequencing. Approximately 1.414 μg of genomic DNA isolated from a single Ae. albopictus pupa of a ninth-generation isofemale line was subjected to whole-genome amplification (67) to produce 243.2 μg of DNA. Amplified DNA was used to construct paired-end short-insert (170,500 and 800 bp in length) and mate-paired long-insert (2 kb, 5 kb, 10 kb, and 20 kb) genomic libraries, and these were sequenced by using the HiSeq.2000 platform.

Data Quality Control and Assembly. The raw sequence data were filtered before assembly by removing duplicated reads caused by gene amplification and reads contaminated by adapters, trimming continuous low-quality bases on 5' ends according to quality graphs, and filtering reads with a significant excess of "N" and low-quality bases. The assembler SOAPdenovo (version 2.04) (68), SSPACE (version 2.0) (69), and Gapcloser (version 1.10) (68) were used for genome assembly. Overlapped pair-end reads from the 170-insert size libraries were connected first to yield long sequences. A 97-bp sequence from the connected long reads was used next to construct contigs. All usable reads from different insert-size libraries then were realigned to the contigs by using SSPACE. The resulting linking information was used to produce the final scaffold construction, and this was followed by gap-filling of the scaffolds. The sequences of Wolbachia pipientis were aligned to the assembly, and the scaffolds matching them were removed to avoid the contamination.

Accuracy of Genome Assembly. The quality of the draft genome was evaluated by assessing the sequencing depth and coverage by using available mRNA and fosmid sequences. All usable sequence reads were realigned to the draft genome by using SOAP2 (70).

Transcriptome Sequencing (RNA-Seq). Transcriptomes were derived from libraries comprising mRNA from seven developmental stages of the Foshan strain: mixed-sex samples of 100 embryos at 0–24 h post deposition (hpd), 100 embryos at 24–48 hpd, a combined pool of 8 first-and second-instar larvae, a combined pool of five third- and fourth-instar larvae, five pupae of all stages, and five each of adult males and sugar-fed adult females. TRizol reagent (Invitrogen) and RNase-free DNase I were used to extract and treat total RNA. Polyadenylated (i.e., poly(A⁺)) mRNA was enriched by using oligo-dT beads, fragmented, and primed randomly during the first-strand synthesis by reverse transcription. Second-strand cDNA was synthesized by using RNase H and DNA polymerase I to create double-stranded fragments. The ds cDNA was applied to 200-bp paired-end RNA-seq libraries per Illumina protocols and sequenced with 90 bp at each end on the Illumina HISEq 2000 platform. The cDNA library was normalized by the duplex-specific nuclease

Fig. 3. Proportion of OBP and OR gene transcripts in each expression level through development. The proportion of (A) OBP and (B) OR genes, containing nonexpressed genes, in each expression level of each stage was calculated. The red, blue, and green columns refer to high (>1), medium (0.1–1) and low (<0.1) RPKM levels, respectively.
method (71) followed by cluster generation on the Illumina HiSeq2000 platform. Transcript reads were mapped by TopHat and analyzed subsequently with custom Perl scripts. Gene expression levels were calculated as reads per kilobase of exon model per million mapped reads (RPKM) (72). Genes expressed differentially between two samples were detected by using a method based on a Poisson distribution, and samples were normalized for differences in the RNA output size, sequencing depth, and gene length. Genes identified in at least one experiment with a minimum twofold difference (RPKM) in two experiments and an false discovery rate of <0.001 were defined as differentially expressed. Enrichment analysis was performed by using Enrich Pipeline (73).

**Gene Annotation.** De novo gene prediction by using RNA-seq data and *Ae. aegypti, D. melanogaster, An. gambiae, and Cx. quinquefasciatus* protein sequences aligned to the *Ae. albopictus* genome with TBLASTN (74) was performed to produce homology-based predictions. Putatively homologous genome sequences were aligned with the matching proteins by using GeneWise (75) to define gene models. Augustus (76) and Gscans (77) were used with appropriate parameters for de novo prediction of coding genes. Homology-based and de novo-derived gene sets were merged to form a comprehensive and nonredundant reference gene set using GLEAN (source-forg.net/projects/glean-gene). The transcriptome reads from the seven different samples were mapped to the reference gene set using TopHat (78) to give RNA-seq-based predictions. TopHat mapping results were combined, and Cufflinks (79) was applied to predict transcript structures. A total of 1,000 intact genes also were selected from the homology-based prediction to pass a fifth-order Markov model, then to predict the ORFs of RNA transcripts based on the hidden Markov model. Finally, the RNA transcripts were integrated with the GLEAN gene set to form the final nonredundant gene set.

Manual annotation of putative diapause-related genes was performed by using Web Apollo (80) to integrate the original GLEAN/Cuff annotations on the scaffolds with Maker annotations (81) based on a comprehensive diapause transcriptome (27–30). Annotated genes included those involved in chromatin remodeling, lipid metabolism, hormonal regulation, circadian rhythms, and other functions. Final annotations were based on the presence of a start codon, stop codon, canonical splice sites, and extended 5′ or 3′ UTRs that were supported by Maker or exonerate (82) alignment of contigs from the transcriptome.

**Gene Functional Annotation.** *Ae. albopictus* protein sequences were aligned by using InterPro (83), Swiss-Prot (84), Kyoto Encyclopedia of Genes and Genomes (KEGG) (85), and TrEMBL (86) to infer their biological functions or their molecular pathways. GO descriptions of gene products were retrieved from InterPro. The symbol of each gene was assigned based on the best match derived from the alignments with Swiss-Prot databases by using BLASTP. Motifs and domains were annotated by InterPro by searching publicly available databases, including Pfam, PRINTS, PANTHER, PROSITE, ProDom, and SMART. Genes also were mapped to KEGG pathway maps by searching KEGG databases and finding the best hit for each gene.

**Gene Family Clustering.** The TreeFam methodology (86) was used to define gene families using data from five mosquito species (*Ae. albopictus, An. gambiae, Ae. aegypti, Cx. quinquefasciatus*, and *An. darlingi*) as references, and the fruit fly *D. melanogaster* was used as the outgroup. BLASTP was used to find all homologous relationships among protein sequences of the six species with e-values <1e-10, and Solar (in-house software, version 0.9.6) was used to conjoin high-scoring segment pairs between each pair of protein homologs. Protein sequence similarity was assessed with bit-score, and protein encoding genes clustered into gene families by a hierarchical clustering algorithm (an implementation included in the TreeFam pipeline, version 0.5.0) with an algorithm analogous to average-linkage clustering with the parameters set to be “-w 5 s 0.33 -m 100000”.

**Phylogenetic Tree Construction and Divergence Time Estimate.** A total of 2,096 single-copy gene families defined as orthologous genes according to the TreeFam pipelines chosen in this analysis were assigned to a coding sequence (CDS) based on the alignment results. All CDSs and the 4d sites (fourfold degenerate synonymous sites) were extracted from each alignment and concatenated to one super gene for the six species. PhyMLv3.0 (parameters: -m HKY85, other default) was used to construct a phylogenetic tree for the six species. The chain length was set to 100,000 (1 sample/100 generations), and the first 1,000 samples were burned in. The transition/transversion ratio was estimated as a free parameter. Divergence time was estimated by using the program MCMCTREE (version 4), which was part of the PAML package. “JC69” models in MCMCTREE program were used in our calculations.

**Expansion and Contraction of Gene Families.** Computational Analysis of gene Family Evolution (version 2.1) (87) was used to detect gene family expansion and contraction in *Ae. albopictus, An. gambiae, Ae. aegypti, Cx. quinquefasciatus, An. darlingi*, and *D. melanogaster* with the parameters “P-value threshold 0.05, number of random 10,000, and search for the λ value.” Genes families with P values <0.05 were analyzed manually.

**Detection of Positively Selected Genes.** BLASTP and TreeFam methodologies were used to define orthologs among *Ae. albopictus, An. gambiae, Ae. aegypti, Cx. quinquefasciatus, An. darlingi*, and *D. melanogaster*. The coding sequences of the orthologs were aligned by using Prank software (88) (code. google.com/prank-msa) with default parameters. The genes were filtered even if the alignment rate of the gene was less than 80% in only one species. Ks (nonsynonymous substitution rates) and Kn (synonymous substitution rates) were calculated for the aligned orthologs by using KaKs calculator software (89) (version 1.2, parameter “-m YN”) with default parameters.

**DNA Loss Analysis in Mosquito Genomes.** The DNA loss rates for neutrally evolved DNA sequences in mosquito genomes were estimated by using a previously described method (90). In brief, the consensus sequences of autonomous non-LTR retrotransposons in the focal mosquito genomes were collected. The consensus sequences for *Ae. aegypti* and *Cx. quinquefasciatus* were downloaded from TFam (tefam.biochem.vt.edu/tefam/index.php). The consensus sequences for *Ae. albopictus* were generated in the present study by using RepeatScout (91). Second, the consensus sequences were trimmed to keep only the protein-coding regions. Third, the consensus sequences after trimming were used as a repeat library to mask their corresponding genomic sequences by RepeatMasker (www.repeatmasker.org/) to generate pairwise aligning files. We used the obtained alignments to eliminate all non-LTR sequences with nonrandom distributions of substitutions against codon positions (χ2 test, P < 0.05) to avoid counting substitutions that occurred along master element lineages. Finally, for each remaining non-LTR element copy, the numbers of insertions, deletions, and substitutions relative to the consensus sequence were obtained based on the RepeatMasker-generated alignment, and the sums of these values for every individual element copy were used to represent the total amounts of DNA gained and lost through small indels (<30 bp) in the focal mosquito genome (base pairs deleted minus base pairs insertionsubstituted).

**Analyses of Specific Gene Features, Gene Families, and Developmentally Regulated Gene Expression.** The specific materials and methods used in the discovery and analysis of TEs, and integrated flavivirus-like sequences, are described in the SI Appendix. Discovery and analysis of gene family members involved in insecticide resistance, diapause, sex determination, immunity, and olfaction are also described in the SI Appendix.

**ACKNOWLEDGMENTS.** This work was supported by National Natural Science Foundation of China Grants U0832004, 81371845, and 8142010100284 (to X.-G.C.); Research Team Program of Natural Science Foundation of Guangdong Grant 2014A0303101619 (to X.-G.C.); Scientific and Technological Program of Guangdong Grant 2013SY05000012 (to X.-G.C.); International Cooperation Program of Guangzhou Grant 201343500016 (to X.-G.C.); National Institute of Allergy and Infectious Diseases Grants AI083202 (to X.-G.C.), D43TW009527 (to G.Y.), and R27AI209746 (to A.J.L.); Marie Curie International Outgoing Fellowship PIOF- GA-2011-303312 (to R.M.W.); and the Leading Scholar Program of Guangdong (G.Y.), W.D. is a postdoctoral fellow of the Fund for Scientific Research Flanders.

Chen et al.

PNAS Early Edition | 7 of 9


