Check for updates

Received: 28 October 2020 | Revised: 5 January 2021 | Accepted: 7 January 2021

DOI: 10.1111/mec.15798

ORIGINAL ARTICLE

MOLECULAR ECOLOGY WILEY

Population genomics in the arboviral vector *Aedes aegypti* reveals the genomic architecture and evolution of endogenous viral elements

Cristina M. Crava[1] | Finny S. Varghese[2] | Elisa Pischedda[1] | Rebecca Halbach[2] | Umberto Palatini[1] | Michele Marconcini[1] | Leila Gasmi[1] | Seth Redmond[3] | Yaw Afrane[4] | Diego Ayala[5] | Christophe Paupy[5] | Rebeca Carballar-Lejarazu[1] | Pascal Miesen[2] | Ronald P. van Rij[2] | Mariangela Bonizzoni[1]

[1]Department of Biology and Biotechnology, University of Pavia, Pavia, Italy

[2]Department of Medical Microbiology, Radboud University Medical Center, Radboud Institute for Molecular Life Sciences, Nijmegen, The Netherlands

[3]Institute of Vector Borne Disease, Monash University, Australia

[4]Department of Medical Microbiology, University of Ghana, Accra, Ghana

[5]MIVEGEC, Univ. Montpellier, IRD, CNRS, Montpellier, France

Correspondence
Mariangela Bonizzoni, Department of Biology and Biotechnology, University of Pavia, via Ferrata 9, 27100 Pavia, Italy.
Email: m.bonizzoni@unipv.it

Present address
Cristina M. Crava, Institute of Biotechnology and Biomedicine, Universitat de València, Burjassot, Spain
Rebeca Carballar-Lejarazu, Department of Molecular Biology and Biochemistry, University of California at Irvine, Irvine, CA, USA

FUNDING INFORMATION
This research was funded by a Human Frontier Science Program Research grant (RGP0007/2017) to R.v.R. and M.B.; by the Italian Ministry of Education, University and Research FARE-MIUR project R1623HZAH5 to M.B.; by a European Research Council Consolidator Grant (ERC-CoG) under the European Union's Horizon 2020 Programme (Grant No. ERC-CoG 682394) to M.B.; by a European Research Council Consolidator Grant (ECR-CoG) under the European Union's Seventh Framework Programme (ERC-CoG 615680) to R.v.R.; by a VICI grant from the Netherlands Organization for Scientific Research (NWO, grant no. 016.VICI.170.090) to R.v.R.; and by the Italian Ministry of Education, University and Research (MIUR): Dipartimenti Eccellenza Program (2018–2022) to the Department of Biology and Biotechnology "L. Spallanzani," University of Pavia.

Abstract

Horizontal gene transfer from viruses to eukaryotic cells is a pervasive phenomenon. Somatic viral integrations are linked to persistent viral infection whereas integrations into germline cells are maintained in host genomes by vertical transmission and may be co-opted for host functions. In the arboviral vector *Aedes aegypti*, an endogenous viral element from a nonretroviral RNA virus (nrEVE) was shown to produce PIWI-interacting RNAs (piRNAs) to limit infection with a cognate virus. Thus, nrEVEs may constitute a heritable, sequence-specific mechanism for antiviral immunity, analogous to piRNA-mediated silencing of transposable elements. Here, we combine population genomics and evolutionary approaches to analyse the genomic architecture of nrEVEs in *A. aegypti*. We conducted a genome-wide screen for adaptive nrEVEs and searched for novel population-specific nrEVEs in the genomes of 80 individual wild-caught mosquitoes from five geographical populations. We show a dynamic landscape of nrEVEs in mosquito genomes and identified five novel nrEVEs derived from two currently circulating viruses, providing evidence of the environmental-dependent

Ronald P. van Rij and Mariangela Bonizzoni jointly supervised the work

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
© 2021 The Authors. Molecular Ecology published by John Wiley & Sons Ltd.

modification of a piRNA cluster. Overall, our results show that virus endogenization events are complex with only a few nrEVEs contributing to adaptive evolution in *A. aegypti*.

KEYWORDS
Aedes aegypti, endogenous viral elements, mosquito genomes, piRNA cluster

1 | INTRODUCTION

Horizontal gene transfer (HGT) provides gene flow between evolutionary remote lineages, which may be instrumental in the adaptation of the recipient organism to novel ecological niches (Keeling & Palmer, 2008). In this context, virus–eukaryote HGT is particularly interesting as a growing body of genomics studies provides evidence that virus-to-host HGT occurs frequently (Gilbert & Cordaux, 2017). Endogenization of viral sequences into germline cells leads to the vertical inheritance of the inserted DNA in the host genome. These new genomic features are expected to be mostly deleterious and thus eliminated in a few generations (Aswad & Katzourakis, 2012). Still, a small proportion of endogenous viral elements (EVEs) are co-opted by the host, allowing the recipient species to restrict and better tolerate cognate viral infections (Blair et al., 2020; Ophinni et al., 2019). In mammals, the best-known EVE-based antiviral mechanism is based on the direct interaction between EVE-encoded proteins and viral or host proteins that control viral replication (Frank & Feschotte, 2017). In insects, EVEs from nonretroviral RNA viruses (nrEVEs) are transcribed as precursors of antiviral PIWI-interacting RNAs (piRNAs) (Arensburger et al., 2011; Palatini et al., 2017; Russo et al., 2019; Suzuki et al., 2017, 2020; Tassetto et al., 2019; ter Horst et al., 2019; Whitfield et al., 2017).

piRNAs are a class of small RNAs of around 25–30 nt that guide PIWI proteins onto complementary target RNAs, resulting in gene silencing at the transcriptional or post-transcriptional level (Ozata et al., 2019). In the model organism *Drosophila melanogaster*, piRNA precursors are generated from genomic loci called piRNA clusters (Czech & Hannon, 2016). These regions are enriched for (remnants of) sequences of transposable elements (TEs) and, consequently, cluster-derived piRNAs show sequence complementarity to TEs. piRNA clusters have been proposed to act as traps for new TE invasions, and the composition of piRNA clusters thus reflects the history of past TE mobilization (Brennecke et al., 2007; Khurana et al., 2011; Parhad & Theurkauf, 2019).

Cluster-derived piRNA precursors are processed into primary piRNAs, which associate with the PIWI protein Aubergine in the cytoplasm and direct cleavage of sense TE RNAs (Brennecke et al., 2007). The cleaved fragments are processed into secondary piRNAs, which associate with the PIWI protein Ago3 and direct cleavage of piRNA precursors, leading to the so-called ping-pong amplification cycle (Brennecke et al., 2007). Due to the bias for uridine at the first nucleotide of primary piRNAs, secondary piRNAs are enriched for an adenine at the tenth nucleotide, referred to as the 1U/10A bias for pong-pong amplified piRNAs (Brennecke et al., 2007).

The genome of the mosquito *Aedes aegypti*, a major arboviral vector, harbours hundreds of nrEVEs, which are mostly embedded in piRNA clusters in association with long terminal repeat (LTR) TEs and produce primary piRNAs in antisense orientation to viral RNA (Arensburger et al., 2011; Palatini et al., 2017; Russo et al., 2019; ter Horst et al., 2019; Whitfield et al., 2017). Endogenization of nonretroviral sequences probably occurs through reverse transcription of viral RNA by host LTR TEs (Tassetto et al., 2019). A recent genome engineering study demonstrated that an nrEVE derived from cell fusing agent virus (CFAV) produces piRNAs that engage in ping-pong amplification with viral target RNA to limit replication of the cognate virus in ovaries (Suzuki et al., 2020).

Based on the production of nrEVE-derived piRNAs, the physical contiguity of nrEVEs and TEs in piRNA clusters, and the recent discovery of piRNA-mediated antiviral activity of selected nrEVEs (Palatini et al., 2017; Suzuki et al., 2020; Tassetto et al., 2019; Whitfield et al., 2017), nrEVEs have been hypothesized to constitute an adaptive antiviral immune system in *A. aegypti*, functionally analogous to the defence function of the piRNA pathway against TEs in *D. melanogaster* (Blair et al., 2020; Whitfield et al., 2017). Hence, like *D. melanogaster* piRNA clusters that record the history of TE mobilization and provide a heritable source of piRNAs to silence active TEs, we expect the repertoire of nrEVEs (nrEVEome) to diversify across wild mosquitoes depending on virus exposure.

Here, we addressed this hypothesis from the bottom up: we first characterized the nrEVEome and piRNA clusters in the newest reference *A. aegypti* genome (AaegL5). We then used a genome-wide analysis to systematically identify novel nrEVEs that are not present in the reference genome by sequencing the genomes of 80 wild mosquitoes from five populations from regions endemic for arboviruses. Finally, we ran a genome-wide screen to identify reference nrEVEs probably involved in adaptive evolution, calculating three different statistics to detect candidate adaptive nrEVEs. We used these data to address three major questions: how widespread are virus endogenization events among wild-collected mosquitoes? Are nrEVEs viral fossils without function or are they maintained in the genome by adaptive evolution? What is the relationship between piRNA clusters and nrEVEs with signs of adaptive evolution?

We identified five novel nrEVEs derived from two currently circulating *A. aegypti*-infecting viruses. These novel nrEVEs occurred

both within and outside reference piRNA clusters, and with and without concomitant insertion of TEs. Our genome-wide screen identified few adaptive nrEVEs across populations. Our analysis is the first systematic search to identify novel nrEVEs in wild mosquitoes and to study adaptive evolution of nrEVEs in *A. aegypti*. Our results suggest a dynamic landscape of viral integrations in mosquito genomes and identify nrEVEs that may mediate adaptive antiviral immunity.

2 | MATERIAL AND METHODS

2.1 | nrEVE annotation in the A. aegypti genome assembly AaegL5

The latest *Aedes aegypti* genome assembly (AaegL5) (Matthews et al., 2018) was used to identify nrEVEs using iterative standalone BLASTX searches. A viral database was composed at the end of August 2018 using ssRNA, dsRNA and unclassified viral amino acid sequences available in the NCBI RefSeq viral database with host flag invertebrates plus two *Aedes* anphevirus (AeAV) sequences available from GenBank (ID AWW13507 and AWW13504) (DataS1). Initial BLASTX was run using the viral database and the *A. aegypti* genome as query with a conservative e-value of 10^{-6}. Multicopy nrEVEs were then merged using the EVE FINDER pipeline (Whitfield et al., 2017). A second BLASTX search (e-value 10^{-6}) with newly identified viral integrations as a query was run against the whole RefSeq amino acid database. Predicted viral integrations that had the highest identity to eukaryotic genes were manually discarded. Viral taxonomy was then assigned to each nrEVE based on the top hit retrieved by BLASTX searches against the nonredundant (NR) database, which was used to extract the corresponding viral family from the NCBI taxonomy database. When a viral family could not be extracted from the NCBI database, that viral integration was annotated as "unclassified."

The BED file containing all TEs annotated in the *A. aegypti* genome (AaegL5 assembly) (Matthews et al., 2018) was parsed with BEDTOOLS to find TEs overlapping each viral integration (Quinlan & Hall, 2010). Maximum observed overlap between nrEVEs and TEs was 100 bp and we removed the overlapping region from the annotation of the nrEVE sequence. BEDTOOLS was further used to identify the closest genomic feature (i.e., a TE, another nrEVE or a coding sequence) to each nrEVE. The presence of Chuviridae-like EVEs within a TE was analysed using a combination of BEDTOOLS closest and BEDTOOLS intersect.

2.2 | piRNA cluster annotation in the A. aegypti genome assembly AaegL5

Clusters were annotated on the current *A. aegypti* genome assembly (AaegL5) (Matthews et al., 2018) using two publicly available small RNAseq data sets from blood-fed female *A. aegypti* germline (ovaries) and female somatic tissues (carcasses) (SRR5961506 and SRR5961505, respectively) (Lewis et al., 2018). Adapters were clipped from reads with CUTADAPT (version 1.18) (Martin, 2011) and the reads were then mapped with BOWTIE version 1.2.2 without allowing for any mismatches over the whole length of the small RNA (Langmead et al., 2009). Therefore, ambiguously mapping reads were randomly distributed across all possible mapping positions (--best --strata -M 1 –seed 123) or discarded to retain only uniquely mapping reads associated with single-copy piRNA loci (-m 1).

piRNA clusters were annotated analogously to the approach used in *Drosophila melanogaster* (Brennecke et al., 2007), with minimal requirements adjusted for mosquito genomes. Briefly, only the first 5′ nucleotide of each piRNA sized read (23–32 nt) was used and normalized to the total number of mapped piRNAs per million (ppm) within each library to account for the higher fraction of piRNAs in germline tissues relative to other small RNA classes. The genome was scanned with nonoverlapping 5-kb sliding windows, applying and optimizing various threshold values such as piRNA density per window, unambiguity of piRNA mapping, as well as size and minimal piRNA density per cluster. For the final cluster annotation all windows with 10 or more ppm (Figure S1) and a maximum distance of 5 kb were merged into a single cluster. Clusters were required to contain at least five single-copy (unique) piRNA loci, and to be covered by at least five uniquely mapping piRNAs per million (Figure S1). The 5′ and 3′ ends of the cluster were defined by the most distant piRNAs within the merged windows. Finally, all clusters that were either very small (< 1 kb) or had a very low read coverage (average coverage < 10 ppm kb^{-1}) were filtered out (Figure S1). Initial piRNA cluster annotation was performed separately for ovary and somatic tissues to recover also clusters that are expressed in only one of soma and germline tissue, and these results were merged to reach the final cluster annotation. As the sRNA populations in the size range of 23–32 nt in the used libraries show general characteristic features of mature, PIWI-bound piRNAs (1U bias, 10 nt 5′–5′ offset, resistance to β-elimination) (Lewis et al., 2018), all reads from this population were used for our analysis. Annotation of piRNA clusters was only guided by piRNA coverage without taking into account strand asymmetry or nucleotide bias. While the latter two characteristics had been used for cluster annotation in Aag2 cells (Whitfield et al., 2017), this leads to exclusion of clusters that do not show these characteristics, such as the satellite repeat-derived piRNA cluster with important functions in embryonic development in culicine mosquitoes (Halbach et al., 2020).

Ping-pong signature for each individual piRNA cluster was evaluated with a published python script (Antoniewski, 2014), and sequence biases were plotted with GGSEQLOGO (Wagih, 2017) with the absolute number of each nucleotide per position as input. For the latter, orientation of the piRNAs was considered relative to annotated features (e.g., genes, repeats, nrEVEs), and thus piRNAs within a cluster that did not map to any feature were not included in the analysis. The z-score at position 10 as well as number of pairs underlying the score, and fraction of 1U-harbouring piRNAs and total number for antisense mapping piRNAs, as well as fraction of 10A-harbouring piRNAs and total number of sense-mapping piRNAs for each cluster is provided in Data S2.

2.3 | Analysis of piRNA production from nrEVEs

The same small RNAseq data sets (SRR5961506 and SRR5961505) used for piRNA cluster prediction were mapped to the *A. aegypti* genome (AaegL5 assembly) using BOWTIE with a minimum seed match of 18 nt. Aligned reads were filtered by length using BBMAP reformat.sh (https://sourceforge.net/projects/bbmap/), keeping only piRNA-sized reads (23–32 nt) (Czech & Hannon, 2016). BEDTOOLS Intersect (Quinlan & Hall, 2010) removed reads mapping outside annotated nrEVEs. Finally, all reads with 100% identity were collapsed with FASTX-TOOLKIT (Hannon, 2009) and used for quantification. Due to sequence similarity and overlap among nrEVEs, it is impossible to quantify reads mapping uniquely to single nrEVEs. To avoid any bias, we used custom scripts to extract piRNA-sized sequence from each original fastq file and then identified all the viral integrations to which each piRNA could be mapped. Counts for each experiment were normalized based on the library size by Quantile-to-Quantile Normalization as implemented in EDGER (McCarthy et al., 2012). For piRNAs mapping to nrEVEs, signs of overlap and ping-pong amplification were assessed and plotted using the small RNA signature tool (Antoniewski, 2014); in pairs, 10A bias was assessed with PINGPONGPRO (Uhrig & Klein, 2019).

2.4 | Geographical samples

A. aegypti mosquitoes were sampled as adults by BG-sentinel traps or as larvae in the summer and autumn of 2017 in Tapachula (Mexico), Franceville (Gabon), Larabanga (Ghana), M'barakani village near Rabai (Kenya), and Tafuna Village, Tutuila Island (America Samoa). Larvae were reared to adulthood *in situ* and ethanol-preserved adults were shipped to the University of Pavia (Italy).

2.5 | Genome sequence generation

Genomic DNA was extracted individually from each mosquito with the Promega Wizard Genomic DNA Purification Kit, according to the manufacturer's protocol. Individual DNA libraries were prepared with TruSeq DNA PCR-free reagents and sequenced to a minimum 20× coverage (average 24×) on the Illumina HiSeq X Ten platform by Macrogen to generate paired-end 150-nt reads. Raw reads were trimmed with TRIMMOMATIC version 0.38 (Bolger et al., 2014).

2.6 | Identification of novel nrEVEs

We used VY-PER followed by VIR to search for novel viral integrations (Forster et al., 2015; Pischedda et al., 2020). Because VY-PER uses BLAT (Kent, 2002) which recognizes sequences of 95% and greater sequence similarity of at least 40 bp in length, we restricted our search for novel viral integrations to 167 viral species already identified as part of the mosquito virome (Data S3). *De novo* assembled novel nrEVEs are available as Data S4.

2.7 | Mining of *A. aegypti* linked-read sequencing data

For those novel nrEVEs for which we could not identify the genome integration site by remapping the flanking regions to the *A. aegypti* genome, we investigated a set of 28 sequences, representing a broad geographical and genetic distribution of *A. aegypti*, that had been sequenced with linked-read (10×) sequencing that generated libraries in which reads that derived from a single strand of DNA are tagged with a unique barcode to map the new nrEVE in the *A. aegypti* genome (Redmond et al., 2020). Comparison of common barcodes allows inference of proximity between reads up to 80 kb apart. Each read set was aligned to novel nrEVEs to identify any reads that might comprise these viral integrations; following detection of novel nrEVEs, reads that were linked to these sequences were then aligned to the AaegL5 genome identifying the sequence flanking the viral integration. Background signal can derive from misalignment or multiple occupancy of 10× droplets; for each 1-MB window we calculated the sequence covered by flanking reads, and positions of viral integrations were determined as those within the 0.999^{th} percentile.

2.8 | Molecular analysis to confirm novel nrEVEs

Novel nrEVEs and their flanking regions were amplified by PCR (polymerase chain reaction) and Sanger-sequenced to confirm their presence in the mosquito genome. PCR was carried out with the DreamTaq Green PCR Master Mix (ThermoFisher) using 1 μl of 1:10 dilution of the DNA that had been used for next-generation sequencing. Amplified bands were purified with an ExoSAP-IT kit (ThermoFisher) and Sanger-sequenced (Macrogen). Sequences were analysed with BIOEDIT (T. Hall, 1999). After confirming the identity of each viral integration, PCR was used to analyse the distribution of these nrEVEs in the tested populations. Primers used are listed in Table S1.

2.9 | Bioinformatic pipeline to detect reference nrEVEs

The presence or absence of each viral integration characterized from the *A. aegypti* reference genome (AaegL5 assembly) was analysed in genomic resequencing of individual mosquitoes using an in-house bioinformatic pipeline (Pischedda et al., 2019). The pipeline allows us to detect the presence of nrEVEs in each tested individual, but not to distinguish between heterozygote and

homozygote status. Hence, we approximated allele frequency as the number of viral integrations normalized by the total number of individuals. Because of the stringency used in the call for nrEVE presence (Pischedda et al., 2019), a short sequence that shares sequence identity with a longer one could be erroneously called as absent in all individuals tested because of reads shared with the longer nrEVE. For this reason, a total of 29 viral integrations, which were called as absent in all individuals, were excluded from further analyses.

2.10 | Analysis of nrEVE polymorphism

nrEVE polymorphism was analysed at two levels. First, the frequency distribution of each nrEVE was analysed by investigating its presence or absence in each tested mosquito. nrEVE distribution across geographical samples was visualized using convex logistic principal components analysis (PCA) from the R package logisticPCA (Landgraf & Lee, 2015). Heterogeneity in the occurrence of nrEVEs among populations was evaluated using a maximum likelihood procedure adapted from a study on the distribution of TEs in *D. melanogaster* (González et al., 2008). Thus, the data for each nrEVE can be described as $\{m1, m2\}$ where $m1$ is the number of individuals in which an nrEVE was present, independently of its genotypic status, and $m2$ is the number of individuals in which that nrEVE is absent. The log-likelihood of observing such data conditional to the frequency p is:

$$ln(L(m1, m2|p)) = m1ln(p) + m2ln(1-p)$$

The $L(p)$ is maximized at the value $\hat{p}$:

$$\hat{p} = \frac{m1}{m1+m2}$$

To determine whether the frequencies were different among populations, a likelihood ratio test (LRT) test was used that compares two models. Under the null hypothesis, we assumed that the frequencies of viral integrations were the same in all populations and estimated $\hat{p}$ using combined data from all populations. Under H1, we assumed that nrEVE frequencies were different among populations and estimated $\hat{p}$ for each population, separately. We then calculated maximum log-likelihood for both $\hat{p}$ and they were compared as:

$$LRT = -2ln\left(\frac{Ho}{H1}\right)$$

Heterogeneity was detected when LTR was greater than 9.49 corresponding to 5% of the χ^2 test with four degrees of freedom.

Second, sequence polymorphism of each nrEVE was estimated in single individuals by analysing their single nucleotide polymorphisms (SNPs) and calculating the level of polymorphism (LoP), as previously described (Pischedda et al., 2019).

2.11 | Signature of selection

Different methods were used to test for signatures of positive selection. The presence of hard selective sweep was predicted using SWEED (Pavlidis et al., 2013) on a window size of 100 kb (50 kb upstream and 50 kb downstream of the nrEVE) using SNP and INDEL data sets called with SAMTOOLS-MPILEUP (H. Li, 2011). For each data set, the composite-likelihood ratio (CLR) was calculated over a grid of 250, which resulted in estimates over ~400 bp. CLR estimates were visualized with RSTUDIO. Candidate nrEVEs harbouring a signature of a selective sweep were selected when their CLR values were higher than the 99th percentile of their corresponding window distribution. Signatures of soft sweep were predicted using the G12 statistics implemented in the SELECTION-HAPSTAT software (Garud et al., 2015) after having identified SNP variants using FREEBAYES (Garrison & Marth, 2012). The H12_2H1.py script was run using 50 SNPs as the window size for each of the three *A. aegypti* chromosomes in each population, with overlaps of 25 bp for each window. We used small and overlapping window sizes to avoid biases from recombination; linkage disequilibrium in *A. aegypti* is estimated to 52–67 kb (Matthews et al., 2018). Windows in the top 15% most extreme G12 values were selected (Table S2) and analysed for the presence of viral integrations (Rech et al., 2019).

Selection in nrEVEs was also tested by calculating Tajima's *D* values (Tajima, 1989), a procedure previously applied to test for selection in fixed TEs (Kofler et al., 2012; Rech et al., 2019). Tajima's *D* values were calculated in nonoverlapping 500-bp windows using VCFTOOLS version 0.1.15 (Danecek et al., 2011). The approach of Rech et al. (2019) was used to identify windows with significantly low Tajima's *D* values. Average Tajima's *D* values were first calculated per chromosome and population and windows with Tajima's *D* values lower than the 5th percentile of the whole chromosome distribution were then analysed (Table S3). Finally, significant windows were screened for the presence of nrEVEs.

3 | RESULTS

3.1 | Atlas of viral integrations in the newest *Aedes aegypti* genome

We annotated the nrEVEome of the highly contiguous AaegL5 reference genome (Matthews et al., 2018) to systematically assess their role in adaptive evolution through a genome-wide screen. We annotated 252 nrEVEs from seven viral families (Flaviviridae, Rhabdoviridae, Xinmoviridae, Chuviridae, Phasmaviridae, Phenuiviridae and Mesoniviridae) plus several viruses that are still unclassified (Figure 1a; Data S5 and available at http://www.nreves.com/). Most nrEVEs derived from Rhabdoviridae (Rhabdo-nrEVEs), followed by Xinmoviridae (Xinmo-nrEVEs), Chuviridae (Chuvi-nrEVEs) and Flaviviridae (Flavi-nrEVEs). The composition of the

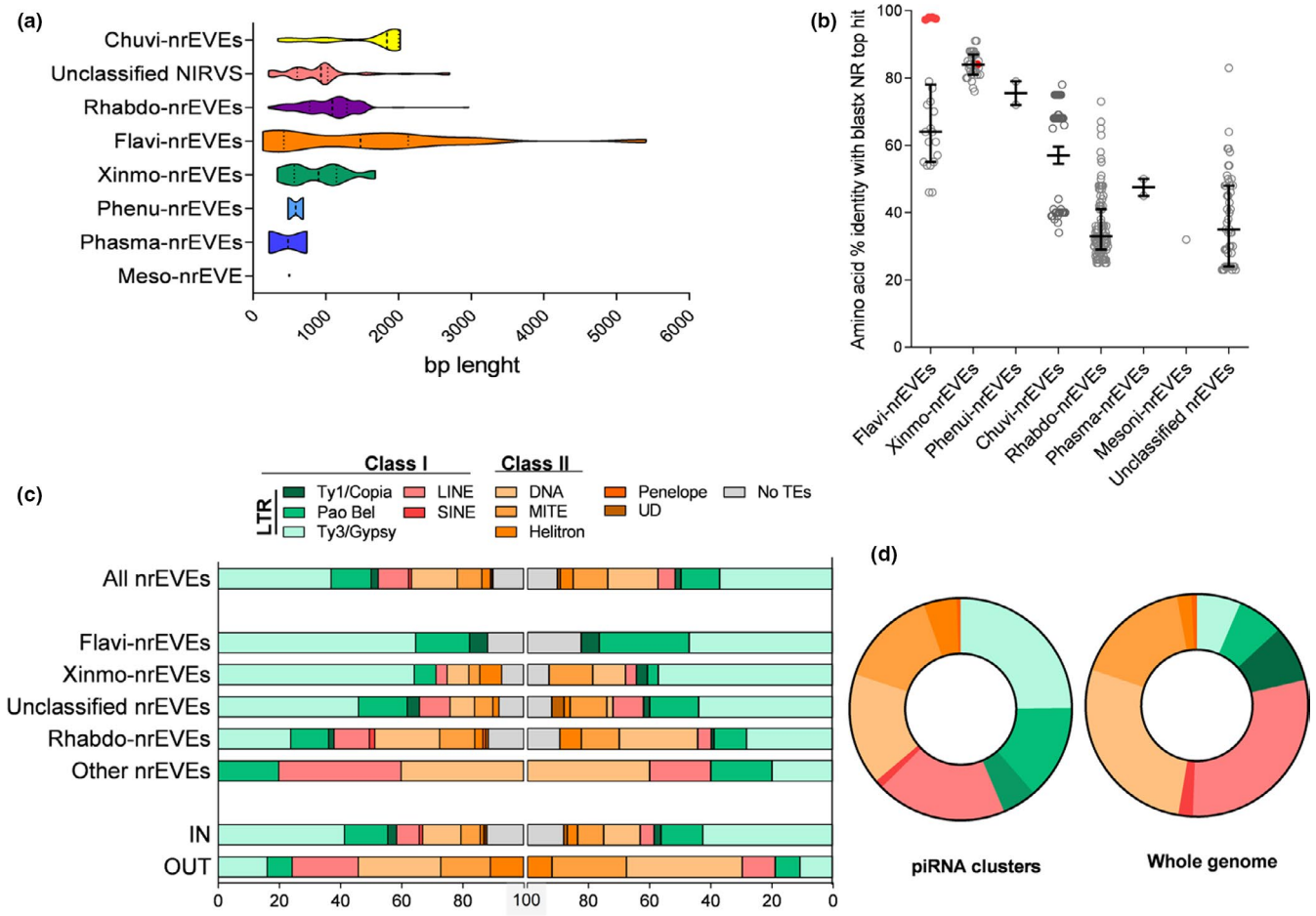


FIGURE 1 Atlas of *Aedes aegypti* nrEVEs. (a) Violin plot showing nrEVEs identified in the *A. aegypti* genome (AaegL5 assembly). (b) Scatter plot representing the amino acid identity of each nrEVE and its best hit retrieved by BLASTX searches against the NR database grouped by viral family. Whiskers represent the median and the interquartile range. Red dots are the novel nrEVEs discovered in the geographical populations. (c) Bar plots showing the type of the closest transposable element (TE) upstream and downstream of all nrEVEs (upper panel), nrEVEs grouped by their viral origin (middle panel), and nrEVEs grouped by their location within (IN) or outside (OUT) piRNA clusters. Abbreviations: LTR (long terminal repeat), UD (unclassified TEs). (d) Pie charts indicating the transposon composition of TEs in the whole genome or piRNA clusters [Colour figure can be viewed at wileyonlinelibrary.com]

AaegL5 nrEVEome resembles those obtained with earlier *A. aegypti* genome assemblies (AaegL3, Aag2) (Palatini et al., 2017; ter Horst et al., 2019; Whitfield et al., 2017), although the uneven quality of assemblies impairs a comparative analysis. nrEVEs distribute evenly on the three chromosomes, without enrichment at telomeric or centromeric regions (Figure S2).

Predicted amino acid identity of nrEVEs to the most similar virus ranges from 23% to 91%, with Xinmo-nrEVEs generally showing higher amino acid identity (average 83.9%) than Flavi-nrEVEs (68.8%), Chuvi-nrEVEs (57%), Rhadbo-nrEVEs (35.4%) and unclassified nrEVEs (38%) (Figure 1b). Low amino acid identities probably reflect ancient integrations of sequences of viruses that are now extinct but may also be due to recent integration events of currently circulating viruses that have not yet been identified. Mapping nrEVEs to a representative viral genome for each family showed that nrEVEs are not evenly distributed across viral genomes (Figure 2; Figures S3–S4). For example, Rhabdo-nrEVEs primarily originated from the nucleoprotein and glycoprotein sequences (Figure S3) whereas Flavi-nrEVEs derived mainly from regions encoding nonstructural proteins, primarily NS1 an

d NS2 (Figure 2). We also identified few nrEVEs (i.e., Fla5, Fla7, Fla16, Fla17 and Rha73) that are composed of adjacent sequences that are not consecutive in the source viral genome (hereafter, composite nrEVEs), which we hypothesize to be due to recombination and circularization events that probably occurred before integration (Tassetto et al., 2019).

nrEVEs annotated in the AaegL5 assembly are flanked by TEs, often LTR retrotransposons (Figure 1c), similar to what was observed in earlier, higher fragmented, genome assemblies (Palatini et al., 2017; Whitfield et al., 2017). Chuvi-nrEVEs are an exception as all 39 members are completely embedded within four different elements of the Bel/Pao family. Strikingly, all Chuvi-nrEVEs derive from viral glycoprotein sequences, which cluster together according to the Bel/Pao element in which they are embedded (Figure S5). LTR retrotransposons normally do not possess an envelope gene (*env*), but can acquire *env*-like genes from disparate viral sources and, through

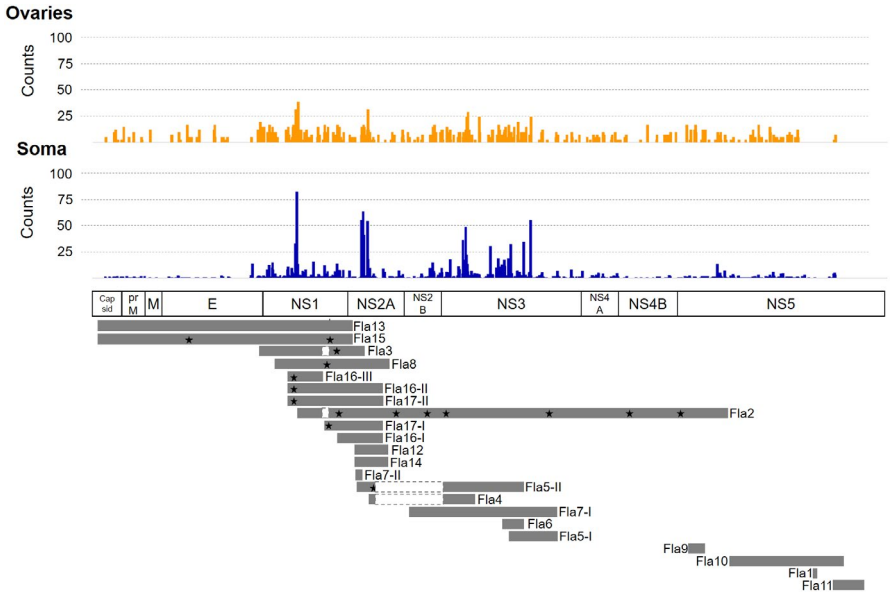


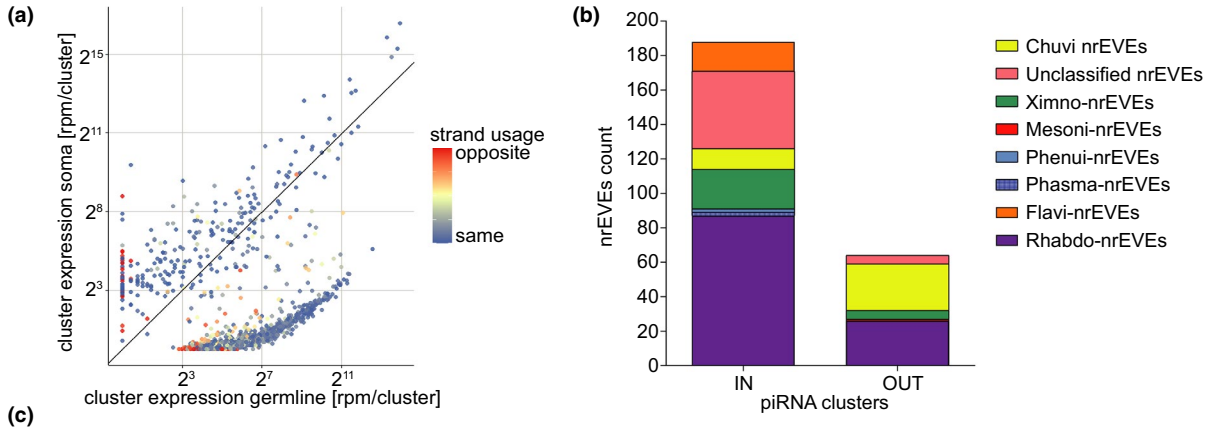
FIGURE 2 Distribution and piRNA coverage of nrEVEs on reference viral genome. Flaviviridae-derived nrEVEs (Flavi-nrEVE) aligned to the Xishuangbanna flavivirus genome (NC_034017.1). Flavi5, Flavi7, Flavi16 and Flavi17 are composed of repeated and not contiguous parts of the viral genome (composite nrEVEs) and thus have been fragmented to map to the corresponding viral sequence. Stars indicate stop codons or small indels that interrupt the viral open reading frame and dotted white boxes indicate large deletions that generate stop codons. Top panels indicate piRNAs mapping to the indicated positions in soma (orange) and ovaries (blue), respectively [Colour figure can be viewed at wileyonlinelibrary.com]

that, acquire properties of infectious retroviruses (Hayward, 2017; Malik et al., 2000). Because a feature unique to Chuvi-nrEVEs is the presence of nrEVEs that contain a whole open reading frame (ORF; Figure S6), we speculate that Bel/Pao TEs might have gained infectiousness through the acquisition of Chuviridae-derived glycoprotein sequences.

3.2 | piRNA clusters are enriched for nrEVEs and LTR retrotransposons

We systematically predicted piRNA clusters in AaegL5 with a methodology analogous to that used in *Drosophila melanogaster* (Brennecke et al., 2007) and using small RNAs from thorax (somatic) and ovaries (germline) of female *A. aegypti* (Lewis et al., 2018) (see Methods). We identified 1,158 clusters occupying less than 0.1% of the *A. aegypti* genome (Data S2). Of these, 108 and 38 clusters are predominantly expressed in germline or somatic tissues, respectively, whereas the rest are expressed in both (Figure 3a). In total, 188 nrEVEs were located within 67 piRNA clusters spanning 277 kb (2% of the total piRNA cluster length of 14 Mb) (Figure 3b), representing a strong nrEVE enrichment in piRNA clusters relative to the whole genome ($\chi^2 = 12 \times 10^6$, df = 1, $p < .00001$) as observed in other *A. aegypti* genome assemblies (Palatini et al., 2017; ter Horst et al., 2019; Whitfield et al., 2017). Five piRNA clusters harbour >10 nrEVEs (Figure 3c), mostly from different viral families, including all Flavi-nrEVEs (Figure 2). There is a clear distinction between piRNA clusters harbouring a single nrEVE and piRNA clusters with three or more nrEVES. The latter (except 1p12.1 and 2p12.17) are active in both soma and ovaries where they tend to produce piRNAs with a 1U bias from one dominant strand (uni-strand clusters) (Figure S7 (e.g., the *flamenco*-like piRNA cluster 2q44.4, Figure 3d). piRNA clusters harbouring a single nrEVE are mainly dual-strand clusters and are mostly active in the germline (e.g., piRNA cluster 3p14.2; Figure 3e), except those harbouring a single Chuvi-nrEVE, which are uni-strand and active in both tissues. In *D. melanogaster*, primary piRNAs produced by ovarian germline cells mostly derive from dual-strand piRNA clusters as opposed to uni-strand clusters in surrounding somatic cells (Théron et al., 2014). This seems to be conserved in *A. aegypti* (Figure S8), but it remains to be elucidated whether there is a link between the presence of nrEVEs in uni- or dual-strand clusters and their function.

nrEVEs are often embedded between Ty3/Gypsy and Bel/Pao elements (Figure 1c). This is possibly due to the specific incorporation of viral RNA into LTR retrotransposon replication complexes, followed by reverse transcription and recombination between transposon and virus sequences (Tassetto et al., 2019). We observed that the relationship between nrEVEs and LTR retrotransposons is linked to the location of an nrEVE within or outside a piRNA cluster. More than half (58.5%) of the TEs adjacent to nrEVEs within piRNA clusters are LTR retroelements, whereas nrEVEs are predominantly flanked by Class II transposons outside piRNA clusters (Figure 1c). Thus, the physical contiguity between nrEVEs and LTR retrotransposons can be explained by the overall enrichment of LTR elements in piRNA clusters (Fisher's exact test $p < .001$) (Figure 1d). This finding implies that integration of viral DNA can equally happen for hybrids between viral DNA and LTR retrotransposon sequences as well as for viral fragments alone.



(c)

Top ten piRNA clusters:

Name	Chr	Start	End	Size [bp]	#nrEVEs	Germline #piRNAs [rpm]	Germline #piRNAs (unique) [rpm]	Germline strand bias (+/-)[%]	Soma #piRNAs [rpm]	Soma #piRNAs (unique) [rpm]	Soma strand bias (+/-)[%]
3q23.13	3	314,095,023	314,219,346	124,323	7	9,889	4,877	100/0	59,285	31,426	100/0
2q44.4	2	461,190,266	461,374,889	184,623	23	14,191	5,652	0/100	39,797	13,592	0/100
3p23.4	3	105,760,031	106,093,134	333,103	13	11,567	6,737	18/82	29,802	13,872	6/94
3p14.9	3	158,078,314	158,083,627	5,313	0	2,692	1,801	99/1	13,556	6,182	100/0
2q44.8	2	469,140,024	469,314,324	174,300	10	3,305	1,997	100/0	9,027	5,850	100/0
3q23.10	3	313,431,501	313,504,782	73,281	7	2,864	1,934	0/100	8,389	5,962	0/100
2p42.8	2	33,865,001	33,874,989	9,988	1	553	69	0/100	6,394	705	0/100
3q33.2	3	344,847,000	344,850,087	3,087	0	5,936	488	100/0	33	26	100/0
1q31.3	1	210,430,129	210,584,987	154,858	15	3,605	1,836	1/99	2,573	1,191	0/100
1q21.5	1	91,784,999	91,844,891	59,892	11	1,328	810	99/1	3,388	1,998	100/0

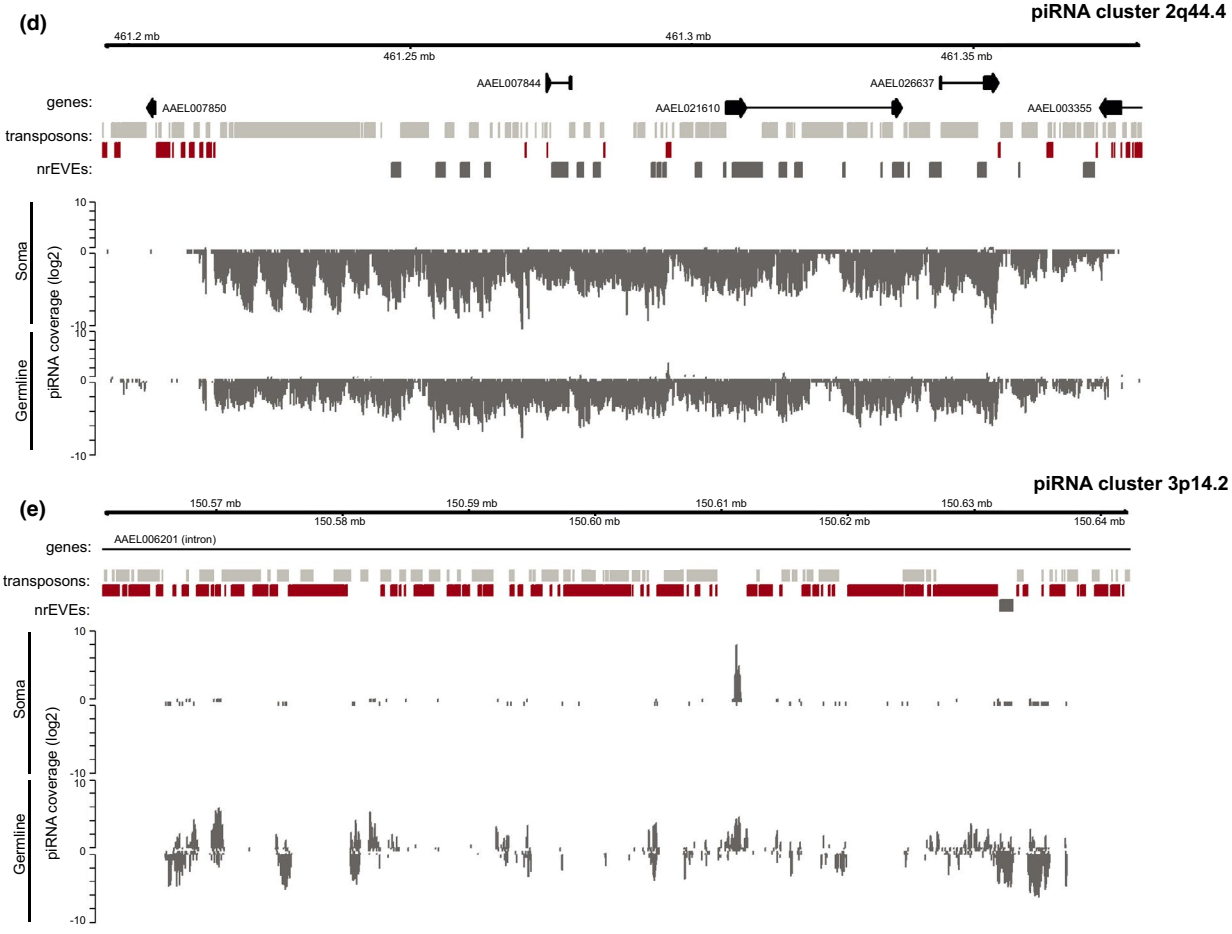


FIGURE 3 *Aedes aegypti* piRNA clusters. (a) Expression of piRNA clusters in germline and somatic tissues. piRNA coverage per million mapped small RNAs (rpm) plus a pseudo-count of 1 is plotted in order to include values of zero. Colour indicates the likelihood of a cluster being expressed with the same strand bias in both tissues. (b) Bar plots showing the distribution of nrEVEs from different viral families within (IN) and outside (OUT) piRNA clusters. (c) Table listing the top-10 most highly expressed piRNA clusters based on their highest overall expression in either germline or soma, with at least 5% uniquely mapping piRNA reads. (d) Coverage plot of a piRNA cluster with strand bias towards expression from one strand (uni-strand). (e) Coverage plot of a piRNA cluster without strong strand bias (dual-strand). Log_2 coverage in both germline (ovaries) and somatic tissues is shown. Genes are indicated with black arrows, transposons are indicated with light grey (plus strand) or red (minus strand) boxes, and nrEVEs are depicted with dark grey boxes (plus strand) [Colour figure can be viewed at wileyonlinelibrary.com]

3.3 | nrEVEs produce different piRNAs in soma and germline tissues

Previous studies revealed that nrEVEs serve as templates for piRNA production (Palatini et al., 2017; Suzuki et al., 2020; Tassetto et al., 2019; Whitfield et al., 2017). We used small RNA populations in the size range of 23–32 nt from libraries from somatic and germline tissues (ovaries), which were demonstrated to show general characteristic features of mature, PIWI-bound piRNAs (1U bias, 10-nt 5′–5′ offset, resistance to β-elimination) (Lewis et al., 2018), to identify nrEVE-derived piRNAs and compare their expression profile between tissues. These piRNAs mapped to all nrEVEs, with the exception of Chu10, Chu16, Xin23, Xin24, Rha17, Rha45, Rha60, Rha69 and Meso1, consistent with their location outside piRNA clusters.

We built a library of nrEVE-derived piRNAs consisting of 56,999 unique sequences (Data S6). The majority of nrEVE-derived piRNAs are expressed in the soma (85.1%), whereas 11.9% of nrEVE-derived piRNAs are germline exclusive, and only a small fraction (3%) are expressed in both tissues (Figure S8). Somatic nrEVE-piRNAs are largely dominated by piRNAs mapping to Flavi-nrEVEs and Xinmo-nrEVEs (51.6% and 42.6% respectively), whereas their fraction decreased to 20.1% in germline where almost half of the nrEVE piRNAs (48.3%) derive from Rhabdo-nrEVEs instead. Only germline nrEVE-derived piRNAs displayed the ping-pong amplification signature (Figure S9). Recent research showed ping-pong amplification signals in ovaries of a CFAV-derived nrEVE in the presence of cognate virus infection (Suzuki et al., 2020). We cannot exclude that some insect-specific viruses that persistently infect mosquitos, such as CFAV (Bolling et al., 2015; R. A. Hall et al., 2016), were affecting the individuals used for sRNA sequencing, albeit we did not see a clear pick of 21-nt viral siRNAs suggesting absence of active infection.

Within each viral family, there are nrEVEs that share 100% nucleotide identity, probably originating from duplications in the host genome after integration (multi-copy nrEVEs), as well as nrEVEs that correspond to the same viral region with nucleotide identity below 60%, probably representing different endogenization events. To assess whether multi-copy nrEVEs produce more piRNAs than single-copy nrEVEs, we mapped Flavi-nrEVEs and Rhabdo-nrEVEs along with their corresponding piRNAs to a representative viral sequence (Figure 2; Figure S3). We observed that piRNAs are not evenly distributed across the virus sequence covered by nrEVEs, but that they spike in distinct portions independently of the number of corresponding nrEVEs. Discontinuous piRNA expression was also observed for piRNA clusters in *Drosophila* and proposed to be defined by the local and long-range sequence context (Muerdter et al., 2012). For Flavi-nrEVEs, piRNA hotspots are in regions corresponding to flavivirus NS1, NS2 and NS3 sequences and piRNA profiles of soma and germline are roughly similar (Figure 2). In contrast, Rhabdo-EVE-derived piRNAs differ between the two tissues: germline piRNAs span the nucleoprotein (N), glycoprotein (G) and polymerase (L) sequences whereas low piRNAs levels are seen mapping to the L sequence in the soma (Figure S3).

3.4 | Endogenization of two currently circulating insect-specific viruses in wild mosquitoes

To test for virus endogenization in a natural system, we analysed the genomes of 80 individual mosquitoes sampled from five geographical populations and searched for nrEVEs that are absent from the *A. aegypti* reference nrEVEome. Samples included three populations from Africa, one from Mexico and one from American Samoa, regions with frequent arboviral outbreaks or considered high-risk zones for arboviral transmission (Cotter et al., 2018; Guerbois et al., 2016; Weetman et al., 2018).

Our genome-wide analysis identified five nrEVEs not present in the reference nrEVEome, which are variably distributed across 49 of the 80 tested genomes (Figure 4). Of these, 11 individuals concurrently harbour two novel nrEVEs. Four novel nrEVEs have similarity to the insect-specific flavivirus CFAV, with a nucleotide identity of over 97% to the Galveston reference strain (Figure 1b). The fifth novel nrEVE has 85% identity to another insect-specific virus, AeAV. Both CFAV and AeAV infections are widespread in *A. aegypti* cell lines, laboratory colonies and wild-caught mosquitoes (Baidaliuk et al., 2020; Parry & Asgari, 2018).

CFAV-derived nrEVEs correspond to different genomic regions of the viral genome and three of them (CFAV-EVE-1, CFAV-EVE-2 and CFAV-EVE-3) probably arose from recombination of CFAV sequences that are not contiguous in the CFAV genome (Figure 4a). The longest novel viral integration (i.e., CFAV-EVE-1) is a 3,909-bp sequence composed by three regions, including the complete NS1 sequence. A second novel CFAV-derived integration (CFAV-EVE-2) is 1,259 bp long and is composed of two parts of similar length that include sequences of the E protein, the glycoprotein NS1 and a central part of the NS3 sequence. A third novel CFAV-derived integration (CFAV-EVE-3) is 734 bp long and is composed of four parts (two of them inserted in

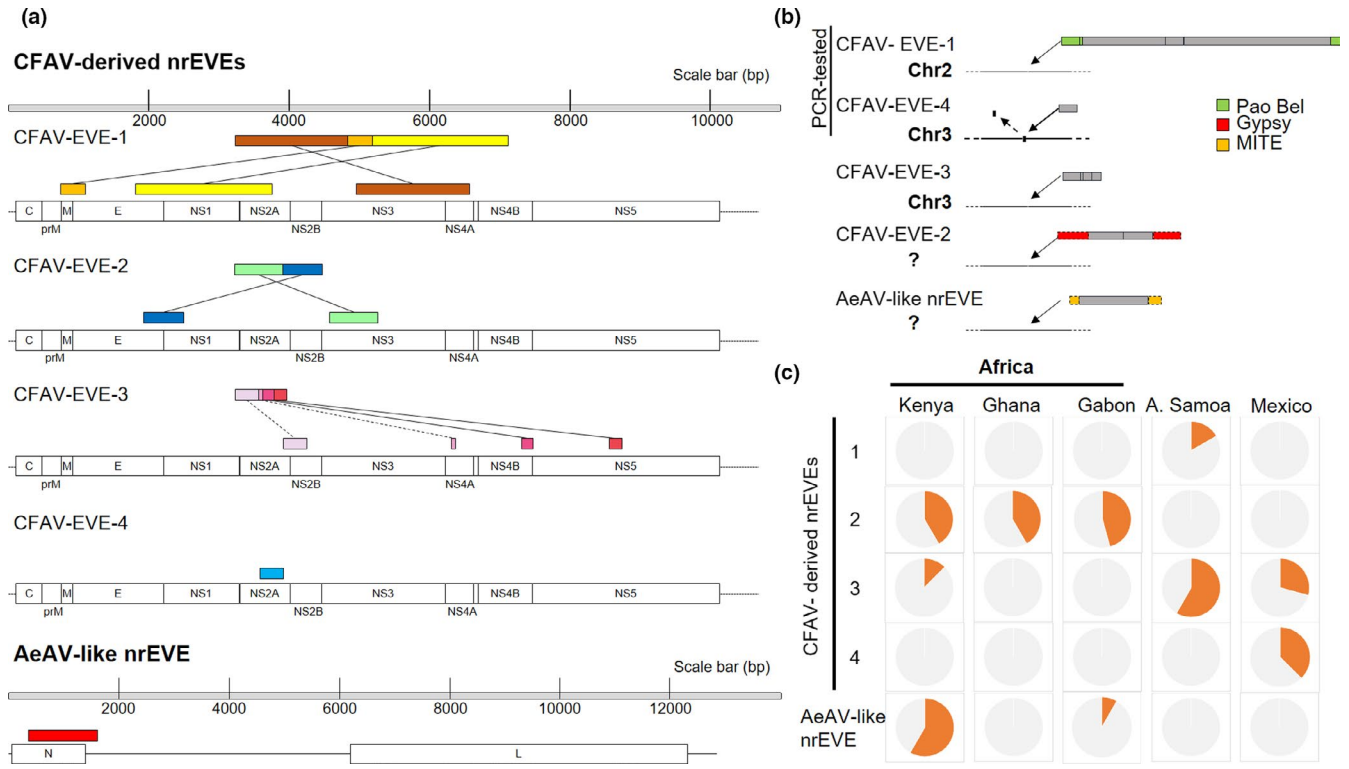


FIGURE 4 Novel viral integrations in wild-collected *Aedes aegypti* mosquitoes. (a) Scheme of the novel nrEVEs with similarity to cell fusing agent virus (CFAV) and Aedes anphevirus (AeAV) identified in the genome of wild-collected mosquitoes. CFAV-EVEs are mapped to the genome of CFAV Galveston strain (NCBI Reference Sequence: NC_001564.2) and the AeAV-like nrEVE to the genome of AeAV strain MRL-12 (MH037149.1). Dotted lines represent part of CFAV-EVE-3 that integrated in the opposite direction compared to the CFAV genome. (b) Scheme of the integration points and endogenized sequences. nrEVE sequences are represented by grey boxes; flanking TE sequences (if any) are represented with colours as indicated. Dotted TE boxes indicate flanking TEs for which we could not distinguish if they integrated together with the viral sequences or were already present in the integration point. (c) Frequency distribution of novel nrEVEs tested with PCR in 24 mosquitoes from each site (Kenya, Ghana, Gabon, American Samoa and Mexico) [Colour figure can be viewed at wileyonlinelibrary.com]

reverse orientation compared to the CFAV genome) spanning part of the NS2A, NS2B, NS4A, NS4B and NS5 coding sequences. CFAV-EVE-3 was recently identified by mining publicly available *A. aegypti* genomic and transcriptomic data, and proven to confer tolerance to CFAV infection in ovaries (Suzuki et al., 2020). The shortest CFAV-derived integration (CFAV-EVE-4) is a 328-bp sequence, corresponding to part of the NS2A coding region (Figure 4a). The novel AeAV-like integration contains a single genomic sequence that corresponds to a portion of the viral nucleoprotein (Figure 4a). None of the novel CFAV-like integrations displays polymorphism or indels among samples and populations. All novel nrEVE sequences were amplified by PCR and Sanger-sequenced, providing independent molecular validation of the bioinformatic-based identification (Figure S10).

For three of the five novel nrEVEs, chromosomal integration sites were deduced *in silico* by *de novo* assembly of sequence reads, and further confirmed by PCR and Sanger sequencing for two of them (Figure 4b). CFAV-EVE-1 is inserted in chromosome 2, at position 294,058,716, and is embedded between fragments of the LTR Bel/Pao 277 element (Matthews et al., 2018), resulting in a hybrid sequence longer than 5,900 bp. CFAV-EVE-3 is inserted in chromosome 3, at position 137,741,235. CFAV-EVE-4 is inserted between genomic positions 106,043,272 and 106,043,299 on chromosome 3 within piRNA cluster 3p23.4 (Figure 5), leading to the loss of a 27-nt sequence. Notably, CFAV-EVE-3 and CFAV-EVE-4 were generated by the insertion of viral sequences without co-integration of LTR TE fragments (Figure 4b). In Aag2 cells, fragments of Sindbis virus are reverse-transcribed during LTR TE retrotransposition, forming hybrid episomes (Tassetto et al., 2019). Our data suggest that viral RNA-derived DNA fragments can integrate into the *A. aegypti* genome either alone (such as CFAV-EVE-3 and 4) or as hybrid sequences with LTR TE fragments (such as CFAV-EVE-1).

Regions flanking CFAV-EVE-2 correspond to Gypsy245, an LTR Ty3/Gypsy TE that is present in multiple copies in AaegL5 assembly (Matthews et al., 2018). The repetitive nature of these sequences along with the lack of knowledge of the TE landscape in our samples prevented mapping of the CFAV-EVE2 integration site. To overcome this issue, we mined data from linked-read (10X Genomics) libraries (Redmond et al., 2020). This analysis identified a clean signal of a presumably single-copy viral insertion around 461 MB on chromosome 2, at the far end of the q arm, in two mosquitoes from Africa (Figure S11). Unexpectedly, in a third mosquito from Gabon, CFAV-EVE-2 is integrated in chromosome

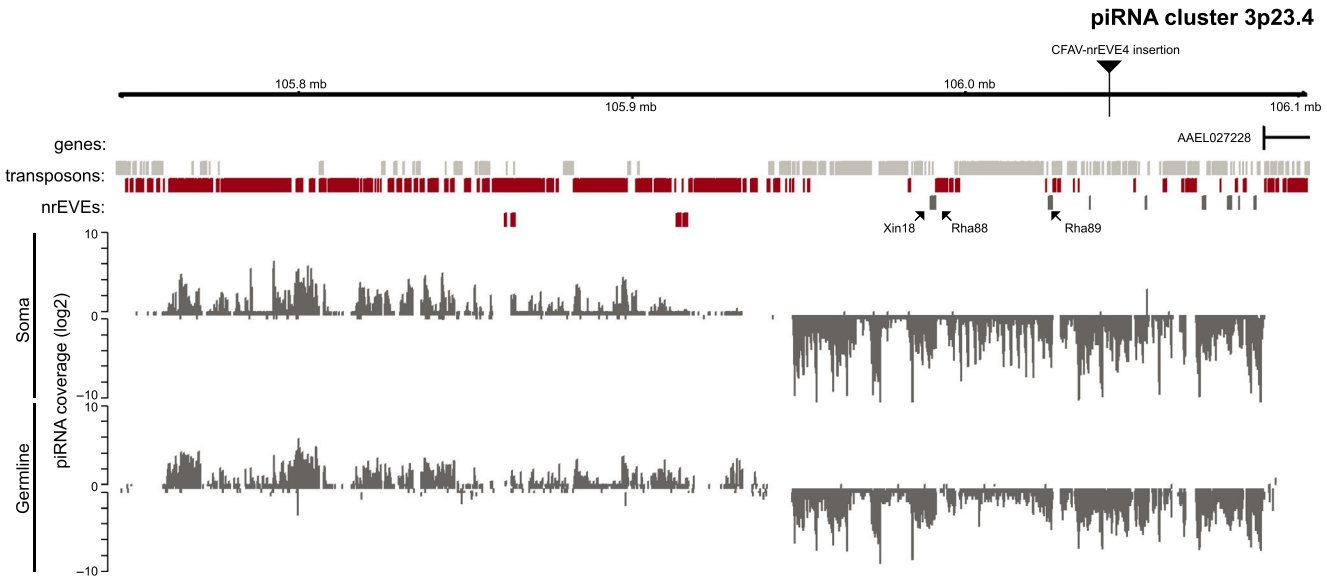


FIGURE 5 Acquisition of a novel nrEVE by a piRNA cluster. Coverage plot of piRNA cluster 3p23.4. Genes are indicated with black arrows, and transposons and nrEVEs are indicated with light and dark grey (plus strand) or light and dark red (minus strand) boxes, respectively. Arrows indicate variably distributed nrEVEs which are present only in some individual mosquitoes [Colour figure can be viewed at wileyonlinelibrary.com]

3, suggesting a possible chromosome rearrangement. The AeAV-like EVE was flanked by mTA element 38c, a DNA transposon of the miniature inverted-repeat transposable element (MITE) family that is present in multiple copies in the genome (Matthews et al., 2018). Similar to CFAV-EVE-2, we mined linked-read libraries to identify the integration site, but AeAV-like EVE sequences were not present in these libraries.

3.5 | Novel viral integrations are mostly population-specific

Novel nrEVEs displayed a population-specific pattern. CFAV-EVE-2 and AevAV-like EVE were present only in samples from Africa (the latter only in samples from Gabon and Kenya). CFAV-EVE-1 was exclusively detected in individuals collected in American Samoa, and CFAV-EVE-4 was only found in mosquitoes from Mexico (Figure 4c). In contrast, CFAV-EVE-3 was present in multiple populations spanning three continents. The same integration was also found in a laboratory population from Thailand (Suzuki et al., 2020), confirming the ubiquitous distribution of this nrEVE.

3.6 | Wild mosquitoes have a variable landscape of reference viral integrations

To gain insights into nrEVE evolution, we analysed patterns of insertional polymorphism of reference nrEVEs in the genomes of wild-collected mosquitoes. We reasoned that, if viral integrations result from fortuitous events and are viral fossils, their distribution should be governed by drift and additionally, their sequence polymorphisms should evolve at a neutral rate (Aswad & Katzourakis, 2012). If viral integrations behave like immunity effectors, their presence/absence pattern is expected to be variable across host genomes and sequence polymorphism to be under natural selection (Frank & Feschotte, 2017).

We first analysed the presence/absence pattern of reference nrEVEs, classifying them as fixed, when they do not display any insertional polymorphism among geographical populations, and variably distributed (VD-nrEVEs) when they are absent in the genome of at least one wild-caught mosquito (Figure 6a). We identified a total of 114 VD-nrEVEs. Their overall frequency distribution separates samples according to their geographical origin (Figure 6b), although population frequencies were not statistically different based on the LRT for any of the VD-nrEVEs (Table S4). Still, the overall level of polymorphism (LoP) of VD-nrEVEs was comparable to that of a set of conserved genes in the *A. aegypti* genome (Pischedda et al., 2019) in all the tested populations (Figure 6c), indicating that some of them may have contributed to adaptation. Sixty-nine viral integrations appear to be shared across all populations (fixed nrEVEs) (Figure 6a). These may represent a conserved core of viral integrations that have reached fixation through processes other than positive selection. However, the LoP of fixed nrEVEs was suggestive of selection in all the populations tested (Figure 6c; Figure S12). Both fixed and VD-nrEVEs have similar LoP values across populations, which is far lower than LoP values of *A. aegypti* fast evolving genes. This result indicates that fixed and VD-nrEVEs may be under similar selective forces and that the different frequency patterns may reflect different integration times (old nrEVEs have reached fixation whereas VD-nrEVEs are younger insertions). There is a strong bias for Flavi-nrEVEs among the fixed nrEVEs as 92% of them are present in all individuals and in all populations (Data S5). Thus, the high frequency of the reference Flavi-nrEVEs may result from strong purifying selection. We analysed the distribution of SNPs vs. singletons (i.e., SNPs

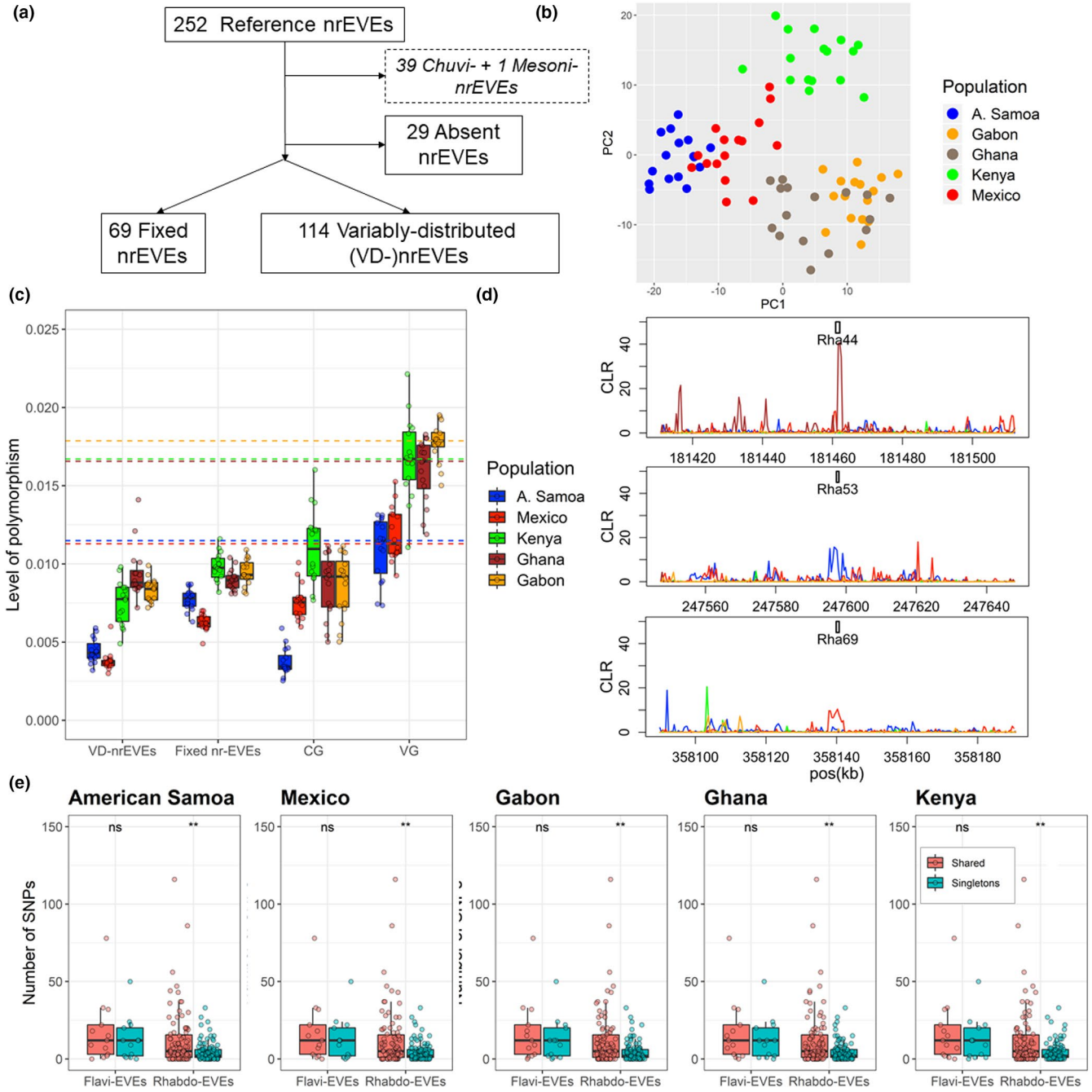


FIGURE 6 Genome-wide screen of reference nrEVEs. (a) Outline of the distribution of reference nrEVEs annotated in the AaegL5 assembly. Fixed nrEVEs have been observed in all populations and variably distributed (VD)-nrEVEs have been observed only in some populations. Mesoni- and Chuvi-nrEVEs are not included in this analysis. (b) Convex logistic principal component analysis (PCA) of VD-nrEVE frequencies based on their geographical origin. Each dot indicates an individual mosquito, colour-coded based on geographical location. (c) Whisker plots comparing the level of nucleotide polymorphism among *Aedes aegypti* populations in conserved genes (CG), fast-evolving genes (VG), fixed nrEVEs and VD-nrEVEs. Each dot represents the average value of an individual mosquito, boxes span the interquartile range, marked lines within the boxes represent the median, and whiskers represent the minimum and the maximum. Dotted lines are colour-coded according to the population they represent and depict the median value of the level of polymorphism of fast-evolving genes. (d) Composite likelihood ratio (CLR) signal around the indicated nrEVEs in mosquito populations. CLR signal around Rha44 in Ghana, Rha53 in American Samoa and Rha69 in Mexico is higher than the 99th percentile of their corresponding window distribution and thus indicative of positive selection. (e) Comparison of the number of singletons (i.e., SNPs found only in one individual) vs. SNPs in Flavi- or Rhabdo-nrEVEs in each population. Statistical differences were established by the Wilcoxon rank sum test (ns, not significant; $*p < .05$; $**p < .01$; $***p < .0001$) [Colour figure can be viewed at wileyonlinelibrary.com]

found only in one mosquito) across nrEVEs in each population. An excess of singletons is consistent with purifying selection (Bourgeois & Boissinot, 2019). This was not observed in Flavi-nrEVEs in any population, but their singletons/SNPs ratio was higher than that of Rhabo-nrEVEs, which have significantly fewer singletons than SNPs in all the tested populations (Figure 6e).

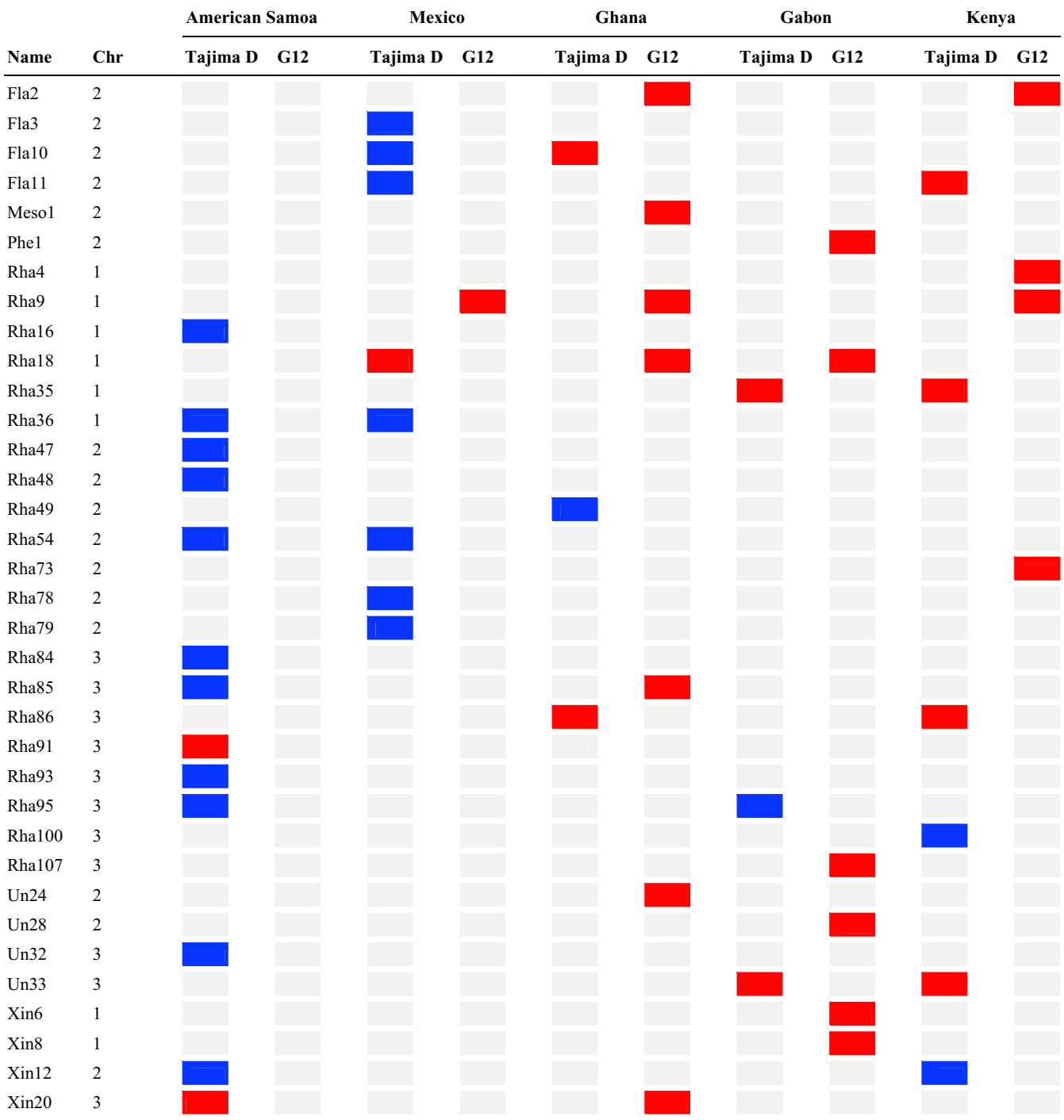


FIGURE 7 nrEVEs with signal of selection. Tajima's *D* and G12 values in the indicated nrEVEs in geographical populations of *Aedes aegypti*. Red indicates lower Tajima's *D* or higher G12 values than the cutoffs for each chromosome in each population (Tables S2 and S3). Blue boxes indicate high Tajima's *D* values (i.e., values higher than the 95% percentile for each chromosome in each population). Light grey boxes indicate nonsignificant tests. Only nrEVEs fixed in all tested populations and with either a significant Tajima's *D* or a significant G12 values in at least one of the tested populations are shown [Colour figure can be viewed at wileyonlinelibrary.com]

3.7 | Signs of positive selection on viral integrations

The spectrum of nucleotide polymorphism spanning a DNA sequence can be used to infer its evolutionary history under the premises of the theory of neutral molecular evolution (Kimura, 1983). The theory states that most of the polymorphism at the DNA sequence level is a neutral balance between two forces: mutation, which introduces new variants, and genetic drift, which randomly eliminates polymorphism. A number of statistical tests are available to compare the observed polymorphism of a region to that expected under neutral evolution given assumptions of population size, demography,

random mating and recombination (for a review, see Booker et al., 2017; Pavlidis & Alachiotis, 2017). Deviations from expectations under the null hypothesis of neutral evolution are interpreted as signs of selection. An increase in the frequency of a variant, associated with reduced variability in the neighbouring region due to genetic hitchhiking, generates a hard-selective sweep, which indicates positive or adaptive selection. The power of detecting signatures of hard selective sweeps using SNPs is sufficiently strong when they occurred less than ~0.1 *Ne* generations ago (Kim & Stephan, 2000; Przeworski, 2002), where *Ne* is the effective population size. Given that *A. aegypti* has around 11 generations per year and an *Ne* of roughly 500 (Saarman et al., 2017), this translates into the ability to detect only very recent hard sweeps, in the order of ~4 to 8 years. Thus, selective sweeps detected are likely to be population-specific and, in the case of nrEVEs, probably associated with locally circulating viruses. Signatures of hard sweeps were detected for Rha44 in Ghana, Rha69 in Mexico and Rha53 in American Samoa (Figure 6d). Rha44 and Rha69 are fixed in the populations where a hard sweep was predicted and Rha53 is close to fixation in American Samoa (Freq = 0.938), supporting the hypothesis that the increase in frequency of these three nrEVEs is probably due to positive selection acting on the population.

Adaptive evolution at nrEVE loci was also tested by estimating the Tajima's *D* statistic genome-wide and analysing whether reference nrEVEs map in windows with significantly negative Tajima's *D* values (Kofler et al., 2012; Rech et al., 2019). Tajima's *D* compares the polymorphism and the segregating sites in a region under the premises of the neutral evolution theory: a low Tajima's *D* value indicates an excess of low-frequency variants resulting from a bottleneck or a selective sweep; whereas a Tajima's *D* value significantly higher than 0 indicates a scarcity of rare variants, interpreted as balancing selection or recent population admixture (Stephan, 2019). Consistent with results of hard sweep, Rha69 showed significantly negative Tajima's *D* values in Mexico (Tajima's *D* = −0.62). Rha44 and Rha53 showed lower than chromosome-average Tajima's *D* values in Ghana and American Samoa, although the values were not statistically significant (Tables S3 and S5). Among these three nrEVEs, only Rha44 maps in a piRNA cluster, 2p21.18, which is active only in the germline (Data S2). Eight additional nrEVEs showed significantly negative Tajima's *D* values in different populations, with Rha35, Rha86 and Un33 consistently in two African populations (Figure 7). These Rhabdo-nrEVEs derive from different viral regions from probably different viruses and all map in piRNA clusters active in both somatic and germline tissues (Data S4).

An allele may be present in a population along with other variants and segregate neutrally, until an environmental change arises that favours its segregation. This situation will result in a soft selective sweep, detectable through *ad hoc* statistics such as the G12 and G1/2 method (Garud et al., 2015; Harris et al., 2018). We calculated the G12 and G1/2 statistics genome-wide and analysed whether nrEVEs occurred in windows with the top 15% most extreme G12 values, following a reference (Rech et al., 2019). Signatures of soft sweeps were identified in 13 fixed nrEVEs, primarily in African populations (Figure 7). Overall, these results highlight nrEVEs present in the reference AaegL5 assembly that are evolving under different scenarios in some selected populations and thus may be considered candidate adaptive nrEVEs.

4 | DISCUSSION

Experimental genetic approaches recently demonstrated that selected nrEVEs mediate piRNA-based antiviral activity in *A. aegypti* (Suzuki et al., 2020; Tassetto et al., 2019), expanding the immunity functions of viral sequences acquired by HGT and suggesting similarities to both prokaryotic CRISPR-Cas9 immunity and piRNA-mediated genome defence against TE movement (Ophinni et al., 2019). A system analogous to CRISPR-Cas9 immunity implies "adaptation," the modification of immune features (i.e., properties of immune cells, genomic loci or effectors) in response to environmental stimuli in a way that influence their subsequent responses to the same stimulus (Natoli & Ostuni, 2019). Here we tested a key feature of both piRNA-mediated genome defence and CRISPR-Cas9 immunity in the genomes of wild-caught mosquitoes: variability in the repertoire of foreign nucleic acids acquired by HGTs (Amitai & Sorek, 2016; Brennecke et al., 2007; Khurana et al., 2011). We demonstrate that mosquitoes have a variable landscape of nrEVEs, including modification of the composition of piRNA clusters, and we provide evidence that nrEVEs are evolving under different selective scenarios.

4.1 | Annotation of the reference nrEVEome of *A. aegypti*

While we were finishing our study, another annotation of the nrEVEome of AaegL5 assembly was published (Russo et al., 2019). The number of nrEVEs and the phylogenetic classification in the two studies are roughly similar (252 vs. 277). However, around one-fifth of the nrEVEs annotated in our study (46) have not been annotated by Russo et al. (2019). Conversely, 43 nrEVEs annotated by Russo are not present in our data set (Data S7). nrEVEs that have been missed by one of the two studies are dominated by unclassified nrEVEs. A wide variety of RNA viruses has been unearthed by metagenomics (C.-X. Li et al., 2015; Zhang et al., 2019), many of which are awaiting classification. Hence, the differences in nrEVEs repertoires of the two studies probably reflect the diversity of viruses included in databases used to annotate nrEVEomes and it is likely that ongoing identification of new mosquito viruses will require regular updates of nrEVE annotation. Of the 277 nrEVEs annotated by Russo et al. (2019), 234 correspond to 206 nrEVEs annotated in our study, which could be attributed to longer nrEVE sequences generated by our annotation pipeline.

Of the 252 nrEVEs annotated in our study, 243 nrEVEs probably produce piRNAs. Among nrEVEs that do not produce piRNAs, Meso1 is the only Mesoniviridae-derived nrEVE (Meso-nrEVE) annotated in

the AaegL5 assembly. This nrEVE is inserted in a sequence encoding a hypothetical protein present in other dipteran species, and a previous study proposed that it probably represents an unusual "nido-like" domain conserved across the Diptera lineage (Russo et al., 2019).

Considering the overall genomic architecture of nrEVEs and their piRNA profile, it is tempting to speculate that nrEVEs are organized through a redundant system with few viral regions being over-represented by multiple overlapping nrEVEs and showing hotspots in their piRNA profile. The majority of nrEVE-derived piRNAs are soma-specific, and only a small fraction (3%) is expressed in both soma and germline. Recent research suggests that the piRNA-mediated antiviral effect of a CFAV-derived nrEVE is strongest in the ovaries (Suzuki et al., 2020). Thus, different piRNA expression patterns between germline and soma could reflect the specialization of some nrEVEs in mediating antiviral immunity specifically in germline or somatic tissue.

4.2 | Identification and timing of novel nrEVEs

Investigation of the genome sequences of 80 wild-collected mosquitoes resulted in the identification of five novel nrEVEs similar to either CFAV or AeAV and occurring both within and outside piRNA clusters. Among the novel CFAV-EVEs identified, only CFAV-EVE-4 inserted within a piRNA cluster (namely 3p23.4), which is highly active in both soma and germline and hosts 13 other viral integrations (eight Rhabdo-nrEVEs, one Xinmo-nrEVE, two Chuvi-nrEVEs and two unclassified nrEVEs) (Figure 5). This is the first demonstration of insertional polymorphism of an nrEVE within a piRNA cluster in nature. This "invasion" occurred in the germline, constituting a vertically inherited trait that is subject to evolutionary selection. Something similar was observed in *Drosophila melanogaster*, where invading soma-specific TEs can be trapped into germline piRNA clusters (Duc et al., 2019). Insertion of an nrEVE outside a piRNA cluster can also trigger piRNA production, as illustrated by CFAV-EVE-3, which produces piRNAs but integrated outside annotated piRNA clusters (Suzuki et al., 2020).

Novel nrEVEs are not fixed in any population and their frequencies range from 8.3% (AevAV-like EVE in Gabon) to 58.3% (CFAV-EVE-3 in American Samoa). This distribution pattern, the high identity between the viral integrations and their corresponding viral genomes, and the absence of polymorphism in the identified nrEVEs, suggest that these novel integrations are recent events. The possibility remains that the lack of polymorphisms is due to selection, although in this case we would have expected novel integrations to be present at higher than the detected frequencies (González et al., 2008; Keightley & Eyre-Walker, 2007). Overall, the detection of only five novel viral integrations across 80 tested genomes suggests integration events are rare.

The absence of polymorphism and the absence of orthologous nrEVEs in *Aedes albopictus* (Palatini et al., 2017; Whitfield et al., 2017) prevents us from precisely dating integration events (Aiewsakun & Katzourakis, 2015). However, the timing of viral endogenization events can be partly deduced from the natural history of *A. aegypti*. Throughout Africa, mosquitoes occur predominantly as a darker form, called *A. aegypti formosus* (Aaf), which feeds on animals and uses natural water collections including tree holes for larval development. Outside of Africa, a lighter, domesticated form of *A. aegypti*, called *A. aegypti aegypti* (Aaa), occurs, which feeds on humans and uses anthropogenic containers as larval breeding sites (Crawford et al., 2017). The divergence between Aaf and Aaa is estimated to have occurred in Africa less than 1,000 years ago, prior to the global Aaa expansion that has been dated to ~400–600 years ago (Crawford et al., 2017; Soghigian et al., 2020). The population-specific occurrence of nrEVEs suggests that these integrations probably occurred after Aaa invasion into new areas rather than representing multiple losses of ancestral nrEVEs. The first records of *A. aegypti* in Mexico are from the 17th century and *A. aegypti* populations from Oceania and surrounding islands are derived from American populations, probably through navigation routes well established by the 19th century (Powell et al., 2018). Thus, CFAV-EVE-4 and CFAV-EVE-1 (detected in Mexico and American Samoa, respectively) should be posterior to the establishment of these two invasive populations. On the same basis, we hypothesize that CFAV-EVE-2 and the AeAV-like EVE (both present only in African populations) endogenized after the first out-of-Africa colonization event. The ubiquitous presence of CFAV-EVE-3 in genomes of mosquitoes from Mexico, American Samoa, Kenya and Thailand (Suzuki et al., 2020) points to an endogenization event prior to the split between Aaf and Aaa. However, the lack of polymorphisms and the low frequency of CFAV-EVE3 in some populations (i.e., 12.5% in Kenya) support the hypothesis that CFAV-EVE3 is not an ancestral integration. The most likely phylogeographical scenario predicts that Aaa arrived in Asia from the New World ~150 years ago (Powell et al., 2018). This agrees with the presence of CFAV-EVE-3 in Mexico, American Samoa and Thailand. African individuals harbouring CFAV-EVE-3 are present exclusively in mosquitoes collected in Rabai (Kenya). Here, Aaa mosquitoes are sympatric to Aaf (Brown et al., 2014). The origin of this unique Aaa form is still debated and previous research hypothesized that it was introduced to coastal East Africa, including Rabai, after the subspeciation event (Brown et al., 2014; Powell & Tabachnick, 2013). The presence of CFAV-EVE-3 in mosquitoes from Rabai, Mexico, Polynesia and Asia is consistent with a common origin of these populations, and thus an endogenization event posterior to the out-of-Africa colonization. However, in Rabai, mosquitoes with a peridomestic behaviour, closer to that of Aaf mosquitoes, were also identified suggesting mosquitoes from this location are mixed (Xia et al., 2020). We received ethanol-preserved mosquitoes from Rabai, and thus could not verify the coloration of the body which is the main visible character differentiating Aaa and Aaf.

4.3 | Selected nrEVEs show signs of positive selection

Our results suggest a dynamic landscape of viral integrations in mosquito genomes and identify adaptive nrEVEs both within and outside piRNA clusters. Overall, our data demonstrate that nrEVEs

are a complex component of the mosquito repeatome being maintained through both drift and selection. Thus, selected, but not all nrEVEs may play important functions in the virus–mosquito arms race, rather than being simply viral fossils in the mosquito genomes. Forward genetic approaches targeting candidate adaptive nrEVEs will be needed to understand their role in adaptive viral immunity.

ACKNOWLEDGEMENTS

We thank Mark Schmaedick from the American Samoa Community College for providing access to mosquitoes from American Samoa. We thank Lino Ometto from the University of Pavia for fruitful discussions.

AUTHOR CONTRIBUTIONS

C.C., R.v.R, and M.B. designed the study. C.R., F.V., R.H., U.P, M.M., L.G., S.R.,Y.A., A.D., C.L.R., and P.M. collected the data. C.R., F.V., R.H., U.P, M.M., L.G., S.R.,Y.A., A.D., C.L.R., P.M., and M.B. analysed the data. R.v.R, and M.B. provided supervision. C.R., R.H., L.G., S.R., P.M., R.v.R. and M.B. wrote the paper with edits from all authors.

DATA AVAILABILITY STATEMENT

Whole genome sequencing data have been submitted at NCBI SRA under BioProject PRJNA609256.

ORCID

Cristina M. Crava https://orcid.org/0000-0003-3774-4567
Diego Ayala https://orcid.org/0000-0003-4726-580X
Mariangela Bonizzoni https://orcid.org/0000-0003-0568-8564

REFERENCES

Aiewsakun, P., & Katzourakis, A. (2015). Endogenous viruses : Connecting recent and ancient viral evolution. *Virology*, *479–480*, 26–37. https://doi.org/10.1016/j.virol.2015.02.011.

Amitai, G., & Sorek, R. (2016). CRISPR-Cas adaptation: insights into the mechanism of action. *Nature Reviews Microbiology*, *14*(2), 67–76. https://doi.org/10.1038/nrmicro.2015.14.

Antoniewski, C. (2014). Computing siRNA and piRNA overlap signatures. *Methods in Molecular Biology (Clifton. N.J.)*, *1173*, 135–146. https://doi.org/10.1007/978-1-4939-0931-5_12.

Arensburger, P., Hice, R. H., Wright, J. A., Craig, N. L., & 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*, https://doi.org/10.1186/1471-2164-12-606.

Aswad, A., & Katzourakis, A. (2012). Paleovirology and virally derived immunity. *Trends in Ecology and Evolution*, *27*(11), 627–636. https://doi.org/10.1016/j.tree.2012.07.007.

Baidaliuk, A., Lequime, S., Moltini-Conclois, I., Dabo, S., Dickson, L. B., Prot, M., Duong, V., Dussart, P., Boyer, S., Shi, C., Matthijnssens, J., Guglielmini, J., Gloria-Soria, A., Simon-Lorière, E., & Lambrechts, L. (2020). Novel genome sequences of cell-fusing agent virus allow comparison of virus phylogeny with the genetic structure of Aedes aegypti populations. *Virus. Evolution*, *6*(1), veaa018. https://doi.org/10.1093/ve/veaa018.

Blair, C. D., Olson, K. E., & Bonizzoni, M. (2020). The Widespread Occurrence and Potential Biological Roles of Endogenous Viral Elements in Insect Genomes. *Current Issues in Molecular Biology*, *34*, 13–30.https://doi.org/10.21775/cimb.034.013 .

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

Bolling, B. G., Weaver, S. C., Tesh, R. B., & Vasilakis, N. (2015). Insect-specific virus discovery: Significance for the arbovirus community. *Viruses*, *7*(9), 4911–4928. https://doi.org/10.3390/v7092851.

Booker, T. R., Jackson, B. C., & Keightley, P. D. (2017). Detecting positive selection in the genome. *BMC Biology*, *15*(1), https://doi.org/10.1186/s12915-017-0434-y.

Bourgeois, Y., & Boissinot, S. (2019). On the population dynamics of junk: A review on the population genomics of transposable elements. *Genes*, *10*(6), https://doi.org/10.3390/genes10060419.

Brennecke, J., Aravin, A. A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., & Hannon, G. J. (2007). Discrete Small RNA-Generating Loci as Master Regulators of Transposon Activity in Drosophila. *Cell*, *128*, 1089–1103. https://doi.org/10.1016/j.cell.2007.01.043.

Brown, J. E., Evans, B. R., Zheng, W., Obas, V., Barrera-Martinez, L., Egizi, A., Zhao, H., Caccone, A., & Powell, J. R. (2014). Human impacts have shaped historical and recent evolution in Aedes aegypti, the dengue and yellow fever mosquito. *Evolution; International Journal of Organic Evolution*, *68*(2), 514–525. https://doi.org/10.1111/evo.12281.

Cotter, C. J., Tufa, A. J., Johnson, S., Matai'a, M., Sciulli, R., Ryff, K. R., Hancock, W. T., Whelen, C., Sharp, T. M., & Anesi, M. S. (2018). Outbreak of Dengue Virus Type 2 — American Samoa, November 2016–October 2018. *MMWR Morb Mortal Wkly Rep*, *67*, 1319–1322. https://doi.org/10.15585/mmwr.mm6747a5

Crawford, J. E., Alves, J. M., Palmer, W. J., Day, J. P., Sylla, M., Ramasamy, R., Surendran, S. N., Black, W. C., Pain, A., & Jiggins, F. M. (2017). Population genomics reveals that an anthropophilic population of Aedes aegypti mosquitoes in West Africa recently gave rise to American and Asian populations of this major disease vector. *BMC Biology*, *15*(1), 1–16. https://doi.org/10.1186/s12915-017-0351-0.

Czech, B., & Hannon, G. J. (2016). One Loop to Rule Them All: The Ping-Pong Cycle and piRNA-Guided Silencing. *Trends in Biochemical Sciences*, *41*(4), 324–337. https://doi.org/10.1016/j.tibs.2015.12.008.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330.

Duc, C., Yoth, M., Jensen, S., Mouniée, N., Bergman, C. M., Vaury, C., & Brasset, E. (2019). Trapping a somatic endogenous retrovirus into a germline piRNA cluster immunizes the germline against further invasion. *Genome Biology*, *20*(1), 1–14. https://doi.org/10.1186/s13059-019-1736-x.

Forster, M., Szymczak, S., Ellinghaus, D., Hemmrich, G., Rühlemann, M., Kraemer, L., Mucha, S., Wienbrandt, L., & Stanulla, M. (2015). VyPER : eliminating false positive detection of virus integration events in next generation sequencing data. *Scientific Reports*, *5*, 11534. https://doi.org/10.1038/srep11534.

Frank, J. A., & Feschotte, C. (2017). Co-option of endogenous viral sequences for host cell function. *Current Opinion in Virology*, *25*, 81–89. https://doi.org/10.1016/j.coviro.2017.07.021.

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *ArXiv*.

Garud, N. R., Messer, P. W., Buzbas, E. O., & Petrov, D. A. (2015). Recent Selective Sweeps in North American Drosophila melanogaster Show Signatures of Soft Sweeps. *PLoS Genetics*, *11*(2), 1–32. https://doi.org/10.1371/journal.pgen.1005004.

Gilbert, C., & Cordaux, R. (2017). Viruses as vectors of horizontal transfer of genetic material in eukaryotes. *Current Opinion in Virology*, *25*, 16–22. https://doi.org/10.1016/j.coviro.2017.06.005.

González, J., Lenkov, K., Lipatov, M., Macpherson, J. M., & Petrov, D. A. (2008). High Rate of Recent Transposable Element-Induced

Adaptation in Drosophila melanogaster. *PLoS Biology*, *6*(10), e251. https://doi.org/10.1371/journal.pbio.0060251.

Guerbois, M., Fernandez-Salas, I., Azar, S. R., Danis-Lozano, R., Alpuche-Aranda, C. M., Leal, G., Garcia-Malo, I. R., Diaz-Gonzalez, E. E., Casas-Martinez, M., Rossi, S. L., Del Río-Galván, S. L., Sanchez-Casas, R. M., Roundy, C. M., Wood, T. G., Widen, S. G., Vasilakis, N., & Weaver, S. C. (2016). Outbreak of Zika Virus Infection, Chiapas State, Mexico, 2015, and First Confirmed Transmission by Aedes aegypti Mosquitoes in the Americas. *The Journal of Infectious Diseases*, *214*(9), 1349–1356. https://doi.org/10.1093/infdis/jiw302.

Halbach, R., Miesen, P., Joosten, J., Taşköprü, E., Rondeel, I., Pennings, B., Vogels, C. B. F., Merkling, S. H., Koenraadt, C. J., Lambrechts, L., & van Rij, R. P. (2020). A satellite repeat-derived piRNA controls embryonic development of Aedes. *Nature*, *580*(7802), 274–277. https://doi.org/10.1038/s41586-020-2159-2.

Hall, R. A., Bielefeldt-Ohmann, H., McLean, B. J., O'Brien, C. A., Colmant, A. M. G., Piyasena, T. B. H., Harrison, J. J., Newton, N. D., Barnard, R. T., Prow, N. A., Deerain, J. M., Mah, M. G. K. Y., & Hobson-Peters, J. (2016). Commensal viruses of mosquitoes: Host restriction, transmission, and interaction with arboviral pathogens. *Evolutionary Bioinformatics*, *12*, 35–44. https://doi.org/10.4137/EBo.s40740.

Hall, T. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, *41*, 95–98.https://doi.org/citeulike-article-id:691774 .

Hannon, G. J. (2009). *FASTX-Toolkit*. http://hannonlab.cshl.edu/fastx_toolkit/index.html .

Harris, A. M., Garud, N. R., & DeGiorgio, M. (2018). Detection and Classification of Hard and Soft Sweeps from Unphased Genotypes by Multilocus Genotype Identity. *Genetics*, *210*(4), 1429–1452. https://doi.org/10.1534/genetics.118.301502.

Hayward, A. (2017). Origin of the retroviruses: when, where, and how? *Current Opinion in Virology*, *25*, 23–27. https://doi.org/10.1016/j.coviro.2017.06.006.

Keeling, P. J., & Palmer, J. D. (2008). Horizontal gene transfer in eukaryotic evolution. *Nature Reviews. Genetics*, *9*(8), 605–618. https://doi.org/10.1038/nrg2386.

Keightley, P. D., & Eyre-Walker, A. (2007). Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics*, *177*(4), 2251–2261. https://doi.org/10.1534/genetics.107.080663.

Kent, W. J. (2002). BLAT - The BLAST-like alignment tool. *Genome Research*, *12*(4), 656–664. https://doi.org/10.1101/gr.229202. Article published online before March 2002.

Khurana, J., Wang, J., Xu, J., Koppetsch, B., Thomson, T., Nowosielska, A., Li, C., Zamore, P., Weng, Z., & Theurkauf, W. (2011). Adaptation to transposon invasion in *Drosophila melanogaster*. *Cell*, *147*(7), 1551–1563. https://doi.org/10.1016/j.cell.2011.11.042.Adaptation.

Kim, Y., & Stephan, W. (2000). Joint effects of genetic hitchhiking and background selection on neutral variation. *Genetics*, *155*(3), 1415–1427.

Kimura, M. (1983). The Neutral Theory of Molecular Evolution. *Cambridge University Press*, https://doi.org/10.1017/CBO9780511623486.

Kofler, R., Betancourt, A. J., & Schlötterer, C. (2012). Sequencing of pooled DNA samples (Pool-Seq) uncovers complex dynamics of transposable element insertions in Drosophila melanogaster. *PLoS Genetics*, *8*(1), https://doi.org/10.1371/journal.pgen.1002487.

Landgraf, A. J., & Lee, Y. (2015). Dimensionality reduction for binary data through the projection of natural parameters. *ArXiv*.

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, *10*(3), R25. https://doi.org/10.1186/gb-2009-10-3-r25.

Lewis, S. H., Quarles, K. A., Yang, Y., Tanguy, M., Frézal, L., Smith, S. A., Sharma, P. P., Cordaux, R., Gilbert, C., Giraud, I., Collins, D. H., Zamore, P. D., Miska, E. A., Sarkies, P., & Jiggins, F. M. (2018). *Pan-arthropod analysis reveals somatic piRNAs as an ancestral defence against transposable elements.*, *2*, 174–181. https://doi.org/10.1038/s41559-017-0403-4.

Li, C.-X., Shi, M., Tian, J.-H., Lin, X.-D., Kang, Y.-J., Chen, L.-J., Qin, X.-C., Xu, J., Holmes, E. C., & Zhang, Y.-Z. (2015). Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. *ELife*, *4*, https://doi.org/10.7554/eLife.05378.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987–2993. https://doi.org/10.1093/bioinformatics/btr509.

Malik, H. S., Henikoff, S., & Eickbush, T. H. (2000). Poised for contagion: Evolutionary origins of the infectious abilities of invertebrate retroviruses. *Genome Research*, *10*(9), 1307–1318. https://doi.org/10.1101/gr.145000.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Joutnal*, *17*, 10–12.

Matthews, J. B., Dudchenko, O., Kingan, S. B., Koren, S., Antoschechkin, I., Crawford, J. E., Glassford, W. J., Herre, M., Redmond, S. N., Rose, N. H., Weedall, G. D., Wu, Y., Batra, S. S., Brito-Sierra, C. A., Buckingham, S. D., Campbell, C. L., Chan, S., Cox, E., Evans, B. R., … Vosshall, L. B. (2018). Improved reference genome of Aedes aegypti informs arbovirus vector control. *Nature*, *563*(7732), 501–507. https://doi.org/10.1038/s41586-018-0692-z.

McCarthy, D. J., Chen, Y., & Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, *40*(10), 4288–4297. https://doi.org/10.1093/nar/gks042.

Muerdter, F., Olovnikov, I., Molaro, A., Rozhkov, N. V., Czech, B., Gordon, A., Hannon, G. J., & Aravin, A. A. (2012). Production of artificial piRNAs in flies and mice. *RNA*, *18*(1), 42–52. https://doi.org/10.1261/rna.029769.111.

Natoli, G., & Ostuni, R. (2019). Adaptation and memory in immune responses. *Nature Immunology*, *20*(7), 783–792. https://doi.org/10.1038/s41590-019-0399-9.

Ophinni, Y., Palatini, U., Hayashi, Y., & Parrish, N. F. (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 .

Ozata, D. M., Gainetdinov, I., Zoch, A., O'Carroll, D., & Zamore, P. D. (2019). PIWI-interacting RNAs: small RNAs with big functions. *Nature Reviews Genetics*, *20*(2), 89–108. https://doi.org/10.1038/s41576-018-0073-3.

Palatini, U., Miesen, P., Carballar-Lejarazu, R., Ometto, L., Rizzo, E., Tu, Z., van Rij, R. P., & Bonizzoni, M. (2017). Comparative genomics shows that viral integrations are abundant and express piRNAs in the arboviral vectors Aedes aegypti and Aedes albopictus. *BMC Genomics*, *18*(1), 1–15. https://doi.org/10.1186/s12864-017-3903-3.

Parhad, S. S., & Theurkauf, W. E. (2019). Rapid evolution and conserved function of the piRNA pathway. *Open Biology*, *9*(1), https://doi.org/10.1098/rsob.18.0181.

Parry, R., & Asgari, S. (2018). Aedes Anphevirus: an Insect-Specific Virus Distributed Worldwide in %3cspan class="named-content genus-species" xml:id="named-content-1"%3eAedes aegypti%3c/span%3e Mosquitoes That Has Complex Interplays with %3cspan class="named-. *Journal of Virology*, *92*(17), e00224–18. https://doi.org/10.1128/JVI.00224-18.

Pavlidis, P., & Alachiotis, N. (2017). A survey of methods and tools to detect recent and strong positive selection. *Journal of Biological Research-Thessaloniki*, *24*(1), 7. https://doi.org/10.1186/s40709-017-0064-0.

Pavlidis, P., Zickovic, D., Stamatakis, A., & Alachiotis, N. (2013). SweeD : Likelihood-Based Detection of Selective Sweeps in Thousands

of Genomes. *Molecular Biology and Evolution*, *30*(9), 2224–2234. https://doi.org/10.1093/molbev/mst112.

Pischedda, E., Crava, C. M., Carlassara, M., Gasmi, L., & Bonizzoni, M. (2020). ViR: a tool to acount for intrasample variability in the detection fo viral integrations. *BioRxiv*. https://doi.org/10.1101/2020.06.16.155119.

Pischedda, E., Scolari, F., Valerio, F., Carballar-lejarazú, R., Catapano, P. L., Waterhouse, R. M., & 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. *Frontiers in Genetics*, *10*, 93. https://doi.org/10.3389/fgene.2019.00093.

Powell, J. R., Gloria-Soria, A., & Kotsakiozi, P. (2018). Recent history of Aedes aegypti: Vector genomics and epidemiology records. *BioScience*, *68*(11), 854–860. https://doi.org/10.1093/biosci/biy119.

Powell, J. R., & Tabachnick, W. J. (2013). History of domestication and spread of Aedes aegypti–a review. *Memórias Do Instituto Oswaldo Cruz*, *108*(October), 11–17. https://doi.org/10.1590/0074-02761 30395.

Przeworski, M. (2002). The signature of positive selection at randomly chosen loci. *Genetics*, *160*(3), 1179–1189.

Quinlan, A. R., & Hall, I. M. (2010). BEDTools : a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033.

Rech, G. E., Bogaerts-Márquez, M., Barrón, M. G., Merenciano, M., Villanueva-Cañas, J. L., Horváth, V., Fiston-Lavier, A. S., Luyten, I., Venkataram, S., Quesneville, H., Petrov, D. A., & González, J. (2019). Stress response, behavior, and development are shaped by transposable element-induced mutations in Drosophila. *PLoS Genetics*, *15*(2), e1007900. https://doi.org/10.1371/journal.pgen.1007900.

Redmond, S. N., Sharma, A., Sharakhov, I., Tu, Z., Sharakhova, M., & Neafsey, D. E. (2020). Linked-read sequencing identifies abundant microinversions and introgression in the arboviral vector Aedes aegypti. *BMC Biology*, *18*(1), 26. https://doi.org/10.1186/s12915-020-0757-y.

Russo, A. G., Kelly, A. G., Enosi Tuipulotu, D., Tanaka, M. M., & White, P. A. (2019). Novel insights into endogenous RNA viral elements in Ixodes scapularis and other arbovirus vector genomes. *Virus. Evolution*, *5*(1), https://doi.org/10.1093/ve/vez010.

Saarman, N. P., Gloria-Soria, A., Anderson, E. C., Evans, B. R., Pless, E., Cosme, L. V., Gonzalez-Acosta, C., Kamgang, B., Wesson, D. M., & Powell, J. R. (2017). Effective population sizes of a major vector of human diseases. *Aedes aegypti*. *Evolutionary Applications*, *10*(10), 1031–1039. https://doi.org/10.1111/eva.12508.

Soghigian, J., Gloria-Soria, A., Robert, V., Le Goff, G., Failloux, A.-B., & Powell, J. R. (2020). Genetic evidence for the origin of Aedes aegypti, the yellow fever mosquito, in the southwestern Indian Ocean. *Molecular Ecology*, *29*(19), 3593–3606. https://doi.org/10.1111/mec.15590.

Stephan, W. (2019). Selective Sweeps. *Genetics*, *211*(1), 5–13. https://doi.org/10.1534/genetics.118.301319.

Suzuki, Y., Baidaliuk, A., Miesen, P., Frangeul, L., Crist, A. B., Merkling, S. H., Fontaine, A., Lequime, S., Moltini-Conclois, I., Blanc, H., van Rij, R. P., Lambrechts, L., & Saleh, M.-C. (2020). Non-retroviral endogenous viral element limits cognate virus replication in Aedes aegypti ovaries. *Current Biology*, https://doi.org/10.1101/2020.03.28.013441.

Suzuki, Y., Frangeul, L., Dickson, L. B., Blanc, H., Verdier, Y., Vinh, J., Lambrechts, L., & Saleh, M.-C. (2017). Uncovering the Repertoire of Endogenous Flaviviral Elements in Aedes Mosquito Genomes. *Journal of Virology*, *91*(15), https://doi.org/10.1128/JVI.00571-17.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, *123*(3), 585–595.

Tassetto, M., Kunitomi, M., Whitfield, Z. J., Dolan, P. T., Sánchez-Vargas, I., Garcia-Knight, M., Ribiero, I., Chen, T., Olson, K. E., & Andino, R. (2019). Control of RNA viruses in mosquito cells through the acquisition of vDNA and endogenous viral elements. *ELife*, *8*, 1–29. https://doi.org/10.7554/eLife.41244.

ter Horst, A. M., Nigg, J. C., Dekker, F. M., & Falk, B. W. (2019). Endogenous Viral Elements Are Widespread in Arthropod Genomes and Commonly Give Rise to PIWI-Interacting RNAs. *Journal of Virology*, *93*(6), e02124–18. https://doi.org/10.1128/JVI.02124-18.

Théron, E., Dennis, C., Brasset, E., & Vaury, C. (2014). Distinct features of the piRNA pathway in somatic and germ cells: from piRNA cluster transcription to piRNA processing and amplification. *Mobile DNA*, *5*(1), 28. https://doi.org/10.1186/s13100-014-0028-y.

Uhrig, S., & Klein, H. (2019). PingPongPro: a tool for the detection of piRNA-mediated transposon-silencing in small RNA-Seq data. *Bioinformatics (Oxford, England)*, *35*(2), 335–336. https://doi.org/10.1093/bioinformatics/bty578.

Wagih, O. (2017). ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics*, *33*(22), 3645–3647. https://doi.org/10.1093/bioinformatics/btx469.

Weetman, D., Kamgang, B., Badolo, A., Moyes, C. L., Shearer, F. M., Coulibaly, M., Pinto, J., Lambrechts, L., & McCall, P. J. (2018). Aedes Mosquitoes and Aedes-Borne Arboviruses in Africa: Current and Future Threats. *International Journal of Environmental Research and Public Health*, *15*(2), 220 https://doi.org/10.3390/ijerph15020220.

Whitfield, Z. J., Dolan, P. T., Kunitomi, M., Andino, R., Tassetto, M., Seetin, M. G., Oh, S., Heiner, C., Paxinos, E., & Andino, R. (2017). The diversity, structure, and function of heritable adaptive immunity sequences in the Aedes aegypti genome. *Current Biology*, *27*(22), 3511–3519.e7. https://doi.org/10.1016/j.cub.2017.09.067.

Xia, S., Cosme, L. V., Lutomiah, J., Sang, R., Ngangue, M. F., Rahola, N., Ayala, D., & Powell, J. R. (2020). Genetic structure of the mosquito *Aedes aegypti* in local forest and domestic habitats in Gabon and Kenya. *Parasites & Vectors*, *13*(1), 417. https://doi.org/10.1186/s13071-020-04278-w.

Zhang, Y.-Z., Chen, Y.-M., Wang, W., Qin, X.-C., & Holmes, E. C. (2019). Expanding the RNA Virosphere by Unbiased Metagenomics. *Annual Review of Virology*, *6*(1), 119–139. https://doi.org/10.1146/annurev-virology-092818-015851.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Crava CM, Varghese FS, Pischedda E, et al. Population genomics in the arboviral vector *Aedes aegypti* reveals the genomic architecture and evolution of endogenous viral elements. *Mol Ecol*. 2021;30:1594–1611. https://doi.org/10.1111/mec.15798