UNIVERSITÀ
DI PAVIA

Dipartimento di Biologia e Biotecnologie "L. Spallanzani"

An improved genome for the Asian tiger mosquito *Aedes albopictus* and its applications in studying endogenous viral sequences

Umberto Palatini

Dottorato di Ricerca in
Genetica, Biologia Molecolare e Cellulare
Ciclo XXXIII – A.A. 2017-2020.

UNIVERSITÀ
DI PAVIA

Dipartimento di Biologia e Biotecnologie "L. Spallanzani"

An improved genome for the Asian tiger mosquito *Aedes albopictus* and its applications in studying endogenous viral sequences

Supervised by Prof. Mariangela Bonizzoni

Umberto Palatini

Dottorato di Ricerca in
Genetica, Biologia Molecolare e Cellulare
Ciclo XXXIII – A.A. 2017-2020.

Abstract

Mosquito-borne diseases, including arboviral diseases such as Dengue, Chikungunya and Zika, have increased their world-wide incidence in the past 50 years and currently account for about 17% of all infectious diseases globally. Urbanization, globalization, increased international mobility and the widespread distribution of the main arboviral vectors, the mosquitoes *Aedes aegypti* and *Aedes albopictus* are all factors that have contributed to the (re)-emergence of arboviral diseases. Effective therapies and vaccines are limited for most arboviruses. Accordingly, vector control is the primary way to prevent transmission of arboviruses to humans. A deep understanding of the interaction and co-evolution between viruses and mosquito vectors is expected to aid in the development of novel transmission control strategies. The capacity of mosquitoes to support viral replication and transmission is called vector competence and is a dynamic and variable trait affected by many factors, suggesting an "arms race" between mosquitoes and viruses. The main arboviral vector in Europe is the invasive species *Ae. albopictus*, which has increasingly received attention from the scientific community due to its quick worldwide spread from south East Asia in the past 50 years. Nonretroviral RNA endogenous viral elements (nrEVEs) with similarities to the non-retroviral RNA viruses of the *Flaviviridae* and *Rhabdoviridae* family have been found with high frequency in *Ae. albopictus* mosquitoes. These viral integrations often interact with the most-recently characterized of the three RNA interference (RNAi) pathways: the PIWI-interacting RNA (piRNA) pathway. Most, but not all, nrEVEs in the genome of *Ae. albopictus* map next to transposable elements (TEs) fragments in piRNA clusters and produce piRNAs, small molecules that associate with Argonaute proteins of the PIWI clade to silence TEs based on sequence complementarity to piRNAs. In addition to its canonical role in preserving genome integrity, the piRNA pathway has antiviral activity in *Aedes* spp. mosquitoes. Despite the abundance of nrEVEs the biology, functional role, and patterns of integration in wild mosquito populations are still relatively unexplored. nrEVEs could be expressed and influence the phenotype of mosquitoes acting as a form of host antiviral immunity. A difficulty in studying the *Ae. albopictus* genome was the absence of a high-quality reference genome. During my PhD, I focused my attention on *Ae. albopictus* to primarily improve knowledge of its genome. I contributed to the sequencing, assembly, and annotation of a new reference genome for *Ae. albopictus* based on long-reads sequencing technologies; I coordinated an international consortium to annotate and characterize genomic features and their expression and produce a physical map of the genome. The availability of this new assembly allowed me to ask more specific questions on the landscape of viral integrations. A newly developed bioinformatic pipeline was combined with molecular biology techniques to identify viral sequences integrated into the genomes of wild-collected mosquitoes from different geographical locations. I also used the new genome assembly to reconstruct and identify RNA viruses in mosquito small-RNA sequencing data. I correlated the viruses identified in the mosquitoes with their population-specific

nrEVE landscape, under the hypothesis that the pattern of nrEVEs is shaped by exposure to viruses. Lastly, I applied the CRISPR-Cas9 genome editing technology on *Ae. albopictus* embryos to modify a viral integration and a piRNA cluster and test the hypothesis that viral integrations have a role as immunity effectors against cognate viral infections. Overall, results gained through my PhD activities will enhance our understanding on the genome structure of *Ae. albopictus* and the importance of repetitive elements like viral integrations in the context of its biology.

Abbreviations

AeFV = Aedes Flavivirus

BF = Blood Fed

CFAV = Cell Fusing Agent Virus

CFP = Cyan Fluorescent Protein

CHIKV = Chikungunya Virus

CRISPR = Clustered Regularly Interspaced Short Palindromic Repeats

DENV = Dengue Virus

EIP = Extrinsic Incubation Period

EVE = Endogenous Viral Element

F-nrEVE= Flavivirus-derived nrEVE

FISH = Fluorescence In Situ Hybridization

FPA = Foshan Pavia A

FPKM = Fragments Per Kilobase Million

HMW = High Molecular Weight

ISFV = Insect Specific Flavivirus

ISV = Insect Specific Virus

KRV = Kamiti River Virus

LD = Linkage Disequilibrium

LTR = Long Terminal Repeats

NBF = Non-Blood Fed

NGS = Next Generation Sequencing

NRV = Non retroviral RNA virus

ORF = Open Reading Frame

PAMP = Pathogen Associated Molecular Pattern

PBM = Post Blood Meal

PRR = Pattern Recognition Receptor

R-nrEVEs = Rhabdovirus-derived nrEVEs

RISC = RNA-Induced Silencing Complex

RNAi = RNA interference

RPMM = Reads per Million miRNAs

TE = Transposable Element

TLR = Toll Transmembrane Receptor

TPM = Transcripts Per Million

WGS = Whole Genome Sequencing

WT = Wild Type

YFV= Yellow Fever Virus

ZIKV = Zika Virus

dsRNA = Double Stranded RNA

miRNA = micro-RNA

nrEVE = Nonretroviral Endogenous Viral Element

piRNA = piwi-interacting RNA

sRNA = small RNA

siRNA = small-interfering RNA

tRNA = RNA transfer

vDNA = Viral DNA

Table of contents

1. Introduction

1.1 General characteristics of *Aedes* spp. mosquitoes

Mosquito is a generic term used to identify insect species sharing common morphological and genetic features, grouped in the Culicidae family, under the Diptera order. The family includes 3,574 existing species classified into two subfamilies, Culicinae and Anophelinae, and 113 genera. The Anophelinae subfamily includes *Anopheles* mosquitoes, which are responsible for the transmission of *Plasmodium* protozoa, the etiological agents of Malaria. Under the Culicinae subfamily, the *Aedes* genus contains the *Stegomyia* subgenus, which include 128 species (Mosquito Taxonomic inventory, as of 2020). Amongst these 128 species, the most important from a medical perspective are undoubtedly *Aedes aegypti* and *Aedes albopictus*. These species are the main vectors for public-health challenging Arthropod-borne viruses (arboviruses) such as Dengue (DENV), Zika (ZIKV), Chikungunya (CHIKV) and Yellow Fever (YFV) viruses. Adult *Aedes* mosquitoes are dark, and their body is divided into a head, thorax, and abdomen. Legs are thin and terminate with a pair of claws. They possess a pair of functional wings and a pair of halteres, knobbed structures derived from wings used to improve flight. *Aedes* spp. show a unique pattern of white/silver scales on the abdomen and thorax and alternating light bands on the legs. Different scale patterns can be used to morphologically distinguish adult *Ae. aegypti* and *Ae. albopictus* mosquitoes. Adult *Ae. albopictus* are 2 to 10 mm long and can be recognized from a median longitudinal line on the thorax formed by silver scales and white/silver scales on tarsi and the palpus. Adult *Ae. aegypti* range from 4 to 7 mm and exhibit a pair of longitudinal white stripes, and a white lyre-shaped marking. The abdomen is composed of 10 segments, but only the first seven or eight are visible. The abdomen is normally thin and elongated, but in females it gets enlarged and oval after a blood-meal and during egg development [1]. The head has a pair of kidney-shaped compound eyes. Between the eyes arises a pair of filamentous and segmented antennae that provide non-visual senses through chemosensory and mechanosensory systems. Apart from being smaller than females, males can be recognized by their plumose antennae that are crucial in locating females [2]. Just below the antennae are a pair of palps (that are longer in the males) and the forward-projecting proboscis, which is efficiently used by females to penetrate the skin of hosts to suck their blood [3]. *Aedes* spp. mosquitoes are holometabolous insects and undergo complete metamorphosis, which means that the organism develops through radically different life stages (egg, larvae, pupae, and adult). In each stage the insect is completely different in morphology, feeding habits and behavior. Adult mosquitoes feed on sugar from rotting fruit, nectar, tree sap and other sources. Adult females require a blood meal to develop eggs.

Females are attracted by warm-blooded vertebrate hosts by a combination of factors including the carbon dioxide emitted by the animal and other chemicals such as lactic acid [4].

A female can lay around 100 eggs close to containers with water, above the waterline. After the water level rises to cover the eggs, eggs hatch into larvae within 48 hours. Larvae feed on microorganisms in the water and undergo four stages (also called instars) of development lasting 2-3 days each before becoming pupae. Pupae do not feed and undergo metamorphosis emerging into adults within 24-48 hours. This life cycle usually completes within 15 days but depending on environmental conditions it can range from 4 days to almost a month [5]. The adult mosquito life span is about 3-4 weeks.

1.2 Geographical distribution of *Aedes albopictus*

Aedes albopictus is a highly invasive species and is currently found throughout tropical and temperate regions of the world ([6] (**Figure 1**).

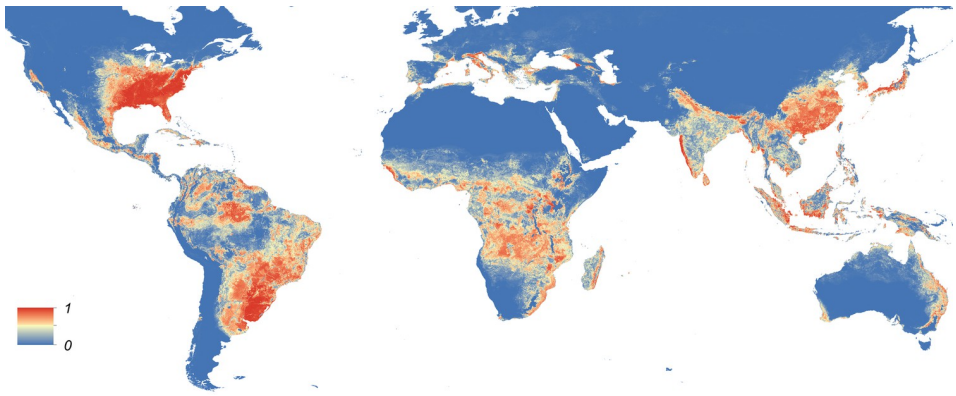


Figure 1: Global map of the predicted distribution of *Ae. albopictus*. The map depicts the probability of occurrence (from 0 blue to 1 red) at a spatial resolution of 5 km × 5 km [10].

The great expansion and evolutionary success of *Ae. albopictus* is largely due to its remarkable ecological plasticity. *Ae. albopictus* can undergo two different types of dormancy: quiescence, and diapause [7]. Quiescence is characterized by slowed metabolism and increased egg resistance to environmental conditions, excessive desiccation in particular. It is a non-seasonal mechanism imposed by harsh

environmental conditions such as high temperature and dry climate. Diapause is a much more complex season-survival strategy, influenced by multiple factors, including the daily amount of light (a sign that winter is imminent). Diapause is genetically determined and hormonally regulated and reduces metabolic activity, morphogenesis, and reproduction. Diapause leads to increased mosquito tolerance to extreme environmental conditions such as the harsh winters of temperate regions [8]. Both quiescence and diapause contribute to the establishment, maintenance and spread of natural mosquito populations, offering a mechanism for surviving unfavorable climate in either very hot or very cold seasons.

Aedes albopictus was originally a zoophilic species living in the tropical forests of East Asia. However, it adapted to easily colonize urban environments where there is the availability of shelter from difficult environmental conditions, water for oviposition and human blood. In the early 1900s *Ae. albopictus* spread to islands of the Indian and Pacific Ocean [9]; a second round of expansion took place in the second half of the twentieth century and rapidly moved *Ae. albopictus* globally and to more temperate regions in Europe, America and Africa [10]. *Aedes albopictus* was first reported in Europe in 1979 in Albania [11]. In 1985 it was reported in Texas (USA) and has since spread in most North American states [12]. From north America, *Ae. albopictus* spread into the south and was detected in Brazil (1986) and Mexico (1988), even though some populations appear to have been introduced from Asia [13, 14]. *Aedes albopictus* was imported to Italy in 1990, through the port of Genoa and since then it spread to different regions, making Italy the most heavily infested country in Europe [15]. Moreover, *Ae. albopictus* populations have become already permanently established in different countries around the Mediterranean, including France, Spain, Greece, Malta, Bosnia, Croatia and even Israel and Turkey (https://ecdc.europa.eu/en/disease-vectors/surveillance-and-disease-data/mosquito-maps). Given the changing climate conditions and given its great capacity of adaptation, it is believed that *Ae. albopictus* will be able to spread also to the northern European countries in the future [6]. As a confirmation of this hypothesis, *Ae. albopictus* has already been detected in southern England [16].

1.3 *Aedes albopictus* as an arboviral vector

Several mosquito species are vectors, organisms that can carry infectious agents from one host to another. The most known and studied mosquito-borne disease is malaria, caused by parasitic protozoa of the *Plasmodium* genus, for which Anopheline mosquitoes are vectors. Aedeine mosquitoes, in primis *Ae. albopictus* and *Ae. aegypti* are vectors for viruses collectively called arboviruses (arthropod-borne viruses). Arboviruses are, with one known exception, RNA viruses. Arboviruses are maintained by a cycle that requires a host (that act as a viral reservoir) and a vector

for transmission to another host [17]. Adult female mosquitoes acquire viral pathogens through an infectious blood. Ingested viral particles must replicate in the midgut and further disseminate throughout the mosquito body to reach salivary glands. Only if/when infection is established in the salivary glands, mosquitoes are able to transmit viruses to a new host, thus continuing the viral replication cycle [18–21] (**Figure 2**).

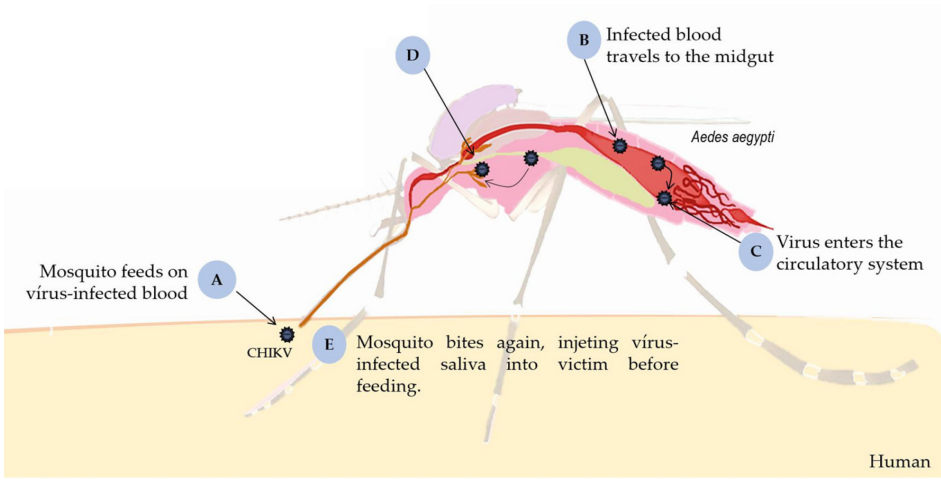


Fig 2: Replication steps of an arbovirus in a mosquito vector [23] .

An infected mosquito remains capable of infecting a host for life. The amount of time between the acquisition of an infected blood-meal to the moment in which the vector is able of transmitting the virus is called extrinsic incubation period (EIP). EIP length depends on multiple factors, including the temperature and the number of virions entering the midgut [17, 22, 23]. In some hosts, low-level viremia inhibits the possibility of a vector getting infected upon blood-feeding on the host. These are called dead-end hosts, and they do not contribute to the viral reservoir [17].

Aedes albopictus has globally emerged as the main vectors for several arboviruses because of its high vector competence for many arboviruses [24, 25]. Vector competence is a parameter that describes the intrinsic capacity of a vector to transmit a specific pathogen. Arboviral vector competence is a complex and variable trait that depends on the relationship between the vector, the host and the arbovirus [24]. Vector competence is influenced by intrinsic factors affecting the host (i.e. mosquito genetics and immune response; the microbiome; salivary gland and midgut barriers) and the virus genetics to reflect a continuous “arms race” between vectors and virus [24, 26].

Vectorial capacity is a wider parameter that describes the likeliness of a mosquito population in a specific area to transmit a specific pathogen [27]. Vectorial capacity is influenced by vector competence and by environmental (abiotic) factors such as temperature, mosquito longevity, density of the mosquito populations, blood-feeding behaviors, and host availability.

Vectorial capacity can be defined mathematically with the formula:

$$VC = ma^2bp^n/-\log_e p$$

where m=number of female mosquitoes per host, a=daily blood feeding rate, b=transmission rate among exposed mosquitoes, p=the probability of daily survival, and n=extrinsic incubation period [27, 28].Vector competence is favored by a high tolerance to infection, as tolerance reduces the negative impact of the virus on the host and permits higher viral levels. On the contrary, vector competence is reduced by an increase in resistance, that blocks viral replication.

RNA viruses lack polymerases with proofreading activity [29]. Their high mutation frequencies, rapid replication, and large population sizes give these viruses a great genomic plasticity and environmental adaptability [29, 30]. However, the alternating host transmission cycle of arboviruses may evolutionarily constrain their ability to adapt to new hosts and vectors as most arboviruses do not evolve quickly [30]. Arboviruses are also subject to fitness declines following repeated bottlenecks [31]. Nevertheless, many aspects of arbovirus evolution and relationship with both their vertebrate host and vector are not clear yet.

1.4 Overview on mosquito-borne arboviruses

Arboviruses are mostly RNA viruses and are assigned to six different viral families: *Flaviviridae*, *Togaviridae*, *Bunyaviridae*, *Rhabdoviridae*, *Orthomyxoviridae* and *Reoviridae* [32, 33]. Currently, a single DNA arbovirus is known: the *Asfarviridae* African swine fever virus. Arboviruses have different genomic and structural features (**Figure 3**) but share the reliance on hematophagous vectors (mosquitoes, ticks and flies) for transmission and the lack of genes encoding for transcriptase and integrase that retroviruses possess, making them non-retroviral RNA viruses (NRVs). The *Flaviviridae* and *Togaviridae* families and the *Bunyavirales* order (formerly the *Bunyaviridae* family) include the most diffused and dangerous mosquito-borne arboviruses. The great heterogeneity in the structural organization and the replication strategies of arboviruses suggest that their dependency on an arthropod-vector for transmission has arisen independently many times during the evolution of arboviruses [25]. Many of the endemic arboviruses in the tropical and subtropical regions of the

world belong to the *Flavivirus* genus in the *Flaviviridae* family. Flaviviruses are enveloped viruses with an icosahedral capsid, about 50 nm in diameter [34]. Their genome is made of single-stranded positive RNA and is 10-11 kb large. The entire genome is translated into a polyprotein that is processed by viral and host proteases to produce structural proteins (required for capsid and envelope formation) and non-structural proteins (involved in replication, immune evasion, and host machinery hijacking).

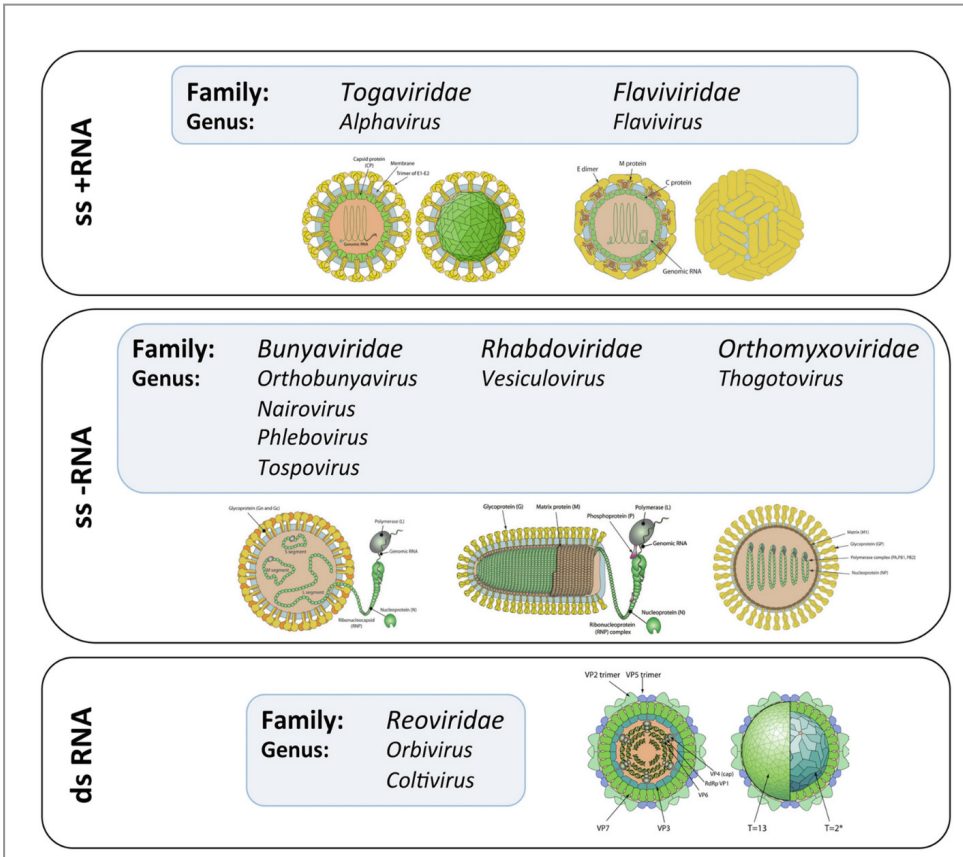


Figure 3: Differences in genome type (positive and negative ssRNA and dsRNA) and physical structures of viral families that include RNA arboviruses. For each genome type, the main arboviral families and genera are shown together with a representative virion for that family [32].

The most public-health relevant *Flavivirus* is Dengue virus (DENV). WHO estimates around 390 million DENV infections per year and predicts around 3.9 billion people worldwide to be at risk of infection with DENV [35, 36]. Although dengue fever is a

historically old disease, only 9 countries registered Dengue epidemics before 1970 [36]. After the quick global expansion of *Aedes* species, the disease is now endemic in more than 100 countries in Africa, the Americas, South-East Asia, the Eastern Mediterranean and the Pacific regions [36]. DENV infection may progress into Dengue hemorrhagic fever that, if not properly managed, can cause circulatory system failure and shock, followed by death. If timely treated, through hospitalization and fluid replacement, mortality may be lower than 1% (CDC Dengue epidemiology, 2020).

Other relevant mosquito-borne flaviviruses are Zika, Yellow Fever, West Nile and Usutu viruses, which cause fever and may provoke severe neurological damages, and the Japanese encephalitis virus.

Viruses grouped in the *Togaviridae* family are enveloped, icosahedral viruses, 65-70 nm in diameter. They have a single stranded positive RNA genome, 10-12 kb large, capped and polyadenylated. The non-structural proteins are expressed as a single polyprotein, while the structural ones are expressed through a subgenomic mRNA [37]. The *Togaviridae* family comprises two genera: *Rubivirus* and *Alphavirus*. The alphavirus Chikungunya virus (CHIKV) is an important human pathogen. In 2016, CHIKV caused 152,796 confirmed cases only in the Americas (PAHO/WHO report 27/01/2017) and a CHIKV outbreak occurred in Italy in 2007, with re-emerging cases almost every year and another outbreak in Lazio-Calabria in 2017 [38]. CHIKV main symptoms are fever, cutaneous rash, fatigue, and a strong pain in the joints that may last weeks or even longer, particularly in old people. Venezuelan equine encephalitis and Eastern equine encephalitis viruses, also belonging to the *Alphavirus* genus, may provoke brain inflammation.

Viruses belonging to the *Bunyavirales* order are enveloped and spherical, 80-120 nm in diameter. They have a negative-stranded RNA linear genome, divided in three segments: L (RNA-dependent RNA polymerase), M (glycoproteins) and S (nucleocapsid). The overall genomic size is around 11-20 kb [39, 40]. The most important arboviruses are in the *Peribunyaviridae* family and are mostly grouped in the *Orthobunyavirus* genus, that includes many mosquito-borne arboviruses such as La Crosse Encephalitis, California Encephalitis, and the Jamestown Canyon viruses. The *Phenuiviridae* family includes few other arboviruses such as Rift Valley fever virus (genus *Phlebovirus*), whose symptoms are mild flu-like illness and hemorrhagic fever in the most severe cases.

1.5 Insect-specific viruses

While arboviruses have a peculiar dual tropism for invertebrate and vertebrate hosts,

several viruses within the same families of arboviruses, are not able to infect vertebrate cells and are restricted to insects. These viral species are called insect-specific viruses (ISVs). The first ISV was identified in 1975 in *Aedes aegypti* and was observed to form cellular syncytia in *Ae. albopictus* cells and thus named Cell Fusing Agent Virus (CFAV) [41]. CFAV belongs to the *Flaviviridae* family and has been detected in several mosquito samples in the last decade. Thanks to the advancement in high-throughput sequencing and metagenomics new ISVs have been discovered in the past few years in different insects, including mosquitoes where they are termed mosquito-specific viruses. The impossibility of ISVs to replicate in vertebrate cells has been assessed through viral inoculation of vertebrate cells and by intracerebral infection of neonatal mice [42, 43]. Host-range restriction can be a consequence of the inability of the virus to enter the cell, deliver their genome, replicate, or build new infectious virus particles. The innate immune system of the vertebrate cell is believed to play a key role in restricting ISVs [44] but there is evidence pointing to restriction at other levels such as viral attachment to the cell, viral RNA replication, virus assembly and release [17, 45]. The main mechanisms of maintenance and transmission of ISVs within insect populations is thought to be vertical transmission, in which the virus is passed transovarially from infected female mosquitoes to their offspring. This is supported by experimental evidence both in laboratory and wild-collected mosquitoes [46–48].

Currently known ISVs are classified in different families, including *Flaviviridae*, *Bunyaviridae*, *Togaviridae* and *Rhabdoviridae* [49, 50]. Also, the discovery of novel ISVs led to the formation of new taxonomical groups such as the genus *Negevirus* [51] and the family *Chuviridae* (ICTV Virus Taxonomy: 2019 Release).

Several observations support the hypothesis that arboviruses are derived from ISVs [52, 53]. Both ISVs and arboviruses infect and replicate in insects and are almost exclusively RNA viruses, lacking an efficient proofreading mechanism which results in higher mutation rates and greater plasticity. Reconstruction of phylogenetic ancestral hosts for the families *Flaviviridae* (**Figure 4**) and *Rhabdoviridae* and for the order *Bunyavirales* provides evidence of ancestral host switching processes. In these taxonomical categories ISVs often branch at the base of the trees, suggesting arthropod origin of arboviruses within the same category [54]. An interesting theory is that ancient insect borne bunyavirus lineage made the jump from insects to mammals and that these viruses comprise the present-day genus *Hantavirus* [55]. Overall, these pieces of evidences suggest that arboviruses could have once been ISVs that overcome the barriers preventing infection of vertebrates (e.g. viral entry blockage, replication and assembly inhibition, interferon response) and expanded their host range to include vertebrate cells [54, 56]. Various studies have shown that ISVs alter the host vectoral capacity through the upregulation of the host antiviral immune responses and through a superinfection exclusion phenomenon, suggesting ISVs may be implemented as new biological control agents against arboviruses [52,

54]. However, most studies on the result of co-infection between arboviruses and ISVs or superinfection are inconclusive or contradict each other, suggesting ISV-arbovirus relationship is species-specific. For instance, the ISV Palm Creek virus and Nhumirim virus can interfere with the replication of the arbovirus West Nile virus, although the responsible mechanisms are not clear yet [54].

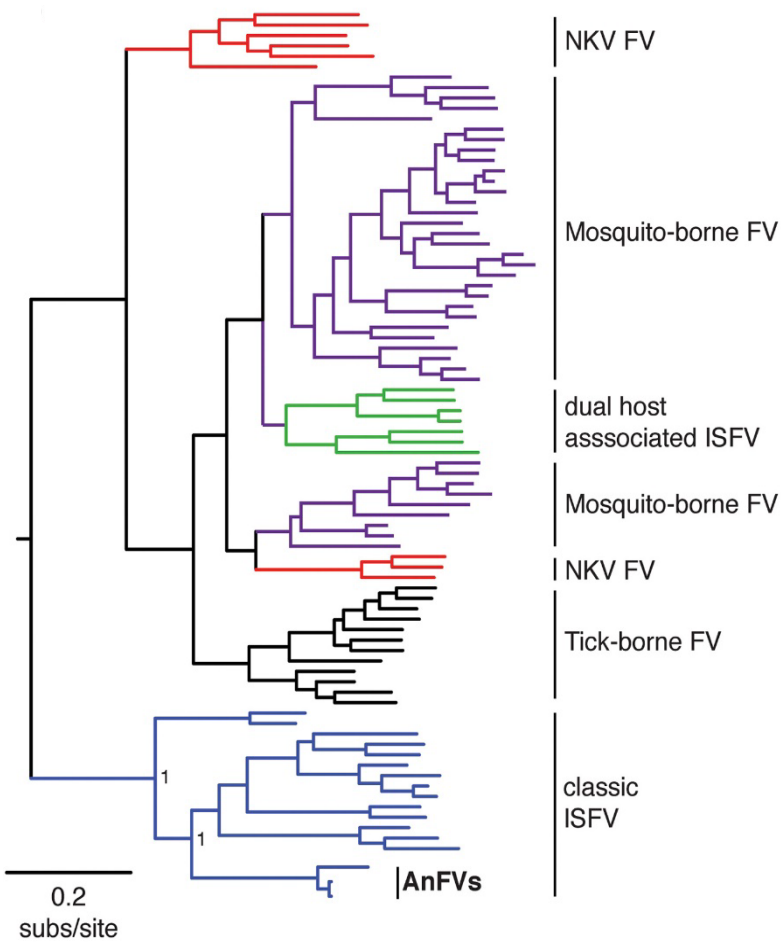


Figure 4. Bayesian phylogeny of Flaviviruses from multiple alignment of NS5 protein sequences. The phylogeny was rooted on the branch to Tamana bat virus (not shown). ISVs within the *Flavivirus* genus are grouped together, separated from arthropod borne *Flavivirus* (FV). NKVs are Flaviviruses for which there is no known vector [54].

Recent studies also demonstrated a robust interference in mosquito cell lines previously infected with two ISVs (from the *Flaviviridae* and/or *Peribunyavridiae* families) against arboviral infection from the same or from closely related viruses [57]. In addition, ISVs could be used to design chimeric, non-replicating viruses to be used as vaccines for arboviruses. The Eilat virus (genus: *Alphavirus*) is an ISV that has been extensively used for the development of vaccines against CHIKV and Venezuelan equine encephalitis virus and for the production of high quality antigens for enzyme-linked immunosorbent assays [42, 58].

1.6 Antiviral response in *Aedes* spp. mosquitoes

One of the most important determinants of vector competence is the immune response of the mosquito vector upon infection with an arbovirus. The first defense barrier for mosquitoes is their mesenteron or midgut, where viral particles are deposited during a blood meal. The mosquito gut is capable of mounting a variety of defenses against invading pathogens, but cells in the midgut can be infected after threshold number of virions, which is specific for both mosquito and virus species or population/strain, is reached [24, 27, 59]. Viral entry is mediated by direct fusion between the plasma membrane of mosquito cells and the virus envelope or through clathrin-independent endocytosis [60]. Viral particles spread to other mesenteric cells and to the hemolymph through the midgut basal lamina. Through the hemolymph the virus can spread to other tissues, including the salivary glands [24]. There is a lack of molecular and biochemical knowledge about the exact mechanisms and barriers against viruses in the midgut and salivary glands [61]. Mosquitoes possess efficient antiviral strategies. The most relevant mechanisms are the Toll and Immune deficiency (Imd) pathways, the Janus Kinase/Signal transduction and activators of transcription (JAK/STAT) and the RNA interference (RNAi) pathway.

The JAK/STAT pathway core mechanism is conserved amongst animals, including mammals, where it elicits interferon production [62]. The pathway involves a signaling cascade initiated by the binding of a ligand to the Domeless (Dome) receptor, which causes the Hop kinase to self-phosphorylate and consequently phosphorylate tyrosine residues on the receptor. The phosphorylated Dome receptor possesses binding site for STATs, which are recruited and phosphorylated leading to STATs dimers that will translocate into the nucleus and promote transcription of immune system effector genes [63]. The JAK/STAT pathway has been demonstrated to be effective against DENV and partially against ZIKV and *Plasmodium* parasites, while is not effective in defending the mosquito from CHIKV and bacteria [64–66].

The Toll pathway is activated by the binding between microbial pathogen-associated molecular patterns (PAMPs) and mosquito pattern recognition receptors (PRRs). This event leads to the activation of a proteolytic cascade resulting in the cleavage of the cytokine Späetzle [67]. Späetzle activates the Toll transmembrane receptor (TLR)

leading to the phosphorylation and degradation of Cactus, a negative regulator of the transcription factor Re11, that promotes the transcription of antimicrobial peptides [64, 67]. An antiviral activity of the Toll pathway has been demonstrated against DENV in *Ae. aegypti* and the pathway is thought to be stimulated by the microbiota activity in the mosquito [67].

The Immune deficiency (Imd) pathway is initiated by PRR-mediated recognition of PAMPs, which causes a cascade involving various caspase-like proteins and kinases and leads to a functional split into two downstream branches [68] . Each branch activates a specific transcription factor triggering the expression of antimicrobial effectors. In Anopheline mosquitoes the pathway guides immune response against *Plasmodium* protozoa [69]. The exact contribution of the Imd pathway to the mosquito antiviral immunity is still not clear. An enhanced transcription of Imd components and effectors following DENV infection has been verified [70] and a study in *Ae. aegypti* observed that the activation of the Imd pathway has an indirect effect on Sindbis virus levels and it is induced by the mosquito microbiota [71].

1.7 RNA interference pathways

RNA interference (RNAi) is the major antiviral defense system in mosquitoes [72]. RNAi takes advantage of small (<30nt) molecules of non-coding RNA to silence genes or undesired DNA sequences. The RNAi family includes three systems: the microRNA (miRNA), small-interfering RNA (siRNA) and piwi-interacting RNA (piRNA) pathways. These pathways differ for the biogenesis and length of small RNAs and for their molecular targets. The most common proteins taking part in these pathways are the Dicer (DCR) and Argonaute (AGO) proteins. Phylogenetic analyses separated eukaryotic Argonautes into four families, two of which are present in insects: The Ago-like family and the PIWI family. Both these protein families are highly diversified across lineages [73]. Proteins of the AGO subfamily bind miRNAs and siRNAs, whereas PIWI proteins bind exclusively piRNAs [74]. Nevertheless, they share a common mechanism involving the formation of RNA-induced silencing complexes (RISCs) formed by small-RNA guides and nuclease proteins [75]. RISC complexes can transcriptionally and post-transcriptionally silence genes by regulating DNA methylation and histone modification and by cleaving mRNA in the cytoplasm, respectively (**Figure 5**).

RNAi related genes appear to be evolving quickly in mosquitoes. In *Ae. aegypti*, both exo-siRNA and miRNA pathway genes have undergone rapid, positive, diversifying selection [76]. Interestingly, the small regulatory RNA pathway of the arboviral vectors *Aedes* and *Culex* are evolving faster than those of the malaria vector, *Anopheles gambiae* [76, 77].

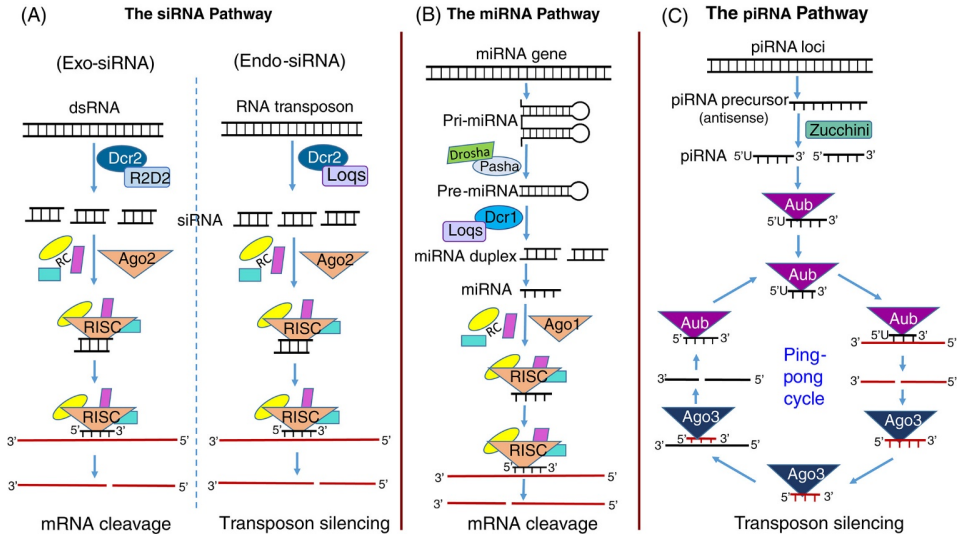


Figure 5. Core enzymes of the three RNA interference (RNAi) pathways in insects [75].

The miRNA pathway is based on endogenous small sRNAs, approximately 22 nt long, called microRNAs (miRNAs) that play a role in regulating gene expression. Usually miRNA precursors (pri-miRNAs) are expressed in the nucleus from specific intergenic and intronic regions. pri-miRNAs are processed by Drosha and Pasha proteins into a pre-miRNA, which is then cleaved by DCR1. The pre-miRNA forms a complex with AGO1 that is capable of actively regulating cellular gene expression by inhibiting translation, cleaving mRNA, and reorganizing chromatin [78]. In *Aedes* species several differentially expressed miRNAs have been identified. Expression of these mRNAs varies upon blood feeding, exposure to arbovirus infection, and at different developmental stages [79].

The siRNA pathway utilizes small dsRNA molecules (21-25 nt in length). Depending on origin of the dsRNA molecule the pathway is distinct into exo-siRNA and endo-siRNA, which are derived from exogenous or degenerated cellular dsRNAs or from endogenous repetitive sequences, sense–antisense pairs and long stem-loop structures, respectively [78]. The loquacious-DCR2 protein complex is required to cut the precursor RNA sequence and produce small molecules. The mature dsRNA is then bound to AGO2 to form the RISC complex. In insects, the siRNA pathway is considered the most effective antiviral system [80]. As most NRVs require a dsRNA step for replication, the exo-siRNA pathway is highly effective against these viruses,

including arboviruses. The production of siRNA derived from DENV and other flaviviruses has been observed in *Aedes* mosquitoes and correlated with resistance to infection [81].

1.7.1 Focus on the piRNA pathway

The P-element Induced Wimpy Testis (PIWI) RNA pathway is ancient and highly conserved amongst metazoans, including arthropods and mammals [82, 83]. The canonic function of the piRNA pathway is to silence transposable elements (TEs) in the germline at transcriptional and post-transcriptional level. A dysfunctional piRNA pathway leads to germ cells death and sterility, a consequence of uncontrolled TE expression and genomic insertion [84]. The piRNA-induced silencing complex (often referred to as piRISC) is formed by a PIWI subfamily protein and a small RNA (24-31 nt), called PIWI-interacting RNA or piRNA, that guides the piRISC towards the target. In *Drosophila melanogaster* three distinct piwi genes are present: *Ago3*, *aubergine*, and *piwi*. One non-functional piwi gene is enough to cause uncontrolled TE movements, resulting in sterility [84].

piRNA precursors are transcribed from genomic loci termed piRNA clusters that are frequently found in heterochromatic regions. In *D. melanogaster* piRNA clusters occupy up to 3.5% of the genome, can be up to 200kb long and produce approximately 80% of all the piRNAs [85]. These loci contain remnants of TE sequences derived from past invasions. A single TE insertion in an active piRNA cluster seems to be enough to repress related TEs located elsewhere in the genome [86, 87]. piRNA clusters are considered genetic traps that integrate TEs entering the genome by horizontal gene transfer. TEs in piRNA clusters subsequently produce piRNAs resulting in silencing of the corresponding TE. This theory is supported by the observation that upon artificial introduction of reporter sequences into piRNA clusters, functional piRNAs targeting these sequences are produced [88]. Indeed, the genomic content of piRNA clusters is variable across different populations of the same species and reflects different exposure to TEs. This variability has been observed in *Flamenco*, a 180-kb heterochromatic piRNA cluster in *D. melanogaster* [87].

Two different pathways lead to the biogenesis of two different piRNA molecules: primary piRNA are directly processed from piRNA clusters, while secondary piRNAs are generated through a mechanism termed ping-pong amplification loop in the cytoplasm [89]. In *D. melanogaster*, long piRNA precursors are transcribed from piRNA clusters and exported to the cytoplasm by several specific proteins [90]. The precursors are cleaved at their 5' end by the endonuclease activity of Piwi or Aub, or by the endonuclease Zucchini [90]. In both cases, slicer proteins are guided by

initiator piRNAs. The 3' ends of these pre-piRNAs are then trimmed to the mature piRNA size by the 3'-5' exonuclease Nibbler and 2'-O-methylated by Hen1 to produce 24–31nt mature primary piRNAs with a 5'-end uridine bias (1U). Intermediates generated through this process are similarly processed to produce trailing or phased piRNAs [91]. Secondary piRNAs are produced in the ping-pong cycle. Aub proteins are loaded with primary piRNAs, antisense to TEs. These Aub-piRISCs complexes cleave transposon transcripts and produce the precursors for sense strand piRNAs bound to Ago3. Ago3 bound to sense strand piRNAs then cleaves antisense piRNA cluster transcripts, producing the precursors of piRNAs that are loaded into Aub, continuing the cycle. Aub-bound primary piRNAs are antisense and possess a 1U bias while sense secondary piRNAs loaded onto Ago3 harbor 10 nt complementarity with Aub-bound piRNAs and a bias towards adenine at the tenth nucleotide (10A) [91]. The peculiar sense and antisense piRNAs 1U/10A pairing and consequent 10nt 5′ overlap is often referred to as the ping-pong signature [89, 91].

Apart from an evident presence in germline cells, PIWI proteins in *D. melanogaster* are expressed also in stem cells and somatic cells [92]. Somatic piRNAs have been identified in different metazoans, often with lineage-specific functions, that include body regeneration in ascidians and polychaetae worms, sex determination in the silkworm and behavior in honey bees [93, 94]. In addition, a growing number of studies have shown that piRNAs and PIWI proteins are abnormally expressed in tumors, but the relationship between cancer and the piRNA pathway is still unclear [95].

An antiviral function for the piRNA pathway was first hypothesized following the observation of virus-derived piRNAs (vpiRNAs) in *D. melanogaster* ovarian cell line persistently infected with multiple ISVs [96]. Nevertheless recent studies in *D. melanogaster* did not yield solid evidence of antiviral effects of piRNAs and no vpiRNAs were observed following viral infection [97]. On the contrary, there is solid proof of a piRNA-based antiviral activity in *Aedes* spp. mosquitoes (**Figure 6**). vpiRNAs derived from arboviruses, including members of the *Flaviviridae, Togaviridae* and *Bunyaviridae* families, have been found both in germ and somatic cells of *Ae. aegypti* and *Ae. albopictus* adult mosquitoes [98]. Mosquito vipiRNAs show a 1U and 10A bias, the hallmark of ping-pong loop amplification, and their abundance is increased upon infection with arboviruses [98, 99].

Comparative genomics studies revealed an expansion of piwi proteins in *Aedes* species (**Figure 6**). In addition to an Ago3 homolog, *Ae. aegypti* and *Ae. albopictus* genomes encode seven and six PIWI subclade proteins, respectively [100, 101]. *Aedes aegypti* PIWI proteins are expressed in germline and somatic cells and several of them are associated with vpiRNAs and viral suppression. Combined knockdown of Ago3 and Piwi1-7 in *Ae. aegypti* cells leads to increased replication of the *Togaviridae* Semliki Forest virus. Ago3, Piwi 5 and Piwi 6 knockdown correlates

with increased replication of *Togaviridae* and a reduction of DENV-specific vpiRNAs. Interestingly, knockdown of Piwi4 results in increased *Flaviviridae* and *Togaviridae* replication but not in a reduction of vpiRNA production, suggesting that Piwi4 has an antiviral function independent from piRNAs [101]. In *Ae. albopictus* the expression of piwi genes is elicited upon infection with DENV and CHIKV.

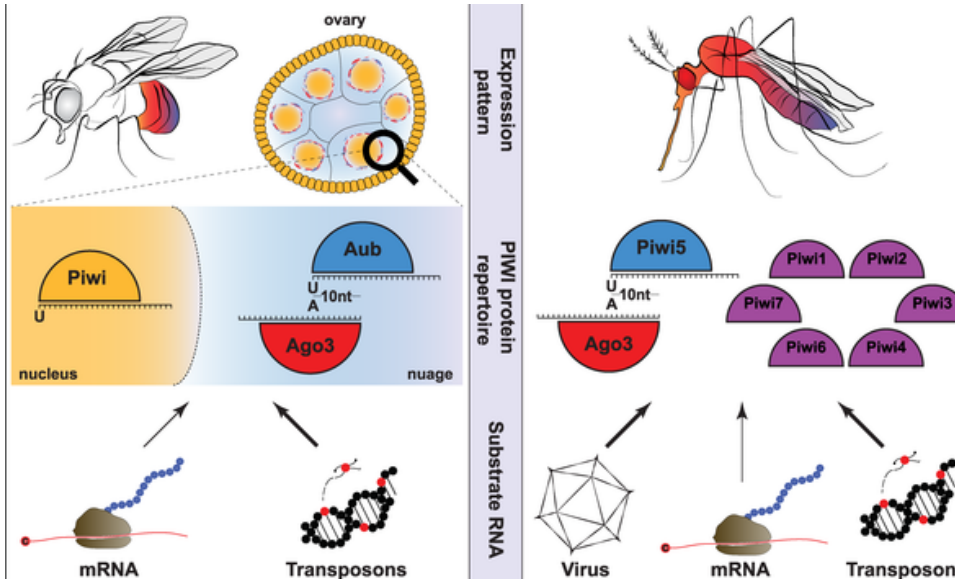


Figure 6. Differences between the composition and activity of the piRNA pathway in *Drosophila* and *Aedes* spp. [98]

For maintenance of infection, DENV seems to rely on Piwi5, while CHIKV relies on Piwi4 and Piwi5 [100]. Overall, these results support the hypothesis that in *Aedes* spp. mosquitoes the evolution of the piRNA pathway might have been driven by the arm-race against fast-evolving RNA viruses. Despite these findings, *Aedes* piRNA cluster biogenesis and expression, as well as the antiviral function of each component of the piRNA pathway are not clear yet. Recent studies reported an antiviral function for PIWI proteins in mammals, though the involvement of piRNAs is not clear [102]. Human Piwi homologs and Ago3-bound tRNAs were found to inhibit Human Immunodeficiency virus 1 in cell cultures [102].

1.8 Endogenous Viral Elements

Eukaryotic genomes contain sequences derived from viruses, collectively called endogenous viral elements (EVEs). Genome integration of viruses is an important factor in the context of the co-evolution between viruses and their hosts and permits to study ancient viral species. Viral sequences can integrate into both somatic and germline cells. Germline integrations are inherited by the progeny and, if not deleterious to the host, they are maintained and even become fixed in a population [56]. Examples of EVEs that have shaped the innate and adaptive immune response are known for many animals [103]. EVEs contribute to genetic novelty in the hosts and EVE-derived proteins are involved in development in mammals, synaptic communication in arthropods and can provide transgenerational antiviral immunity in some species [104]. Additionally, viral integrations can favor the persistence of infection with cognate viruses [105]. The most abundant EVEs are endogenous retroviruses, as Retroviruses are obliged to produce dsDNA intermediates that integrate into host genomic DNA. Unlike retroviruses, NRVs do note code for reverse transcriptase and integrase and their genomes remain exclusively in episomal form during replication. Unexpectedly, it was recently shown that, fragments of viral RNA genomes are converted into viral DNA (vDNA) through the reverse transcriptase activity of endogenous retrotransposons in *Aedes* species [106]. These vDNAs increase the RNAi-mediated antiviral response and help in establishing persistent viral infection [106, 107]. When vDNA production is inhibited, a reduction in the levels of virus-specific small RNA can be observed (including vpiRNAs) resulting in the mosquito death from a loss of tolerance towards viral infection [106].

Considering the nature of nonretroviral RNA viruses, it was unexpected to find sequences from single-stranded and double-stranded NRVs stably integrated into many metazoan genomes, including humans. These integrations are referred to as nonretroviral RNA EVEs (nrEVEs) or nonretroviral integrated RNA virus sequences (NIRVS). The first nrEVEs were identified in *Ae. aegypti* and *Ae. albopictus* in 2004. These sequences were amplified using degenerate primers for flaviviruses and bared similarity to the ISV CFAV and Kamiti River virus (KRV) [108].

The increase accessibility to Next Generation Sequencing (NGS) techniques led to an increased bioinformatical identification of nrEVEs in insect and other animal genomes [109]. Some nrEVEs have been shown to directly express proteins affecting the replication of cognate viruses. As an example, a protein encoded by a Bornavirus-derived nrEVEs in the genome of the thirteen-lined ground squirrel efficiently blocks infection and replication of an extant Bornavirus [110]. nrEVEs can be found in most insect genomes, suggesting similar mechanisms of biogenesis and integration across the lineage. nrEVEs in mosquito genomes, with a focus on *Aedes* species, were characterized by us and other groups [111–114]. Currently known insect nrEVEs appear to be derived from viruses belonging to the *Flavivirus*, *Benyvirus*, and

Quaranjavirus genera; the *Reoviridae*, *Rhabdoviridae* and *Peribunyaviridae* families and the recently characterized family *Chuviridae*. Several nrEVEs are similar to unclassified viruses, that are often grouped as phlebovirus-like and virga/nege-like viruses [109]. Among all the insect species screened for viral integrations, *Ae. aegypti* and *Ae. albopictus* are particularly notable for their high numbers of nrEVEs and the variety of viral origin of nrEVEs. In these species, nrEVEs related to flaviviruses, rhabdoviruses, reoviruses, bunyaviruses, phleboviruses, and quaranjaviruses and to the *Chuviridae* and *Virgaviridae* families have been identified [112–114]. The number of annotated nrEVEs in *Ae. aegypti* and *Ae. albopictus* is different in each study, as there are differences in the bioinformatic methods, and the reference genomes used. Our comparative genomics analyses identified 122 nrEVEs in *Ae. aegypti* and 72 in *Ae. albopictus*, but the limited quality and very high fragmentation of the *Ae. albopictus* genome [115] may have led to an underestimation of the number of nrEVEs in this species. In any case, *Aedes* species harbor 10-fold more nrEVEs than any other mosquito species analyzed, including Anophelines. This difference in nrEVEs number is increased to 30-fold if considering flavivirus-derived nrEVEs (F-nrEVEs). Flaviviruses include both medically relevant arboviruses such as ZIKV, DENV and YFV as well as ISVs like KRV and CFAV. Considering that *Aedes* spp. are mostly vectors for RNA viruses while *Anopheles* spp. are parasitic vectors, nrEVEs may have a role in shaping vector competence for arboviruses in *Aedes* species [109]. A comparative analysis of the nrEVEs landscape across five *Ae. albopictus* geographic populations worldwide revealed a variability in nrEVEs distribution [116]. From this study, a difference between F-nrEVEs and Rhabdovirus-derived nrEVEs (R-nrEVEs) emerged: R-nrEVEs are more widespread and include more ancient integrations based on accumulation of sequence changes than F-nrEVEs, even though the *Rhabdoviridae* family is thought to be evolutionarily younger than the *Flaviviridae* family [116]. In both *Aedes* species, nrEVEs are frequently located in regions rich in repetitive DNA and are associated with long terminal repeat (LTR) retroelements. In particular, F-nrEVEs and R-nrEVEs are often flanked by or engulfed within LTRs of the Pao/Bel, Ty3/Gypsy and Copia families [112, 113]. nrEVEs are not randomly distributed in the genomes of *Ae. aegypti* and *Ae. albopictus*, but are statistically significantly enriched in piRNA clusters [112]. Additionally, nrEVEs produce primary piRNAs (but not siRNAs) that are most often in antisense orientation with respect to the viral RNA genome from which they are derived. Additionally, a 10A bias of sense piRNAs, particularly in *Ae. albopictus* indicates the existence of R-nrEVEs producing secondary piRNAs through ping-pong amplification [111, 112]. nrEVEs -derived small RNAs are enriched in immunoprecipitations (IP) of Piwi5 and Piwi6, suggesting an actual association with PIWI proteins [112]. A nrEVEs enrichment in piRNA clusters is true for other arthropods as well: analysis of 48 genomes including mosquitos, flies, and moths, showed that nrEVEs are enriched within piRNA clusters in 30 genomes [111].

These observations led to a hypothetical model for nrEVEs production and

mechanism [56, 109]. The integration of nrEVEs into insect genomes depends on the formation of vDNA from the non-retroviral RNA template by the action of endogenous reverse transcriptases during infection [106]. Integration events in insect genomes may occur through non-homologous recombination between the vDNAs of infecting NRVs and LTR retroelements during their reverse transcription [109]. There are additional observations supporting this hypothesis [109]: a) a similar mechanism has been found in mammalian genomes [117]; b) reverse transcription of LTR retroelements occurs in the cytoplasm, where most NRVs complete their life-cycle; c) LTR retrotransposons are active in germline cells, favoring heritable integration events and d) some retroelements harbor viral-like domains, such as viral-like helicases from ISVs [118]. Overall, the functional significance of nrEVEs may be to produce piRNAs targeting cognate viruses, and thus provide heritable, sequence-specific antiviral immune memory in arboviral vectors, in a similar way as CRISPR-Cas immunity in prokaryotes (**Figure 7**) [101, 109].

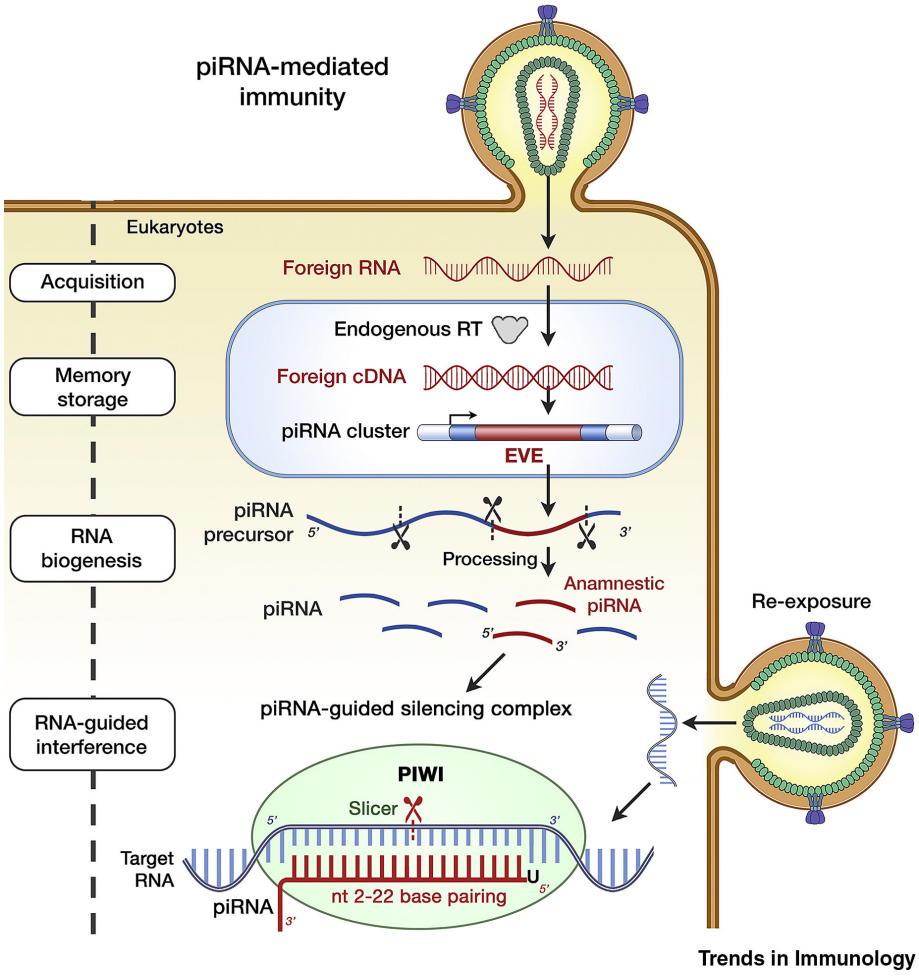


Figure 7. In eukaryotes, RNA virus genetic material may be captured by endogenous RT with a mechanism similar to bacterial CRISPR-Cas9. If the resulting cDNA is integrated as an EVE into a piRNA cluster (acting as a CRISPR locus), piRNAs with memory of the previous virus infection, termed anamnestic piRNAs, might be produced and may target cognate viruses by associating with PIWI proteins [101].

2. Aims of the thesis

Aedes albopictus is globally expanding and has become the prime vector for human arboviruses in Europe. Along with its close relative *Ae. aegypti*, it is responsible for most human infections of mosquito-borne viral diseases, including Dengue, Zika and Chikungunya. *Aedes albopictus* can tolerate temperate climates, placing a much larger portion of the human population at risk for arboviral diseases. Currently, vector control is the primary approach to prevent mosquito-borne diseases, because of the absence of antiviral drugs and limited vaccines available. As traditional methods such as insecticides have become less effective, alternative methods of vector control are being pursued, including approaches that involve genetic manipulation of mosquitoes. A deep knowledge of mosquito genomes, genetic variability, antiviral immunity, and virus co-evolutionary history is essential for the effective development of these newer approaches to vector control. The backbone for any genetic study is a reliable and accurate complete genome sequence.

My long-term objective is to understand the interplay between nrEVEs and the piRNA pathway in the context of the potential impact of viral integrations on mosquito immunity and, consequently, their vector competence. Because during my master thesis I demonstrated that nrEVEs are found in repetitive regions of the genome [112] and available genome assemblies for *Ae. albopictu*s were built using the Illumina short-reads sequencing technology and produced highly fragmented assemblies consisting of >150,000 scaffolds [115, 119], the first aim of my PhD program was to re-sequence the *Ae. albopictus* genome using long-reads strategies and build a less fragmented assembly (Chapters 3.1, 4.1 and 5.1).

The availability of a new assembly allowed me to ask more specific questions on the landscape of viral integrations, including identifying viral sequences integrated into the genomes of wild-collected mosquitoes (Chapters 3.2, 4.2 and 5.2). I also used the new genome assembly for a metagenomic analysis to identify RNA viruses in mosquito small-RNA sequencing data and correlate the mosquito virome with their population-specific nrEVE landscape, under the hypothesis that the pattern of nrEVEs is shaped by exposure to viruses. Finally, I used the genome assembly to identify a viral integration in a piRNA cluster with unique characteristics (i.e. it can be easily and uniquely identified by PCR) to investigate its genetic modification using the CRISPR-CAS9 gene editing technology (Chapters 3.3, 4.3 and 5.3).

3. Materials and Methods

3.1 *De novo* assembly of the *Aedes albopictus* genome

We teamed up with world-wide experts in genome assembly, *in situ* hybridization, smallRNA pathways and mosquito biology. I managed the progression of the project, constantly interacting with all collaborators and travelling to Yale University and to the Virginia Polytechnic Institute. I personally contributed to several bioinformatic analyses relative to the quality control of the genome assembly and the characterization of nrEVEs. For clarity of exposition, I will describe all aspects of the genome project, citing each collaborator when appropriate.

3.1.1 FPA strain creation and samples sequencing

Foshan Pavia A (FPA) mosquitoes (derived from the Foshan strain) were sequenced to produce the new assembly of the *Ae. albopictus* genome. The Foshan strain was received from Dr. Chen of the Southern Medical University of Guangzhou (China) in 2013 and mosquitoes have been reared at the insectary of the University of Pavia since 2013. Insectary conditions are 28°C temperature with 70-80% relative humidity and a 12:12 h (L:D) photoperiod. Larvae are fed on red fish food (Teramin, Tetra Werke, Germany) while adults are provided cotton soaked with 20% sugar. Females are blood-fed using a membrane feeding apparatus (Hemotek, UK) and commercially available mutton blood supplied at a temperature of 37°C. Single matings between a male and a female of the Foshan strain were established, and their progeny was forced again into single matings. We repeated this operation for six consecutive times to generate the FPA strain. The progeny of the sixth single mating was let interbreed and the strain has been maintained ever since. High molecular weight (HMW) DNA was extracted from mosquitoes of the FPA strain in the first 3 generations of its existence and sequenced by the Genomic Sequencing Laboratory of the California Institute for Quantitative Biosciences, University of California, Berkeley. To obtain HMW DNA, fresh frozen pupae from the $2^{nd}/3^{rd}$ generation of the FPA strain (around 80 male sibling pupae) were disassembled using a Pyrex mortar in 2 mL ATL with 4ul RNase. Samples were than incubated at 37°C for 30 minutes with a parafilm cover and gentle agitation (300 rpm). After addition of 100-200 µl proteinase K, samples were incubated overnight at 37°C overnight. DNA was then purified from proteins using a standard phenol-chloroform extraction protocol, followed by precipitation in 100% ice-cold ethanol. DNA was then wash with 70% ethanol at room temperature (max speed spins 10 min).

Purified DNA was resuspended in elution buffer with no EDTA and samples were left in slowly rotation overnight at 4°C to resuspend.

A total of 30 SMRT cells were sequenced using a PacBio RSII instrument (Pacific Bioscience, USA) generating 108 Gb of data for a final 70X coverage. 130 Gb of Hi-C chromosome conformation capture sequencing data were produced by Arima Genomics (San Diego, California) for a final 237X coverage. Raw reads are publicly available on SRA under the BioProject accession ID PRJNA530512.

3.1.2 Flow cytometry

Genome size of *Ae. albopictus* mosquitoes from different strains was estimated by flow cytometry as previously described [120]. Briefly, nuclei were released from the heads of a mosquito and a *Drosophila virilis* standard (1C – 328 Mb) in 1 mL of cold Galbraith buffer using 15 strokes of a pestle in a 2 ml Kontes Dounce Tissue grinder (Kimble, USA). The released nuclei were filtered through 40µm nylon mesh, stained with 25 ml of 1 mg/ml propidium iodide, allowed to stain for three hours in the cold and dark and then scored for relative red (PI) fluorescence using a Cytoflex flow cytometer (Beckman-Coulter, USA). The 1C genome size of the sample was estimated as the ratio of the relative fluorescence of the 2C peaks of the sample and standard multiplied times the 1C amount of DNA in the standard. A minimum of 1000 nuclei were scored under each peak. All scored peaks were symmetric with a CV below 2.0. This experiment was carried out by Spencer Johnston at the Texas A&M University, College Station (USA).

3.1.3 Contig Assembly and Polishing

Initial contigs from the data generated as described above were assembled using Canu v1.7.1 [121] with the following parameters: 'genomeSize=2g' 'correctedErrorRate=0.105' 'corMinCoverage=4' 'corOutCoverage=all' 'corMhapSensitivity=normal' 'gridOptions=--time=72:00:00 --partition=norm' 'stageDirectory=/lscratch/$SLURM_JOBID' 'gridEngineStageOption=--gres=lscratch:100' 'ovlMerThreshold=500'.

Following the first assembly, all contigs were polished for consensus using the Arrow polishing algorithm provided with the PacBio SMRTAnalysis suite, v. 5.1.0.26412 (https://www.pacb.com/products-and-services/analytical-software/smrt-analysis/). Reads were mapped using Minimap2 v2.11 [122] and pbbamify (https://github.com/PacificBiosciences/pbbam). As the large draft assembly size was

a consequence of the retention of secondary alleles introduced from sequencing a pool of individuals, two approaches were used to remove them from the assembly. First, Purge Haplotigs (commit: 6414f68101103af33d47650ea84623a26343bda1) was run using Minimap2 with the "-ax map-pb --secondary=no" options for mapping the reads. Reads were processed by Purge Haplotigs using SAMtools view, sort and index. The purge_haplotigs contigcov command was launched with the settings "-l 5 -m 33 -h 55 -j 200". Following Purge Haplotigs, an additional custom program was run to identify BUSCO genes present on multiple contigs (https://github.com/skingan/HomolContigsByAnnotation). All BUSCO results were generated using BUSCO v3 [123] using the "diptera_odb9" ortholog database. Contigs flagged by either Purge Haplotigs or the BUSCO gene duplication analysis were considered secondary alleles and removed from the primary assembly. The primary assembly contigs were then scaffolded using SALSA v2.0 with the options "-c 10000 -i 20 -e GATC -m yes".

The scaffolded primary contigs were considered the AalbF2 assembly, available on the NCBI repository under the GenBank ID GCA_006496715.1 (RefSeq ID GCF_006496715.1). The alternative alleles were left as unscaffolded contigs but included in the submitted assembly for completeness and uploaded on NCBI under the GenBank ID GCA_006516635.1.

This process was conducted primarily by the group of Adam Philippy from NIH.

3.1.4 Comparative analysis of AalbF2 versus AaloF1

Length metrics for the AalbF2 and AaloF1 assemblies were calculated using Quast v5 [124]. Single-Copy Orthologs genes were identified using BUSCO v3 [123] as described above. Whole Genome Sequencing (WGS) and Total RNA sequencing data from uninfected *Ae. albopictus* Foshan mosquitoes were downloaded from NCBI Sequence Reads Archive (SRA). The data were produced by our team [116] and uploaded under the BioProject ID PRJNA475859. Reads quality and cleanliness was assessed with FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). WGS and RNA-seq reads were separately aligned to AalbF2 and AaloF1. WGS and RNA-seq reads were aligned using MAGIC-Blast [125] and HISAT2 [126] with default parameters, respectively. Alignments were converted to BAM with SAMtools View and sorted with SAMtools Sort [127]. Alignment statistics were calculated with BAMtools Stats [128].

The numbers of rRNA genes in the AalbF2 and AaloF1 genome assemblies was estimated using Barrnap v. 0.9 (https://github.com/tseemann/barrnap) with HMMER v. 3.1 with "eukaryotes" setting.

3.1.5 *In situ* hybridization and physical map construction

A new mapping approach based on amplification of DNA probes using cDNA instead of Bacterial Artificial Chromosome (BAC) clones was used to map scaffolds on chromosomes. DNA probes that derived from the largest genomic scaffolds, 18S rDNA, Prophenoloxidase (PPO) genes, and the nrEVE Canu-Flavi19 were mapped to the chromosomes using Fluorescent In-Situ Hybridization (FISH). Transcripts of *Ae. albopictus* C6/36 cell lines [129] were aligned against AalbF2 to identify genes that could be used for the physical mapping of the *Ae. albopictus* genome. DNA fragments were amplified by PCR using a Q5 high-fidelity DNA polymerase (New England Biolabs, USA). cDNA or genomic DNA fragments were used as templates to amplify transcript fragments or large exons, respectively. Total RNA was obtained from mosquito ovaries with the Zymo Research Direct-Zol DNA/RNA mini prep kit (Zymo Research Corporation, USA). cDNA was synthesized using ~200 ng RNA and primed with oligo (dT) using the Thermofisher Scientific Superscript III first-stand synthesis system (Thermofisher, USA). Laboratory protocols for preparing and performing FISH on *Aedes* spp. have been described earlier [130, 131]. Transcript fragments or large exons with minimal length of 3.8 kb were used as probes for FISH. PCR amplified DNA was labeled with two fluorescence dyes Cy3- or Cy5-dUTP (Enzo Life Sciences, USA) by nick-translation. A pair of DNA probes was hybridized simultaneously to the chromosomes [130, 132]. Slides of mitotic chromosomes were prepared from imaginal discs of 4th instar larvae from the FPA strain following the published protocols [130, 131, 133]. Chromosomes were stained with a YOYO-1 dye (Thermofisher, USA) and slides were mounted using a Prolog Gold reagent (Thermofisher, USA). FISH results were analyzed using a Zeiss LSM 880 Laser Scanning Microscope (Zeiss Microscopy, USA) at 600X magnification. Chromosome idiograms were developed using previously described protocols [131, 133]. Chromosome proportions, such as relative chromosome length and centromeric index (relative length of the p arm) were calculated based on measurements of 60 chromosomes. The statistical analysis was performed using the JPM Pro 15 software program at 95% confidence intervals [134]. One-way ANOVA was used to calculate p-values for comparing chromosome proportions between *Ae. albopictus* and *Ae. aegypti*. *Aedes albopictus* chromosomes were subdivided into 96 bands with 4 different intensities. *In situ* experiments were conducted by the group of Maria Sharakova at Virgina Tech.

3.1.6 Pair-wise comparison between *Aedes aegypti* chromosomes and *Aedes albopictus* scaffolds

The *Ae. aegypti* AaegL5 genome assembly was downloaded from VectorBase (https://www.vectorbase.org/). The first 58 scaffolds of AalbF2 (corresponding to the L75 of the entire assembly) were aligned to *Ae. aegypti* chromosomes using the minimap2 aligner [122]. Only hits with more than 40% of identity between *Ae. albopictus* and *Ae. aegypti* were retained. Alignment results were summarized and visualized as a comparative genome dot plot using D-GENIES [135].

3.1.7 Annotation of *Ae. albopictus* nrEVEs

The AalbF2 genome assembly was screened for integrations from NRVs using an approach based on the Basic Local Alignment Search Tool (BLAST) [136]. To this purpose, a database of viral proteins was created to include all complete amino acids sequences belonging to ssRNA, dsRNA and unclassified RNA viruses with a tropism for arthropods registered in the NCBI RefSeq database as of November 2019 (**Table 1**). The database was expanded including the *Xinmoviridae* and *Phenuiviridae* families. Candidate viral integrations were identified running BLAST against the AalbF2 genome assembly (used as query) against the viral database and with the BLASTx algorithm with an e-value threshold (-evalue option) of $1e^{-6}$ [136]. Resulting BLASTx hits were merged and refined with the EveFinder Pipeline [113]. Putative viral integrations were blasted against all proteins available in the NCBI RefSeq and in the more comprehensive Non-Redundant (NR) database. A custom pipeline was used to recognize and remove false positives, including sequences with a high degree of homology to eukaryotic or bacterial proteins. Additionally, contiguous viral integrations closer than 100 bp to each other and derived from the same viral species were joined. Each viral integration was assigned to a viral family based on its most similar virus in the NR database. The upstream and downstream 1 kb regions of each viral integration were inspected for Repeated Elements using a custom script based on BLASTn and a database of *Ae. albopictus* repeats predicted using RepeatModeler with default settings (http://www.repeatmasker.org/RepeatModeler/). This database was used to run RepeatMasker (http://www.repeatmasker.org) with default parameters to find and classify transposable elements. The correspondence between AaloF1 and AalboF2 viral integrations was evaluated using a BLASTn-based script (Additional file 2: Table S5). Unique viral integrations annotated in AalbF2 were also used to test whether the haplotig purging pipeline effectively separated the alternative haplotypes from the primary assembly. BLASTn was used to find known nrEVEs annotated in the primary assembly in the secondary assembly.

Table 1. Number of viral species included for each family or taxonomical category.

Family	Species	Family	Species
Birnaviridae	1	*Phenuiviridae*	43
Carmotetraviridae	1	*Polycipiviridae*	8
Chuviridae	1	*Qinviridae*	4
Chuviridae-Like	953	*Reoviridae*	54
Cruliviridae	1	*Rhabdoviridae*	87
Dicistroviridae	26	*Roniviridae*	1
Euroniviridae	3	*Solinviviridae*	2
Flaviviridae	87	*Tobaniviridae*	1
Iflaviridae	36	*Togaviridae*	31
Lispiviridae	1	*Tospoviridae*	12
Medioniviridae	1	*Totiviridae*	3
Mesoniviridae	6	*Tymoviridae*	35
Mypoviridae	1	*unassigned Mononegavirales*	2
Nairoviridae	13	*unclassified Picornavirales*	8
Nodaviridae	10	*unclassified Riboviria*	55
Nyamiviridae	8	*Unclassified Viruses*	1
Orthomyxoviridae	2	*Xinmoviridae*	2
Peribunyaviridae	52	*Yueviridae*	2
Permutotetraviridae	1		

Unclassified and Chuviridae-like nrEVEs were excluded from the analysis because these integrations are too similar among themselves to provide reliable results. Hits were retained when at least 98% of the query length was present with a minimum percentage of identity of 95%.

3.1.8 Annotation of piRNA clusters

One-week-old female Foshan mosquitoes were blood-fed with 2mL of rabbit blood at the Institute Pasteur (Paris). A total of 60 fully engorged females were kept at 28°C and fed on a 10% sucrose solution. Two groups of thirty females were collected at 14- and 21-days post blood-meal (PBM), respectively, and dissected to separate ovaries and carcasses, that were pooled in two groups of 15 for each time point. In parallel, 60 sugar-fed, virgin, female mosquitoes were processed as described before. Total RNA was extracted from each pool using the Nucleospin miRNA kit (Macherey Nagel, Germany). Custom small RNA sequencing libraries were prepared and sequenced by the Beijing Genomics Institute (BGI), China, on a BGI-SEQ 500 to obtain 40 million reads SE50 per sample. Sequencing data were deposited to the SRA archive (BioProject PRJNA607026). Small RNA reads were aligned with bowtie (v1.2.2) [137]. Ambiguous (multi-mapping) reads from the small RNA-seq libraries described above were either randomly distributed over all possible mapping positions (--best –strata -M1) or excluded (-m1). For piRNA clusters annotation, reads in the 25-30 bp size range were normalized to one million mapped piRNAs (PPM) to reduce the impact of the lower relative amount of piRNAs in comparison to other sRNA types, and piRNAs were trimmed to their 5' terminal nucleotide. Clusters were annotated with a similar approach to the one used in fruit flies [85], optimizing the settings for *Ae. albopictus* larger and more repetitive genome (**Figure 8**). The parameters that were evaluated were: 1) Influence of the requirement for minimal number of piRNAs per 5 kb window, and the maximum distance between windows to be merged (**Figure 8A**); 2) the requirement for a minimal coverage of uniquely mapping piRNAs, and for the number of single-loci piRNA positions on the fraction of piRNAs incorporated in clusters (**Figure 8B** left panel), or the fraction of the genome covered by clusters (**Figure 8B** right panel).

The genome was scanned with non-overlapping 5 kb windows and clusteres were formed merging windows covered by more than 10 ppm piRNA reads and with a maximum distance of 5 kb. Clusters covered by at least 5 unique PPM piRNAs mapping to at least 5 different positions were then selected. The borders of the clusters were defined by the two furthest piRNAs. Clusters smaller than 1 kb and large clusters with piRNA density < 10 PPM/kb were excluded. Germinal and somatic piRNA clusters were annotated in two separate datasets to avoid averaging out clusters only expressed in one tissue. The final annotation of piRNA clusters was obtained by merging the two datasets, and manually excluding two clusters that were exclusively determined by rRNA reads. piRNA clusters annotation in AalbF2 was solely guided by piRNA coverage of the respective genomic regions and did not include assumptions on nucleotide biases or strand asymmetry, because mosquitoes may encode developmentally relevant piRNAs without 1U bias [138].

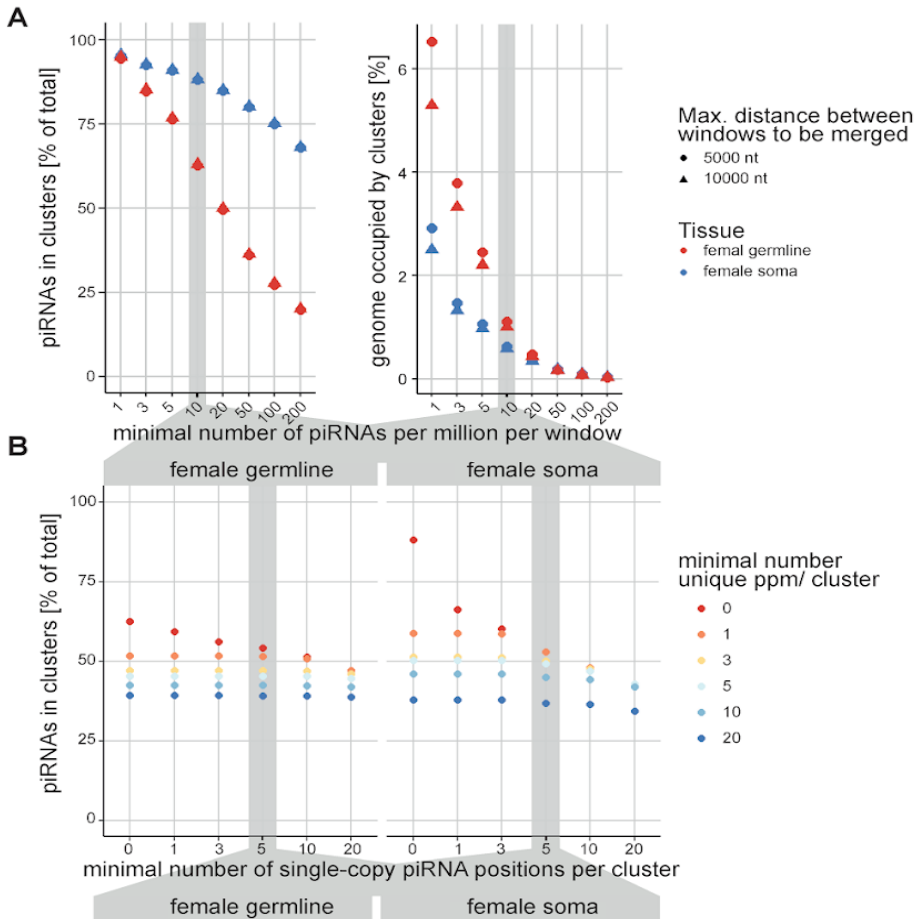


Figure 8. Impact of arbitrary thresholds on cluster annotation.

piRNA expression of clusters was confirmed and quantified using additional small RNA libraries than the ones used for the initial cluster annotation. Expression was normalized to one million mapped piRNAs to compare somatic and germline tissues with different proportions of piRNAs among total small RNAs, or to million mapped small RNAs to plot coverage of the clusters. Enrichment of repeat classes was calculated as the quotient of the genomic fraction of nucleotides annotated with the respective repeat in clusters compared to the whole genome.

3.1.9 miRNA predictions and expression analysis

Small RNA libraries produced from samples collected 14 days PBM were mapped to the AalbF2 assembly using bowtie (v1.2.2) [137] without allowing mismatches (option -n 0). Mapped small RNA reads were filtered by a size of 18-24 bp for a total of 23,644,778 reads which were used as input for the mapping module of miRDeep2 [139], accessed through the Mississippi Galaxy (Galaxy Version 2.0.0.8.) instance available at http://mississippi.fr. miRDeep2 was run with the settings: –k 19 –m –p –r 100. The output was passed to the miRDeep2 module together with a list of known precursor and mature miRNA sequences from the *Ae. aegypti* genome, downloaded from the miRBase database in May 2019 [140]. In addition, precursor miRNAs from *Anopheles gambiae*, *Culex quinquefasciatus*, *D. melanogaster*, *Apis mellifera*, and *Bombyx mori* were added to the list of possible precursor miRNAs. The resulting output file was split into three lists that were processed differently:

1) known and predicted miRNAs based on the *Ae. aegypti* reference datasets [141]. The list of known miRNAs was inspected for miRNA predictions in which the known 3' miRNA was mapped on a 5' arm of a putative hairpin and *vice versa*. The isoforms having very low miRDeep scores compared to the true copy (3p miRNA mapped on 3' arm and/or 5p miRNA mapped on 5' arm) were manually removed.

2) known miRNAs that were unpredicted using the reference dataset. From this list, only miRNAs that were supported by at least 10 mature miRNA counts were considered. Their genomic origin was determined using the NCBI BLASTn algorithm [142] with the pre-miRNA sequences from the *Ae. aegypti* genome and the AalbF2 assembly as query and subject inputs.

3) previously unknown but predicted miRNAs including predictions supported by the reference data from other insect species provided as well as entirely new predictions. The list of novel miRNA predictions was manually curated using stringent parameters described in the literature [99]. At least 80% of the mature miRNAs were required to have the same 5' end on the precursor. Additionally, at least 80% of the predicted miRNA star reads were required to start and end at nucleotide positions predicted to be cut by Drosha/Dicer, allowing a margin of +/- 1 bp at both the 5' and 3' ends. Predictions that were not supported by any predicted miRNA star read were excluded unless the precursor showed high similarity to a known insect miRNA and was supported by > 1000 mapped reads. Precursors with > 1000 BLASTn hits were also excluded.

miRNA expression analysis was performed on Galaxy [143] using the small RNA datasets described above. Small RNAs were mapped to the AalbF2 assembly and their genomic positions were intersected with the location of known and predicted pre-miRNAs obtained from the miRDeep2 analysis using BEDtools intersect

intervals (Galaxy Version 2.29.0; settings: *same* strand, -wo,-abam) [144]. The output was filtered for overlapping small RNA reads and miRNA precursors in the size range of 18-24 bp. Raw counts of each pre-miRNA were normalized to the total number of miRNA reads in each dataset and expressed as reads per million miRNAs (RPMM). Expression data were transformed to $\log_2$(RPMM+1) and plotted in GraphPad Prism. piRNA cluster annotation and miRNA analyses were conducted by the group of Prof. Ronald van Rij from the Radboud University.

3.1.10 Generation of RefSeq geneset annotation

The NCBI Eukaryotic Genome Annotation Pipeline (https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/) was used to annotate genes, transcripts, and proteins on the primary assembly of AalbF2, Aalbo_primary.1 (NCBI accession GCF_006496715.1). The masking of the assembly was done with RepeatMasker using a collection of repeats generated with RepeatModeler [129] and WindowMasker [145] and resulted in 74% of the genome being masked. RNA-Seq reads from 170 *Ae. albopictus* samples were retrieved from SRA and aligned to the masked genome using BLAST [142] followed by Splign [146] and combined with 366 known RefSeq transcripts, 6,046 GenBank transcripts, and 302,415 ESTs from the *Aedes* genus. A set of proteins was aligned to the masked genome to infer the gene models' structures and boundaries. The dataset consisted of 30,044 known RefSeq proteins from *D. melanogaster*; 27,814 predicted RefSeq proteins from *Ae. aegypti*; 100,517 GenBank proteins from Insects; 1084 known RefSeq proteins from *Nasonia vitripennis* and 528 known RefSeq proteins from *A. mellifera*. *Ab initio* extension and joining/filling of partial ORFs of sufficient length and in partial frame but with no supporting alignment was performed by Gnomon (https://www.ncbi.nlm.nih.gov/genome/annotation_euk/gnomon/). Gnomon was run with a hidden Markov model (HMM) trained on *Ae. albopictus* in the putative loci where alignments did not define a complete model, but the coding propensity of the region was sufficiently high to predict a coding gene with confidence. tRNAs were predicted with tRNAscan-SE:1.23 [147] and small non-coding RNAs were predicted by searching the RFAM 12.0 HMMs for eukaryotes using the cmsearch tool from the Infernal package [148]. The annotation of the Aalbo_primary.1 assembly, *Ae. albopictus* Annotation Release 102 or AR102 (https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Aedes_albopictus/102/) resulted in 26,856 protein-coding genes (84% fully supported by experimental evidence, and 12% with more than 5% *ab initio*), 9,530 non-coding genes and 4,108 pseudogenes for a total of 40,494 genes and pseudogenes. The annotation of the new *Ae. albopictus* assembly was done by the group of Francois-Thibaud Niessen at the National Institute of Health (NIH).

3.1.11 Artifacts and gene duplication detection in AalbF2

To identify putative duplications and highly similar sequences in the AalbF2 assembly a BLASTp all_vs_all with an e-value threshold of $1e^{-40}$ was run on the 40,086 peptide sequences annotated in the previously described Aalbo_primary.1 AR 102 genome annotation. Ignoring self-alignments, we extracted the sequences of suspected gene duplications including the 500 bp and 1000 bp in both the 5' and 3' flanking regions of the coding sequence. A BLASTn analysis using as queries the 500 bp-gene regions and 1000 bp-gene dataset against the new assembly was then performed (**Figure 9**). All matches with 100% coverage over the entire sequence and 98% of identity were filtered and collapsed into candidate artifact pairs list.

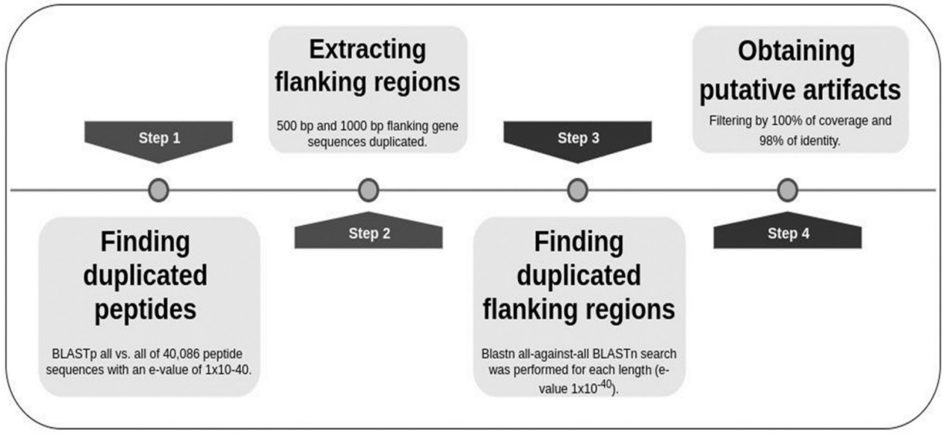


Figure 9. Summary of the bioinformatic pipeline used to detect putative artifacts duplicated in the new assembly of *Ae. albopictus*. After selecting peptides with identical sequences (step 1), the pipeline extracted (step 2) and compared (step 3) the regions flanking the genes expressing the peptides to confirm extensive duplications of the genes and the surrounding regions (step 4).

3.1.12 Identification of immunity genes and manual curation of their annotation

A protein sequence homology analysis pipeline was developed by the group of Philippos Papathanos at the Hebrew University of Jerusalem to identify immune-related genes in AalbF2. A dataset of 417 manually curated protein sequences from

27 immune functions of *Ae. aegypti* [123] was used as query for a BLASTp local alignment against a database of peptides based on the AalbF2 gene annotation (GCF_006496715.1). BLASTp hits were filtered to extract results with an e-value cutoff of 1e-20 and more than 60% of identity. Selected sequences were extracted, and isoform sequences were collapsed. Comparative analyses and manual curation were used to map synteny, phylogeny and sequence identity. The orthologous immune-related genes of *Ae. aegypti* were used as reference to detect gene duplication events. A genome mapping analysis was executed to uncover paralogous generated by tandem duplications. The evolutionary history of each expanded immune family protein was then inferred using the Phylogeny.fr platform [149] with Maximum Likelihood method and the default programs including MUSCLE [150] and Gblocks [151]. Phylogenies were improved after removing divergent and ambiguously aligned blocks from protein sequence alignments [152] and TreeDyn was used to visualize and annotate a robust phylogenetic tree from the phylogenies built as described above [153]. The trees obtained were edited in the iTOL platform (https://itol.embl.de/login.cgi).

Orthogroups, orthologues and single-copy gene clusters of immune-related genes across multiple species were defined by clustering the immune-related peptides of *Ae. albopictus* and the complete peptides of *Ae. aegypti* (Liverpool-AaegL5 assembly), *An. gambiae* (PEST-AgamP4) and *D. melanogaster* (Dmel-r6.26). Two approaches were performed using OrthoVenn2 [154] and OrthoFinder [155]. Parameters for OrthoVenn2 considered e-value cutoff for all-to-all protein similarity comparisons, the Inflation value for the generation of orthologous clusters using the Markov Cluster Algorithm (e-value=$1x10^{-2}$ and Inflation value=1.5).

3.1.13 Analyses of the sex-determining M locus

The annotated *Ae. albopictus nix* transcript (XM_019669557), predicted on the C6/36 cell line genome annotation [129] was used as query to perform a BLASTn against AalbF2 with e-value cutoff of $1e^{-5}$. The *Ae. albopictus* annotated NIX protein sequence (XP_019525102) was used as a query for a tBLASTn run against AalbF2, revealing a possibly duplicated copy in addition to the gene LOC109397226 annotated in both the C6/36 cell line and AalbF2. Male specificity of both LOC109397226 and the duplicated *nix* genome sequences was confirmed by using the Chromosome Quotient analysis [156] with Illumina reads obtained from Foshan strain male and female single mosquitoes [116]. The *Ae. aegypti* myo-sex protein sequence [157] was used in a tBLASTn search with BLAST-EXPLORER to identify the *Ae. albopictus* homologues of myo-sex [158]. Phylogenetic analysis was conducted using Phylogeny.fr [149] with default parameters and programs as described above.

3.1.14 Analyses of genome wide polymorphism and Linkage Disequilibrium

We processed WGS datasets of mosquitoes derived from the Foshan strain (China) and collected in La Reunion Island and Tapachula (Mexico) [100, 116] to detect single nucleotide polymorphisms (SNPs) and derive estimates of linkage disequilibrium (LD) and other population genetics parameters. Paired-end reads were aligned to AalbF2 using BWA-MEM v. 0.7.17 [159]. Unmapped reads and reads with mapping quality (mapQ) below 30 were discarded using SAMtools v. 1.9 [127]. The paired- and single-end pseudo read alignments were merged and sorted with SAMtools. GATK v. 3.8. [160] was used to perform realignments around indels. Second, Picard tools v. 2.9.0 (https://broadinstitute.github.io/picard/) was run to remove optical and PCR duplicates. An uncompressed BCF was generated using SAMtools mpileup v. 1.3.1 with indel calling disabled, skipping bases with baseQ/BAQ less than 30, and with mapQ adjustment (-C) set to 30. A final VCF file was produced using bcftools v. 1.5. and low quality SNPs were removed with SNPcleaner version 2.4.1 [161]. Sites that had a total depth across all individuals less than 1,500 reads or had less than 10 individuals with at least two reads each were purged from the VCF. Finally, additional sites were filtered out based on the default settings within the SNPcleaner script. This pipeline produced a dataset of robust sites for each population comprising the sites that passed all the filtering thresholds. Downstream analyses were restricted to these sites using ANGSD v. 0.929-21 [162] -sites option. Only uniquely mapped reads with minimum map quality and base quality thresholds of 30 and 20, respectively, were used. Linkage disequilibrium (LD) was calculated with ANGSD genotype likelihoods to directly estimate decay using ngsLD version 1.1.0 [163]. Global Weir and Cockerham F_{ST} [164] between populations and diversity (π) within populations were calculated directly from the estimated allele frequencies from the sequencing read data. Approximately 359 million robust sites per population resulted after filtering. F_{ST} and π were estimated through a sliding window analysis across all scaffolds of the new genome with 50,000 bp windows and 10,000 bp steps, with a total of 85,844 windows. Windows with at least 2,000 sites where plotted with each window being a point in the plots. The ngsLD [163] package was used to estimate the pairwise LD avoiding hard call genotypes entirely and using genotype likelihoods (GLs). The program has two algorithms to estimate LD levels from GLs. One is a maximum likelihood approach to estimate the haplotype frequencies between pairs of sites to estimate D, D' and r^2 and the other is based on the squared Pearson correlation (r^2) between expected genotypes using their posterior probabilities. All LD estimates were done with 100 bootstraps and testing different bin sizes until small confidence intervals were found. The LD pairwise comparisons were estimated for all sites and 0.01% of the comparisons were randomly picked to run the ngsLD algorithms for fitting and plotting. The 0.01% sampling data points represents at least 1.5 million r^2

comparisons. New and previously published SNP-chip data from *Ae. aegypti* were combined to estimate LD for this species based on the last reference genome available [141] and compare this LD data to our results. Plots were generated in R using the built-in functions and the packages ggplot2 [165], Sushi [166] and qqman [167].

3.1.15 Developmental profile analyses

RNA was extracted from wild-type *Ae. albopictus* mosquitoes from San Gabriel Valley, Los Angeles County, CA. Mosquito rearing, total RNA isolation and RNA-Seq were carried out as previously described [168]. RNA-seq libraries were aligned to the AalbF2 assembly using STAR aligner [169]. The gene annotation was downloaded from the NCBI repository (GCF_006496715.1_Aalbo_primary.1_genomic.gtf) and expression was quantified with featureCounts [170]. Transcripts Per Million (TPM) and Fragments Per Kilobase Million (FPKM) values were calculated from count data using Perl scripts. All Sequencing data has been made publicly available at NCBI SRA under BioProject PRJNA563095 (genomic) and PRJNA563095 (transcriptomic).

3.2 The landscape of nrEVEs in wild collected *Aedes albopictus* mosquitoes and their virome

3.2.1 Natural populations strains and wild sample sequencing

Four strains established from eggs collected in the wild were used for a preliminary analysis of nrEVE geographic distribution. Laboratory strains were derived from eggs collected in Chiang Mai (CM) from the Lampang Province in the northern part of Thailand; St. Pierre (StP) in the west coast of La Reunion Island; Tapachula (Tap) in the southern Mexican state of Chiapas and Crema (Cr) from the Lombardia region in northern Italy. Wild-collected eggs were hatched in the University of Pavia insectary and maintained as previously described ever since. Along with eggs, we also received wild-collected mosquitoes preserved in ethanol from each of the above-mentioned localities. DNA was extracted from 3 pools of 40 adult wild-collected mosquitoes of each of the four populations using the DNeasy Blood and Tissue Kit (Qiagen, Germany) following manufacturer's protocol. Genomic DNA was sent to the "Polo D'Innovazione Genomica, Genetica e Biologica" (Italy) for quality control, DNA-seq library preparation and sequencing. The TruSeq DNA PCR-Free library (Illumina, USA) was used to produce paired-end reads. Samples were sequenced on an Illumina HiSeq 2500 instrument ensuring a minimum coverage of 20X.

Adult *Ae. albopictus* individuals and eggs were collected from the wild in Mexico, La Reunion island and southern China. In China, also pupae and larvae from breeding sites were collected. Samples from Mexico were collected in 2016 from the Chiapas region through an agreement with Dr. Mauricio Casa Martinez.

Samples from La Reunion island were collected in April 2017 from different locations on the eastern and westerns sides of La Reunion island based on the epidemiology of arboviral infections in the island [171]. From the individuals collected in La Reunion island, pools of 10 female mosquitoes each were stored at 4°C in DNA/RNA-shield (Zymo Research, USA) and transported back to Italy after having established a Material Transfer Agreement with CIRAD (La Reunion). DNA was extracted from mosquito pools using the DNeasy Blood and Tissue Kit (Qiagen, Germany), according to manufacturer's protocol and stored at -20 °C. RNA was extracted using a custom protocol based on Trizol (Ambion/Invitrogen, USA), treated with DNAse I (Sigma-Aldrich, USA) and stored at -80 °C. DNA and RNA from La Reunion mosquito pools were sequenced by Biodiversa srl (Italy). Whole genome and Small RNA libraries were prepared using the Nextera DNA Library Preparation Kit for paired-end reads (Illumina, USA) and NEBNext Small RNA Library Prep Set (New England Biolabs, USA), respectively. DNA and small-RNA Libraries were sequenced on an Illumina HiSeq 2500 instrument. A 20X minimum coverage was ensured for DNA sequencing. More than 20 Million single-pair 75bp reads were produced from each sample. Additionally, 24 single mosquitoes from Mexico and 23 from La Reunion Island were sequenced individually at Verily, USA, with an Illumina HiSeq 4000 sequencer producing more than 20X coverage for each sample.

I personally collected *Ae. albopictus* individuals in Guangzhou, Guangdong province (China) in November and December 2017. Adult mosquitoes were retrieved using an aspirator while larvae and pupae were collected from breeding sites found in parks and residential areas. Adults, larvae, and pupae were divided in pools of 10 individuals and stored at 4° in DNA/RNA-shield (Zymo Research, USA). Pools were processed at Prof. Xiao Guang Chen laboratory at Southern Medical University, Guangzhou. DNA was extracted using the universal genomic DNA extraction kit (Takara, Japan) with the supplier protocol and stored at -20 °C. RNA was extracted using a custom protocol based on Trizol (Ambion/Invitrogen, USA), treated with DNAse I (Takara, Japan) and stored at -80 °C. DNA and RNA pools of 30 individuals were sequenced in Wuhan by Beijing Genomics Institute (BGI), China. DNA WGS libraries were prepared with the TruSeq Small RNA Library Preparation Kit (Illumina, USA) and were sequenced on an Illumina HiSeq X instrument. Each library was sequenced with a 60X minimum coverage. Small RNA libraries were prepared with the NEBNext Small RNA Library Prep Set and sequenced an Illumina HiSeq 4500 sequencer. A minimum of 20 million small RNAs sequences were obtained from each pool.

3.2.2 Identification of novel viral integrations

WGS data of adult female mosquitoes were analyzed for the presence of additional viral integrations different from nrEVEs characterized in the AalbF2 assembly with Vy-PER v. 0.3 [172], followed by ViR [173]. Pools of mosquitoes were available from Thailand, Mexico, La Reunion, and Italy as described above, in addition to single mosquitoes from La Reunion and Mexico. Vy-PER was originally developed to accurately detect integrations from viruses like hepatitis B and HIV in human genomes and genome integrations in cancer patients [172]. Briefly, the pipeline aligned WGS data on the AalbF2 genome assembly using BWA v. 0.7.15 aln followed by BWA sampe [159] and extracted unmapped reads with SAMtools [127]. Low-complexity reads were discarded using Phobos v. 3.3.12 [174] and a custom script. The remaining unmapped reads were screened for viral sequences with BLAT v. 36x1 [175], using a custom viral genomes database including arboviruses and ISVs as query. Reads exhibiting similarity to one or more viruses and representing putative integrations in the genome are given a position based on the alignment of their mapped pairs. The results were then processed by ViR [173], a custom pipeline composed of four scripts created in our laboratory. The first script of the pipeline, ViR, RefineCandidates.sh script, is used to identify the best candidate viral integrations from the list of chimeric reads identified running Vy-PER through subsequent filtering steps. After this step, the second script of the ViR pipeline, SolveDispersion.sh, is run to group chimeric reads having the viral read of the pair that corresponds to the same viral species and the host read of the pair mapping to regions of the mosquito genome with the same sequences (i.e. equivalent genomic regions). Equivalent genomic regions result from the presence of the same repeated sequences across the genome (due to repetitions or mis annotations) and, in case of pooled samples, variability within individuals. Reads supporting the same viral integrations were realigned with BLASTn v. 2.9.0+ [136] to identify all mapping positions and grouped together when mapping in the same genomic region with the third script of the pipeline (AligToGroup.sh). When the viral portion of the identified integration is different than already annotated nrEVEs, the fourth script of the ViR pipeline (LTFinder.sh) is used with the initial raw WGS data to produce *de-novo* assemblies starting from the putative viral integration to extend its sequence. Assemblies are elongated using multiple iterations of Trinity v. 2.7.0 [176], adapted to be used on DNA rather than RNA sequencing data.

Bioinformatic predictions of each novel viral integration were molecularly tested by PCR using the DreamTaq Green PCR Master Mix 2X (ThermoFisher, USA) with specific primers (**Table 2**). PCR products were directly purified or cloned into *E. Coli* TOP10 competent cells using the TOPO TA Cloning Kit for Sequencing (Thermo Fisher, USA). Plasmids containing the PCR product insertions were extracted with the QIAprep Spin Miniprep Kit (Qiagen, Germany). Plasmids and purified PCR products were Sanger sequenced by Macrogen Europe (The Netherlands).

Table 2. Primers designed for the amplification of the novel nrEVEs. F=forward, R=reverse

nrEVE	Primer Sequence	direction	*Location*
nrEVEnew-1	CGACAGCCTGTTCGAATGC	F	*Mosquito genome*
	CATCAGCCTTTCCGTAGTTCC	R	viral integration
nrEVEnew-2	CCGCGCTCTCACTCAGTA	F	*Mosquito genome*
	CCATCAGCACAAGATCATCAGT	R	viral integration
nrEVEnew-3	AAGTTCTCGCGACTAACCCA	R	viral integration
	GCCATCCAACTTGAACCGAT	F	*Mosquito genome*
nrEVEnew-4	CCGCGTTGGTCCCTTCTG	F	viral integration
	GTGAGTGCCCTATACGTTAGCA	R	*Mosquito genome*
nrEVEnew-5	GACTCGCTTACCAGAACATCG	R	viral integration
	TGAGTTTGGGCAAGCATGAG	F	*Mosquito genome*
nrEVEnew-6	CTGGAGAAGCTAAGGGGTCG	F	viral integration
	GGTGGCTTGAGTTTGGGCAA	R	*Mosquito genome*
nrEVEnew-7	CGTTTGTACACCCGCTTTTGT	F	*Mosquito genome*
	TCCAAGTGAATGAGTCCGCG	R	viral integration

3.2.3 Virome analysis

Raw small-RNA sequencing data was retrieved from Biodiversa Srl and BGI following the sequencing process described above. Reads were quality checked with FastQC v. 0.11.9 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and the presence of adapters was confirmed with DNApy v. 1.1 [177]. Adapters and bases with baseQ inferior to 20 were trimmed from the reads using Cutadapt v. 2.9 [178].

Low complexity reads were defined as reads having >80% of the sequence composed by duplets or triplets and were removed with the Duster.pl script from the NGS TOOLBOX release 2 (https://sourceforge.net/p/ngs-toolbox/wiki/Home/). A custom pipeline to filter out reads mapping to *Ae. albopictus* genome and to produce and analyze the unaligned reads was run on each sample. Briefly, all nrEVEs were masked from the *Ae. albopictus* AalbF2 assembly using BEDtools maskfasta v. 2.28 [144]. Small-RNA reads were mapped on the masked assembly with bowtie v. 1.1.2

[137] optimizing the parameters for small-RNA reads (-n 1 -l 12 --best). The reads that did not map to the *Ae. albopictus* genome were extracted with SAMtools view v. 1.4 [127] (with the -f 4 setting) and converted to fastq with SAMtools fastq.

RNA sequences were assembled from small-RNA reads using the Oases pipeline [179] with Velvet v.1.2.10 (https://www.ebi.ac.uk/~zerbino/velvet/) testing k-mer lengths from 13 to 31. The transcripts assembled by Oases were merged and filtered to remove nrEVEs-derived transcripts using BLASTn v. 2.6.0 [136] to compare the transcripts against the nrEVEs annotated in AalbF2. Transcripts with a percentage of identity higher than 90% and covering at least 50% of an existing nrEVE were removed. Redundant and duplicated transcripts were clustered using CD-HIT v. 4.8.0 [180] with the following options: -c 0.9 -n 8 -d 60 -g 1. The clustered transcripts were screened against the entire NCBI nr peptides database (https://www.ncbi.nlm.nih.gov/refseq/about/nonredundantproteins/, downloaded in march 2020) using DIAMOND v. 0.9.31 [181] blastx algorithm with default setting. DIAMOND hits were loaded on MEGAN6 community edition v. 6.19.9 [182]. MEGAN6 was run with the naive LCA algorithm requesting a 70% minimum percent to cover. Minimum score and maximum e-values were set on 1.0 and 0.01, respectively. Transcripts assigned to viruses were extracted and screened again against the nr database using BLASTn [136] with an e-value threshold of 1e-5. For each transcript, the hit with the highest score was selected and was used to assign a viral species and viral protein to the transcript.

3.3 Genetic modification of the *Ae. albopictus* genome

3.3.1 Mosquito rearing and egg collection

Eggs of the previously described FPA strain were shipped to the laboratory of Prof. Rasgon at Pennsylvania State University in State College, PA (USA) where a colony was established. Mosquitoes were reared at 27°C and 70% humidity, with a 12h/12h light and dark cycle. Larvae and adults were maintained as described for mosquitoes in the University of Pavia insectary. Females were blood-fed on anonymous human blood (Biospecialty, USA) using a water-jacketed membrane feeding system. Mosquitoes were blood-fed for approximately 1-2 hours twice in the 5 days preceding embryo injection experiments. Blood-fed mosquitoes were provided with vials for oviposition, consisting of cylindrical plastic tubes containing a disc of chromatography filter paper (Whatman, UK), kept wet by placing it over wet cotton. Six to ten blood-fed females were transferred to each oviposition vial. Each vial was kept in an obscured container for 30-60 minutes and, if eggs were present, the filter

paper was moved to a petri dish and kept wet all the time to protect the eggs from desiccation. After egg collection, female mosquitoes were transferred to a separate cage.

3.3.2 Injection needle fabrication

Quartz needles for embryo injections were prepared using a Sutter P-2000 micropipette puller (Sutter Instruments, USA). Two different programs were used, coded 68 and 69 respectively, with the following settings:

- Program 68: Heat:705, Fil:4, Vel:40, Del:125, Pull:130. Needles pulled with this program were used to inject light-colored, softer, eggs.

- Program 69: Heat:750, Fil:4, Vel:40, Del:150, Pull:165. Needles pulled with this program are more resistant and were used to inject dark-colored and harder, eggs.

After their fabrication, quartz needles were laid on a soft paste and stashed in a petri dish. As *Ae. albopictus* eggs are easier to inject than *Anopheles* or *Culex* spp. eggs, quartz needles did not require beveling.

3.3.3 Injection mix preparation

The injection mix was prepared combining the following reagents:

- 1x Injection solution: 5mM KCl, 0.1mM sodium phosphate, pH 6. 8

- Cas9 protein supplied by Integrated DNA Technologies (USA) at a concentration of 300 ng μL^{-1}.

- Guide RNA (gRNA): Double stranded DNA templates for the gRNAs were synthesized by PNAbio (USA). gRNAs were tested by PCR and *in-vitro* transcribed to RNA with the MegaScript Kit (Invitrogen, USA), purified using the the MegaClear kit (Invitrogen, USA) and concentrated through ethanol precipitation. Both the Cas family proteins and the gRNAs were tested for an effective endonuclease activity towards the target DNA sequence by an In-Vitro cleavage assay.

- Donor Plasmid DNA: The Insert plasmid for injection was prepared by Dr. Macias and cloned in competent *E. coli* cells. Plasmid DNA was isolated using a MaxiPrep plasmid extraction kit (Qiagen, Germany) and amplified

by PCR and sequenced by an in-house sequencing facility before being used for injection. The injection mix was incubated at room temperature for 15 min before injection into *Ae. albopictus* eggs 90–120 min post-oviposition.

3.3.4 Injection mix preparation

Fresh eggs were transferred from oviposition vials to wet filter paper. Light-grey to darkish-grey eggs were manipulated with forceps and paintbrush to align and orient them in same direction, exposing the posterior pole to the needle. Embryos were transferred to a glass slide covered in double-sided adhesive tape by gently pressing the edge of the slide on the filter paper containing them. Eggs were immediately covered with a 1:1 halocarbon oil 700 and halocarbon oil 27 mixture to prevent further desiccation. Using an Eppendorf Microloader tip (Eppendorf, Germany) on a 0.1-10 ul micropipette, 2-4ul of injection mix was inserted in a quartz microinjection needle prepared as previously described. Microinjections were performed under an Olympus SZX16 wide zoom stereo microscope (Olympus, Japan). An Eppendorf Injectman micromanipulator was used to move the needle. Injection was controlled with an Eppendorf Femtojet Express injector adjusting the parameters (time, injection pressure and standard pressure) to inject 0.2-0.5 nl of solution. Following injection of the embryos, the glass slide was washed with deionized water to remove halocarbon oil.

3.3.5 Post-Injection procedures and matings of injected mosquitoes (G0)

The eggs were picked with a fine brush and transferred on wet filter paper kept in insectary conditions as described above. After 4 days, filter papers containing the injected embryos (Generation zero or G0) were laid down on a plastic pan and covered in deionized water. Upon hatching, G0 injected larvae were reared in insectary conditions as described above. G0 pupae were individually collected and moved to a cup until emergence of the adult mosquito. G0 males and females were separated and two different protocols for crossing them with wild-type adults were followed.

Each G0 transgenic adult male was added to a separate cup containing two adult wild-type virgin females of the same age and a few days were allowed for them to mate. After mating, mosquitoes from five cups (for a total of 5 G0 males and 10 wild-type females) were pooled in a cage. Pools of 10 G0 virgin females were transferred to a cage together with 12-15 wild-type males. Both G0 female + WT male and G0 male + WT female pools were permitted to mate and were blood-fed and provided with an oviposition cup. G1 eggs were collected, kept wet for ~2-3 days and then dried for another ~2-3 days and moved to a plastic pan and covered in deionized water for

hatching.

3.3.6 G1 to G3 screening

G1 larvae were screened for the expression of the Cyan Fluorescent Protein (CFP) inserted in the construct using an Olympus Fluoview 1000 fluorescence microscope (Olympus, Japan). As no expression of CFP was detected in the larvae, the following PCR primers were designed to amplify the entire CFP ORF producing a 971 bp amplicon:

3XP3 For 5'-CGCCCGGGGATCTAATTCAA-3'

CFP Rev 5'-CTTGTACAGCTCGTCCATGC-3'

A single leg was removed from female mosquitoes keeping the insect alive. DNA was extracted using the Phire Tissue Direct PCR Master Mix kit (Thermo Fisher, USA). A PCR on this DNA template was performed using the Phire and Phusion Enzymes (Thermo Fisher, USA) at an annealing temperature of 64°C. Positive individuals for the CFP sequence were pooled together, allowed some time to mate and blood-fed. Eggs were collected, and the process was repeated until G3.

4. Results

4.1 *De novo* assembly of the *Aedes albopictus* genome

4.1.1 Genome assembly metrics, quality assessment and annotation

The genome length of *Ae. albopictus* was estimated to be similar to that of *Ae. aegypti*. Total genome size measured with cytofluorimetry was established to be between 1.190-1.275 Gb, across populations from the native home range (Thailand, Malaysia, Singapore), old-colonized regions (La Reunion Island), and recently invaded areas (Italy, USA and Mexico) (**Figure 10**).

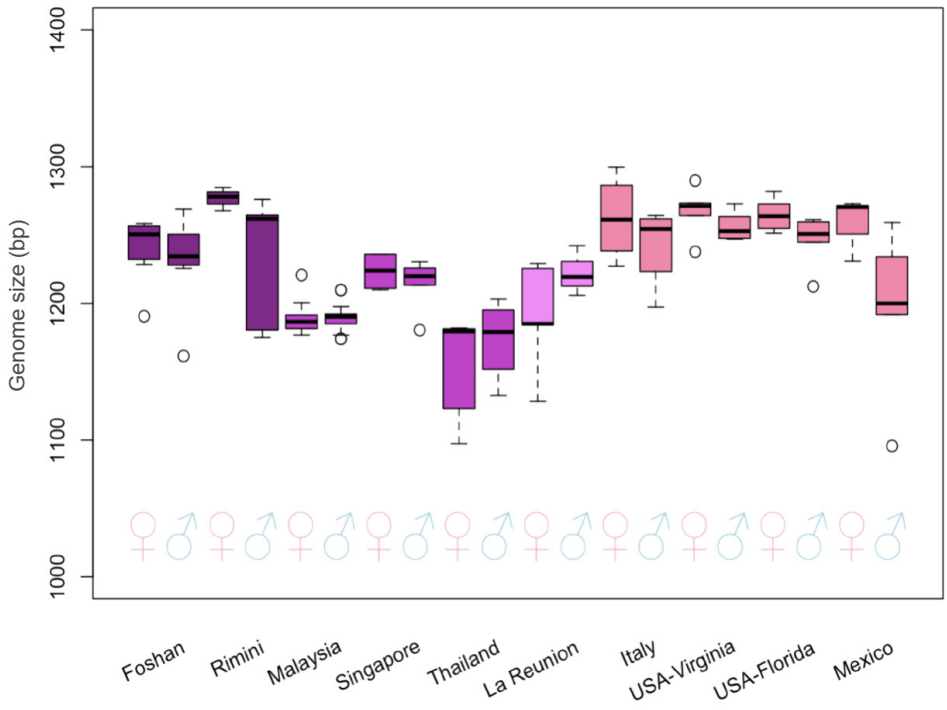


Figure 10. Boxplot graph showing cytofluorimetric-based estimates of the genome size of *Ae. albopictus* strains. Median values are represented by black lines. Whiskers represent the 1st and 3rd quartiles, and white circles represent outliers.

The Foshan strain was used to produce the AaloF1 assembly [115] so we chose this strain to re-sequence the *Ae. albopictus* genome and produce the new AalbF2 assembly [183]. A sequencing strategy based on the combination of long-read sequencing and chromosome conformation capture was used to produce enough data to generate a high-quality assembly with a reduced number of scaffolds. A single mosquito did not yield enough DNA to use this strategy so the genetically homogeneous FPA strain was created as described in material and methods. After six consecutive rounds of single sister-brother matings of individuals from the FPA strain, HMW DNA was extracted from 40 pupae. More than 82 Gb of PacBio SMRT single molecule long reads were produced by the Genomic Sequencing Laboratory of the California Institute for Quantitative Biosciences, University of California, Berkeley. The PacBio reads had a mean length of 10 kb and an N50 length (half of the data comprises sequences of this length or longer) of 18 kb. Additionally, a Hi-C proximity ligation library was prepared from ten adult mosquitoes of the FPA strain and was sequenced on an Illumina instrument yielding 135 Gb of sequencing data. The long-read PacBio data were assembled with Canu [121] and the resulting contigs were polished with Arrow (https://www.pacb.com/products-and-services/analytical-software/smrt-analysis/) using the raw PacBio signal data. This initial assembly totaled 5.17 Gb, far exceeding the expected haploid genome size (~1.25 Gb), suggesting the presence of alleles that failed to collapse in the assembly. This original unfiltered and unsplit assembly had a BUSCO [123] completeness of 97.6% with 81.8% duplication, indicating that the majority of genes were represented in the combined assembly by more than one allele. This could be explained by a high level of heterozygosity in the pool of sequenced mosquitoes, resulting in multiple allelic variants assembled separately, as has been previously noted in long-read assemblies [184]. The large genome size of *Aedes* spp. mosquitoes and the high variability even across individuals of the same strain [116] required the removal of haplotypic duplications prior to the creation of haploid reference scaffolds [141]. To produce a homogeneous, non-redundant genome, this initial assembly was partitioned into primary and alternative contig sets. Contig alignments and depth of coverage were analyzed with Purge Haplotigs [184] along with BUSCO single-copy orthologs [123] to determine which contigs were likely to be redundant and should be designated as alternative alleles. Haplotig purging reduced the size of the primary assembly to 2.54 Gb, which was then scaffolded via the Hi-C chromosome conformation capture data using SALSA2 [185]. We decided to call the final assembly AalbF2 (*Ae. albopictus* Foshan 2) [183]. The primary AalbF2 assembly consists of 2,197 scaffolds with an N50 length of 55.7 Mb (**Table 3**). This represents an increase in scaffold size of two orders of magnitude compared to the previous genome assembly, AaloF1, which had a scaffold N50 of 201 kb [115]. AaloF1 was a highly fragmented genome, being composed of 154,782 scaffolds. A significant improvement of AalbF2 is that more than 50% (L50) of the genome assembly sequence is contained within the 13 largest scaffolds (**Table 3**).

Table 3. Data table listing, comparatively between AalbF2 and AaloF1, assembly, BUSCO and Barrnap statistics; genome and transcriptome alignment rates. [1]L50 is the smallest number of contigs whose length makes up 50% of the genome size. [2]Alignment of 16 singly-sequenced Foshan mosquitoes using MagicBlast. [3]Alignment of published RNA-seq data using Hisat2.

	AaloF1	AalbF2
Assembly Statistics		
Total assembly size	1,923,476,627 bp	2,538,387,871 bp
GC (%)	40.05%	40.40%
N. contigs	355061	5556
N. scaffolds	154782	2197
Scaffold N50	195,500 bp	55,702,539 bp
L50[1]	2578	13
Max scaffold length	1,305 Mb	196,395 Mb
BUSCO Statistics		
Complete BUSCOs (C)	2620 (93.6%)	2610 (93.2%)
Complete and single-copy BUSCOs (S)	1984 (70.9 %)	2218 (79.2%)
Complete and duplicated BUSCOs (D)	636 (22.7%)	392 (14.0%)
Fragmented BUSCOs (F)	94 (3.4%)	70 (2.5%)
Missing BUSCOs (M)	85 (3.0%)	119 (4.3%)
Total BUSCO groups searched	2799 (100%)	2799 (100%)
Barrnap Statistics		
Predicted rRNA genes	22	484
Genome alignment Statistics[3]		
Alignment rate	82.76%	84.85%
Properly paired reads	63.75%	67.15%
Both Reads mapped	65.00%	72.53%
Singletons	17.76%	12.32%
Transcriptome Data Statistics[4]		
Alignment rate	82.04%	86.54%
Properly paired reads	70.92%	78.49%
Both Reads mapped	76.37%	82.02%
Singletons	5.68%	4.52%

This increase in continuity provides a wider view of the genomic organization of *Ae. albopictus* and allows for a more accurate annotation of genes and other features. A comparative analyses of single-copy orthologues via BUSCO [123] in AalbF2 and AaloF1 showed an 8.3% increase in the percentage of complete, single-copy BUSCOs (**Table 3**) in AalbF2. Although being larger, AalbF2 has a BUSCO duplication of 14.0% against the 22.7% BUSCO duplication of AaloF1. Similarly, Barrnap (https://github.com/tseemann/barrnap) estimated 484 ribosomal RNA gene sequences in AalbF2 (compared to 22 in AaloF1), a value close to the number (430) independently estimated from an *Ae. albopictus* haploid genome [186]. The percentage of alignment of DNA and RNA sequencing data from published resources [100, 116, 187] and the percentage of properly paired reads were also analyzed and confirmed the quality and continuity of AalbF2 (Additional file 2: Table S1). The higher continuity of AalbF2 is also shown by the annotation of TEs and other repeated elements (**Table 4, Table 5**). We detected a TE occupancy of 55.03% of the genome size, a value comparable to that (54.85%) of the most recent assembly of the *Ae. aegypti* genome, AaegL5 (**Table 4**). These results support our hypothesis that even if the primary AalbF2 assembly retains duplications, these are mostly from intergenic and repeated regions and are present in the smaller scaffolds and do not affect the usableness of the genome assembly.

Table 4. Comparison between the TE repertoire of the AalbF2, AaloF1 and AaegL5 genome assemblies.

	AalbF2	**AaloF1**	**AaegL5**
DNA	15.05%	8.52%	15.06%
LINE	15.16%	34.67%	16.09%
SINE	1.92%	0.07%	1.16%
LTR	5.60%	16.21%	11.66%
Other	17.30%	8.85%	10.88%
TOTAL TEs	55.03%	68.33%	54.85%

Table 5. Complete repeated elements annotation in the AalbF2 genome assembly.

	Number of elements	Occupancy (bp)	Percentage of sequence
Retroelements	**1430646**	**575946434**	**22.69%**
SINEs:	208123	48743703	1.92%
Penelope	103794	18576486	0.73%
LINEs:	866027	384927593	15.16%
CRE/SLACS	0	0	0.00%
L2/CR1/Rex	81301	32399143	1.28%
R1/LOA/Jockey	320615	167497304	6.6%
R2/R4/NeSL	8369	2745389	0.11%
RTE/Bov-B	317328	149941249	5.91%
L1/CIN4	9665	4479453	0.18%
LTR elements:	**356496**	**142275138**	**5.6%**
BEL/Pao	93848	39116362	1.54%
Ty1/Copia	81033	32119258	1.27%
Gypsy/DIRS1	180988	69910987	2.75%
Retroviral	0	0	0.00%
DNA transposons	**1485602**	**382032412**	**15.05%**
hobo-Activator	121429	30954936	1.22%
Tc1-IS630-Pogo	113870	30939226	1.22%
En-Spm	0	0	0.00%
MuDR-IS905	0	0	0.00%
PiggyBac	9040	2702468	0.11%
Tourist/Harbinger	15859	3420295	0.13%
Other (Mirage, P-element, Transib)	7028	1202716	0.05%
Rolling-circles	67687	19651620	0.77%
Unclassified:	1721247	439014757	17.3%
Total interspersed repeats:		1396993603	55.03%
Satellites:	382834	164062446	6.46%
Simple repeats:	562796	193327114	7.62%
Low complexity:	22198	1161877	0.05%

The NCBI Eukaryotic Genome Annotation Pipeline (https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/) was used to produce the NCBI *Ae. albopictus* Annotation Release 102.

A total of 26,856 protein-coding sequences were predicted in AalbF2, plus non-coding and pseudogenes up to a total of 40,494. Using RNA-seq data, 35,721 fully supported messenger RNAs (mRNAs) and 8,437 non-coding RNAs were detected. A detailed report of the gene annotation in AalbF2 and all the related gene model, mRNAs and other genomic features annotation files can be viewed and downloaded on https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Aedes_albopictus/102/. To help distinguish between artifacts and genuine gene duplications, which are resistant to proper assembly, and mitigating the heterozygosity effect from the original pooled DNA, we developed a pipeline based on the assumptions that selection acts mainly on the coding sequence of a gene and that homology between highly related paralogs drops in the flanking untranslated sequences. To perform the analysis, we compared 500 bp or 1,000 bp of the flanking regions at the 5' and 3' ends of all candidate gene duplicates with an all-against-all BLASTn search with an e-value of $1x10^{-40}$ for each flanking region. We found 1,329 (8.05% of total) genes with high similarity within 500 bp of their 5' and 3' flanking regions mapping on 452 of the 2,196 scaffolds. When we considered the extended 1,000 bp regions, the number of candidates duplicated was lower (808 mapping on 300 scaffolds: 4.89 % of total). Most of these artifacts involved a single duplicated gene (twins) and the number decreased with increasing copies (**Table 6**).

Table 6. Number of duplicated genes identified in the AalbF2 assembly by comparing the 3' flanking 500 and 1000 bp, respectively.

Number of Duplications	3' 500	3' 1000
di-	540	329
tri-	66	32
tetra-	8	6
penta-	3	4
hexa-	4	0
hepta-	2	1
octa-	0	2
nona-	2	0
deca-	0	0
undeca-	2	2
dodeca-	1	1

4.1.2 Construction of a physical map for *Aedes albopictus*

Cytogenetic comparison (**Table 7**) between *Ae. albopictus* and *Ae. aegypti* demonstrated that the total chromosome length is 4.9 µm or 16.4% longer in *Ae. albopictus* (P<0.0001), which indicates a slightly larger genome size in this species, as supported by cytofluorimetry.

We developed the first physical genome map of the *Ae. albopictus* genome using fluorescence in situ hybridization (FISH) on mitotic chromosomes covering 57% of the genome assembly by targeting twenty of the largest genomic scaffolds and three minor scaffolds (**Appendix 1).** A total of 4, 9, and 10 scaffolds were assigned to the chromosome 1, 2 and 3, respectively. The probes designed from scaffolds 15, 48, and 55 hybridized to places already covered with other large scaffolds. To provide an independent confirmation of the accuracy of the in-situ hybridization results, D-Genies was used with MiniMap2 to align AalbF2 scaffolds to the *Ae. aegypti* genome, which is assembled at chromosome-level [141]. This bioinformatic analysis revealed that the positions of all tested transcripts were consistent with their predicted positions on the *Ae. aegypti* chromosomes. Combining FISH, probe-mapping to the *Ae. aegypti* genome and homology between *Ae. aegypti* and *Ae. albopictus* chromosomes (**Figure 11**), we bioinformatically assigned the 58 longest scaffolds covering 75% of the genome (coinciding to the L75 metric) to *Ae. albopictus* chromosomes (**Appendix 2).**

Table 7. Comparative cytogenetic analysis between *Ae. albopictus* and *Ae. aegypti* chromosome.

Chromosome length/proportions	***Ae. albopictus***		***Ae. aegypti***
Chromosome 1			
Average length, µm	8		7.1
Relative length, %; P-value*	26.8	P<0.0001	28.4
Centromeric index, %; P-value	46.7	P=0.0044	46.9
Chromosome 2			
Average length, µm	11.7		9.5
Relative length, %; P-value	39.1	P<0.0001	38
Centromeric index, %; P-value	46.9	P=0.0051	48.6
Chromosome 3			
Average length, µm	10.2		8.4
Relative length, %; P-value	34.1	P<0.0001	33.6
Centromeric index, %; P-value	47.2	P=0.0079	47.4

There were also differences between the two species in chromosome proportions, such as relative chromosome and arm lengths. In *Ae. albopictus,* chromosome 1 was shorter, while chromosome 2 was longer relative to *Ae. aegypti*. Besides positioning and orienting the largest scaffolds, the 18S rDNA and other genomic features described below were physically mapped (**Figure 12**). These features include the largest viral integration we annotated in the genome and representative immunity genes. The 18S rDNA mapped in the region of the secondary constriction in region 1q22. The intensity of the signal significantly varied among chromosomes from individual mosquitoes suggesting variations in numbers of ribosomal genes.

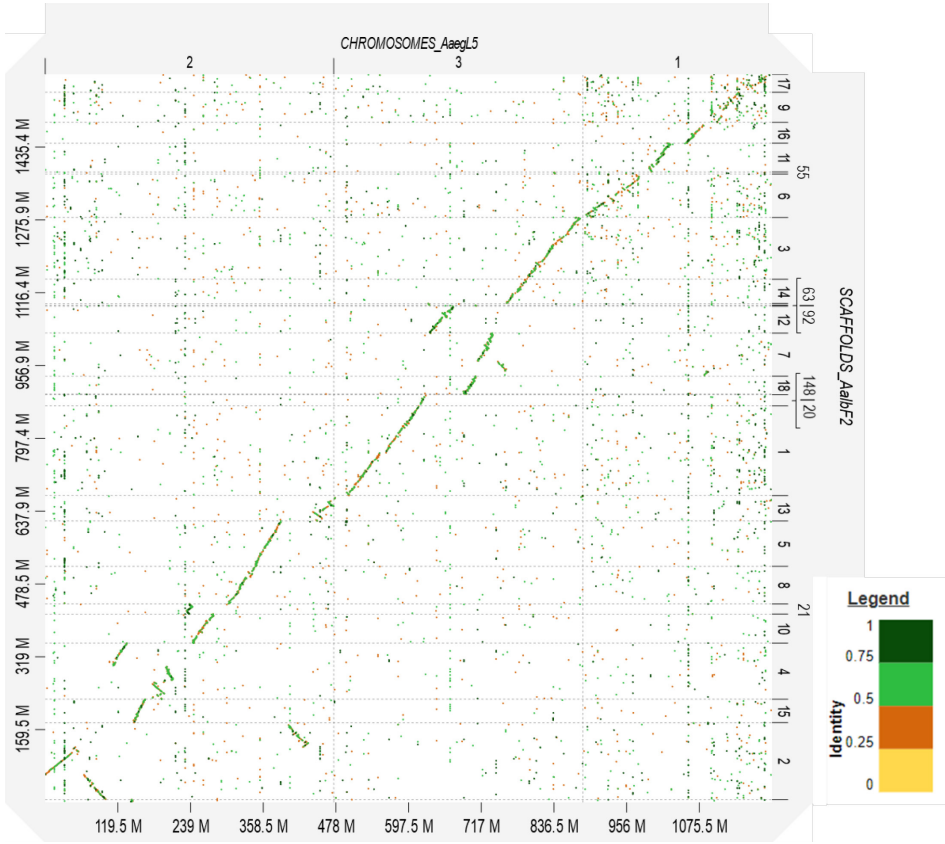


Figure 11. Comparative mapping on *Ae. aegypti* chromosomes (shown on the X axis) of the 18 largest *Ae. albopictus* scaffolds (shown on the Y axis), including those anchored to the chromosomes by in situ hybridization. The green dots and lines represent genome regions with highly similar sequences between the genome assemblies.

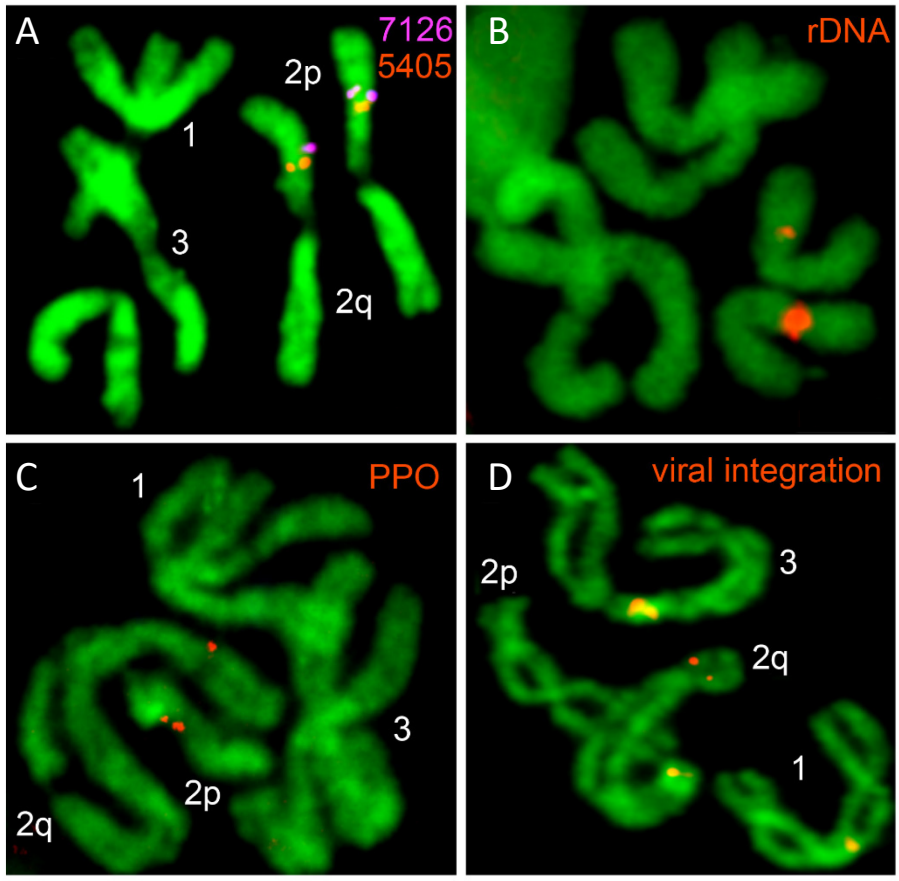


Figure 12. Examples of fluorescence in situ hybridization. Chromosomal locations of (**A**) transcripts XM_019675405 and XM_020077126 from scaffold 4 and 48, respectively; (**B**) rDNA; (**C**) prophenoloxidase (PPO) gene clusters; and (**D**) the largest viral integration in the genome (Canu-Flavi19) are demonstrated. Transcripts are indicated on a figure by the last four digits of their accession numbers.

4.1.3 The landscape of endogenous viral elements

In my 2017 article [112] I demonstrated the presence of viral integrations in *Ae. albopictus* and *Ae. aegypti* and observed that nrEVEs are 10-fold more abundant in these species compared to mosquitoes from the *Anopheles* and *Culex* genera.

I identified hundreds of nrEVEs in the genome of *Ae. albopictus* that are often contained in piRNA clusters and produce piRNAs. An approached based on BLASTx and a series of custom filtering scripts was used to screen the new AalbF2 genome assembly for sequences derived from 1,563 viral species included in 34 different families and 4 unclassified categories (**Table 1**). A total of 456 nrEVEs were identified and classified into ten taxonomic categories (**Figure 13**). The nrEVEs can be browsed on the online database of nrEVEs which is maintained by our laboratory (http://nreves.com/).

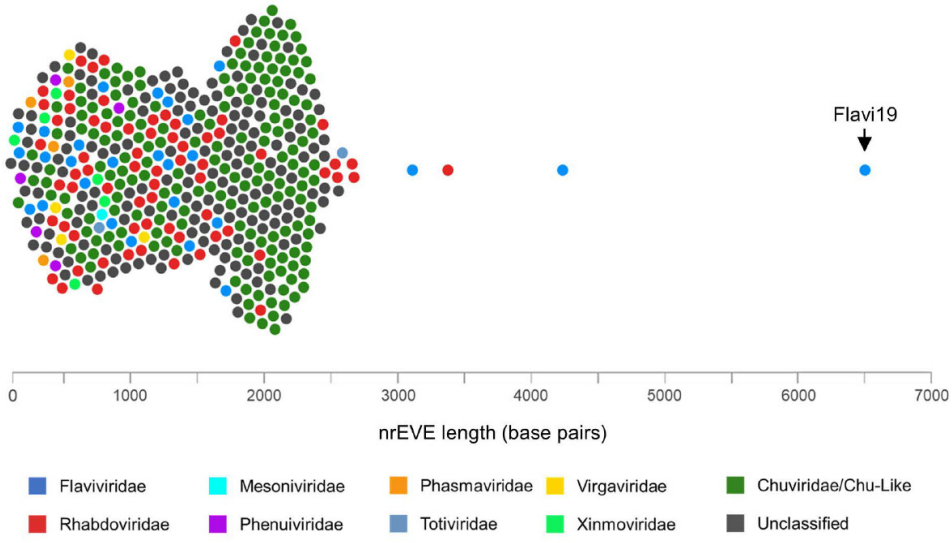


Figure 13. Beeswarm plot showing viral integrations identified in AalbF2. Each dot is a nrEVE, plotted according to its length and color-coded based on its viral origin. nrEVEs range in length from 131 to 6593 nt, with an average of 1289 nt. Arrow points to Canu-Flavi19, the longest nrEVE.

In total, viral integrations occupied 585,891 bp (0.02% of the total genome), spread across 115 out of the 2,197 assembled scaffolds. The most represented viral families from which nrEVEs originated are *Chuviridae*, *Rhabdoviridae* and *Flaviviridae*, but most nrEVEs are from recently discovered unclassified viruses, some of which are similar to classified chuviruses (**Figure 13, Figure 14**). We confirmed that the majority of nrEVEs in the *Ae. albopictus* genome have similarities to known ISVs. F-EVEs have similarities to known insect specific flaviviruses that are at the base of the phylogenetic tree of flaviviruses [188].

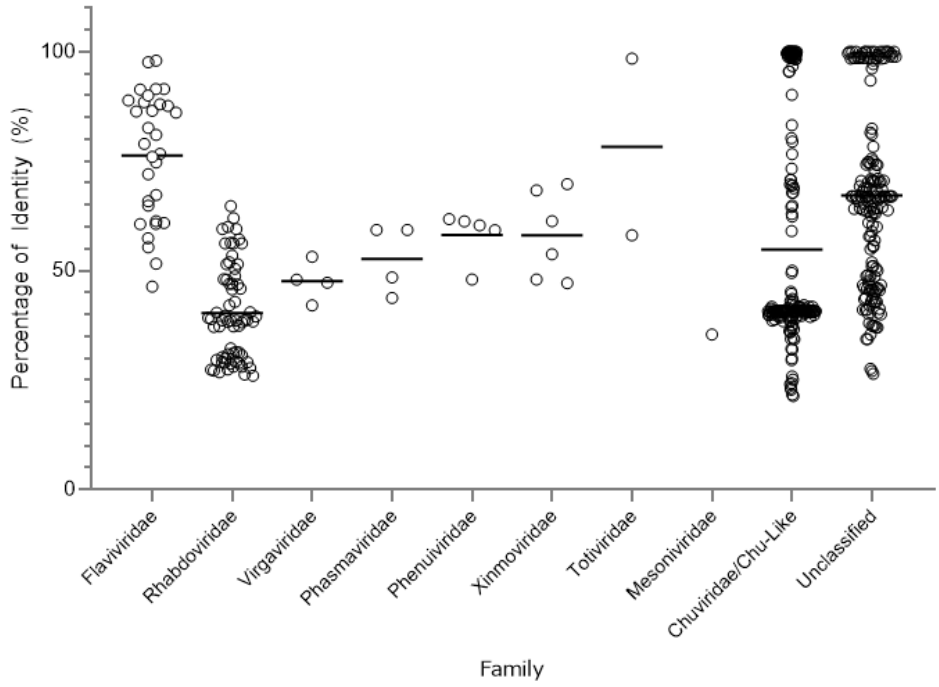


Figure 14. Scatter plot representing the amino acid identity of each nrEVE and its best hit retrieved by blastx searches against NR database grouped by viral family. The average is shown by a line.

When all F-EVEs are aligned against a representative flavivirus genome, they cover the whole genome with overrepresentation of the E (envelope) and NS5 (non-structural protein 5) coding regions (**Figure 15**). R-EVEs are similar to the N (nucleoprotein), G (glycoprotein) and L (RNA-dependent RNA polymerase) coding regions of different rhabdoviruses (**Figure 15**). Both F- and R-derived nrEVEs often tend to map less than 10kb to each other, generating clusters. Clusters may include both F- and R-EVEs or be composed uniquely of nrEVEs from one viral family. Integrations in a cluster that are derived from the same viral species are often rearranged or duplicated in comparison to the original viral genome.

To provide correspondence between the nrEVEs annotation in AaloF1 and AalbF2, the 72 nrEVEs identified in AaloF1 [112] were aligned against the 456 nrEVEs identified in AalbF2 with BLASTn [136].

The difference in the number of nrEVEs annotated in AaloF1 and AalbF2 can be explained by the different annotation methods and viral databases used. Of the nrEVEs annotated in AaloF1, 23 out of the 30 F-nrEVEs and 31 out of the 42 R-nrEVEs showed correspondance to one or more viral integrations identified in AalbF2 (**Appendix 3**). The remaining 7 F-EVEs and 11 R-EVEs were not correlated with any nrEVE identified in AalbF2.

CanuFlavi19, the largest viral integration identified in AalbF2 (6593bp) was not present in AaloF1. We validated the existence of this nrEVE, and we located it on chromosome 3 with FISH (**Figure 12**).

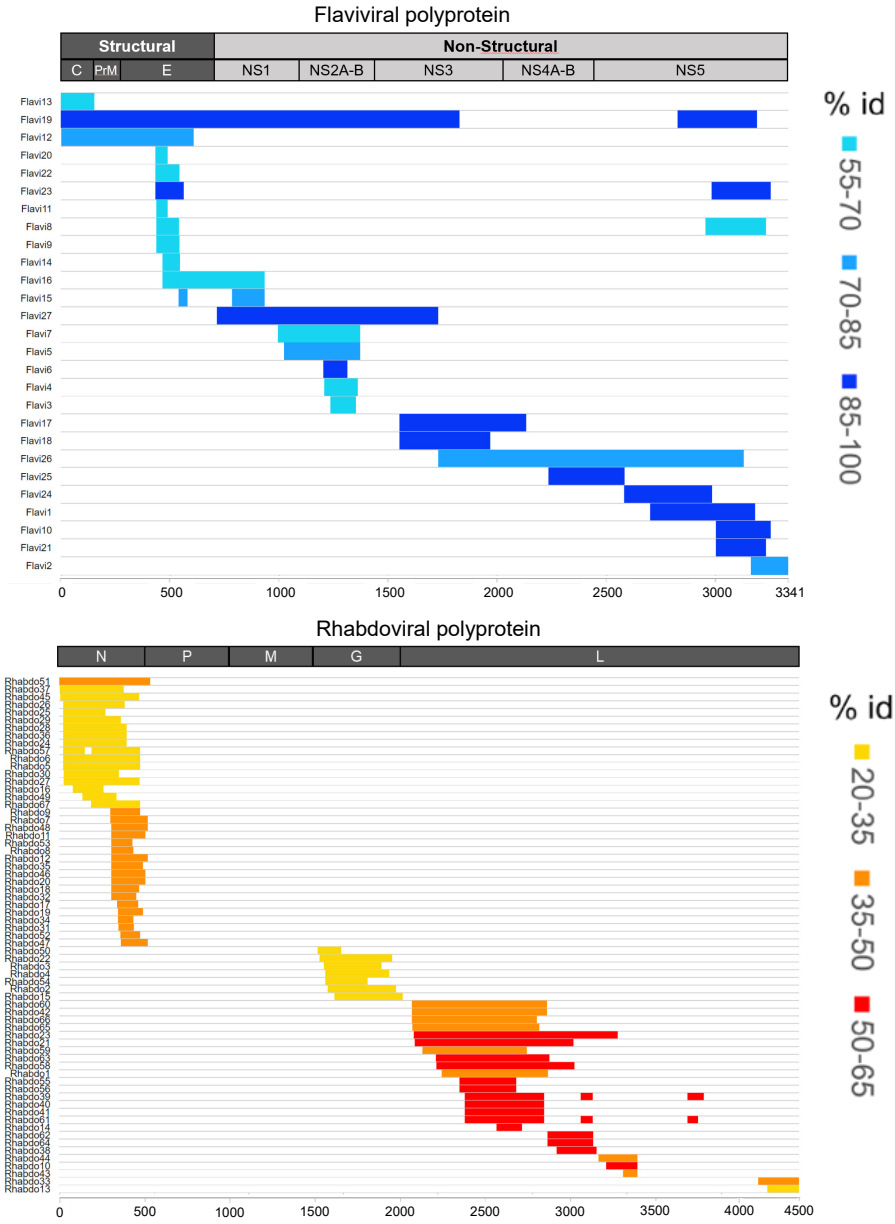


Figure 15. AalbF2 nrEVEs sequences overlapping on representative *Flaviviridae* and *Rhabdoviridae* genomes.

4.1.4 Distribution and structure of piRNA clusters

We and others have already demonstrated that in *Aedes* species piRNAs are produced from nrEVEs integrated in piRNA clusters [111–113]. An annotation of *Ae. albopictus* piRNA clusters had already been done on the AaloF1 assembly [189] but the number (643) and the maximum length (10kb) of the clusters were discordant to what was observed in *D. melanogaster* and other insects where piRNA clusters generally span 50–100 kb [85, 190]. This discrepancy was probably caused by the high fragmentation of the previous reference genome, assembled with short reads [115].

To produce a reliable piRNA cluster annotation for *Ae. albopictus*, we generated small RNA libraries from somatic tissues (female carcasses) as well as germline tissues (ovaries) because the piRNA pathway in mosquitoes is also active in the soma [191]. Each library comprised of more than 23 million miRNA-sized 18-24nt reads. A total of 1441 piRNA clusters were annotated with an average size of 10.9 kb and a maximum size of 139.9 kb (**Table 8, Figure 16**), covering 0.62% of the genome. This result is close to what it was found with the same approach in the latest *Ae. aegypti* assembly [141]. In contrast, using the same annotation pipeline on the highly fragmented *Ae. albopictus* AaloF1 genome assembly, a higher number of piRNA clusters (2467) smaller in size (average size: 5.9 kb; maximum size 64.2 kb) were detected (**Table 8**).

Only 31.8% and 47.3% of all piRNAs in the germline and soma, respectively, were mapped inside AalbF2 piRNA clusters. This fraction was nearly twice as large in *Ae. aegypti* (**Table 8**).

Table 8. Summary statistics on annotated piRNA clusters using the genome assemblies for *Ae. aegypti* AaegL5 and *Ae. albopictus* AaloF1 or AalbF2

	AaegL5	**AaloF1**	**AalbF2**
#clusters	1158	2467	1441
average size [kb]	11.9	5.9	10.9
max. size [kb]	333.1	46.2	139.9
piRNAs in clusters [%]			
Soma	71.3	49.9	47.3
germline	41.7	36.8	31.8
soma (unique)	86.4	69.5	72.9
germline (unique)	70.5	63.8	59.1

This is likely a result of the 14% duplication in the assembly as predicted by BUSCO. This may be leading to the exclusion of piRNA clusters without or with very few uniquely mapping piRNAs, required to annotate piRNA clusters. Indeed, when only considering unambiguously mapping piRNAs, the fraction of piRNAs included in clusters increases to 59.1 % and 72.9 % in germline and soma, respectively. The vast majority of all clusters expressed piRNAs from one single strand (uni-strand clusters) while only approximately one fifth of all clusters expressed piRNAs from both strands (dual-strand clusters). Dual strand clusters were mostly expressed in the germline. In contrast to *Drosophila*, relative piRNA expression from clusters varied substantially between somatic and germline tissues, with some clusters showing a soma-dominant expression and others being predominantly expressed in the germline. Blood-feeding had little impact on cluster expression. Analysis of publicly available small RNA libraries derived from the widely used *Ae. albopictus* C6/36 and U4.4 cell lines showed piRNA production from both somatic and germline clusters.

Drosophila piRNA clusters are highly enriched with TEs [85] unlike *Ae. aegypti* mosquitoes [192] even though total genomic content of TEs is much higher in *Ae. aegypti* than in *D. melanogaster*. Similarly, few *Ae. albopictus* piRNAs were derived from endogenous TEs [189], and repetitive sequences were not abundant in piRNA clusters except for hellions and LTR-retrotransposons (**Figure 17**). Interestingly, nrEVEs were enriched compared to the rest of the genome (**Figure 17**): 138 out of 456 viral integrations were contained within piRNA clusters, suggesting strong evolutionary pressure to integrate, or maintain, viral sequences into piRNA clusters.

This strengthens the hypothesis of a CRISPR-like immune function for piRNA clusters and the piRNA pathway [101].

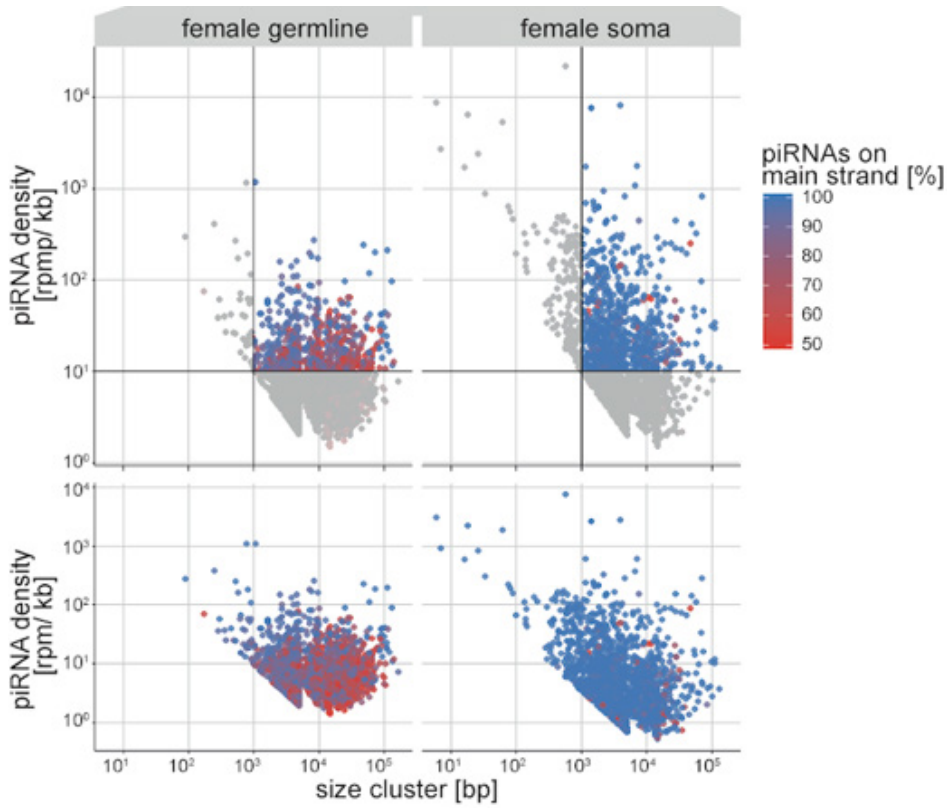


Figure 16. Scatterplots indicating the sizes and piRNA densities of annotated clusters, with clusters falling below the minimal length and density requirement highlighted with shades. Colours display the bias of piRNA expression towards one dominant strand (with 50% corresponding to an unbiased piRNA expression from both strands, dual-strand clusters). rpmp, reads per million mapped piRNAs. For comparative reasons, cluster density when normalizing to total small RNAs (rpm, piRNA reads per million mapped reads) is indicated in the lower panels

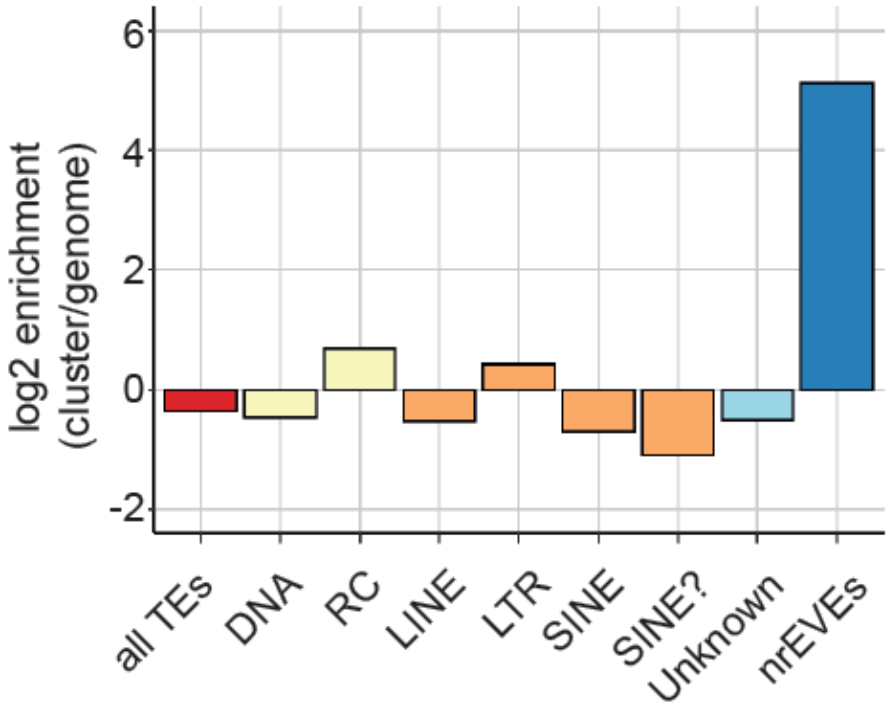


Figure 17. Enrichment of TEs and nrEVEs in piRNA clusters compared to the whole AalbF2 genome. Class I TEs (DNA and RC) and Class II TEs (LINE, LTR, SINE) are shown.

4.1.5 miRNA annotation

A comprehensive inventory of *Ae. albopictus* miRNAs is an important resource for investigating small RNA function in vector biology and mosquito antiviral immunity. Currently, *Ae. albopictus* miRNAs are not included in the most comprehensive depository of miRNA genes across all species, miRBase [140]. To bridge this gap, miRNA genes in AalbF2 were annotate using the miRDeep2 algorithm [139] on data from the small RNA libraries described above. The most abundant and robust miRNA-sized reads were derived from carcass samples, which is expected as small RNA libraries prepared from ovary samples are more biased towards piRNAs. Initially, miRDeep2 predicted 473 pre-miRNA loci in AalbF2, which after stringent filtering and manual inspection was reduced to 229 loci representing 121 distinct pre-miRNA species.

Amongst these, 92 predicted miRNAs loci were previously annotated in the *Ae. aegypti* genome, three were predicted based on conservation to miRNAs in other insect species, and 26 were entirely novel miRNA genes. Using these predictions, the expression of miRNAs in ovaries and carcasses were characterized and the changes induced by blood-feeding were analyzed. These analyses revealed that the most abundant miRNAs have a similar expression pattern between ovaries and carcasses (**Figure 18**).

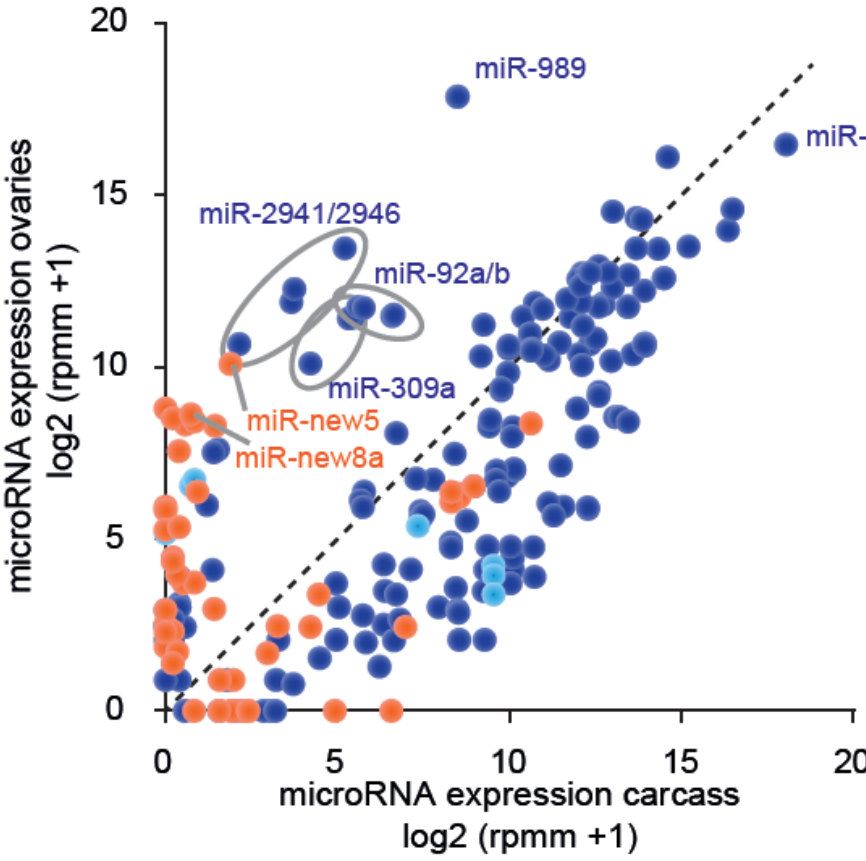


Figure 18. miRNA abundance in *Ae. albopictus* carcass and ovary samples. Counts for individual miRNAs were normalized to the total number of miRNAs in each dataset and expressed as log2 transformed reads per million miRNAs (rpmm) + 1. The mean of two independent libraries for each condition is shown. Blue and azure dots represent miRNA predicted on *Ae. aegypti* and other insects species, respectively. Orange dots are newly predicted miRNAs.

Nevertheless, a group of miRNAs, including miR-2941, miR-2946, miR-989, miR-309a, miR-92a/b (already annotated in *Ae. aegypti*) and a newly predicted miRNA, miR-new5, had a high expression level (>1000 reads per million miRNAs or RPMM) exclusively in the ovaries (**Figure 18**).

These findings are coherent with previous studies that demonstrated a specific expression in *Ae. aegypti* ovaries for the clustered miRNAs miR2941/2946 [193]. Similarly, miR-989 is known to be amongst the most abundant miRNAs in mosquito ovaries, both in *Anopheline* and *Aedes* spp. mosquitoes [194, 195]. miR-309 was found to be predominantly expressed in *Ae. aegypti* ovary tissue. miR-309 was furthermore shown to be strongly induced upon blood feeding both in *Aedes* and *Anopheles spp.* mosquitoes [196, 197], a behavior we also observed comparing sugar and blood-fed *Ae. albopictus* mosquitoes (**Figure 19**).

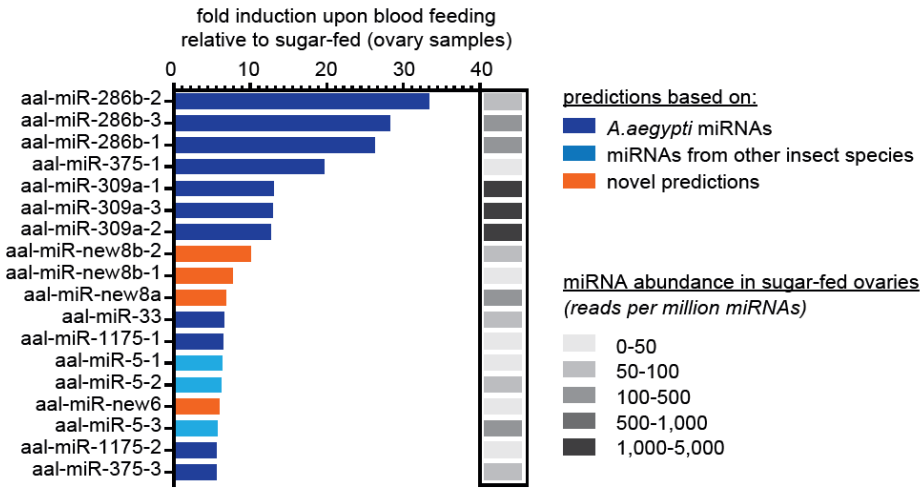


Figure 19. Fold induction of miRNA levels in blood-fed ovary samples compared to sugar-fed samples. Basal expression of each miRNA in the sugar-fed samples is indicated in grey scale.

Likewise, miR-286b and miR-375, were found to be strongly induced upon blood meal, as it has previously been shown in *Anopheles stephensi* and *Ae. aegypti* [197, 198].

Overall, our results support the observations that an orchestrated miRNA response to blood feeding in vector mosquitoes is present and that it is conserved between different mosquito species.

Most of the newly predicted miRNAs were predominantly expressed in ovaries (**Figure 18**), which likely reflects a sampling bias of previous studies that did not focused on predicting miRNAs from dissected ovary samples. Some of the newly predicted miRNA species are relatively abundant and are differentially expressed upon blood feeding, suggesting important functions in regulating the physiological processes that are induced upon blood meal.

4.1.6 Curation of immunity repertoire

Immune system functions are broadly classified into three main categories: recognition, signal transduction and effectors [199–201]. To catalog genes encoding the immune repertoire of *Ae. albopictus*, BLASTp was used to predict immune-related peptides in the AalbF2 annotated transcripts dataset using as a query 417 manually curated proteins of *Ae. aegypti* from ImmunoDB [199]. Our pipeline identified 663 putative immune-related genes encoding 979 predicted proteins, belonging to 27 functional groups (**Table 9**). A manual inspection of the 663 putative immune-related genes using a custom pipeline comparing their 5' and 3' flanking regions identified a set of 78 suspiciously duplicated genes that were distributed in half of the immune gene families (Table 2 and Additional file 8), reducing the total number of predicted immune genes to 622. This value is close to what is annotated in the AaloF1 assembly (521) and confirms the finding that the immune repertoire of *Ae. albopictus* is larger than that of other dipteran species [115, 199].

A detailed analysis of the immune repertoire of *Ae. albopictus* revealed extensive expansions in 16 of the 27 functional groups already annotated in *Ae. aegypti.* In mosquitoes, several signaling cascade pathways exist, including the Janus kinase-signal transducer and activator of transcription (JAK-STAT), Toll and immune deficiency (Imd) pathways. We detected an augmentation of genes involved in recognition and in genes associated with the Toll and IMD and Toll-1/Spz signal transduction pathways. On the contrary, immune effector genes did not display similar family-wide expansion. Five cecropins (CEC) genes encoding for antimicrobial peptides are known in *Ae. aegypti* but intriguingly only one CEC gene was found in the AalbF2 assembly. We detected an increase in the number of genes involved in all immune phases of the melanization pathway, a prominent pathway in arthropods that involves the synthesis of melanin to encapsulate pathogens [200]. The most extreme expansion event regards the CLIP proteases, which have multiple purposes in innate immune responses[202]. In AalbF2, 118 CLIP family members were identified, compared to 67 and 56 genes reported for *Ae. aegypti* and *An. gambiae,* respectively (**Table 9**).

Table 9. The repertoire of the immune genes manually annotated in The AalbF2 *Ae. albopictus* genome assembly, compared with the genes annotated in AaloF1, *Ae. aegypti* (Aae), *Anopheles gambiae* (Ag) and *Drosophila melanogaster* (Dm).

Gene Family or Pathway	AalbF2	AaloF1	Aae	Ag	Dm
Antimicrobial Peptides (AMPs)	6	11	16	11	7
Autophagy Pathway Members (APHAGs)	30 (28)	21	20	21	20
Caspase Activators (CASPAs)	3	4	4	2	5
Caspases (CASPs)	24 (18)	12	10	15	7
Catalases (CATs)	2	2	2	1	2
CLIP-Domain Serine Proteases (CLIP)	118 (109)	107	82	64	48
C-Type Lectins (CTLs)	66 (63)	48	44	27	35
Fibrinogen-Related proteins (FREPs)	52 (50)	49	38	53	17
Galectins (GALEs)	12	15	12	8	5
Gram-Negative Binding Proteins (GNBPs)	13	13	7	7	3
Heme Peroxidases (HPXs)	36 (34)	20	23	23	21
IMD Pathway Members	16 (13)	11	10	7	8
Inhibitors of Apoptosis (IAPs)	5	5	5	7	4
JAK/STAT Pathway Members (JAKSTATs)	3	4	3	3	3
Lysozymes (LYSs)	7 (6)	9	7	8	11
MD2-like Proteins (MLs)	36	26	27	18	10
Peptidoglycan Recognition P. (PGRPs)	18 (16)	13	10	7	13
Prophenoloxidase (PPOs)	23 (20)	16	14	9	3
Rel-like NFkappa-B Proteins (RELs)	4	4	3	2	3
Scavenger receptor (SCRs)	28 (27)	20	18	16	22
Serine Protease Inhibitors (SRPNs)	46 (45)	30	26	17	29
Small Reg. RNA Pathway Mem. (SRRPs)	51 (49)	41	39	26	24
Superoxide dismutase (SOD)	9	9	6	4	4
Spaetzle (SPZ)	13	13	9	6	6
Thioester (TE)-containing proteins (TEP)	7(5)	3	6	10	6
Toll Pathway Members (TOLLPATHs)	10	7	6	5	5
Toll Receptors (TOLLs)	26 (23)	14	12	9	9
Total	**663(622)**	**527**	**459**	**386**	**344**

Another interesting case involves the prophenoloxidases (PPOs), precursors for the activated phenoloxidases (POs) that oxidize phenolic molecules to produce melanin around invading pathogens and wounds [203]. The *Ae. aegypti* PPO gene family includes six tandemly arrayed genes: PPO4, PPO8, PPO7, PPO5, PPO1, and PPO2. The entire cluster of six genes was found to be locally duplicated twice in *Ae. albopictus*, resulting in 18 genes. The PPO clusters triplication was confirmed using *in situ* hybridization (**Figure 12C**). Duplication of PPO genes is not common in insects [204] and in mosquitoes the number of genes is strikingly higher than other insects. The high conservation of the PPO-organization and order in the array in both *Ae. aegypti* and *Ae. albopictus* supports the hypothesis that these duplications are ancient events that occurred 71.4 Mya before the split between the two species [115]. In *Ae. aegypti*, melanization is more complex than in other insects as there are at least two different pathways, immune and tissue melanization, that are controlled by distinct modules of CLIPs and serpins [200]. The PPOs expansion in *Ae. albopictus* suggests an even more complex melanization pathway which may have additional functions. Future studies focusing on dissecting the functional importance of specific family expansions in *Ae. albopictus* will be important to understand their significance for immunity, vector competence and ecological adaptation.

4.1.7 The sex-determining M locus

Insects employ different genetic pathways to establish sex-determination but all mechanisms share a common final step: the sex-specific splicing of the transcripts of two conserved genes, *doublesex* and *fruitless* which results in sex-specific isoforms of the DSX and FRU proteins that program sexual differentiation. In *Aedes* species, the male phenotype is conferred by a dominant sex-determining locus (M locus or M factor) that resides on one copy of chromosome 1 [205]. The M locus gene *Nix* is the only gene regulating the male-specific splicing of *doublesex* and *fruitless* and was first discovered in *Ae. aegypti* [205]. Recently it was demonstrated that the *Ae. albopictus nix* splice profile is more complicated and expressed more isoforms than *Ae. aegypti nix* but it was not possible to further characterize the gene [189]. In the AalbF2 assembly, *nix* was found in an approximately 917 kb scaffold (NW_021838423.1). By applying chromosome quotient analysis [156] using Illumina reads obtained from male and female mosquitoes of the Foshan strain [116] we demonstrated that *nix* is male-specific. A portion of the *nix* gene was previously identified in *Ae. albopictus* [205, 206] and its full-length sequence was described in the *Ae. albopictus* C6/36 cell line genome assembly [129]. The *nix* gene annotated in AalbF2 (LOC109397226 or XM_019669557.1), is composed of two exons flanking a small intron, a similar structure to that reported in *Ae. aegypti* [141]. Interestingly, there is an apparently defective copy of *nix* approximately 22 kb away from LOC109397226.

This *nix* pseudogene does not have an intact ORF and has a ~70% amino acid identity to XM_019669557.1.

In *Ae. aegypti*, no similar duplications were observed [141, 205]. In *Ae. aegypti* a second gene encoding a myosin heavy chain protein named *myo-sex* [157] has been demonstrated to be located in the M locus together with *nix* [141]. Knocking out *Myo-sex* in *Ae. aegypti* leads to flight-deficient males [207]. Two *myo-sex* homologs (XM_019707039.1 or XP_019562584.1; **Figure 20**) were found in two separate contigs (NW_021838603.1 and NW_021838542.1). It is unclear whether the gene that encodes XP_019562584.1 is also located in the M locus in *Ae. albopictus*, as the Chromosome Quotient analysis [156] was complicated by the presence of highly similar autosomal paralogs (e.g., AALF000603 and XP_019560880).

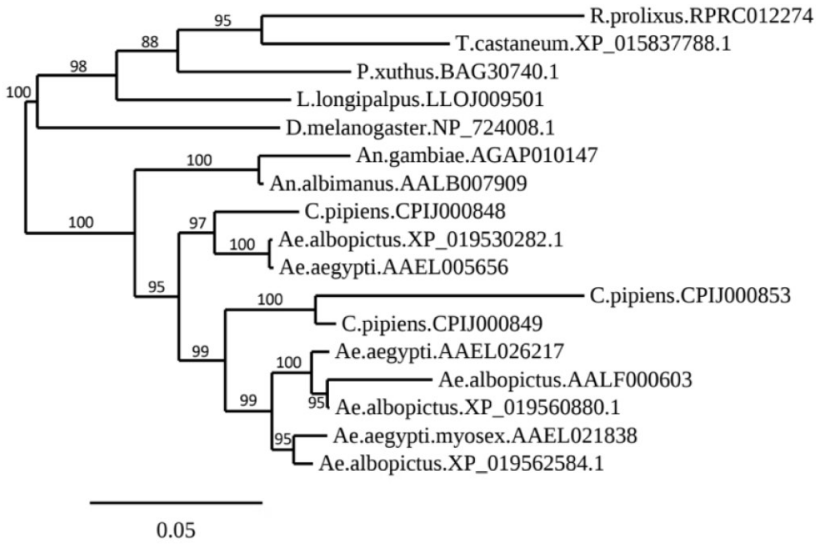


Figure 20. PhyML Tree shows inferred phylogeny of *Ae. albopictus* and select mosquito myosin family deduced protein sequences. Species names are followed by sequence ID. Bootstrap Proportion values are indicates at the node. The scale bar represents 5% nucleotide divergence. The *Ae. albopictus* sequence XP_019562584.1 appears to be the ortholog to *Ae. aegypti myo-sex.*

4.1.8 Genome-wide polymorphism and linkage disequilibrium

We used WGS data of mosquitoes from La Reunion island and Mexico [100], in which we searched for novel viral integrations as described above, and from the Foshan strain to show the usefulness of AalbF2 in analyzing the genomic diversity of *Ae. albopictus* populations. We obtained lower genetic diversity (π) estimates for the laboratory strain than for the wild populations, which is the hallmark of a population bottleneck in the laboratory strain (**Figure 21**).

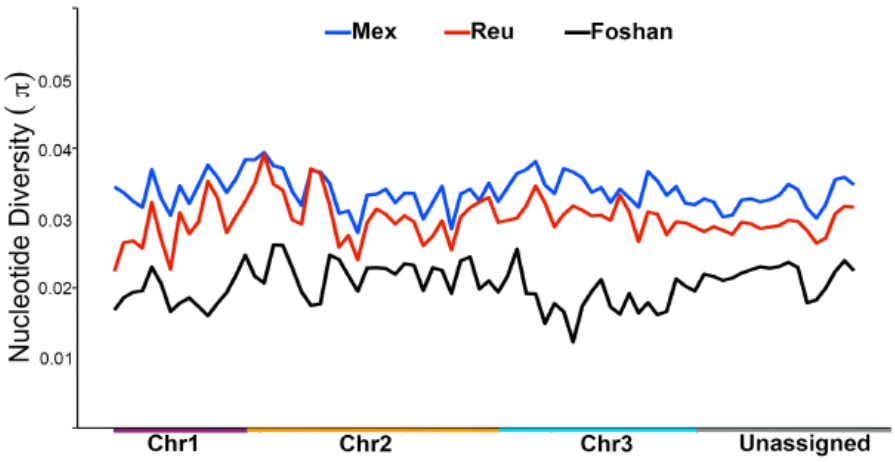


Figure 21. Overview of the pattern of nucleotide diversity (π) across the genome. Scaffolds were assigned to chromosome according to the chromosomal mapping previously described. Nucleotide diversity was measured using 50-kb-long sliding windows and 10-kb steps.

Genetic diversity is slightly higher for the recent invasive Mexican population than the old population from La Reunion. Global estimates of genetic differentiation (F_{ST}) among the three samples were 0.13 (Mexico vs Reunion), 0.21 (Mexico vs Foshan) and 0.18 (Reunion vs Foshan). Foshan was the most differentiated although being a laboratory strain. The levels of genetic diversity for the two wild populations and the Foshan strain varied greatly. Sliding window analyses across the genome showed regions having high and low genetic differentiation hotspots between the two wild populations (**Figure 22A**). LD analysis (**Figure 22B**) estimated the r^2 Max/2 across the three samples to be approximately 1.3 kb. These estimates are far smaller than the values calculated in *Ae. aegypti*, which range between 34 to 101 kb [141]. Technically, comparing these LD estimates can be difficult because of the differences in data generation (short-read WGS for *Ae. albopictus* and SNP-chip for *Ae. aegypti*).

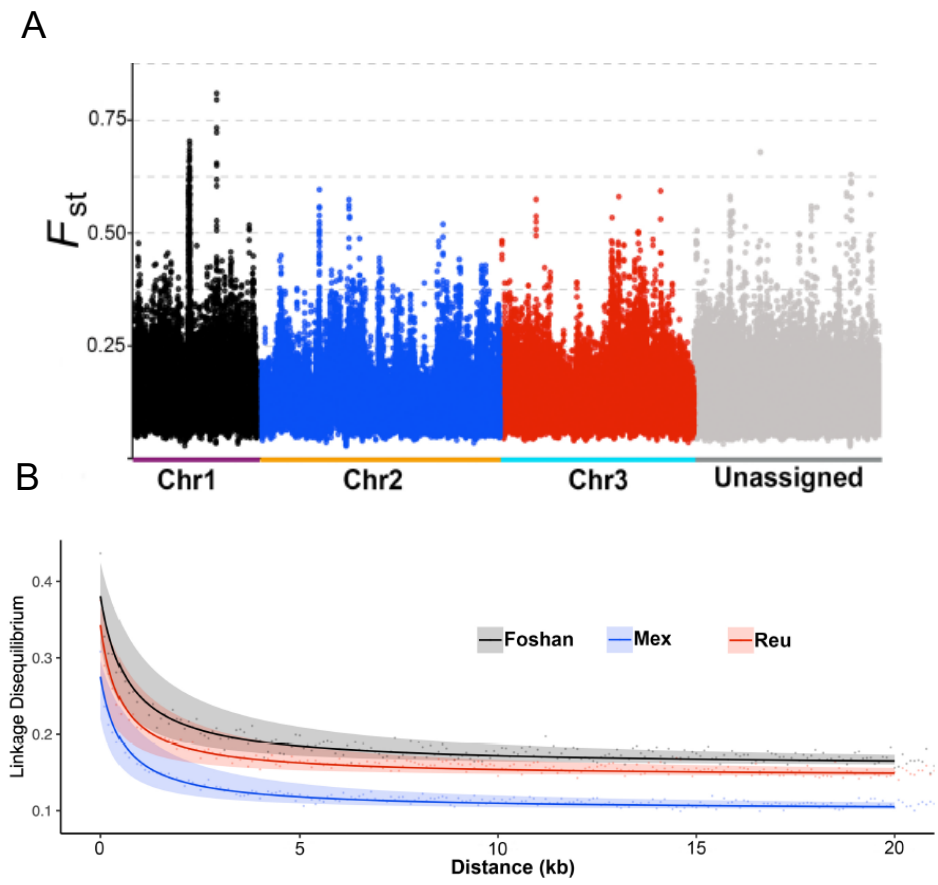


Figure 22. (**A**) Overview of the pattern of nucleotide diversity across the genome. Nucleotide diversity was measured using 50Kb long sliding windows and 10kb steps. (**B**) Linkage disequilibrium (r2). Red, green, and blue lines represent the fitting curve estimated with the ngsLD package and shaded areas around the lines represent confidence intervals from 100 bootstraps.

Still, this high difference makes sense in the context of the different colonization histories of *Ae. aegypti* and *Ae. albopictus* populations [208, 209]. *Aedes aegypti* experienced a slow colonization process that started in the seventh century, while *Ae. albopictus* suddenly dispersed in the past 50 years.

Also, *Ae. albopictus* chaotic dispersion resulted in genetic admixture among the invasive populations [13, 15, 210]. The age of mutations can affect LD estimates. Younger mutations generate higher LD values and it is possible that SNP-chip data differ from WGS in average age of mutations, as SNPs are estimated across the whole-genome with no or prior analyses with WGS approaches [211, 212]. Further experiments aiming to understand the spatial context of genetic signals and long-range patterns will be necessary and these types of analysis will be helped by the improved continuity of AalbF2.

4.1.9 Developmental Transcriptional Profile

As other insects, mosquitoes must regulate several complex processes like holometabolous metamorphosis, diapause, immunity, and blood-feeding. Understanding the network of expression throughout development could provide insights into biological functions implicated in the adaptation of this invasive species to different environments. Moreover, the knowledge on tissue- and time-specific expression in *Ae. albopictus* and the elements that regulates such expression could be identified from the analyses of transcriptional profiles and serve as bases to develop novel genetic-based strategies of vector control. The AalbF2 genome assembly and its predicted gene annotation (GCF_006496715.1, Annotation release 102) were used to obtain a comprehensive global view of gene expression dynamics throughout *Ae. albopictus* developmental phases taking advantage of recently produced Illumina RNA sequencing data from 47 unique samples representing 34 distinct stages of mosquito development [168]. Developmental time-course included 34 stages spanning the major developmental groups and both non blood-fed (NBF) and post blood meal (PBM) samples : M (male testes, male carcass), Fc (NBF carcass, and multiple timepoints PBM: 12hr, 24hr, 36hr, 48hr, 60hr, and 72hr), Ov (NBF ovaries, and multiple ovarian timepoints PBM: 12hr, 24hr, 36hr, 48hr, 60hr, and 72hr), D (diapause at multiple timepoints: 0-1wk, 1-2wk, 2-3wk, and 3-4wk), Emb (embryo at multiple timepoints: 0-1hr, 0-2hr, 2-4hr, 4-8hr, 8-12hr, 12-16hr, 16-20hr, 20hr-24hr, 24-28hr, 28-32hr, 32-36hr, 36-40hr, 40-44hr, 44-48hr, 48-52hr, 52-56hr, 56-60hr, 60-64hr, 64-68hr, and 68-72hr embryos), L (larvae 1st, 2nd, 3rd, and 4th instar larvae stages), and P (pupae, early male and female, and late male and female pupae stages). In total, 1.56 billion reads corresponding to total sequence output of 78.19 Gb were used to build the RNA-sequencing dataset. The short reads alignment rate on AalbF2 was of the 94.1%. The number of uniquely mapped reads increased significantly in comparison to the C6/36 cells genome assembly [129] suggesting that most of the coding regions were correctly assembled in the genome.

As an additional confirmation of the increased completeness and continuity of AalbF2, the number of spliced alignments increased substantially from 39,991,260 in the assembly of the C6/36 *Ae. albopictus* cell line (canu_80X_arrow2.2, *17*) to 56,243,825 in AalbF2 (40.64% increase).

Gene expression profiles across all developmental time points were computed and expression was calculated in Transcripts Per Million (TPM). The number of expressed genes (Transcripts Per Million ≥1) was observed to gradually increases through embryogenesis, reaching its highest peak at 68-72 hrs after the fecundation. As previously observed, number of expressed genes increased during the early pupal stages and the male germline exhibited the highest number of genes expressed among all samples [168]. After a blood meal, female mosquitoes undergo a series of physiological changes to metabolize the ingested blood and support oogenesis. Accordingly, in blood-fed female ovaries, the number of genes expressed in the germline changes dramatically from 12 hrs to 36 hrs post blood-feeding (PBM) PBM. Pairwise correlation analysis revealed that almost every developmental stage is most highly correlated with its adjacent stage and is very similar to what was previously found [168]. To visualize the various patterns of gene expression and the relationships between the samples, hierarchical clustering and Principal Component Analyses was performed (**Figure 23**).

These analyses revealed that embryos, PBM ovaries, pupae, larvae, and PBM female carcass samples tend to cluster closer together. This is an expected result since their gene expression profiles are similar as these are developmentally related samples. Two notable exceptions include the male testes and early embryos (0-1 hrs, 0-4 hrs, and 4-8 hrs), likely due to transcripts related to the maternal-to-zygotic transition (**Figure 23**). The male testes sample clusters away from all other samples, reflecting a distinguishing difference between this sample as compared to other samples sequenced.

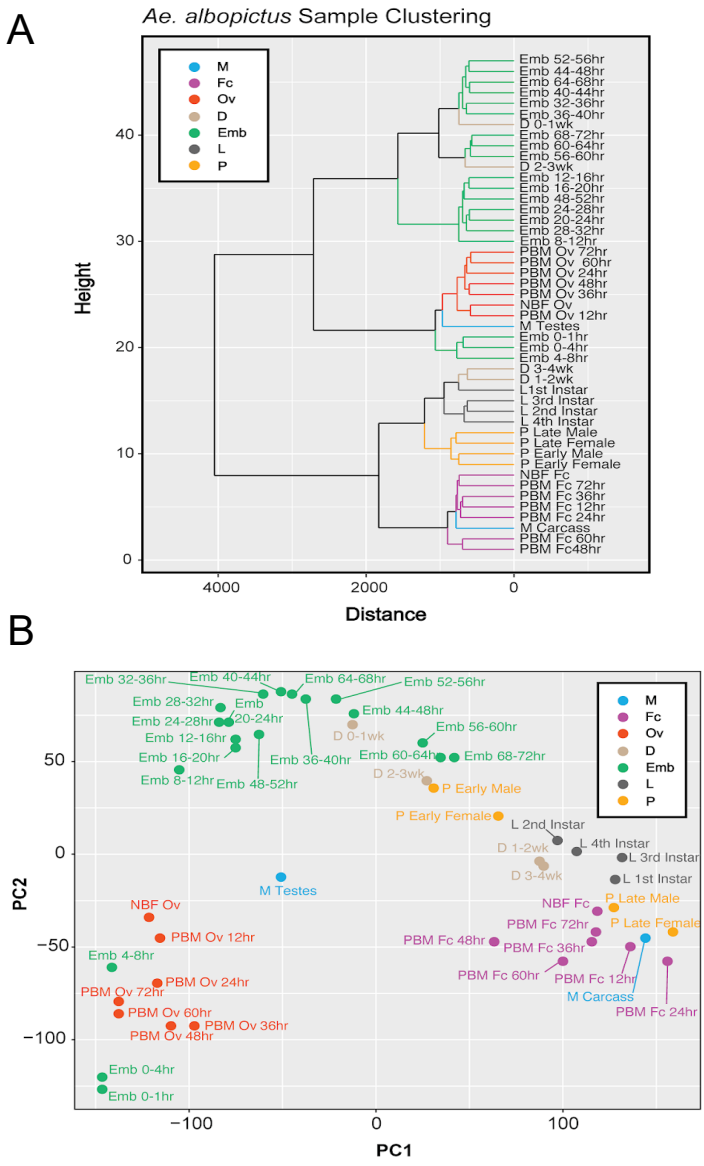


Figure 23. (A) Dendrogram of *Ae. albopictus* samples clustering similar life stages closer together. The plot depicts the close relationship between all developmental samples. **(B)** PCA clustering of *Ae. albopictus* samples depicting clustering of life stages which show close similarity.

4.2 The landscape of novel nrEVEs in wild collected *Ae. albopictus* mosquitoes

At the start of my PhD, in 2017, the identification of viral integrations using WGS data from wild-collected *Ae. albopictus* was a challenging task because: 1) repetitions are abundant (>50% of the genome) in *Aedes* spp. mosquitoes [115, 141]; 2) Vy-PER and comparable methods to identify viral integrations using WGS data were generated in the context of cancer genetics, thus they are geared towards the human genome [213]; 3) The AaloF1 *Ae. albopictus* genome assembly was highly fragmented. Repetitions and small scaffolds in a genome cause biased read alignments, thus preventing unambiguous identification of nrEVEs. A more reliable detection of novel viral integrations was made possible by the creation of the improved AalbF2 genome assembly as described above, which is made by considerably longer scaffolds. Furthermore the ViR pipeline [173], that was developed in our laboratory, allowed for a more precise prediction of novel nrEVEs, including from WGS of pooled individuals where there is a high variability in read sequences. From mosquitoes collected in China and La Reunion, I extracted both DNA and RNA. DNA was used for WGS, while RNA was used for small-RNA sequencing to study the smallRNA profile of nrEVEs and concomitantly identify viral fragments in mosquito pools. Thus, the purpose of this bioinformatical analysis was not to fully characterize the virome of *Ae. albopictus* (for which RNA sequencing data would have been more indicated) but rather to start gaining insights on the possible correlation between nrEVEs and infecting viruses.

4.2.1 Novel nrEVEs discovery *in Ae. albopictus*

4.2.1.1 Novel viral integrations in mosquito natural populations worldwide

The landscape of nrEVEs was already observed to be variable across worldwide mosquito populations, and even among individuals of the same population [116].

To further confirm this observation, I analyzed the presence and absence of the F- and R-nrEVEs annotated in the AalbF2 genome assembly in single mosquitoes collected in La Reunion island (a relatively old population) and Tapachula, Mexico (an invasive population). As expected, some nrEVEs were observed with a high frequency in both populations, while others were absent or rare only in a specific population (**Figure 24**)

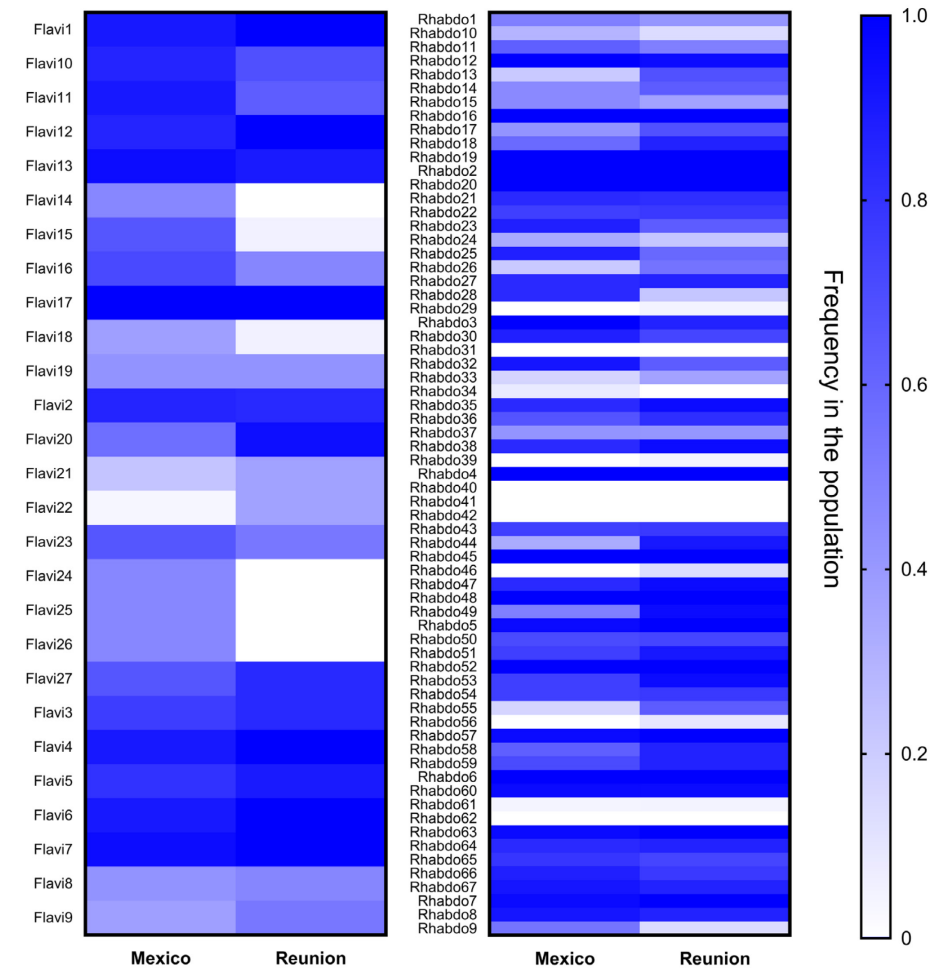


Figure 24. Frequency for the identification of AalbF2 F- and R-nrEVEs in samples from Mexico and La Reunion island.

A total of 13 selected nrEVEs annotated in the AaloF1 assembly were found, by PCR, to be differentially distributed in native (China and Thailand), old (La Reunion Island) and recent (United States and Italy) *Ae. albopictus* populations so that a tree built from a matrix of shared-allele distances proved able to differentiate mosquito populations in accordance with the historical records of *Ae. albopictus* invasive process [116].

This result, together with the abundance of nrEVEs in piRNA clusters [112], led to hypothesize that nrEVEs are a dynamic component of the mosquito repeatome. On this basis, we further hypothesized that novel viral integrations could arise in wild mosquitoes depending on their viral exposure [109]. To validate this hypothesis, wild-caught mosquitoes were initially collected in Thailand, La Reunion Island, Mexico, and Italy (**Table 10**).

Table 10. WGS Samples available for the search of novel nrEVEs.

Country	Location	Sample type	Number of samples
Thailand	Chiang Mai	Pool, 40	3
Italy	Crema	Pool, 40	3
Mexico	Tapachula	Pool, 40	3
	Chiapas region	Single Female	23
La Reunion	Saint Pierre	Pool, 40	3
	Saint Paul	Pool, 30	2
	Le Tampon	Pool, 30	2
	Bras-Panon	Pool, 30	2
	Saint Rose	Pool, 30	2
	Le Tampon	Single Female	24
China	SMU campus (Guangzhou)	Pool, 30	3
	Baiyun district (Guangzhou)	Pool, 30	1

From each of these four localities we formed and sequenced three pools of 40 individuals. Given the importance of La Reunion Island during the *Ae. albopictus* expansion out of its native home range, additional pools were collected in different localities of the island (**Figure 25**) and sequenced. Two pools of 30 individuals each were collected from each of the following locations: Saint-Paul and Le Tampon on the west side of La Reunion island; Bras-Panon and Saint-Rose from the east side. I also collected pools of 30 samples in Guangzhou (**Figure 25**), the largest city in the Southern China province of Guangdong which is an endemic area for DENV.

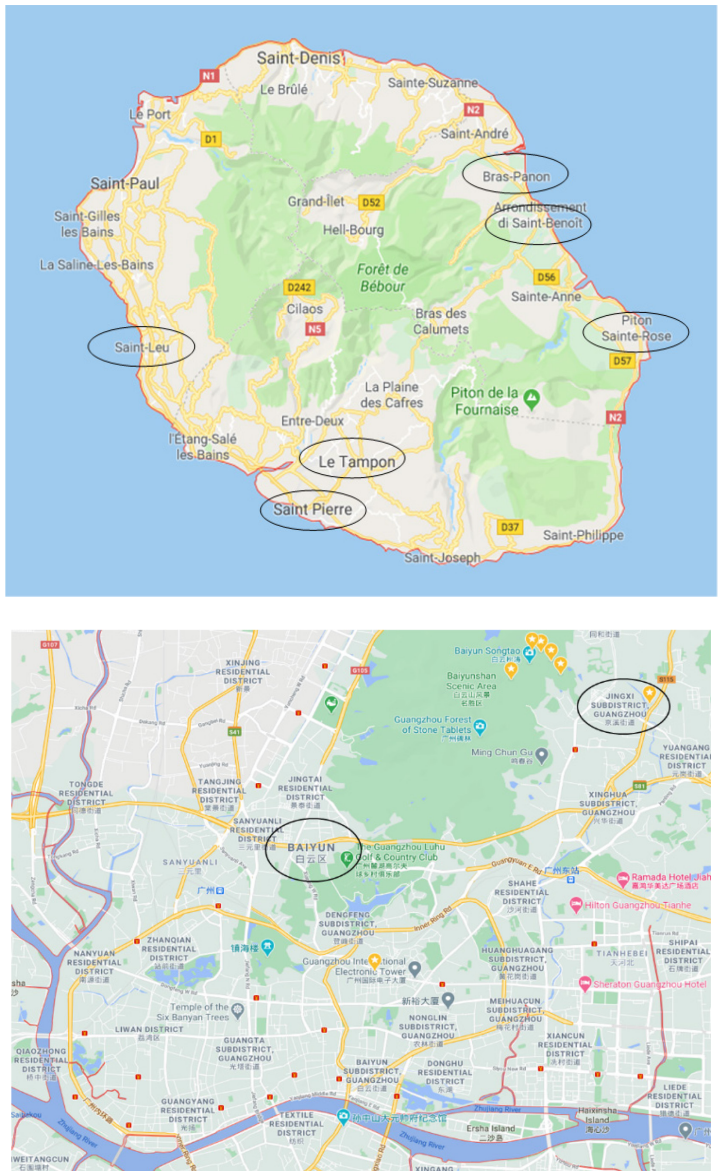


Figure 25. *Ae. albopictus* samples collection locations in La Reunion Island (upper picture) and Guangzhou (lower picture).

Three samples were collected inside the campus of the Southern Medical University (SMU) in Guangzhou while one sample was collected in the Baiyun district, a partially wooded area of the city where a high number of DENV cases were registered.

Lastly, 23 and 24 individual mosquitoes emerged from wild-collected eggs in the Chiapas region in Mexico and Tampon, in the island of La Reunion, respectively, were sequenced to have single individual genomes in addition to pooled samples.

All these samples were sequenced and inspected with the Vy-PER and ViR pipelines for new nrEVEs in collaboration with Elisa Pischedda, our bioinformatician. As described in the Methods and Materials section, Vy-PER and ViR are pipelines designed to identify viral sequences integrated in a genome from WGS data, given a reference genome and a database of viruses. Briefly, the pipeline aligns WGS paired-end reads on the reference genome, identifies unmapped reads containing viral sequences and predicts an insertion locus for these reads according to the position of their mapped pairs, which must derive from the same short fragments as the unmapped reads [172, 173]. For this part of my thesis, I worked in close contact with a Master student in Molecular Biology and Genetics, Annamaria Mattia, who contributed to the molecular validation of the novel viral integrations which were predicted bioinformatically.

Initially, a total of 33 novel nrEVEs were identified across all the samples. Some of these were shared between samples from different countries and were grouped together resulting in 13 distinct novel nrEVEs. Two novel nrEVEs were later merged into a single one (nrEVEnew-4/6) because we discovered that they represent the 5' and 3' end of the same integration, thus giving a total of 12 novel nrEVEs (**Table 11**). These viral integrations were absent in both the AaloF1 and AalbF2 genome assemblies and while some were bioinformatically identified in mosquitoes collected from multiple locations (i.e. nrEVEnew-2, 3, 4/6, 5 8 and 11), others were exclusive of one population (i.e. nrEVEnew-1, 7, 9, 10, 12, 13). All novel nrEVEs appear to be derived from ISVs of the *Flaviviridae* family and derive from different proteins (**Figure 26A**). The ViR_LTFinder script was used to extend the upstream and downstream ends of the putative novel nrEVEs to identify the integration sites. All the novel nrEVEs are flanked by repeated sequences. With the exception of nrEVEnew-10, 12 and 13 it was possible to identify at least the upstream or downstream flanking sequences of the viral integrations. Analogously to the annotated nrEVEs in AalbF2, all novel nrEVEs were flanked by repeated sequence. For 5 out of the 12 novel nrEVEs the flanking regions were classified as unknown TEs. Being flanked by these elements, which are very abundant in *Ae. albopictus* [183], makes it impossible to determine the precise genomic location of the viral integrations. It was possible to molecularly confirm in mosquito samples the first six novel nrEVEnews (counting nrEVEnew-4 and 6 as a single nrEVE) (**Figure 26B**).

nrEVEnew-8 was supported by multiple reads, but it was not amplified by PCR due to its rarity or to difficulties in designing unique primer pairs. nrEVEnew-9 to -13 are currently being tested by PCR

Table 11. WGS Samples available for the search of novel nrEVEs.

Candidate name	Virus	Size (bp)	Locations
nrEVEnew-1	KRV	295	Saint Rose (Reunion) Le Tampon (Reunion)
nrEVEnew-2	CFAV	57	Saint Rose (Reunion) St. Pierre (Reunion) Chiapas (Mexico)
nrEVEnew-3	KRV	306	Bras-Panon (Reunion) Saint Rose (Reunion) Le Tampon (Reunion) Baiyun (China)
nrEVEnew-4/6	KRV	534	Saint Rose (Reunion) Saint Paul (Reunion) Saint Pierre (Reunion) Le Tampon (Reunion) Crema (Italy) Chiapas (Mexico)
nrEVEnew-5	KRV	880	Saint Paul (Reunion) Le Tampon (Reunion) Chiapas (Mexico) Chiang Mai (Thailand)
nrEVEnew-7	AeFV	318	Saint Rose (Reunion)
nrEVEnew-8	CFAV	180	Saint Rose (Reunion) Tapachula (Mexico)
nrEVEnew-9	AeFV	117	Chiang Mai (Thailand)
nrEVEnew-10	KRV	138	Chiang Mai (Thailand)
nrEVEnew-11	KRV	946	Chiapas (Mexico) Baiyun (China)
nrEVEnew-12	AnFV	123	Tapachula (Mexico)
nrEVEnew-13	CFAV	104	Chiapas (Mexico)

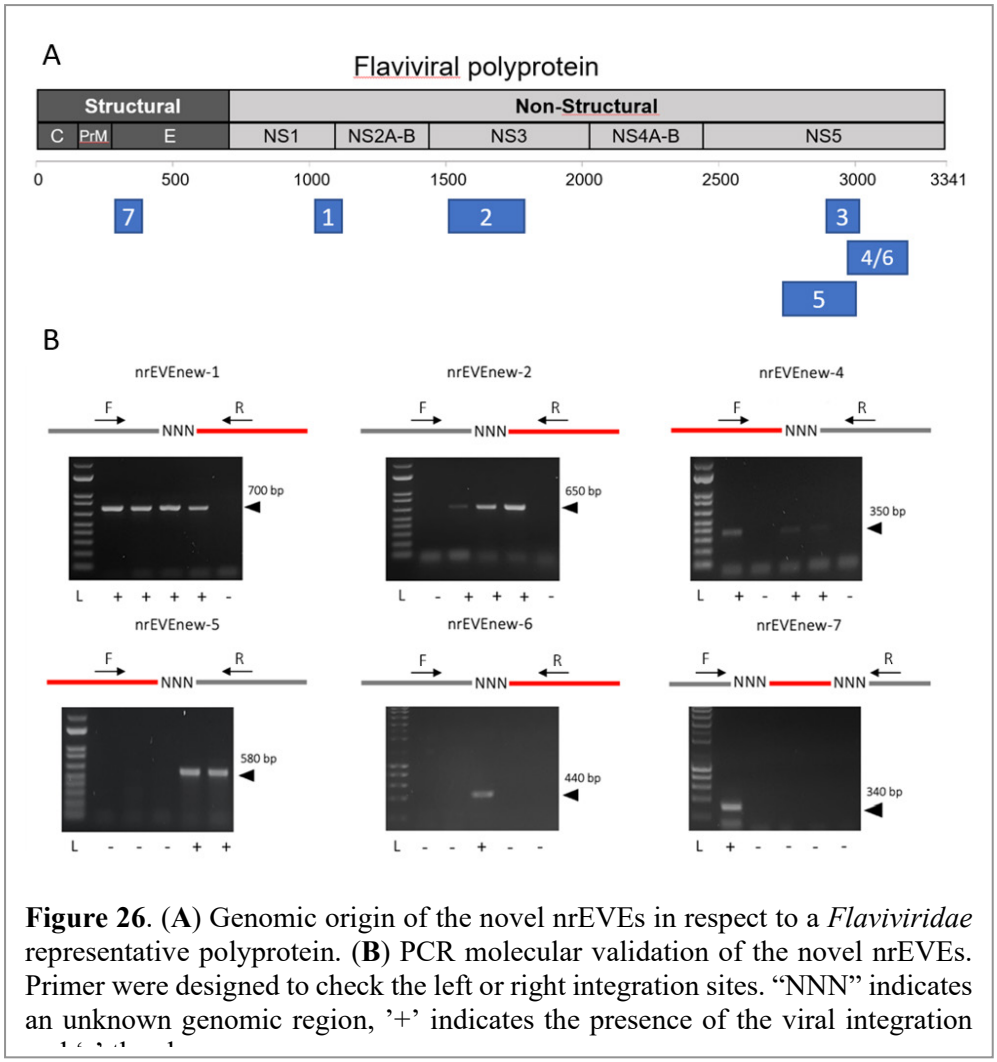


Figure 26. (**A**) Genomic origin of the novel nrEVEs in respect to a *Flaviviridae* representative polyprotein. (**B**) PCR molecular validation of the novel nrEVEs. Primer were designed to check the left or right integration sites. "NNN" indicates an unknown genomic region, '+' indicates the presence of the viral integration

4.2.1.2 Analyses of the widespread of nrEVEs in different Ae. albopictus strains

The specific primers used to molecularly validate novel viral integrations (**Table 2**) were used to test mosquitoes for nrEVEs presence and to get a view of the frequency of the distribution of these nrEVEs.

This analysis was conducted using mosquitoes from Canton, Chiang Mai, Crema, Tapachula and La Reunion. Novel nrEVEs were found with a frequency between 4% to 90% (**Figure 27**).

Notably, some nrEVEs were found in more locations than the ones in which they had been bioinformatically predicted. For example, nrEVE-new3 was identified in WGS of mosquitoes from La Reunion and Canton but was found also in Crema and Chiang Mai. Novel nrEVEs of *Ae. albopictus* appear broadly distributed across mosquitoes from different origin, unlike the novel viral integrations we identified in *Ae. aegypti* that were mostly population-specific [214]. This difference probably reflects the quick and chaotic invasion history of *Ae. albopictus* [6, 15].

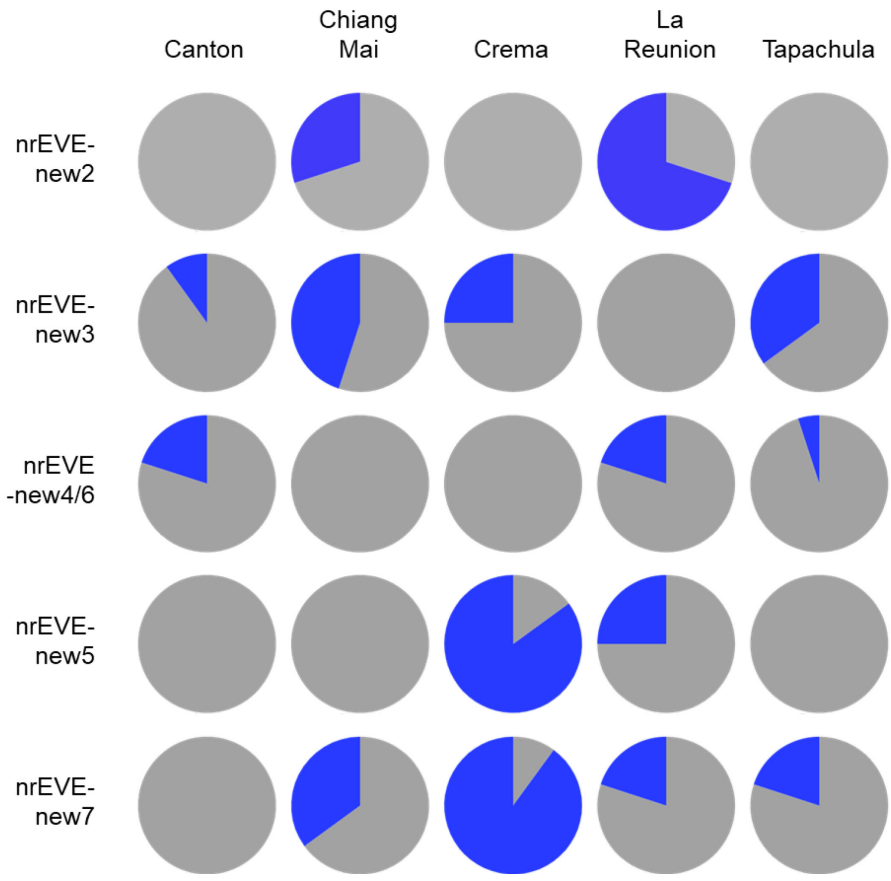


Figure 27. Frequency of novel nrEVEs in different *Ae. albopictus* laboratory strains. Blue and grey indicates the presence and absence of the novel nrEVEs, respectively.

4.2.2 Analysis of the RNA virome from small RNA data

In addition to WGS, also small-RNA data were produced from samples collected in La Reunion and Guangzhou (**Table 12**). In Guangzhou, I also collected *Ae. albopictus* larvae and pupae and these samples were used for small-RNA sequencing.

These small-RNA data contain mostly small molecules produced from the mosquitoes in the context of cellular activities (e.g. miRNA) and antiviral immunity (piRNA, siRNA) and thus they represent a picture of the RNA viruses eliciting a small-RNA mediated immune response in the samples rather than a total, unbiased, virome. In the context of nrEVEs these data may be useful to correlate ongoing infections in wild populations with nrEVEs of that sample.

Table 12. Samples used for WGS and Small RNA-seq analyses.

Country	Location	Sample type	Number of samples
La Reunion	Saint Paul	Pool, 30	2
	Le Tampon	Pool, 30	2
	Bras-Panon	Pool, 30	2
	Saint Rose	Pool, 30	2
China (Guangzhou)	SMU campus	Pool, 30	1
	SMU campus	Pool, 30 (Larvae)	1
	SMU campus	Pool, 30 (Pupae)	1
	Baiyun district	Pool, 30	1

As previously described, a pipeline to remove mosquito reads, assemble transcripts and extract transcripts with viral identities was run on the samples. The pipeline identified 122 and 126 viral transcripts for a total of 31,141 and 39,345 bp in the samples from La Reunion and China, respectively.

Transcripts derived from the same viral species were clustered, revealing 12 viral species which were found in samples from both locations and 16 and 14 location-specific viruses for China and La Reunion, respectively. (**Figure 28**). As expected, most of the identified viruses were ISVs. Remarkably, five viruses belonging to the *Rhabdoviridae* family were found in multiple samples from La Reunion island while only one rhabdovirus was found in a single Chinese sample.

Insect specific flaviviruses were found in some samples from both China and La Reunion. In particular, KRV and AeFV were found in both locations while CFAV was found only in La Reunion. These viruses are relevant because most of the novel nrEVEs identified in the wild-collected mosquitoes described in the previous section have a high similarity with these viral species. Indeed, when using BLASTX to compare the novel nrEVEs against all the proteins in the NCBI nr database, KRV, CFAV and AeFV were (with one exception) the best hit (**Table 11**).

Bunyavirales species were detected in all samples from China, including larvae and pupae, suggesting that these two viruses persistently infect the population, and were absent in samples from La Reunion. Similarly, the only *Totiviridae* found by this analysis was widespread in La Reunion samples but absent in Chinese samples. Some of the viral species that were found only in the Chinese samples (e.g. Wuhan Mosquito Virus 1, Shinobi tetravirus, Wenzhou sobemo-like virus 4) were previously identified in metagenomics analysis from Chinese arthropods [215].

Overall, the number of viruses found in the Chinese samples increased in subsequent developmental stages (**Figure 28**): 4 in larvae, 10 in pupae, 12 and 17 in adults. This may be a consequence of an increased risk of exposure to viral agents as the life stages of an individual progresses. Adults collected in the Baiyun district had more viral species, consistently with being in a semi-wooded area where few vector control methods have been applied.

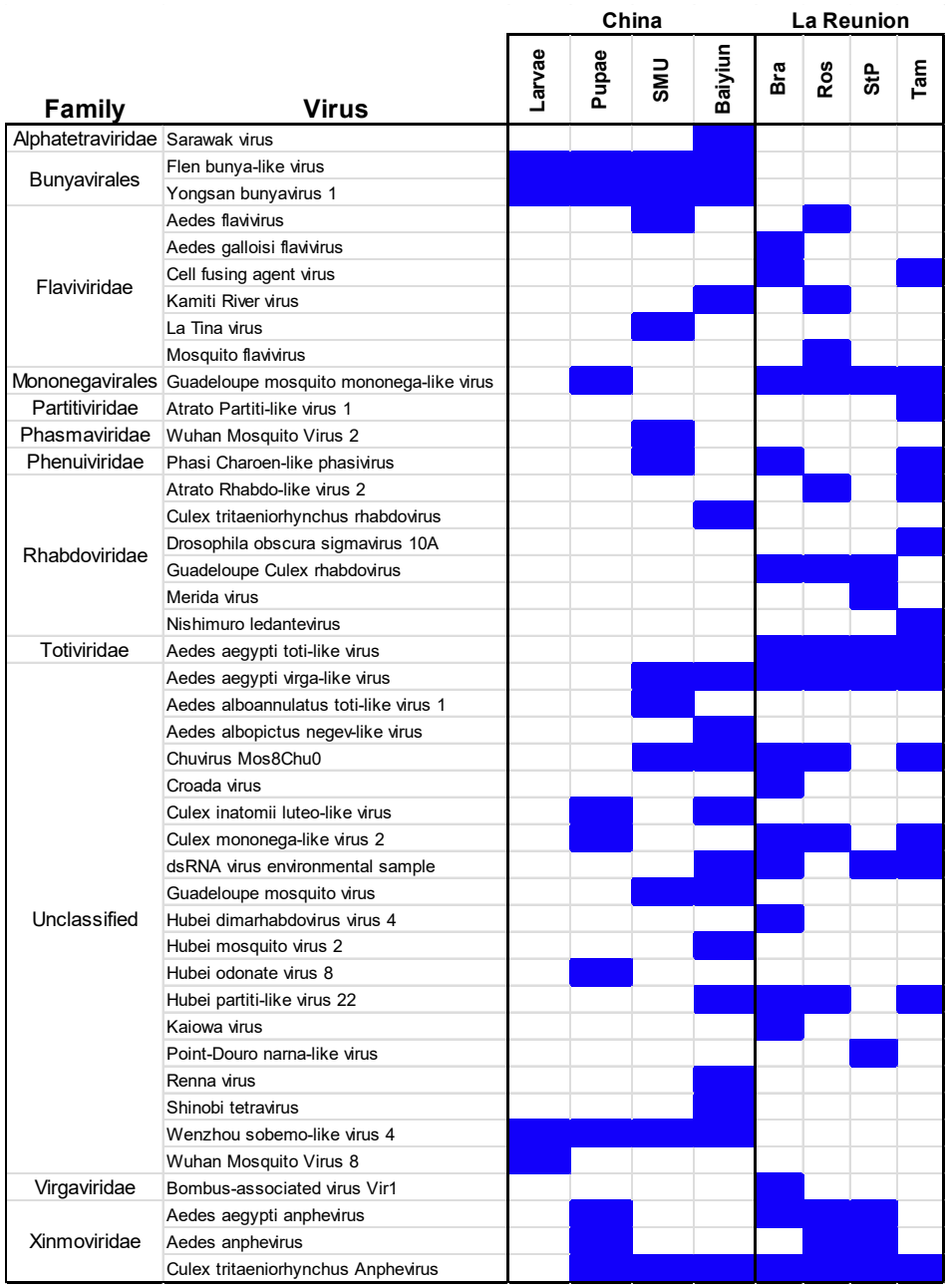


Figure 28. Comparison between viral species identified in samples from China and La Reunion island. SMU=Southern Medical University; Bra=Bras-panon; Ros=Saint Rose; Stp=Saint Paul; Tam=Le Tampon.

4.3 Genetic modification of the *Ae. albopictus* genome

This task was carried out in the laboratory of Prof. Jason Rasgon at the Penn State University where I spent two months in 2019. My work at Prof. Rasgon's lab was supervised by his postdoctoral researcher Dr. Vanessa Macias. My visit to Prof. Rasgon laboratory was supported by a Peer-to-peer fellowship assigned to me by the Insect Genetic Technologies Research Coordination Network (IGTRCN) and financed by the US National Science Foundation. A laboratory colony of the FPA strain sequenced to produce the AalbF2 genome assembly was established in Prof. Rasgon laboratory and maintained in the insectary conditions described above.

4.3.1 Experimental design

The process of acquisition of TE and nrEVEs sequences within piRNA clusters is not well understood and TEs and nrEVEs are also found outside of piRNA clusters. Given the importance of the piRNA pathway in *Aedes* spp. antiviral response, it is important to understand the mechanisms of piRNA cluster biogenesis and antiviral activity of nrEVEs inserted in clusters. We decided to use CRISPR-Cas9 to create a transgenic line with a modified piRNA cluster to test if 1) canonical mRNA expression can be achieved from piRNA clusters, and if this is possible, under which circumstances it occurs; 2) synthetic piRNAs can be derived from a transgenic construct to exert antiviral activity against cognate viruses.

We took advantage of the completeness of the AalbF2 assembly to design a construct to be inserted into the genome by CRISPR-Cas9-triggered homologous recombination. As previously described, 1441 piRNA clusters were identified in AalbF2 and 53 of these clusters contain one or more nrEVEs. The small-RNA expression profile of piRNA clusters was analyzed using small-RNA sequencing data from FPA mosquitoes ovaries and carcasses (BioProject PRJNA607026) [183]. piRNA cluster 347 was chosen as a good candidate for genome editing because it expresses a high number of unique piRNAs, and it contains the Flavi-nrEVEs Flavi12. Flavi12 is similar to the AeFV E protein and, consistently with being in a cluster, it is the template for several unique piRNAs (**Figure 29**). Furthermore Flavi19, another F-nrEVE, contains a portion derived from the same AeFV viral protein but the sequence harbors polymorphism making it distinguishable by PCR from Flavi12; additionally, Flavi19 is not integrated in a piRNA cluster, rather it maps on an intergenic region of scaffold 20 (NW_021838154.1), and has a reduced piRNAs production (**Figure 29**). These characteristics make Flavi12 and Flavi19 ideal targets for genetic modification to test whether piRNA production depends on the sequence or the position of the DNA fragment within a piRNA cluster.

On these bases a CRISPR-Cas9 construct including a portion of Flavi19 and other viral fragments (**Figure 29C, Appendix 4**) was designed to be inserted into the sequence of Flavi12 in piRNA cluster 347.

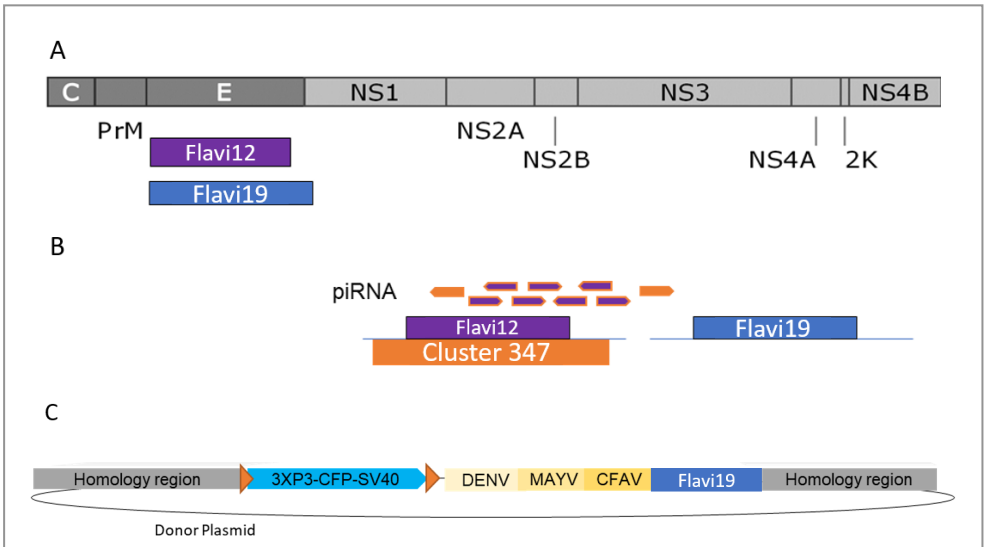


Figure 29. (A) Mapping position of Flavi12 and Flavi19 with respect to the AeFV polyprotein. **(B)** piRNA production from Flavi12. **(C)** simplified structure of the CRISPR-Cas9 construct.

Because I know the nrEVEs landscape of the FPA strain, I was able to design the construct to include fragmented sequences with no sequence similarity to any nrEVEs annotated in FPA. Chimeric viral sequences included in the construct represent fragments from Dengue, Mayaro and Cell Fusing Agent viruses and the coding sequence for the fluorescent marker CFP under the control of the 3XP3 promoter and insulated by gypsy sequences [216] (**Appendix 4**). The fluorescent marker transgene will test whether mRNA (and protein) expression can be achieved within a piRNA cluster if a complete ORF, a functional promoter and gypsy insulators are provided. The inclusion of viral fragments will allow me to test whether synthetic piRNAs can be derived from a transgenic construct to affect mosquito antiviral response with respect to cognate viruses.

4.3.2 Embryo injection

The endonuclease efficiency of the commercially available Cas9 protein supplied by Integrated DNA Technologies (USA) was tested with an *in-vitro* cleavage assay together with a specific gRNA synthesized and a PCR amplicon of the target sequence. Most of the target sequence was cleaved, indicating a successful endonucleasic activity of the Cas9-sgRNA complex. Microinjections were performed using eggs collected from two different generations of FPA mosquitoes through one month. Dark-colored eggs were not used for microinjection as the endochorion darkening via melanization leads to the formation of a resistant serosal cuticle that makes piercing with microneedle difficult [217]. Eggs were injected with a mix of Cas9, sgRNA, construct plasmid and injection buffer as previously described.

In total, ~900 *Ae. albopictus* FPA eggs were injected over three weeks. Approximately 150 eggs hatched, indicating a post-injection survival rate of 17%. After hatching, G0 Pupae were collected and crossed to wild-type mosquitoes. Eggs were collected and hatched to create the first generation (G1) of the transgenic line CFP12-19.

4.3.3 CFP expression assay and knock-in validation

CFP is a cyan variant of the GFP protein derived from the jellyfish *Aequoria victoria*. In the construct the CFP ORF was preceded by the 3XP3 promoter and insulated by gypsy sequences to promote and stabilize transgene expression [216, 218].

More than 200 larvae from the G1 of the CFP12-19 strain were tested for CFP expression using a fluorescent microscope and no visible fluorescence was detected in any individual, suggesting three possible hypothesis: 1) knock-in of the construct was not successful or the plasmid remained as an episomal sequence; 2) the transgene was integrated into G0 but lost in the G1; 3) CFP mRNA is transcriptionally or post-transcriptionally repressed.

Legs were removed from ~150 adults of the G1 and tested for the 3XP3 and CFP ORF by PCR. This portion of the transgene was identified in ~30% of the tested adult individuals (**Figure 30**). This result indicates that the transgene was correctly knocked-in, but CFP expression is repressed. Being the transgene inserted into a piRNA cluster, it is possible that the canonical mRNA expression is inhibited because the transgene is used as template for precursor piRNAs. I will test this hypothesis in the future by performing small-RNA sequencing and RNA sequencing on transgenic mosquito samples.

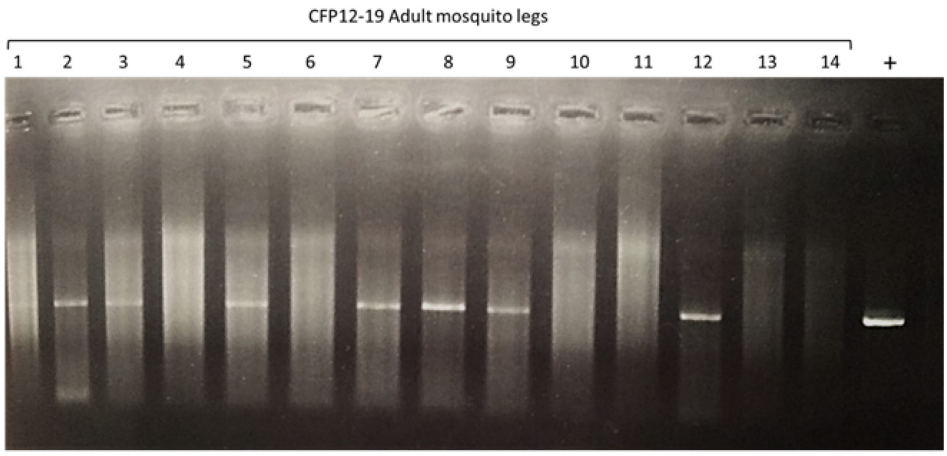


Figure 30. PCR amplification of the 3XP3 and CFP ORFs on transgenic mosquito legs. The last lane shows the positive control for the 970 bp expected amplicon.

Single adult mosquitoes from the G1 were tested by PCR and the individuals carrying the transgenes were selected and intercrossed for three generations until G4. G4 eggs were collected and sent to the University of Pavia. I received the G4 eggs at the end of Sept. 2020 and established a colony in the insectary. I am rearing the mosquitoes to collect material for DNA-seq to verify if the whole construct was properly integrated, RNA-seq and small RNA-seq to check whether the CFP ORF is expressed as mRNAs or piRNAs and whether piRNAs are produced from Flavi19 and the synthetic fragments from DENV, Mayaro virus and CFAV. The ultimate goal is to check for differential production of piRNAs in normal conditions and upon infection with cognate viruses.

5. Discussion

5.1 *De novo* assembly of the *Aedes albopictus* genome

Aedes albopictus has an exceptionally complex genome (filled with repeats, small inversions, etc.) presenting a challenge to any attempt to assemble an animal genome with accuracy. The importance of a quality genome assembly is clear, as this species is the main arboviral vector in Europe, North America, and Asia. Thus, the initial genome assembly of *Ae. albopictus* represented a fundamental achievement. However, due to very high levels of repetitive DNA and reliance solely on short-read sequencing, this initial assembly remains highly fragmented.

I worked with a consortium of scientists and combined a cytofluorimetric approach with PacBio and Hi-C sequencing to generate a new high-quality assembly. The AalbF2 genome assembly and its associated gene set, databases of nrEVEs, miRNAs and piRNA clusters are collective resources that will enable great advances in *Ae. albopictus* biology. All of these resources are available for the scientific community and will be beneficial for any project involved in *Ae. albopictus* genomics, immunology, population genetics, interaction with microorganisms and for genetic manipulation of mosquitoes. AalbF2 has currently been set as the representative genome for *Ae. albopictus* by NCBI (https://www.ncbi.nlm.nih.gov/genome/45).

Additionally, we developed the first physical map of *Ae. albopictus,* which consists of fifty DNA markers that cover the largest genomic scaffolds that alone comprise more than half of the genome. We physically located rDNA, PPO gene clusters, and the largest viral integration in the genome (CanuFlavi19). Overall, FISH data were consistent with the assembled genome, confirming its large-scale structural accuracy. Combining *in situ* and bioinformatic approaches based on synteny between *Ae. albopictus* and *Ae. aegypti*, we anchored to the three *Aedes* chromosomes the 58 larger scaffolds, whose length sum makes 75% of the genome (L75 parameter). Analyses of mitotic chromosomes also showed that the *Ae. albopictus* chromosomes are slightly longer than *Ae. aegypti* ones, which is consistent with cytofluorimetry results.

Small RNA analyses identified 121 miRNAs including 26 novel miRNAs, some of which are strongly induced upon blood-feeding, suggesting important functions for these miRNAs in reproduction and development. piRNA cluster annotation has provided a high confidence set of piRNA clusters. Given the peculiarities of the piRNA pathway in *Aedes* species, this information is crucial to infer the origin of piRNAs in sequenced mosquitoes.

Even more importantly, knowledge of piRNA clusters and piRNAs expression and

features will set the stage for piRNA cluster modification to understand their functions and to explore their exploitation for preventing arbovirus transmission. This idea is strengthened by the enrichment of nrEVE sequences in piRNA clusters which provides support for the hypothesis that they may provide a potential inherited antiviral defense system [98, 112, 113]. Indeed, the improved assembly quality of the genome allowed us to annotate precisely 456 nrEVEs. The majority of the nrEVEs I identified appeared to be derived from unclassified viruses, many of which have recently been identified.

The annotated nrEVEs and their position in the genome can be visualized and downloaded on an online database maintained by our laboratory (http://nreves.com/). They can also be compared to the viral integrations annotated in the *Ae. aegypti* AaegL5 genome using the same pipeline and viral database [214].

Curation of immunity genes revealed an expansion in several immunity-associated genes that suggest a specialized immune system in comparison to other insect species. In addition to the immunity genes, the M locus manual curation and the predicted 26,856 protein-coding sequences will enable insights into the genetics and the molecular pathways of *Ae. albopictus* and provide avenues for novel genetic-based strategies of control, including those for population suppression based on gene-drive systems creating male-biased populations [219]. The genome model, together with datasets for the predicted transcripts and proteins, is publicly available on the NCBI repositories (https://www.ncbi.nlm.nih.gov/assembly/GCF_006496715.1). The developmental transcriptome analysis described here demonstrates that the new genome assembly has produced a significantly more complete gene set with fewer potentially erroneous gene duplications as compared to the previously available genome. The quantification data across developmental timepoints and multiple tissues will provide the community with an invaluable resource for further exploration of *Ae. albopictus* biology.

5.2 The landscape of novel nrEVEs in wild collected *Ae. albopictus* mosquitoes

Eukaryotic genomes contain abundant viral integrations, most of which derive from retroviruses encoding for retrotranscriptases and integrases. It was unexpected to find nonretroviral EVEs in many organisms, including mosquitoes. The mechanism of integration is often unclear. However, viral transcript reverse transcription by host-encoded reverse transcriptase (RT) has been implicated [109].

Previous work conducted in our laboratory revealed the presence of nrEVEs with similarities to flaviviruses and rhabdoviruses in *Ae. albopictus* and *Ae. aegypti* but

not in *Anopheles* species and *Culex quinquefasciatus* [112]. The new genome assembly described in this thesis, coupled with an expanded database of viruses, permitted a more precise and complete annotation of nrEVEs in a reference genome revealing an increased number of viral integrations. It is known that viral integrations are differentially distributed within and across mosquito populations, with a conserved set of evolutionarily old R-nrEVEs [116]. Despite the abundance of nrEVEs in *Ae. albopictus* mosquitoes, the biology, functional role, and pattern of integration in wild mosquito populations are still relatively unexplored.

The AalbF2 genome assembly was generated using genome sequencing data obtained from mosquitoes derived from the Foshan laboratory strain. This line was first collected in 1981 in the Guangdong Province (China), part of the original habitat of *Ae. albopictus* and has been maintained in laboratory ever since. As such, the hypothesis that nrEVEs formation is a dynamic and continuous process shaped by viral exposure was confirmed by the identification of novel nrEVEs in wild samples of *Ae. albopictus* from different geographical regions worldwide. These nrEVEs were found using bioinformatical methods and are absent in the AalbF2 and AaloF1 genome assemblies and in mosquitoes of the Foshan strain, indicating that the integration event may have occurred after the beginning of the global expansion of *Ae. albopictus.*

Using bioinformatics and molecular biology techniques, 12 novel nrEVEs were identified in samples from Italy, Mexico, Thailand, China and La Reunion island. Populations from China and Thailand are native and are characterized by a genetic connectivity which supports their ancestral origin [15]. Populations from La Reunion island are old, the island was colonized in the 18t^{h} century [220]. A more complex and world-wide invasion process occurred in the past 50-60 years, moving *Ae. albopictus* to all continents except Antarctica [208]. Nowadays, while populations of the old-colonized regions have reached a certain genetic stability, the ones of the newly-invaded regions can be considered as an admixture of genetic material from different populations [15]. These new populations include the ones from Mexico and Italy, which were colonized in the last 30 years [14, 15]. Therefore, great inter-population variability is observed among *Ae. albopictus* populations from geographically close countries.

The bioinformatical distribution of the novel nrEVEs supports this data, as population specific novel nrEVEs were rarely observed in *Ae. albopictus* samples. Even though most viral integrations were bioinformatically identified in samples from La Reunion and Mexico, and only few were found in China and Thailand, the frequency distribution of five molecularly validated novel-nrEVEs (nr-EVEnew-1 to 7) was between 4% and 90% in at least one native population (Canton, Chiang Mai) and/or an old-colonized region (St. Pierre in La Reunion).

Three of the five analyzed integrations were detected in the newly colonized regions too (Crema and Tapachula).

This result supports the hypothesis that the new integration events occurred before the recent spread of *Ae. albopictus* worldwide. However, considering the chaotic dispersion of *Ae. albopictus*, it is possible that some of these integrations originated in one of the old-colonizing populations that was globally spread by the human movement.

On average, novel nrEVEs had a higher average of identity in comparison to the annotated nrEVEs. This may suggest more recent integrations events in comparison to most of the nrEVEs found in the AalbF2 assembly. However, the hypothesis that the annotated nrEVEs originated from extinct or unidentified viruses with sequences similar to circulating viral species cannot be excluded. In comparison to the novel nrEVEs found in *Ae. aegypti* [214], those found in *Ae. albopictus* were shorter and they did not show rearrangements in the viral sequence. In most cases it was not possible to identify the full viral integration, probably due to the still fragmented nature of the *Ae. albopictus* current genome assembly and to its high repeats content.

Overall, these results demonstrate the existence of novel viral integrations in wild-collected samples of *Aedes* mosquitoes. The number of novel viral integrations identified support the hypothesis that integration of sequences from non-retroviral RNA viruses is a rare event. Both the new and the annotated integrations were primarily from ISVs.

The metagenomic screening for viral species using a *de novo* assembly approach on small-RNA reads revealed ISVs in wild-collected mosquitoes. My results may be biased by the used sequencing strategy (small RNA-seq), which might have favored the identification of viruses against which mosquitoes mount an immune response. However, this is a useful result, demonstrating that mosquitoes mount an immune response against the same viral species from which viral integrations are derived. This is a further proof that the endogenization of a viral sequence depends on viral exposure and that it may have a utility for mosquito antiviral immunity. The antiviral mechanisms of mosquitoes are usually studied in the context of arbovirus infection, but it is becoming clear that the majority of the viruses infecting mosquitoes are ISVs. Using genome engineering techniques on both viruses and mosquitoes, selected viral integrations were shown to exert antiviral activity with respect to cognate viral infections [107, 221]. It is hypothesized that nrEVEs constitute a memory of past viral infections and interact with the piRNA pathway to confer a form of adaptive immunity in mosquitoes similar to CRISPR-Cas9 in bacteria [101]. If this is the case, ISV-derived nrEVEs would be abundant as a consequence of the exposure of mosquitoes to these viruses.

5.3 Genetic modification of the *Ae. albopictus* genome

This part of my project combined cutting-edge bioinformatics and genome editing techniques to provide new insights into unexplored functions of the piRNA pathway.

Similar to many processes that are involved in pathogen transmission in mosquitoes, the piRNA pathway is sufficiently specialized that the closest model organism, *D. melanogaster,* offers little for extrapolation to mosquitoes [222]. The transgenic mosquito lines that I generated will constitute an important resource to further test the biogenesis and the behavior of piRNA clusters, as well as the antiviral functions of nrEVEs contained in piRNA clusters upon infection. By including promoters and insulators in the constructs that was inserted in the piRNA clusters I will be able to test whether, and under which circumstances, canonical mRNA expression can be achieved from piRNA clusters.

In addition, my experiment was the first knock-in manipulation of *Ae. albopictus* and demonstrated that genetically modifying the genome (and in particular repeat-dense regions like piRNA clusters) of this species is feasible. Genetic manipulation of piRNA clusters can also allow us to overcome the challenges of genes suppression and ablation in the germline where the piRNA pathway is greatly active, but which is mostly inaccessible to RNAi-based injection methods of knockdown and where much of the gene function is necessary for reproduction [223, 224].

Therefore, besides expanding our knowledge on the biogenesis of piRNA clusters in *Aedes* species, mosquito lines and tools generated in this project could pave the way for the development of novel transmission blocking strategies that may be based on manipulation of piRNA clusters and piRNAs. This project will continue even after the end of my PhD. Indeed, I plan to use the transgenic line I created at the Pennsylvania State University to do small-RNA and RNA sequencing experiments to assess the production of mRNAs, piRNAs and other small RNAs from transgenic mosquitoes both in normal conditions and upon infection with arboviruses and other viruses.

References

1. Rudin W, Hecker H. Functional morphology of the midgut of *Aedes aegypti* L. (Insecta, Diptera) during blood digestion. Cell Tissue Res. 1979;200:193–203.

2. Göpfert MC, Briegel H, Robert D. Mosquito hearing: sound-induced antennal vibrations in male and female Aedes aegypti. J Exp Biol. 1999;202 Pt 20:2727–38.

3. Kong XQ, Wu CW. Mosquito proboscis: an elegant biomicroelectromechanical system. Phys Rev E Stat Nonlin Soft Matter Phys. 2010;82 1 Pt 1:11910.

4. Geier M, Sass H, Boeckh J. A search for components in human body odour that attract females of *Aedes aegypti*. Ciba Found Symp. 1996;200:132-138,178-183.

5. Hawley WA. The biology of *Aedes albopictus*. Journal of the American Mosquito Control Association. Supplement. 1988;1:1–39.

6. Kraemer MUG, Reiner RC, Brady OJ, Messina JP, Gilbert M, Pigott DM, et al. Past and future spread of the arbovirus vectors *Aedes aegypti* and *Aedes albopictus*. Nat Microbiol. 2019;4:854–63.

7. Diniz DFA, de Albuquerque CMR, Oliva LO, de Melo-Santos MAV, Ayres CFJ. Diapause and quiescence: dormancy mechanisms that contribute to the geographical expansion of mosquitoes and their evolutionary success. Parasit Vectors. 2017;10:310.

8. Denlinger DL, Armbruster PA. Mosquito Diapause. Annu Rev Entomol. 2014;59:73–93.

9. Gimonneau G, Delatte H, Triboire A, Fontenille D. Influence of Temperature on Immature Development, Survival, Longevity, Fecundity, and Gonotrophic Cycles of *Aedes albopictus*, Vector of Chikungunya and Dengue in the Indian Ocean . J Med Entomol. 2009;46:33–41.

10. Kraemer MUG, Sinka ME, Duda KA, Mylne AQN, Shearer FM, Barker CM, et al. The global distribution of the arbovirus vectors *Aedes aegypti* and *Ae. albopictus*. Elife. 2015;4:e08347–e08347.

11. Adhami J, Reiter P. Introduction and establishment of Aedes (Stegomyia) albopictus skuse (Diptera: Culicidae) in Albania. J Am Mosq Control Assoc. 1998;14:340–3.

12. Hahn MB, Eisen RJ, Eisen L, Boegler KA, Moore CG, McAllister J, et al. Reported Distribution of Aedes (Stegomyia) aegypti and Aedes (Stegomyia) albopictus in the United States, 1995-2016 (Diptera: Culicidae). J Med Entomol. 2016;53:1169–75.

13. Kotsakiozi P, Richardson JB, Pichler V, Favia G, Martins AJ, Urbanelli S, et al. Population genomics of the Asian tiger mosquito, *Aedes albopictus*: insights into the recent worldwide invasion. Ecol Evol. 2017;7:10143–57.

14. Vega-Rúa A, Marconcini M, Madec Y, Manni M, Carraretto D, Gomulski LM, et al. Vector competence of *Aedes albopictus* populations for chikungunya virus is shaped by their demographic history. Commun Biol. 2020;3:326.

15. Manni M, Guglielmino CR, Scolari F, Vega-Rua A, Failloux A-B, Somboon P, et al. Genetic evidence for a worldwide chaotic dispersion pattern of the arbovirus vector, *Aedes albopictus*. PLoS Negl Trop Dis. 2017;11:e0005332.

16. Medlock JM, Vaux AGC, Cull B, Schaffner F, Gillingham E, Pfluger V, et al. Detection of the invasive mosquito species *Aedes albopictus* in southern England. Lancet Infect Dis. 2017;17:140.

17. Weaver SC, Barrett ADT. Transmission cycles, host range, evolution and emergence of arboviral disease. Nat Rev Microbiol. 2004;2:789–801.

18. Cox J, Mota J, Sukupolvi-Petty S, Diamond MS, Rico-Hesse R. Mosquito Bite Delivery of Dengue Virus Enhances Immunogenicity and Pathogenesis in Humanized Mice. J Virol. 2012.

19. Moser LA, Lim P-Y, Styer LM, Kramer LD, Bernard KA. Parameters of Mosquito-Enhanced West Nile Virus Infection. J Virol. 2016.

20. Agarwal A, Parida M, Dash PK. Impact of transmission cycles and vector competence on global expansion and emergence of arboviruses. Reviews in Medical Virology. 2017.

21. Liu J, Swevers L, Kolliopoulou A, Smagghe G. Arboviruses and the challenge to establish systemic and persistent infections in competent mosquito vectors: The interaction with the RNAi mechanism. Frontiers in Physiology. 2019.

22. Black WC, Bennett KE, Norma G, Carolina B, GorrochóteguiEscalante N, BarillasMury C, et al. Flavivirus susceptibility in *Aedes aegypti*. Arch Med Res. 2002;33:379–88.

23. Monteiro VVS, Navegantes-Lima KC, de Lemos AB, da Silva GL, de Souza Gomes R, Reis JF, et al. Aedes-Chikungunya Virus Interaction: Key Role of Vector Midguts Microbiota and Its Saliva in the Host Infection. Front Microbiol. 2019;10:492.

24. Hardy JL, Houk EJ, Kramer LD, Reeves WC. Intrinsic Factors Affecting Vector Competence of Mosquitoes for Arboviruses. Annu Rev Entomol. 1983;28:229–62.

25. Weaver SC, Reisen WK. Present and future arboviral threats. Antiviral Res. 2010;85:328–45.

26. Beerntsen BT, James AA, Christensen BM. Genetics of mosquito vector

competence. Microbiol Mol Biol Rev. 2000;64:115–37.

27. Kramer LD, Ciota AT. Dissecting vectorial capacity for mosquito-borne viruses. Curr Opin Virol. 2015;15:112–8.

28. MacDonald G. Epidemiologic models in studies of vectorborne diseases. Public Heal reports (Washington, DC 1896). 1961;76:753–64.

29. Steinhauer DA, Domingo E, Holland JJ. Lack of evidence for proofreading mechanisms associated with an RNA virus polymerase. Gene. 1992;122:281–8.

30. Coffey LL, Vasilakis N, Brault AC, Powers AM, Tripet F, Weaver SC. Arbovirus evolution in vivo is constrained by host alternation. Proc Natl Acad Sci U S A. 2008;105:6970–5.

31. Forrester NL, Coffey LL, Weaver SC. Arboviral bottlenecks and challenges to maintaining diversity and fitness during mosquito transmission. Viruses. 2014;6:3991–4004.

32. Young PR. Arboviruses: A Family on the Move BT - Dengue and Zika: Control and Antiviral Treatment Strategies. In: Hilgenfeld R, Vasudevan SG, editors. Singapore: Springer Singapore; 2018. p. 1–10.

33. Ciota AT, Kramer LD. Insights into arbovirus evolution and adaptation from experimental studies. Viruses. 2010;2:2594–617.

34. Barrows NJ, Campos RK, Liao K-C, Prasanth KR, Soto-Acosta R, Yeh S-C, et al. Biochemistry and Molecular Biology of Flaviviruses. Chem Rev. 2018;118:4448–82.

35. Bhatt S, Gething PW, Brady OJ, Messina JP, Farlow AW, Moyes CL, et al. The global distribution and burden of dengue. Nature. 2013;496:504–7.

36. World Health Organization. Global Strategy for Dengue Prevention and Control 2012–2020. World Heal Organiszation. 2012;:43.

37. Chen R, Mukhopadhyay S, Merits A, Bolling B, Nasar F, Coffey LL, et al. ICTV virus taxonomy profile: Togaviridae. J Gen Virol. 2018;99:761–2.

38. Vairo F, Mammone A, Lanini S, Nicastri E, Castilletti C, Carletti F, et al. Local transmission of chikungunya in Rome and the Lazio region, Italy. PLoS One. 2018;13:e0208896–e0208896.

39. Elliott RM. Molecular biology of the Bunyaviridae. Journal of General Virology. 1990;71:501–22.

40. Briese T, Calisher CH, Higgs S. Viruses of the family Bunyaviridae: Are all available isolates reassortants? Virology. 2013;446:207–16.

41. Stollar V, Thomas VL. An agent in the *Aedes aegypti* cell line (Peleg) which causes fusion of Aedes albopictus cells. Virology. 1975;64:367–77.

42. Nasar F, Palacios G, Gorchakov R V., Guzman H, Travassos Da Rosa AP, Savji N, et al. Eilat virus, a unique alphavirus with host range restricted to insects by RNA replication. Proc Natl Acad Sci U S A. 2012;109:14622–7.

43. Huhtamo E, Cook S, Moureau G, Uzcátegui NY, Sironen T, Kuivanen S, et al. Novel flaviviruses from mosquitoes: mosquito-specific evolutionary lineages within the phylogenetic group of mosquito-borne flaviviruses. Virology. 2014;464–465:320–9.

44. Tree MO, McKellar DR, Kieft KJ, Watson AM, Ryman KD, Conway MJ. Insect-specific flavivirus infection is restricted by innate immunity in the vertebrate host. Virology. 2016;497:81–91.

45. Nasar F, Gorchakov R V, Tesh RB, Weaver SC. Eilat virus host range restriction is present at multiple levels of the virus life cycle. J Virol. 2015;89:1404–18.

46. Agboli E, Leggewie M, Altinli M, Schnettler E. Mosquito-Specific Viruses-Transmission and Interaction. Viruses. 2019;11.

47. Cook S, Bennett SN, Holmes EC, De Chesse R, Moureau G, de Lamballerie X. Isolation of a new strain of the flavivirus cell fusing agent virus in a natural mosquito population from Puerto Rico. J Gen Virol. 2006;87 Pt 4:735–48.

48. Saiyasombat R, Bolling BG, Brault AC, Bartholomay LC, Blitvich BJ. Evidence of efficient transovarial transmission of Culex flavivirus by *Culex pipiens* (Diptera: Culicidae). J Med Entomol. 2011;48:1031–8.

49. Cook S, Chung BYW, Bass D, Moureau G, Tang S, McAlister E, et al. Novel virus discovery and genome reconstruction from field rna samples reveals highly divergent viruses in dipteran hosts. PLoS One. 2013;8:1–22.

50. Li CX, Shi M, Tian JH, Lin XD, Kang YJ, Chen LJ, et al. Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. Elife. 2015;2015:1–26.

51. Vasilakis N, Forrester NL, Palacios G, Nasar F, Savji N, Rossi SL, et al. Negevirus: a proposed new taxon of insect-specific viruses with wide geographic distribution. J Virol. 2013;87:2475–88.

52. Bolling BG, Weaver SC, Tesh RB, Vasilakis N. Insect-Specific Virus Discovery : Significance for the Arbovirus Community. Viruses. 2015;7 July:4911–28.

53. Halbach R, Junglen S, van Rij RP. Mosquito-specific and mosquito-borne viruses: evolution, infection, and host defense. Current Opinion in Insect Science. 2017;22 May:16–27.

54. Öhlund P, Lundén H, Blomström AL. Insect-specific virus evolution and potential effects on vector competence. Virus Genes. 2019;55:127–37.

55. Plyusnin A, Sironen T. Evolution of hantaviruses: co-speciation with reservoir

hosts for more than 100 MYR. Virus Res. 2014;187:22–6.

56. Olson KE, Bonizzoni M. Nonretroviral integrated RNA viruses in arthropod vectors: an occasional event or something more? Curr Opin Insect Sci. 2017;22:45–53.

57. Schultz MJ, Frydman HM, Connor JH. Dual Insect specific virus infection limits Arbovirus replication in Aedes mosquito cells. Virology. 2018;518:406–13.

58. Reynaud JM, Kim DY, Atasheva S, Rasalouskaya A, White JP, Diamond MS, et al. IFIT1 Differentially Interferes with Translation and Replication of Alphavirus Genomes and Promotes Induction of Type I Interferon. PLoS Pathog. 2015;11:1–32.

59. Saraiva RG, Kang S, Simões ML, Angleró-Rodríguez YI, Dimopoulos G. Mosquito gut antiparasitic and antiviral immunity. Dev Comp Immunol. 2016;64:53–64.

60. Lim HY, Ng ML. A different mode of entry by dengue-2 neutralisation escape mutant virus. Arch Virol. 1999;144:989–95.

61. Kumar A, Srivastava P, Sirisena P, Dubey SK, Kumar R, Shrinet J, et al. Mosquito Innate Immunity. Insects. 2018;9:95.

62. Fu XY, Schindler C, Improta T, Aebersold R, Darnell JE. The proteins of ISGF-3, the interferon alpha-induced transcriptional activator, define a gene family involved in signal transduction. Proc Natl Acad Sci. 1992;89:7840 LP – 7843.

63. Souza-Neto JA, Sim S, Dimopoulos G. An evolutionary conserved function of the JAK-STAT pathway in anti-dengue defense. Proc Natl Acad Sci U S A. 2009;106:17841–6.

64. Angleró-Rodríguez YI, MacLeod HJ, Kang S, Carlson JS, Jupatanakul N, Dimopoulos G. *Aedes aegypti* Molecular Responses to Zika Virus: Modulation of Infection by the Toll and Jak/Stat Immune Pathways and Virus Host Factors. Front Microbiol. 2017;8:2050.

65. Jupatanakul N, Sim S, Angleró-Rodríguez YI, Souza-Neto J, Das S, Poti KE, et al. Engineered *Aedes aegypti* JAK/STAT Pathway-Mediated Immunity to Dengue Virus. PLoS Negl Trop Dis. 2017;11:e0005187–e0005187.

66. Bahia AC, Kubota MS, Tempone AJ, Araújo HRC, Guedes BAM, Orfanó AS, et al. The JAK-STAT pathway controls Plasmodium vivax load in early stages of *Anopheles aquasalis* infection. PLoS Negl Trop Dis. 2011;5:e1317–e1317.

67. Xi Z, Ramirez JL, Dimopoulos G. The *Aedes aegypti* Toll Pathway Controls Dengue Virus Infection. PLOS Pathog. 2008;4:e1000098.

68. Liu T, Xu Y, Wang X, Gu J, Yan G, Chen X-G. Antiviral systems in vector mosquitoes. Dev Comp Immunol. 2018;83:34–43.

69. Garver LS, Bahia AC, Das S, Souza-Neto JA, Shiao J, Dong Y, et al. *Anopheles* Imd pathway factors and effectors in infection intensity-dependent anti-Plasmodium action. PLoS Pathog. 2012;8:e1002737–e1002737.

70. Luplertlop N, Surasombatpattana P, Patramool S, Dumas E, Wasinpiyamongkol L, Saune L, et al. Induction of a peptide with activity against a broad spectrum of pathogens in the *Aedes aegypti* salivary gland, following Infection with Dengue Virus. PLoS Pathog. 2011;7:e1001252–e1001252.

71. Barletta ABF, Nascimento-Silva MCL, Talyuli OAC, Oliveira JHM, Pereira LOR, Oliveira PL, et al. Microbiota activates IMD pathway and limits Sindbis infection in *Aedes aegypti*. Parasit Vectors. 2017;10:103.

72. Leggewie M, Schnettler E. RNAi-mediated antiviral immunity in insects and their possible application. Curr Opin Virol. 2018;32:108–14.

73. Swarts DC, Makarova K, Wang Y, Nakanishi K, Ketting RF, Koonin E V, et al. The evolutionary journey of Argonaute proteins. Nat Struct Mol Biol. 2014;21:743–53.

74. Kim VN, Han J, Siomi MC. Biogenesis of small RNAs in animals. Nat Rev Mol Cell Biol. 2009;10:126–39.

75. Cooper AMW, Silver K, Zhang J, Park Y, Zhu KY. Molecular mechanisms influencing efficiency of RNA interference in insects. Pest Manag Sci. 2019;75:18–28.

76. Bernhardt SA, Simmons MP, Olson KE, Beaty BJ, Blair CD, Black WC. Rapid intraspecific evolution of miRNA and siRNA genes in the mosquito *Aedes aegypti*. PLoS One. 2012;7:e44198–e44198.

77. Campbell CL, Black WC, Hess AM, Foy BD. Comparative genomics of small RNA regulatory pathway components in vector mosquitoes. BMC Genomics. 2008;9:425.

78. Ylla G, Fromm B, Piulachs M-D, Belles X. The microRNA toolkit of insects. Sci Rep. 2016;6:37736.

79. Asgari S. Role of microRNAs in arbovirus/vector interactions. Viruses. 2014;6:3514–34.

80. Karlikow M, Goic B, Saleh M-C. RNAi and antiviral defense in Drosophila: Setting up a systemic immune response. Dev Comp Immunol. 2014;42:85–92.

81. Wu X, Hong H, Yue J, Wu Y, Li X, Jiang L, et al. Inhibitory effect of small interfering RNA on dengue virus replication in mosquito cells. Virol J. 2010;7:270.

82. Vagin V V, Sigova A, Li C, Gvozdev V, Zamore PD. A distinct small RNA pathway silences selfish genetic elements in the germline. Science (80-). 2006;313 July:320–5.

83. Grimson A, Srivastava M, Fahey B, Woodcroft BJ, Chiang HR, King N, et al. Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. Nature. 2008;455:1193–7.

84. Thomson T, Lin H. The Biogenesis and Function of PIWI Proteins and piRNAs: Progress and Prospect. Annu Rev Cell Dev Biol. 2009;25:355–76.

85. Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, et al. Discrete Small RNA-Generating Loci as Master Regulators of Transposon Activity in *Drosophila*. Cell. 2007;128:1089–103.

86. Josse T, Teysset L, Todeschini A-L, Sidor CM, Anxolabéhère D, Ronsseray S. Telomeric trans-silencing: an epigenetic repression combining RNA silencing and heterochromatin formation. PLoS Genet. 2007;3:1633–43.

87. Zanni V, Eymery A, Coiffet M, Zytnicki M, Luyten I, Quesneville H, et al. Distribution, evolution, and diversity of retrotransposons at the flamenco locus reflect the regulatory properties of piRNA clusters. Proc Natl Acad Sci U S A. 2013;110:19842–7.

88. Yamamoto Y, Watanabe T, Hoki Y, Shirane K, Li Y, Ichiiyanagi K, et al. Targeted gene silencing in mouse germ cells by insertion of a homologous DNA into a piRNA generating locus. Genome Res. 2013;23:292–9.

89. Czech B, Hannon GJ. One Loop to Rule Them All: The Ping-Pong Cycle and piRNA-Guided Silencing. Trends Biochem Sci. 2016;41:324–37.

90. Huang X, Tóth KF, Aravin AA. piRNA Biogenesis in Drosophila melanogaster. Trends Genet. 2017;33:882–94.

91. Han BW, Wang W, Li C, Weng Z, Zamore PD. piRNA-guided transposon cleavage initiates Zucchini-dependent, phased piRNA production. Science (80-). 2015;348:817 LP – 821.

92. Barckmann B, El-Barouk M, Pélisson A, Mugat B, Li B, Franckhauser C, et al. The somatic piRNA pathway controls germline transposition over generations. Nucleic Acids Res. 2018;46:9524–36.

93. Tang W, Seth M, Tu S, Shen E-Z, Li Q, Shirayama M, et al. A Sex Chromosome piRNA Promotes Robust Dosage Compensation and Sex Determination in C. elegans. Dev Cell. 2018;44:762-770.e3.

94. Gibson JD, Arechavaleta-Velasco ME, Tsuruda JM, Hunt GJ. Biased Allele Expression and Aggression in Hybrid Honeybees may be Influenced by Inappropriate Nuclear-Cytoplasmic Signaling. Front Genet. 2015;6:343.

95. Liu Y, Dou M, Song X, Dong Y, Liu S, Liu H, et al. The emerging role of the piRNA/piwi complex in cancer. Mol Cancer. 2019;18:123.

96. Wu Q, Luo Y, Lu R, Lau N, Lai EC, Li W-X, et al. Virus discovery by deep

sequencing and assembly of virus-derived small silencing RNAs. Proc Natl Acad Sci. 2010;107:1606 LP – 1611.

97. Petit M, Mongelli V, Frangeul L, Blanc H, Jiggins F, Saleh M-C. piRNA pathway is not required for antiviral defense in Drosophila melanogaster. Proc Natl Acad Sci. 2016;113:E4218–27.

98. Miesen P, Joosten J, van Rij RP. PIWIs Go Viral: Arbovirus-Derived piRNAs in Vector Mosquitoes. PLoS Pathog. 2016;12:e1006017–e1006017.

99. Miesen P, Ivens A, Buck AH, van Rij RP. Small RNA Profiling in Dengue Virus 2-Infected Aedes Mosquito Cells Reveals Viral piRNAs and Novel Host miRNAs. PLoS Negl Trop Dis. 2016;10:1–22.

100. Marconcini M, Hernandez L, Iovino G, Houé V, Valerio F, Palatini U, et al. Polymorphism analyses and protein modelling inform on functional specialization of Piwi clade genes in the arboviral vector *Aedes albopictus*. PLoS Negl Trop Dis. 2019;13:e0007919.

101. Ophinni Y, Palatini U, Hayashi Y, Parrish NF. piRNA-Guided CRISPR-like Immunity in Eukaryotes. Trends Immunol. 2019;40:998–1010.

102. Peterlin BM, Liu P, Wang X, Cary D, Shao W, Leoz M, et al. Hili Inhibits HIV Replication in Activated T Cells. J Virol. 2017;91.

103. Holmes EC. The Evolution of Endogenous Viral Elements. Cell Host Microbe. 2011;10:368–77.

104. Johnson WE. Origins and evolutionary consequences of ancient endogenous retroviruses. Nat Rev Microbiol. 2019;17:355–70.

105. Flegel TW. Research progress on viral accommodation 2009 to 2019. Dev Comp Immunol. 2020;112:103771.

106. Goic B, Stapleford KA, Frangeul L, Doucet AJ, Gausson V, Blanc H, et al. Virus-derived DNA drives mosquito vector tolerance to arboviral infection. Nat Commun. 2016;7:12410.

107. Tassetto M, Kunitomi M, Whitfield ZJ, Dolan PT, Sánchez-Vargas I, Garcia-Knight M, et al. Control of RNA viruses in mosquito cells through the acquisition of vDNA and endogenous viral elements. Elife. 2019;8.

108. Crochu S, Cook S, Attoui H, Charrel RN, De Chesse R, Belhouchet M, et al. Sequences of flavivirus-related RNA viruses persist in DNA form integrated in the genome of Aedes spp. mosquitoes. J Gen Virol. 2004;85:1971–80.

109. Blair CD, Olson KE, Bonizzoni M. The Widespread Occurrence and Potential Biological Roles of Endogenous Viral Elements in Insect Genomes. Curr Issues Mol Biol. 2020;34:13–30.

110. Fujino K, Horie M, Honda T, Merriman DK, Tomonaga K. Inhibition of Borna disease virus replication by an endogenous bornavirus-like element in the ground squirrel genome. Proc Natl Acad Sci U S A. 2014;111:13175–80.

111. ter Horst AM, Nigg JC, Dekker FM, Falk BW. Endogenous Viral Elements Are Widespread in Arthropod Genomes and Commonly Give Rise to PIWI-Interacting RNAs. J Virol. 2019;93:e02124-18.

112. Palatini U, Miesen P, Carballar-Lejarazu R, Ometto L, Rizzo E, Tu Z, et al. Comparative genomics shows that viral integrations are abundant and express piRNAs in the arboviral vectors *Aedes aegypti* and *Aedes albopictus*. BMC Genomics. 2017;18:1–15.

113. Whitfield ZJ, Dolan PT, Kunitomi M, Tassetto M, Seetin MG, Oh S, et al. The Diversity, Structure, and Function of Heritable Adaptive Immunity Sequences in the *Aedes aegypti* Genome. Curr Biol. 2017;27:3511-3519.e7.

114. Suzuki Y, Frangeul L, Dickson LB, Blanc H, Verdier Y, Vinh J, et al. Uncovering the Repertoire of Endogenous Flaviviral Elements in Aedes Mosquito Genomes. J Virol. 2017;91:e00571-17.

115. Chen XG, Jiang X, Gu J, Xu M, Wu Y, Deng Y, et al. Genome sequence of the Asian tiger mosquito, *aedes albopictus*, reveals insights into its biology, genetics, and evolution. Proc Natl Acad Sci U S A. 2015;112:E5907–15.

116. Pischedda E, Scolari F, Valerio F, Carballar-Lejarazu R, Catapano PL, Waterhouse RM, et al. Insights Into an Unexplored Component of the Mosquito Repeatome: Distribution and Variability of Viral Sequences Integrated Into the Genome of the Arboviral Vector *Aedes albopictus*. Front Genet. 2019;10:93.

117. Geuking MB, Weber J, Dewannieux M, Gorelik E, Heidmann T, Hengartner H, et al. Recombination of retrotransposon and exogenous RNA virus results in nonretroviral cDNA integration. Science. 2009;323:393–6.

118. Lazareva E, Lezzhov A, Vassetzky N, Solovyev A, Morozov S. Acquisition of Full-Length Viral Helicase Domains by Insect Retrotransposon-Encoded Polypeptides . Frontiers in Microbiology . 2015;6:1447.

119. Dritsou V, Topalis P, Windbichler N, Simoni A, Hall A, Lawson D, et al. A draft genome sequence of an invasive mosquito: an Italian *Aedes albopictus*. Pathog Glob Health. 2015;109:207–20.

120. Johnston JS, Bernardini A, Hjelmen CE. Genome Size Estimation and Quantitative Cytogenetics in Insects. Methods Mol Biol. 2019;1858:15–26.

121. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017;27:722–36.

122. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.

123. Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, et al. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. Mol Biol Evol. 2017;35:543–8.

124. Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. Versatile genome assembly evaluation with QUAST-LG. Bioinformatics. 2018;34:i142–50.

125. Boratyn GM, Thierry-Mieg J, Thierry-Mieg D, Busby B, Madden TL. Magic-BLAST, an accurate RNA-seq aligner for long and short reads. BMC Bioinformatics. 2019;20:405.

126. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37:907–15.

127. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.

128. Barnett DW, Garrison EK, Quinlan AR, Strömberg MP, Marth GT. BamTools: a C++ API and toolkit for analyzing and managing BAM files. Bioinformatics. 2011;27:1691–2.

129. Miller JR, Koren S, Dilley KA, Puri V, Brown DM, Harkins DM, et al. Analysis of the *Aedes albopictus* C6/36 genome provides insight into cell line utility for viral propagation. Gigascience. 2018;7.

130. Timoshevskiy VA, Sharma A, Sharakhov I V, Sharakhova M V. Fluorescent in situ hybridization on mitotic chromosomes of mosquitoes. J Vis Exp. 2012;:e4215–e4215.

131. Timoshevskiy VA, Severson DW, DeBruyn BS, Black WC, Sharakhov I V, Sharakhova M V. An Integrated Linkage, Chromosome, and Genome Map for the Yellow Fever Mosquito *Aedes aegypti*. PLoS Negl Trop Dis. 2013;7:e2052.

132. Sharakhova M V, Artemov GN, Timoshevskiy VA, Sharakhov I V. Physical Genome Mapping Using Fluorescence In Situ Hybridization with Mosquito Chromosomes. Methods Mol Biol. 2019;1858:177–94.

133. Sharakhova M V, Timoshevskiy VA, Yang F, Demin SI, Severson DW, Sharakhov I V. Imaginal Discs – A New Source of Chromosomes for Genome Mapping of the Yellow Fever Mosquito *Aedes aegypti*. PLoS Negl Trop Dis. 2011;5:e1335.

134. Jabeen R, Iftikhar T, Mengal T, Khattak M. A comparative chromosomal count and morphological karyotyping of three indigenous cultivars of Kalongi (Nigella sativa L.). Pakistan J Bot. 2012;44:1007–12.

135. Cabanettes F, Klopp C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. PeerJ. 2018;6:e4958–e4958.

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

137. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10:R25.

138. Halbach R, Miesen P, Joosten J, Taşköprü E, Pennings B, Vogels CBF, et al. An ancient satellite repeat controls gene expression and embryonic development in *Aedes aegypti* through a highly conserved piRNA. bioRxiv. 2020;:2020.01.15.907428.

139. Friedländer MR, Mackowiak SD, Li N, Chen W, Rajewsky N. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Res. 2011;40:37–52.

140. Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. Nucleic Acids Res. 2018;47:D155–62.

141. Matthews BJ, Dudchenko O, Kingan SB, Koren S, Antoshechkin I, Crawford JE, et al. Improved reference genome of *Aedes aegypti* informs arbovirus vector control. Nature. 2018;563:501–7.

142. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.

143. Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Čech M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res. 2018;46:W537–44.

144. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.

145. Morgulis A, Gertz EM, Schäffer AA, Agarwala R. WindowMasker: window-based masker for sequenced genomes. Bioinformatics. 2005;22:134–41.

146. Kapustin Y, Souvorov A, Tatusova T, Lipman D. Splign: algorithms for computing spliced alignments with identification of paralogs. Biol Direct. 2008;3:20.

147. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997;25:955–64.

148. Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, et al. Rfam 12.0: updates to the RNA families database. Nucleic Acids Res. 2015;43 Database issue:D130–7.

149. Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, et al. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res.

2008;36 suppl_2:W465–9.

150. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–7.

151. Talavera G, Castresana J. Improvement of Phylogenies after Removing Divergent and Ambiguously Aligned Blocks from Protein Sequence Alignments. Syst Biol. 2007;56:564–77.

152. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Syst Biol. 2010;59:307–21.

153. Chevenet F, Brun C, Bañuls A-L, Jacq B, Christen R. TreeDyn: towards dynamic graphics and annotations for analyses of trees. BMC Bioinformatics. 2006;7:439.

154. Xu L, Dong Z, Fang L, Luo Y, Wei Z, Guo H, et al. OrthoVenn2: a web server for whole-genome comparison and annotation of orthologous clusters across multiple species. Nucleic Acids Res. 2019;47:W52–8.

155. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 2015;16:157.

156. Hall AB, Qi Y, Timoshevskiy V, Sharakhova M V, Sharakhov I V, Tu Z. Six novel Y chromosome genes in *Anopheles* mosquitoes discovered by independently sequencing males and females. BMC Genomics. 2013;14:273.

157. Hall AB, Timoshevskiy VA, Sharakhova M V, Jiang X, Basu S, Anderson MAE, 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. 2014;6:179–91.

158. Dereeper A, Audic S, Claverie J-M, Blanc G. BLAST-EXPLORER helps you building datasets for phylogenetic analysis. BMC Evol Biol. 2010;10:8.

159. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.

160. DePristo MA, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43:491–8.

161. Fumagalli M, Vieira FG, Linderoth T, Nielsen R. ngsTools: methods for population genetics analyses from next-generation sequencing data. Bioinformatics. 2014;30:1486–7.

162. Korneliussen TS, Albrechtsen A, Nielsen R. ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics. 2014;15:356.

163. Fox EA, Wright AE, Fumagalli M, Vieira FG. ngsLD: evaluating linkage disequilibrium using genotype likelihoods. Bioinformatics. 2019;35:3855–6.

164. Weir BS, Cockerham CC. Estimating F-Statistics for the Analysis of Population Structure. Evolution (N Y). 1984;38:1358–70.

165. Wickham H. ggplot2. 2nd edition. Cham: Springer International Publishing; 2016.

166. Phanstiel DH, Boyle AP, Araya CL, Snyder MP. Sushi.R: flexible, quantitative and integrative genomic visualizations for publication-quality multi-panel figures. Bioinformatics. 2014;30:2808–10.

167. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7.

168. Gamez S, Antoshechkin I, Mendez-Sanchez SC, Akbari OS. The Developmental Transcriptome of *Ae. albopictus*, a Major Worldwide Human Disease Vector. G3 Genes|Genomes|Genetics. 2020;:g3.401006.2019.

169. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2012;29:15–21.

170. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2013;30:923–30.

171. Santé Publique France. Surveillance de la dengue à la Réunion. Point épidémiologique au 19 mars 2019. 2019;2019:2018–9.

172. Forster M, Szymczak S, Ellinghaus D, Hemmrich G, Rühlemann M, Kraemer L, et al. Vy-PER: eliminating false positive detection of virus integration events in next generation sequencing data. Sci Rep. 2015;5:11534.

173. Pischedda E, Crava C, Carlassara M, Gasmi L, Bonizzoni M. ViR: a tool to account for intrasample variability in the detection of viral integrations. bioRxiv. 2020;:2020.06.16.155119.

174. Mayer C. Phobos 3.3.11. 2010.

175. Kent JK. BLAT—The BLAST-Like Alignment Tool. Genome Res. 2002;12:656–64.

176. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29:644–52.

177. Tsuji J, Weng Z. DNApi : A De Novo Adapter Prediction Algorithm for Small RNA Sequencing Data. 2016;:1–10.

178. Martin M. Cutadapt removes adapter sequences from high-throughput

sequencing reads. EMBnet.journal; Vol 17, No 1 Next Gener Seq Data Anal. 2011.

179. Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics. 2012;28:1086–92.

180. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22:1658–9.

181. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12:59–60.

182. Huson DH, Beier S, Flade I, Górska A, El-Hadidi M, Mitra S, et al. MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. PLOS Comput Biol. 2016;12:e1004957.

183. Palatini U, Masri RA, Cosme L V, Koren S, Thibaud-Nissen F, Biedler JK, et al. Improved reference genome of the arboviral vector *Aedes albopictus*. Genome Biol. 2020;21:215.

184. Roach MJ, Schmidt SA, Borneman AR. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics. 2018;19:460.

185. Ghurye J, Rhie A, Walenz BP, Schmitt A, Selvaraj S, Pop M, et al. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. PLOS Comput Biol. 2019;15:e1007273.

186. Gale K, Crampton J. The ribosomal genes of the mosquito, *Aedes aegypti*. Eur J Biochem. 1989;185:311–7.

187. Xu J, Su X, Bonizzoni M, Zhong D, Li Y, Zhou G, et al. Comparative transcriptome analysis and RNA interference reveal CYP6A8 and SNPs related to pyrethroid resistance in *Aedes albopictus*. PLoS Negl Trop Dis. 2018;12:e0006828.

188. Moureau G, Cook S, Lemey P, Nougairede A, Forrester NL, Khasnatinov M, et al. New Insights into Flavivirus Evolution, Taxonomy and Biogeographic History, Extended by Analysis of Canonical and Alternative Coding Sequences. PLoS One. 2015;10:e0117849.

189. Liu P, Dong Y, Gu J, Puthiyakunnon S, Wu Y, Chen X. Developmental piRNA profiles of the invasive vector mosquito *Aedes albopictus*. Parasit Vectors. 2016;:1–15.

190. Lau NC, Seto AG, Kim J, Kuramochi-Miyagawa S, Nakano T, Bartel DP, et al. Characterization of the piRNA Complex from Rat Testes. Science (80-). 2006;313:363 LP – 367.

191. Lewis SH, Quarles KA, Yang Y, Tanguy M, Frézal L, Smith SA, et al. Pan-arthropod analysis reveals somatic piRNAs as an ancestral defence against

transposable elements. Nat Ecol Evol. 2018;2:174–81.

192. Arensburger P, Hice RH, Wright JA, Craig NL, Atkinson PW. The mosquito *Aedes aegypti* has a large genome size and high transposable element load but contains a low proportion of transposon-specific piRNAs. BMC Genomics. 2011;12:606.

193. Hu W, Criscione F, Liang S, Tu Z. MicroRNAs of two medically important mosquito species: *Aedes aegypti* and *Anopheles stephensi*. Insect Mol Biol. 2015;24:240–52.

194. Feng X, Zhou S, Wang J, Hu W. microRNA profiles and functions in mosquitoes. PLoS Negl Trop Dis. 2018;12:e0006463.

195. Akbari OS, Antoshechkin I, Amrhein H, Williams B, Diloreto R, Sandler J, et al. The Developmental Transcriptome of the Mosquito *Aedes aegypti* , an Invasive Species and Major Arbovirus Vector. G3: Genes|Genomes|Genetics. 2013;3:1493–509.

196. Zhang X, Aksoy E, Girke T, Raikhel AS, Karginov F V. Transcriptome-wide microRNA and target dynamics in the fat body during the gonadotrophic cycle of *Aedes aegypti*. Proc Natl Acad Sci. 2017;114:E1895 LP-E1903.

197. Jain S, Rana V, Shrinet J, Sharma A, Tridibes A, Sunil S, et al. Blood feeding and Plasmodium infection alters the miRNome of *Anopheles stephensi*. PLoS One. 2014;9:e98402.

198. Hussain M, Walker T, O'Neill SL, Asgari S. Blood meal induced microRNA regulates development and immune associated genes in the Dengue mosquito vector, *Aedes aegypti*. Insect Biochem Mol Biol. 2013;43:146–52.

199. Waterhouse RM, Kriventseva E V, Meister S, Xi Z, Alvarez KS, Bartholomay LC, et al. Evolutionary Dynamics of Immune-Related Genes and Pathways in Disease-Vector Mosquitoes. Science (80-). 2007;316:1738 LP – 1743.

200. Zou Z, Shin SW, Alvarez KS, Kokoza V, Raikhel AS. Distinct Melanization Pathways in the Mosquito *Aedes aegypti*. Immunity. 2010;32:41–53.

201. Myllymäki H, Valanne S, Rämet M. The Drosophila Imd Signaling Pathway. J Immunol. 2014;192:3455 LP – 3462.

202. Nakhleh J, Christophides GK, Osta MA. The serine protease homolog CLIPA14 modulates the intensity of the immune response in the mosquito *Anopheles gambiae*. J Biol Chem. 2017;292:18217–26.

203. Dudzic JP, Kondo S, Ueda R, Bergman CM, Lemaitre B. *Drosophila* innate immunity: regional and functional specialization of prophenoloxidases. BMC Biol. 2015;13:81.

204. Xia X, Yu L, Xue M, Yu X, Vasseur L, Gurr GM, et al. Genome-wide

characterization and expression profiling of immune genes in the diamondback moth, *Plutella xylostella* (L.). Sci Rep. 2015;5:9877.

205. Hall AB, Basu S, Jiang X, Qi Y, Timoshevskiy VA, Biedler JK, et al. A male-determining factor in the mosquito *Aedes aegypti*. Science (80-). 2015;348:1268 LP – 1270.

206. Gomulski LM, Mariconti M, Di Cosimo A, Scolari F, Manni M, Savini G, et al. The Nix locus on the male-specific homologue of chromosome 1 in *Aedes albopictus* is a strong candidate for a male-determining factor. Parasit Vectors. 2018;11:647.

207. Aryan A, Anderson MAE, Biedler JK, Qi Y, Overcash JM, Naumenko AN, et al. Nix alone is sufficient to convert female *Aedes aegypti* into fertile males and myo-sex is needed for male flight. Proc Natl Acad Sci U S A. 2020;117:17702–9.

208. Bonizzoni M, Gasperi G, Chen X, James AA. The invasive mosquito species *Aedes albopictus*: Current knowledge and future perspectives. Trends in Parasitology. 2013;29:460–8.

209. Powell JR, Gloria-Soria A, Kotsakiozi P. Recent History of *Aedes aegypti*: Vector Genomics and Epidemiology Records. Bioscience. 2018;68:854–60.

210. Pichler V, Kotsakiozi P, Caputo B, Serini P, Caccone A, della Torre A. Complex interplay of evolutionary forces shaping population genomic structure of invasive *Aedes albopictus* in southern Europe. PLoS Negl Trop Dis. 2019;13:e0007554.

211. Rannala B, Reeve JP. Joint Bayesian estimation of mutation location and age using linkage disequilibrium. Pac Symp Biocomput. 2003;:526–34.

212. Toomajian C, Ajioka RS, Jorde LB, Kushner JP, Kreitman M. A Method for Detecting Recent Selection in the Human Genome From Allele Age Estimates. Genetics. 2003;165:287 LP – 297.

213. Chen X, Kost J, Li D. Comprehensive comparative analysis of methods and software for identifying viral integrations. Brief Bioinform. 2019;20:2088–97.

214. Crava C, Varghese FS, Pischedda E, Halbach R, Palatini U, Marconcini M, et al. Immunity to infections in arboviral vectors by integrated viral sequences: an evolutionary perspective. bioRxiv. 2020;:2020.04.02.022509.

215. Shi C, Liu Y, Hu X, Xiong J, Zhang B, Yuan Z. A Metagenomic Survey of Viral Abundance and Diversity in Mosquitoes from Hubei Province. PLoS One. 2015;10:e0129845.

216. Gdula DA, Gerasimova TI, Corces VG. Genetic and molecular analysis of the gypsy chromatin insulator of Drosophila. Proc Natl Acad Sci. 1996;93:9378–83.

217. Farnesi LC, Vargas HCM, Valle D, Rezende GL. Darker eggs of mosquitoes resist more to dry conditions: Melanin enhances serosal cuticle contribution in egg resistance to desiccation in *Aedes*, *Anopheles* and *Culex* vectors. PLoS Negl Trop

Dis. 2017;11:e0006063.

218. Kokoza V, Ahmed A, Wimmer EA, Raikhel AS. Efficient transformation of the yellow fever mosquito *Aedes aegypti* using the piggyBac transposable element vector pBac[3xP3-EGFP afm]. Insect Biochem Mol Biol. 2001;31:1137–43.

219. Galizi R, Doyle LA, Menichelli M, Bernardini F, Deredec A, Burt A, et al. A synthetic sex ratio distortion system for the control of the human malaria mosquito. Nat Commun. 2014;5:3977.

220. Paupy C, Girod R, Salvan M, Rodhain F, Ailloux A. Population structure of *Aedes albopictus* from La Réunion Island (Indian Ocean) with respect to susceptibility to a dengue virus. Heredity (Edinb). 2001;87:273–83.

221. Suzuki Y, Baidaliuk A, Miesen P, Frangeul L, Crist AB, Merkling SH, et al. Non-retroviral endogenous viral element limits cognate virus replication in *Aedes aegypti* ovaries. bioRxiv. 2020;:2020.03.28.013441.

222. Nouzova M, Clifton ME, Noriega FG. Mosquito adaptations to hematophagia impact pathogen transmission. Curr Opin insect Sci. 2019;34:21–6.

223. Huvenne H, Smagghe G. Mechanisms of dsRNA uptake in insects and potential of RNAi for pest control: A review. J Insect Physiol. 2010;56:227–35.

224. Balakrishna Pillai A, Nagarajan U, Mitra A, Krishnan U, Rajendran S, Hoti SL, et al. RNA interference in mosquito: understanding immune responses, double-stranded RNA delivery systems and potential applications in vector control. Insect Mol Biol. 2017;26:127–39.

Appendixes

Appendix 1

Transcript positions in the *Ae. albopictus* and *Ae. aegypti* (AaegL5) genome assemblies and in the chromosomes of *Ae. albopictus*. Transcripts are indicated by accession number.

Scaffold ID (AalbF2)	Transcript (C6/36 assembly)	Scaff.	Start position	Location	Location in *Ae. aegypti*	Status with respect to *Ae. aegypti*
NW_021837045.1	XM_019703761.1	1	141,912	3p34	3p:24,232,751	Cons.
NW_021837045.1	XM_019687494.1	1	78,206,953	3p44	3p:67,490,043	Cons
NW_021837045.1	XM_019689702.1	1	102,316,324	3p34	3p:89,374,076	Cons
NW_021837045.1	XM_019684861.1	1	115,572,240	3p32	3p:94,611,521	Cons
NW_021837045.1	XM_019682463.1	1	129,449,579	3p32	3p:102,088,493	Cons
NW_021837045.1	XM_019690818.1	1	132,846,558	3p32	3p:104,272,700	Cons
NW_021838153.1	XM_019691989.1	2	8,394,996	2p35	2p:93,350,227	Cons
NW_021838153.1	XM_019675272.2	2	32,875,176	2p32	2p:76,267,897	Cons
NW_021838153.1	XM_019682530.1	2	129,936,557	2q44	2q:422,726,667	Cons
NW_021838465.1	XM_019704755.1	3	18,369,974	3q31	3q:325,406,225	Cons
NW_021838465.1	XM_019701500.1	3	114,597,945	3q43	3q:392,144,268	Cons
NW_021838576.1	XM_019675405.1	4	15,276,711	2p22	2p:126,200,520	Inv.
NW_021838576.1	XM_019673127.1	4	24,157,064	2p25	2p:121,906,299	Cons
NW_021838576.1	XM_019677471.1	4	24,157,064	2p25	2p:121,906,299	Cons
NW_021838576.1	XM_019681895.1	4	53,051,265	2p12	2p:198,995,083	Cons
NW_021838576.1	XM_019698846.1	4	68,135,981	2p12	2p:203,995,757	Cons
NW_021838576.1	XM_019698741.1	4	77,060,065	2p12	2q:210,256,925	Cons
NW_021838687.1	XM_019703202.1	5	2,060,636	2q34	2q:387,886,655	Cons
NW_021838687.1	XM_019694026.1	5	21,213,967	2q33	2q:378,098,440	Cons
NW_021838687.1	XM_019694057.1	5	23,813,996	2q33	2q:378,806,902	Cons
NW_021838687.1	XM_019694256.1	5	56,553,773	2q31	2q:360,884,653	Cons

NW_021838798.1	XM_019696654.1	6	28,914,960	1p25	1p:73,017,099	Cons
NW_021838798.1	XM_019706593.1	6	69,371,979	1p34	1p:27,332,552	Cons
NW_021838909.1	XM_019676257.1	7	81,139,711	3q14	3q:255,525,249	Cons
NW_021838909.1	XM_019705822.1	7	8,146,758	1q21	1q:203,878,017	Cons
NW_021839020.1	XM_019670988.1	8	23,795,330	2q24	2q:316,771,716	Cons
NW_021839020.1	XM_019695499.1	8	60,172,454	2q24	2q:336,171,722	Cons
NW_021839130.1	XM_019686202.1	9	16,379,504	1q31	1q:242,666,370	Cons
NW_021839130.1	XM_019686203.1	9	16,379,504	1q31	1q:242666370	Cons
NW_021839130.1	XM_019685481.1	9	58,786,903	1q33	1q:223,480,890	Cons
NW_021837267.1	XM_019704970.1	12	109,742	3p11	3p:178,234,261	Cons
NW_021837267.1	XM_019674355.1	12	4,413,385	3p13	3p:160,039,183	Cons
NW_021837267.1	XM_019677377.1	12	47,026,645	3p12	3p:189,773,252	Cons
NW_021837378.1	XM_019674838.1	13	20,304,874	2q46	2q:439,707,785	Cons
NW_021837489.1	XM_019699815.1	14	489,131	3q23	3q:313,741,920	Cons
NW_021837489.1	XM_019707152.1	14	49,492,391	3q23	3q:286,848,677	Cons
NW_021837600.1	XM_019691051.1	15	49,408,719	2p22	2p:147,096,234	Cons
NW_021837711.1	XM_019698166.1	16	2,841,814	1q12	1q:171,186,806	Cons
NW_021837822.1	XM_019698650.1	17	27,654,978	1q44	1q:292,319,975	Cons
NW_021837931.1	XM_019682829.1	18	20,015,332	3q11	3q:225,565,535	Cons
NW_021838154.1	XM_019699074.1	20	2,199,384	3p21	3p:142,620,468	Cons
NW_021838154.1	XM_019697550.1	20	22,883,182	3p14	3p:151,341,908	Cons
NW_021838233.1	XM_019703990.1	21	12,395,103	2q12, CM1,3	2q: 238,557,210	Cons
NW_021838233.1	XM_019703499.1	21	15,046,841	2q11	2q:241,488,879	Cons
NW_021838665.1	XM_020077126.1	48	3,448,080	2p24	2p:143,591,914	Cons
NW_021838743.1	XM_019702030.1	55	2,960,159	3q34	3q:362,330,504	Cons
NW_021838832.1	XM_019675517.1	63	1,551,340	3p13	3p:156,723,676	Cons
NW_021839153.1	XM_019696917.1	92	864,056	3q13	3q:234,093,089	Cons
NW_021837578.1	XM_019691611.1	148	98,118	2q26	2q:344,922,562	Cons

Appendix 2

Bioinformatically mapping the first 58 scaffolds (L75) to the chromosomes of *Aedes albopictus* using mapping alignments to *Aedes aegypti* chromosomes. Bioinformatic-based and in situ mapped scaffolds are shown.

Scaffold ID	Scaffold Number	FISH-mapped chromosome	Dgenuies mapped Chr.	Scaffold Len	Q-start	Q-stop	T-len	T-start	T-stop
NW_021837045.1	1	3p32	3	196395033	4836	196394362	409777670	818751	409343898
NW_021838153.1	**2**	2p32	2	168827982	97908	168459680	474425716	31540	472179123
NW_021838465.1	3	3q31	3	135305655	75172	135305593	409777670	282374	409231705
NW_021838576.1	4	2p12	2	122869687	65214	122845322	474425716	709956	472518817
NW_021838687.1	5	2q31	2	99254364	7308	99184464	474425716	567355	472224827
NW_021838798.1	6	1p25	1	95072813	14066	95026473	310827022	740768	309548875
NW_021838909.1	7	3q14**	3	94263231	94841	94255626	409777670	1163769	408456253
NW_021839020.1	8	2q24	2	82511891	8220	82459432	474425716	188809	472227243
NW_021839130.1	9	1q31	1	65883261	50050	65882908	310827022	1159548	309433748
NW_021837046.1	10		2	63746563	2478	63720061	474425716	709956	472224467
NW_021837156.1	11		1	62838808	42717	62811988	310827022	2319182	310163075
NW_021837267.1	12	3p11	3	58853413	33637	58828551	409777670	930522	408972797
NW_021837378.1	13	2q46	2	55702539	29171	55692883	474425716	2001175	474378467
NW_021837489.1	14	3q23	3	52942089	77635	52901422	409777670	235555	409232253
NW_021837600.1	15	2p22	2	51173165	109619	51172994	474425716	1572995	472518811
NW_021837711.1	16	1q12	1	45635565	10781	45246087	310827022	2318946	310009407

NW_021837822.1	17	1q44	1	39427437	36112	39206771	310827022	2426604	310163813
NW_021837931.1	18	3q11	3	39279557	45612	39141352	409777670	1508614	408454682
NW_021838042.1	19		2	27194535	30173	27178416	474425716	2994068	472227746
NW_021838154.1	20	3p14	3	24405320	67845	24370121	409777670	1914801	407228286
NW_021838233.1	21	2q11	2	21999845	79	21990515	474425716	2937402	472178394
NW_021838343.1	22		2	21387314	549	21371175	474425716	1506814	472227710
NW_021838388.1	23		2	15820435	5307	15769690	474425716	2937575	472222761
NW_021838399.1	24		2	14680195	484	14678688	474425716	6939110	472518817
NW_021838410.1	25		1	12700773	40729	12687724	310827022	1782533	309430400
NW_021838421.1	26		1	11967382	13254	11790126	310827022	3841683	309431875
NW_021838432.1	27		3	11907018	114409	11812395	409777670	1508614	395666574
NW_021838443.1	28		2	10504575	41351	10468552	474425716	13387261	472518817
NW_021838454.1	29		2	10455035	4631	10406668	474425716	13387261	467193453
NW_021838466.1	30		2	10003331	24750	9997291	474425716	1407962	470318076
NW_021838477.1	31		2	8923808	3285	8794727	474425716	197573	472177009
NW_021838488.1	32		3	8738692	2196	8738198	409777670	7429763	409232253
NW_021838499.1	33		1	8696722	112	8474377	310827022	23506573	310061255
NW_021838510.1	34		3	8296403	28976	8271757	409777670	44516148	409232286
NW_021838521.1	35		1	8124246	43774	8103026	310827022	10519127	309421208
NW_021838532.1	36		2	7801424	88675	7709837	474425716	2994046	464470419
NW_021838543.1	37		1	7739647	6875	7688847	310827022	6436673	300445293
NW_021838554.1	38		3	7645035	631	7644566	409777670	1915529	407252759

NW_021838565.1	39		1	7506809	204	7443195	310827022	15938023	307812612
NW_021838577.1	40		3	7114296	197923	7010106	409777670	1914801	408454655
NW_021838588.1	41		2	7078098	44447	6848836	474425716	3271702	472040095
NW_021838599.1	42		3	6377459	61750	6157821	409777670	1508614	397802913
NW_021838610.1	43		1	6209708	170821	6160254	310827022	5576272	309430965
NW_021838621.1	44		1	6075262	72854	5967581	310827022	2424055	306235135
NW_021838632.1	45		1	6042954	25107	5988754	310827022	4643554	309430216
NW_021838643.1	46		2	6008651	7869	5960461	474425716	16012949	472224467
NW_021838654.1	47		3	5997344	9512	5931893	409777670	7649196	407228286
NW_021838665.1	48	2p24	2	5928163	30231	5921784	474425716	17069773	472224805
NW_021838676.1	49		1	5616786	4674	5595548	310827022	13708366	309430965
NW_021838688.1	50		1	5603677	11	5565027	310827022	453107	305074731
NW_021838699.1	51		1	5568893	103110	5483763	310827022	13711300	307812681
NW_021838710.1	52		1	5428130	61260	5279199	310827022	17362205	304230677
NW_021838721.1	53		2	4868804	19270	4821996	474425716	12717555	468601915
NW_021838732.1	54		1	4850854	1455	4768656	310827022	16239465	306128849
NW_021838743.1	55	3q34	1	4826477	29747	4762033	310827022	14142692	307812680
NW_021838754.1	56		1	4637188	35013	4606591	310827022	15936991	307490695
NW_021838765.1	57		2	4606087	17708	4515852	474425716	188809	472177152
NW_021838776.1	58		2	4348337	12679	4343083	474425716	5578639	466433958

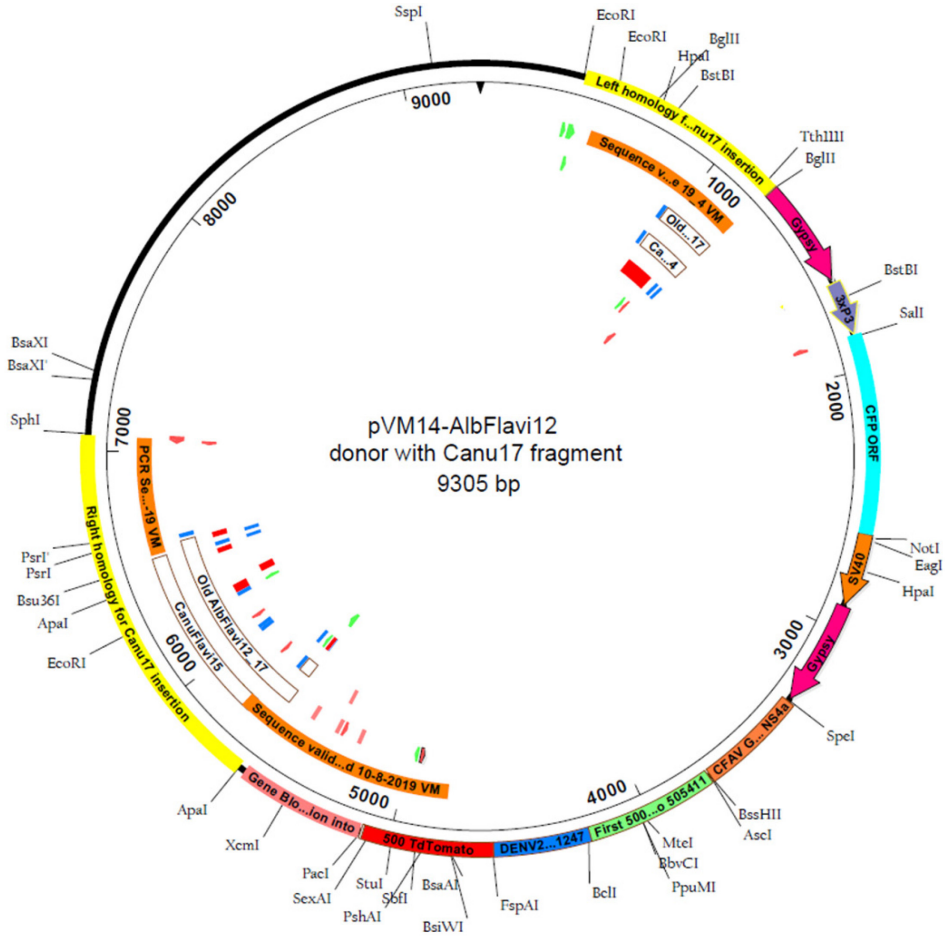
Appendix 3

Data table showing the association between the viral integrations annotated in AaloF1 and AalbF2 assembly.

AaloF1	AalbF2 ID > 99%	ID < 99%*
AlbFlavi1	0	Flavi11, Flavi9, Flavi8, Flavi20, Flavi22
AlbFlavi10	Flavi26, Flavi24	0
AlbFlavi12_17	Flavi8	Flavi10, Flavi21, Flavi9, Flavi11, Flavi23, Flavi22, Flavi23, Flavi20
AlbFlavi18	0	0
AlbFlavi19	0	0
AlbFlavi2	Flavi12	0
AlbFlavi20	0	0
AlbFlavi22	0	Flavi4, Flavi7, Flavi3, Flavi6, Flavi5
AlbFlavi23	Flavi6, Flavi7	Flavi4, Flavi3, Flavi5
AlbFlavi24	Flavi27	0
AlbFlavi25	0	Flavi3, Flavi7
AlbFlavi26	0	Flavi7, Flavi5, Flavi4, Flavi3, Flavi6
AlbFlavi27	0	Flavi4, Flavi7, Flavi3, Flavi6, Flavi5
AlbFlavi28	0	0
AlbFlavi3	Flavi12	0
AlbFlavi31	0	Flavi27
AlbFlavi32	0	Flavi27
AlbFlavi33	Flavi27	0
AlbFlavi34	0	Flavi27, Flavi26, Flavi17, Flavi18
AlbFlavi36	Flavi25, Flavi26	0
AlbFlavi37	0	Flavi7, Flavi6, Flavi3, Flavi5, Flavi4
AlbFlavi38	0	0
AlbFlavi39	0	0
AlbFlavi4	Flavi13	0
AlbFlavi40	0	0
AlbFlavi41	Flavi15	Flavi16
AlbFlavi42	Flavi5, Flavi7	0
AlbFlavi6	0	Flavi1
AlbFlavi7	0	Flavi1
AlbFlavi8	Flavi15	Flavi16
AlbRha1	0	Rhabdo2
AlbRha10	0	0
AlbRha11	Un144	0
AlbRha12	Rhabdo37, Rhabdo26	Rhabdo30, Rhabdo25
AlbRha14	Xinmo3	0

Appendix 4

Detailed map of the CRISPR-Cas9 construct inserted in piRNA cluster 347 to create the CFP12-19 transgenic line. The first 500bp sequences of DENV, MAYV and CFAV are indicated in blue, green and orange respectively.



Original manuscripts

Cristina Crava, Finny S Varghese, Elisa Pischedda, Rebecca Halbach, **Umberto Palatini**, Michele Marconcini, Annamaria Mattia, et al. 2020. "Immunity to Infections in Arboviral Vectors by Integrated Viral Sequences: An Evolutionary Perspective." *BioRxiv*. https://doi.org/10.1101/2020.04.02.022509.

Umberto Palatini, Reem A Masri, Luciano V Cosme, Sergey Koren, Françoise Thibaud-Nissen, James K Biedler, Flavia Krsticevic, et al. 2020. "Improved Reference Genome of the Arboviral Vector *Aedes albopictus*." *Genome Biology* 21 (1): 215. https://doi.org/10.1186/s13059-020-02141-w.

Youdiil Ophinni, **Umberto Palatini**, Yoshitake Hayashi, and Nicholas F Parrish. 2019. "PiRNA-Guided CRISPR-like Immunity in Eukaryotes." *Trends in Immunology* 40 (11): 998–1010. https://doi.org/https://doi.org/10.1016/j.it.2019.09.003.

Michele Marconcini, Luis Hernandez, Giuseppe Iovino, Vincent Houé, Federica Valerio, **Umberto Palatini**, Elisa Pischedda, et al. 2019. "Polymorphism Analyses and Protein Modelling Inform on Functional Specialization of Piwi Clade Genes in the Arboviral Vector *Aedes Albopictus*." *PLOS Neglected Tropical Diseases* 13 (12): e0007919.