Virus Evolution, 2019, 5(1): vez010

doi: 10.1093/ve/vez010
Research article

Novel insights into endogenous RNA viral elements in *Ixodes scapularis* and other arbovirus vector genomes

Alice G. Russo, Andrew G. Kelly, Daniel Enosi Tuipulotu†, Mark M. Tanaka, and Peter A. White*

School of Biotechnology and Biomolecular Sciences, Faculty of Science, University of New South Wales, Sydney, New South Wales, Australia

*Corresponding author: E-mail: p.white@unsw.edu.au
†http://orcid.org/0000-0002-6442-4633

Abstract

Many emerging arboviruses are not transmitted by traditional mosquito vectors, but by lesser-studied arthropods such as ticks, midges, and sand flies. Small RNA (sRNA) silencing pathways are the main antiviral defence mechanism for arthropods, which lack adaptive immunity. Non-retroviral integrated RNA virus sequences (NIRVS) are one potential source of sRNAs which comprise these pathways. NIRVS are remnants of past germline RNA viral infections, where viral cDNA integrates into the host genome and is vertically transmitted. In *Aedes* mosquitoes, NIRVS are widespread and produce PIWI-interacting RNAs (piRNAs). These are hypothesised to target incoming viral transcripts to modulate viral titre, perhaps rendering the organism a more efficient arbovirus vector. To explore the NIRVS landscape in alternative arbovirus vectors, we validated the NIRVS landscape in *Aedes* spp. and then identified novel NIRVS in six medically relevant arthropods and also in *Drosophila melanogaster*. We identified novel NIRVS in *Phlebotomus papatasi*, *Culicoides sonorensis*, *Rhipicephalus microplus, Anopheles gambiae*, *Culex quinquefasciatus*, and *Ixodes scapularis*. Due to their unexpected abundance, we further characterised NIRVS in the black-legged tick *I. scapularis* ($n = 143$). Interestingly, NIRVS are not enriched in *R. microplus*, another hard tick, suggesting this is an *Ixodes*-specific adaptation. *I. scapularis* NIRVS are enriched in bunya- and orthomyxo-like sequences, reflecting that ticks are a dominant host for these virus groups. Unlike in mosquitoes, *I. scapularis* NIRVS are more commonly derived from the non-structural region (replicase) of negative-sense viruses, as opposed to structural regions (e.g. glycoprotein). Like other arthropods, *I. scapularis* NIRVS preferentially integrate into genomic piRNA clusters, and serve as a template for primary piRNA production in the commonly used embryonic *I. scapularis* ISE6 cell line. Interestingly, we identified a two-fold enrichment of non-long terminal repeat (non-LTR) retrotransposons, in genomic proximity to NIRVS, contrasting with studeis in *Ae. aegypti*, where LTR retrotransposons are instead associated with NIRVS formation. We characterised NIRVS phylogeny and integration patterns in the important vector, *I. scapularis*, revealing they are distinct from those in *Aedes* spp. Future studies will explore the possible antiviral mechanism conferred by NIRVS to *I. scapularis*,which may help the transmission of pathogenic arboviruses. Finally, this study explored NIRVS as an untapped wealth of viral diversity in arthropods.

Key words: endogenous viral element; arthropod; paleovirology; viral diversity; *Ixodes scapularis*.

1. Introduction

Over 100 million people contract arthropod-borne viruses (arboviruses) each year (WHO 2014). Many arboviruses are classified as emerging diseases, meaning their incidence is rapidly increasing; additionally, a large proportion lack preventative measures or treatment (Jones et al. 2008; LaBeaud, Bashir, and

King 2011). Consequently, managing these diseases is essential to limit the morbidity and mortality they cause in human and animal populations.

Two classes of arthropods transmit arboviruses to mammals: insects and arachnids (Mellor 2000). All insect vectors are classified into the order Diptera (true flies). Of these, mosquitoes (Culicidae) within the genera *Aedes* and *Culex* are the most significant arbovirus vectors, transmitting, e.g. flaviviruses (e.g. Dengue virus (DENV), yellow fever virus (YFV) and Zika virus (ZIKV)) and togaviruses (e.g. Chikungunya virus (CHIKV) and Ross River virus (RRV)) (Table 1). Collectively, these viruses cause over 400 million human infections per year and over 20,000 deaths (LaBeaud, Bashir, and King 2011; Bhatt et al. 2013; WHO 2014). Sand flies (Psychodidae) and midges (Ceratopogonidae) also transmit pathogenic viruses of humans and livestock, mostly in Southern Europe (Table 1) (Mellor, Boorman, and Baylis 2000; Depaquit et al. 2010). Of the arachnids, ticks (order Ixodida) are the only group known to transmit arboviruses and are second only to mosquitoes in the diversity of viruses they can transmit (Table 1) (Mansfield et al. 2017). Since 2000, several emerging tick-borne viruses have caused human outbreaks in Europe, China, North America, India, and the Middle East (Mansfield et al. 2017). Compared with mosquitoes, tick-borne viral infections are under-sampled and understudied. In addition, no specific treatments are available, and vaccines exist only for three tick-borne viruses (Louping-ill virus, tick-borne encephalitis virus, and Kyasanur Forest disease virus) but are not widely distributed (Mansfield et al. 2017).

In contrast to vertebrates, arthropods tolerate arboviral infections well; they do not usually develop disease once infected and maintain the infection for life. This suggests an innate ability to control viral titre, while still permitting arboviral transmission to the next host (Mellor 2000; Blair 2011). However, not all haematophagous arthropods are efficient vectors for the same virus. The ability of a particular virus to replicate within a host is referred to as vector competence (Hardy et al. 1983). Understanding the factors that influence vector competence is of practical interest, as modifying these can foster the development of targeted arboviral control strategies (Weiss and Aksoy 2011).

Vector competence is driven by the complex interaction between the virus and the host immune system. Arthropods lack an adaptive immune response and rely entirely on innate immunity to control viral infections (Keene et al. 2004; Sanchez-Vargas et al. 2009). Small RNA (sRNA) silencing pathways (sRNA SPs), such as RNA interference (RNAi), are a major component of arthropod immunity (Bronkhorst and van Rij 2014). RNAi is a gene-silencing mechanism modulated by sRNAs, the best-described family of which are small interfering RNAs (siRNAs, 20–25 nt in length). The siRNA pathway is triggered by the accumulation of dsRNA which is cleaved to generate siRNAs. The interaction of siRNAs with proteins of the Argonaute family triggers the silencing of target RNAs (e.g. viral transcripts) mediated by the RNA-induced silencing complex. In mosquitoes, suppression of RNAi enhances the accumulation of viral RNA in *Anopheles gambiae* (O'nyong nyong virus) and *Ae. aegypti* (Sindbis virus) and in some cases causes increased mortality (Keene et al. 2004; Campbell et al. 2008; Myles et al. 2008). In ticks, knockdown of key proteins involved in RNAi results in a significant increase of tick-borne encephalitis virus RNA in *Ixodes* cells (Weisheit et al. 2015), suggesting a similar defence mechanism.

PIWI-interacting RNAs (piRNAs) (25–30 nt) were discovered in *Drosophila*, where they defend against the deleterious effects of transposable elements (TEs) in the germ-line (Brennecke et al. 2007). piRNAs are distinct from other sRNA classes as they interact with a distinct set of proteins (Piwi proteins), and they arise from a ssRNA, not a dsRNA, precursor (Siomi et al. 2011). There are two piRNA biogenesis pathways. Primary piRNAs are generated from host transcripts (usually arising from transposon-rich genomic piRNA clusters) that are cleaved to generate 25–30 nt piRNAs. These are amplified via the ping-pong cycle, leading to the cleavage of target RNA (e.g. active transposons), producing secondary piRNAs. While piRNA sequences are not conserved between organisms, primary piRNAs unequivocally have a uridine at their 5′ end (1 U bias), and secondary piRNAs have an adenine bias at the tenth nt from their 5′ end (10 A bias) (Aravin, Hannon, and Brennecke 2007). Unlike in *Drosophila*, mosquitoes can use piRNAs to mount an antiviral defence in cooperation with other sRNA SPs in both somatic and germ cells (Morazzani et al. 2012). Mosquito piRNAs can be directly produced from exogenous viral RNA, indicating that piRNAs may be derived from both endogenous and exogenous sources, and may play an antiviral role (Vodovar et al. 2012; Miesen, Girardi, and van Rij 2015). Although recent studies indicate widespread conservation of the piRNA pathway among arthropods (Lewis et al. 2018), less is known about whether they could play an antiviral role in non-mosquito arthropods.

One source of endogenous piRNAs is non-retroviral integrated RNA virus sequences (NIRVS) (Vijayendran et al. 2013). NIRVS arise from the reverse transcription of viral RNA into cDNA and its integration into the genome of a host germ cell, followed by vertical transmission to offspring (Katzourakis and Gifford 2010). RNA viruses do not integrate into the host genome as part of their lifecycle, so it was initially thought that integration of non-retrotranscribing viruses was serendipitous—especially in vertebrate genomes which contain relatively few insertions (Katzourakis and Gifford 2010). Yet recent evidence has highlighted the abundance of NIRVS in some arthropod genomes—e.g., *Ae. aegypti* and *Ae. albopictus* contain >100 NIRVS from over eight RNA virus families encompassing +ssRNA, −ssRNA, and dsRNA viral groups (Palatini et al. 2017; Whitfield et al. 2017). Additionally, large-scale sequence analysis has shown that many of these NIRVS cluster within piRNA producing genomic regions and produce primary piRNAs (Whitfield et al. 2017; Ter Horst et al. 2018). This suggests that NIRVS have a functional role, perhaps involving the piRNA pathway.

NIRVS are also a rich source of information about long-term viral evolution, diversity, and host range. In its exogenous state, an RNA virus will typically evolve at 10^{-3} substitutions/site/year (Jenkins et al. 2002), but when integrated into a eukaryotic genome, will evolve several orders of magnitude more slowly (Aiewsakun and Katzourakis 2015). NIRVS represent near-intact ancient viral sequences, and so can be useful to estimate the timescale of long-term viral evolution, e.g., dating the emergence of modern viral families (Belyi, Levine, and Skalka 2010). NIRVS are also a useful tool to expand the known host range of viral families, as finding a particular NIRVS family in a genome indicates an ancestral infection event of that host lineage (Aiewsakun and Katzourakis 2015). The true spectrum of arthropod-infecting RNA viruses is only recently coming to light (Shi et al. 2016) and further information on host range and viral diversity can be assisted by the identification and analysis of NIRVS in arthropod genomes.

Most studies thus far have focussed on the abundance and diversity of NIRVS in *Aedes* genomes, due to the high quality of available genomes and the clinical significance of these vectors

Table 1. Summary of arthropods included in this study and their role as vectors of disease in humans and other vertebrates.

Arthropod genus (order; family)	Pathogenic vertebrate viruses transmitted by this host			Representative organism (s) (common name)	Genome assembly/ strain	Assembly level	References
	Viral family (genome type)	Viral species	Known host range				
Culex **(Diptera; Culicidae)**	*Flaviviridae* (+ssRNA)	WNV, ZIKV, St. Louis *Encephalitis virus*	Humans	*Cx. quinquefasciatus* (Southern house mosquito)	CpipJ2/Johannesburg	Scaffold	(Hammon et al. 1941; Sudia et al. 1967; Monath and Tsai 1987; Turell et al. 2005; Guo et al. 2016)
	Togaviridae (+ssRNA)	*Western equine encephalitis virus*, RRV	Horses, humans				
Culicoides **(Diptera; Ceratopogonidae)**	*Reoviridae (dsRNA)*	*Bluetongue virus*, *African horse sickness virus*	Sheep, deer, cows	*C. sonorensis* (biting midge)	Cson_2.0/PIR-s-3	Scaffold	(Foster, Jones, and McCrory 1963; Mellor, Boorman, and Baylis 2000; Ruder et al. 2015; Morales-Hojas et al. 2018)
	Peribunyaviridae (-ssRNA)	*Schmallenberg virus*, *Akabane virus*, *Oropouche virus*	Cows, sheep, goats				
	Rhabdoviridae (-ssRNA)	*Bovine ephemeral fever virus*	Cows				
Phlebotomus **(Diptera; Psychodidae)**	*Phenuiviridae (-ssRNA)*	*Sandfly fever Naples virus*, *Sandfly fever Sicilian virus*	Humans	*P. papatasi* (sandfly)	Ppapl1/Israel	Scaffold	(Depaquit et al. 2010)
	Rhabdoviridae (-ssRNA)	*Chandipura virus*	Humans				
	Reoviridae (dsRNA)	*Changuinola virus*	Humans				
Ixodes **(Ixodida: Ixodidae)**	*Flaviviridae* (+ssRNA)	*Powassan virus* *Deer tick virus*	Humans	*I. scapularis* (Deer tick)	IscaW1/Wikel	Scaffold	(Anderson and Armstrong 2012; Gulia-Nuss et al. 2016)
Rhipicephalus **(Ixodida, Ixodidae)**	*Orthomyxoviridae* (−ssRNA)	*Thogoto virus*	Humans	*R. microplus* (Asian cattle tick)	ADMZ02/Deutsch	Scaffold	(Zhang et al. 2012)
	Nairoviridae (−ssRNA)	*Crimean-Congo Haemorrhagic Fever virus*	Humans				
	Phenuiviridae (−ssRNA)	*Severe fever with thrombocytopaenia virus*	Humans				
Aedes **(Diptera, Culicidae)**	*Flaviviridae* (+ssRNA)	WNV, ZIKV, DENV, YFV	Humans	*Ae. aegypti* (yellow fever mosquito)/ *Ae. albopictus* (Asian tiger mosquito)	AaegL5.0/ Canu_80X_ arrow2.2	Chromosome/ Scaffold	(Nene et al. 2007; Chen et al. 2015; Matthews et al. 2018)
	Togaviridae (+ssRNA)	CHIKV, RRV	Humans				
Anopeles **(Diptera, Culicidae)**	*Togaviridae* (+ssRNA)	*O'nyong'nyong virus*	Humans	*An. gambiae* (malaria mosquito)	AgamP4/PEST	Scaffold	(Holt et al. 2002; Keene et al. 2004)
Drosophila **(Diptera; Drosophilidae)**	None	None	None	*D. melanogaster* (fruit fly)	Release 6/dm6	Chromosome	(Hoskins et al. 2015)

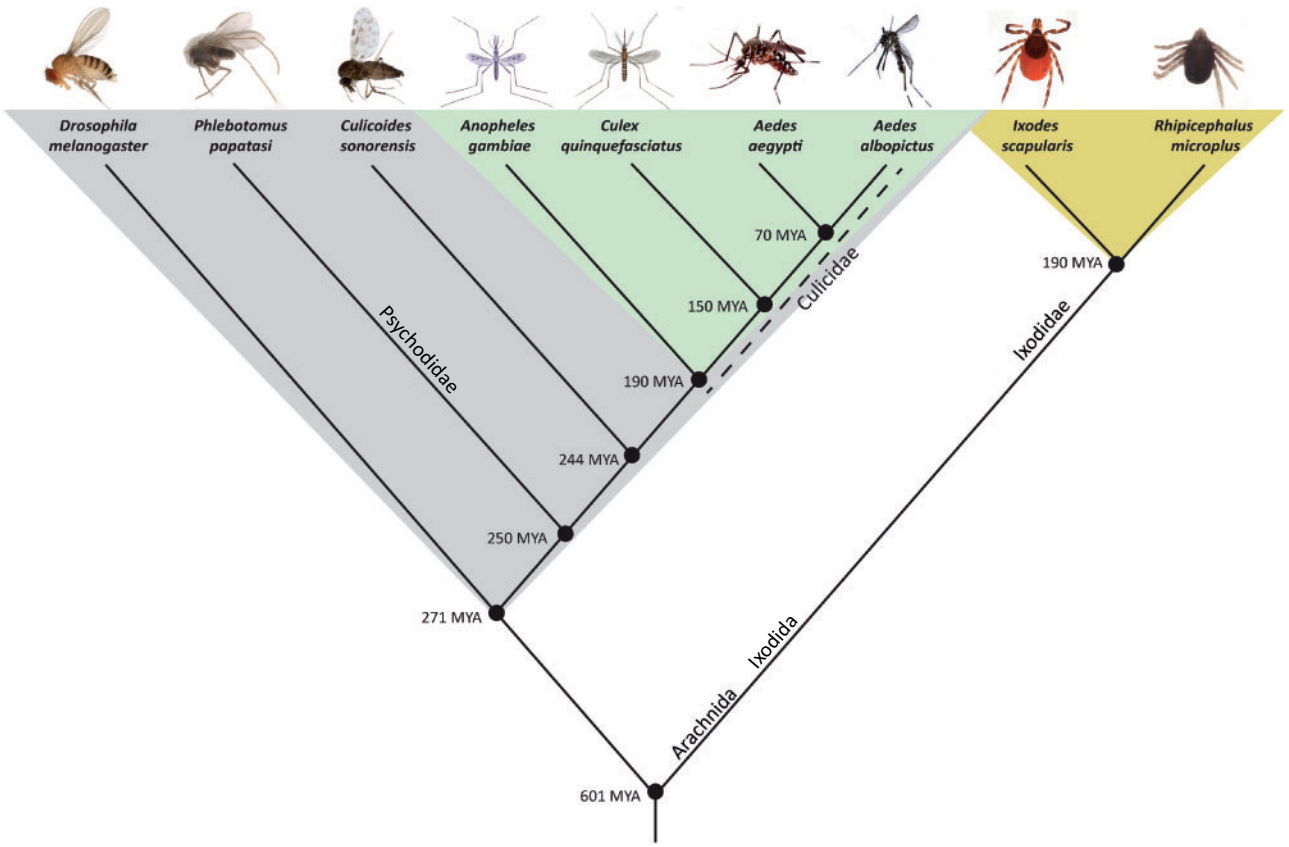

Figure 1. Phylogenetic relationship, divergence times, and taxonomic classifications of arbovirus vectors screened for the presence of NIRVS in this study. Divergence times, in millions of years ago (MYA), between arthropod lineages are indicated according to TimeTree (Kumar et al. 2017). Taxonomy is labelled at the class (Insecta, Arachnida), order (Diptera, Ixodida), family (Drosophilidae, Culicidae, Ixodidae, Psychodidae), and species level (indicated under the image of the arthropod).

(Crochu et al. 2004; Roiz et al. 2009; Palatini et al. 2017; Whitfield et al. 2017). Other studies have also indicated the presence of NIRVS in non-dipteran lineages (Shi et al. 2016; Ter Horst et al. 2018). As the incidence of disease caused by arboviruses rises it will become important to generate a deeper understanding of NIRVS diversity among non-mosquito vectors.

We aimed to examine the spectrum of NIRVS in arbovirus vectors other than *Aedes* mosquitoes, which represent a threat in the context of emerging viruses. Using the well-studied genomes of *Aedes* mosquitoes to validate our analysis, we employed a bioinformatic pipeline to characterise NIRVS in seven non-*Aedes* arbovirus vectors that have representative genomic sequences (Table 1 and Fig. 1).

2. Methods

2.1 Query viral sequence and arthropod genome retrieval for NIRVS identification

To identify NIRVS in arthropod genomes, a list of query viral protein sequences was assembled. Type species for each viral family as defined by the International Committee for Virus Taxonomy (ICTV) Master Species List 2017 were used (ICTV 2017). These included all proteins from reference viral genomes of dsRNA viruses, +ssRNA viruses and −ssRNA viruses. Protein sequences from viruses recently identified in arthropods through deep RNA sequencing studies were also included to increase the likelihood of finding sequences from divergent viruses (Cook et al. 2013; Chandler, Liu, and Bennett 2015; Li et al. 2015; Shi et al. 2016). This dataset comprised 3,694 protein sequences representing 1,933 distinct viruses (Supplementary Table S1).

The arthropod genomes used for analysis are detailed in Table 1. The genome of *Drosophila melanogaster* (Release 6) (Adams et al. 2000; Hoskins et al. 2015) was also searched for NIRVS. This is the best-assembled and annotated arthropod genome and contains no known NIRVS. Therefore, this allowed us to assess whether our pipeline was likely to generate false positives.

2.2 NIRVS identification in arthropod genomes

Identification of NIRVS was based on methods outlined previously (Katzourakis and Gifford 2010) which used Basic Local Alignment Search Tool (BLAST). A tBLASTn search was undertaken separately for each organism in Geneious (v10.2.3) (Kearse et al. 2012), with all viral protein sequences queried against each arthropod genome. BLAST hits were filtered by E-value ($\leq 1e^{-3}$), and duplicate BLAST hits (covering the same region of a host genome) were removed manually. For BLAST hits representing the same NIRVS, the hit containing the maximum genomic coverage was retained. To verify that the BLAST hits were truly viral-derived sequences, a reciprocal search was performed by using translated BLAST hits as query sequences in a tBLASTn search against the NCBI non-redundant (*nr*) database. Matches to retroviruses, viral cloning vectors, and non-specific matches to host loci or other arthropod loci were discarded.

2.3 Mapping of NIRVS to viral genomes

To visualise which region of the viral genome NIRVS were derived from, NIRVS were aligned individually to the genome of a closely related extant virus using ClustalW (v2.0), which does not penalise end-gaps in the nucleotide alignment (Larkin et al. 2007). Alignments were concatenated, and duplicates of the reference sequence were removed. Alignments were visualised in GraphPad Prism (v7.0) based on the numerical mapping position of the NIRVS to the reference sequence. For alignments and trees containing mononegaviruses and mononega-like viruses, the *Chuviridae* was included. The *Chuviridae* is not technically part of this order but shares a close phylogenetic relationship with mononegaviruses (Li et al. 2015).

2.4 Detection of NIRVS duplicates within arthropod genomes

To calculate which proportion of NIRVS had significant identity to other NIRVS within the same genome, indicating that they were probably duplicated sequences, a ClustalW (v2.0) alignment of all NIRVS within each genome was performed in Geneious (v10.2.3). The distance matrix from this alignment was used to identify NIRVS with >98 per cent nt identity to any other NIRVS.

2.5 Phylogenetic analysis of specific NIRVS in *I. scapularis*

To create phylogenetic trees broadly representing each group of NIRVS in this study, a collection of representative nucleotide sequences corresponding to the respective RNA-dependent RNA polymerase (RdRp) region of each group was downloaded from NCBI (Supplementary Table S3). Translated nt sequences, which were of similar sequence lengths, were aligned using MAFFT (v7.017) (Katoh et al. 2002) and trimmed using trimAl (v1.4.1) (Capella-Gutierrez, Silla-Martinez, and Gabaldon 2009). Maximum-likelihood (ML) phylogenetic trees were created in RAxML (v8.2.8) (Stamatakis 2014).

To perform phylogenetic analysis of specific NIRVS from *I. scapularis*, twenty-five closely related viral sequences were identified by a tBLASTn search against the *nr* database. Nt sequences of the same gene from which the NIRVS was derived (e.g. replicase) were downloaded for these viruses, as well as a more distantly-related outgroup sequence. Sequences were aligned and trimmed as above. ML phylogenetic trees were created in RAxML (v8.2.8) (Stamatakis 2014), with an automated aa substitution model and rapid bootstrapping with 500 non-parametric replicates. Eukaryotic hosts of each viral species included in the phylogenetic analyses were deduced from study metadata and manually indicated on the tree.

2.6 piRNA cluster prediction in the *I. scapularis* genome

Analysis of a publicly available sRNA dataset from *I. scapularis* was performed on The University of Queensland Galaxy server (https://usegalaxy.org.au) (Afgan et al. 2018). Raw sequencing reads were downloaded from NCBI [BioProject Accession PRJNA315659], read quality was assessed with FastQC (Andrews 2010), and Illumina sRNA adapters were trimmed with Trim Galore! (Galaxy v0.4.3.1). This dataset, along with the *I. scapularis* IscaW1 genome assembly (Gulia-Nuss et al. 2016), was used to predict piRNA-producing loci with proTRAC (v2.4.2) (Rosenkranz and Zischler 2012) applying custom parameters as previously defined (Ter Horst et al. 2018). NIRVS residing completely within predicted piRNA clusters were identified using a Megablast search implemented in Geneious (v10.2.3), with BLAST matches exhibiting both 100 per cent nt identity and query cover considered to reside within the cluster. The probability of integration into a piRNA cluster was defined as the percentage of the genome occupied by piRNA clusters. A cumulative binomial distribution was used to estimate whether NIRVS had preferentially integrated into piRNA clusters, as opposed to randomly within the genome.

2.7 TE analysis in *I. scapularis* NIRVS integration sites

To judge TE enrichment in regions around NIRVS, genomic sequence lying 5 kb either side of each NIRVS was extracted. For NIRVS that did not contain more than 5 kb flanking sequence due to the size of the scaffold, the sequence up until the end of the genomic scaffold was retained. RepeatMasker (v4.0.7) (Smit, Hubley, and Green 2018) was used to estimate the TE composition of these regions, with a list of predicted TEs identified in the IscaW1 genome assembly as a reference. These proportions were then compared with the genome-wide TE composition (Gulia-Nuss et al. 2016). Where two annotations encompassed more than 100 nt of overlapping genomic region, the annotation with the highest Smith–Waterman score was retained. We also analysed an NIRVS 'hotspot', where multiple NIRVS from different families were present in nearby loci. This was scaffold DS826508, nt position 110,866–186,277, which contained NIRVS from the *Orthomyxoviridae* (n = 2), Mononegavirales (n = 2), and Bunyavirales (n = 1) in close proximity (Supplementary Table S2).

2.8 Analysis of sRNA datasets from *I. scapularis*

The above sRNA dataset [PRJNA315659] was used to further characterise NIRVS-derived sRNAs. For evaluation of the NIRVS sequence content of total sRNA, the forward reads from all four sequencing runs were merged and HISAT2 (Galaxy v2.1.0) (Kim, Langmead, and Salzberg 2015) was used to align the reads to a reference FASTA file containing all *I. scapularis* NIRVS. Aligned reads were extracted and quantified with Salmon (v0.8.2) with custom parameters: (–kmerLen 19; –unmatedReads; incompatPrior 0.0) (Patro et al. 2017). Read counts were then grouped by the NIRVS' viral family of origin, and the proportion of total reads corresponding to each family was calculated.

To assess NIRVS-derived sRNA abundance by sequence length, the aligned reads (forward and reverse strand) for each k-mer (18–30 nt) were quantified with Salmon (v0.8.2) and then divided by the total number of sequences comprising that dataset to calculate the relative number of NIRVS-derived sRNAs for each k-mer. To assess nt bias at each position of the sRNA sequences, mapped sRNA reads (18–30 nt) were trimmed to 10 nt from the 3′ end and used as input in WebLogo (v3.0) (Crooks et al. 2004).

The sRNA dataset used contained two replicates each of mock infection, and two of infection with *Anaplasma phagocytophilum*, the causative agent of human granulocytic anaplasmosis. Although this was not a viral infection, we examined whether NIRVS transcription was upregulated upon general immune responses to bacterial infection. Forward reads from the two infection datasets (SRA ID SRR3236780-SRR3236781) were mapped to all NIRVS and quantified with Salmon (v0.8.2). This was repeated for the mock datasets (SRA ID SRR3236782-SRR3236783). Differential expression counts of the sRNAs were

calculated with DESeq2 (Galaxy v2.11.39) (Love, Huber, and Anders 2014).

3. Results

3.1 Abundance and Baltimore classification of NIRVS in arbovirus vectors

Recent literature has described a large number of *Aedes* NIRVS, so we initially analysed the genomes of *Ae. aegypti* and *Ae. albopictus* to confirm the robustness of our pipeline. Both species had a similar abundance of NIRVS (*Ae. aegypti*, *n* = 276, *Ae. albopictus*, *n* = 276). We then detected NIRVS in a further six arthropods that are known to transmit pathogenic vertebrate viruses. The *I. scapularis* genome harboured the next largest number of NIRVS (*n* = 143), followed by *Cx. quinquefasciatus* (*n* = 28), *An. gambiae* (*n* = 24), *C. sonorensis* (*n* = 4), *P. papatasi* (*n* = 2), and *R. microplus* (*n* = 1) (Supplementary Table S2). Integrations from negative-sense viruses dominated in most organisms: *Ae. aegypti*, *n* = 243 (88%), *Ae. albopictus*, *n* = 233 (84%), *I. scapularis* *n* = 142 (99%), *Cx. quinquefasciatus* *n* = 28 (100%), *An. gambiae* *n* = 24 (100%), and *C. sonorensis* *n* = 4 (100%) (Fig. 2A). In *P. papatasi* and *R. microplus* there was one −ssRNA derived NIRVS (Fig. 2A; Supplementary Table S2). All other NIRVS were related to either +ssRNA or dsRNA viral sequences (Figs 3 and 4). No NIRVS were identified in the negative control genome of *D. melanogaster*.

3.2 NIRVS are often duplicated within arthropod genomes

Many arthropods contain a high proportion of TEs in their genomes; *Ae. albopictus* and *Ae. aegypti* contain about 68 and 47 per cent TEs, respectively (Nene et al. 2007; Chen et al. 2015). Due to the important role of repeat sequences in arthropod evolution (Peccoud et al. 2017), we suspected that many of the NIRVS identified were generated due to post-insertion duplication and not by distinct integration events. Using a ClustalW alignment and distance matrix, we calculated whether NIRVS shared high sequence identity with other NIRVS within the genome. Of 276 NIRVS in *Ae. aegypti*, 126 (46%) shared at least 98 per cent nt identity with at least one other NIRVS. In *Ae. albopictus*, this number was 196 (71%), in *I. scapularis* it was 26 (18%); 20 in *Cx. quinquefasciatus* (71%), and 11 in *An. gambiae* (44%) (Supplementary Table S4). There was no evidence of duplicated NIRVS in other organisms with multiple NIRVS (*C. sonorensis* and *P. papatasi*) (Supplementary Table S4). In *Ae. aegypti* and *I. scapularis*, we visually represented duplicated NIRVS within the groups Mononegavirales, Bunyavirales, and *Orthomyxoviridae* (Fig. 5).

3.3 Phylogenetic classification of NIRVS

3.3.1 −ssRNA virus derived NIRVS

The identified NIRVS from negative-sense RNA viruses ('−ssRNA NIRVS') comprised three phylogenetic groups: mononega-, bunya-, and orthomyxo-like (Fig. 2B; Supplementary Fig. S1). In *Aedes* mosquitoes, −ssRNA NIRVS were dominated by unclassified mononegaviruses (*Ae. aegypti*, *n* = 113 [47%], *Ae. albopictus*, *n* = 151 [65%]), and the *Rhabdoviridae* (*Ae. aegypti*, *n* = 108 [44%], *Ae. albopictus*, *n* = 58 [25%]). Less abundant −ssRNA NIRVS were related to the *Chuviridae* (*Ae. aegypti*, *n* = 15 [6%], *Ae. albopictus* *n* = 11 [5%]), *Phasmaviridae* (*Ae. aegypti*, *n* = 2 [0.8%], *Ae. albopictus*, *n* = 6 [3%]), *Phenuiviridae* (*Ae. aegypti*, *n* = 1 [0.4%], *Ae. albopictus*, *n* = 3 [1%]), and *Orthomyxoviridae* (both *n* = 4 [2%]) (Fig. 2; Supplementary Fig. S1).

A similar pattern was observed in *I. scapularis* where most −ssRNA NIRVS were classified in the Mononegavirales in either the *Rhabdoviridae* (*n* = 44 [31% of −ssRNA NIRVS]) or among unclassified mononegaviruses (*n* = 48, 34%) (Fig. 2). NIRVS related to bunyaviruses (*n* = 29, 20%) were notably more abundant than in dipterans, and evenly distributed among the *Phenuiviridae* (*n* = 11, 8%) and *Nairoviridae* (*n* = 18, 13%) (Fig. 2). Orthomyxo-like NIRVS (*n* = 19, 13%) were the next most prevalent, followed by the *Chuviridae* (*n* = 2, 1%) (Fig. 2). In *Cx. quinquefasciatus* and *An. gambiae*, the −ssRNA NIRVS were related to unclassified mononegaviruses (*n* = 18 and *n* = 13 [64 and 54% −ssRNA NIRVS, respectively]), chuviruses (both *n* = 9 [32 and 38%, respectively]), or rhabdoviruses (*Cx. quinquefasciatus* *n* = 1, 4%) (Fig. 2). *An. gambiae* also contained both one phasma- and one phenui-like insertion (Supplementary Table S2).

C. sonorensis contained three NIRVS related to the *Phasmaviridae* and one to the *Chuviridae*, while the only −ssRNA NIRVS in *P. papatasi* was related to Wuchang Cockroach Virus 1 [KM817748] (Bunyavirales, *Phenuiviridae*) (Fig. 2). In *R. microplus* the only −ssRNA NIRVS was related to American dog tick phlebovirus [KM589348] (Bunyavirales, *Phenuiviridae*) (Supplementary Table S2 and Fig. S1).

3.4 Structural and non-structural region integration skew among mononega-like viral insertions

In our dataset, all dipterans possessed more integrations from structural viral regions (nucleoprotein [NP], glycoprotein [GP], and matrix [M] regions) over non-structural regions (replicase [L] region) among the mononega-like insertions (rhabdoviruses, unclassified mononegaviruses, and chuviruses). Specifically, *Ae. aegypti* had 213 structural of 236 total mononega-like NIRVS (90%) and *Ae. albopictus* possessed 197 of 221 (89%) (Supplementary Table S2). This pattern was consistent for *Cx. quinquefasciatus* (*n* = 28/28, 100%) and *An. gambiae* (*n* = 22/22, 100%) (Fig. 2A). Notably, *I. scapularis* possessed more non-structural [L] region integrations from mononega-like viruses (*n* = 76/94, 81%), with the remainder from structural regions (*n* = 18/94, 19%) (Fig. 2A). To visually demonstrate this difference, a comparative alignment between negative-sense NIRVS in *Ae. aegypti* and *I. scapularis* was created, demonstrating the skew towards either structural or non-structural integration in each organism (Fig. 5A).

In terms of specific structural integrations from mononega-like viruses, GP-derived NIRVS were overrepresented in *Ae. albopictus* (*n* = 158/221 mononega-like insertions, 76%), *Cx. quinquefasciatus* (*n* = 26/28, 93%), and *An. gambiae* (*n* = 22/22, 100%), with either fewer or no NP-derived NIRVS in these organisms (*n* = 48/209 [23%], *n* = 2/28 [7%], and *n* = 0/24 [0%], respectively). Alternatively, NP-derived NIRVS were more abundant in *Ae. aegypti* (*n* = 122/236, 52%) with fewer GP-derived integrations (*n* = 90/236, 38%), and a single integration was derived from the M gene (Fig. 5A) and another from a protein of unknown function (Supplementary Table S2). Similarly, structural protein-derived mononega-like NIRVS in *I. scapularis* were mostly derived from NP sequences (*n* = 40/42, 95%), with only two GP-related NIRVS (5%) (Fig. 5A).

3.4.1 +ssRNA virus derived NIRVS

The Hepe-Virga clade and the *Flaviviridae* were the two major lineages comprising NIRVS from +ssRNA viruses (Fig. 3; Supplementary Fig. S2). +ssRNA NIRVS related to the *Flaviviridae* were the most abundant, but were only observed in

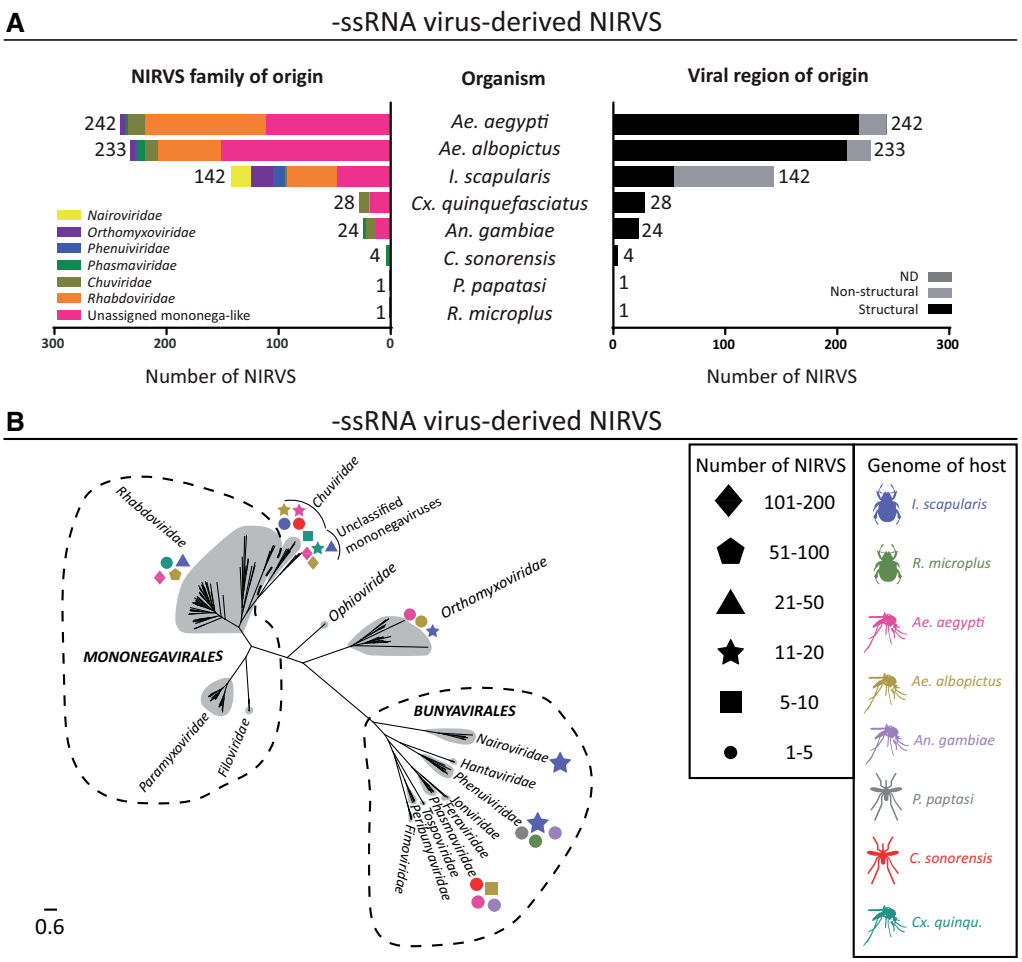

Figure 2. Characteristics of negative-sense RNA viral genome insertions in eight arthropod genomes. (A) NIRVS from eight organisms were analysed, which included number of NIRVS, viral family of origin (left-hand side), and viral structural or non-structural region of genome acquired (right-hand side). NIRVS were identified with a BLAST-based search, and the family of their virus of origin was deduced using aa identity to sequences from extant viral relatives. The total number of negative-sense RNA virus-derived NIRVS for each organism is shown next to the column bar. Bar colour is not visible for *P. papatasi* (*Phasmaviridae*, green) and *R. microplus* (*Phenuiviridae*, dark blue). (B) Phylogeny of negative-sense NIRVS in eight arthropods. Phylogeny was based on closest extant relative to NIRVS. Creation of a near-complete phylogeny of negative-sense RNA viruses was based on an alignment of the replicase (Large protein) region (2,192 aa positions derived from 176 RdRp sequences, as listed in Supplementary Table S3). Shapes correspond to the abundance of NIRVS from that group as shown in the legend, and colour represents the organism analysed (right-hand panels). Viral orders are encircled with dotted lines, and families are shaded grey. The scale bar represents aa substitutions per site.

Aedes mosquitoes (*Ae. aegypti*, *n* = 21 [75% +ssRNA NIRVS], *Ae. albopictus*, *n* = 29, 81%) (Fig. 3). This was also the case for Nege-like viruses (*Ae. aegypti*, *n* = 4 [14%], *Ae. albopictus*, *n* = 5 [14%]), and Hepe-Virga-like viruses (*Ae. aegypti*, *n* = 3 [11%], *Ae. albopictus*, *n* = 2 [6%]) (Fig. 3; Supplementary Fig. S2). No +ssRNA virus-derived NIRVS were detected in *P. papatasi*, *R. microplus*, or *D. melanogaster*. Putative Nido-like NIRVS were initially detected in *Ae. albopictus* (*n* = 2), and once each in *Ae. aegypti*, *Cx. quinquefasciatus*, *An. gambiae*, *C. sonorensis*, and *D. melanogaster*; however, the identity of this as a true NIRVS was disputed (see further discussion below, section 3.5).

3.4.2 dsRNA virus derived NIRVS

dsRNA virus-derived NIRVS were the least abundant across all organisms (*n* = 14 total; Supplementary Fig. S3). dsRNA NIRVS could be classified as either Partiti-like or Toti-like, i.e. relatives of the fungal and protozoan-infecting families *Partitiviridae* and *Totiviridae*, respectively (Fig. 4). In *Ae. aegypti* and *Ae. albopictus*, all dsRNA-like NIRVS (*n* = 5 and *n* = 7, respectively) were Toti-like (Fig. 4). *I. scapularis* and *P. papatasi* each contained one Partiti-like NIRVS only, whose closest relatives were Norway partiti-like virus 1 [MF141076] and Hubei partiti-like virus 31 [KX884162], respectively (Fig. 4; Supplementary Table S2 and Fig. S3).

3.5 Analysis of a protein with an unusual nido-like domain in six dipteran genomes

When searching for NIRVS, false positive hits can arise from viral query proteins that have chance homology to eukaryotic proteins (e.g. heat shock proteins). These can usually be eliminated with a reverse tBLASTn search against the *nr* database, where no viral hits are produced, as hits to eukaryotes are much stronger. We observed an unusual BLAST hit in *Ae. aegypti*, *Ae. albopictus*, *An. gambiae*, *Cx. quinquefasciatus*, *C. sonorensis*, and *D. melanogaster* with homology to the ORF1a (replicase) protein of mosquito-specific viruses in the *Mesoniviridae* (order Nidovirales) (Table 2; Table S2 and Fig. S2C). Reverse tBLASTn searches of the translated sequence against the *nr*

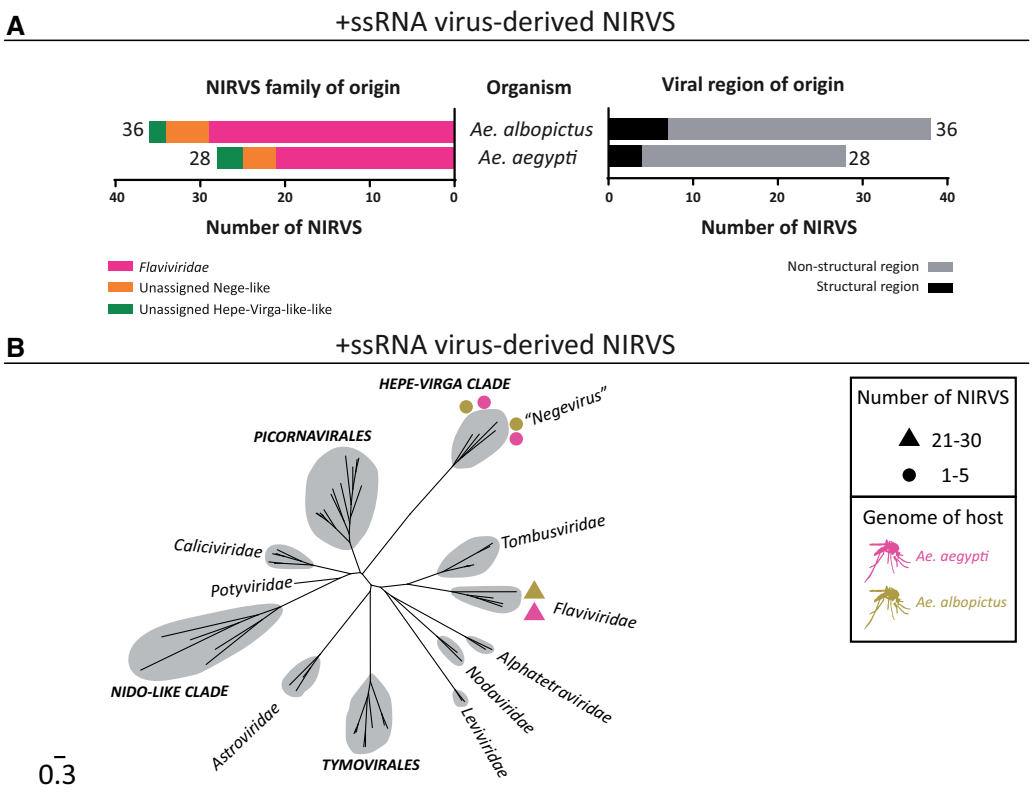

Figure 3. Characteristics of positive-sense RNA viral genome insertions in the *Ae. aegypti* and *Ae. albopictus* genomes. (A) NIRVS from the two mosquitoes were analysed, which included number of NIRVS, viral family of origin (left-hand side), and viral structural or non-structural region of genome acquired (right-hand side). NIRVS were identified with a BLAST-based search, the family of their virus of origin was deduced using aa identity to sequences from extant viral relatives. The total number of positive-sense RNA virus-derived NIRVS for each organism is shown next to the column bar. (B) Phylogeny of positive-sense NIRVS in *Ae. aegypti* and *Ae. albopictus*, based on closest relative as determined by BLAST search. Creation of a near-complete phylogeny of positive-sense RNA viruses was based on an alignment of the replicase region (4,967 aa positions, including gaps, derived from 56 RdRp sequences as listed in Supplementary Table S3). Shapes correspond to the abundance of NIRVS from that phylogenetic group as shown in the legend, and colour indicates the organism analysed (right-hand panels). Viral orders are encircled with dotted lines, and families are shaded grey. The scale bar represents aa substitutions per site.

database generated a list of close matches to hypothetical proteins from related insect species, in addition to more distantly related viral hits, so we did not fully eliminate this as a false positive hit.

We next analysed the flanking regions of this putative NIRVS, to see whether it was part of a functional protein coding sequence. The putative NIRVS was part of an open reading frame (ORF) spanning approximately 1,100 nt in all surveyed organisms, which encoded an annotated hypothetical protein in *Ae. aegypti*, *Ae. albopictus*, *An. gambiae*, *Cx. quinquefasciatus*, and *D. melanogaster* (Table 2). Therefore, it seemed that this hit represented a conserved protein containing a 'nido-like' domain (~600 nt identity, ranging from 26.1 to 45.3%) (Supplementary Table S2). We performed a BLASTp search of this conserved protein to find potential homologues but generated a list of results similar to the initial tBLASTn search (Supplementary Table S5).

3.6 Further phylogenetic analysis of NIRVS from *I. scapularis*

Owing to the emerging threat of tick-borne viruses, it is crucial to sample viral diversity in ticks. NIRVS can represent novel phylogenetic lineages which have not been sampled in the exogenous form. Therefore, we performed phylogenetic analysis on six *I. scapularis* NIRVS from different taxonomic groups.

This encompassed a wide range of genetic diversity, while using the longest sequence from each group ensured high phylogenetic resolution (Table 3 and Fig. 6).

The single partiti-like NIRVS (218 nt) was derived from the conserved RdRp region, and clusters monophyletically with Norway partiti-like virus 1 [MF141076] that was isolated from *I. ricinus* (Fig. 6A). These two sequences are nested within a larger clade which includes viruses sequenced from insects and crustaceans, including viruses from odonates/dragonflies (Hubei partiti-like virus 56 [KX884129] and 57 [KX884116]) and crustaceans (Beihai partiti-like virus 11 [KX884054]/Changjiang partiti-like virus 1 [KX884088]). Overall, these arthropod-specific partiti-like viruses form a clade related to, but distinct from, the *Partitiviridae* family (Fig. 6A). However, due to the short sequence length there is low bootstrap support (<50%) for most tree nodes (Fig. 6A).

The longest chuvirus-like NIRVS (509 nt, genomic scaffold DS620777) was related to the chuviral glycoprotein (M) gene. It sits basally to a larger cluster of arthropod-specific chuviruses; which includes a tick-specific cluster (Suffolk virus [KM460042], Bole tick virus 3 [KM817593], Changping tick virus 2 [KM817594], Lonestar tick chuvirus 1 [KU230451], and Wuhan tick virus 2 [KM817611]), and a mosquito-specific lineage including Gambie virus [KX148553] and Culex mononega-like virus 1 [MF176245] (Fig. 6B). Orthomyxo-like NIRVS in *I. scapularis* were more abundant than in any dipteran, and the longest (1079 nt, derived from the PB1 protein coding region), clustered in a group of

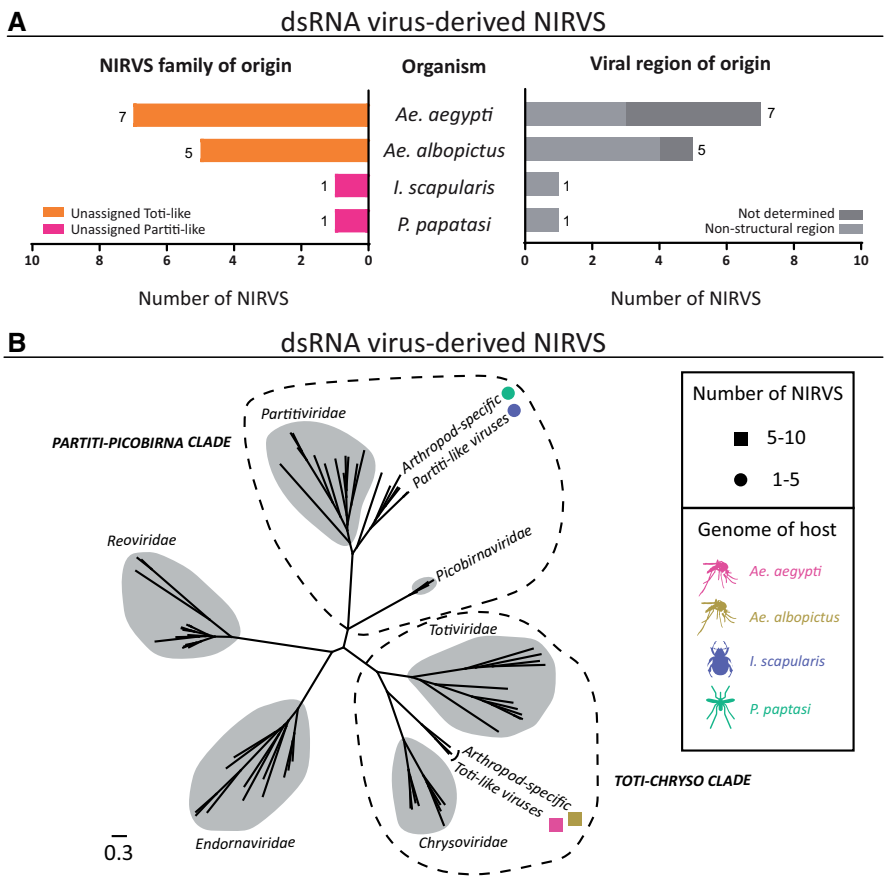

Figure 4. Characteristics of double-stranded RNA viral genome insertions in four arthropod genomes (*Ae. aegypti*, *Ae. albopictus*, *I. scapularis*, and *P. papatasi*). (A) NIRVS from four arthropods were analysed, including number of NIRVS, viral family of origin (left-handed side), and viral structural or non-structural region of genome acquired (right-handed side). NIRVS were identified with a BLAST-based search, the family of their virus of origin was deduced using aa identity to sequences from extant viral relatives. The total number of dsRNA virus-derived NIRVS for each organism is shown next to the column bar. (B) Phylogeny of double-stranded RNA virus NIRVS in the four arthropods, determined using the closest extant relative as determined by tBLASTn search. Creation of a near-complete phylogeny of dsRNA viruses was based on an alignment of the RdRp region (4,412 aa positions, including gaps, maintained from 72 RdRp sequences, as listed in Supplementary Table S3). Shapes correspond to the abundance of NIRVS from that phylogenetic group, as shown in the legend, and colour represents the organism (right-hand panels). Viral orders are encircled with dotted lines, and families are shaded grey. The scale bar represents aa substitutions per site.

tick- and avian-infecting orthomyxo-like viruses. These include Wellfleet bay virus [KM114304], Johnston Atoll virus [FJ861697], Quaranfil virus [FJ861695], and Tjuloc virus [JQ928944] (Fig. 6C).

The selected phenui-like NIRVS (derived from scaffold DS731648) clustered with newly described *Ixodes*-derived phenuiviruses including Norway phlebovirus 1 [MF141050] and Blacklegged tick phlebovirus 1 [KJ746873] and 2 [KJ746874]. This cluster is distinct from sandfly borne human pathogens (e.g. Rift Valley Fever phlebovirus), and from the well-described tick-specific Uukuniemi virus clade (Fig. 6D). The NP-derived nairo-like NIRVS falls into an *Ixodes*-specific cluster within the *Nairoviridae*, distinct from several clusters infecting either hard ticks (Ixodidae) or soft ticks (Argasidae) (Fig. 6E). Lastly, the rhabdo-like NIRVS which was particularly long (3,950 nt) can be classified into a tick-specific rhabdo-like cluster, where its closest relative is Norway mononegavirus 1 (Fig. 6F). This classification notably did not align with its closest BLAST result which was the equine-infecting vesicular stomatitis Indiana virus (Table 3).

3.7 NIRVS in the genome of *I. scapularis* are disproportionately present in piRNA loci

Using proTRAC (v2.4.2), 1,774 piRNA clusters were predicted in *I. scapularis* which comprised a total of 8,129,613 bp (0.46% of the genome). Thirty-two NIRVS (22%) resided totally within these clusters, with the remaining 111 either partially present or absent from the clusters. The cumulative binomial probability that at least 32 NIRVS integrated into piRNA clusters was $P < 0.000001$, indicating a strong preference for integration into these sites.

3.8 Viral integration sites in *I. scapularis* are enriched for non-long-terminal repeat retrotransposons

In *Ae. aegypti*, LTR retrotransposons (Pao_Bel/Ty3_gypsy) are associated with NIRVS integration (Palatini et al. 2017). In *I. scapularis*, genomic regions flanking NIRVS were associated with a two-fold enrichment of fragments of non-long-terminal repeat (non-LTR) class I retrotransposons (6.70% in the whole genome vs. 11.55% in NIRVS flanking sites), but not of LTR retrotransposons (0.64 vs. 0.87%) or DNA transposons (3.06 vs. 3.70%) (Table 4). Among the non-LTR retrotransposon fragments, the most relatively enriched were R1 elements (0.00 vs. 0.78%), *I* elements (0.53 vs. 4.71%), and L1 elements (2.09 vs. 4.30%) (Table 4). The NIRVS 'hotspot' on scaffold DS826508 also contained a high proportion of non-LTR retrotransposons (9.02%), but not as many *I* elements (0.88%) or R1 elements (0.00%) (Table 4).

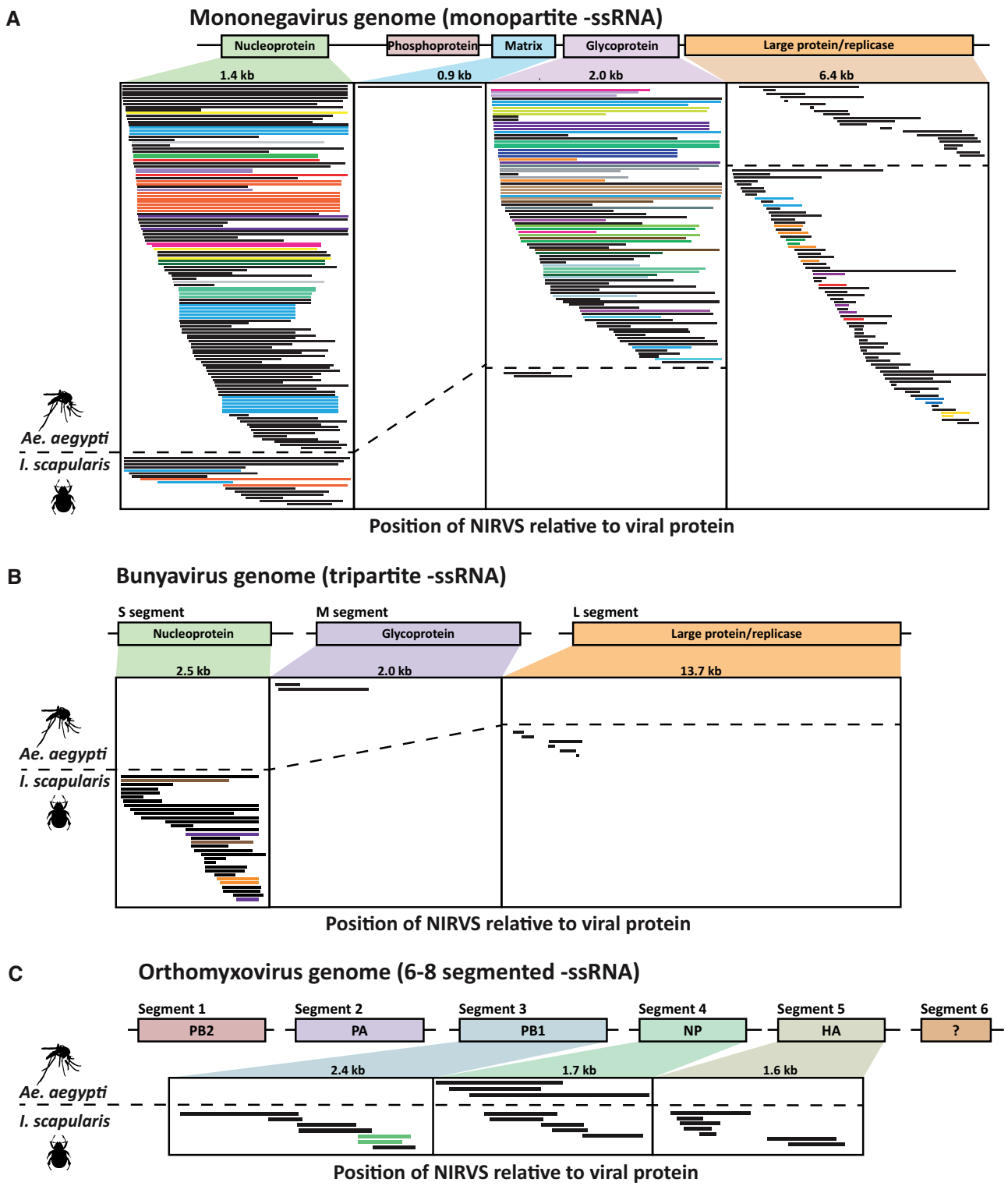

Figure 5. Comparative diagram showing the relative viral genomic position of negative-sense NIRVS from two arthropod species (*Ae. aegypti* and *I. scapularis*). NIRVS were aligned individually to a representative viral reference sequence from each family using ClustalW and the position of the NIRVS relative to the viral genome with GraphPad Prism. The approximate size of each genomic region or segment is shown below in kb and each is colour coded. The dotted line divides *Ae. aegypti* NIRVS from those of *I. scapularis*. The x-axis represents the coverage of the reference viral genome as a percentage, whilst the y-axis shows the identified NIRVS. (A) Comparative genomic alignment of mononega-like NIRVS to a reference genome. Mononegaviruses, as defined here, include the order Mononegavirales and unclassified relatives, including the *Chuviridae* family. (B) Comparative genomic alignment of bunyavirus-like NIRVS to a reference bunyavirus (order *Bunyavirales*) and (C) comparative genomic alignment of orthomyxo-like NIRVS to a reference orthomyxovirus (family *Orthomyxoviridae* and unclassified relatives). NIRVS are the same colour if they represent duplications of the same sequence (>95% identity over >200 nt). NIRVS derived from insertions with no duplication are coloured black. ORF, open reading frame; S/M/L segment, small/medium/large segment; PB1/2, polymerase basic 1/2; PA, polymerase acidic; NP, nucleoprotein; HA, haemagglutinin. Structural proteins are NP, matrix protein, phosphoprotein, GP, HA. Non-structural proteins are replicase/large protein, PB1, PB2, and PA.

Table 2. Details of a protein of unknown function with a nidovirus-like domain present in the genomes of six dipterans.

Organism	CDS/Gene ID of Nido-like protein	Genomic scaffold/chr ID (nt position)	ORF length (spliced) (nt)	Closest viral relative [virus]	Closest hit (BLASTp) (% aa identity)	Length of region of similarity (aa)
An. gambiae	AGAP007003	chr2L_CRA_x9P1GAV591D/ (8,770,648–8,771,802)	1146	ORF1a [Nse virus]	30.7	199
Ae. albopictus	XP_019547702.1/	NW_017856913/ (2,945,395–2,945,982)	1,260	pp1a polyprotein [Dak	34.4	195
	XP_019534630.1	NW_017858044/ (769,840–769,256)	1,299	Nong virus]	35.4	195
D. melanogaster	CG7504	chr3L/ (8,137,217–8,131,997)	4,254	No match upon re-blasting the translated protein sequence against the *nr* database.	–	–
Cx. quinquefasciatus	CPIJ017672	supercont3.1009/ (77,064–78,058)	930	ORF1 [Alphamesonivirus 1]	39.3	196
Ae. aegypti	LOC5779059	chr2/ (102,012,425–102,013,560)	1038	ORF1a [Nse virus]	34.2	
C. sonorensis	N/A	scaffold3214/ (6,204–6,554)	351	pp1a polyprotein [Nam Dinh virus]	34.8	112

Table 3. Basic characteristics of six NIRVS which represent the longest of each phylogenetic group from *I. scapularis*.

I. scapularis genomic scaffold of origin	Closest relative as determined by BLAST search [GenBank accession]	Eukaryotic host of closest viral relative as determined by BLAST search	NIRVS length (nt)	Phylogenetic group	Viral region of origin
DS710489	Norway partiti-like virus 1 [MF141076]	*I. ricinus*	218	Partiti-like viruses	RdRp
DS620777	Blacklegged tick chuvirus 2 [MF360789]	*I. scapularis*	509	Chu-like viruses	Glycoprotein
DS806788	Wellfleet Bay virus [KM114305]	*Somateria mollissima* (common eider)	1,079	Orthomyxo-like viruses	PB1/Polymerase basic 1
DS731648	Blacklegged tick phlebovirus 3 [KU230449]	*I. scapularis*	980	Phenui-like viruses	Large protein/Replicase
DS810239	South Bay virus [KJ746878]	*I. scapularis*	1,580	Nairo-like viruses	NP
DS847572	Vesicular stomatitis Indiana virus strain [AF473864]	*Equus* spp.	3,950	Rhabdo-like viruses	Large protein/Replicase

3.9 sRNAs originating from NIRVS are present in *I. scapularis* ISE6 cells and resemble primary piRNAs

It has been shown that piRNAs are produced by NIRVS in a variety of arthropods (Palatini et al. 2017; Ter Horst et al. 2018). To determine whether this was the case in our own dataset, we analysed a publicly available *I. scapularis* sRNA dataset for the presence of piRNAs derived from NIRVS loci (BioProject accession PRJNA315659). Out of 132,048,347 sRNA reads originating from four independent sequencing runs, 68,823 (0.05%) mapped exactly to at least one NIRVS. These originated primarily from NIRVS related to rhabdoviruses (35.6%), unassigned mononegaviruses (24.2%), orthomyxoviruses (18.5%), phenuiviruses (16.2%), nairoviruses (3.6%), and chuviruses (1.9%); these proportions are very similar to the abundance of each NIRVS family within the genome (Fig. 7A).

The most abundant NIRVS-derived sRNAs were 28 nt in length (0.046% total 28-mers), followed by 29 nt (0.045%), 27 nt (0.039%), 26 nt (0.035%), 30 nt (0.034%), 25 nt (0.028%), 24 nt (0.017%), 23 nt (0.012%), 20 nt (0.008%), 21 nt (0.007%), 22 nt (0.004%), and 19 nt (0.003%) (Fig. 7B). Additionally, NIRVS-derived sRNAs were predominantly antisense (20,538 reads, 90.9%) as opposed to sense (2,067 reads, 9.1%) (Fig. 7B). The length of the NIRVS-derived sRNAs (primarily 25–30 nt sRNAs) their predominantly antisense nature, and the strong bias for a uridine at their 5′ position make them strongly reminiscent of primary piRNAs (Fig. 7B and C).

In the range of 18–30 nt, the most abundantly expressed NIRVS were DS841316 (Rhabdo-like, L protein, 2,349 reads), DS884533_1 (Rhabdo-like, L protein, 1,511 total reads), DS731648_2 (Phenui-like, L protein, 1,908 reads), DS810239_2 (Nairo-like, N protein, 1,572 reads), and DS826508_3 (Orthomyxo-like, N protein, 1,192 reads) (Supplementary Table S6). No sRNAs were significantly up- or down-regulated (adjusted *P*-value <0.05) upon infection with *A. phagocytophilum*.

4. Discussion

Arthropods that transmit viruses to vertebrate hosts maintain a fine balance between carrying a viral infection for an extended

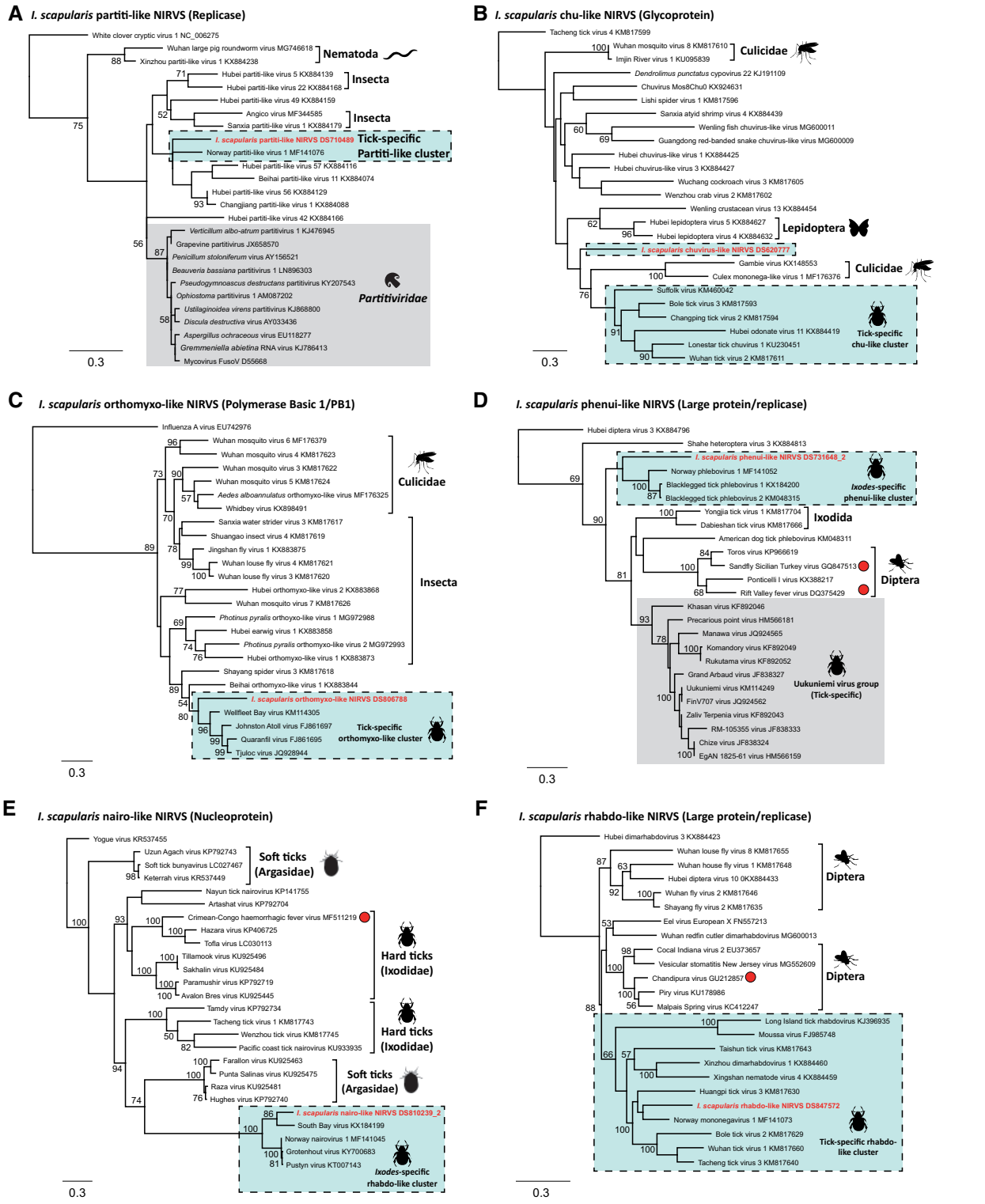

Figure 6. ML phylogenetic trees of six NIRVS from the blacklegged tick *I. scapularis*. Following identification of 143 NIRVS within the *I. scapularis* genome, the longest from each family of negative-sense viruses was selected for phylogenetic analysis ($n = 5$). A dsRNA NIRVS was also selected as it was the unique NIRVS of this type identified (A). First, a tBLASTn search was performed using the NIRVS against the NCBI nr database to generate a list of close matches. The top 25 most closely related sequences were chosen, as well a more distantly related outgroup sequence. Sequences were aligned with MAFFT (v7.017) and trimmed with trimAl (v1.4.1). A RAxML (v8.2.8) ML tree was constructed with gamma model rate heterogeneity and rapid bootstrapping (500 replicates). Bootstrap support (%) is shown next to the node with values below 50 per cent not displayed. (A) *Partitiviridae* and partiti-like viruses (46 aa positions maintained after trimming), (B) *Chuviridae* NP-like NIRVS (110 aa positions maintained after trimming), (C) *Orthomyxoviridae* PB1-like NIRVS (230 aa positions maintained after trimming), (D) *Phenuiviridae* L/replicase-like NIRVS (226 positions maintained after trimming), (E) *Nairoviridae* NP-like NIRVS (379 aa positions maintained after trimming), (F) *Rhabdoviridae* L/replicase-like NIRV (812 aa positions maintained after trimming). If a monophyletic lineage could be assigned to a particular group of hosts, it was annotated. Known human pathogens are indicated with a red circle. Scale bars represent aa substitutions per site.

Table 4. Transposable element (TE) occupancy in the entire *I. scapularis* genome, in regions with NIRVS, and in a region containing multiple NIRVS from three different viral groups (located on scaffold DS826508).

	TE occupancy (%)/TE content (bp)		
	IscaW1 genome assembly (1,765,382,190 bp)	NIRVS-containing regions (±5 kb) (487,490 bp)	DS826508 'hotspot' region (± 5kb) (75,412 bp)
LTR retrotransposons (Class I)	**0.64/11,383,395**	**0.87/4,229**	**0.0/0**
Pao_Bel	0.01/194,086	0.03/140	0.0/0
Ty3_gypsy	0.63/11,189,309	0.84/4,089	0.0/0
Non-LTR retrotransposons (Class I)	**6.70/118,212,063**	**11.55/56,329**	**9.02/6,803**
CR1	1.50/26,561,455	1.55/7,554	2.79/2,106
I	0.53/9,402,964	4.71/22,954	0.88/660
L1	2.09/36,843,465	3.66/17,853	3.39/2,558
L2	0.66/11,639,922	0.85/4,145	1.96/1,479
R1	0.00/61,781	0.78/3,823	0.0/0
DNA transposons (Class II)	**3.06/54,005,181**	**3.70/18,038**	**3.92/2,954**
P	0.28/4,859,952	0.51/2,482	0.75/565
piggyBac	1.20/21178514	1.52/7,388	0.58/438
hAT	0.42/7,362,901	0.97/4,718	1.76/1,328
Tc1mariner	0.75/13,289,414	0.29/1,402	0.32/240
PIF	0.42/7,362,901	0.42/2,048	0.51/383
MITEs	**4.96/87,535,895**	**2.79/13,607**	**1.49/1,122**
Penelope	**1.08/19,113,444**	**1.56/7,613**	**0.08/60**
Unclassified TEs	**0.33/5,849,509**	**0.42/2,040**	**0.17/129**

period and avoiding the deleterious effects of viral infection. NIRVS are proposed to be a heritable antiviral defence mechanism, interacting with sRNA SPs to attenuate transcripts of an infecting virus (Palatini et al. 2017). It is now well-established that NIRVS accumulation is a feature of multiple arthropod classes (Ter Horst et al. 2018). Yet besides *Aedes* spp., an in-depth analysis of NIRVS in emerging arbovirus vectors is lacking. To address this gap, we selected six non-*Aedes* arbovirus vectors for NIRVS analysis. We uncovered more NIRVS in the mosquitoes *Cx. quinquefasciatus* and *An. gambiae* than reported previously, and we characterised novel NIRVS in the non-mosquito dipterans *C. sonorensis* and *P. papatasi*, and in the Asian blue tick *R. microplus*. The focus of our study, however, became the preponderance of NIRVS in the deer tick *I. scapularis*, which exhibit unusual characteristics and are likely to be associated with the piRNA pathway.

4.1 Validation of methodology with *Aedes* mosquitoes

Ae. aegypti and *Ae. albopictus* are the best-studied arthropods in the context of NIRVS and were therefore used as validation organisms for our NIRVS identification pipeline. Our count in *Ae. aegypti* (n = 276) (Figs 2–4) was considerably more than Palatini et al. (2017) (n = 122) who used a similar BLAST-based approach, and a recent version of the same genome (AaegL3.0) (Palatini et al. 2017). Our results may reflect a more extensive range of query viruses (n = 1,933 vs. n = 425). Alternatively, Whitfield et al. (2017) reported 472 NIRVS in *Ae. aegypti*, but used a long-read based genome assembly for NIRVS identification. Since the current AaegL5.0 assembly is based on short-read technology, highly repetitive NIRVS-rich regions may be masked (Dudchenko et al. 2017; Whitfield et al. 2017). These variable figures suggest that the number of NIRVS reported is not definitive but is contingent on filtering parameters and the quality of host genome assembly.

In *Ae. albopictus*, we reported 276 NIRVS (Figs 2–4), also higher than Palatini et al. (n = 72), yet they used a genome from a different source (Foshan strain as opposed to C6/36 cell line used in our study). However, the true number of NIRVS in *Ae. albopictus* is probably as high as or higher than *Ae. aegypti*, as it contains an even higher proportion of TEs (Chen et al. 2015), including LTR retrotransposons which may be responsible for NIRVS formation. Despite the discrepancies, both studies report similar NIRVS characteristics. These NIRVS were dominated by insertions from the structural gene-coding regions (N and G) of negative-sense viruses (predominantly mononegaviruses), with fewer insertions from +ssRNA and dsRNA viruses. This indicates that the methodology of the current study, which identified NIRVS without generating excessive false positive hits, is robust.

4.2 Enriching the repertoire of NIRVS in non-*Aedes* dipterans

We observed that *Cx. quinquefasciatus* and *An. gambiae* harboured approximately twenty NIRVS each (Figs 2–4; Supplementary Table S2), which is inconsistent with Palatini et al. (2017) who identified no NIRVS in *An. gambiae*, and only one NIRVS in *Cx. quinquefasciatus*. These mosquitoes therefore possess more NIRVS than previously thought, with similar characteristics to *Aedes* NIRVS, i.e. dominated by integrations from structural regions of negative-sense viruses (Fig. 2A; Supplementary Table S2).

In this study, we included two non-mosquito dipterans which transmit emerging mammalian viruses (Table 1). Analysis revealed NIRVS in *P. papatasi* (n = 2) (bunya- and partiti-like) and *C. sonorensis* (n = 4) (chu- and bunya-like) (Figs 2

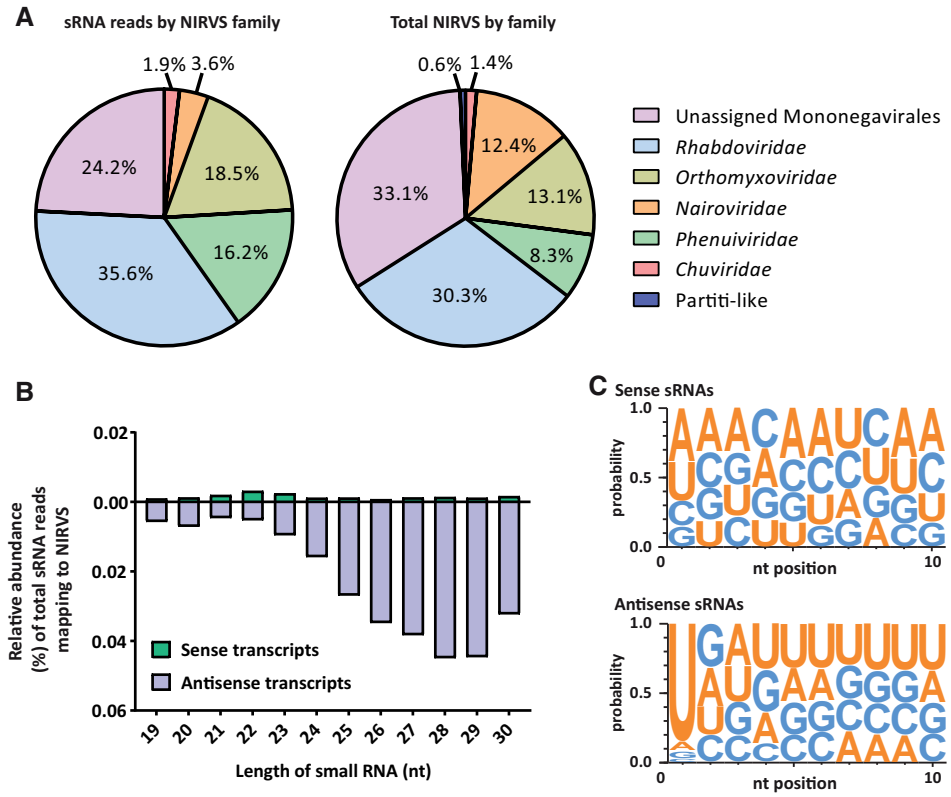

Figure 7. sRNAs in an *I. scapularis* cell line are derived from almost all groups of NIRVS present within the genome and resemble primary piRNAs. (A) sRNA reads derived from a publicly available dataset [BioProject Accession PRJNA315659] map to genomic NIRVS organised by family (left pie chart). This is compared with the abundance of different families of NIRVS within the genome (right pie chart). (B) Abundance of different lengths of NIRVS-derived sRNAs (19–30 nt) for each k-mer, originating from both the sense (green) and antisense (purple) genomic strand. (C) Sequence logo depicting ribonucleotide preference (probability) among the first ten nucleotides of NIRVS-derived sRNAs, from both the sense and antisense genomic strands.

and 4; Supplementary Table S2). These were related to insect-specific viruses discovered in metatranscriptomic studies, yet from a wide range of hosts and all with distant homology (mean 31.4% aa identity) (Supplementary Table S2), demonstrating that these NIRVS probably originated from highly divergent viruses. The virome of these organisms has not been extensively sampled, yet our results imply the presence of circulating, divergent, and arthropod-specific viruses in these species, which may be important for future studies. Overall, our results suggest that NIRVS accumulation is primarily a feature of mosquitoes, and not other dipterans.

4.3 A protein with an unusual nido-like domain is present in multiple dipteran genomes

NIRVS are not usually present as protein coding regions, but as fragmented sequences containing stop codons. However, viral proteins can be repurposed by their hosts (exapted) in remarkable instances of horizontal evolution (Koonin and Krupovic 2018). A BLAST hit in six dipteran genomes revealed the presence of a hypothetical protein containing a nido-like domain (Table 2; Supplementary Table S5). A BLASTp search of this protein generated hits to other insect genomes suggesting widespread conservation among insect lineages (Supplementary Table S5). Although hits to nidoviruses still arose, homology to these was generally weaker than to other insect proteins (e.g. in *Ae. aegypti*), the hypothetical protein LOC5579059 exhibited 52 matches to uncharacterised proteins from other insects (E-values ranging from 0 to $2.07E^{-19}$) before a viral hit occured (E-value $2.43E^{-19}$) (Supplementary Table S5). The pp1a nidovirus protein, to which BLAST matches were generated, is a multifunctional replicase enzyme (Ziebuhr 2006). We speculated that this region of similarity could signify a conserved enzymatic function between virus and eukaryote, e.g. helicase; however, no functional protein motifs were observed in this viral region. Due to the nature of the putative NIRVS it could represent an exapted viral protein which integrated prior to the divergence of the hosts, now serving a novel function in present-day lineages. This possibility is especially interesting as arthropods are predicted to be the origin of some present-day viral groups (Marklewitz et al. 2015). Alternatively, it could represent a case of host-to-virus exaptation, or protein mimicry by the virus (Alcami 2003), although these possibilities are unlikely in the case of an RNA virus. Further analysis of these NIRVS are beyond the scope of this manuscript, but these possibilities present an intriguing opportunity for future analysis.

4.4 Distinct and unusual patterns of viral integration in the deer tick *I. scapularis*

We revealed an abundance of NIRVS in *I. scapularis* (n = 143), which is broadly consistent with other studies of *Ixodes* spp. who report >100 NIRVS (Ter Horst et al. 2018). The majority of our NIRVS generated a BLAST match to viruses isolated from hard ticks (Supplementary Table S2), and phylogenetic analysis of six NIRVS supported their classification into tick-specific viral clades (Fig. 6).

We found both similarities and differences between NIRVS in *I. scapularis* and dipterans. Like the dipterans, *I. scapularis* NIRVS were dominated by negative-sense viruses (n = 142, 99%) which consisted primarily of insertions from non-segmented negative-sense (NNS) viruses (i.e. mononegaviruses) (n = 92, 64%) (Figs 2 and 5; Supplementary Table S2). Yet among the NNS viral insertions, we observed a bias towards non-structural region integration (L protein) as opposed to structural regions (G and N), which are conversely more abundant in mosquitoes (Fig. 5A). This bias towards NIRVS from the structural region of negative-sense viruses is thought to reflect relative mRNA abundance during infection, where the NP is the most abundant and L the least (Holmes 2011). Here, the unexpected abundance of L integrations could reflect an unusual template preference for the endogenous reverse transcriptases (RTs) and integrases in *I. scapularis* which catalyse NIRVS formation.

Another unusual aspect of NIRVS in *I. scapularis* was the overrepresentation of integrations from segmented negative-sense (SNS) RNA viruses (orthomyxo-like and bunya-like) (n = 51, 36%) (Figs 2 and 5). In contrast, all other surveyed organisms contained few SNS virus-like NIRVS (*Ae. aegypti* n = 6, *Ae. albopictus* n = 13, *An. gambiae* n = 2, *C. sonorensis* n = 3, *R. microplus* n = 1) (Fig. 2; Supplementary Table S2). Specifically, the orthomyxo-like NIRVS in *I. scapularis* were exclusively related to tick-borne viruses in the genus *Quaranjavirus* (Presti et al. 2009). Orthomyxoviruses replicate in the nucleus (Ruigrok et al. 2010), where proximity to the host genome may predispose them to NIRVS formation. However, studies on orthomyxovirus replication are heavily biased towards influenza viruses, and mechanisms of quaranjavirus replication are little explored. Similar observations were made for the bunya-like insertions, again abundant in *I. scapularis* (n = 30) despite being infrequent in other surveyed organisms. These could be classified as phenui-like or nairo-like (related to the *Phenuiviridae* or *Nairoviridae*). In contrast, phasma-like (*Phasmaviridae*-related) insertions were confined to *Ae. aegypti*, *Ae. albopictus*, *P. papatasi*, and *An. gambiae* (Supplementary Table S2). This mirrors the known host range of these families, whereby the *Nairoviridae* is tick-borne, the *Phenuiviridae* infects insects and arachnids, and the *Phasmaviridae* is insect-borne (Lasecka and Baron 2014). Overall, the detection of multiple SNS viral insertions here suggests that ticks are the dominant hosts of these viral groups.

Curiously, *I. scapularis* lacked NIRVS from positive-sense RNA viruses, which comprise about one tenth of NIRVS in *Aedes* mosquitoes (Fig. 3; Supplementary Table S2). In particular, flavivirus-derived NIRVS are the most abundant positive-sense group in mosquitoes, yet despite the existence of a tick-borne *Flavivirus* clade, we did not observe any flavi-like NIRVS in *I. scapularis*. This is surprising, as tick-borne flaviviruses are vertically transmitted in nature (Brackney and Armstrong 2016). Their absence could reflect that positive-sense viral genomes are unfavourable templates for ixodid NIRVS formation. Alternatively, since most tick-borne arboviruses are confined to certain groups (Bunyavirales, *Orthomyxoviridae*, *Flaviviridae*), tick viruses outside these groups are significantly under-sampled (Junglen 2016). Accordingly, identification of tick-specific viruses in other groups may reveal positive-sense NIRVS which are too divergent to be detected here.

4.5 Differential NIRVS landscape among the hard ticks

Analysis of the *R. microplus* genome revealed only one NIRVS, a much lower figure than *I. scapularis* (n = 143) (Fig. 2; Supplementary Table S2). This is an interesting disparity, both ticks contain a high proportion of repetitive elements (~70%) (Barrero et al. 2017). Our data suggest that NIRVS accumulation is perhaps confined to *Ixodes* ticks, an observation validated by another study which also identified numerous NIRVS in *I. ricinus* (Ter Horst et al. 2018). This difference could reflect factors such as inefficient vertical viral transmission among *Rhipicephalus* spp., or an alternative TE landscape which does not favour NIRVS formation. Importantly, genome sequencing of more hard and soft ticks will help to further characterise tick NIRVS.

4.6 Insights into the vertical transmission of *I. scapularis* viruses in nature

A requirement for the formation of an NIRVS is infection of a germ cell, permitting transmission of the integrated viral sequence to offspring (Aiewsakun and Katzourakis 2015). Each NIRVS is the footprint of a germline viral infection, suggesting vertical transmission of all the viral NIRVS groups identified here. Although vertical transmission of tick viruses is proposed to occur in nature (Labuda and Nuttall 2004), information is limited. Our data suggest that in *I. scapularis*, vertical transmission of mononegaviruses, bunyaviruses, and orthomyxoviruses may occur regularly, giving them ample opportunity to integrate.

Most of the NIRVS here appeared to originate from tick-specific viral groups (Fig. 6). An exception was NIRVS from the genus *Quaranjavirus* (*Orthomyxoviridae*) (Fig. 6C). Quaranjaviruses appear to have a dual vertebrate-arthropod host range; e.g., Wellfleet Bay virus can cause mass mortality in wild bird populations (Allison et al. 2015). The presence of NIRVS derived from this group therefore implies the potential for vertical transmission of quaranjaviruses, which has implications for the management of these viruses in wild tick populations.

4.7 NIRVS are associated with the piRNA pathway in *I. scapularis*

sRNA SPs are the key antiviral response in insects (Blair and Olson 2015), yet current research is heavily biased towards mosquitoes. Several studies indicate involvement of an sRNA response in tick flaviviral infection (Schnettler et al. 2014; Weisheit et al. 2015), yet the area is understudied. In particular, the observation that NIRVS are proximal to piRNA clusters for a variety of arthropods suggests that NIRVS either (1) integrate into these clusters preferentially or (2) are positively selected for post-integration (Palatini et al. 2017; Whitfield et al. 2017; Ter Horst et al. 2018). Interestingly, despite the identification of multiple NIRVS in *I. scapularis* in another study, no association with piRNA clusters was noted (Ter Horst et al. 2018). Repetition of this analysis for our own dataset uncovered a strong preference for integration into predicted piRNA sites in the *I. scapularis* genome ($P \ll 0.000001$). Despite this discrepancy, our observation of many NIRVS-derived piRNAs, as discussed below, suggests that our clustering analysis was robust.

The second line of evidence for association with the piRNA pathway is the production of NIRVS-specific piRNAs. To investigate this phenomenon in the blacklegged tick, we analysed datasets from *I. scapularis* ISE6 cells (n = 4) for the presence of NIRVS-derived sRNAs. We found that NIRVS-derived sRNAs from all families (except a partiti-like insertion) were produced proportionally to their abundance within the genome (Fig. 7A). Further analysis revealed that these sRNAs exhibited characteristics of primary piRNAs; specifically; they were primarily 25–0 nt in length, antisense, and had the canonical 1U piRNA bias (Fig. 7B and C). This strengthens the hypothesis that the

NIRVS-piRNA association exists in *I. scapularis* as it does in other arthropods. Further studies can elucidate whether NIRVS-derived piRNAs are upregulated upon viral infection, and whether suppressing the piRNA pathway confers an antiviral effect. Additionally, virus-specific piRNA production is known to differ between germ-line (the origin of the ISE6 cell line) and somatic tissues in most arthropods (Lewis et al. 2018). Therefore, it would be interesting to determine whether these NIRVS-derived piRNAs are also produced in somatic tissues where an infecting virus encounters the tick's main antiviral defences.

4.8 Non-LTR retrotransposons are the predominant TE class associated with NIRVS integration in *I. scapularis*

The production of cDNA from viral RNA is critical to the formation of an NIRVS. This requires RT activity, which is likely provided by the retroelements common in eukaryotic genomes (Holmes 2011). Although the association between NIRVS and the piRNA pathway is clear for a variety of arthropods, analysis of the TE classes associated with NIRVS only exists for *Ae. aegypti*, where LTR retrotransposons are increased two-fold in NIRVS-rich regions (Palatini et al. 2017). Our analysis indicated that non-LTR retrotransposons were instead two-fold enriched (6.70 vs. 11.55%) in NIRVS-associated regions in the *I. scapularis* genome, while LTR retrotransposon sequences were not increased (0.64 vs. 0.87%) (Table 4). This association corresponds to the diminished proportion of LTR retrotransposons in ticks (0.64% of the genome) compared with mosquitoes (about 15% in the genomes of *Ae. albopictus*, *Ae. aegypti*, and *Cx. quinquefasciatus*) (Chen et al. 2015; Gulia-Nuss et al. 2016). Our data suggest that the RT activity of non-LTR retrotransposons, particularly *I*, L1, and R1 elements, which were increased in NIRVS-rich regions (Table 4), might be responsible for NIRVS formation in ixodid ticks. However, further studies of the reverse transcription of viral RNA in tick cells is warranted to support this hypothesis.

5. Conclusion

Here, we have provided novel insights into NIRVS in the genomes of important arbovirus vectors. Our data suggest that NIRVS in *I. scapularis* are distinct in both phylogeny and viral region of origin. Furthermore, our data indicate a link to the piRNA pathway in *I. scapularis* and an association with non-LTR retrotransposons. Due to the increased incidence of tick-borne viruses, further studies into how this affects tick antiviral immunity are warranted.

Data availability

Supplementary tables accompanying this text are available at https://arusso1.github.io/NIRVS/.

Supplementary data

Supplementary data are available at *Virus Evolution* online.

Acknowledgements

This research includes computations using the Linux computational cluster Katana supported by the Faculty of Science, UNSW, Australia. A.G.R. and D.E.T. acknowledge support through Australian Government Research Training Program Scholarships. All photographic images in Fig. 1 are free for use under Creative Commons and have been modified. Image credits: *C. sonorensis*, *P. papatasi*, *D. melanogaster*, *I. scapularis*, *R. microplus*: CBG Photography Group, Centre for Biodiversity Genomics, *Ae. albopictus*: US Department of Agriculture, *Cx. quinquefasciatus*: CDC, *An. gambiae*: CDC/Dr. Darsie, *Ae. aegypti*: CDC/Prof. Frank Hadley Collins, Dir., Cntr. for Global Health and Infectious Diseases, University of Notre Dame.

Authors' contributions

A.G.R., A.G.K., D.E.T., M.M.T., and P.A.W. conceived project and designed experiments. A.G.R., and A.G.K. performed experiments and analysed data. A.G.R. and P.A.W. wrote the manuscript. All authors read, edited, and approved the manuscript before submission.

Conflict of interest: None declared.

References

Adams, M. D. et al. (2000) 'The Genome Sequence of *Drosophila melanogaster*', *Science*, 287: 2185–95.

Afgan, E. et al. (2018) 'The Galaxy Platform for Accessible, Reproducible and Collaborative Biomedical Analyses: 2018 Update', *Nucleic Acids Research*, 46: W537–44.

Aiewsakun, P., and Katzourakis, A. (2015) 'Endogenous Viruses: Connecting Recent and Ancient Viral Evolution', *Virology*, 479–480: 26–37.

Alcami, A. (2003) 'Viral Mimicry of Cytokines, Chemokines and Their Receptors', *Nature Reviews Immunology*, 3: 36–50.

Allison, A. B. et al. (2015) 'Cyclic Avian Mass Mortality in the Northeastern United States Is Associated with a Novel Orthomyxovirus', *Journal of Virology*, 89: 1389–403.

Anderson, J. F., and Armstrong, P. M. (2012) 'Prevalence and Genetic Characterization of Powassan Virus Strains Infecting *Ixodes scapularis* in Connecticut', *The American Journal of Tropical Medicine and Hygiene*, 87: 754–9.

Andrews, S. (2010) *FastQC: A Quality Control Tool for High Throughput Sequence Data* <http://www.bioinformatics.babraham.ac.uk/projects/fastqc> accessed 8 Feb 2019.

Aravin, A. A., Hannon, G. J., and Brennecke, J. (2007) 'The Piwi-piRNA Pathway Provides an Adaptive Defense in the Transposon Arms Race', *Science (New York, N.Y.)*, 318: 761–4.

Barrero, R. A. et al. (2017) 'Gene-Enriched Draft Genome of the Cattle Tick *Rhipicephalus microplus*: Assembly by the Hybrid Pacific Biosciences/Illumina Approach Enabled Analysis of the Highly Repetitive Genome', *International Journal for Parasitology*, 47: 569–83.

Belyi, V. A., Levine, A. J., and Skalka, A. M. (2010) 'Sequences from Ancestral Single-Stranded DNA Viruses in Vertebrate Genomes: The Parvoviridae and Circoviridae Are More than 40 to 50 Million Years Old', *Journal of Virology*, 84: 12458–62.

Bhatt, S. et al. (2013) 'The Global Distribution and Burden of Dengue', *Nature*, 496: 504–7.

Blair, C. D. (2011) 'Mosquito RNAi is the Major Innate Immune Pathway Controlling Arbovirus Infection and Transmission', *Future Microbiology*, 6: 265–77.

——, and Olson, K. E. (2015) 'The Role of RNA Interference (RNAi) in Arbovirus-Vector Interactions', *Viruses*, 7: 820–43.

Brackney, D. E., and Armstrong, P. M. (2016) 'Transmission and Evolution of Tick-Borne Viruses', *Current Opinion in Virology*, 21: 67–74.

Brennecke, J. et al. (2007) 'Discrete Small RNA-Generating Loci as Master Regulators of Transposon Activity in *Drosophila*', *Cell*, 128: 1089–103.

Bronkhorst, A. W., and van Rij, R. P. (2014) 'The Long and Short of Antiviral Defense: Small RNA-Based Immunity in Insects', *Current Opinion in Virology*, 7: 19–28.

Campbell, C. L. et al. (2008) '*Aedes aegypti* Uses RNA Interference in Defense Against Sindbis Virus Infection', *BMC Microbiology*, 8: 47.

Capella-Gutierrez, S., Silla-Martinez, J. M., and Gabaldon, T. (2009) 'trimAl: A Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses', *Bioinformatics*, 25: 1972–3.

Chandler, J. A., Liu, R. M., and Bennett, S. N. (2015) 'RNA Shotgun Metagenomic Sequencing of Northern California (USA) Mosquitoes Uncovers Viruses, Bacteria, and Fungi', *Frontiers in Microbiology*, 6: 185.

Chen, X. G. et al. (2015) 'Genome Sequence of the Asian Tiger Mosquito, *Aedes albopictus*, Reveals Insights into Its Biology, Genetics, and Evolution', *Proceedings of the National Academy of Sciences*, 112: E5907–15.

Cook, S. et al. (2013) 'Novel Virus Discovery and Genome Reconstruction from Field RNA Samples Reveals Highly Divergent Viruses in Dipteran Hosts', *PLoS One*, 8: e80720.

Crochu, S. et al. (2004) 'Sequences of Flavivirus-Related RNA Viruses Persist in DNA Form Integrated in the Genome of *Aedes* Spp. mosquitoes', *Journal of General Virology*, 85: 1971–80.

Crooks, G. E. et al. (2004) 'WebLogo: A Sequence Logo Generator', *Genome Research*, 14: 1188–90.

Depaquit, J. et al. (2010) 'Arthropod-Borne Viruses Transmitted by Phlebotomine Sandflies in Europe: A Review', *Eurosurveillance*, 15: 40–7.

Dudchenko, O. et al. (2017) '*De Novo* Assembly of the *Aedes aegypti* Genome Using Hi-C Yields Chromosome-Length Scaffolds', *Science*, 356: 92–5.

Foster, N. M., Jones, R. H., and McCrory, B. R. (1963) 'Preliminary Investigations on Insect Transmission of Bluetongue Virus in Sheep', *American Journal of Veterinary Research*, 24: 1195–200.

Gulia-Nuss, M. et al. (2016) 'Genomic Insights into the *Ixodes scapularis* Tick Vector of Lyme Disease', *Nature Communications*, 7: 10507.

Guo, X. X. et al. (2016) '*Culex pipiens quinquefasciatus*: A Potential Vector to Transmit Zika Virus', *Emerging Microbes & Infections*, 5: 1–5.

Hammon, W. M. et al. (1941) 'Isolation of the Viruses of Western Equine and St Louis Encephalitis from *Culex tarsalis* Mosquitoes', *Science*, 94: 328–30.

Hardy, J. L. et al. (1983) 'Intrinsic Factors Affecting Vector Competence of Mosquitos for Arboviruses', *Annual Review of Entomology*, 28: 229–62.

Holmes, E. C. (2011) 'The Evolution of Endogenous Viral Elements', *Cell Host & Microbe*, 10: 368–77.

Holt, R. A. et al. (2002) 'The Genome Sequence of the Malaria Mosquito *Anopheles gambiae*', *Science*, 298: 129–49.

Hoskins, R. A. et al. (2015) 'The Release 6 Reference Sequence of the *Drosophila melanogaster* Genome', *Genome Research*, 25: 445–58.

ICTV (2017) 'Changes to Taxonomy and the International Code of Virus Classification and Nomenclature Ratified by the International Committee on Taxonomy of Viruses', *Archives of Virology*, 162: 2505.

Jenkins, G. M. et al. (2002) 'Rates of Molecular Evolution in RNA Viruses: A Quantitative Phylogenetic Analysis', *Journal of Molecular Evolution*, 54: 156–65.

Jones, K. E. et al. (2008) 'Global Trends in Emerging Infectious Diseases', *Nature*, 451: 990–3.

Junglen, S. (2016) 'Evolutionary Origin of Pathogenic Arthropod-Borne Viruses - A Case Study in the Family *Bunyaviridae*', *Current Opinion in Insect Science*, 16: 81–6.

Katoh, K. et al. (2002) 'MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform', *Nucleic Acids Research*, 30: 3059–66.

Katzourakis, A., and Gifford, R. J. (2010) 'Endogenous Viral Elements in Animal Genomes', *PLoS Genetics*, 6: e1001191.

Kearse, M. et al. (2012) 'Geneious Basic: An Integrated and Extendable Desktop Software Platform for the Organization and Analysis of Sequence Data', *Bioinformatics*, 28: 1647–9.

Keene, K. M. et al. (2004) 'RNA Interference Acts as a Natural Antiviral Response to O'nyong-Nyong Virus (Alphavirus; *Togaviridae*) Infection of *Anopheles gambiae*', *Proceedings of the National Academy of Sciences of the United States of America*, 101: 17240–5.

Kim, D., Langmead, B., and Salzberg, S. L. (2015) 'HISAT: A Fast Spliced Aligner with Low Memory Requirements', *Nature Methods*, 12: 357–60.

Koonin, E. V., and Krupovic, M. (2018) 'The Depths of Virus Exaptation', *Current Opinion in Virology*, 31: 1–8.

Kumar, S. et al. (2017) 'TimeTree: A Resource for Timelines, Timetrees, and Divergence Times', *Molecular Biology and Evolution*, 34: 1812–9.

LaBeaud, A. D., Bashir, F., and King, C. H. (2011) 'Measuring the Burden of Arboviral Diseases: The Spectrum of Morbidity and Mortality from Four Prevalent Infections', *Population Health Metrics*, 9: 1.

Labuda, M., and Nuttall, P. A. (2004) 'Tick-Borne Viruses', *Parasitology*, 129: S221–S45.

Larkin, M. A. et al. (2007) 'Clustal W and Clustal X Version 2.0', *Bioinformatics (Oxford, England)*, 23: 2947–8.

Lasecka, L., and Baron, M. D. (2014) 'The Molecular Biology of Nairoviruses, an Emerging Group of Tick-Borne Arboviruses', *Archives of Virology*, 159: 1249–65.

Lewis, S. H. et al. (2018) 'Pan-Arthropod Analysis Reveals Somatic piRNAs as an Ancestral Defence against Transposable Elements', *Nature Ecology & Evolution*, 2: 174–81.

Li, C. X. et al. (2015) 'Unprecedented Genomic Diversity of RNA Viruses in Arthropods Reveals the Ancestry of Negative-Sense RNA Viruses', *eLife*, 4: e05378.

Love, M. I., Huber, W., and Anders, S. (2014) 'Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2', *Genome Biology*, 15: 550.

Mansfield, K. L. et al. (2017) 'Emerging Tick-Borne Viruses in the Twenty-First Century', *Frontiers in Cellular and Infection Microbiology*, 7: 298.

Marklewitz, M. et al. (2015) 'Evolutionary and Phenotypic Analysis of Live Virus Isolates Suggests Arthropod Origin of a Pathogenic RNA Virus Family', *Proceedings of the National Academy of Sciences*, 112: 7536–41.

Matthews, B. J. et al. (2018) 'Improved Reference Genome of *Aedes aegypti* Informs Arbovirus Vector Control', *Nature*, 563: 501–7.

Mellor, P. S. (2000) 'Replication of Arboviruses in Insect Vectors', *Journal of Comparative Pathology*, 123: 231–47.

——, Boorman, J., and Baylis, M. (2000) '*Culicoides* Biting Midges: Their Role as Arbovirus Vectors', *Annual Review of Entomology*, 45: 307–40.

Miesen, P., Girardi, E., and van Rij, R. P. (2015) 'Distinct Sets of PIWI Proteins Produce Arbovirus and Transposon-Derived piRNAs in *Aedes aegypti* Mosquito Cells', *Nucleic Acids Research*, 43: 6545–56.

Monath, T. P., and Tsai, T. F. (1987) 'St-Louis Encephalitis - Lessons from the Last Decade', *American Journal of Tropical Medicine and Hygiene*, 37: S40–59.

Morales-Hojas, R. et al. (2018) 'The Genome of the Biting Midge *Culicoides sonorensis* and Gene Expression Analyses of Vector Competence for Bluetongue Virus', *BMC Genomics*, 19: 624.

Morazzani, E. M. et al. (2012) 'Production of Virus-Derived Ping-Pong-Dependent piRNA-like Small RNAs in the Mosquito Soma', *PLoS Pathogens*, 8: e1002470.

Myles, K. M. et al. (2008) 'Alphavirus-Derived Small RNAs Modulate Pathogenesis in Disease Vector Mosquitoes', *Proceedings of the National Academy of Sciences of the United States of America*, 105: 19938–43.

Nene, V. et al. (2007) 'Genome Sequence of *Aedes aegypti*, a Major Arbovirus Vector', *Science (New York, N.Y.)*, 316: 1718–23.

Palatini, U. et al. (2017) 'Comparative Genomics Shows That Viral Integrations Are Abundant and Express piRNAs in the Arboviral Vectors *Aedes aegypti* and *Aedes albopictus*', *BMC Genomics*, 18: 512.

Patro, R. et al. (2017) 'Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression', *Nature Methods*, 14: 417–9.

Peccoud, J. et al. (2017) 'Massive Horizontal Transfer of Transposable Elements in Insects', *Proceedings of the National Academy of Sciences*, 114: 4721–6.

Presti, R. M. et al. (2009) 'Quaranfil, Johnston Atoll, and Lake Chad Viruses Are Novel Members of the Family *Orthomyxoviridae*', *Journal of Virology*, 83: 11599–606.

Roiz, D. et al. (2009) 'Detection of Novel Insect Flavivirus Sequences Integrated in *Aedes albopictus* (Diptera: Culicidae) in Northern Italy', *Virology Journal*, 6: 93.

Rosenkranz, D., and Zischler, H. (2012) 'proTRAC - a Software for Probabilistic piRNA Cluster Detection, Visualization and Analysis', *BMC Bioinformatics*, 13: 5.

Ruder, M. G. et al. (2015) 'Transmission and Epidemiology of Bluetongue and Epizootic Hemorrhagic Disease in North America: Current Perspectives, Research Gaps, and Future Directions', *Vector-Borne and Zoonotic Diseases*, 15: 348–63.

Ruigrok, R. W. H. et al. (2010) 'Towards an Atomic Resolution Understanding of the Influenza Virus Replication Machinery', *Current Opinion in Structural Biology*, 20: 104–13.

Sanchez-Vargas, I. et al. (2009) 'Dengue Virus Type 2 Infections of *Aedes aegypti* Are Modulated by the Mosquito's RNA Interference Pathway', *PLoS Pathogens*, 5: e1000299.

Schnettler, E. et al. (2014) 'Induction and Suppression of Tick Cell Antiviral RNAi Responses by Tick-Borne Flaviviruses', *Nucleic Acids Research*, 42: 9436–46.

Shi, M. et al. (2016) 'Redefining the Invertebrate RNA Virosphere', *Nature*, 540: 539–43.

Siomi, M. C. et al. (2011) 'PIWI-Interacting Small RNAs: The Vanguard of Genome Defence', *Nature Reviews. Molecular Cell Biology*, 12: 246–58.

Smit, A. F. A., Hubley, R., and Green, P. (2018) *RepeatMasker Open-4.0.8* <http://repeatmasker.org> accessed 11 Feb 2019.

Stamatakis, A. (2014) 'RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies', *Bioinformatics (Oxford, England)*, 30: 1312–3.

Sudia, W. D. et al. (1967) 'St Louis Encephalitis Vector Studies in Houston Texas 1964', *Journal of Medical Entomology*, 4: 32–6.

Ter Horst, A. M. et al. (2018) 'Endogenous Viral Elements are Widespread in Arthropod Genomes and Commonly Give Rise to piRNAs', *Journal of Virology*, 93: e02124–18.

Turell, M. J. et al. (2005) 'An Update on the Potential of North American Mosquitoes (Diptera: Culicidae) to Transmit West Nile Virus', *Journal of Medical Entomology*, 42: 57–62.

Vijayendran, D. et al. (2013) 'Arthropod Viruses and Small RNAs', *Journal of Invertebrate Pathology*, 114: 186–95.

Vodovar, N. et al. (2012) 'Arbovirus-Derived piRNAs Exhibit a Ping-Pong Signature in Mosquito Cells', *PLoS One*, 7: e30861.

Weisheit, S. et al. (2015) '*Ixodes scapularis* and *Ixodes ricinus* Tick Cell Lines Respond to Infection with Tick-Borne Encephalitis Virus: Transcriptomic and Proteomic Analysis', *Parasites & Vectors*, 8: 599.

Weiss, B., and Aksoy, S. (2011) 'Microbiome Influences on Insect Host Vector Competence', *Trends in Parasitology*, 27: 514–22.

Whitfield, Z. J. et al. (2017) 'The Diversity, Structure, and Function of Heritable Adaptive Immunity Sequences in the *Aedes aegypti* Genome', *Current Biology*, 27: 3511–9.

WHO (2014) 'A Global Brief on Vector-Borne Diseases' (Technical Report, World Health Organization).

Zhang, Y. Z. et al. (2012) 'The Ecology, Genetic Diversity, and Phylogeny of Huaiyangshan Virus in China', *Journal of Virology*, 86: 2864–8.

Ziebuhr, J. (2006) 'The Coronavirus Replicase: Insights into a Sophisticated Enzyme Machinery', *Advances in Experimental Medicine and Biology*, 581: 3–11.