frontiers in Genetics

ORIGINAL RESEARCH
published: 12 February 2019
doi: 10.3389/fgene.2019.00093

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*

*Elisa Pischedda[1], Francesca Scolari[1], Federica Valerio[1], Rebeca Carballar-Lejarazú[2], Paolo Luigi Catapano[1], Robert M. Waterhouse[3] and Mariangela Bonizzoni[1]**

[1] Department of Biology and Biotechnology, University of Pavia, Pavia, Italy, [2] Department of Microbiology & Molecular Genetics, University of California, Irvine, Irvine, CA, United States, [3] Department of Ecology and Evolution, University of Lausanne and Swiss Institute of Bioinformatics, Lausanne, Switzerland

OPEN ACCESS

Edited by:
Fulvio Cruciani, Sapienza University of Rome, Italy

Reviewed by:
Joao Trindade Marques, Federal University of Minas Gerais, Brazil
Jinbao Gu, Southern Medical University, China

***Correspondence:**
Mariangela Bonizzoni
m.bonizzoni@unipv.it

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *17 October 2018*
Accepted: *29 January 2019*
Published: *12 February 2019*

Citation:
Pischedda E, Scolari F, Valerio F, Carballar-Lejarazú R, Catapano PL, Waterhouse RM and Bonizzoni M (2019) 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. 10:93. doi: 10.3389/fgene.2019.00093

The Asian tiger mosquito *Aedes albopictus* is an invasive mosquito and a competent vector for public-health relevant arboviruses such as Chikungunya (*Alphavirus*), Dengue and Zika (*Flavivirus*) viruses. Unexpectedly, the sequencing of the genome of this mosquito revealed an unusually high number of integrated sequences with similarities to non-retroviral RNA viruses of the *Flavivirus* and *Rhabdovirus* genera. These Non-retroviral Integrated RNA Virus Sequences (NIRVS) are enriched in piRNA clusters and coding sequences and have been proposed to constitute novel mosquito immune factors. However, given the abundance of NIRVS and their variable viral origin, their relative biological roles remain unexplored. Here we used an analytical approach that intersects computational, evolutionary and molecular methods to study the genomic landscape of mosquito NIRVS. We demonstrate that NIRVS are differentially distributed across mosquito genomes, with a core set of seemingly the oldest integrations with similarity to *Rhabdoviruses*. Additionally, we compare the polymorphisms of NIRVS with respect to that of fast and slow-evolving genes within the *Ae. albopictus* genome. Overall, NIRVS appear to be less polymorphic than slow-evolving genes, with differences depending on whether they occur in intergenic regions or in piRNA clusters. Finally, two NIRVS that map within the coding sequences of genes annotated as *Rhabdovirus* RNA-dependent RNA polymerase and the nucleocapsid-encoding gene, respectively, are highly polymorphic and are expressed, suggesting exaptation possibly to enhance the mosquito's antiviral responses. These results greatly advance our understanding of the complexity of the mosquito repeatome and the biology of viral integrations in mosquito genomes.

Keywords: mosquitoes, viral integrations, immunity, piRNA pathway, domestication, repeatome

Abbreviations: FC, fold-change; FGs, fast-evolving genes; F-NIRVS, NIRVS with similarity to *Flaviviruses;* LoP, level of polymorphism; N-Gs, genes harboring NIRVS; NIRVS, non-retroviral integrated RNA virus sequences; R-Gs, Genes of the RNAi pathway; R-NIRVS, NIRVS with similarity to *Rhabdoviruses;* SGs, slow-evolving genes; SSM, single-sequenced mosquitoes.

INTRODUCTION

The amount and the type of repeated DNA sequences, collectively called the "repeatome," affect the size, organization and evolution of eukaryotic genomes (Maumus and Quesneville, 2014). Transposable elements (TEs) are the major and most-studied components of the repeatome because of their potential mutagenic effects (Gilbert and Feschotte, 2018). TEs evolve through a "burst and decay model" whereby newly acquired TEs can multiply rapidly in a genome. The "burst" phase is followed by low amplification periods, the "decay" moments, when TEs tend to accumulate mutations and become inactive (Maumus and Quesneville, 2016). In eukaryotes, TE mobilization during germline formation is counterbalanced by the activity of the PIWI-interacting RNA (piRNA) pathway, the most recently identified of three small RNA-based silencing mechanisms (Brennecke et al., 2007; Guzzardo et al., 2013; Gainetdinov et al., 2017). Briefly, Argonaute proteins of the PIWI-subfamily associate with small RNAs of 25-30 nucleotides, called PIWI-interacting RNAs (piRNAs), and together they silence TEs based on sequence complementarity (Tóth et al., 2016). piRNAs arise from genomic regions called piRNA clusters, which contain fragmented sequences of previously acquired TEs.

Unexpectedly, besides TE fragments, piRNA clusters contain sequences from non-retroviral RNA viruses, which produce piRNAs, in the genome of arboviral vectors like the mosquitoes *Aedes aegypti* and *Aedes albopictus* (Arensburger et al., 2011; Olson and Bonizzoni, 2017; Palatini et al., 2017; Whitfield et al., 2017). This observation is in line with recent experimental evidence that extend the role of the piRNA pathway to immunity against viruses in *Aedes* mosquitoes, differently than in *Drosophila melanogaster* (Miesen et al., 2016; Petit et al., 2016) and show that piRNAs from integrated viral sequences are differentially expressed following viral challenge of *Ae. albopictus* (Wang et al., 2018). As such, NIRVS have been proposed as novel immunity factors of arboviral vectors (Olson and Bonizzoni, 2017; Palatini et al., 2017; Whitfield et al., 2017). However, the organization, stability and mode of action of NIRVS in mosquito genomes are poorly understood.

The landscape of viral integrations in the genome of *Ae. aegypti* and *Ae. albopictus* mosquitoes is rather complex. *Aedes* species are rare examples within the animal kingdom because they harbor dozens of NIRVS from different viruses, such as *Flaviviruses* and Mononegavirales, primarily *Rhabdoviruses* and poorly characterized *Chuviruses* (Katzourakis and Gifford, 2010; Fort et al., 2012; Palatini et al., 2017; Whitfield et al., 2017). In all other animals in which NIRVS have been identified, including mammals, birds and ticks, NIRVS appear to be mainly from one viral type and tend to be found in low numbers (<20) (Belyi et al., 2010; Katzourakis and Gifford, 2010; Holmes, 2011; Kryukov et al., 2018). NIRVS identified in the *Ae. aegypti* and *Ae. albopictus* genomes are not homologous, indicating independent integration events. However, NIRVS of both species encompass fragmented viral open reading frames (ORFs). In *Ae. albopictus,* we characterized 32 NIRVS with similarities to *Flaviviruses* (F-NIRVS) and 40 NIRVS similar to *Rhabdoviruses* (R-NIRVS). These NIRVS are enriched in piRNA clusters and within coding sequences (Palatini et al., 2017). Taken together these findings support the hypothesis that NIRVS contribute to host biology. However, because NIRVS have been identified by *in silico* analyses of the currently available *Ae. albopictus* genome assembly, which was built from the DNA of a single pupa of the Foshan strain (Chen et al., 2015) and we verified the overall conservation of NIRVS within this strain (Palatini et al., 2017), their widespread occurrence in wild mosquitoes, whether all NIRVS or some are functionally active elements, and what is the relative importance of each of them, are all still unexplored questions.

Here we addressed the following questions: is the pattern of NIRVS within mosquitoes of the Foshan strain the same as across geographic samples? If the landscape of NIRVS is variable, could NIRVS be co-opted as novel molecular markers for population genetic studies? Does NIRVS age differ depending on their viral origin? How does the LoP of NIRVS compare with that of fast- and slow-evolving mosquito genes?

Using an analytical approach that intersects computational, evolutionary, and molecular approaches we show that NIRVS are a dynamic component of the *Ae. albopictus* repeatome. The landscape of NIRVS is variable within mosquitoes of the Foshan strains and among geographic samples. The LoP of NIRVS is heterogeneous. R-NIRVS appear more widespread and older integrations than those with similarities F-NIRVS. NIRVS annotated in intergenic regions appear more variable than those mapping within piRNA clusters or gene exons. Among NIRVS identified within gene exons, six are fixed and stably expressed, albeit showing different levels of polymorphism and domestication cannot be excluded for AlbRha52 and AlbRha12, which are part of genes annotated as RNA-dependent RNA polymerase and nucleocapsid-encoding genes of *Rhabdovirus*, respectively.

Overall these results greatly advance our understanding of the widespread occurrence of NIRVS in nature. Additionally, a detailed analysis of NIRVS distribution and polymorphism within the *Ae. albopictus* genome paves the way for choosing candidate NIRVS for functional studies.

MATERIALS AND METHODS

Mosquitoes

Mosquitoes of the Foshan strain have been reared at the insectary of the University of Pavia since 2013 (Palatini et al., 2017). Upon arrival in Pavia, mosquitoes were checked for infection using *Flavivirus* degenerate primers (Crochu et al., 2004). No infection was detected. Mosquitoes are reared at 28°C and 70–80% relative humidity with 12/12 h light/dark cycles. Larvae are reared in pans and fed on finely ground fish food (Tetramin, Tetra Werke, Melle, Germany). Adults are kept in 30-cm^3 cages and allowed access to a cotton wick soaked in 0.2 g/ml sucrose as a carbohydrate source. Adult females are blood-fed using a membrane feeding apparatus and commercially available mutton blood. Sixteen Foshan mosquitoes, eight males and eight females, were sampled and forced in single mating. Progeny from each single mating was collected. DNA was extracted from single individuals, including

parents and their progeny, using the DNeasy Blood and Tissue Kit (Qiagen, Hilden, Germany).

Southern Blotting

Genomic DNA (∼10 mg) from pools of 10–20 adult mosquitoes of the Foshan strain were digested with restriction enzymes (Thermo Scientific, Vilnius, Lithuania) chosen to specifically target individual F-NIRVS and separated on a 0.8% agarose gel. DNA was transferred to nylon membranes (Hybond-N+) (Amersham, Buckinghamshire, United Kingdom) and immobilized by UV irradiation. Random-primed DNA probes (**Supplementary Data 1**) were labeled with [α-^{32}P] dATP/ml and [α-^{32}P] dCTP/ml (3,000 Ci/mmol; 1 Ci- 37 GBq) by using the Megaprime labeling kit (Amersham, Buckinghamshire, United Kingdom). Hybridizations were carried out at 65°C.

Real Time PCR (qPCR) to Test for NIRVS Copy Number

PCR primers were designed using PRIMER3 (Rozen and Skaletksy, 2000) within F-NIRVS to test for their copy number based on real-time PCR (Bubner and Baldwin, 2004; Yuan et al., 2007) (**Supplementary Data 2**). Reaction mixtures were prepared containing 10 μL QuantiNova Sybr Green PCR Master Mix (Qiagen, Hilden, Germany), 1 μL of each 10 μM primer, and template DNA diluted in distilled H_2O up to 20 μL total reaction volume. Template genomic DNA used in the reactions was extracted from individual adult mosquitoes following a standard protocol (Baruffi et al., 1995). Real-time PCR reactions were performed in a two-step amplification protocol consisting of 2 min at 95°C, followed by 40 cycles of 95°C for 5 s and 60°C for 10 s. Reactions were run on a RealPlex Real-Time PCR Detection System (Eppendorf, Hamburg, Germany). The single-copy gene *piwi6* (AALF016369) was used as reference after having verified the region of the primers does not harbor variability. F-NIRVS copy numbers were estimated comparing the relative quantification of NIRVS loci with respect to that of the reference genes using the ΔCt method (Pfaffl, 2006), after having verified that the efficiencies of PCR reactions with primers for F-NIRVS and the reference gene were the same. Support for using relative quantification without an internal calibrator came from a preliminary test where we cloned one NIRVS (AlbFlavi4) and we verified that estimates of its copy number by absolute vs. relative quantification were the same.

qPCR to Estimate NIRVS Expression Levels

Total RNA was extracted using TRIzol (Life Technologies, Madrid, Spain) from pools including 10–20 mosquitoes at different developmental stages such as larvae, pupae, adult males, sugar-fed females and females sampled 48 h after blood feeding. After DNaseI (Sigma-Aldrich, Schnelldorf, Germany) treatment, a total of 100 ng of RNA from each pool was used for reverse transcription using the qScript cDNA SuperMix (Quanta Biosciences, Leuven, Belgium). Expression of the eight N-Gs and always detected in Foshan (i.e., AALF005432, AALF025780, AALF000476, AALF000477, AALF020122, AALF004130, and AALF025779) was quantified using real-time qPCRs following the protocol described above. Expression values were normalized to mRNA abundance levels of the *Ae. albopictus nap* gene (Reynolds et al., 2012) (**Supplementary Data 3**). The qbase+software (Hellemans et al., 2007) was used to compare expression profiles across samples, and Morpheus[1] was used to visualize the data.

Selection of Genes With Slow and High Evolutionary Rates

Orthologous genes across 27 insect species within the Nematocera sub-order were identified in OrthoDB v9.1 (Zdobnov et al., 2016). Levels of sequence divergence were computed for each orthologous group as the average of interspecies amino acid identified normalized to the average identity of all interspecies best-reciprocal-hits, computed from pairwise Smith-Waterman alignments of protein sequences (**Supplementary Table 1**). We selected the 0.1% of the genes (n = 14, number comparable to that of our NIRVS groups) at each tail of the distribution as representative of the conserved and variable categories, the left and right tails respectively. Orthologs of these genes were searched in the *Ae. albopictus* genome (AaloF1 assembly).

NIRVS in Natural Populations

PCR primers were designed using PRIMER3 (Rozen and Skaletksy, 2000) to test for NIRVS polymorphism in *Ae. albopictus* geographic samples (**Supplementary Data 4**). Considering the level of NIRVS sequence similarity, their copy number and heterogeneous presence in Foshan mosquitoes, we selected seven F-NIRVS (AlbFlavi2, AlbFlavi4, AlbFlavi8-41, AlbFlavi10, AlbFlavi36, AlbFlavi1, and AlbFlavi12-17) and six R-NIRVS (AlbRha1, AlbRha7, AlbRha14, AlbRha36, AlbRha52, AlbRha85) that gave unambiguous PCR results, have similarities to different viral ORFs and are distributed in different genomic regions including piRNA clusters, intergenic or coding regions. Natural mosquito samples derive from a world-wide collection available at the University of Pavia and previously analyzed with microsatellite markers (Manni et al., 2017). PCR reactions were performed in a final volume of 25 μL using DreamTaq™ Green PCR Master Mix 2x (Thermo Scientific, Vilnius, Lithuania) and the following cycle conditions: 94°C for 3 min, 40 cycles at 94°C for 30 s, 58–60°C for 45 s, 72°C for 1 min, and a final extension at 72°C for 10 min. Amplification products were electrophoresed on 1–1.5% agarose gels and purified using ExoSAP-IT^{M} PCR product Cleanup Reagent (Thermo Scientific, Vilnius, Lithuania). When the NIRVS were detected, at least five amplification products per population per locus were sent to be sequenced by Macrogen (Barcellona, Spain), following the company's requirements.

Non-retroviral Integrated RNA Virus Sequences alleles were first scored based on their occurrence in each population and their size. A Neighbor-joining tree was built after 1000 bootstrap resampling of the original data set and the calculation of a matrix

[1]https://software.broadinstitute.org/morpheus

of shared allele distances (DAS) using POPULATIONS version 1.2.31 (Langella, 1999).

Estimates of Integration Time

Non-retroviral Integrated RNA Virus Sequences sequences from geographic samples were aligned in Ugene version 1.26.1 (Okonechnikov et al., 2012) with MAFFT (Yamada et al., 2016). Default parameters with five iterative refinements were applied for the alignment. Alignments were manually curated to verify frameshifts, truncations, deletions, and insertions. All positions including gaps were filtered out from the analysis. The following formula was used to estimate the time of integration in years assuming that all mutations are neutral:

$$\text{Mean Mutations/Seq} = \frac{\text{Tot. Obs. Mutations}}{\text{N. Seqs* Seq. Length}}$$

$$\text{Age in Years} = \frac{\text{Mean Mutations/Seq}}{\text{(MR* Seq. Length* GpY)}}$$

Mutation rate (MR) were assumed to be comparable to those of *D. melanogaster* genes in range 3.5-8.4 $\times$ 10^{-09} (Haag-Liautard et al., 2007; Keightley et al., 2009). A range of 4–17 number of generations per year (GpY) was tested considering mosquitoes of temperate or tropical environments (Manni et al., 2017).

Phylogenetic Inference and Timetrees

Deduced NIRVS protein sequences were aligned with subsets of corresponding proteins from *Flavivirus* and *Rhabdovirus* genomes using MUSCLE (Edgar et al., 2004). The timetrees were generated using the RelTime method (Tamura et al., 2012) after having generated the maximum likelihood tree, with 100 bootstrap replicates. Divergence times for all branching points in the topology were calculated using the maximum likelihood method and implementing the best fitted amino acids substitution model. Phylogenies were estimated in MEGA7 (Kumar et al., 2016). The JTT matrix-based model was used for the L protein of *Rhabdovirus* (Jones et al., 1992). In this case, the estimated log likelihood value was −116005.08. A discrete Gamma distribution was used to model evolutionary rate differences among sites [2 categories (+G, parameter = 0.8331)]. The rate variation model allowed for some sites to be evolutionarily invariable ([+I], 0.21% sites). The analysis involved 49 amino acid sequences. There was a total of 2319 positions in the final dataset. For the G protein of *Rhabdoviruses*, the Whelan and Goldman model was implemented (Whelan and Goldman, 2001). In this case, the estimated log likelihood value was −3719.06. A discrete Gamma distribution was used to model evolutionary rate differences among sites [2 categories (+G, parameter = 2.1095)]. The analysis involved 40 amino acid sequences. All positions with less than 95% site coverage were eliminated. That is, fewer than 5% alignment gaps, missing data, and ambiguous bases were allowed at any position. There was a total of 56 positions in the final dataset. The LG model was used for the NS3 protein of *Flaviviruses* (Le and Gascuel, 2008). In this case, the estimated log likelihood value is −6360.35. A discrete Gamma distribution was used to model evolutionary rate differences among sites [2 categories (+G, parameter = 0.8640)]. The analysis involved 30 amino acid sequences. All positions containing gaps and missing data were eliminated. There was a total of 180 positions in the final dataset. For NIRVS with similarities to the NS5 protein of *Flaviviruses* the JTT matrix-based model was used (Jones et al., 1992). The estimated log likelihood value of the topology shown was −26019.44. A discrete Gamma distribution was used to model evolutionary rate differences among sites [2 categories (+G, parameter = 1.0058)]. The rate variation model allowed for some sites to be evolutionarily invariable ([+I], 10.15% sites). The analysis involved 33 amino acid sequences. There was a total of 984 positions in the final dataset. In each case, trees were drawn to scale, with branch lengths measured in the relative number of substitutions per site. The coding sequences for proteins of the Potato Yellow Dwarf virus (PYDV) were used as outgroup for trees of R-NIRVS, considering PYDV belongs to the highly divergent *Nucleorhabdovirus* genus (Dietzgen et al., 2017). To derive the genealogy of F-NIRVS, outgroups were protein sequences from Tamana Bat Virus (TABV) (de Lamballerie et al., 2002).

Bioinformatic Pipeline to Study the Polymorphisms of NIRVS, Fast- and Slow-Evolving Genes

Whole genome sequencing data of 16 singly sequenced (i.e., single-sequenced mosquitoes or SSMs) as previously described (Palatini et al., 2017) was used for the analyses of NIRVS polymorphism. NIRVS presence in a sample was established imposing a more stringent criteria than previously used in Palatini et al. (2017). Here to the "minimum of five reads of depth of coverage," we added a minimum of 30 consecutive nucleotides with that depth of coverage (**Supplementary Table 2**). This more stringent criteria resulted in a difference of one in the number of NIRVS called as absent (AlbRha43). We molecularly validated bioinformatic predictions based on this criterion (**Supplementary Figure 1**). The ratio between the number of R-NIRVS present in a sample and the total R-NIRVS of Foshan (40) was used to estimate R-NIRVS prevalence. The same calculation was done for F-NIRVS. The polymorphism of NIRVS and that of selected FGs and slow-evolving genes (i.e., SGs) was then estimated using a custom pipeline organized into different steps. In the first step, the DepthOfCoverage function of the GATK tool (Mckenna et al., 2010) is used to evaluate the coverage of the region of interest limiting to reads with Phred mapping quality greater than 20. Following read coverage analyses, four different Variant Callers i.e., GATK UnifiedGenotyper (Mckenna et al., 2010), Freebayes (Garrison and Marth, 2012), Platypus (Rimmer et al., 2014), and Vardict (Lai et al., 2016), were implemented to identify SNPs and INDELs within the regions of interest. The search of SNPs and INDELS by different variant callers allowed to increase the pool of variants and reduce the number of false positive. Custom scripts were then used to filter data, retain only variants having allele frequency higher than 0.1 or variants called by at least two

programs. The LoP of the region of interest was calculated as the total number of SNPs and INDELs identified averaged based on its length.

Follow-up statistical analyses were computed and visualized in R studio (RStudio Team, 2015). RStudio: Integrated Development for R. RStudio, Inc. (Boston, 2015). The Kolmogorov–Smirnov test was used to test the significance of the difference of LoP distributions of NIRVS, RNAi genes (R-Gs), N-Gs and FGs with respect to that of SGs (**Supplementary Table 3**). SG LoP was the median of the LoPs of the tested SGs. The threshold of significance was adjusted with the Bonferroni correction and loci were separated according to the adjusted significance of the test. Results of ratio between the LoP of each locus and the median LoP of SGs (fold change [FC]) that were different from 0 were visualized in a volcano plot. For each locus, FC was calculated as the ratio of the median LoP of the locus and that of the SG. The hypergeometric test was applied to test whether the group of NIRVS always identified across SSMs was enriched in (1) F- or R-NIRVS; (2) any viral ORFs; (3) NIRVS shorter or longer than 500 bp; (4) NIRVS mapping in exons, piRNA clusters or intergenic regions.

Search for Novel Viral Integrations

Sequences supported by the presence of soft-clipped reads were molecularly tested by PCR assays using DNA from individual mosquitoes of the Foshan strain (**Supplementary Data 5**). The Vy-PER pipeline (Forster et al., 2015) was applied to WGS data from the 16 SSMs to search for viral integrations that had not been previously identified in genome of the Foshan strain (AaloF1 assembly). Vy-PER was run using 540 viral genomes from VIPERdb (Carrillo-Tripp et al., 2009), including species of the Togaviridae, Flaviviridae, Bunyaviridae, Rhabdoviridae, Orthomyxoviridae, Reoviridae, Bornaviridae, Filoviridae, Nyamiviridae, Paramyxoviridae families. Paired-end reads identified by Vy-PER in which one read maps to the reference mosquito genome (i.e., AaloF1) and its pair maps to one of the tested viral genomes were manually inspected. Candidates including low complexity sequence (i.e., sequence showing more than 80% in mono- and di-nucleotides) or with viral portion shorter than 50 nucleotides were considered false positive and were filtered out.

RESULTS

We use read depth of coverage and variant calling tools to study NIRVS (Non-retroviral Integrated RNA Virus Sequences) widespread occurrence and their polymorphism within the genomes of mosquitoes of the Foshan strain and to look for novel viral integrations. Additionally, we studied the distribution of a selected subset of NIRVS in geographic samples.

NIRVS Are Variably Distributed in SSMs

We used the sequenced genomes of 16 mosquitoes (i.e., single-sequenced mosquitoes or SSMs) (Palatini et al., 2017) and we compared their NIRVS pattern with the list of viral integrations characterized from the Foshan genome assembly (AaloF1). Eleven NIRVS (i.e., AlbFlavi19, AlbFlavi31, AlbFlavi32, AlbFlavi33, AlbFlavi38, AlbFlavi39, AlbFlavi40, AlbRha43, AlbRha79, AlbRha80, AlbRha95) were absent in all 16 SSMs (**Supplementary Table 2**). A total of 20 NIRVS were found in all SSMs, with a statistical enrichment for NIRVS with similarities to *Rhabdovirus* (R-NIRVS) (Hypergeometric test, $p = 0.022$) and NIRVS mapping in gene exons (Hypergeometric test, $p = 0.006$) (**Figure 1A**). This "core" of 20 NIRVS included R-NIRVS identified within the coding sequence of genes (i.e., AlbRha12, AlbRha15, AlbRha28, AlbRha52, AlbRha85 and AlbRha9) and piRNA clusters (i.e., AlbRha14 and AlbRha36). Conversely, NIRVS with similarities to *Flaviviruses* (F-NIRVS) were variably distributed among SSMs. Of note is AlbFlavi4, a 512bp sequence with similarity to the capsid gene of *Aedes flavivirus* (Palatini et al., 2017). AlbFlavi4 is annotated within the second exon of AALF003313 and is also included in piRNA cluster 95 (Liu et al., 2016). AlbFlavi4 produces vepi4730383, a piRNA that is upregulated upon dengue infection (Wang et al., 2018). In SSMs and *Ae. albopictus* geographic samples, variants were identified for AALF003313, only one of which includes AlbFlavi4 (**Figures 1B,C**).

Overall, mean base pairs (bp) occupied by F-NIRVS and R-NIRVS are 12095 and 19293 bp, respectively (**Figure 1D**). Taken together, these results demonstrate that, with an average genome occupancy of 31389 bp, NIRVS represent quantitatively a limited fraction of the mosquito repeatome. However, the enrichment of NIRVS in piRNA clusters (Palatini et al., 2017) and the fact that the pattern of NIRVS is variable in host genomes support the hypothesis that NIRVS are a dynamic component of the repeatome.

NIRVS Distribution in Geographic Populations

To verify if NIRVS are variably distributed in natural samples besides in the Foshan strain, we choose seven F-NIRVS (AlbFlavi2, AlbFlavi4, AlbFlavi8-41, AlbFlavi10, AlbFlavi36, AlbFlavi1, and AlbFlavi12-17) and six R-NIRVS (AlbRha1, AlbRha7, AlbRha14, AlbRha36, AlbRha52, AlbRha85) based on their unique occurrence in different regions of the mosquito genome and their similarity to various viral ORFs. AlbRha52 and AlbRha85 are annotated as unique exons of AALF020122 and AALF004130, respectively. We tested the presence of these NIRVS in native (China and Thailand), old (La Reunion Island) and new (United States and Italy) *Ae. albopictus* populations (Manni et al., 2017). NIRVS alleles were differentially distributed across geographic populations so that a tree built from a matrix of shared-allele distances (DAS) proved able to differentiate mosquito populations in accordance with the historical records of *Ae. albopictus* invasive process when considering all thirteen NIRVS, only F-NIRVS or NIRVS mapping in intergenic regions (**Figures 2A–C** and **Supplementary Data 6**). On the contrary, when data from exclusively R-NIRVS or NIRVS identified in piRNA clusters, were analyzed, bootstrap values differentiating populations were below 50% (**Figures 2D,E**). This result agrees with the observation that the higher abundant R-NIRVS are also more prevalent than F-NIRVS.

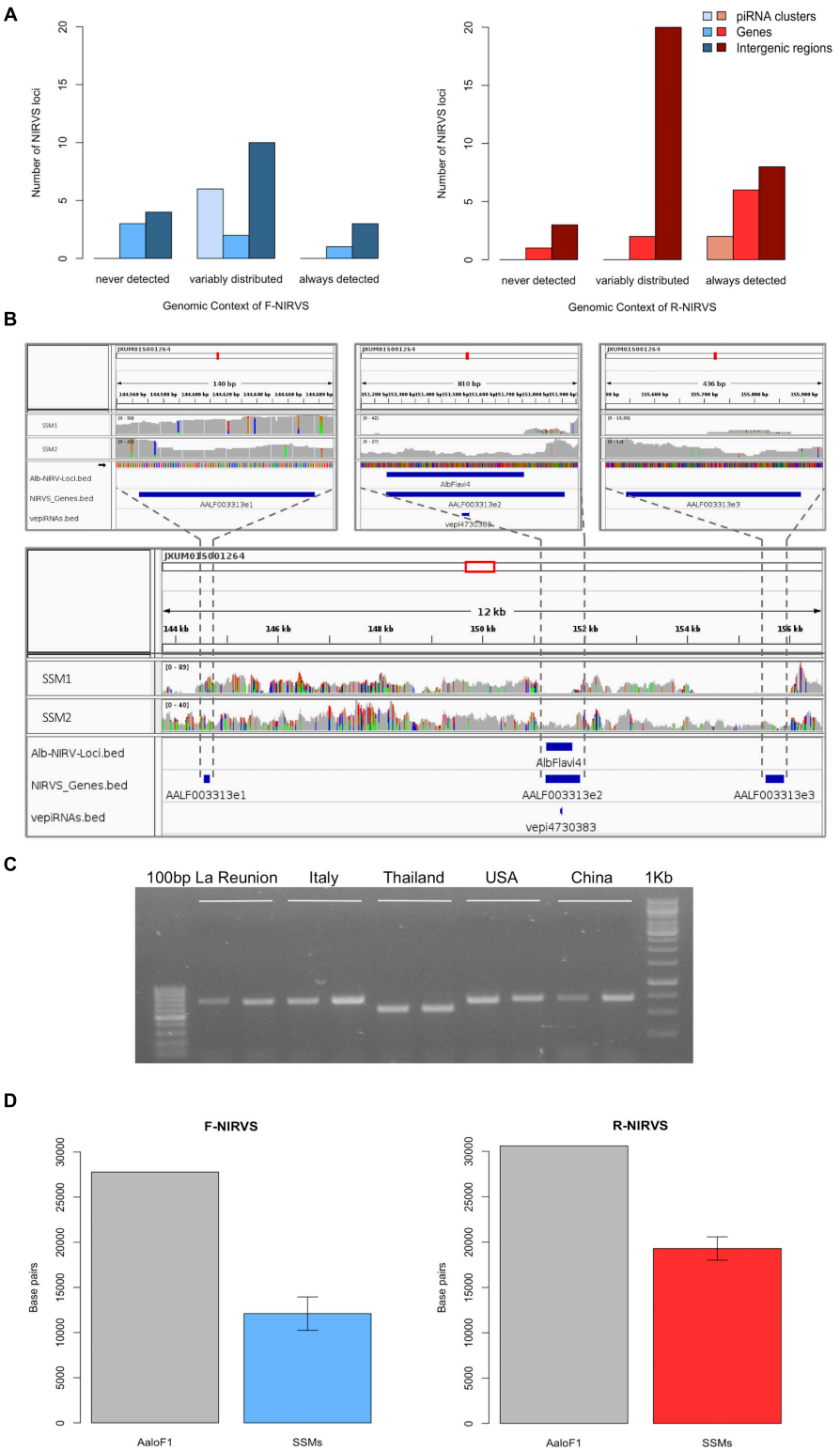

FIGURE 1 | NIRVS are variably distributed in SSMs. **(A)** Number of *Flavivirus* (F-NIRVS) and *Rhabdovirus* (R-NIRVS) loci mapping within genes, piRNA clusters or intergenic regions, classified on the basis of read-coverage across SSMs. **(B)** IGV screen shot showing read-coverage at AALF003313 in SSM1 and SSM2. Positions of the three AALF003313 exons, AlbFlavi4 and vepi4730383 are indicated by blue bars. **(C)** PCR amplification of AALF003313 exon2 in ten *Ae. albopictus* geographic samples. **(D)** F-NIRVS and R-NIRVS loci occupancy in the 16 single-sequenced mosquitoes (SSMs) of the Foshan strain is about half of that expected based on the annotated sequences of the reference genome assembly (AaloF1). F-NIRVS are in blue, R-NIRVS are in red.

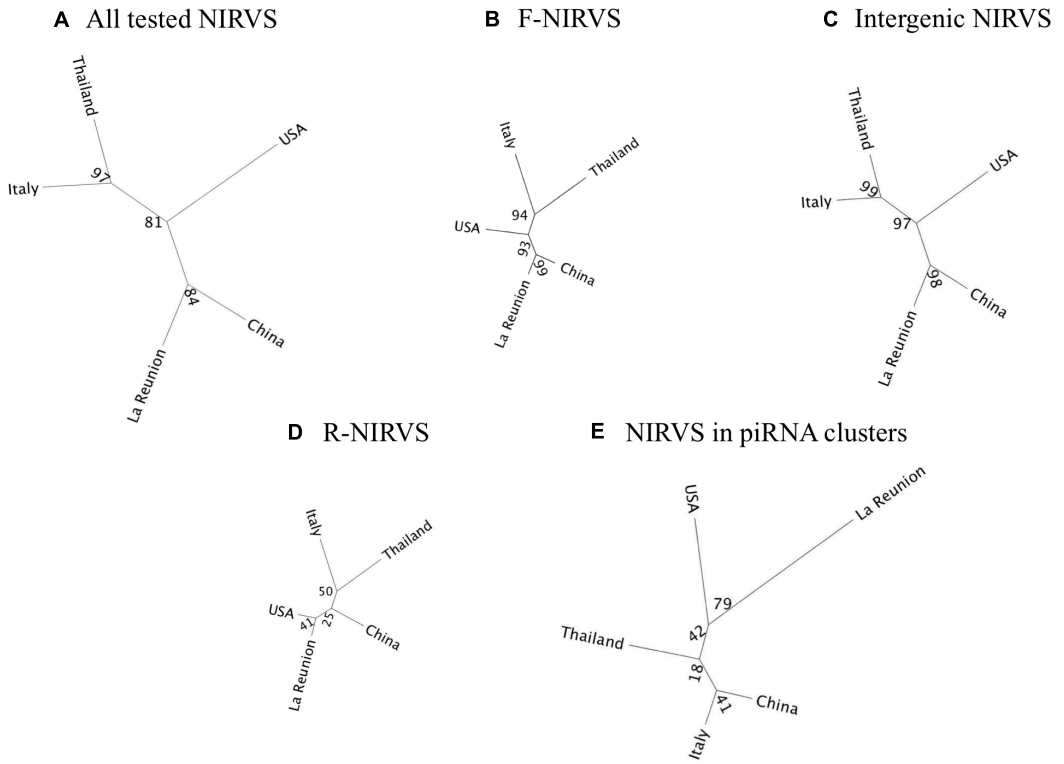

FIGURE 2 | Phylogeographic distribution of NIRVS. Genetic relationships among five mosquito populations shown by a Neighbor-joining trees based on shared-allele distance using data from all 13 genotyped NIRVS **(A)**, only F-NIRVS **(B)**, NIRVS mapping in intergenic regions **(C)**, only R-NIRVS **(D)** and NIRVS mapping in piRNA clusters **(E)**. Bootstrap values are shown at the corresponding nodes.

R-NIRVS Appear to Be Older Integrations Than F-NIRVS

The higher prevalence of R-NIRVS with respect to F-NIRVS suggests R-NIRVS are older integrations (**Figure 3**). To verify this hypothesis, we sequenced alleles of NIRVS identified in the five tested populations and we estimated integration times, assuming comparable mutation rates between *Ae. albopictus* and *D. melanogaster*, that is 3.5–8.4 $\times$ 10^{-9} per site per generation (Haag-Liautard et al., 2007; Keightley et al., 2009), and a range of generations per year between 4 and 17, accounting for mosquitoes from temperate and tropical environments, respectively (Manni et al., 2017). Under these conditions, R-NIRVS integrated between 36 thousand and 2.7 million years ago (mya) and F-NIRVS between 7.4 thousand and 2.4 mya (**Figure 4**). This large window supports the conclusion that integration of viral sequence is a dynamic process occurring occasionally at different times. As shown in **Figure 4**, estimates of integration times varied greatly depending on the genomic context of NIRVS. NIRVS annotated within gene exons appear statistically more recent than NIRVS of piRNA clusters (ANOVA, $^{***}P < 0.001$). Besides reflecting a different integration time, this result is consistent with the hypothesis that integrations within exons are under rapid evolution, a hallmark of domestication (Frank and Feschotte, 2017).

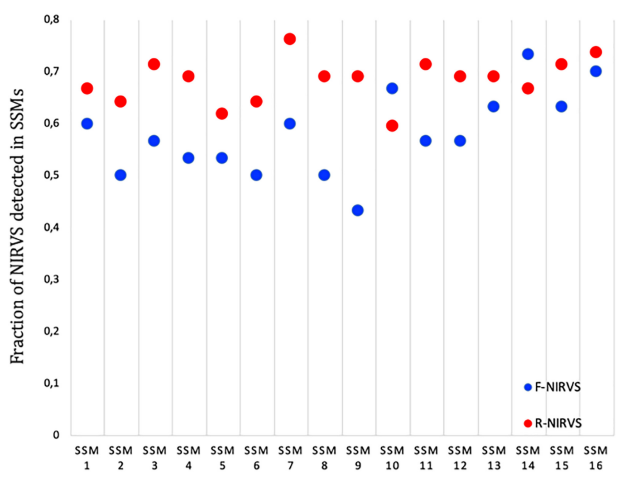

FIGURE 3 | Prevalence of R-NIRVS (red) and F-NIRVS (blue) in each SSMs. Prevalence was calculated in each sample as the ration between detected NIRVS and annotated NIRVS for both R- and F-NIRVS.

Additionally, we tested the genealogy of R-NIRVS and F-NIRVS in comparison to circulating *Rhabdoviruses* and *Flaviviruses*. Relative timetrees were generated for (i) F-NIRVS and corresponding NS3 and NS5 viral proteins from representative *Flaviviruses*, and (ii) R-NIRVS and corresponding L and G proteins of representative *Rhabdoviruses*. Timetrees

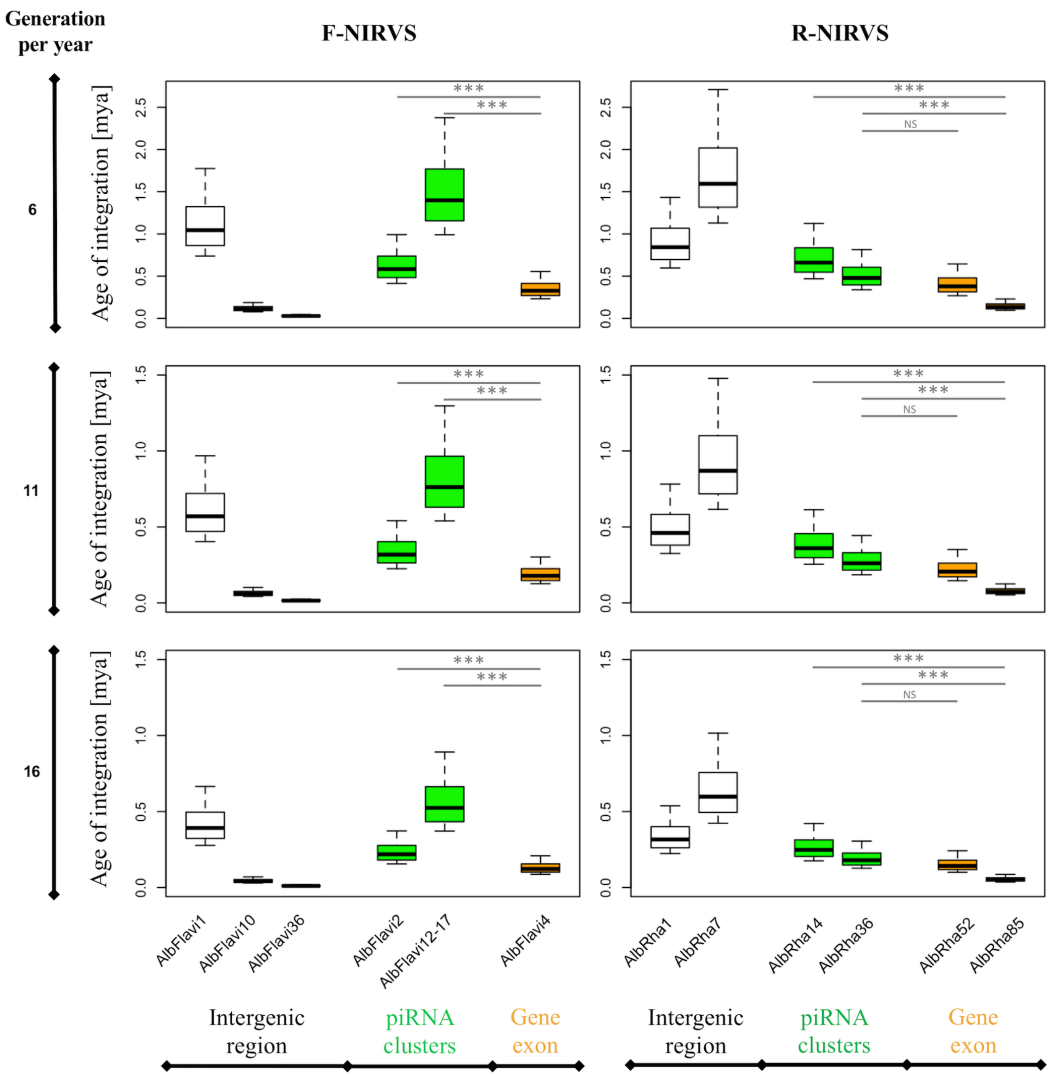

FIGURE 4 | NIRVS integration times. Boxplots showing the integration times for the NIRVS whose variability was studied across five geographic populations. Estimates are based on the *D. melanogaster* mutation rate, i.e., 3.5–8.4 $\times$ 10^{-9} per site per generation (Haag-Liautard et al., 2007; Keightley et al., 2009), and a range of generations per year between 6 and 16 to include mosquitoes from temperate and tropical regions (Manni et al., 2017). NIRVS of piRNA clusters are statistically older than NIRVS mapping in gene exons (ANOVA, ***$P < 0.001$).

showed shorter divergence times between F-NIRVS and viral proteins than R-NIRVS and viral proteins. This clearly indicated multiple integration events and a tendency of R-NIRVS to be older integrations (**Supplementary Figure 2**).

NIRVS Are Heterogeneously Polymorphic at the Sequence Level, With the Majority Being Less Variable Than Slow-Evolving Genes

We selected genes having low and high evolutionary rates in *Ae. albopictus* and we compared their levels of polymorphism (LoP) with that of NIRVS in WGS data from our 16 SSMs. LoP was evaluated as the ratio between the number of total mutations (SNPs and INDELs) found in the locus and its length. We expanded the analyses to include also R-Gs, for which intraspecific rapid evolution has been observed in *Ae. aegypti* (Bernhardt et al., 2012), and the 13 N-Gs in their coding sequence or UTRs (Palatini et al., 2017).

Estimates of gene evolutionary rates were derived from comparisons of levels of protein sequence divergence within groups of orthologous genes across 27 insect species of the Nematocera sub-order (**Supplementary Table 1**). The first and last 0.1% of the genes from the evolutionary rate distribution were selected as slow and fast evolving genes, respectively, and their single-copy orthology status with respect to *Ae. albopictus* was verified. We did not select genes based on their length because of the wide length range of NIRVS, which includes partial viral ORFs of between 151 and 3206 bp (Palatini et al., 2017). SGs that met the above

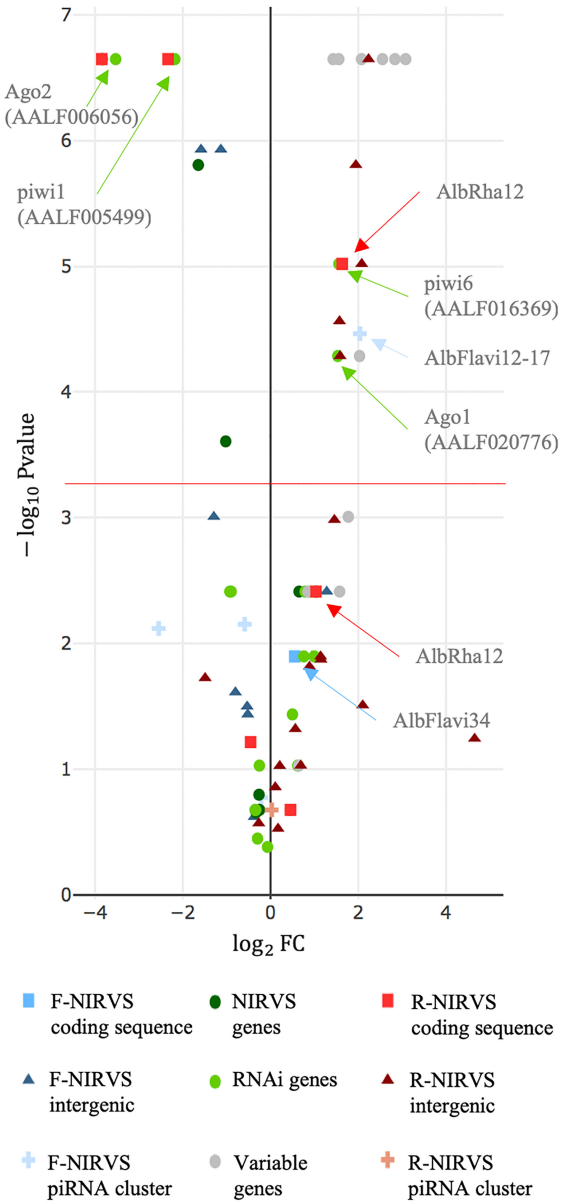

FIGURE 5 | NIRVS polymorphism. Volcano plot showing LoP comparison between SGs and NIRVS, genes encompassing NIRVS (Palatini et al., 2017), genes of the RNAi pathway and FGs. Entities with LoPs statistically different than that of conserved genes are above the red line [adjusted significance of the test (–log10 0.05/99 = 3.32)]. Entities on the left side of the panel (log2FC < 0) have smaller LoP than conserved genes. The opposite for entities on right side of the panel (log2FC > 0).

criteria include genes with hypothetical protein transporter or vesicle-mediated transport activity (*i.e.,* AALF003606, AALF014156, AALF014287; AALF014448; AALF004102), structural activity (AALF005886, annotated as tubulin alpha chain), signal transducer activity (AALF026109), protein and DNA binding activity (AALF027761, AALF028431), SUMO transferase activity (AALF020750), the homothorax homeobox encoding gene AALF019476, the tropomyosin invertebrate gene (AALF0082224), the Protein yippee-like (AALF018378) and autophagy (AALF018476). FGs include genes with unknown functions (AALF004733, AALF009493, AALF009839, AALF012271, AALF026991, AALF014993, AALF017064, AALF018679), proteolysis functions (AALF010748) a gene associated with transcriptional (AALF022019), DNA-binding (AALF019413, AALF024551), structural (AALF028390) and proteolytic (AALF010877) activities. Median LoP of SGs within mosquitoes of the Foshan strain is 0.0071, a value higher than that observed across 63.3% of the detected NIRVS (**Supplementary Figure 3**). Eleven out of fourteen FGs were more variable than SGs, with seven appearing also statistically more polymorphic than SGs (Kolmogorov-Smirnov test, $^*P < 0.05$) (**Figure 5** and **Supplementary Table 3**). This result further supports our selection of SGs and FGs.

Genes of the RNAi pathway are heterogeneously polymorphic (**Figure 5**), with *Ago1* (AALF020776) and *piwi6* (AALF016369) being statistically more polymorphic than SGs; the opposite result was obtained for *piwi1* and *3* (AALF005499, AALF005498), and *Ago2* (AALF006056) (**Figure 5**). LoP of NIRVS is heterogeneous both among SSMs and. NIRVS identified within piRNA clusters (Liu et al., 2016) are all less polymorphic than SGs, with the exception of AlbFlavi12-17 that has a median LoP value of 0.0258. This large LoP may be due by the fact that AlbFlavi12-17 is composed of four small viral sequences nested one next to the other (Palatini et al., 2017). Unlike NIRVS from piRNA clusters, NIRVS spanning gene exons are more heterogeneous; three (i.e., AlbFlavi34, AlbRha12, and AlbRha52) have LoP values higher than those of SGs, while others (*i.e.,* AlbFlavi24, AlbRha28, AlbRha85) are less polymorphic than SGs. AlbFlavi24, AlbFlavi34, AlbRha12, and AlbRha28 are annotated as the only exons of AALF023281, AALF005432, AALF025780, AALF000478, respectively.

NIRVS Identified Within Coding Sequences Are Expressed

The observed LoP for AALF020122 with AlbRha52, AALF025780 with AlbRha12 and AALF005432 with AlbFlavi34 is analogous to that of rapidly evolving genes, suggesting co-option for immunity functions (Frank and Feschotte, 2017). Because domestication of exogenous sequences is a multi-step process, including persistence, immobilization and stable expression of the newly acquired sequences besides rapid evolution (Joly-Lopez and Bureau, 2018), we analyzed the distribution and expression pattern of these genes. Expression analyses were extended to all other N-Gs (AALF025779 with a unique exon containing AlbRha9, AALF000476 with a unique exon corresponding to AlbRha15, AALF000477, and AALF004130 in which the unique exons are contained within AlbRha18 and AlbRha85, respectively) that are fixed within the Foshan strain, but have LoP levels comparable to or lower than those of conserved genes (**Figure 5**). AlbFlavi34 had been previously studied and showed to be expressed in pupae and adult males more than in larvae (Palatini et al., 2017). Genes with NIRVS (N-Gs) form

TABLE 1 | Characteristics of genes with NIRVS in their coding sequence.

Gene ID	NIRVS	Viral ORF	PfamID	Median LoP
AALF000476[a]	AlbRha15	*Rhabdovirus* nucleocapsid protein	PF00945	0.0086
AALF000477[a]	AlbRha18	*Rhabdovirus* nucleocapsid protein	PF00945	0.0052
AALF000478[a,c]	AlbRha28	*Rhabdovirus* nucleocapsid protein	PF00945	0.0004
AALF025780[a]	AlbRha12	*Rhabdovirus* nucleocapsid protein	PF00945	0.0129
AALF025779[a]	AlbRha9	*Rhabdovirus* nucleocapsid protein	PF00945	0.0031
AALF004130[b]	AlbRha85	*Rhabdovirus* RNA dependent RNA polymerase	PF00946	0.0020
AALF020122[b]	AlbRha52	*Rhabdovirus* RNA dependent RNA polymerase	PF00946	0.0196
AALF005432	AlbFlavi34	*Flavivirus* NS2A, NS2B, NS3	PF00949, PF00271, PF07652	0.0099

[a]*Paralogous genes;* [b]*paralogous genes;* [c]*no expression data.*

two groups of paralogs, with similarity to the *Rhabdovirus* RNA-dependent RNA polymerase (RdRPs) and the nucleocapsid-encoding gene, respectively (**Table 1**). As shown in **Figure 6**, apart from AALF00477, all other genes are expressed throughout *Ae. albopictus* development with a similar profile, but at different levels. None of the genes showed sex-biased expression or tissue-specific expression in the ovaries; on the contrary highest expression was observed in sugar- and blood-fed females.

Additional NIRVS Variants Are Found in the Genome of Foshan Mosquitoes

We verified the presence of novel NIRVS alleles by investigating soft-clipped reads. Soft-clipped reads support the contiguity of AlbFlavi6 and AlbFlavi7, that were annotated in separated regions of the same contig (**Figure 7A**). This newly resolved arrangement revealed the existence of a viral ORF of 1191 bp, corresponding to a partial non-structural protein 5 (NS5) of *Flaviviruses*. Additionally, soft-clipped reads supported longer than annotated alleles in AlbFlavi10, AlbFlavi2 and AlbRha4 (**Figure 7B**). We further looked for the presence of novel viral integrations using Vy-PER (Forster et al., 2015). No viral integrations different than the ones identified *in silico* from the Foshan genome (Palatini et al., 2017) were found in the 16 SSMs.

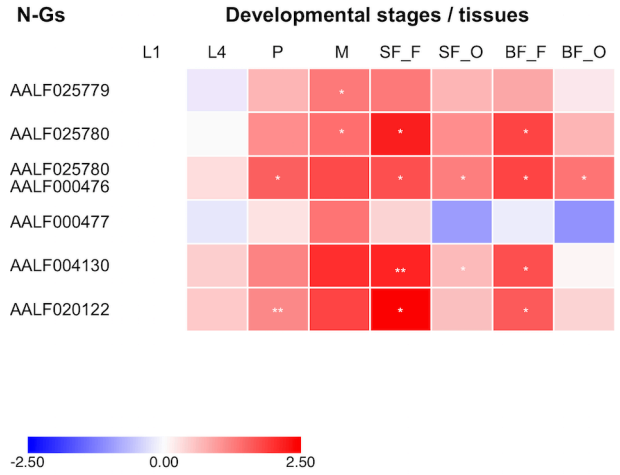

FIGURE 6 | Expression of NIRVS mapping in coding sequences. Heatmap of the expression profiles of genes with NIRVS in the coding sequence (N-Gs) across developmental stages and body tissues (L1-L4: 1st-4th instar larvae; P: pupae; M: male whole body; SF_F/BF_F: sugar/blood fed female whole body; SF_O/BF_O: ovaries from sugar/blood fed females). Color key expresses the log10-fold change relative to larva 1st instar (calibrator). Asterisks indicate significant differences in transcript abundances (Unpaired two-tailed *t*-tests, $*P < 0.05$, $**P < 0.01$).

DISCUSSION

Repetitive DNA is a major source of genome variability and there are also examples of repetitive sequences being co-opted for cellular functions (Gilbert and Feschotte, 2018). Besides having an impact on the evolution, organization and behavior of eukaryotic genomes, variations in repeat sequences or their copy numbers have been exploited for taxonomic and phylogenetic studies (Djadid et al., 2006; Wang-Sattler et al., 2007; Lammers et al., 2017). Therefore, knowledge of the repeatome assists in understanding the plasticity of eukaryotic genomes and may provide markers for population genetic studies (Goubert et al., 2016, 2017). Repetitive DNA represents 68% of the genome of *Ae. albopictus* and include dozens of still-poorly characterized sequences from non-retroviral RNA viruses (Chen et al., 2015; Palatini et al., 2017). Here we studied the widespread occurrence of NIRVS in relation to the geographic distribution of the species and the variability of NIRVS in comparison to that of mosquito genes showing slow and high evolutionary rates. We clearly show that the landscape of viral integrations is variable within and across geographic populations, with a core set of seemingly the oldest integrations from *Rhabdoviruses*. Additionally, the polymorphism of viral integrations is heterogeneous and depends primarily on their location within the genome. Overall, results of this study emphasize the complexity of the composition and structure of the mosquito repeatome and provide an objective strategy to identify viral integrations that most probably affect mosquito biology.

Biological Significance of NIRVS Variable Genomic Landscape and Their Polymorphism in SSMs

The landscape of viral integrations is variable among SSMs, longer than annotated alleles are identified through the analyses

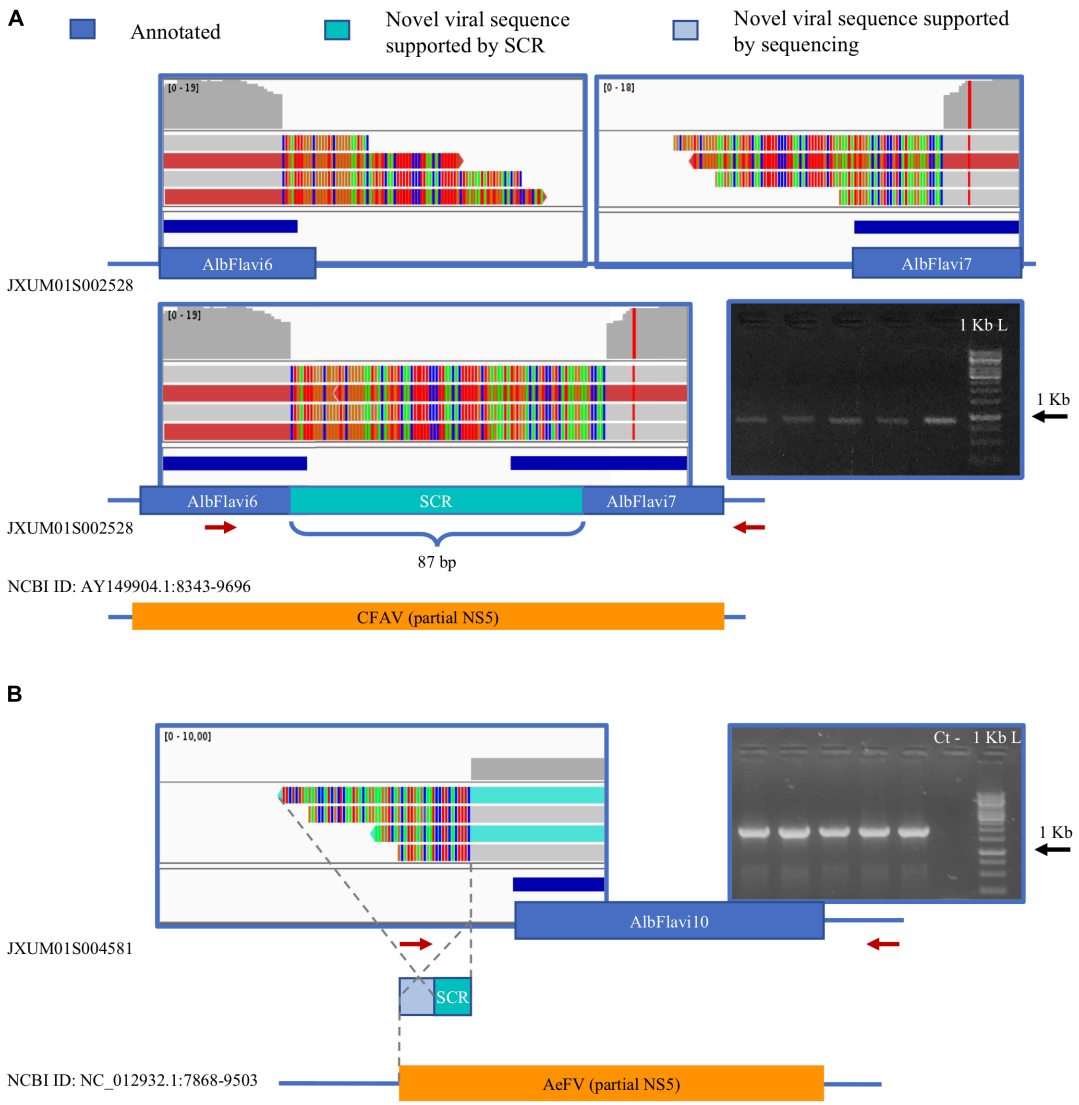

FIGURE 7 | Soft clipped reads (SCR) support novel arrangements and longer than annotated viral integrations. **(A)** AlbFlavi6 and AlbFlavi7 are annotated on the same contig, but 5000 bp apart from each other. Soft-clipped reads (in light blue) and PCR experiments support their contiguity, with a unique ORF with similarity to the *Flavivirus* NS5. **(B)** A sequence of 212 bp extending AlbFlavi10 on its left side was identified investigating SCRs and confirmed by further sequencing. Red arrows indicate primer positions.

of soft-clipped reads, but no additional viral sequences, different than the ones characterized from the Foshan-based genome assembly, are found in SSMs. NIRVS are considered rare events following viral infections. In *Aedes* spp. mosquitoes and mosquito cells short segments of cDNA of viral origin (vDNA) are synthetized upon infection with arboviruses of different genera (*i.e.*, *Flavivirus*, *Alphavirus* and *Bunyavirus*) by the reverse transcriptase of endogenous retrotransposons (Goic et al., 2016; Nag and Kramer, 2017). These vDNAs are composed of fragmented viral sequences, from different regions of the viral genome, next to sequences of TEss (Goic et al., 2016; Nag and Kramer, 2017). Because the composition of vDNAs is analogous to that of NIRVS, vDNAs have been proposed to be the substrate for NIRVS (Olson and Bonizzoni, 2017; Palatini et al., 2017). The SSMs analyzed in this study are from the Foshan strain. The Foshan strain derives from mosquitoes collected in the Chinese city of Foshan in the early '1980 and have since been kept in laboratory settings with no viral exposure (Chen et al., 2015). Under this scenario, the absence of novel viral integrations in SSMs is not unexpected. However, the identification of a variable landscape among SSMs with a core set of NIRVS, which is enriched for integrations with similarity to *Rhabdoviruses* and NIRVS mapping in coding sequences, is significant because it demonstrates that viral integrations are a dynamic component of the repeatome and not all viral integrations are dispensable genomic elements. Interestingly, when compared to fast- and slow-evolving mosquito genes, NIRVS polymorphism was not homogeneous. NIRVS identified

within piRNA clusters were less polymorphic than SGs. Selection constraints on sequences within piRNA clusters have been previously identified in both flies and mice (Chirn et al., 2015). This is despite piRNAs have an incredible sequence diversity and their biogenesis and processing do not appear to be linked to common sequences or structural motifs (Huang et al., 2017). In *D. melanogaster*, piRNA clusters are dynamic loci and their composition has been linked to their regulatory abilities. For instance, the ability of the *D. melanogaster* master piRNA locus *flamenco* to control transposons such as *gypsy*, *ZAM* and *Idefix* was shown to be dependent on frequent chromosomal rearrangements, loss or gain of fragmented TE sequences (Zanni et al., 2013; Guida et al., 2016). Additionally, variations in the composition of subtelomeric piRNA clusters were observed upon adaptation to laboratory conditions of *D. melanogaster* wild collected flies (Asif-Laidin et al., 2017). Importantly, structural differences in subtelomeric piRNA clusters did not impair host genome integrity and occurred with the maintenance of conserved groups of sequences, which could be alternatively distributed among different strains (Asif-Laidin et al., 2017). Data on the geographic distribution of NIRVS mapping in piRNA clusters studied here (i.e., AlbFlavi2, AlbFlavi4, AlbFlavi12-17, AlbRha14, and AlbRha36) show a situation analogous to that identified with TE fragments of the *flamenco* locus in *D. melanogaster*. On this basis, it is tempting to propose that the analogy between *D. melanogaster* and *Ae. albopictus* in the dynamic composition of piRNA clusters extends to their function so that the pattern of viral integrations within piRNA clusters influence mosquito susceptibly to viral infection. If proven, this hypothesis may help explain the observed variability in vector competence across mosquito populations and could be adapted into novel genetic-based strategies of vector control.

Among NIRVS encompassing gene exons, three appeared more variable than FGs and are also expressed; two of these (AlbRha52 and AlbRha12) are also persistent suggesting exaptation (Joly-Lopez and Bureau, 2018). AlbRha52 and AlbRha12 have similarity to the RdRPs and nucleocapsid-encoding genes of *Rhabdovirus*, respectively. RdRPs are ancient enzymes, essential for RNA viruses (de Farias et al., 2017). While the existence of RdRP genes in insects is still debated, cellular RdRP activity has been observed in plants, fungi and *Caenorhabditis elegans* in association with RNA silencing functions (Zong et al., 2009; de Farias et al., 2017; Pinzon et al., 2018). An RdRP of viral origin was recently described in a bat species of the *Eptesicus* clade (Horie et al., 2016) and exaptation of a viral nucleocapsid gene was shown in Afrotherians (Kobayashi et al., 2016). On this basis, further experiments to characterize the functions of the *Ae. albopictus* genes AALF020122 and AALF025780 are on-going.

Biological Significance of NIRVS Variable Genomic Landscape in Geographic Populations

To start gaining insights into the natural widespread occurrence of NIRVS, a set of 13 viral integrations representative of both R- and F-NIRVS and mapping within piRNA clusters, intergenic regions and gene exons were selected and both their occurrence and their sequence polymorphism was analyzed in mosquitoes from five geographic populations. Populations were selected following the invasion history of *Ae. albopictus* out of its native home range in south East Asia and included samples from China, Thailand, La Reunion island and newly colonized areas such as Italy and United States. Distributions of NIRVS in these populations was consistent with results from SSMs as R-NIRVS were more frequently detected than F-NIRVS. Additionally, R-NIRVS appeared on overage older integrations than F-NIRVS.

The difference in the number and age of the integration events among sequences from *Rhabdoviruses* and *Flaviviruses* is intriguing because Mononegavirales, including *Rhabdoviruses*, are considered evolutionary more recent than *Flaviviridae* (Koonin et al., 2015). The *Rhabdovirus* genus contains viruses that are extremely variable in both their genomic organization and host preferences, with viruses infecting vertebrates, invertebrates and plants (Dietzgen et al., 2017; Geoghegan et al., 2017). Additionally, *Rhabdoviruses* have been shown to frequently transfer horizontally among host species based on their ecological and geographic proximity (Geoghegan et al., 2017). Thus, the ecological diversity and the wide geographic distribution range of *Rhabdoviruses* may favor their integrations into mosquito genomes. Alternatively, the promiscuous nature of *Rhabdoviruses* with frequent horizontal transfers could select for the emergence of generalist protection mechanisms, of which integrations could be part of.

The variable landscape of NIRVS across geographic populations should be interpreted with caution. The rapid global invasion of *Ae. albopictus* from South-East Asia, which happened over the past 50–60 years, was human-mediated and occurred through the movement of propagules (Manni et al., 2017), creating a situation of genetic admixture. Mosquito populations from newly invaded areas, such as Italy and United States, lack isolation by distance and appear genetically mixed (Kotsakiozi et al., 2017; Manni et al., 2017; Maynard et al., 2017). The occurrence of frequent bottlenecks followed by interbreeding can partly explain the variable NIRVS landscape observed here. However, the enrichment for R-NIRVS, the variable distribution of NIRVS within piRNA clusters and their heterogenous polymorphism indicate that evolutionary forces other than genetic drift and gene flow have played a role in the distribution of NIRVS and suggests a multifaceted impact of NIRVS on mosquito physiology.

DATA AVAILABILITY STATEMENT

Whole Genome Sequencing data alignments have been deposited to the SRA archive under accession number from SAMN09759672 to SAMN09759687.

AUTHOR CONTRIBUTIONS

MB and EP conceived and designed the study, analyzed the data, and drafted the manuscript. EP and RW contributed to

bioinformatic analyses, analyzed the results, and revised the manuscript. FS, FV, PC, and RC-L collected and analyzed molecular data and revised the manuscript. All authors read and approved the final manuscript.

FUNDING

This research was funded by a European Research Council Consolidator Grant (ERC-CoG) under the European Union's Horizon 2020 Program (Grant Number ERC-CoG 682394) to MB, by the Italian Ministry of Education, University and Research FARE-MIUR project R1623HZAH5 to MB, by the Italian Ministry of Education, University and Research (MIUR): Dipartimenti di Eccellenza Program (2018–2022) Department of Biology and Biotechnology "L. Spallanzani," University of Pavia, and by the Swiss National Science Foundation grant PP00P3_170664 to RW.

ACKNOWLEDGMENTS

The authors thank Ruth Monica Waghchoure for mosquito maintenance and Lino Ometto for fruitful discussions. A previous version of the study was published as a preprint here: https://www.biorxiv.org/content/early/2018/08/06/385666.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00093/full#supplementary-material

REFERENCES

Arensburger, P., Hice, R. H., Wright, J. A., Craig, N. L., and Atkinson, P. W. (2011). 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* 12:606. doi: 10.1186/1471-2164-12-606

Asif-Laidin, A., Delmarre, V., Laurentie, J., Miller, W. J., Ronsseray, S., and Teysset, L. (2017). Short and long-term evolutionary dynamics of subtelomeric piRNA clusters in *Drosophila*. *DNA Res.* 24, 459–472. doi: 10.1093/dnares/dsx017

Baruffi, L., Damiani, G., Guglielmino, C. R., Bandi, C., Malacrida, A. R., and Gasperi, G. (1995). Polymorphism within and between populations of Ceratitis: comparison between RAPD and multilocus enzyme electrophoresis data. *Heredity* 74, 425–437. doi: 10.1038/hdy.1995.60

Belyi, V. A., Levine, A. J., and Skalka, A. M. (2010). Unexpected inheritance: multiple integrations of ancient bornavirus and ebolavirus/marburgvirus sequences in vertebrate Genomes. *PLoS Pathog.* 6:e1001030. doi: 10.1371/journal.ppat.1001030

Bernhardt, S. A., Simmons, M. P., Olson, K. E., Beaty, B. J., Blair, C. D., and Black, W. C. (2012). Rapid intraspecific evolution of miRNA and siRNA Genes in the mosquito *Aedes aegypti*. *PLoS One* 7:e44198. doi: 10.1371/journal.pone.0044198

Boston, M. (2015). *R Studio*. Available at: https://www.rstudio.com

Brennecke, J., Aravin, A. A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., et al. (2007). Discrete small RNA-Generating loci as master regulators of transposon activity in *Drosophila*. *Cell* 128, 1089–1103. doi: 10.1016/j.cell.2007.01.043

Bubner, B., and Baldwin, I. T. (2004). Use of real-time PCR for determining copy number and zygosity in transgenic plants. *Plant Cell Rep.* 23, 263–271. doi: 10.1007/s00299-004-0859-y

Carrillo-Tripp, M., Shepherd, C. M., Borelli, I. A., Venkataraman, S., Lander, G., Natarajan, P., et al. (2009). VIPERdb 2: an enhanced and web API enabled relational database for structural virology. *Nucleic Acids Res.* 37, 436–442. doi: 10.1093/nar/gkn840

Chen, X.-G., Jiang, X., Gu, J., Xu, M., Wu, Y., Deng, Y., et al. (2015). Genome sequence of the Asian Tiger mosquito, *Aedes albopictus*, reveals insights into its biology, genetics and evolution. *Proc. Natl. Acad. Sci. U.S.A.* 112, E5907–E5915. doi: 10.1073/pnas.1516410112

Chirn, G. W., Rahman, R., Sytnikova, Y. A., Matts, J. A., Zeng, M., Gerlach, D., et al. (2015). Conserved piRNA expression from a distinct set of piRNA cluster loci in Eutherian mammals. *PLoS Genet.* 11:e1005652. doi: 10.1371/journal.pgen.1005652

Crochu, S., Cook, S., Attoui, H., Charrel, R. N., De Chesse, R., Belhouchet, M., et al. (2004). Sequences of flavivirus-related RNA viruses persist in DNA form integrated in the genome of Aedes spp. *mosquitoes*. *J. Gen. Virol.* 85, 1971–1980. doi: 10.1099/vir.0.79850-0

de Farias, S. T., dos Santos, A. P. Jr., Rêgo, T. G., and José, M. V. (2017). Origin and evolution of RNA-dependent RNA Polymerase. *Front. Genet.* 8:125. doi: 10.3389/fgene.2017.00125

de Lamballerie, X., Crochu, S., Billoir, F., Neyts, J., de Micco, P., Holmes, E. C., et al. (2002). Genome sequence analysis of Tamana bat virus and its relationship with the genus *Flavivirus*. *J. Gen. Virol.* 83, 2443–2454. doi: 10.1099/0022-1317-83-10-2443

Dietzgen, R. G., Kondob, H., Goodinc, M. M., Kurathd, G., and Vasilakise, N. (2017). The family *Rhabdoviridae*: mono - and bipartite negative-sense RNA viruses with diverse genome organization and common evolutionary origins. *Virus Res.* 227, 158–170. doi: 10.1016/j.virusres.2016.10.010.

Djadid, N. D., Gholizadeh, S., Aghajari, M., Zehi, A. H., Raeisi, A., and Zakeri, S. (2006). Genetic analysis of rDNA-ITS2 and RAPD loci in field populations of the malaria vector, Anopheles stephensi (Diptera: Culicidae): implications for the control program in Iran. *Acta Trop.* 97, 65–74. doi: 10.1016/j.actatropica.2005.08.003

Edgar, R. C., Drive, R. M., and Valley, M. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Forster, M., Szymczak, S., Ellinghaus, D., Hemmrich, G., Rühlemann, M., Kraemer, L., et al. (2015). Vy-PER: eliminating false positive detection of virus integration events in next generation sequencing data. *Sci. Rep.* 5:11534. doi: 10.1038/srep11534

Fort, P., Albertini, A., Van-Hua, A., Berthomieu, A., Roche, S., Delsuc, F., et al. (2012). Fossil rhabdoviral sequences integrated into arthropod genomes: ontogeny, evolution, and potential functionality. *Mol. Biol. Evol.* 29, 381–390. doi: 10.1093/molbev/msr226

Frank, J. A., and Feschotte, C. (2017). Co-option of endogenous viral sequences for host cell function. *Curr. Opin. Virol.* 25, 81–89. doi: 10.1016/j.coviro.2017.07.021

Gainetdinov, I., Skvortsova, Y., Kondratieva, S., Funikov, S., and Azhikina, T. (2017). Two modes of targeting transposable elements by piRNA pathway in human testis. *RNA* 23, 1614–1625. doi: 10.1261/rna.060939.117

Garrison, E., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv*:1207.3907 [Preprint].

Geoghegan, J. L., Duchêne, S., and Holmes, E. C. (2017). Comparative analysis estimates the relative frequencies of co-divergence and cross-species transmission within viral families. *PLoS Pathog.* 13:e1006215. doi: 10.1371/journal.ppat.1006215

Gilbert, C., and Feschotte, C. (2018). Horizontal acquisition of transposable elements and viral sequences: patterns and consequences. *Curr. Opin. Genet. Dev.* 49, 15–24. doi: 10.1016/j.gde.2018.02.007

Goic, B., Stapleford, K. A., Frangeul, L., Doucet, A. J., Gausson, V., Blanc, H., et al. (2016). Virus-derived DNA drives mosquito vector tolerance to arboviral infection. *Nat. Commun.* 7:12410. doi: 10.1038/ncomms12410

Goubert, C., Henri, H., Minard, G., Valiente Moro, C., Mavingui, P., Vieira, C., et al. (2017). High-throughput sequencing of transposable element insertions suggests adaptive evolution of the invasive Asian tiger mosquito towards temperate environments. *Mol. Ecol.* 26, 3968–3981. doi: 10.1111/mec.14184

Goubert, C., Minard, G., Vieira, C., and Boulesteix, M. (2016). Population genetics of the Asian tiger mosquito Aedes albopictus, an invasive vector of human diseases. *Heredity* 117, 125–134. doi: 10.1038/hdy.2016.35

Guida, V., Cernilogar, F. M., Filograna, A., De Gregorio, R., Ishizu, H., Siomi, M. C., et al. (2016). Production of small noncoding RNAs from the *flamenco*locus is regulated by the *gypsy*retrotransposon of *Drosophila melanogaster*. *Genetics* 204, 631–644. doi: 10.1534/genetics.116.187922

Guzzardo, P. M., Muerdter, F., and Hannon, G. J. (2013). The piRNA pathway in flies: highlights and future directions. *Curr. Opin. Genet. Dev.* 23, 44–52. doi: 10.1016/j.gde.2012.12.003

Haag-Liautard, C., Dorris, M., Maside, X., Macaskill, S., Halligan, D. L., Charlesworth, B., et al. (2007). Direct estimation of per nucleotide and genomic deleterious mutation rates in *Drosophila*. *Nature* 445, 82–85. doi: 10.1038/nature05388

Hellemans, J., Mortier, G., De Paepe, A., Speleman, F., and Vandesompele, J. (2007). qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data. *Genome Biol.* 8:R19. doi: 10.1186/gb-2007-8-2-r19

Holmes, E. C. (2011). The evolution of endogenous viral elements. *Cell Host Microbe* 10, 368–377. doi: 10.1016/j.chom.2011.09.002

Horie, M., Kobayashi, Y., Honda, T., Fujino, K., Akasaka, T., Kohl, C., et al. (2016). An RNA-dependent RNA polymerase gene in bat genomes derived from an ancient negative-strand RNA virus. *Sci. Rep.* 6:25873. doi: 10.1038/srep25873

Huang, X., Fejes Tóth, K., and Aravin, A. A. (2017). piRNA biogenesis in *Drosophila melanogaster*. *Trends Genet.* 33, 882–894. doi: 10.1016/j.tig.2017.09.002

Joly-Lopez, Z., and Bureau, T. E. (2018). Exaptation of transposable element coding sequences. *Curr. Opin. Genet. Dev.* 49, 34–42. doi: 10.1016/j.gde.2018.02.011

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). The rapid generation of mutation data matricies from protein sequences. *Comput. Appl. Biosci.* 8, 275–282. doi: 10.1093/bioinformatics/8.3.275

Katzourakis, A., and Gifford, R. J. (2010). Endogenous viral elements in animal genomes. *PLoS Genet.* 6:e1001191. doi: 10.1371/journal.pgen.1001191

Keightley, P. D., Trivedi, U., Thomson, M., Oliver, F., Kumar, S., and Blaxter, M. L. (2009). Analysis of the genome sequences of three *Drosophila melanogaster*spontaneous mutation accumulation lines. *Genome Res.* 19, 1195–1201. doi: 10.1101/gr.091231.109

Kobayashi, Y., Horie, M., Nakano, A., Murata, K., Itou, T., and Suzuki, Y. (2016). Exaptation of bornavirus-Like nucleoprotein elements in Afrotherians. *PLoS Pathog.* 12:e1005785. doi: 10.1371/journal.ppat.1005785

Koonin, E. V., Dolja, V. V., and Krupovic, M. (2015). Origins and evolution of viruses of eukaryotes: the ultimate modularity. *Virology* 47, 2–25. doi: 10.1016/j.virol.2015.02.039

Kotsakiozi, P., Richardson, J. B., Pichler, V., Favia, G., Martins, A. J., Urbanelli, S., et al. (2017). Population genomics of the Asian tiger mosquito, *Aedes albopictus*: insights into the recent worldwide invasion. *Ecol. Evol.* 7, 10143–10157. doi: 10.1002/ece3.3514

Kryukov, K., Ueda, M. T., Imanishi, T., and Nakagawa, S. (2018). Systematic survey of non-retroviral virus-like elements in eukaryotic genomes. *Virus Res.* [Epub ahead of print] doi: 10.1016/j.virusres.2018.02.002,

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for Bigger Datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi: 10.1093/molbev/msw054

Lai, Z., Markovets, A., Ahdesmaki, M., Chapman, B., Hofmann, O., Mcewen, R., et al. (2016). VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res.* 44:e108. doi: 10.1093/nar/gkw227

Lammers, F., Gallus, S., Janke, A., and Nilsson, M. A. (2017). Phylogenetic conflict in bears identified by automated discovery of transposable element insertions in low-coverage genomes. *Genome Biol. Evol.* 9, 2862–2878. doi: 10.1093/gbe/evx170

Langella, O. (1999). *Population 1.2.31*. Available at: http://bioinformatics.org/populations/

Le, S. Q., and Gascuel, O. (2008). An improved general amino acid replacement matrix. *Mol. Biol. Evol.* 25, 1307–1320. doi: 10.1093/molbev/msn067

Liu, P., Dong, Y., Gu, J., Puthiyakunnon, S., Wu, Y., and Chen, X. G. (2016). Developmental piRNA profiles of the invasive vector mosquito *Aedes albopictus*. *Parasit. Vectors* 9:524. doi: 10.1186/s13071-016-1815-8

Manni, M., Guglielmino, C. R., Scolari, F., Vega-Rúa, A., Failloux, A. B., Somboon, P., et al. (2017). Genetic evidence for a worldwide chaotic dispersion pattern of the arbovirus vector, *Aedes albopictus*. *PLoS Negl. Trop. Dis.* 11:e0005332. doi: 10.1371/journal.pntd.0005332

Maumus, F., and Quesneville, H. (2014). Deep investigation of *Arabidopsis thaliana*junk DNA reveals a continuum between repetitive elements and genomic dark matter. *PLoS One* 9:e94101. doi: 10.1371/journal.pone.0094101

Maumus, F., and Quesneville, H. (2016). Impact and insights from ancient repetitive elements in plant genomes. *Curr. Opin. Plant Biol.* 30, 41–46. doi: 10.1016/j.pbi.2016.01.003

Maynard, A. J., Ambrose, L., Cooper, R. D., Chow, W. K., Davis, J. B., Muzari, M. O., et al. (2017). Tiger on the prowl: invasion history and spatio-temporal genetic structure of the Asian tiger mosquito *Aedes albopictus*(Skuse 1894) in the Indo-Pacific. *PLoS Negl. Trop. Dis.* 11:e0005546. doi: 10.1371/journal.pntd.0005546

Mckenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit?: a MapReduce framework for analyzing next-generation DNA sequencing data the genome analysis toolkit?: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110

Miesen, P., Joosten, J., and van Rij, R. P. (2016). PIWIs go viral: arbovirus-derived piRNAs in vector mosquitoes. *PLoS Pathog.* 12:e1006017. doi: 10.1371/journal.ppat.1006017

Nag, D. K., and Kramer, L. D. (2017). Patchy DNA forms of the Zika virus RNA genome are generated following infection in mosquito cell cultures and in mosquitoes. *J. Gen. Virol.* 98, 2731–2737. doi: 10.1099/jgv.0.000945

Okonechnikov, K., Golosova, O., Fursov, M., Varlamov, A., Vaskin, Y., Efremov, I., et al. (2012). Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* 28, 1166–1167. doi: 10.1093/bioinformatics/bts091

Olson, K. E., and Bonizzoni, M. (2017). Nonretroviral integrated RNA viruses in arthropod vectors: an occasional event or something more? *Curr. Opin. Insect Sci.* 22, 45–53. doi: 10.1016/j.cois.2017.05.010

Palatini, U., Miesen, P., Carballar-Lejarazu, R., Ometto, L., Rizzo, E., Tu, Z., et al. (2017). Comparative genomics shows that viral integrations are abundant and express piRNAs in the arboviral vectors *Aedes aegypti*and *Aedes albopictus*. *BMC Genomics* 18:512. doi: 10.1186/s12864-017-3903-3

Petit, M., Mongelli, V., Frangeul, L., Blanc, H., Jiggins, F., and Saleh, M.-C. (2016). piRNA pathway is not required for antiviral defense in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. U.S.A.* 113, E4218–E4227. doi: 10.1073/pnas.1607952113

Pfaffl, M. W. (2006). "Relative quantification' in real-time PCR," in *"Relative Quantification" in Real-Time PCR*, ed. T. Dorak (La Jolla, CA: International University Line), 63–82.

Pinzon, N., Bertrand, S., Subirana, L., Busseau, I., Escriva, H., and Seitz, H. (2018). Functional lability of RNA-dependent RNA polymerases in animals. *bioRXiv* [Preprint]. doi: 10.1101/339820

Reynolds, J. A., Poelchau, M. F., Rahman, Z., Armbruster, P. A., and Denlinger, D. L. (2012). Transcript profiling reveals mechanisms for lipid conservation during diapause in the mosquito, *Aedes albopictus*. *J. Insect Physiol.* 58, 966–973. doi: 10.1016/j.jinsphys.2012.04.013

Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., and Twigg, S. R. F. (2014). Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* 46, 912–918. doi: 10.1038/ng.3036

Rozen, S., and Skaletksy, H. (2000). Primer3 on the WWW for general uses and for biologist programmers. *Methods Mol. Biol.* 132, 365–386.

RStudio Team (2015). *RStudio: Integrated Development for R*. Boston, MA: RStudio, Inc.

Tamura, K., Battistuzzi, F. U., Billing-Ross, P., Murillo, O., Filipski, A., and Kumar, S. (2012). Estimating divergence times in large molecular phylogenies. *Proc. Natl. Acad. Sci. U.S.A.* 109, 19333–19338. doi: 10.1073/pnas.1213199109

Tóth, K. F., Pezic, D., Stuwe, E., and Webster, A. (2016). The piRNA pathway guards the germline genome against transposable elements. *Adv. Exp. Med. Biol.* 886, 51–77. doi: 10.1007/978-94-017-7417-8_4

Wang, Y., Jin, B., Liu, P., Li, J., Chen, X., and Gu, J. (2018). PiRNA profiling of dengue virus type 2-infected Asian tiger mosquito and midgut tissues. *Viruses* 10:E213. doi: 10.3390/v10040213

Wang-Sattler, R., Blandin, S., Ning, Y., Blass, C., Dolo, G., Touré, Y. T., et al. (2007). Mosaic genome architecture of the anopheles gambiae species complex. *PLoS One* 2:e1249. doi: 10.1371/journal.pone.0001249

Whelan, S., and Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* 18, 691–699. doi: 10.1093/oxfordjournals.molbev.a003851

Whitfield, Z. J., Dolan, P. T., Kunitomi, M., Tassetto, M., Seetin, M. G., Oh, S., et al. (2017). The diversity, structure, and function of heritable adaptive immunity sequences in the *Aedes aegypti*genome. *Curr. Biol.* 27, 3511–3519. doi: 10.1016/j.cub.2017.09.067

Yamada, K. D., Tomii, K., and Katoh, K. (2016). Application of the MAFFT sequence alignment program to large data - Reexamination of the usefulness of chained guide trees. *Bioinformatics* 32, 3246–3251. doi: 10.1093/bioinformatics/btw412

Yuan, J. S., Burris, J., Stewart, N. R., Mentewab, A., and Stewart, C. N. (2007). Statistical tools for transgene copy number estimation based on real-time PCR. *BMC Bioinformatics* 8:S6. doi: 10.1186/1471-2105-8-S7-S6

Zanni, V., Eymery, A., Coiffet, M., Zytnicki, M., and Luyten, I. (2013). Retrotransposons at the *flamenco*locus reflect the regulatory properties of piRNA clusters. *Proc. Natl. Acad. Sci. U.S.A.* 110, 19842–19847. doi: 10.1073/pnas.1313677110

Zdobnov, E. M., Tegenfeldt, F., Kuznetsov, D., Waterhouse, R. M., Simao, F. A., Ioannidis, P., et al. (2016). OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Res.* 45, D744–D749. doi: 10.1093/nar/gkw1119

Zong, J., Yao, X., Yin, J., Zhang, D., and Ma, H. (2009). Evolution of the RNA-dependent RNA polymerase (RdRP) genes: duplications and possible losses before and after the divergence of major eukaryotic groups. *Gene* 447, 29–39. doi: 10.1016/j.gene.2009.07.004

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.