

RESEARCH ARTICLE

Open Access



# Comparative genomics shows that viral integrations are abundant and express piRNAs in the arboviral vectors *Aedes aegypti* and *Aedes albopictus*

Umberto Palatini<sup>1†</sup>, Pascal Miesen<sup>2†</sup>, Rebeca Carballar-Lejarazu<sup>1</sup>, Lino Ometto<sup>3</sup>, Ettore Rizzo<sup>4</sup>, Zhijian Tu<sup>5</sup>, Ronald P. van Rij<sup>2</sup> and Mariangela Bonizzoni<sup>1\*</sup> 

## Abstract

**Background:** Arthropod-borne viruses (arboviruses) transmitted by mosquito vectors cause many important emerging or resurging infectious diseases in humans including dengue, chikungunya and Zika. Understanding the co-evolutionary processes among viruses and vectors is essential for the development of novel transmission-blocking strategies. Episomal viral DNA fragments are produced from arboviral RNA upon infection of mosquito cells and adults. Additionally, sequences from insect-specific viruses and arboviruses have been found integrated into mosquito genomes.

**Results:** We used a bioinformatic approach to analyse the presence, abundance, distribution, and transcriptional activity of integrations from 425 non-retroviral viruses, including 133 arboviruses, across the presently available 22 mosquito genome sequences. Large differences in abundance and types of viral integrations were observed in mosquito species from the same region. Viral integrations are unexpectedly abundant in the arboviral vector species *Aedes aegypti* and *Ae. albopictus*, in which they are approximately ~10-fold more abundant than in other mosquito species analysed. Additionally, viral integrations are enriched in piRNA clusters of both the *Ae. aegypti* and *Ae. albopictus* genomes and, accordingly, they express piRNAs, but not siRNAs.

**Conclusions:** Differences in the number of viral integrations in the genomes of mosquito species from the same geographic area support the conclusion that integrations of viral sequences is not dependent on viral exposure, but that lineage-specific interactions exist. Viral integrations are abundant in *Ae. aegypti* and *Ae. albopictus*, and represent a thus far underappreciated component of their genomes. Additionally, the genome locations of viral integrations and their production of piRNAs indicate a functional link between viral integrations and the piRNA pathway. These results greatly expand the breadth and complexity of small RNA-mediated regulation and suggest a role for viral integrations in antiviral defense in these two mosquito species.

**Keywords:** Mosquitoes, Arboviruses, Non-retroviral integrated RNA viruses, Endogenous viral elements, piRNA

\* Correspondence: m.bonizzoni@unipv.it

†Equal contributors

<sup>1</sup>Department of Biology and Biotechnology, University of Pavia, via Ferrata 9, 27100 Pavia, Italy

Full list of author information is available at the end of the article



## Background

Nearly one-quarter of emerging or resurging infectious diseases in humans are vector-borne [1]. Hematophagous mosquitoes of the *Culicidae* family are the most serious vectors in terms of their worldwide geographic distribution and the public health impact of the pathogens they transmit. The *Culicidae* is a large family whose members separated between 180 to 257 million years ago (mya) into the *Culicinae* and *Anophelinae* subfamilies [2]. Mosquitoes of the *Aedes* and *Culex* genera within the *Culicinae* subfamily are the primary vectors of RNA viruses. These viruses include taxa with different RNA genomic structures and replication strategies, but all are non-retroviral viruses [3]. Collectively, these viruses are referred to as arthropod-borne (arbo-) viruses. Within the *Aedes* genus, *Aedes aegypti* and *Aedes albopictus* are the main arboviral vectors due to their broad geographic distribution, adaptation to breed in human habitats, and the wide number of viral species from different genera that they can vector [4, 5]. These two mosquito species are able to efficiently transmit arboviruses of the genera *Flavivirus* (e.g. dengue [DENV], Zika [ZKV], Usutu, Japanese encephalitis and yellow fever virus), *Alphavirus* (e.g. chikungunya virus [CHIKV]), viruses of the Venezuelan equine encephalitis [VEE] and eastern equine encephalitis [EEE] complexes), *Orthobunyavirus* (e.g. Potosi, Cache Valley and La Crosse virus [LACV]), *Phlebovirus* (e.g. Rift Valley fever virus [RVFV]) and *Orbivirus* (e.g. Orungo Virus) [5–7]. Mosquitoes of the *Culex pipiens* complex, such as *Cx. pipiens pipiens* and *Cx. quinquefasciatus*, are the most prominent *Culex* vectors because of their wide distribution and close association with humans [7]. These mosquito species are primary vectors of encephalitic flaviviruses, such as West Nile virus (WNV) and Japanese encephalitis virus, and they can also vector RVFV [7, 8]. The only arbovirus known to be transmitted by *Anophelinae* is the alphavirus O'nyong-nyong [9]. Recently, additional RNA viruses have been identified from wild mosquitoes, but their virulence to humans and their impact on vector competence is still uncertain [9–11].

Mosquito competence for arboviruses is a complex and evolving phenotype because it depends on the interaction of genetic factors from both mutation-prone RNA viruses and mosquito vectors with environmental variables [12–15]. Not surprisingly, large variation exists in vector competence not only among mosquito species, but also across geographic populations within a species [16, 17]. Understanding the genetic components of vector competence and how these genetic elements are distributed in natural populations and interact with environmental factors is essential for predicting the risk of arboviral diseases and for developing new transmission-blocking strategies [12]. Genomic and functional studies, primarily in *Drosophila melanogaster* and *Aedes* mosquitoes,

have shown that RNA interference (RNAi) is the main antiviral mechanism in insects [18–20]. In this pathway, small RNAs are used to guide a protein-effector complex to target RNA based on sequence-complementarity. Three RNA silencing mechanisms exist: the microRNA, small interfering RNA (siRNA) and PIWI-interacting RNA (piRNA) pathways. They can be distinguished based on the mechanism of small RNA biogenesis and the effector protein complex to which these small RNAs associate [18, 19]. While the role of the siRNA pathway in restricting arboviral infection has been widely studied and appears universal across mosquitoes, recent studies highlight the contribution of the piRNA pathway in antiviral immunity of *Aedes* mosquitoes [21]. Although important aspects of piRNA biogenesis and function in mosquitoes remains to be elucidated, it is clear that endogenous piRNAs arise from specific genomic loci called piRNA clusters, as was originally observed in *D. melanogaster* [22]. These piRNA clusters contain repetitive sequences, remnants of transposable elements and, in *Ae. aegypti*, virus-derived sequences [23].

Recent studies have shown that the genomes of some eukaryotic species, including mosquitoes, carry integrations from non-retroviral RNA viruses [24–32]. Viral integrations are generally referred to as Endogenous Viral Elements (EVEs) [33] or, if they derive from non-retroviral RNA viruses, as Non-Retroviral Integrated RNA Virus Sequences (NIRVS) [29, 34]. Integration of non-retroviral sequences into host genomes is considered a rare event because it requires reverse transcription by an endogenous reverse transcriptase, nuclear import and genomic insertion of virus-derived DNA (vDNA) [35]. During infection with DENV, WNV, Sindbis virus, CHIKV and LACV, fragments of RNA virus genomes are converted into vDNA by the reverse transcriptase activity of endogenous transposable elements (TEs) in cell lines derived from *D. melanogaster*, *Culex tarsalis*, *Ae. aegypti*, and *Ae. albopictus*, as well as in adult mosquitoes. The episomal vDNA forms produced by this mechanism reside in the nucleus and it has been proposed that they contribute to the establishment of persistent infections through the RNAi machinery [20, 36, 37]. These recent studies not only show that reverse transcription of RNA viruses occurs in *Culicinae*, they also suggest the functional involvement of RNAi. If vDNAs fragments from arboviruses integrate into host genomes, the NIRVS landscape should be different across mosquito species depending both on their viral exposure and the activity of their RNAi pathways. To test this hypothesis, we used a bioinformatics approach to analyse the presence, abundance, distribution, and transcriptional activity of NIRVS across the currently available 22 mosquito genome sequences. We probed these genomes for integrations from 425 non-retroviral viruses, including 133 arboviruses. We observed a ten-fold

difference in the number of NIRVS between *Aedes* and the other tested mosquitoes. NIRVS were not evenly distributed across *Aedes* genomes, but occurred preferentially in piRNA clusters and, accordingly, they produced piRNAs. Among the viral species tested, integrations had the highest similarities to rhabdoviruses, flaviviruses and bunyaviruses, viruses that share the same evolutionary origin [38].

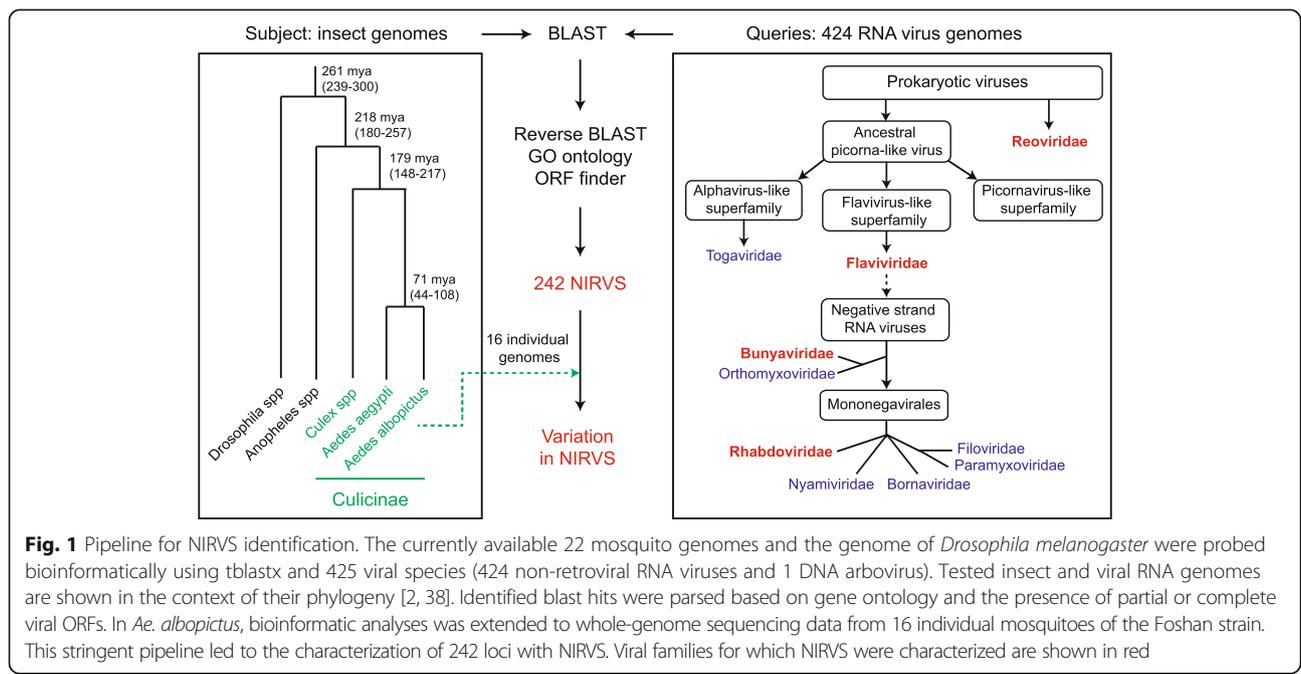
The larger number of NIRVS identified in *Ae. aegypti* and *Ae. albopictus*, their genome locations and their production of piRNAs show that in these species genomic integration of viral sequences is a more pervasive process than previously thought. Based on the interplay between viral integrations and the piRNA pathway, we propose that viral integrations represent heritable immune signals and contribute to shape vector competence.

**Results**

**NIRVS are unevenly distributed across mosquito species**

Twenty-five genome assemblies from 22 *Culicinae* species, along with the genome of *D. melanogaster*, were searched bioinformatically for sequence integrations derived from all 424 non-retroviral RNA viruses for which a complete genome sequence is currently available. Additionally, we tested the genome of African swine fever virus, the only known DNA arbovirus [3], giving a total of 133 arboviruses (Fig. 1, Additional file 1: Tables S1 and Additional file 2: Table S2). The genomes of 16 individual *Ae. albopictus* Foshan mosquitoes were sequenced to further validate NIRVS in this species. Retrieved sequences longer than 100 base pairs (bp) were filtered based on gene

ontology and the presence of partial or complete open reading frames (ORFs) of viral proteins. This stringent pipeline led to the characterization of a total of 242 loci harboring NIRVS across the genome of 15 mosquitoes (Table 1, Fig. 2). NIRVS loci were unevenly distributed across species. *Anopheleline* species had a maximum of seven NIRVS-loci, one NIRVS-locus was found in *Cx. quinquefasciatus*, 122 NIRVS were detected in *Ae. aegypti*, and 72 were found in *Ae. albopictus*. NIRVS landscape was highly variable across the 16 sequenced *Ae. albopictus* genomes with extensive differences in the number of NIRVS and in their length, suggesting that NIRVS are frequently rearranged (Fig. 3). A total of ten viral integrations detected in the AaloF1 assembly had no read coverage in any of the 16 sequenced *Ae. albopictus* genomes (Additional file 3: Table S3). The percentage of mapped reads and coverage was comparable across libraries excluding insufficient sequence depth as an explanation for the lack of coverage in these ten NIRVS loci (Additional file 4: Table S4). It is currently unclear if these ten NIRVS are rare integrations or result from mis-assembly of the reference genome. Among the eleven viral families tested, NIRVS had sequence similarities exclusively with viruses of the *Rhabdoviridae*, *Flaviviridae*, *Bunyaviridae* and *Reoviridae* families, including currently circulating viruses (Table 1). *Reoviridae* and *Bunyaviridae*-like integrations were similar to recently characterized viruses [39, 40] and were rare, with no more than one integration per species (Fig. 2). Phylogenetic analyses showed that viral integrations from *Reoviridae* were separated from currently known virus species in this family (Fig. 4a,b). Integrations from *Bunyaviridae* were at the base of the phylogenetic tree and



**Table 1** Number of viral integrations (NIRVS) detected for each of the viral families tested across the 22 mosquito genomes. A total of 424 non-retroviral RNA viruses with complete genomes were analyzed. The genome of African swine fever virus, the only known DNA arbovirus was also included in the analyses, but no NIRVS were found for this virus. Mosquito species and viral families for which NIRVS were detected are in bold

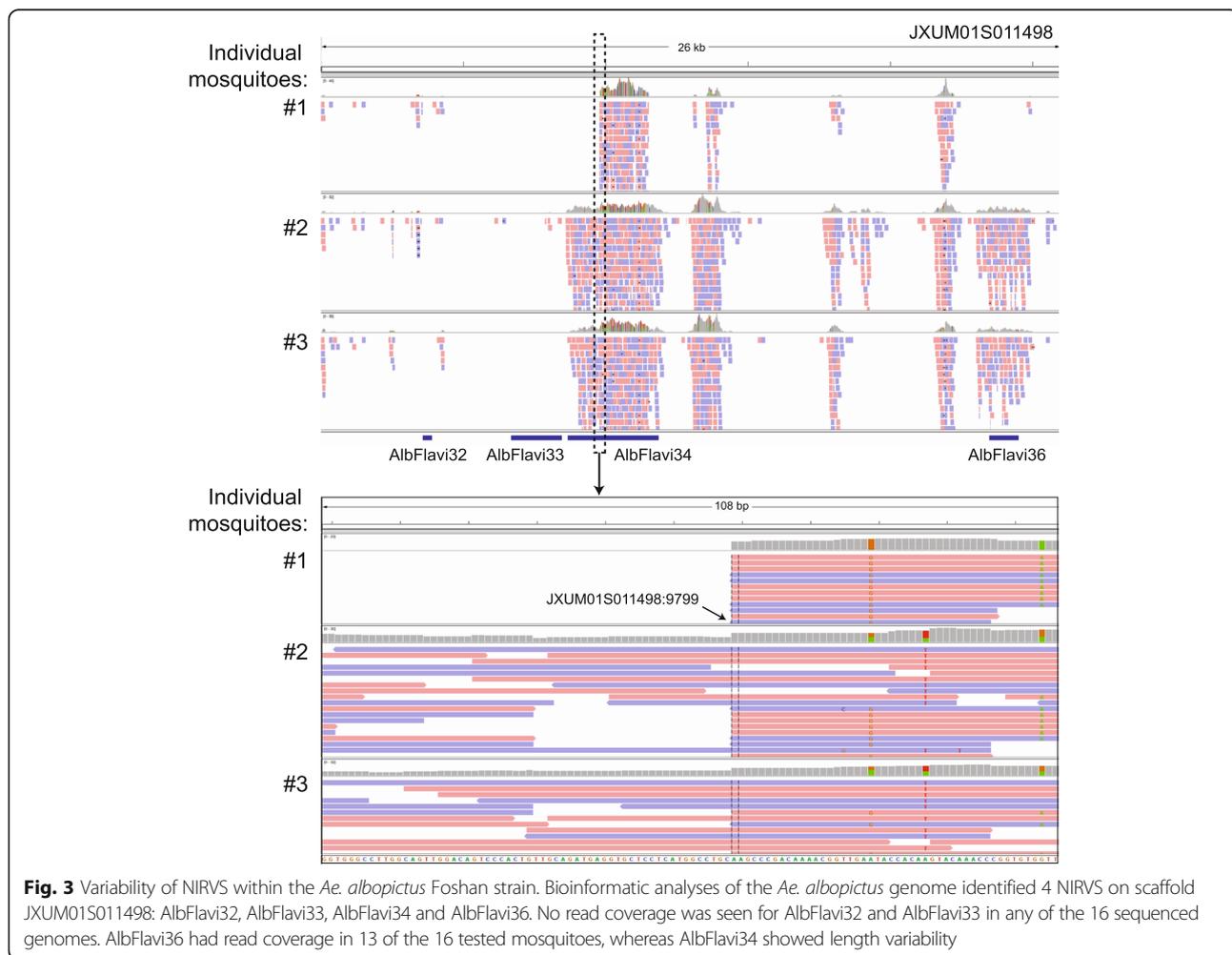
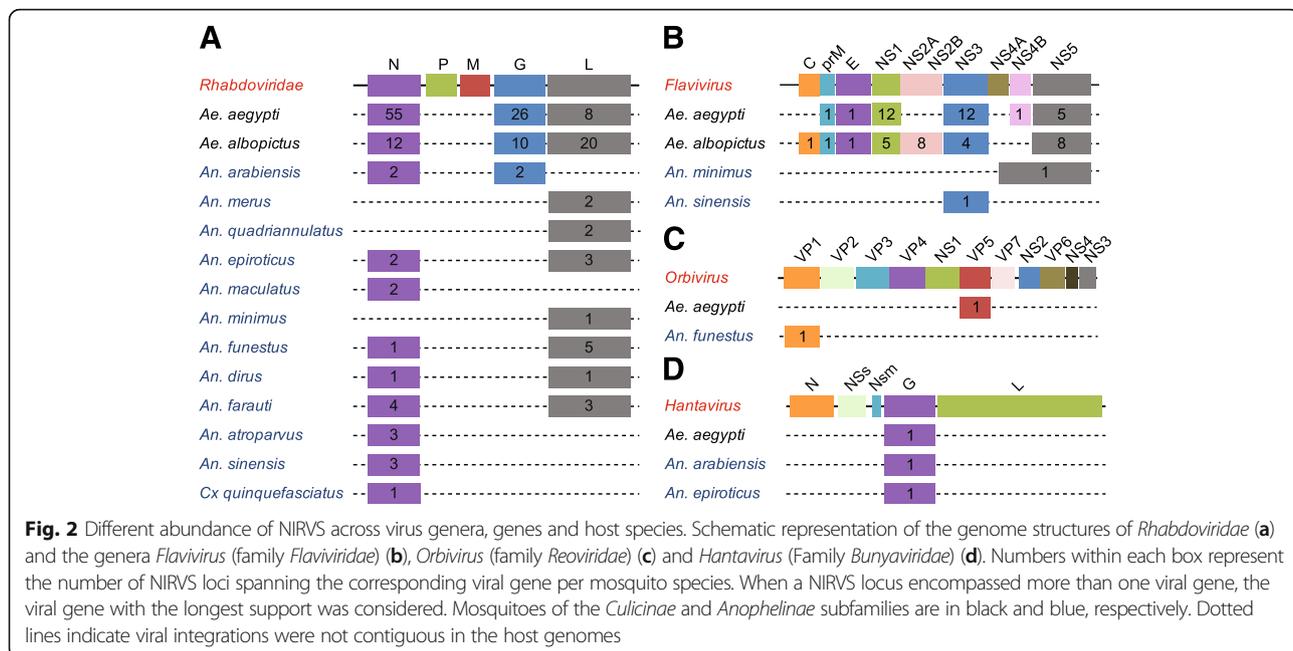
Mosquito species	Families of tested non-retroviral RNA viruses (N. species)									
	Toga <sup>a</sup> (24)	Flavi <sup>a</sup> (92)	Bunya <sup>a</sup> (59)	Reo <sup>a</sup> (70)	Orthomyxo <sup>a</sup> (4)	Rhabdo <sup>a</sup> (93)	Borna (6)	Filo (8)	Nyami (4)	Paramyxo (64)
<i>Aedes aegypti</i>	0	<b>32</b>	<b>1</b>	<b>1</b>	0	<b>88</b>	0	0	0	0
<i>Aedes albopictus</i>	0	<b>30</b>	0	0	0	<b>42</b>	0	0	0	0
<i>Culex quinquefasciatus</i>	0	0	0	0	0	<b>1</b>	0	0	0	0
<i>Anopheles christy</i>	0	0	0	0	0	0	0	0	0	0
<i>Anopheles gambiae</i>	0	0	0	0	0	0	0	0	0	0
<i>Anopheles coluzzi</i>	0	0	0	0	0	0	0	0	0	0
<i>Anopheles arabiensis</i>	0	0	<b>1</b>	0	0	<b>4</b>	0	0	0	0
<i>Anopheles melas</i>	0	0	0	0	0	0	0	0	0	0
<i>Anopheles merus</i>	0	0	0	0	0	<b>2</b>	0	0	0	0
<i>Anopheles quadrianulatus</i>	0	0	0	0	0	<b>2</b>	0	0	0	0
<i>Anopheles epiroticus</i>	0	0	0	0	0	<b>7</b>	0	0	0	0
<i>Anopheles stephensi</i>	0	0	0	0	0	0	0	0	0	0
<i>Anopheles maculatus</i>	0	0	0	0	0	<b>2</b>	0	0	0	0
<i>Anopheles culicifacies</i>	0	0	0	0	0	0	0	0	0	0
<i>Anopheles minimus</i>	0	<b>1</b>	0	0	0	<b>1</b>	0	0	0	0
<i>Anopheles funestus</i>	0	0	0	<b>1</b>	0	<b>7</b>	0	0	0	0
<i>Anopheles dirus</i>	0	0	0	0	0	<b>4</b>	0	0	0	0
<i>Anopheles farauti</i>	0	0	0	0	0	<b>7</b>	0	0	0	0
<i>Anopheles atroparvus</i>	0	0	0	0	0	<b>3</b>	0	0	0	0
<i>Anopheles sinensis</i>	0	<b>1</b>	0	0	0	<b>1</b>	0	0	0	0
<i>Anopheles albimanus</i>	0	0	0	0	0	0	0	0	0	0
<i>Anopheles darlingi</i>	0	0	0	0	0	0	0	0	0	0

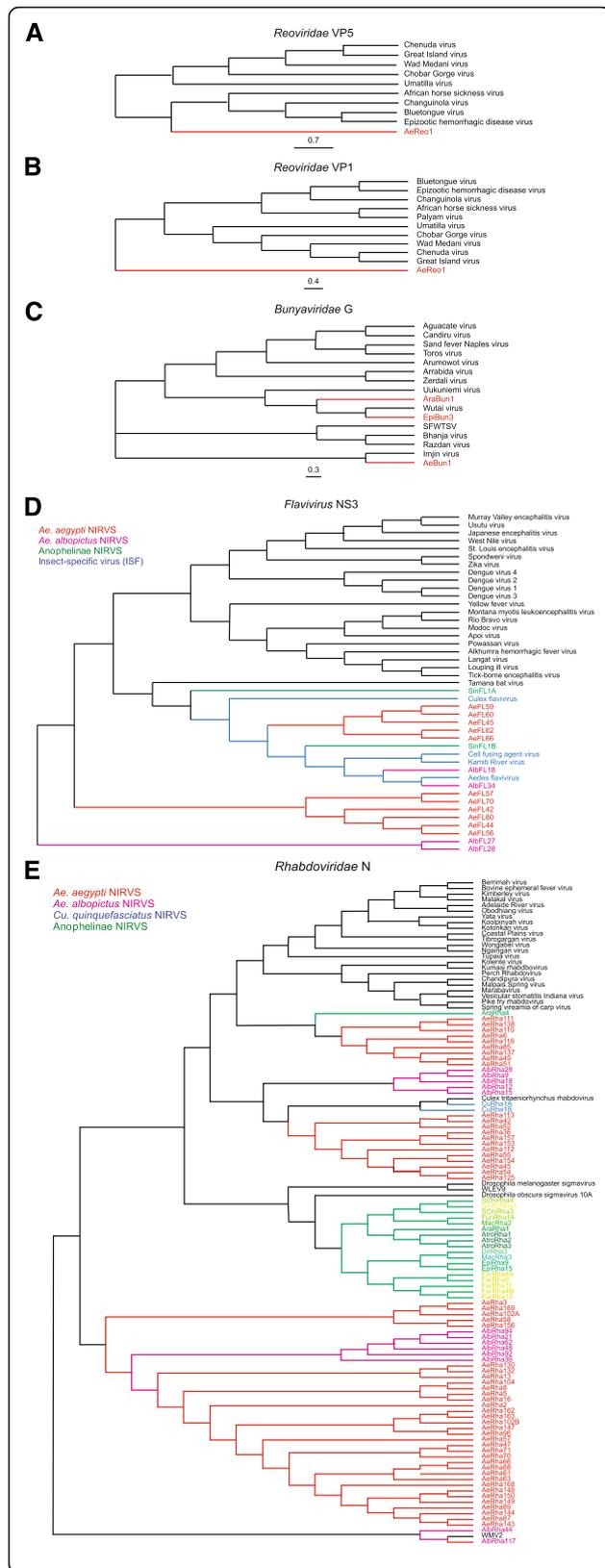
<sup>a</sup>Virus families that contain arboviruses

clustered with newly identified viruses such as Imjin virus and Wutai mosquito virus [40, 41] (Fig. 4c). In contrast, we observed numerous integrations from viruses of different genera within the *Rhabdoviridae* family and from viruses of the *Flavivirus* genus in multiple mosquito species, predominantly in *Ae. aegypti* and *Ae. albopictus* (Fig. 4d, e). *Rhabdoviridae*-like NIRVS (R-NIRVS) had similarities to genes encoding Nucleoprotein (N), Glycoprotein (G) and the RNA-dependent RNA polymerase (L), the relative abundance of which differed across mosquito species. We did not detect integrations corresponding to the matrix (M) or phosphoprotein (P) genes, consistent with observations in other arthropods [28]. R-NIRVS from *Culicinae* and *Anophelinae* formed separate clades in phylogenetic trees, supporting the conclusion that independent integrations occurred in the two mosquito lineages (Fig. 4e). *Flavivirus*-like NIRVS (F-NIRVS) with similarities to structural genes (envelope [E], membrane [prM] and capsid [C]) were less frequent than integrations corresponding to

non-structural genes (Fig. 2). Some NIRVS within one mosquito genome were highly similar to each other (nucleotide identity >90%) (Additional file 5: Table S5), which suggests that these were duplicated in the genome after a single integration event. This interpretation is also supported by the genomic proximity of several of these NIRVS (Fig. 5). Some identical NIRVS in *Ae. aegypti*, however, were found at locations that are physically unlinked (i.e. AeRha138, AeRha110 and AeRha111). Thus, we cannot determine whether these identical NIRVS represent recent independent integration events or arose from duplication or ectopic recombination after integration.

Generally, NIRVS were most similar to insect-specific viruses (ISVs), which replicate exclusively in arthropods, but are phylogenetically-related to arboviruses [10, 42] (Fig. 4d). However, we observed integrations that were most similar to arboviruses of the *Vesiculovirus* genus (*Rhabdoviridae*) in both *Ae. aegypti* and *Ae. albopictus* (Additional file 3: Table S3).



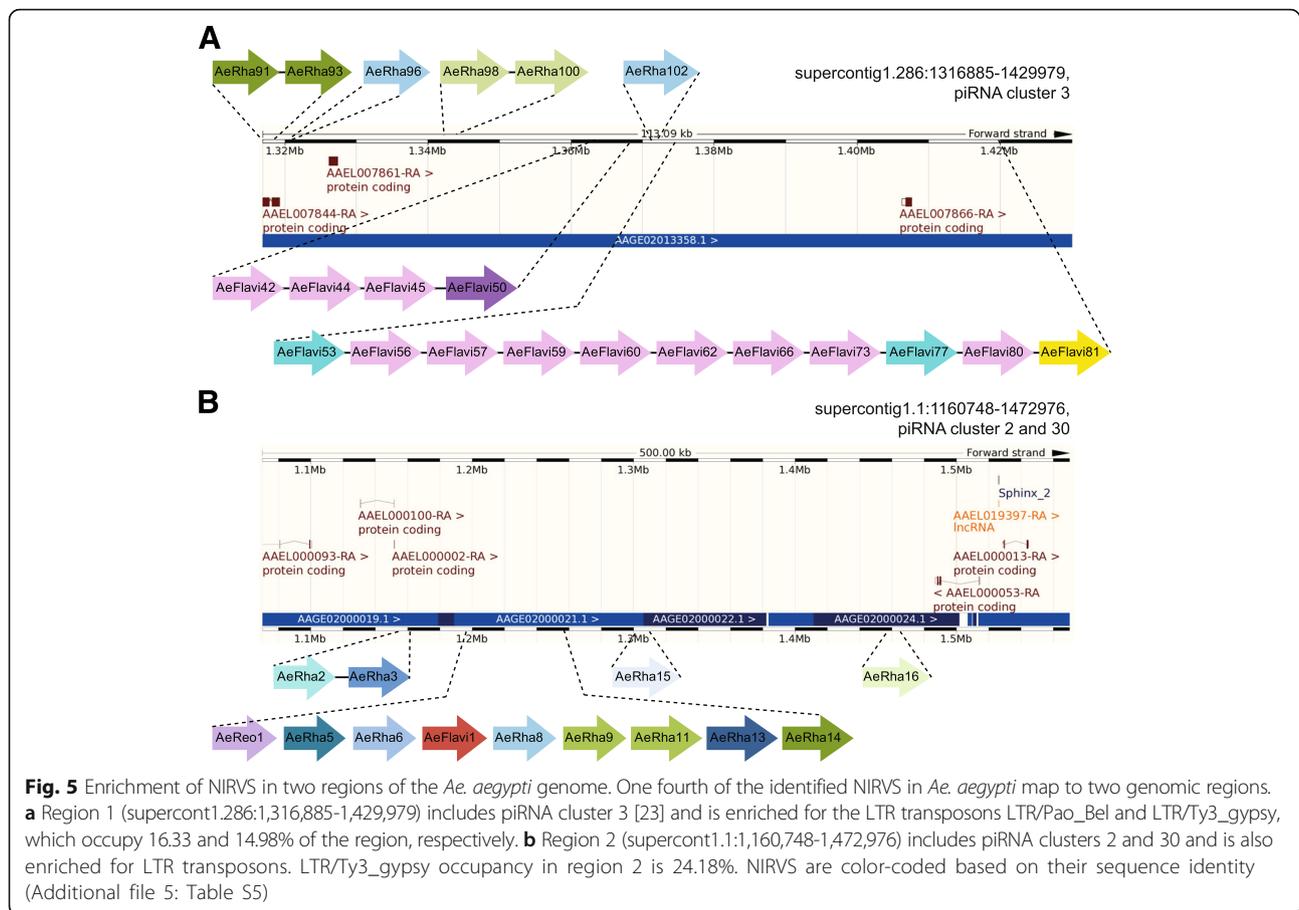


**Fig. 4** Phylogenetic analyses of *Reoviridae*, *Bunyaviridae*, *Flavivirus*, and *Rhabdoviridae*-like integrations. Phylogenetic relationships of NIRVS with similarity to the *Reoviridae* VP5 (a), *Reoviridae* VP1 (b), *Bunyaviridae* G (c), *Flavivirus* NS3 (d), and *Rhabdoviridae* N (e) genes. The evolutionary history was inferred using the Maximum Likelihood method. The trees with the highest log likelihood are shown. Support for tree nodes was established after 1000 bootstraps

**NIRVS produce piRNAs and map in piRNA clusters more frequently than expected by chance**

To better understand the mechanisms of integration, we analysed in greater detail the genomic context of NIRVS in *Ae. aegypti* and *Ae. albopictus*, the mosquitoes with the largest number of identified NIRVS. Previously, uncharacterized viral sequences were identified as piRNA producing loci in *Ae. aegypti* [23, 43], and these observations prompted us to analyze whether NIRVS are enriched in piRNA clusters. Currently annotated piRNA clusters represent 1.24% of the *Ae. aegypti* genome and 0.61% of *Ae. albopictus* genome [23, 44]. Remarkably, 44% and 12.5% of all NIRVS map to these genomic loci, which is significantly higher than expected by chance (Table 2). Enrichment of NIRVS in piRNA clusters in *Ae. aegypti* was driven by two regions that harbored one fourth of all NIRVS loci and span three piRNA clusters (Fig. 5). Region1 maps on scaffold 1.286, between 1,316,885 and 1,429,979 bp and includes piRNA cluster 3; region 2 is located on scaffold 1.1, between 1,160,748–1,472,976 bp, and includes piRNA clusters 2 and 30 [23]. In these two regions, NIRVS span partial ORFs with similarities to different *Rhabdovirus* and *Flavivirus* genes, with instances of duplications as well as unique viral integrations. NIRVS also were enriched in regions annotated as exons in *Ae. albopictus*, but not in *Ae. aegypti* (Table 2).

The presence of NIRVS in piRNA clusters prompted us to analyze the expression of NIRVS-derived small RNAs. Therefore, we used deep-sequencing data from published resources and mapped small RNAs on NIRVS sequences after collapsing those elements that shared identical sequences (Additional file 5: Table S5). Small RNAs in the size range of piRNAs (25–30 nucleotides), but not siRNAs (21-nucleotides) mapped to NIRVS in both *Ae. aegypti* and *Ae. albopictus*, independently of genomic localization and viral origin (Fig. 6a,b). Generally, piRNAs derived from individual NIRVS sequences are not highly abundant. Of all tested NIRVS, 43% ( $n = 33$ ) and 11% ( $n = 6$ ) had at least 10 piRNA reads per million genome-mapped reads in *Ae. aegypti* and *Ae. albopictus*, respectively. In *Ae. aegypti*, the highest piRNA counts were a few hundred reads per million genome-mapped reads. In *Ae. albopictus*, the maximum piRNA counts per NIRVS were about 10 fold lower, suggesting that NIRVS piRNA are less efficiently produced



or retained in this species. In both mosquito species, R-NIRVS showed higher piRNA coverage than F-NIRVS (Fig. 6e). NIRVS-derived piRNAs were biased for uridine at position 1 and primarily in antisense orientation to the predicted viral ORF, establishing the potential to target viral mRNA (Fig. 6a-d). Yet, a 10A bias of sense piRNAs, particularly in *Ae. albopictus* indicates that some NIRVS produce piRNAs through ping-pong amplification. Interestingly, ping-pong dependent secondary piRNAs seem to be almost exclusively (100% in *Ae. aegypti* and >99.5% in *Ae. albopictus*) derived from

R-NIRVS (Fig. 6e). The basis for this specific induction of secondary piRNAs from Rhabdoviral sequences is currently unknown.

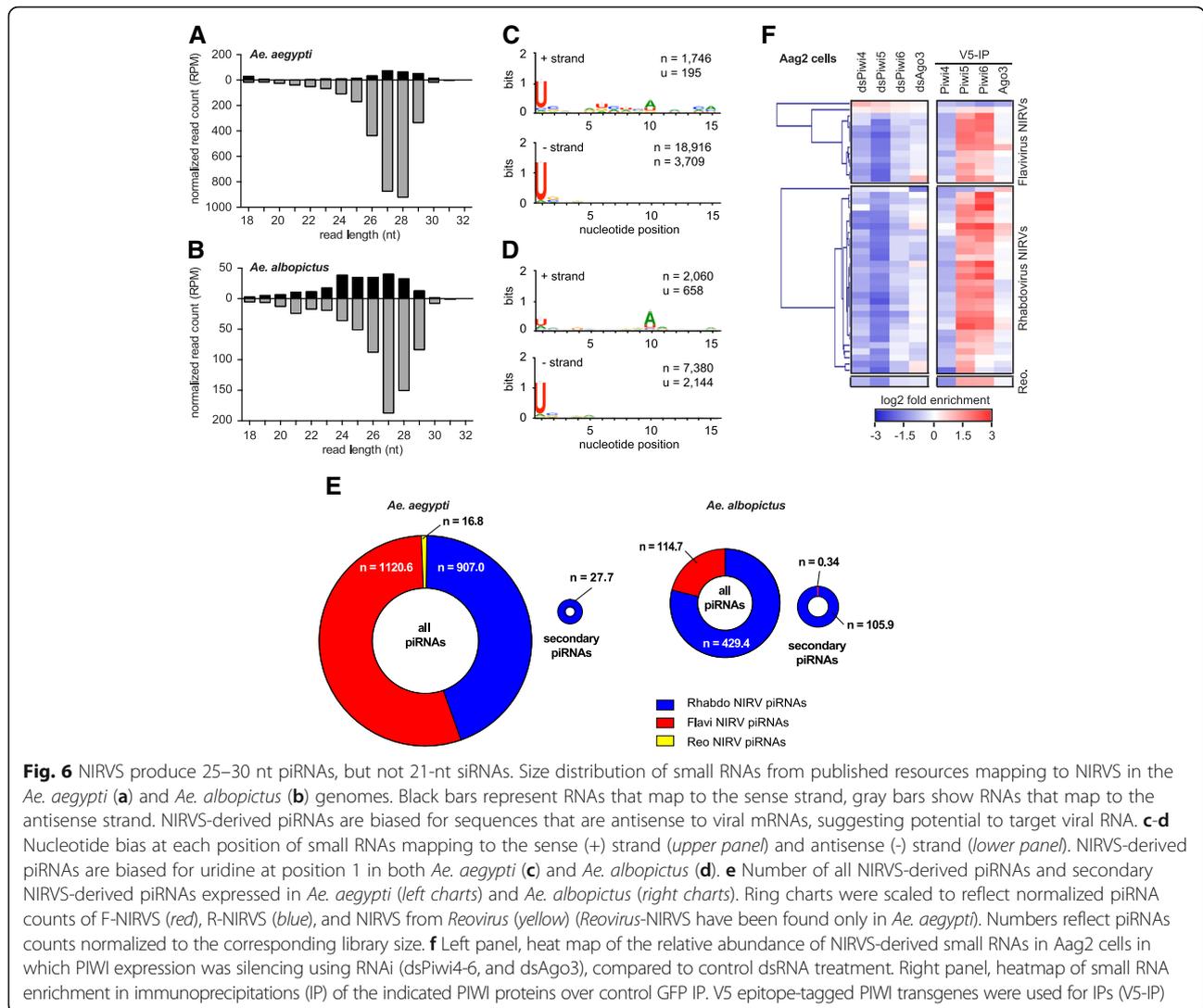
We next analysed the dependency on and association with PIWI proteins of NIRVS-derived small RNAs in Aag2 cells [45] and found that small RNA expression was reduced by knockdown of Piwi5 and, to a lesser extent, Piwi4 and Piwi6 (Fig. 6f), with only few exceptions. Consistent with this finding, NIRVS-derived small RNAs were enriched in immunoprecipitations (IP) of Piwi5 and Piwi6 (Fig. 6f). Together, these data indicate that

**Table 2** Clustering of viral integrations (NIRVS) in piRNA loci of the *Ae. aegypti* and *Ae. albopictus* genomes

Host	Genomic region	Length (bp)	% genome	k integrations <sup>a</sup>	P-value
<i>Ae. aegypti</i>	piRNA cluster	17,000,000	1.24	54	< 10 <sup>-10</sup>
	Coding genes	286,538,182	20.82	24	0.66
	Intergenic regions	1,072,461,818	77.94	44	1
<i>Ae. albopictus</i>	piRNA cluster	1,926,670	0.61	9	< 10 <sup>-10</sup>
	Coding genes	163,407,667	8.26	14	2.08 10 <sup>-3</sup>
	Intergenic regions	1,803,592,333	91.14	49	1

The probability (P) of observing k NIRVS loci in piRNA clusters, coding genes and intergenic regions. P was estimated using cumulative binomial distribution; a value of P < 0.05 indicates a statistically significant enrichment of NIRVS in the corresponding genomic region

<sup>a</sup>Six integrations in the *Ae. aegypti* genome were in exons of genes within piRNA clusters; in these analyses they were attributed to piRNA clusters. Statistical significance did not change when these integrations were assigned to coding genes (P changed from 0.66 to 0.180)



NIRVS produce piRNAs, the majority of which have the characteristics of primary piRNAs. Yet, secondary piRNA biogenesis as indicated by a 10A bias and association with Ago3, seems to occur specifically from R-NIRVS.

**NIRVS and transposable elements**

piRNA clusters in *D. melanogaster* are enriched for remnants of TE sequences, and it is likely that vDNA is produced by the reverse transcriptase activity of TEs [36, 37]. Moreover, NIRVS-derived piRNAs resemble the characteristics of TE-derived primary piRNAs in their antisense 1 U bias and enrichment in Piwi5 and Piwi6 protein complexes. We therefore analysed the transposon landscape of NIRVS loci by systematically identifying all annotated TEs in the 5 and 10 kb genomic regions flanking each side of the NIRVS integration. We observed that NIRVS were predominantly associated with

long terminal repeat (LTR) retrotransposons. Within LTR-retrotransposons, we observed enrichment of members of the Ty3\_gypsy family (Table 3). Ty3\_gypsy enrichment was even more pronounced in the two regions in *Ae. aegypti* where 40% of NIRVS reside (Fig. 5). While LTR retrotransposon occupancy was 12.34% across the entire *Ae. aegypti* genome, it reached 23.60–25.88%, 31.35%, and 30.55% in regions flanking all NIRVS-loci, region 1, and region 2, respectively. More strikingly, whereas the occupancy of the Ty3\_gypsy family of LTR retrotransposons was 2.58% across the entire *Ae. aegypti* genome, it reached 14.7–17.5%, 14.98% and 24.18% in regions flanking all NIRVS-loci, region 1, and region 2, respectively (Table 3). Nine full-length TEs were found flanking NIRVS-loci, seven of which are Ty3\_gypsy retrotransposons. For example, 3 copies and 1 copy of the full-length Ty3\_gypsy\_Ele58 (TF000321) were found in regions 1 and 2, respectively. Moreover, one viral integration in *Ae. aegypti*

**Table 3** NIRVS and transposable elements (TEs)

TE group <sup>a</sup>	TE occupancy (%)			
	AaegL3 <sup>b</sup>	NIRVS <sup>c</sup>	Region 1 <sup>d</sup>	Region 2 <sup>e</sup>
LTR retrotransposons	12.34	23.06 (25.88)	31.35	30.56
LTR/Pao_Bel	4.42	6.9 (6.49)	16.33	4.15
LTR/Ty1_copia	5.34	1.46 (1.90)	0.04	2.22
LTR/Ty3_gypsy	2.58	14.7 (17.50)	14.98	24.18
non-LTR retrotransposons	12.81	3.91 (5.54)	0	3.43
SINEs	1.14	0.16 (0.22)	0	0
DNA transposons	6.96	3.29 (3.49)	3.38	2.72
MITEs	12.81	8.03 (7.34)	0.26	2.34
Helitrons	1.2	2.01 (2.32)	0	2.04
Penelope	0.42	0.2 (0.28)	0.26	0.8

TE occupancy in the entire *Ae. aegypti* genome (AaegL3), in regions harboring NIRVS (NIRVS), in region1 and in region 2

<sup>a</sup>For consistency with previous publications and for unambiguous classification, only TEs annotated in TEfam were used. TE occupancy was defined as the number of bases in the genomic sequence that match TEs

<sup>b</sup>The genome assembly described in [71] is slightly different from AaegL3 (*Aedes-aegypti*-Liverpool\_SCAFFOLDS\_AaegL3.fa), which is used in this analysis. For better comparison with viral integration sites, a new RepeatMasker analysis was performed using the AaegL3 assembly under default parameters

<sup>c</sup>The NIRVS sites plus 5 kb or 10 kb (in brackets) of sequence flanking each side were retrieved for the analysis. These results may under-estimate the actual TE occupancy because NIRVS sequences are included in the analyses

<sup>d</sup>Supercontig 1.286 between positions 1,316,885 bp and 1,429,979 bp (plus 5 kb flanking each side)

<sup>e</sup>Supercontig 1.1 between positions 1,160,748 bp and 1,472,976 bp (plus 5 kb flanking each side)

(i.e. AeBunya1) was found embedded within a full-length TE of the Pao-Bel family.

### NIRVS transcriptional activity

All NIRVS encompassed partial viral ORFs, with the exception of AlbFlavi34. AlbFlavi34 corresponds to a portion of the first *Flavivirus*-like sequence characterized in mosquitoes and includes a complete ORF for NS3 [24]. Two alleles of different length were seen for AlbFlavi34 in the 16 sequenced *Ae. albopictus* genomes (Fig. 3). The short allele, which interrupts the NS3 ORF, had a frequency of 53% (Additional file 6). Based on recent experimental data showing that NIRVS are transcriptionally active even if they do not encode a complete ORF [24, 28, 32, 46] we analysed NIRVS expression using published RNA-seq data from poly(A) selection protocols. Expression levels were <5 reads per kilobase per million mapped reads (RPKM) for >92% of all tested NIRVS, including NIRVS that produce piRNAs (Additional file 7: Table S6). Similar to small RNA profiles, expression levels of R-NIRVS were higher than those of F-NIRVS (Additional file 7: Table S6). Even though RNA-seq data showed limited transcriptional activity for AlbFlavi34 (RPKM values ranging from 0.009 to 0.013), we analysed its expression in different developmental stages by qRT-PCR using primers that amplify both the short and long

alleles. Cycle threshold (Ct) values ranged from 27 (in pupae) to 39.34 (in ovaries of blood-fed females), with the highest expression in pupae and adult males (Additional file 8: Figure S1). Yet, 60% of the samples had Ct values >30, confirming that AlbFlavi34 is expressed at low levels. Together, these data support the conclusion that steady-state RNA levels of most NIRVS are rather low or even undetectable. Yet, the production of piRNAs indicates that they must be transcriptionally active. Whether piRNA precursor transcripts are non-polyadenylated or rapidly processed into piRNAs remains to be established.

### Discussion

The genomes of mosquitoes and several eukaryotic species carry integrations from non-retroviral RNA viruses, including arboviruses. To shed light on the occurrence and biological significance of this phenomenon, we analysed the presence, distribution and transcriptional activity of integrations from 424 non-retroviral RNA viruses and one DNA arbovirus in 22 mosquito genomes, in context of both their phylogeny and mosquito vector competence. We showed that the genomes of the arboviral vector species *Ae. aegypti* and *Ae. albopictus* contain ten-fold more integrations than all other tested mosquitoes. Moreover, we found that viral integrations produce piRNAs and occur predominantly in piRNA clusters. Our results support the conclusion that the abundance of viral integrations is not dependent on viral exposure, but seems to correlate with the TE landscape and piRNA pathway of the mosquito species.

### Viral origin of NIRVS

Across all 425 viral species tested, viral integrations had similarities primarily to ISVs of the families *Bunyaviridae*, *Reoviridae* and, predominantly, *Rhabdoviridae* and *Flaviviridae*. Notably, although the *Togaviridae* family contains mosquito-borne members as well as insect-specific viruses, we identified no integrations from viruses of this family. Further studies are required to clarify whether this result is due to a sampling bias or to the different evolutionary history of *Alphaviruses* versus *Flaviviruses* [38]. For instance, Eilat virus and Taï Forest alphavirus are thus far the only insect-specific alphaviruses (family *Togaviridae*) identified, suggesting low abundance of mosquito-specific viruses among *Togaviridae* [47], unlike the families *Bunyaviridae*, *Reoviridae*, *Rhabdoviridae* and *Flaviviridae* in which many ISVs have been identified [42]. An alternative explanation may be based on different interactions of these viruses with the piRNA machinery. For example, whereas both alphaviruses and flaviviruses produce viral piRNAs in *Aedes*, the distribution of piRNAs across the viral genome is not comparable between these genera, suggesting that piRNA biogenesis might be different

[21]. Both alphaviruses (Sindbis virus, CHIKV) and flaviviruses (DENV, WNV) have been shown to produce nuclear episomal vDNA forms after infection of mosquitoes [20, 36, 37]. These vDNA forms do not arise uniformly from the viral genome and their profile may be different between alphaviruses and flaviviruses [37]. If episomal vDNAs are the source of viral integrations in the genome, this would explain differences in NIRVS landscapes of these viruses.

ISVs of the families *Bunyaviridae*, *Flaviviridae* and *Rhabdoviridae* are ancient and diversified within their hosts, and they seem to be maintained in mosquito populations through transovarial transmission [10, 42]. Additionally, mounting phylogenetic evidence implicate ISVs as precursors of arboviruses [48], for which vertical transmission occurs at a lower frequency than horizontal transmission through a vertebrate host [49]. Vertical transmission provides access to the mosquito germ-line, a mechanism through which NIRVS could be maintained within vector populations. Thus, the observed higher incidence of NIRVS from ISVs than arbovirus may be linked to differences in the frequency of their transovarial transmission.

NIRVS from *Bunyaviridae* and *Rhabdoviridae* have been identified in insects other than mosquitoes, including different *Drosophila* species and the tick *Ixodes scapularis* [26–28]. In contrast, NIRVS from flaviviruses have been found only in mosquitoes, predominantly in *Ae. aegypti* and *Ae. albopictus* [2, 26, 32]. Interestingly, vertebrates that may be part of the arbovirus transmission cycle do not have integrations from arboviruses, but a low number (<10) of integrations from bornaviruses and/or filoviruses have been identified in the genomes of humans, squirrel, microbat, opossum, lemur, wallaby and medaka [25–27]. Finally, several *Anophelinae* mosquitoes analysed here were sampled in the same geographic area as *Ae. albopictus*, but showed ten times fewer NIRVS than *Ae. albopictus* and *Ae. aegypti*. Overall, these data indicate that viral exposure is not a determinant of NIRVS, but that virus-host lineage-specific interactions play a crucial role in how their genomes co-evolve. Additionally, our comparative analysis shows that *Aedes* mosquitoes acquire and retain fragments of infecting non-retroviral RNA viruses of primarily the *Flaviviridae* and *Rhabdoviridae* families, more frequently than other tested arthropods and vertebrates. A deeper understanding of the evolution of viruses within these large and diverse families along with insights into the variability of the genomes of mosquito populations is warranted to elucidate the dynamic species-specific interactions between RNA viruses and *Aedes* mosquitoes.

#### Genomic context of NIRVS

NIRVS are significantly enriched in piRNA clusters in both *Ae. aegypti* and *Ae. albopictus*, which could be the

result of positive selection favoring the retention of those NIRVS that integrated by chance in these genomic loci [50]. However, we also observed NIRVS in intergenic and coding sequences and found that NIRVS expressed piRNAs independently of their genomic localization. These observations suggest that additional piRNA clusters exist [23, 44] or that other features in these NIRVS loci prime piRNAs production. For example, a piRNA trigger sequence (PTS) was recently found to drive piRNA production from a major piRNA cluster (named *Flamenco*) in *Drosophila* [51]. We analysed the mosquito genome sequences, but we did not find orthologous sequences to the *Flamenco* PTS in neither *Ae. aegypti* nor *Ae. albopictus*. It remains to be established whether other PTS sequences exist that may explain piRNA production from non-cluster associated NIRVS.

Analyses of the integration sites showed that NIRVS are primarily associated with LTR transposons of the Gypsy and Pao families, which are the most abundant TE families in both the *Ae. aegypti* and *Ae. albopictus* genomes [2]. Additionally, full-length TEs, primarily Ty3\_gypsy retrotransposons, were found to flank NIRVS-loci. This organization is compatible with recent experimental data showing that vDNA forms are produced by retrotransposon-derived reverse transcriptase, likely by template switching [20, 37]. This arrangement is also favorable for ectopic recombination, a mechanism proposed for both NIRVS biogenesis and piRNA cluster evolution [52]. Ectopic recombination would be a more parsimonious explanation than independent integrations from the same viral source for our finding of several not physically-linked, but identical NIRVS in *Ae. aegypti*. Despite many remaining uncertainties due to the highly repetitive and complex structure of the regions in which NIRVS map, these data confirm a functional link among NIRVS, TEs, and the piRNA pathway.

#### NIRVS and mosquito immunity

Our data indicate that NIRVS in *Ae. aegypti* and *Ae. albopictus* do not encode proteins that interfere *in trans* with viral proteins as was observed for bornavirus-derived NIRVS in vertebrates [53]. Rather our data suggest that NIRVS may be part of a piRNA-based antiviral response. Only one of the characterized NIRVS had a complete viral ORF, which showed two alleles of different length within the 16 individuals of the Foshan strain that we sequenced. The short variant interrupted the NS3 ORF. We cannot exclude that this is due to lack of purifying selective pressure as the *Ae. albopictus* Foshan strain has been reared under standard laboratory conditions without infection challenges for more than 30 years [2]. However, the enrichment of NIRVS within piRNA clusters and their small RNA profile suggest that their

transcriptional activity is geared to produce piRNA precursors.

Our results show a basal expression of NIRVS-derived primary piRNAs that are antisense to viral mRNA. These piRNAs could block novel infections with cognate viruses or they could interact with RNAi mechanisms to contain replication of incoming viruses at a level that does not become detrimental to mosquitoes. Albeit leading to opposite effects on vector competence, both mechanisms display functional similarities to the CRISPR-Cas system of prokaryotic adaptive immunity. Even if further studies are essential to clarify the effect of NIRVS-derived piRNAs on mosquito immunity, our study clearly demonstrates that *Ae. aegypti* and *Ae. albopictus* have a high number of NIRVS in their genome, which confers heritable immune signals.

The higher number of NIRVS in *Aedine* than in *Anopheline* mosquitoes correlates with competence for a larger number of arboviruses of *Aedine* mosquitoes. In this regard, *Cx. quinquefasciatus* shows an interesting intermediate phenotype because it is phylogenetically closer to *Aedine* mosquitoes, but vectors a smaller range of arboviruses, and, like *Anophelinae*, it can vector more protozoans and nematodes than *Aedine* [54]. Additionally, *Cx. quinquefasciatus* has a number of NIRVS and TE load comparable to *Anopheline*, but an expanded PIWI gene family like *Ae. aegypti* [2, 55, 56].

## Conclusions

NIRVS are regarded as viral fossils, occurring as occasional events during the long co-evolutionary history of viruses and their hosts [33, 35]. The high abundance and diversity of NIRVS in the genomes of *Ae. aegypti* and *Ae. albopictus* and the observation that NIRVS produce piRNAs and reside in piRNA clusters support the intriguing hypothesis that the formation and maintenance of NIRVS is coupled with the evolution of the PIWI pathway in these two species. This may have led to functional specialization of the expanded PIWI gene family, PIWI expression in the *soma*, and a role for the piRNA pathway in antiviral immunity [21, 45]. This hypothesis is compatible with two scenarios. First, NIRVS formation is an occasional event, which occurs more frequently in *Aedine* mosquitoes than in *Culex* and *Anophelinae* because of the higher abundance of retrotransposons in the genome of *Aedine* mosquitoes [2]. NIRVS that have integrated by chance into piRNA clusters produce transcripts that are shuttled into the piRNA pathway. PIWI proteins loaded with viral sequences may target incoming viruses, possibly conferring selective advantage. Thus, an occasional event linked to a particular TE landscape may be the trigger for the functional specialization of PIWI proteins. This scenario remains compatible with the possibility that NIRVS outside of piRNA loci encode

protein products that compete *in trans* with virus replication, thereby affecting vector competence [57]. Second, it has been hypothesized that PIWI proteins actively interact with incoming viruses and that they are loaded with episomal vDNAs and integrate them into piRNA clusters [58]. Under this scenario, the selective pressure favoring PIWI protein specialization would come primarily from viruses. Taken together our data show that the interaction between viruses and mosquitoes is a more dynamic process than previously thought and that this interplay can lead to heritable changes in mosquito genomes.

## Methods

### In silico screening of viral integrations

Genome assemblies of *D. melanogaster* and 22 currently available mosquito species were screened in silico using tBLASTx and a library consisting of genome sequences of 424 non-retroviral RNA viruses and one DNA arbovirus (Additional file 1: Table S1).

Tested mosquito species were classified in arboviral vectors (*Aedes aegypti*, *Aedes albopictus*, *Culex quinquefasciatus*) and protozoan vectors (*Anopheles gambiae*, *Anopheles albimanus*, *Anopheles arabiensis*, *Anopheles darlingi*, *Anopheles stephensi*, *Anopheles funestus*, *Anopheles atroparvus*, *Anopheles coluzzii*, *Anopheles culicifascies*, *Anopheles dirus*, *Anopheles epiroticus*, *Anopheles farauti*, *Anopheles maculatus*, *Anopheles melas*, *Anopheles merus*, *Anopheles minimus*, *Anopheles sinensis*), depending on whether they most efficiently transmit arboviruses or protozoans to humans, respectively (Additional file 2: Table S2). The non-vector *Anopheles christyi* and *Anopheles quadriannulatus* were also included in the analyses [59].

Host genome sequences of at least 100 bp and with high identity (e-values <0.0001) to viral queries were extracted from the respective insect genomes using custom scripts. When several queries mapped to the same genomic region, only the query with the highest score was retained. Blast hits were considered different when they mapped to genomic positions at least 100 bp apart from each other, otherwise they were included in the same NIRVS locus.

All putative viral integrations were subjected to a three-step filtering process before being retained for further analyses to reduce the chance of false positives and ensure that the identified sequences are from non-retroviral RNA viruses [25]. Filtering steps included 1) a reverse-search against all nucleotide sequences in the NCBI database using the BLAST algorithm, 2) a search for ORFs encompassing viral proteins based on NCBI ORFfinder, and 3) a functional annotation based on Argot<sup>2</sup> [60].

While our search expanded the range of viral integrations identified in *Ae. albopictus* and *Ae. aegypti* [2, 26, 28], it is likely that refinements of the current genome annotations of the species analysed, especially in repeat regions, the application of alternative bioinformatic pipelines, and the identification of novel viral species will lead to the characterization of additional integrations. Additionally, to reduce chance of false positives, our bioinformatics pipeline focused on sequences in which we could unambiguously identify viral ORFs, thus excluding viral sequences coming from UTRs or non coding regions.

#### Genomic data from 16 *Ae. albopictus* mosquitoes

The genomes of 16 individual mosquitoes of the *Ae. albopictus* Foshan strain were sequenced. The strain was received from Dr. Chen of the Southern Medical University of Guangzhou (China) in 2013. Since 2013, the Foshan strain has been reared in an insectary of the University of Pavia at 70–80% relative humidity, 28 °C and with a 12–12 h light–dark photoperiod. Larvae are fed on finely-ground fish food (Tetramin, Tetra Werke, Germany). A membrane feeding apparatus and commercially available mutton blood is used for blood-feeding females.

DNA was extracted from single mosquitoes using the DNeasy Blood and Tissue Kit (Qiagen, Hilden Germany) following the manufacturer's protocol. DNA was shipped to the Polo D'Innovazione Genomica, Genetica e Biologia (Siena, Italy) for quality control, DNA-seq library preparation and sequencing on an Illumina HiSeq 2500. After quality control, retrieved sequences were aligned to the *Ae. albopictus* Foshan reference genome (AaloF1 assembly) using the Burrows-Wheeler Aligner (BWA) [61] and marking identical read copies. The resulting indexed BAM files were used to calculate the counts of alignments, with mapping quality score above 10, that overlapped intervals of *Ae. albopictus* NIRVS using BEDTools [62]. Alignment files were visualized using the Integrative Genomics Viewer [63].

#### Phylogenetic analysis

Deduced NIRVS protein sequences were aligned with subsets of corresponding proteins from *Flavivirus*, *Rhabdovirus*, *Reovirus* and *Bunyavirus* genomes using MUSCLE. Maximum likelihood (ML) phylogenies were estimated in MEGA6 [64], implementing in each case the best fitting substitution model. Statistical support for inferred tree nodes was assessed with 1000 bootstrap replicates. Figures were generated using FIGTREE (v.1.4) (<http://tree.bio.ed.ac.uk/software/figtree/>).

#### Bioinformatic analysis of integration sites

Clustering of viral integrations in piRNA loci was estimated using cumulative binomial distribution, where the

probability of integration was assumed to equal the fraction of the genome occupied by the respective genomic region. Genomic regions considered were piRNA clusters, coding regions and intergenic regions as previously defined [2, 23, 44]. A value of  $P < 0.05$  suggests a statistically significant enrichment of integration events in the corresponding genomic region (Table 2).

Analysis of TE enrichment in all NIRVS sites and region 1 and region 2 of *Ae. aegypti* were based on RepeatMasker (version open-4.0.3, default parameters) using *A. aegypti* TEs retrieved from TEfam (<http://tefam.biochem.vt.edu/tefam/>), which was manually annotated. Percent TE occupancy (percent of bases in the genomic sequence that match TEs) was used as a measure for possible enrichment of certain TEs. TE copy numbers were not used because it is likely that some TEs are broken into multiple fragments and counted multiple times. We retrieved sequences of the viral integration sites plus 5 kb sequences flanking each side of the integration for the analysis. In addition, to identify potentially full-length TEs, 10 kb sequences flanking each side of the viral integration were analysed by RepeatMasker (version open-4.0.3). The presence of full-length TEs was verified by comparing the length of masked sequences with the length of the annotated TEs.

#### Analysis of piRNAs production from NIRVS

Small RNA deep-sequencing data of female *Ae. aegypti* (methoprene treated; SRX397102) [65], *Ae. albopictus* mosquitoes (sugar-fed; SRX201600) [66], and PIWI knockdown and IP libraries in Aag2 cells (SRA188616) [45] were downloaded from the European Nucleotide Archive. Subsequently, small RNA datasets were manipulated using the programs available in the Galaxy toolshed [67]. After removal of the 3' adapter sequences, small RNAs were mapped to NIRVS sequences that were oriented in the direction of the predicted ORF, using Bowtie permitting one mismatch in a 32 nt seed [68]. From the mapped reads, size profiles were generated. For the analysis of nucleotide biases, the 25–30 nt reads were selected and separated based on the strand. The FASTA-converted sequences of small RNA reads were then trimmed to 25 nt and used as input for the Sequence-Logo generator (Galaxy version 0.4 based on Weblogo 3.3) [69]. piRNA counts on individual NIRVS were generated by mapping to NIRVS sequences after collapsing (near-) identical sequences (Additional file 5: Table S5). Bowtie was used to map the small RNAs allowing one mismatch in a 32 nt seed. Only uniquely mapping reads were considered and the –best and the –strata options were enabled. From the mapped reads, 25–30 nt small RNAs were selected. To identify secondary piRNAs, reads in sense orientation to viral ORFs that had an adenine at position 10 were selected. To avoid

taking piRNAs into consideration that coincidentally contain a 10A, the population of 10A sense piRNAs was required to make up at least 50% of all sense piRNAs derived from the NIRVS. If this criterion was not met, sense reads from the corresponding NIRVS did not qualify as secondary piRNAs. Total piRNA counts and secondary piRNA counts were determined for F-NIRVS, R-NIRVS and *Reovirus* NIRVS and normalized to the corresponding library size. The size of the ring-graph was scaled to reflect the normalized read counts. piRNA counts on individual NIRVS was also determined from acetone treated female *Ae. aegypti*, male *Ae. aegypti*, blood-fed *Ae. albopictus* and male *Ae. albopictus* mosquitoes. The data were obtained from the same studies as described above.

To identify the PIWI dependency of NIRVS-derived piRNAs, we analysed libraries from Aag2 cells transfected with double stranded RNA (dsRNA) targeting the somatic PIWI genes (Piwi4-6, Ago3) and a non-targeting control (dsRNA targeting luciferase, dsLuc) [45]. These datasets were mapped against the collapsed NIRVS dataset as described above. Since small RNA profiles were dominated by piRNA-sized reads, no further size selection was performed. The mean fold change in small RNA read counts was calculated for each PIWI knock-down condition compared to the negative control. To identify the PIWI proteins that NIRVS piRNAs associate with, we analysed the IP libraries of PIWI proteins in Aag2 cells previously published in the same study. For the different PIWI IPs the enrichment of small RNA counts compared to a control GFP-IP was determined. Hierarchical clustering of NIRVS based on the combined fold changes of PIWI knockdowns and IPs was performed using multiple experiment viewer (Tm4). Clustering was based on Pearson correlation and performed independently for F-NIRVS and R-NIRVS.

### NIRVS transcriptional activity

RNA deep-sequencing data of *Ae. albopictus* and *Ae. aegypti* mosquitoes, including both DENV-infected and non-infected mosquitoes were downloaded from NCBI Sequence Read Archive. Libraries analysed correspond to data SRA438038 for *Ae. albopictus*, and SRA058076, SRX253218, SRX253219 and SRX253220 for *Ae. aegypti*. RNA-seq reads were mapped using BWA [61] to NIRVS, after collapsing identical sequences (Additional file 5: Table S5), and read counts were converted into RPKM using custom scripts.

To analyze AlbFlavi34 expression in different *Ae. albopictus* developmental stages, total RNA was extracted using Trizol reagent (Life Technologies) from 3 pools of 5 entities for each condition (eggs, larvae, adult males, blood-fed and sugar-fed females). From each pool, a total of 100 ng of RNA was used for reverse transcription using

the qScript cDNA SuperMix following the manufacturer's protocol (Quanta Biosciences). AlbFlavi34 expression was quantified in a 20  $\mu$ L final reaction volume containing 10  $\mu$ L of QuantiNova SYBR Green PCR Master Mix (Qiagen), 700 nM of each forward (5'-CTTGCGACCCATGGTCTTCT-3') and reverse (5'-GTCCTCGGCGCTGATCATA-3') primers and 5.0  $\mu$ L cDNA sample on an Eppendorf RealPlex Real-Time PCR Detection System (Eppendorf). We used a two-step amplification protocol consisting of 40 cycles of amplification (95 °C for 5 s, 60 °C for 10 s) after an initial denaturation of 2 min at 95 °C. AlbFlavi34 expression values were normalized to mRNA abundance levels of the *Ae. albopictus* Ribosomal Protein L34 (RPL34) gene [70]. QBASE+ software was used to visualise data and compare expression profiles across samples. Absence of *Flavivirus* infection was verified using a published protocol [24] on all samples before qRT-PCR.

### Additional files

**Additional file 1: Table S1.** List of Dipteran species analysed. (XLSX 34 kb)

**Additional file 2: Table S2.** List of viral genomes used for in silico screening of Dipteran genomes. The only DNA arbovirus is in yellow. (XLSX 38 kb)

**Additional file 3: Table S3.** Description of all identified NIRVS (XLSX 132 kb)

**Additional file 4: Table S4.** DNA-seq library statistics. Total number of reads, percentage of mapped reads and percentage of properly paired reads from DNA-seq of 16 individual *Ae. albopictus* mosquitoes (Foshan strain). (DOCX 20 kb)

**Additional file 5: Table S5.** Viral integrations (NIRVS) in the *Ae. aegypti* and *Ae. albopictus* genomes. Classification was based on their sequence identity. Integrations with sequence identity >90% were grouped together. (DOCX 20 kb)

**Additional file 6:** Details of AlbFlavi34. (DOCX 17 kb)

**Additional file 7: Table S6.** (A) piRNA coverage at NIRVS loci. (B) NIRVS transcriptional activity. (XLSX 90 kb)

**Additional file 8: Figure S1.** qRT-PCR based expression of AlbFlavi34 in different developmental stages of *Ae. albopictus*. (PDF 102 kb)

### Abbreviations

BWA: Burrows-Wheeler Aligner; CHIKV: chikungunya virus; Ct: cycle threshold; DENV: dengue viruses; EEE: eastern equine encephalitis; EVEs: Endogenous Viral Elements; G: Glycoprotein; IP: immunoprecipitations; ISV: Insect Specific Virus; L: RNA-dependent RNA polymerase; LACV: La Crosse virus; LTR: long terminal repeat (LTR) retrotransposons; ML: Maximum likelihood; N: Nucleoprotein; NIRVS: Non-Retroviral Integrated RNA Viruses Sequences; ORF: Open Reading Frame; piRNA: PIWI-interacting RNA; PTS: piRNA trigger sequence; RPKM: Reads per kilobase per million mapped reads; RVFV: Rift Valley fever virus; siRNA: small interfering RNA; TE: transposable elements; VEE: Venezuelan equine encephalitis; WNV: West Nile virus; ZIKV: Zika virus

### Acknowledgments

We thank Professor Anthony A. James (University of California at Irvine) for helpful comments on the manuscript. We thank Patrizia Chiari for insectary work.

### Funding

The project was supported by a European Research Council Consolidator Grant (ERC-CoG) under the European Union's Seventh Framework Programme (grant number ERC CoG 615,680) to R.P.v.R. and an ERC-CoG under the European Union's Horizon 2020 Programme (Grant Number ERC CoG 682,394) to M.B. The

fundes had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

#### Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

#### Authors' contributions

UP performed bioinformatic analyses and analysed the data; PM performed bioinformatic analyses, analysed the data and wrote the manuscript; RCL performed qRT-PCR analyses and analysed the data; LO performed bioinformatic analyses and analysed the data; ER performed bioinformatic analyses and analysed the data; ZT performed bioinformatic analyses, analysed the data and wrote the manuscript; RvR, conceived the study and wrote the manuscript; MB conceived the study, analysed the data and wrote the manuscript. All authors read and approved the final manuscript.

#### Ethics approval and consent to participate

Not applicable.

#### Consent for publication

Not applicable.

#### Competing interests

Not applicable.

#### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

#### Author details

<sup>1</sup>Department of Biology and Biotechnology, University of Pavia, via Ferrata 9, 27100 Pavia, Italy. <sup>2</sup>Department of Medical Microbiology, Radboud University Medical Center, Radboud Institute for Molecular Life Sciences, P.O. Box 9101, 6500 HB Nijmegen, Nijmegen, The Netherlands. <sup>3</sup>Independent Researcher, Mezzocorona, Trento, Italy. <sup>4</sup>Engenome srl, Via Ferrata 5, Pavia, Italy. <sup>5</sup>Department of Biochemistry and Fralin Life Sciences Institute, Virginia Tech, Blacksburg, VA 24061, USA.

Received: 14 March 2017 Accepted: 26 June 2017

Published online: 05 July 2017

#### References

- Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, et al. Global trends in emerging infectious diseases. *Nature*. 2008;451:990–3.
- Chen XG, Jiang X, Gu J, Xu M, Wu Y, et al. Genome sequence of the Asian Tiger mosquito, *Aedes albopictus*, reveals insights into its biology, genetics, and evolution. *Proc Natl Acad Sci U S A*. 2015;112:E5907–15.
- Weaver SC, Reisen WK. Present and future arboviral threats. *Antivir Res*. 2010;85:328–45.
- Powell JR, Tabachnick WJ. History of domestication and spread of *Aedes aegypti* – a review. *Mem Inst Oswaldo Cruz*. 2010;108:11–7.
- Paupy C, Delatte H, Bagny L, Corbel V, Fontenille D. *Aedes albopictus*, an arbovirus vector: from the darkness to the light. *Microbes Infect*. 2009;11:1177–85.
- Bonizzoni M, Gasperi G, Chen X, James AA. The invasive mosquito species *Aedes albopictus*: current knowledge and future perspectives. *Trends Parasitol*. 2013;29:460–8.
- Farajollahi A, Fonseca DM, Kramer LD, Marm KA. "bird biting" mosquitoes and human disease: a review of the role of *Culex pipiens* Complex mosquitoes in epidemiology. *Infect Genet Evol*. 2011;11:1577–85.
- Linthicum KJ, Britch SC, Anyamba A. Rift Valley fever: an emerging mosquito-borne disease. *Annu Rev Entomol*. 2016;61:395–415.
- Carissimo G, Eiglmeier K, Reveillaud J, Holm I, Diallo M, et al. Identification and characterization of two novel RNA viruses from *Anopheles gambiae* species complex mosquitoes. *PLoS one*. 2016;11(5):e0153881.
- Bolling BG, Weaver SC, Tesh RB, Vasilakis N. Insect-specific virus Discovery: significance for the arbovirus community. *Viruses*. 2015;7:4911–28.
- Fauver JR, Grubaugh ND, Krajacich BJ, Weger-Lucarelli J, Lakin SM. West African *Anopheles gambiae* mosquitoes harbor a taxonomically diverse virome including new insect-specific flaviviruses, mononegaviruses, and totiviruses. *Virology*. 2016;498:288–99.
- Tabachnick WJ. Nature, nurture and evolution of intra-species variation in mosquito arbovirus transmission competence. *Int J Environ Res Public Health*. 2013;10:249–77.
- Stapleford KA, Coffey LL, Lay S, Borderia AV, Duong V, et al. Emergence and transmission of arbovirus evolutionary intermediates with epidemic potential. *Cell Host Microbe*. 2014;15:706–16.
- Dennison NJ, Jupatanakul N, Dimopoulos G. The mosquito microbiota influences vector competence for human pathogens. *Curr Opin Insect Sci*. 2014;3:6–13.
- Cheng G, Liu Y, Wang P, Xiao X. Mosquito defense strategies against viral infection. *Trends Parasitol*. 2016;32:177–86.
- Chouin-Carneiro T, Vega-Rua A, Vazeille M, Yebakima A, Girod R, et al. Differential susceptibilities of *Aedes aegypti* and *Aedes albopictus* from the Americas to Zika virus. *PLoS Negl Trop Dis*. 2016;10(3):e0004543.
- Weger-Lucarelli J, Ru ckert C, Chotiwan N, Nguyen C, Garcia Luna SM, et al. Vector Competence of American Mosquitoes for Three Strains of Zika Virus. *PLoS Negl Trop Dis*. 2016;10(10):e0005101.
- Bronkhorst AW, van Rij RP. The long and short of antiviral defense: small RNA-based immunity in insects. *Curr Opin Virol*. 2014;7:19–28.
- Olson KE, Blair C. Arbovirus-mosquito interactions: RNAi pathway. *Curr Opin Virol*. 2015;15:119–26.
- Goic B, Stapleford KA, Frangeul L, Doucet AJ, Gausson V, et al. Virus-derived DNA drives mosquito vector tolerance to arboviral infection. *Nat Commun*. 2016;7:12410.
- Miesen P, Joosten J, van Rij RP. PIWIs go viral: arbovirus-derived piRNAs in vector mosquitoes. *PLoS Pathog*. 2016;12(12):e1006017.
- Handler D, Meixner K, Pizka M, Lauss K, Schmied C, Gruber FS, Brennecke J. The genetic makeup of the *Drosophila* piRNA pathway. *Mol Cell*. 2013;50(5):762–77.
- Arensburger P, Hice RH, Wright JA, Craig NL, Atkinson PW. The mosquito *Aedes aegypti* has a large genome size and high transposable element load but contains a low proportion of transposon-specific piRNAs. *BMC Genomics*. 2011;12:606.
- Crochu S, Cook S, Attoui H, Charrel RN, De Chesse R, et al. Sequences of flavivirus-related RNA viruses persist in DNA form integrated in the genome of *Aedes* spp. mosquitoes. *J Gen Virol*. 2004;85:1971–80.
- Belyi VA, Levine AJ, Skalka AM. Unexpected inheritance: multiple integrations of ancient bornavirus and ebolavirus/marburgvirus sequences in vertebrate genomes. *PLoS Pathog*. 2010;6:e1001030.
- Katzourakis A, Gifford RJ. Endogenous viral elements in animal genomes. *PLoS Genet*. 2010;6:e1001191.
- Horie M, Honda T, Suzuki Y, Kobayashi Y, Daito T, Oshida T, Ikuta K, Gojobori T, Coffin JM, Tomonaga K. Endogenous non-retroviral RNA virus elements in mammalian genomes. *Nature*. 2010;463:84–7.
- Fort P, Albertini A, Van-Hua A, Berthomieu A, Roche S, et al. Fossil rhabdoviral sequences integrated into arthropod genomes: ontogeny, evolution, and potential functionality. *Mol Biol Evol*. 2012;29:381–90.
- Ballinger MJ, Bruenn JA, Taylor DJ. Phylogeny integration and expression of sigma virus-like genes in *Drosophila*. *Mol Phylogenet Evol*. 2012;65:251–8.
- Ballinger MJ, Bruenn JA, Hay J, Czechowski D, Taylor DJ. Discovery and evolution of bunyavirids in arctic phantom midges and ancient bunyavirid-like sequences in insect genomes. *J Virol*. 2014;88:8783–94.
- Geisler C, Jarvis DL. Rhabdovirus-like endogenous viral elements in the genome of *Spodoptera frugiperda* insect cells are actively transcribed: implications for adventitious virus detection. *Biologicals*. 2016;44:219–25.
- Lequime S, Lambrechts L. Discovery of flavivirus-derived endogenous viral elements in *Anopheles* mosquito genomes supports the existence of *Anopheles*-associated insect-specific flaviviruses. *Virus Evol* 2017; 3:vev035.
- Aiweksakun P, Katzourakis A. Endogenous viruses: connecting recent and ancient evolution. *Virology*. 2015;479:480:26–37.
- Tromas N, Zwart MP, Forment J, Elena SF. Shrinkage of genome size in a plant RNA virus upon transfer of an essential viral Gene into the host genome. *Genome Biol Evol*. 2014;6:538.
- Aswad A, Katzourakis A. Palaeovirology and virally derived immunity. *Trends Ecol Evol*. 2012;27:627–36.
- Goic B, Vodovar N, Mondotte JA, Monot C, Frangeul L, Blanc H, Gausson V, Vera-Otarola J, Cristofari G, Saleh MC. RNA-mediated interference and reverse transcription control the persistence of RNA viruses in the insect model *Drosophila*. *Nat Immunol*. 2013;14(4):396–403.

37. Nag DK, Brecher M, Kramer LD. DNA forms of arboviral RNA genomes are generated following infection in mosquito cell cultures. *Virology*. 2016;498:164–71.
38. Koonin EV, Dolja W, Krupovic M. Origin and evolution of viruses of eukaryotes: the ultimate modularity. *Virology*. 2015;479–480:2–25.
39. Attoui H, Jaafar MF, Belhouchet M, Tao S, Chen B, et al. Liao ning virus, a new Chinese seadornavirus that replicates in transformed and embryonic mammalian cells. *J Gen Virol*. 2006;87:199–208.
40. Song LW, Kang HJ, Gu SH, Moon SS, Bennet SN, et al. Characterization of Imjin Virus, a Newly Isolated Hantavirus from the Ussuri White-Toothed Shrew (*Crocidura lasiura*). *J Virol*. 2009;83:6184–91.
41. Li CX, Shi M, Tian JH, Lin XD, Kang YJ, et al. Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. *elife*. 2015;4:e05378.
42. Vasilakis N, Tesh RB. Insect-specific viruses and their potential impact on arbovirus transmission. *Curr Opin Virol*. 2015;15:69–74.
43. Girardi E, Miesen P, Pennings B, Frangeul L, Saleh MC, van Rij RP. Histone-derived piRNA biogenesis depends on the ping-pong partners Piwi5 and Ago3 in *Aedes aegypti*. *Nucleic Acids Res*. 2017;5;45(8):4881–892.
44. Liu P, Dong Y, Gu J, Puthiyakunnon S, Wu Y, Chen XG. Developmental piRNA profiles of the invasive vector mosquito *Aedes albopictus*. *Parasit Vectors*. 2016;9:524.
45. Miesen P, Girardi E, van Rij RP. Distinct sets of PIWI proteins produce arbovirus and transposon-derived piRNAs in *Aedes aegypti* Mosquito cells. *Nucleic Acids Res*. 2015;43:6545–56.
46. Maori E, Tanne E, Sela I. Reciprocal sequence exchange between non-retrovirus and hosts leading to the appearance of new host phenotypes. *Virology*. 2007;362:342–9.
47. Hermanns K, Zirkel F, Kopp A, Marklewitz M, Rwego IB, et al. Discovery of a novel alphavirus related to Eilat virus. *J Gen Virol*. 2017;98(1):43–9.
48. Marklewitz M, Zirkel F, Kurth A, Drosten C, Junglen S. Evolutionary and phenotypic analyses of live virus isolates suggests arthropod origin of a pathogenic RNA virus family. *Proc Natl Acad Sci U S A*. 2015;112:7536–41.
49. Grunill M, Boots M. How important is vertical transmission of dengue viruses by mosquitoes (Diptera: Culicidae)? *J Med Entomol*. 2016;53:1–19.
50. Parrish NF, Fujino K, Shiromoto Y, Iwasaki YW, Ha H, Xing J, Makino A, Kuramochi-Miyagawa S, Nakano T, Siomi H, Honda T. Tomonaga K piRNAs derived from ancient viral processed pseudogenes as transgenerational sequence-specific immune memory in mammals. *RNA*. 2015;21:1–13.
51. Homolka D, Pandey RR, Goriaux C, Brassat E, Vaury C, Sachidanandam R, Fauvarque MO, Pillai RS. PIWI slicing and RNA elements in precursors instruct directional primary piRNA biogenesis. *Cell Rep*. 2015;12:418–28.
52. Assis R, Kondrashov AS. Rapid repetitive element-mediated expansion of piRNA clusters in mammalian evolution. *Proc Natl Acad Sci U S A*. 2009;106:7079–82.
53. Fujino K, Horie M, Honda T, Merriman DK, Tomonaga K. Inhibition of Borna disease virus replication by an endogenous bornavirus-like element in the ground squirrel genome. *Proc Natl Acad Sci U S A*. 2014;111:13175–80.
54. Bhattacharya S, Basu P. The southern house mosquito, *Culex quinquefasciatus*: profile of a smart vector. *J Entomol Zool Studies*. 2016;4:73–81.
55. Bartholomay L, Waterhouse RM, Mayhew GF, Campbell CL, Michel K, et al. Pathogenomics of *Culex quinquefasciatus* and meta-analysis of infection responses to diverse pathogens. *Science*. 2010;330:88–90.
56. Lewis SH, Salmela H, Obbard DJ. Duplication and diversification of dipteran argoate genes and the evolutionary divergence of piwi and augergine. *Genome Biol Evol*. 2016;8:507–18.
57. Honda T, Tomonaga K. Endogenous non-retroviral RNA virus elements evidence a novel type of antiviral immunity. *Mobile Genetic Elements*. 2016; 6:e1165785.
58. Blair CD, Olson KE. Targeting Dengue virus replication in Mosquitoes. In: Aksoy S, Wikel S, Dimopoulos G, editors. *Arthropod Vector: The Controller of Transmission* Vol. 1. Elsevier. Pages 201 - 217.
59. Neafsey DE, Waterhouse RM, Abai MR, Aganezov SS, Allen JE. Mosquito genomics. Highly evolvable malaria vectors: the genomes of 16 Anopheles mosquitoes. *Science* 2015; 347(6217):1258522.
60. Falda M, Toppo S, Pescarolo A, Lavezzo E, Di Camillo B, Facchinetti A, Cilia E, Velasco R, Fontana P. Argot2: a large scale function prediction tool relying on semantic similarity of weighted Gene ontology terms. *BMC Bioinformatics*. 2012;13:S14.
61. Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*. 2010;26:589–95.
62. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26:841–2.
63. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. *Nat Biotechnol*. 2011;29:24–6.
64. Tamura K, Stecher G, Peterson D, Filipksi A, Kumar S. MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol*. 2013;30:2725–9.
65. Hu W, Criscione F, Liang S, Tu Z. MicroRNAs of two medically important mosquito species: *Aedes aegypti* and *Anopheles stephensi*. *Insect Mol Biol*. 2015;24(2):240–52.
66. Gu J, Hu W, Wu J, Zheng P, Chen M, James AA, et al. miRNA Genes of an Invasive Vector Mosquito, *Aedes albopictus*. *PLoS One*. 2013;8:e67638.
67. Blankenberg D, Gordon A, von Kuster G, Coraor N, Taylor J, Nekrutenko A. Manipulation of FASTQ data with Galaxy. *Bioinformatics*. 2010;26:1783–5.
68. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 2009;10(3):R25.
69. Crooks GE, Hon G, Chandonia JM, Brenner S. WebLogo: A sequence logo generator. *Genome Res*. 2004;14:1188–90.
70. Reynolds JA, Poelchau MF, Rahman Z, Armbruster PA, Denlinger DL. Transcript profiling reveals mechanisms for lipid conservation during diapause in the mosquito, *Aedes albopictus*. *J Insect Physiol*. 2012;58:966–73.
71. Nene V, Wortman JR, Lawson D, Haas B, Kodira C, Tu ZJ, Loftus B, Xi Z, Megy K, Grabherr M, Ren Q, Zdobnov EM, Lobo NF, Campbell KS, Brown SE, Bonaldo MF, Zhu J, Sinkins SP, Hogenkamp DG, Amedeo P, Arensburger P, Atkinson PW, Bidwell S, Biedler J, Birney E, Bruggner RV, Costas J, Coy MR, Crabtree J, Crawford M, DeBruyn B, Decaprio D, Eiglmeier K, Eisenstadt E, El-Dorry H, Gelbart WM, Gomes SL, Hammond M, Hannick LI, Hogan JR, Holmes MH, Jaffe D, Johnston JS, Kennedy RC, Koo H, Kravitz S, Kriventseva EV, Kulp D, Labutti K, Lee E, Li S, Lovin DD, Mao C, Mauceli E, Menck CF, Miller JR, Montgomery P, Mori A, Nascimento AL, Naveira HF, Nusbaum C, O'leary S, Orvis J, Perteu M, Quesneville H, Reidenbach KR, Rogers YH, Roth CW, Schneider JR, Schatz M, Shumway M, Stanke M, Stinson EO, Tubio JM, Vanzee JP, Verjovski-Almeida S, Werner D, White O, Wyder S, Zeng Q, Zhao Q, Zhao Y, Hill CA, Raikhel AS, Soares MB, Knudson DL, Lee NH, Galagan J, Salzberg SL, Paulsen IT, Dimopoulos G, Collins FH, Birren B, Fraser-Liggett CM, Severson DW. Genome sequence of *Aedes aegypti*, a major arbovirus vector. *Science*. 2007 Jun 22;316(5832):1718–23. Epub 2007 May 17. PubMed PMID: 17510324; PubMed Central PMCID: PMC2868357.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at  
www.biomedcentral.com/submit

