Discovery of flavivirus-derived endogenous viral elements in two *Anopheles* mosquito genomes supports the existence of *Anopheles*-associated insect-specific flaviviruses

Sebastian Lequime[1,2,3], Louis Lambrechts[1,2]*

[1]Insect-Virus Interactions Group, Department of Genomes and Genetics, Institut Pasteur, Paris, France

[2]Centre National de la Recherche Scientifique, Unité de Recherche Associée 3012, Paris, France

[3]University Pierre et Marie Curie, Cellule Pasteur UPMC, Paris, France

[*]Corresponding author. Address: Insect-Virus Interactions, Institut Pasteur, 28 rue du Docteur Roux, 75724 Paris Cedex 15, France. E-mail: sebastian.lequime@gmail.com

Abstract

The *Flavivirus* genus encompasses several arboviruses of public health significance such as dengue, yellow fever, and Zika viruses. It also includes insect-specific flaviviruses (ISFs) that are only capable of infecting insect hosts. The vast majority of mosquito-infecting flaviviruses have been associated with mosquito species of the *Aedes* and *Culex* genera in the Culicinae subfamily, which also includes most arbovirus vectors. Mosquitoes of the Anophelinae subfamily are not considered significant arbovirus vectors, however flaviviruses have occasionally been detected in field-caught *Anopheles* specimens. Whether such observations reflect occasional spillover or laboratory contamination or whether *Anopheles* mosquitoes are natural hosts of flaviviruses is unknown. Here, we provide *in silico* and *in vivo* evidence of transcriptionally active, flavivirus-derived endogenous viral elements (EVEs) in the genome of *Anopheles minimus* and *Anopheles sinensis*. Such non-retroviral endogenization of RNA viruses is consistent with a shared evolutionary history between flaviviruses and *Anopheles* mosquitoes. Phylogenetic analyses of the two newly described EVEs support the existence of a distinct clade of *Anopheles*-associated ISFs.

Introduction

Flaviviruses are positive-sense, single-stranded RNA viruses that infect a broad range of hosts including vertebrates (e.g., birds, primates) and arthropods (e.g., ticks, mosquitoes). In addition to major arboviruses of public health significance such as dengue, Zika, West Nile and yellow fever viruses, the *Flavivirus* genus also includes vertebrate-specific (not known vector; NKV) and insect-specific (insect-specific flaviviruses; ISFs) members (Moureau et al. 2015). The majority of mosquito-infecting flaviviruses have been associated with members of the Culicinae subfamily, mainly from the *Culex* and *Aedes* genera. *Anopheles* mosquitoes in the Anophelinae subfamily are well known for their role in the transmission of human malaria parasites, but they are not considered important vectors of arboviruses in general, and of flaviviruses in particular. Nevertheless, several studies have detected flaviviruses in field-caught *Anopheles* mosquitoes from different parts of the world. In North America, West Nile virus (WNV) was detected in *Anopheles punctipennis* (Bernard et al. 2001; Kulasekera et al. 2001). In Asia, Japanese encephalitis virus was detected in *Anopheles sinensis* (Feng et al. 2012; Liu et al. 2013). In Europe, *Anopheles maculipennis* was found positive for WNV (Filipe 1972; Kemenesi et al. 2014), Usutu virus (Calzolari et al. 2013) and Batai virus (Calzolari et al. 2010). Interestingly, some ISFs were also detected in *An. sinensis* (Zuo et al. 2014; Liang et al. 2015) and *Anopheles atroparvus* (Aranda et al. 2009). It is unknown whether these detections reflect occasional spillover or laboratory contamination, or whether *Anopheles* mosquitoes are in fact natural hosts of flaviviruses.

Endogenous viral elements (EVEs) are chromosomal integrations of partial or full viral genetic material into the genome of a host species. Not only retroviruses, whose replication cycle includes incorporation of a DNA form of the RNA viral genome into the host cell genome, but virtually all types of eukaryotic viruses can become endogenous (Feschotte and Gilbert 2012). Non-retroviral EVEs have been documented in the genome of a wide variety of host species, including vertebrates and arthropods (Feschotte and Gilbert 2012). Unlike detection of exogenous viruses, subject to possible laboratory or environmental contamination, EVEs are likely to reflect a long-lasting evolutionary relationship between an RNA virus and its natural host. This is because endogenization, for a single-stranded RNA virus, requires (1) reverse transcription, (2) integration of virus-derived DNA into the genome of germinal host cells and (3) fixation of the integrated sequence in the host population (Holmes 2011; Aiewsakun and Katzourakis 2015). The low probability of this combination of events makes endogenization exceedingly unlikely to occur unless the viral infection is common in the host population over long evolutionary times. For example, flavivirus-derived EVEs have been reported in the genome of *Aedes aegypti* (Crochu et al. 2004; Katzourakis and Gifford 2010) and *Aedes albopictus* (Roiz et al. 2009; Chen et al. 2015). These EVEs are phylogenetically related to the clade of *Aedes*-associated ISFs (Crochu et al. 2004; Roiz et al. 2009; Katzourakis and Gifford 2010; Chen et al. 2015), which is consistent with the ancient evolutionary relationship between *Aedes* mosquitoes and ISFs.

Here, we report the discovery of two flavivirus-derived EVEs in the genomes of *An. minimus* and *An. sinensis* mosquitoes. We screened 24 publicly available *Anopheles* genomes (Holt et al. 2002; Zhou et al. 2014; Logue et al. 2015; Neafsey et al. 2015) for flaviviral sequences, and validated *in silico* hits both at the DNA and RNA levels *in vivo*. The two newly described flavivirus-derived EVEs are phylogenetically related to ISFs, and support the existence of a previously unsuspected *Anopheles*-associated clade of ISFs.

Material and Methods

1. *In silico* survey

1.1. Genome screen

Twenty-four *Anopheles* genomes (full list and accession numbers are provided in Table 1) were screened by tBLASTn (BLAST+ 2.2.28) (Camacho et al. 2009) for the presence of flavivirus-derived EVEs using a collection of 50 full flavivirus polyprotein queries (full list and accession numbers are provided in Table S1) representing the currently known diversity of the *Flavivirus* genus. The sequences of hits whose E-value was > E^{-4} were extracted using an in-house bash shell script. In order to reconstruct putative flavivirus-derived EVEs, BLAST hits were clustered using CD-HIT v.4.6.1 (Li and Godzik 2006) and aligned using MAFFT v7.017 (Katoh et al. 2002). Putative EVEs were extracted and used as query for a reciprocal tBLASTn (BLAST+ 2.2.30) against the National Center for Biotechnology Information (NCBI) nucleotide database (E-value > 10^{-4}). Genetic features of identified EVE were analyzed using the NCBI Conserved Domain Database (Marchler-Bauer et al. 2015). Nucleotide sequence data reported are available in the Third Party Annotation Section of the DDBJ/ENA/GenBank databases under the accession numbers TPA: BK009978-BK009980.

1.2. Phylogenetic analyses

Translated EVE sequences were aligned to the corresponding sections of several flavivirus polyproteins (Table S1) with MAFFT v7.017 and phylogenetically uninformative positions were trimmed using TrimAI v.1.3 (Capella-Gutiérrez et al. 2009) accessed through the webserver Phylemon 2 (Sánchez et al. 2011). The trimmed alignments were used to construct phylogenetic trees with PhyML Best AIC Tree (Sánchez et al. 2011). Best substitution models were Blosum62+I+G+F for *An. minimus* and Blosum62+I+G for *An. sinensis*.

1.3. Transcriptome screen

Published RNA sequencing (RNA-seq) data were retrieved from NCBI Sequence Read Archive (Leinonen et al. 2011) and explored for the presence of previously identified EVE sequences. Only one *An. minimus* transcriptome sequence read archive was found under accession number SRX265162. Six *An. sinensis* transcriptome sequence read archives were found under accession numbers SRX448985, SRX449003, SRX449006 and SRX277584 for experiments using Illumina sequencing technology, and SRX218691 and SRX218721 for experiments using Roche 454 sequencing technology. RNA-seq reads were mapped to the EVE nucleotide sequence using Bowtie2 v2.1.0 (Langmead and Salzberg 2012). The alignment file was converted, sorted and indexed with Samtools v0.1.19 (Li et al. 2009). Coverage was assessed using bedtools v2.17.0 (Quinlan and Hall 2010).

2. *In vivo* validation

2.1. Mosquitoes

Anopheles minimus and *An. sinensis* mosquitoes were obtained through BEI Resources (www.beiresources.org), National Institute of Allergy and Infectious Diseases, National Institutes of Health (*An. minimus* MINIMUS1, MRA-729; *An. sinensis* SINENSIS, MRA-1154). *Anopheles minimus* and *An. sinensis* specimens came from the 132nd and 65th generations of laboratory colonization, respectively. Eggs were hatched in filtered tap water, reared in 24×34×9 cm plastic trays and fed with fish food (TetraMin, Tetra, Melle, Germany). Adults were maintained in 30×30×30 cm screened cages under controlled insectary conditions (28±1°C, 75±5% relative humidity, 12:12 hour light-dark cycle). They were provided with cotton soaked in a 10% (m/v) sucrose solution *ad libitum*. *Anopheles stephensi* nucleic acids, used as a reaction control, were kindly provided by the Genetics and Genomics of Insect Vectors unit, Institut Pasteur, Paris.

2.2. EVE genomic integration

Mosquitoes were homogenized in pools of 10 separated by sex in 300 µL of Dulbecco's phosphate-buffered saline (DPBS) during two rounds of 30 sec at 5,000 rpm in a mixer mill (Precellys 24, Bertin Technologies, Montigny le Bretonneux, France). DNA was extracted using All Prep DNA/RNA Mini Kit (Qiagen, Hilden, Germany) following the manufacturer's instructions. EVE presence in genomic DNA was assessed by 35 cycles of PCR using Taq Polymerase (Invitrogen, Thermo Fisher Scientific, Waltham, MA, USA) (Table S2). PCR primers were designed to generate an amplicon spanning part of the EVE sequence and a section of the flanking host sequence. Identity of the EVE sequence was confirmed by Sanger sequencing of the PCR product.

2.3. EVE transcription level

Mosquitoes were homogenized in pools of 5 separated by sex or development stage in 300 µL of DPBS during two rounds of 30 sec at 5,000 rpm in a mixer mill (Precellys 24). RNA was extracted from mosquito homogenates separated by sex using TRIzol Reagent (Life Technologies, Thermo Fisher Scientific, Waltham, MA, USA) following the manufacturer's instructions. Samples were treated with Turbo DNA-free kit (Life Technologies) and reverse transcribed using random hexamers and M-MLV reverse transcriptase (Invitrogen). Complementary DNA was amplified with 35 cycles of PCR for *An. minimus* and 40 cycles of PCR for *An. sinensis*, respectively, using DreamTaq polymerase (Thermo Fisher Scientific) and primers located within the EVE sequence (Table S2). To verify that RNA samples were free of DNA contamination, two sets of primers spanning exons 3 and 4 of the RPS7 gene of both *Anopheles* species (under VectorBase annotation number AMIN008193 and ASIC017918 for *An. minimus* and *An. sinensis*, respectively) were designed (Table S2). Because the corresponding DNA sequence includes intron 3, DNA contamination is expected to result in a larger PCR product. The length of intron 3 is 252 base pairs (bp) for *An. minimus* and 295 bp for *An. sinensis*.

Results

The *in silico* screen of 24 *Anopheles* genomes identified two flavivirus-derived EVEs, one in the reference genome sequence of *An. minimus* and one in both genome sequences available for *An. sinensis* (Table 1).

An. minimus EVE

The *An. minimus* EVE is 1,881 bp long (627 amino acid residues) with Nienokoue virus as the closest BLAST hit (44% amino-acid identity). The integrated sequence spans non-structural protein 4A (NS4A), NS4B and NS5 (Figure 1A) without any stop codon. Conserved domain search identified the NS5-methyltransferase domain involved in RNA capping and part of the RNA-directed RNA-polymerase domain. Phylogenetically, the *An. minimus* EVE is sister to the ISF clade (Figure 1B). The *An. minimus* genome supercontig where the EVE was detected is 2,043 bp long and consists almost exclusively of the EVE sequence.

Presence of the EVE was verified *in vivo* by PCR on genomic DNA (Figure 2A), followed by amplicon sequencing to confirm identity. The *An. minimus* EVE was found in both male and female genomic DNA, and was transcriptionally expressed for all combinations of sex and development stages tested (Figure 2B). Evidence for transcriptional activity of the *An. minimus* EVE was confirmed in published RNA-seq data (Figure S1A).

An. sinensis EVE

The *An. sinensis* EVE was detected in two distinct genome sequences that are available for this mosquito species, derived from a Korean and a Chinese strain, respectively. The EVE is 792 bp long (264 amino acid residues) and 799 bp long (266 amino acid residues) for the Korean and Chinese strains, respectively (Table 1). The closest BLAST hit is *Culex* flavivirus (45% amino-acid identity for the Korean strain, 44% amino-acid identity for the Chinese strain). The integrated sequence corresponds to the middle part of NS3 (Figure 3A) and contains six and eight stop codons in the Korean and Chinese strains, respectively. Conserved domain search identified the presence of a P-loop containing the nucleoside triphosphate hydrolase domain found in the NS3 protein of exogenous flaviviruses. Phylogenetically, the *An. sinensis* EVE is sister to the ISF clade (Figure 3B).

The *An. sinensis* supercontig containing the EVE is 2,797 bp long for the Korean strain and 7,380 bp long for the Chinese strain. Analyses of flanking regions revealed the presence of another EVE in the same orientation, upstream of the flavivirus-derived EVE in both the Korean and the Chinese strains. The closest BLAST hit of this additional EVE is Xincheng mosquito virus (43% and 42% amino-acid identity for the Korean and Chinese strains, respectively), an unclassified, negative-sense, single-stranded RNA virus detected in a pool of Chinese mosquitoes including *An. sinensis* specimens (Table S3). BLAST and conserved domain search identified a class II Mariner-like transposase close to a mariner mos1 element, approximately 1,000 bp downstream of the flavivirus-derived EVE (Table S3). This was only the case for the Chinese strain because the supercontig of the Korean strain was not long enough.

Presence of the *An. sinensis* EVE was verified *in vivo* by PCR on genomic DNA from the Korean strain (Figure 4A), followed by amplicon sequencing to confirm identity. The *An. sinensis* EVE was found in both male and female genomic DNA, and was transcriptionally expressed, although less abundantly than the *An. sinensis* EVE, for all combinations of sex and development stages tested, especially in L_4 larvae (Figure 4B). Low expression observed for the *An. sinensis* EVE is consistent with barely detectable transcriptional activity in published RNA-seq data (Figure S1B).

Discussion

ISFs have attracted substantial interest in recent years after some of them were shown to enhance or suppress the replication of medically important flaviviruses in co-infected mosquito cells (Blitvich and Firth 2015). Over a dozen of ISFs have been formally identified to date, mainly in *Aedes* and *Culex* genera of the Culicinae subfamily (Blitvich and Firth 2015). ISFs were also reported in *Anopheles* mosquitoes of the Anophelinae subfamily (Aranda et al. 2009; Zuo et al. 2014; Liang et al. 2015). However, these *Anopheles*-associated ISFs are thought to infect a broad range of hosts including several mosquito species, mainly in the *Culex* genus, and are phylogenetically related to *Culex*-associated ISFs. Therefore, it is unclear whether *Anopheles* mosquitoes are true natural hosts of flaviviruses. Detection of ISFs in field-caught mosquitoes could result from incidental infection, or from a laboratory artifact. In this study, we discovered flavivirus-derived EVEs in the genomes of two *Anopheles* species. Phylogenetic analyses indicated that both EVEs are related to ISFs but belong to a clade that is distinct from *Aedes*-associated and *Culex*-associated ISFs.

Presence of flavivirus-derived EVEs in *Anopheles* genomes supports the hypothesis that *Anopheles* mosquitoes are natural hosts of flaviviruses. Endogenization of non-retroviral RNA viruses is unlikely to occur in the absence of recurrent host-virus interactions over a long evolutionary time scale. Endogenization requires reverse transcription, germ line integration and fixation in the host population, three steps whose combined frequency is exceedingly rare (Holmes 2011; Aiewsakun and Katzourakis 2015). The species-wide frequency of the *An. minimus* EVE is unknown because our *in silico* and *in vivo* analyses were based on the same mosquito strain. Presence of the *An. sinensis* EVE in two mosquito strains from different geographical locations, however, suggests that it could be fixed at the species level. Thus, our discovery of flavivirus-derived EVEs in *Anopheles* genomes is consistent with a long-lasting host-virus interaction between flaviviruses and mosquitoes of the Anophelinae subfamily.

ISFs are thought to be mainly maintained through vertical transmission from an infected female to its offspring (Blitvich and Firth 2015). Vertical transmission is likely to favor co-divergence of pathogens and hosts (Jackson and Charleston 2004), as illustrated by the existence of *Aedes*-associated and *Culex*-associated clades of ISFs (Moureau et al. 2015). Although extrapolation is limited by the scarcity of data on ISF host range and diversity, phylogenetic position of *Anopheles*-associated ISFs as sister to all other known ISFs is consistent with the co-divergence hypothesis. During the evolutionary history of mosquitoes, the Anophelinae diverged from the Culicinae prior to the separation of *Culex* and *Aedes* genera (Reidenbach et al. 2009). Further investigations will be necessary to determine whether an *Anopheles*-associated clade of exogenous ISFs exists, or existed.

Non-retroviral EVEs are thought to be generated by interaction of exogenous viruses with endogenous retro-elements, either with or without long terminal repeats (LTR) (Holmes 2011). The short size of the supercontigs containing the *An. minimus* and *An. sinensis* EVEs limited our ability to investigate the integration mechanism(s). Another EVE sequence that we identified close to the flavivirus-derived EVE in *An. sinensis* may point to an EVE hotspot.

Sequence conservation of the *An. minimus* EVE (i.e., absence of stop codon across 1,881 bp) is consistent with a recent integration or a more ancient integration followed by non-neutral evolution. Our observation that the *An. minimus* EVE is abundantly transcribed may reflect a selective advantage for the host (Holmes 2011). Transcriptionally active EVEs have been suggested to confer protection or tolerance against related exogenous viruses (Flegel 2009; Holmes 2011; Aswad and Katzourakis 2012; Bell-Sakyi and Attoui 2013; Fujino et al. 2014). Despite the lack of empirical evidence so far, flavivirus-derived EVEs could contribute to antiviral immunity and arbovirus vector competence in mosquitoes.

Acknowledgements

This work was supported by the French Government's Investissement d'Avenir program Laboratoire d'Excellence Integrative Biology of Emerging Infectious Diseases grant ANR-10-LABX-62-IBEID, and the City of Paris Emergence(s) program in Biomedical Research. S.L. was supported by a doctoral fellowship from University Pierre and Marie Curie. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

The authors thank Inge Holm and Guillaume Carissimo for providing *Anopheles stephensi* nucleic acids, Albin Fontaine for help with the *in silico* screen, Davy Jiolle and Elliott Miot for technical assistance, Clément Gilbert for advice and critical reading of an earlier version of the manuscript, and members of the Lambrechts lab for insightful comments and discussions.

References

Aiewsakun P, Katzourakis A. 2015. Endogenous viruses: Connecting recent and ancient viral evolution. Virology 479-480:26–37.

Aranda C, Sánchez-Seco MP, Cáceres F, Escosa R, Gálvez JC, Masià M, Marqués E, Ruíz S, Alba A, Busquets N, et al. 2009. Detection and monitoring of mosquito flaviviruses in Spain between 2001 and 2005. Vector Borne Zoonotic Dis. 9:171–178.

Aswad A, Katzourakis A. 2012. Paleovirology and virally derived immunity. Trends Ecol. Evol. (Amst.) 27:627–636.

Bell-Sakyi L, Attoui H. 2013. Endogenous tick viruses and modulation of tick-borne pathogen growth. Front Cell Infect Microbiol 3:25.

Bernard KA, Maffei JG, Jones SA, Kauffman EB, Ebel G, Dupuis AP, Ngo KA, Nicholas DC, Young DM, Shi PY, et al. 2001. West Nile virus infection in birds and mosquitoes, New York State, 2000. Emerging Infect. Dis. 7:679–685.

Blitvich BJ, Firth AE. 2015. Insect-specific flaviviruses: a systematic review of their discovery, host range, mode of transmission, superinfection exclusion potential and genomic organization. Viruses 7:1927–1959.

Calzolari M, Bonilauri P, Bellini R, Albieri A, Defilippo F, Tamba M, Tassinari M, Gelati A, Cordioli P, Angelini P, et al. 2013. Usutu virus persistence and West Nile virus inactivity in the Emilia-Romagna region (Italy) in 2011. PLoS ONE 8:e63978.

Calzolari M, Bonilauri P, Bellini R, Caimi M, Defilippo F, Maioli G, Albieri A, Medici A, Veronesi R, Pilani R, et al. 2010. Arboviral survey of mosquitoes in two northern Italian regions in 2007 and 2008. Vector Borne Zoonotic Dis. 10:875–884.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973.

Chen X-G, Jiang X, Gu J, Xu M, Wu Y, Deng Y, Zhang C, Bonizzoni M, Dermauw W, Vontas J, et al. 2015. Genome sequence of the Asian Tiger mosquito, Aedes albopictus, reveals insights into its biology, genetics, and evolution. Proc. Natl. Acad. Sci. U.S.A. 112:E5907–E5915.

Crochu S, Cook S, Attoui H, Charrel RN, De Chesse R, Belhouchet M, Lemasson J-J, de Micco P, De Lamballerie X. 2004. Sequences of flavivirus-related RNA viruses persist in DNA form integrated in the genome of Aedes spp. mosquitoes. J. Gen. Virol. 85:1971–1980.

Feng Y, Fu S, Zhang H, Li M, Zhou T, Wang J, Zhang Y, Wang H, Tang Q, Liang G. 2012. Distribution of mosquitoes and mosquito-borne viruses along the China-Myanmar border in Yunnan Province. Jpn. J. Infect. Dis. 65:215–221.

Feschotte C, Gilbert C. 2012. Endogenous viruses: insights into viral evolution and impact on host biology. Nat. Rev. Genet. 13:283–296.

Filipe AR. 1972. Isolation in Portugal of West Nile virus from Anopheles maculipennis mosquitoes. Acta Virol. 16:361.

Flegel TW. 2009. Hypothesis for heritable, anti-viral immunity in crustaceans and insects. Biol. Direct 4:32.

Fujino K, Horie M, Honda T, Merriman DK, Tomonaga K. 2014. Inhibition of Borna disease virus replication by an endogenous bornavirus-like element in the ground squirrel genome. Proc. Natl. Acad. Sci. U.S.A. 111:13175–13180.

Holmes EC. 2011. The evolution of endogenous viral elements. Cell Host Microbe 10:368–377.

Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, Nusskern DR, Wincker P, Clark AG, Ribeiro JMC, Wides R, et al. 2002. The genome sequence of the malaria mosquito Anopheles gambiae. Science 298:129–149.

Jackson AP, Charleston MA. 2004. A cophylogenetic perspective of RNA-virus evolution. Mol. Biol. Evol. 21:45–57.

Katoh K, Misawa K, Kuma K-I, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30:3059–3066.

Katzourakis A, Gifford RJ. 2010. Endogenous viral elements in animal genomes. PLoS Genet. 6:e1001191.

Kemenesi G, Krtinić B, Milankov V, Kutas A, Dallos B, Oldal M, Somogyi N, Nemeth V, Banyai K, Jakab F. 2014. West Nile virus surveillance in mosquitoes, April to October 2013, Vojvodina province, Serbia: implications for the 2014 season. Euro Surveill. 19:20779.

Kulasekera VL, Kramer L, Nasci RS, Mostashari F, Cherry B, Trock SC, Glaser C, Miller JR. 2001. West Nile virus infection in mosquitoes, birds, horses, and humans, Staten Island, New York, 2000. Emerging Infect. Dis. 7:722–725.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9:357–U54.

Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database Collaboration. 2011. The sequence read archive. Nucleic Acids Res. 39:D19–D21.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659.

Liang W, He X, Liu G, Zhang S, Fu S, Wang M, Chen W, He Y, Tao X, Jiang H, et al. 2015. Distribution and phylogenetic analysis of Culex flavivirus in mosquitoes in China. Arch.

Virol. 160:2259–2268.

Liu H, Lu H-J, Liu Z-J, Jing J, Ren J-Q, Liu Y-Y, Lu F, Jin N-Y. 2013. Japanese encephalitis virus in mosquitoes and swine in Yunnan province, China 2009-2010. Vector Borne Zoonotic Dis. 13:41–49.

Logue K, Small ST, Chan ER, Reimer L, Siba PM, Zimmerman PA, Serre D. 2015. Whole-genome sequencing reveals absence of recent gene flow and separate demographic histories for Anopheles punctulatus mosquitoes in Papua New Guinea. Mol. Ecol. 24:1263–1274.

Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, Geer RC, He J, Gwadz M, Hurwitz DI, et al. 2015. CDD: NCBI's conserved domain database. Nucleic Acids Res. 43:D222–D226.

Moureau G, Cook S, Lemey P, Nougairede A, Forrester NL, Khasnatinov M, Charrel RN, Firth AE, Gould EA, De Lamballerie X. 2015. New insights into flavivirus evolution, taxonomy and biogeographic history, extended by analysis of canonical and alternative coding sequences. PLoS ONE 10:e0117849.

Neafsey DE, Waterhouse RM, Abai MR, Aganezov SS, Alekseyev MA, Allen JE, Amon J, Arcà B, Arensburger P, Artemov G, et al. 2015. Mosquito genomics. Highly evolvable malaria vectors: the genomes of 16 Anopheles mosquitoes. Science 347:1258522–1258522.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842.

Reidenbach KR, Cook S, Bertone MA, Harbach RE, Wiegmann BM, Besansky NJ. 2009. Phylogenetic analysis and temporal diversification of mosquitoes (Diptera: Culicidae) based on nuclear genes and morphology. BMC Evol. Biol. 9:298.

Roiz D, Vázquez A, Seco MPS, Tenorio A, Rizzoli A. 2009. Detection of novel insect flavivirus sequences integrated in Aedes albopictus (Diptera: Culicidae) in Northern Italy. Virol. J. 6:93.

Sánchez R, Serra F, Tárraga J, Medina I, Carbonell J, Pulido L, de María A, Capella-Gutiérrez S, Huerta-Cepas J, Gabaldón T, et al. 2011. Phylemon 2.0: a suite of web-tools for molecular evolution, phylogenetics, phylogenomics and hypotheses testing. Nucleic Acids Res. 39:W470–W474.

Zhou D, Zhang D, Ding G, Shi L, Hou Q, Ye Y, Xu Y, Zhou H, Xiong C, Li S, et al. 2014. Genome sequence of Anopheles sinensis provides insight into genetics basis of mosquito competence for malaria parasites. BMC Genomics 15.

Zuo S, Zhao Q, Guo X, Zhou H, Cao W, Zhang J. 2014. Detection of Quang Binh virus from mosquitoes in China. Virus Res. 180:31–38.

Figures

Figure 1: Discovery of a flavivirus-derived EVE in *An. minimus*. (A) EVE location in a generic *Flavivirus* genome. Positioning is based on the genome sequence of Nienokoue virus (GenBank accession no. JQ957875). C=capsid protein, E=envelope glycoprotein, M=membrane glycoprotein, NS1=non-structural glycoprotein 1; NS2A=non-structural protein 2A; NS2B= non-structural protein 2B; NS3=non-structural protein 3 (protease/helicase); NS4A=non-structural protein 4A; NS4B=non-structural protein 4B; NS5=non-structural protein 5 (RNA-dependent RNA polymerase). (B) Phylogenetic relationships of the newly discovered *An. minimus* flavivirus-derived EVE with exogenous flaviviruses. Maximum likelihood trees were constructed based on the translated EVE sequence. Clades are color-coded according to known host specificity: green, ISFs; purple, tick-borne arboviruses; black, ‘not known vector’ (vertebrate specific); blue, mosquito-borne arboviruses; red: EVEs. Scale bar indicates the number of substitutions. Node values represent Shimodaira-Hasegawa (SH)-like branch support (only values > 0.8 are shown).

Figure 2: *In vivo* detection of the *An. minimus* flavivirus-derived EVE. (A) EVE detection in genomic DNA. Lane 1: size marker; lane 2: amplified genomic DNA from a pool of 10 *An. minimus* adult females; lane 3: amplified genomic DNA from a pool of 10 *An. minimus* adult males; lane 4: amplified genomic DNA from a pool of 10 *An. sinensis* adult females; lane 5: amplified genomic DNA from a pool of 10 *An. sinensis* adult males; lane 6: amplified DNA from a pool of 10 *An. stephensi* females; lane 7: no template control (NTC). (B) EVE detection in total RNA. Lane 1: size marker; lanes 2 and 3: amplified cDNA from pools of 5 adult females; lanes 4 and 5: amplified cDNA from pools of 5 adult males; lanes 6 and 7: amplified cDNA from pools of 5 L_4 larvae; lane 8: amplified DNA from a pool of 10 females; lane 9: amplified cDNA from a pool of 5 *An. stephensi* females; lane 10: DNA contamination

control (no reverse transcription) using the same pool of 5 adult females as lane 2; lane 11: no template control (NTC). First row: EVE; second row: RPS7 (control gene). The RPS7 target DNA sequence includes an intron, so that DNA contamination is expected to result in a larger PCR product.

Figure 3: Discovery of a flavivirus-derived EVE in *An. sinensis*. (A) EVE location in a generic *Flavivirus* genome. Positioning is based on the genome sequence of *Culex* flavivirus (GenBank accession no. JQ308188). C=capsid protein, E=envelope glycoprotein, M=membrane glycoprotein, NS1=non-structural glycoprotein 1; NS2A=non-structural protein 2A; NS2B= non-structural protein 2B; NS3=non-structural protein 3 (protease/helicase); NS4A=non-structural protein 4A; NS4B=non-structural protein 4B; NS5=non-structural protein 5 (RNA-dependent-RNA polymerase). (B) Phylogenetic relationships of the newly discovered *An. sinensis* flavivirus-derived EVEs with exogenous flaviviruses. Maximum likelihood trees were constructed based on the translated EVE sequence. Clades are color-coded according to known host specificity: green, ISFs; purple, tick-borne arboviruses; black, 'not known vector' (vertebrate specific); blue, mosquito-borne arboviruses; red: EVEs. Scale bar indicates the number of substitutions. Node values represent Shimodaira-Hasegawa (SH)-like branch support (only values > 0.8 are shown).

Figure 4: *In vivo* detection of the *An. sinensis* flavivirus-derived EVE. (A) EVE detection in genomic DNA from the Korean strain of *An. sinensis*. Lane 1: size marker; lane 2: amplified genomic DNA from a pool of 10 *An. minimus* adult females; lane 3: amplified genomic DNA from a pool of 10 *An. minimus* adult males; lane 4: amplified genomic DNA from a pool of 10 *An. sinensis* adult females; lane 5: amplified genomic DNA from a pool of 10 *An. sinensis* adult males; lane 6: amplified DNA from a pool of 10 *An. stephensi* females; lane 7: no template control (NTC). (B) EVE detection in total RNA from the Korean strain of

An. sinensis. Lane 1: size marker; lanes 2 and 3: amplified cDNA from pools of 5 adult females; lanes 4 and 5: amplified cDNA from pools of 5 adult males; lanes 6 and 7: amplified cDNA from pools of 5 L_4 larvae; lane 8: amplified DNA from a pool of 10 females; lane 9: amplified cDNA from a pool of 5 *An. stephensi* females; lane 10: DNA contamination control (no reverse transcription) using the same pool of 5 adult females as lane 2; lane 11: no template control (NTC). First row: EVE; second row: RPS7 (control gene). The RPS7 target DNA sequence includes an intron, so that DNA contamination is expected to result in a larger PCR product.

434 **Tables**

435 **Table 1:** *Anopheles* genomes screened in this study.

Species	**GenBank WGS Project**	**Assembly**	**GenBank Assembly ID**
Anopheles albimanus	APCK01	AalbS1	GCA_000349125.1
Anopheles arabiensis	APCN01	AaraD1	GCA_000349185.1
Anopheles atroparvus	AXCP01	AatrE1	GCA_000473505.1
Anopheles christyi	APCM01	AchrA1	GCA_000349165.1
Anopheles coluzzii	ABKP02	AcolM1	GCA_000150765.1
Anopheles culicifacies A	AXCM01	AculA1	GCA_000473375.1
Anopheles darlingi	ADMH02	AdarC3	GCA_000211455.3
Anopheles dirus A	APCL01	AdirW1	GCA_000349145.1
Anopheles epiroticus	APCJ01	AepiE1	GCA_000349105.1
Anopheles farauti	AXCN01	AfarF1	GCA_000473445.1
Anopheles funestus	APCI01	AfunF1	GCA_000349085.1
Anopheles gambiae	AAAB01	AgamP4	GCA_000005575.2
Anopheles koliensis	JXXB01	AKwgs3	GCA_000956275.1
Anopheles maculatus B	AXCL01	AmacM1	GCA_000473185.1
Anopheles melas	ACXO01	AmelC1	GCA_000473525.1
Anopheles merus	AXCQ01	AmerM1	GCA_000473845.1
Anopheles minimus A	APHL01	AminM1	GCA_000349025.1
Anopheles nili	ATLZ01	Anili1	GCA_000439205.1
Anopheles punctulatus	JXXA01	APwgs2	GCA_000956255.1
Anopheles quadriannulatus A	APCH01	AquaS1	GCA_000349065.1
Anopheles sinensis: Sinensis (Korean strain)	AXCK02	AsinS2	GCA_000472065.2
Anopheles sinensis: China (Chinese strain)	ATLV01	AsinC2	GCA_000441895.2

Anopheles stephensi: Indian	ALPR002	AsteI2	GCA_000300775.2
Anopheles stephensi: SDA-500	APCG01	AsteS1	GCA_000349045.1

WGS: Whole-genome shotgun sequencing

Table 2: Newly described flavivirus-derived EVEs in *An. minimus* and *An. sinensis* genomes.

Element	GenBank accession no.	Protein identity	Supercontig accession no.	Supercontig length (bp)	Coordinates			
					Supercontig		Viral genome*	
					Start	End	Start	End
An. minimus EVE	BK009978	NS4-NS5	KB664005.1	2,043	107	1,987	6,567	8,432
An. sinensis EVE (Korean strain)	BK009979	NS3	AXCK02024744	2,797	822	1,613	5,214	5,822
An. sinensis EVE (Chinese strain)	BK009980	NS3	ATLV01019207.1	7,380	3,578	4,376	5,214	5,822

* Viral genomes coordinates are based on the closest tBLASTn hits: Nienokoue virus (GenBank accession no. JQ957875) for the *An. minimus* EVE and *Culex* flavivirus (GenBank accession no. JQ308188) for the *An. sinensis* EVE.

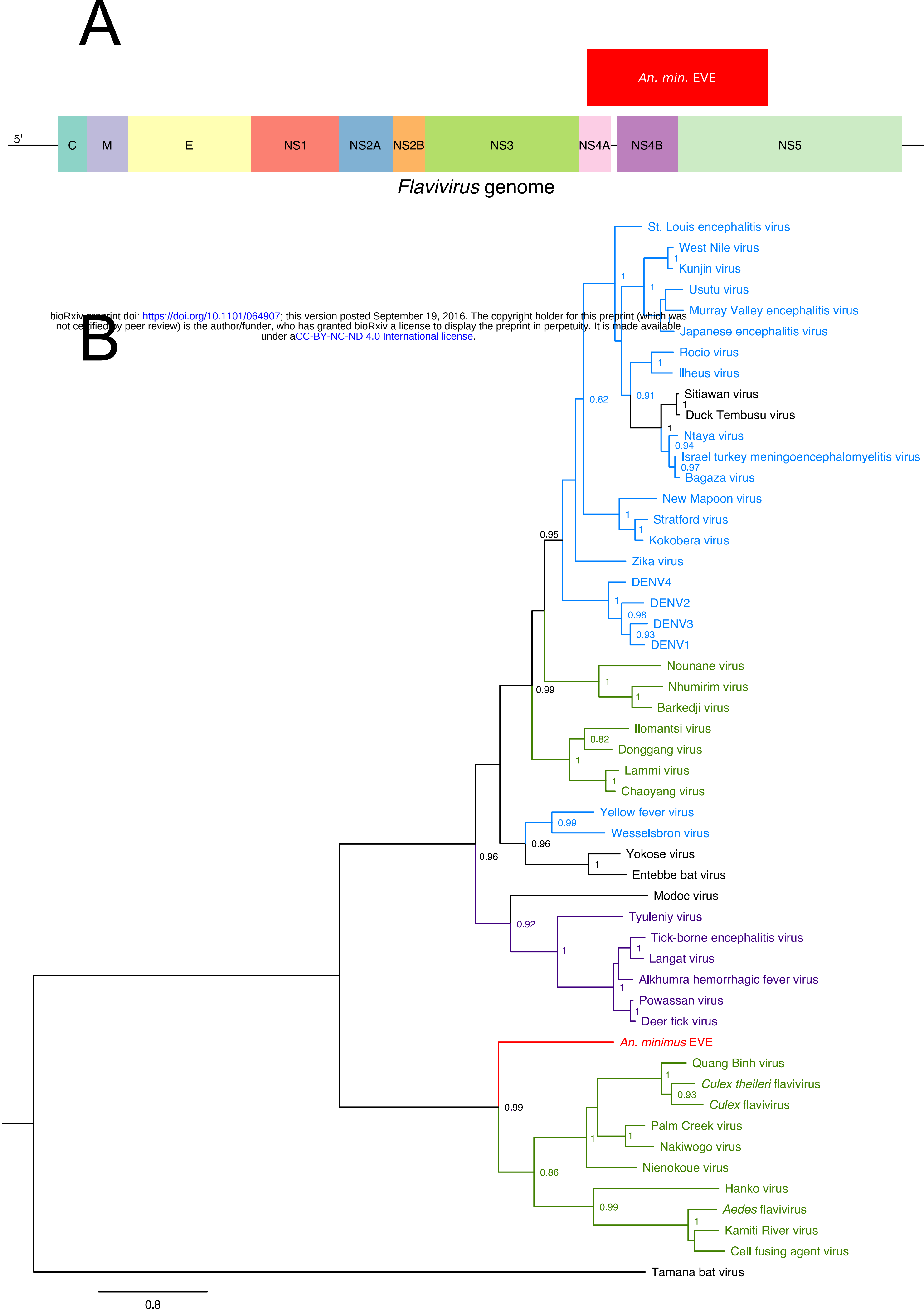
Supporting information

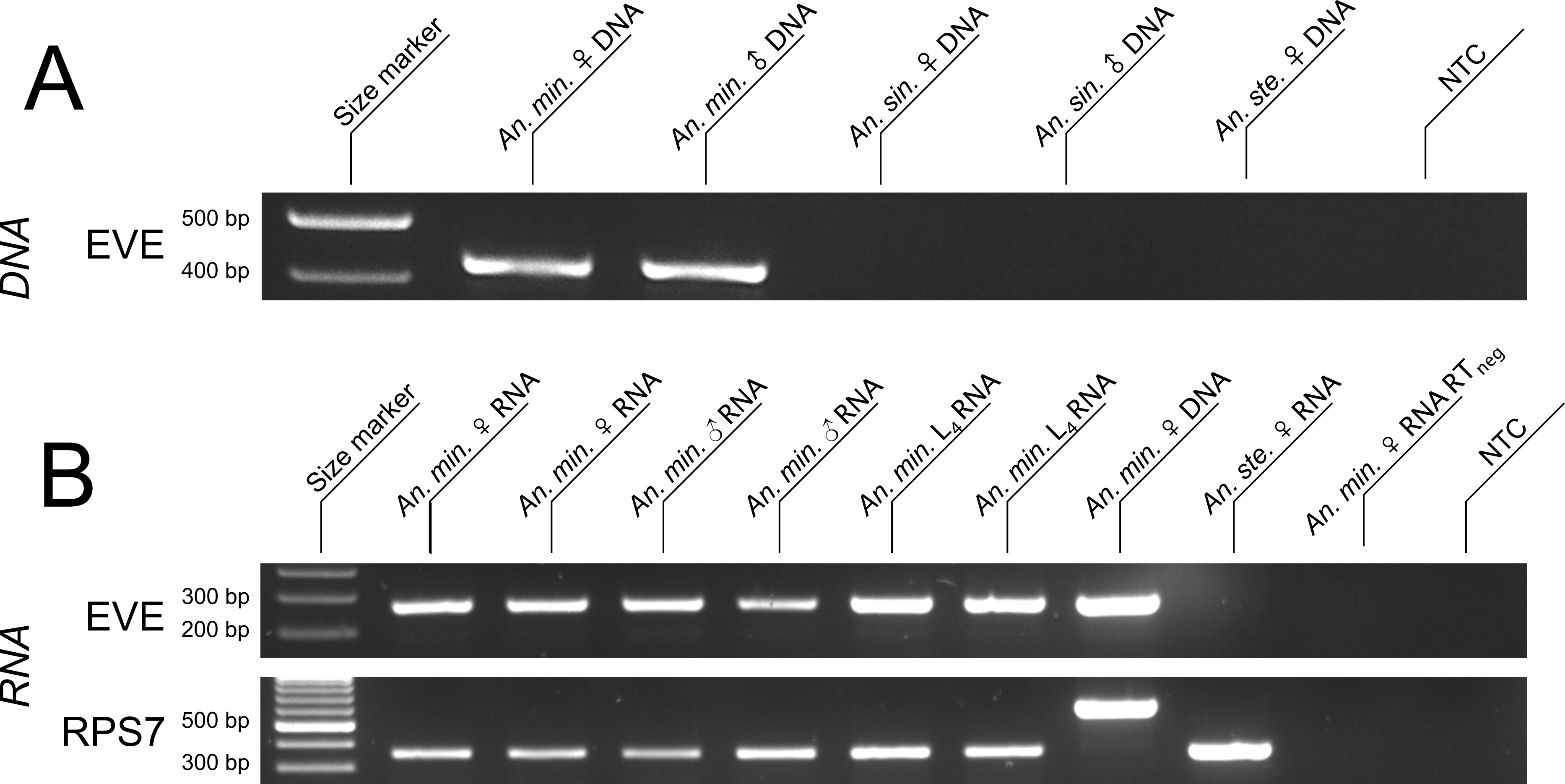
Table S1: Names and accession numbers of flavivirus polyprotein sequences used as queries for the *Anopheles* genome screen and included in the EVE phylogenetic analyses.

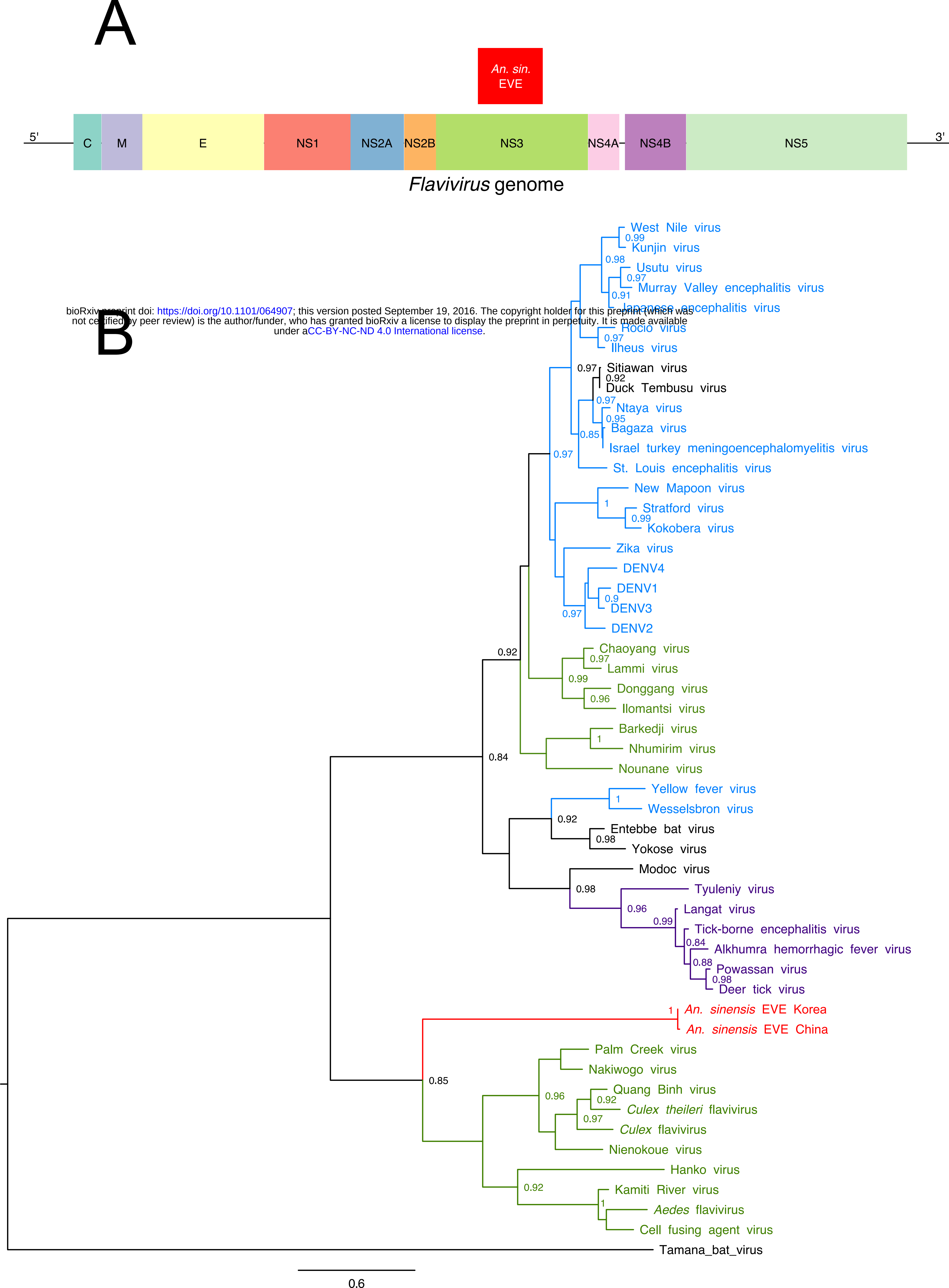
Table S2: PCR primers and reaction conditions used to amplify EVE and RPS7 targets from DNA or cDNA templates.

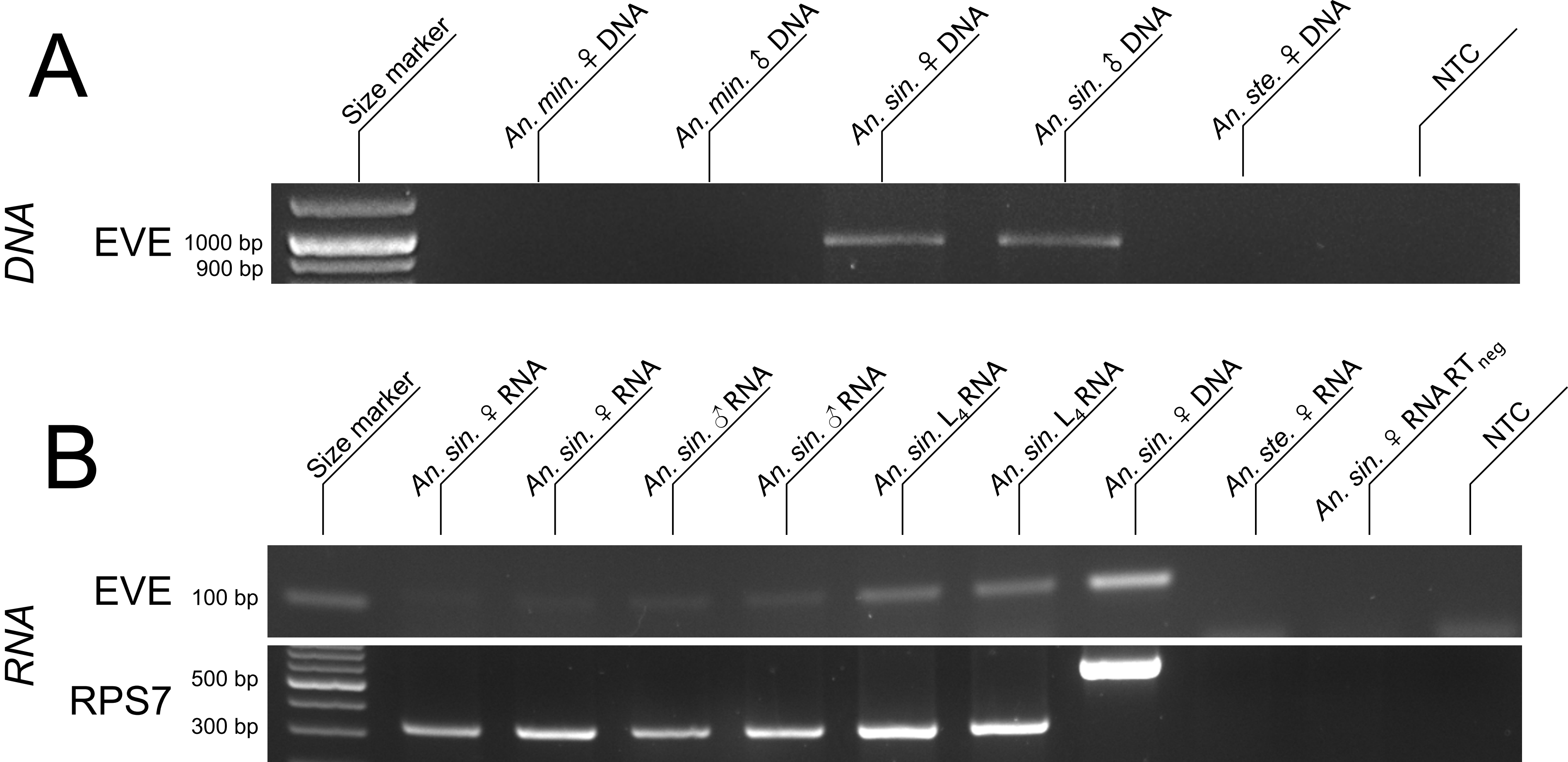
Table S3: Other genetic integrations detected in the *An. sinensis* genome supercontigs containing the flavivirus-derived EVE.

Figure S1: Sequencing coverage of (A) *An. minimus* and (B) *An. sinensis* EVEs in published RNA-seq experiments. For each experiment, the Sequence Read Archive (NCBI) accession number is indicated above the graph.


A
An. min. EVE
5'
C
M
E
NS1
NS2A
NS2B
NS3
NS4A
NS4B
NS5
3'
Flavivirus genome
B
St. Louis encephalitis virus
West Nile virus
Kunjin virus
Usutu virus
Murray Valley encephalitis virus
Japanese encephalitis virus
Rocio virus
Ilheus virus
Sitiawan virus
Duck Tembusu virus
Ntaya virus
Israel turkey meningoencephalomyelitis virus
Bagaza virus
New Mapoon virus
Stratford virus
Kokobera virus
Zika virus
DENV4
DENV2
DENV3
DENV1
Nounane virus
Nhumirim virus
Barkedji virus
Ilomantsi virus
Donggang virus
Lammi virus
Chaoyang virus
Yellow fever virus
Wesselsbron virus
Yokose virus
Entebbe bat virus
Modoc virus
Tyuleniy virus
Tick-borne encephalitis virus
Langat virus
Alkhumra hemorrhagic fever virus
Powassan virus
Deer tick virus
An. minimus EVE
Quang Binh virus
Culex theileri flavivirus
Culex flavivirus
Palm Creek virus
Nakiwogo virus
Nienokoue virus
Hanko virus
Aedes flavivirus
Kamiti River virus
Cell fusing agent virus
Tamana bat virus
1
1
1
1
0.82
0.91
1
1
0.94
0.97
1
1
0.95
1
0.98
0.93
1
1
0.99
0.82
1
1
0.99
0.96
1
0.96
0.92
1
1
1
1
1
0.93
0.99
1
1
0.86
0.99
1
0.8


A
Size marker
An. min. ♀ DNA
An. min. ♂ DNA
An. sin. ♀ DNA
An. sin. ♂ DNA
An. ste. ♀ DNA
NTC
DNA
EVE
500 bp
400 bp
B
Size marker
An. min. ♀ RNA
An. min. ♀ RNA
An. min. ♂ RNA
An. min. ♂ RNA
An. min. L_4 RNA
An. min. L_4 RNA
An. min. ♀ DNA
An. ste. ♀ RNA
An. min. ♀ RNA RT_{neg}
NTC
RNA
EVE
300 bp
200 bp
RPS7
500 bp
300 bp


A
An. sin.
EVE
5'
C
M
E
NS1
NS2A
NS2B
NS3
NS4A
NS4B
NS5
3'
Flavivirus genome
B
West Nile virus
0.99
Kunjin virus
0.98
Usutu virus
0.97
Murray Valley encephalitis virus
0.91
Japanese encephalitis virus
Rocio virus
0.97
Ilheus virus
0.97
Sitiawan virus
0.92
Duck Tembusu virus
0.97
Ntaya virus
0.95
Bagaza virus
0.85
Israel turkey meningoencephalomyelitis virus
0.97
St. Louis encephalitis virus
New Mapoon virus
1
Stratford virus
0.99
Kokobera virus
Zika virus
DENV4
DENV1
0.9
DENV3
0.97
DENV2
0.92
Chaoyang virus
0.97
Lammi virus
0.99
Donggang virus
0.96
Ilomantsi virus
Barkedji virus
1
Nhumirim virus
0.84
Nounane virus
Yellow fever virus
1
Wesselsbron virus
0.92
Entebbe bat virus
0.98
Yokose virus
Modoc virus
0.98
Tyuleniy virus
0.96
Langat virus
0.99
Tick-borne encephalitis virus
0.84
Alkhumra hemorrhagic fever virus
0.88
Powassan virus
0.98
Deer tick virus
1
An. sinensis EVE Korea
An. sinensis EVE China
Palm Creek virus
0.85
Nakiwogo virus
Quang Binh virus
0.96
0.92
Culex theileri flavivirus
0.97
Culex flavivirus
Nienokoue virus
Hanko virus
0.92
Kamiti River virus
1
Aedes flavivirus
Cell fusing agent virus
Tamana_bat_virus
0.6


A
Size marker
An. min. ♀ DNA
An. min. ♂ DNA
An. sin. ♀ DNA
An. sin. ♂ DNA
An. ste. ♀ DNA
NTC
DNA
EVE
1000 bp
900 bp
B
Size marker
An. sin. ♀ RNA
An. sin. ♀ RNA
An. sin. ♂ RNA
An. sin. ♂ RNA
An. sin. L_4 RNA
An. sin. L_4 RNA
An. sin. ♀ DNA
An. ste. ♀ RNA
An. sin. ♀ RNA RT_{neg}
NTC
EVE
100 bp
RNA
RPS7
500 bp
300 bp