The Widespread Occurrence and Potential Biological Roles of Endogenous Viral Elements in Insect Genomes

Carol D. Blair¹, Ken E. Olson¹ and Mariangela Bonizzoni²*

¹Arthropod-borne and Infectious Diseases Laboratory, Department of Microbiology, Immunology and Pathology, Colorado State University, Fort Collins, CO, USA.
²Department of Biology and Biotechnology, University of Pavia, Pavia, Italy.
*Correspondence: mariangela.bonizzoni@unipv.it

https://doi.org/10.21775/cimb.034.013

Abstract
Modern genomic sequencing and bioinformatics approaches have detected numerous examples of DNA sequences derived from DNA and RNA virus genomes integrated into both vertebrate and insect genomes. Retroviruses encode RNA-dependent DNA polymerases (reverse transcriptases) and integrases that convert their RNA viral genomes into DNA proviruses and facilitate proviral DNA integration into the host genome. Surprisingly, DNA sequences derived from RNA viruses that do not encode these enzymes also occur in host genomes. Non-retroviral integrated RNA virus sequences (NIRVS) occur at relatively high frequency in the genomes of the arboviral vectors Aedes aegypti and Aedes albopictus, are not distributed randomly and possibly contribute to mosquito antiviral immunity, suggesting these mosquitoes could serve as a model system for unravelling the function of NIRVS. Here we address the following questions: What drives DNA synthesis from the genomes of non-retroviral RNA viruses? How does integration of virus cDNA into host DNA occur, and what is its biological function (if any)? We review current knowledge of viral integrations in insect genomes, hypothesize mechanisms of NIRVS formation and their potential impact on insect biology, particularly antiviral immunity, and suggest directions for future research.

Introduction: definition of endogenous viral elements and the surprising discovery of those from non-retroviral RNA viruses
Interactions between viruses and their hosts occur at many levels. Viruses evolve rapidly through genetic drift and/or natural selection and can influence the genetic structures of host populations, especially if viral infection results in disease and mortality. A classic example of virus and host co-evolution is the natural experiment initiated by the release of highly pathogenic myxoma virus into the naïve Australian wild rabbit population in 1950, followed by rapid development of resistance in rabbits coincident with natural attenuation of the virus (Best and Kerr, 2000; Kerr, 2012). Another mechanism through which viruses affect the genetics of the host is viral genome integration. Integrations of viral nucleic acid sequences can occur in both somatic and germline cells. Somatic integrations are frequently associated with host genome instability (e.g. Chen et al., 2014).
Additionally, complete viral sequence integration into host genomes can favour the persistence of the infection (Cohn et al., 2015). If integrations occur in germline cells, they can be inherited by the next generation. The persistence and outcome of these integrations in host populations depend on their effects on host fitness. If deleterious, integrations are likely lost. Alternatively, viral sequences can be functionally adopted by the host and exert beneficial functions (Frank and Feschotte, 2017). For instance, the product of the murine retrovirus restriction gene, *Fv1*, protects mice against infection with murine leukaemia virus (MLV). *Fv1* genes in *Mus musculus* and other *Mus* spp. were derived from the *gag* genes of ancient endogenous retroviruses, some of which are distantly related to MLV, and protect against other retroviruses, including lentiviruses and spumaviruses (Yap et al., 2014). A third outcome is found when integrations are in chromosomal locations that are not transcribed or lack regulatory regions. In this case, viral integrations may persist but accumulate mutations at the host rate, which is typically much slower than that of exogenous viruses. These endogenous viral elements (EVEs) thus represent a magnifying lens into the past, allowing researchers to more closely examine the evolutionary history of viruses and virus–host co-evolution (Aiiewsakun and Kat sourakis, 2015; Katzourakis, 2017).

Integration of genome sequences from DNA viruses and retroviruses is a relatively common phenomenon, as evidenced by the abundance of virus-derived sequences in the genomes of various organisms. For instance, genetic code from retroviruses constitutes about 8% of the human genome (Griffiths et al., 2001). Genome integrations are essential events in the retrovirus life cycle. Because non-retroviral RNA viruses lack coding for reverse transcriptase, the machinery needed for successful conversion to DNA and integration into host genomes, their potential to endogenize has been considered minimal. However, an increasing number of studies show genome integrations into both somatic and germline host cells from non-retroviral RNA viruses, including both single-stranded (positive and negative) and double-stranded RNA viruses. The acronym non-retroviral integrated RNA virus sequences (NIRVS) has been proposed to emphasize the unusual viral origin of these EVEs (Ballinger et al., 2012; Kondo et al., 2013; Ballinger et al., 2014; Olson and Bonizzoni, 2017; Palatini et al., 2017).

Here we describe current knowledge on the origin, widespread occurrence and genomic context of EVEs in insect genomes. Additionally, we report examples of the influence of EVEs on both vertebrate and insect host physiology and immunity.

**Integration of genomes from DNA viruses**

Polydnaviruses (PDV) provide a well-known example of DNA virus integration into insect genomes (Table 2.1A; for a complete, authoritative review of polydnaviruses and their interactions with their insect hosts, see Chapter 8). PDV are recognized by the International Committee on Taxonomy of Viruses (ICTV; https://talk.ictvonline.org/ictv-reports/ictv_9th_report/dsdna-viruses-2011/w/dsdna_viruses/127/polydnarviidae) as the family *Polydnarviidae*, divided into the genera *Bracovirus* (BV) and *Ichnovirus* (IV), which are associated with braconid and ichneumonid parasitoid wasps, respectively (Strand and Burke, 2015; Gauthier et al., 2018). Bracoviruses have been domesticated by wasps, and their genomes are integrated as proviral segments dispersed throughout the genome in every cell of both female and male wasps (Desjardins et al., 2008). They replicate by producing defective virus-like particles (VLPs) in the calyx of the wasp ovary. Bracovirus genomes in VLPs consist of multiple, diverse, circular dsDNAs depleted of replication genes, with one DNA molecule per VLP. Parasitoid wasps reproduce by depositing their eggs in or on other insects, which are consumed by juvenile stages of the wasp as they develop (Burke et al., 2014). The defective VLPs are deposited into the wasp’s hemipteran or dipteran host along with its eggs, their defective genomes migrate to the host cell nucleus, where they subsequently express virulence factors that alter host physiology and allow the development of wasp progeny (Strand and Burke, 2015; Gauthier et al., 2018). PDVs cannot be transmitted horizontally; rather, the integrated proviruses are transmitted vertically in wasps. Transcript studies suggest that BVs evolved from a single whole genome integration of an ancient ancestor of the insect virus family *Nudiviridae* (Bezier et al., 2009; Burke and Strand, 2012). Lineages of ichneumonid wasps with integrated IVs are polyphyletic.
and integrated IVs are diverse, suggesting multiple evolutionary integration events (Gauthier et al., 2018).

Analyses of a recently discovered endogenization involving an alphanudivirus and the braconid wasp Fopius arisanus are shedding light on the mechanisms of integration. Alphanudiviruses are members of the family Nudiviridae with a single circular dsDNA genome that usually cause virulent infections in insects. Fopius arisanus belongs to the Opinae subfamily of Braconidae. Wasp species in this subfamily usually lack integrations from bracoviruses (Burke et al., 2018). Consequently, the identification of integrated alphanudivirus sequences in the genome of F. arisanus suggests that this integration is a recent event. This event shows that endogenization is accompanied with gene-loss in the viral genome, and that changes in its architecture occur before expression of the endogenized virus genes falls under the control of the wasp genome (Burke et al., 2018).

Genome endogenization has also been observed for the betanudivirus HzNV-1, which naturally infects the corn earworm Helicoverpa zea. HzNV-1 can exhibit a latent phase of infection during which its DNA is integrated into the host genome (Gauthier et al., 2018). Moreover, HzNV-1 genome integrations were observed in persistently infected cells of other lepidopteran species (Trichoplusia ni and Spodoptera frugiperda). Interestingly, cells of the fall armyworm carrying endogenous HzNV-1 appeared resistant to superinfections (Lin et al., 1999).

Integration of non-retroviral RNA virus genomes: discovery and association with retroelements

The first discovery of integrations of non-retroviral RNA virus genome sequences in insect genomes was the unexpected finding in 2004 of four regions similar to RNAs of cell fusing agent (CFAV) and Kamiti River (KRV) viruses in the genomes of Aedes aegypti and Aedes albopictus. These sequences, which were amplified using degenerate flavivirus primers, were called cell silent agents (CSA) because of their similarity to CFAV (Crochu et al., 2004). Both CFAV and KRV are insect-specific viruses (ISVs) that are genetically related to arboviruses but exclusively infect insect cells. In 2009, additional unexpected integrations from ISVs were detected in wild-caught Aedes and Ochlerotatus genomes by PCR using flavivirus-specific primers (Roiz et al., 2009; Rizzo et al., 2009; Sanchez-Seco et al., 2010).

Descriptions of RNA virus genome sequence integrations into insect genomes increased with the advent of next-generation sequencing technologies and metagenomic analyses in the early 2000s. Bioinformatics-based identification of viral integrations uses BLAST+ in the form of BLASTx or tBLASTn (Altschul et al., 1990). Researchers parse positive BLAST hits using cut-off values of between $10^{-3}$ and $10^{-6}$ (Kondo et al., 2019; Whitfield et al., 2017; ter Horst et al., 2018). Additional criteria (i.e. presence of viral open reading frames or minimum size restrictions) are applied to further reduce false positive results that occur primarily due to low-complexity reads with short tandem repeats or homopolymers. The complexity of the viral database used to search for viral integrations may influence the number of identified BLAST hits. Researchers have used a number of virus genome databases to identify viral integrations in host genomes. These viral sequence databases include sequences from single virus families (Fort et al., 2012; Chen et al., 2015), from plant viruses transmitted by insects (Cui and Holmes, 2012), all known RNA viruses, including those most-recently identified (i.e. Whitfield et al., 2017; Palatini et al., 2017) and from a mixture of DNA and RNA viruses (Katzourakis and Gifford, 2010; ter Horst et al., 2018). These analyses provided the number of viral integrations in sequenced insect genomes and insight into the diversity and richness of their viral origins. The results were influenced by the breadth of viral genomes analysed and the stringency of the criteria used to define an insect sequence as a viral integration (i.e. cut off BLAST e-value; reverse BLAST; analyses of low complexity sequences; length of sequences; inclusion of newly discovered and unclassified viruses).

Currently-known insect genome viral integrations are largely from non-retroviral RNA viruses with sequences also derived from single-stranded DNA viruses, e.g. from densoviruses (Parvoviridae) in the genome of Acyrthosiphon pisum (ter Horst et al., 2018). NIRVS have been characterized from the genomes of both haematophagous and non-bloodsucking insects (Table 2.1B and C). A review
<table>
<thead>
<tr>
<th>References</th>
<th>Expression Discovery Method</th>
<th>Genome with Host</th>
<th>N</th>
<th>Host</th>
<th>Family</th>
<th>Genus, Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>(2012) Chu and Homes</td>
<td>in silico screening: PCR</td>
<td>C6/36</td>
<td>2</td>
<td>Plant RNA viruses</td>
<td>Cu. gumba族ricus</td>
<td>Cu. gumba族ricus with mosquitos</td>
</tr>
<tr>
<td>(2012) Fort et al.</td>
<td>in silico screening: PCR</td>
<td>Y</td>
<td>112 in Ae. aegypti, 1 in ND</td>
<td>Rhoddo. R</td>
<td>Cu. gumba族ricus</td>
<td>Ae. aegypti, Rhoddo</td>
</tr>
<tr>
<td>(2012) Vassquez et al.</td>
<td>PCR</td>
<td>ND</td>
<td>c</td>
<td>ND</td>
<td>Havana</td>
<td>Havana</td>
</tr>
<tr>
<td>(2009) Ruiz et al.</td>
<td>PCR</td>
<td>ND</td>
<td>143</td>
<td>Havana</td>
<td>Havana</td>
<td>Havana</td>
</tr>
<tr>
<td>(2009) Sanchez-Seco et al.</td>
<td>PCR</td>
<td>ND</td>
<td>5</td>
<td>Havana</td>
<td>Havana</td>
<td>Havana</td>
</tr>
<tr>
<td>(2009) Ruiz et al.</td>
<td>PCR</td>
<td>ND</td>
<td>1</td>
<td>Havana</td>
<td>Havana</td>
<td>Havana</td>
</tr>
<tr>
<td>(2009) Crocchi et al.</td>
<td>DNA Repetitions</td>
<td>Y</td>
<td>6</td>
<td>Havana</td>
<td>Havana</td>
<td>Havana</td>
</tr>
</tbody>
</table>

### Table 2.1

Characteristics of EVEs that have been identified in insect genomes including parasitic wasps (A), hemipterans (B), and non-hemipterans (C) and non-bloodsucking wasps.
<table>
<thead>
<tr>
<th>Reference</th>
<th>Method</th>
<th>Species</th>
<th>Type</th>
<th>Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ballinger et al. (2012)</td>
<td>In silico screening</td>
<td>Drosophilidae</td>
<td>Viruses</td>
<td>Plant</td>
</tr>
<tr>
<td>Ballinger et al. (2012)</td>
<td>In silico screening</td>
<td>Drosophilidae</td>
<td>Viruses</td>
<td>Plant</td>
</tr>
<tr>
<td>Cui et al. (2012)</td>
<td>In silico screening</td>
<td>Drosophilidae</td>
<td>Viruses</td>
<td>Plant</td>
</tr>
<tr>
<td>Fort et al. (2012)</td>
<td>In silico screening</td>
<td>Drosophilidae</td>
<td>Viruses</td>
<td>Plant</td>
</tr>
<tr>
<td>Fort et al. (2012)</td>
<td>In silico screening</td>
<td>Drosophilidae</td>
<td>Viruses</td>
<td>Plant</td>
</tr>
<tr>
<td>16 (2012)</td>
<td>PCR</td>
<td>Drosophilidae</td>
<td>Repetitive</td>
<td>1</td>
</tr>
<tr>
<td>16 (2012)</td>
<td>PCR</td>
<td>Drosophilidae</td>
<td>Repetitive</td>
<td>1</td>
</tr>
<tr>
<td>16 (2012)</td>
<td>PCR</td>
<td>Drosophilidae</td>
<td>Repetitive</td>
<td>1</td>
</tr>
<tr>
<td>Wu et al. (2017)</td>
<td>In silico screening</td>
<td>Culicidae</td>
<td>DNA</td>
<td>Repetitive</td>
</tr>
<tr>
<td>Wu et al. (2017)</td>
<td>In silico screening</td>
<td>Culicidae</td>
<td>DNA</td>
<td>Repetitive</td>
</tr>
<tr>
<td>Wu et al. (2017)</td>
<td>In silico screening</td>
<td>Culicidae</td>
<td>DNA</td>
<td>Repetitive</td>
</tr>
<tr>
<td>Partiali et al. (2017)</td>
<td>In silico screening</td>
<td>Culicidae</td>
<td>DNA</td>
<td>Repetitive</td>
</tr>
<tr>
<td>Suzuki et al. (2017)</td>
<td>In silico screening</td>
<td>Culicidae</td>
<td>DNA</td>
<td>Repetitive</td>
</tr>
<tr>
<td>Lamertins et al. (2017)</td>
<td>In silico screening</td>
<td>Culicidae</td>
<td>DNA</td>
<td>Repetitive</td>
</tr>
<tr>
<td>Li et al. (2019)</td>
<td>In silico screening</td>
<td>Culicidae</td>
<td>DNA</td>
<td>Repetitive</td>
</tr>
<tr>
<td>Kondo et al. (2019)</td>
<td>In silico screening</td>
<td>Culicidae</td>
<td>DNA</td>
<td>Repetitive</td>
</tr>
<tr>
<td>Virus Family</td>
<td>Genus</td>
<td>Species</td>
<td>Host</td>
<td>Context</td>
</tr>
<tr>
<td>-------------</td>
<td>-------</td>
<td>---------</td>
<td>------</td>
<td>---------</td>
</tr>
<tr>
<td>Spodoptera higgedea</td>
<td>Philodromus-like</td>
<td>R</td>
<td>Chui, Mononen</td>
<td>12</td>
</tr>
<tr>
<td>Philodromus-like</td>
<td>Quasars</td>
<td>R</td>
<td>Chui, Mononen</td>
<td>23</td>
</tr>
<tr>
<td>Leiodopera dorsipalpus</td>
<td>Gymnosoma braunus</td>
<td>R</td>
<td>Chui, Mononen</td>
<td>4-201</td>
</tr>
<tr>
<td>Gymnosoma braunus</td>
<td>Leiodopera dorsipalpus</td>
<td>R</td>
<td>Chui, Mononen</td>
<td>3</td>
</tr>
<tr>
<td>Gymnosoma braunus</td>
<td>Acromyrmex echinataris</td>
<td>R</td>
<td>Chui, Mononen</td>
<td>8</td>
</tr>
<tr>
<td>Hemimetopera</td>
<td>Alle cephaletes</td>
<td>R</td>
<td>Chui, Mononen</td>
<td>9</td>
</tr>
<tr>
<td>Hemimetopera</td>
<td>Gymnosoma braunus</td>
<td>R</td>
<td>Chui, Mononen</td>
<td>4-201</td>
</tr>
<tr>
<td>Zootermopsis nevadensis</td>
<td>Colopera</td>
<td>R</td>
<td>Chui, Burya</td>
<td>3</td>
</tr>
<tr>
<td>Drosophilidae, sp.</td>
<td>Thriophilus saeumann</td>
<td>R</td>
<td>Chui, Burya</td>
<td>3</td>
</tr>
</tbody>
</table>

**Table 2.1 Continued**


**Reoviridae**

- Reoviridae viruses are typically double-stranded RNA viruses. They are negative-strand RNA viruses that undergo RNA-dependent RNA polymerase replication. Reoviridae viruses are known for their complex genome structure, consisting of 10 segments.

**Flavi**

- Flavi viruses are a group of single-stranded positive-sense RNA viruses. They are found in a variety of hosts, including humans, and are known for their role in disease. Flavi viruses include viruses such as hepatitis C virus (HCV) and yellow fever virus (YFV).

**Peribunyaviridae**

- Peribunyaviridae viruses are a group of positive-sense单-stranded RNA viruses. They are known for their role in causing infections in a variety of hosts, including humans. Peribunyaviridae viruses include viruses such as Rift Valley fever virus (RVFV) and California serogroup virus (CSMV).

**D**

- The D genome is one of the 10 segments that make up the Reoviridae virus genome. It contains the viral polymerase gene and is essential for viral replication.

**Integration**

- Integration refers to the process by which the viral genome is incorporated into the host genome. This can occur through the action of viral integrases or other enzymes.

**Genomic viruses**

- Genomic viruses are viruses that replicate within the host cell's nucleus. They include viruses such as retroviruses and DNA viruses. Genomic viruses are known for their replication strategies, which often involve reverse transcription or transcription from a DNA intermediate.

**B.**

- B. refers to the genus of viruses within the Reoviridae family. This genus includes viruses such as reovirus and orbivirus.

**N.**

- N. refers to the genus of viruses within the Peribunyaviridae family. This genus includes viruses such as bunyavirus and Rift Valley fever virus.

**Nasonia vitripennis**

- Nasonia vitripennis is a species of parasitoid wasp that has been studied extensively for its role in viral research. It has a complex life cycle that involves the transmission of viruses to its host, Drosophila melanogaster.

**X-linked priming**

- X-linked priming is a mechanism by which the viral genome is replicated in the female lineage. It involves the use of the X chromosome as a template for viral genome replication.

**Various viruses**

- Various viruses are viruses that are not yet well-understood or classified. They include viruses such as the mosquito-borne viruses and the insect-borne viruses. These viruses are known for their role in causing disease in a variety of hosts, including humans.

**RNA**

- RNA is a type of nucleic acid that is used by viruses for replication. It is typically single-stranded and is transcribed from DNA or other RNA molecules.

**DNA**

- DNA is a type of nucleic acid that is used by viruses for replication. It is typically double-stranded and is transcribed from RNA or other DNA molecules.

**NIRVS**

- NIRVS refers to the National Institute of Research on Viral and Rickettsial Diseases. It is a research institute that focuses on the study of viral and rickettsial diseases.

**N.**

- N. refers to the genus of viruses within the Peribunyaviridae family. This genus includes viruses such as bunyavirus and Rift Valley fever virus. 
of the widespread occurrence of NIRVS in insect genomes suggests similar mechanisms of biogenesis and integration occurred across different insect species; however, striking differences were found in the number of NIRVS and the viral origins of those in Aedes mosquitoes relative to other insects.

In contrast to endogenization of polydnavirus sequences originating from entire rearranged viral genomes, NIRVS consist of complete or interrupted viral RNA open-reading frames (ORFs). Most NIRVS have similarity to coding regions of non-retroviral nucleoproteins, glycoproteins, and/or RNA-dependent-RNA polymerases (Crochu et al., 2004; Katzourakis and Gifford, 2010; Fort et al., 2012; Ballinger et al., 2012; Cui and Holmes, 2012; Kondo et al., 2013; Ballinger et al., 2014; Geisler and Jarvis, 2016; Lequime and Lambrechts, 2017; Suzuki et al., 2017; Palatini et al., 2017; Whitfield et al., 2017).

A striking similarity among NIRVS detected across different insect species and from different virus families is that integrations are frequently found in regions of repetitive DNA, in association with long terminal repeat (LTR) retroelements. Retrotransposons of the Pao/Ninja, BEL/Pao, Copia and Ty3/Gypsy lineages and, less frequently, Tc1/mariner, flank NIRVS derived from flaviviruses and rhabdoviruses in both Culicidae and Drosophilidae (Crochu et al., 2004; Ballinger et al., 2012; Goic et al., 2013; Lequime and Lambrechts, 2017; Palatini et al., 2017; Whitfield et al., 2017). Retroelements also flank NIRVS from plant RNA viruses and newly-characterized alphanavirus-like ISVs in examples from Hemiptera, Hymenoptera, Thysanoptera, Homoptera, Lepidoptera, and from Drosophilidae (Cui and Holmes, 2012; Kondo et al., 2013, 2019).

The origin of NIRVS in insect genomes requires the formation of cDNA from the non-retroviral RNA template with the help of reverse transcriptases. Since NIRVS come from viruses that lack this enzyme, retrotransposons are a likely source within insect cells. Naturally occurring cDNAs of non-retroviral RNA sequences, vDNAs, have been found in cells infected with these viruses (Goic et al., 2013; Nag and Kramer, 2017). A reduction in formation of Chikungunya virus (CHIKV; Alphavirus) vDNA in Ae. albopictus treated with a reverse transcriptase inhibitor has been observed (Goic et al., 2016). This finding and the observed physical contiguity between NIRVS and retroelements in genomes of insects has led to the hypothesis that integration events in insect genomes occur through non-homologous recombination between the genomes of infecting non-retroviral RNA viruses and LTR retroelements during their reverse transcription (Ballinger et al., 2012; Goic et al., 2013). The hypothesis is supported by similar findings in mammalian genomes (Geuking et al., 2009) and by the fact that LTR retroelement reverse transcription occurs in the cytoplasm, where most non-retroviral RNA viruses complete their life-cycle (Havecker et al., 2004; Geuking et al., 2009; Horie et al., 2010; Ballinger et al., 2012). Furthermore, LTR retroelements are active in germline cells, which could favour inheritance of newly acquired viral sequences (Morozov et al., 2017).

The association between NIRVS and retroelements is also supported by the recent finding that some retroelements harbour viral-like domains (Ballinger et al., 2012). Viral-like helicase domains of recently characterized ISVs were found as part of TRAS and Jockey non-LTR retrotroposon proteins in the genomes of Lepidoptera (i.e. Plutella xylostella), Hemiptera (i.e. Nilaparvata lugens, Homalodisca vitripennis) Orthoptera (species of the Ceuthorhithus genus), Hymenoptera (Leptopilina boulardi) and Culicidae (Ae. aegypti) species (Lazareva et al., 2015; Shi et al., 2016; Morozov et al., 2017). Interestingly, expression of viral-insect fused proteins in some cases increased the evolutionary fitness of the harbouring retroelements by functioning as viral suppressors of RNA silencing (Lazareva et al., 2015). Helicases containing a viral motif could stabilize retroposon transcripts and increase their transposition efficiency. An excess of stabilizing helicase activity could also negatively impact non-retroviral RNA virus infection by altering the balance between virus transcription and replication (Morozov et al., 2017).

Integration of non-retroviral RNA virus genomes: viral origin and number of NIRVS in insect genomes

Because most current discoveries of NIRVS come from in silico studies of sequenced insect genomes with limited data from natural specimens,
knowledge of the widespread occurrence of NIRVS in insect genomes is biased by comparisons across species as compared to among populations within species. With this limitation in mind, ISVs and other viruses known to establish persistent infections are well represented in the list of NIRVS (Table 2.1B).

Currently known insect NIRVS are similar to the sequences of viruses from the Flavivirus, Beny-virus, and Quaranjavirus genera; the Reoviridae, Rhabdoviridae and Peribunyaviridae families and the recently characterized family Chuviridae and phlebovirus-like and virga/nege-like viruses. Chuviridae is a newly identified monophyletic family of negative-sense RNA viruses, with diverse genome structures (i.e. unsegmented, bi-segmented and circular forms) (Li et al., 2015). Quaranjavirus and phlebovirus-like viruses have negative-sense, segmented RNA genomes, belong to the Orthomyxoviridae and Phenuviridae families, respectively, and have been recently identified in arthropods (Li et al., 2015). ‘Virga/nege-like viruses’ refers to a still poorly characterized group of positive-strand RNA viruses that appear to belong to the alphavirus-like superfamily (Kondo et al., 2019). Virgaviridae is a family of plant alpha-like viruses, while negevirus refers to newly recognized ISVs from sandflies and mosquitoes, which are distinctly phylogenetically-related to plant cileviruses (Vasilakis et al., 2013; Carapeta et al., 2015; Nunes et al., 2017).

NIRVS from flaviviruses (F-NIRVS) have been found almost exclusively in the genomes of culicine mosquitoes (Table 2.1). The Culicidae is a large mosquito family that includes arboviral vectors such as Aedes spp., Ochlerotatus spp. and Culex spp. and protozoan vectors such as those in the subfamily Anophelinae (Chen et al., 2015). F-NIRVS are found in the genomes of medically important arboviral vectors (Ae. aegypti and Ae. albopictus), in the genomes of the minor arboviral vectors Ochlerotatus spp. (i.e. Oc. caspius and Oc. detritus) and A. vexans, but not in the genome of Cx. quinquefasciatus (Vasquez et al., 2012, Palatini et al., 2017; Whitfield et al., 2017; Blagrove et al., 2017). Among the 19 species of Anophelinae for which a genome sequence is available, F-NIRVS were identified only in Anopheles minimus and Anopheles sinensis (Palatini et al., 2017; Lequime and Lambrechts, 2017). Interestingly, the number of F-NIRVS in the Ae. aegypti and Ae. albopictus genomes is thirty times greater than that detected in any other culicine genome. Although the precise number of F-NIRVS will vary with improvements in genome annotations and with sequencing of novel culicine genomes, the large number of F-NIRVS in these two mosquito species was obtained by applying the same conservative bioinformatics pipeline across the genomes of Aedes, Culex and Anopheles spp., and hence the analysis was methodologically unbiased.

Current data, which come from studies using different bioinformatics approaches and analysing different insect species, show that NIRVS from reovirus, bunyavirus, sigmavirus, benyvirus, orthomyxovirus and phlebovirus genomes are sporadic, as they are found in both limited numbers and distribution across insect species (Table 2.1). Even with the paucity of data currently available, it is interesting to note that haematophagous species like Ae. aegypti and the kissing bug Rhodnius prolixus harbour not only arbovirus-derived NIRVS, but also NIRVS with similarity to the genomes of plant RNA viruses (Cui and Holmes, 2012; Kondo et al., 2013). Non-blood-sucking insect species harbour NIRVS primarily from Rhabdoviridae and Chuviridae, which are also in the genomes of haematophagous insects (Table 2.1). The Rhabdoviridae family includes viruses that infect vertebrates, invertebrates and plants and that are extremely variable in their genomic organization (Dietzgen et al., 2017; Geoghegan et al., 2017). The ecological diversity of rhabdoviruses is also reflected in their frequent cross-species transmission between hosts (Geoghegan et al., 2017). While the ecology of Chuviridae is not completely understood, viruses in this family were found in arthropod species that share the same ecological niche, indicative of possible horizontal transfer (Li et al., 2015). Integrations from Rhabdoviridae and Chuviridae may appear to be more frequent than NIRVS from other viruses simply because these viruses are ecologically diverse and have wide geographical distributions. Alternatively, the infection capabilities of these viruses and their capacity to transmit between host species, genera, and even host kingdoms in the case of rhabdoviruses, depending on their ecological and geographic proximity, could select for the emergence of generalist protection mechanisms, which may include viral sequence integrations.
Integration of RNA virus genomes: focus on Aedes spp.

Among all the insect species in which NIRVS have been identified to date, Ae. aegypti and Ae. albopictus are notable for their high number of NIRVS and the diversity of their viral origins. In the genomes of these two mosquito species, NIRVS derived from genomes of flaviviruses, rhabdoviruses, reoviruses, bunyaviruses, phleboviruses, and quaranjavirus have been identified and NIRVS have been characterized from the Virgaviridae and Chuviridae virus families. (Crochu et al., 2004; Roiz et al., 2009; Katzourakis and Gifford, 2010; Fort et al., 2012; Cui and Holmes, 2012; Li et al., 2015; Suzuki et al., 2017; Palatini et al., 2017; Whitfield et al., 2017; ter Horst et al., 2018). All these studies agree on hundreds of individual NIRVS in Aedes spp., although the exact number of NIRVS from each viral type differs among studies depending on the bioinformatics parameters, sequencing and assembly method used, and source of data (i.e. genome sequence from the Aag2 cell line or adult mosquitoes). This richness and diversity of NIRVS in Ae. aegypti and Ae. albopictus is striking because these species are among the most medically important vectors of arboviruses such as dengue, Zika, yellow fever and chikungunya. This observation has prompted ongoing research efforts to understand the distribution of NIRVS in natural populations and whether NIRVS influence mosquito physiology and arbovirus infection and transmission.

NIRVS are not distributed randomly in the genomes of Ae. aegypti and Ae. albopictus, but rather are statistically significantly enriched in regions corresponding to PIWI-interacting RNA (piRNA)-generating loci, called piRNA clusters (Palatini et al., 2017; Whitfield et al., 2017). NIRVS are also differentially distributed in geographic populations (Pischedda et al., 2018). A difference between F-NIRVS and NIRVS with similarity to rhabdoviruses (R-NIRVS) was seen in the first whole-genome survey of NIRVS across five geographical populations of Ae. albopictus: R-NIRVS were more widespread and included more ancient integrations based on accumulation of sequence changes than F-NIRVS (Pischedda et al., 2018), an interesting result considering that negative-strand RNA viruses like rhabdoviruses are thought to be evolutionarily more recent than positive-strand flaviviruses (Koonin et al., 2015).

Exogenous virus resistance mediated by endogenous viral elements

The most comprehensive current data on the effects of NIRVS on host physiology are from studies of endogenous bornavirus-like nucleoprotein-encoding sequences (EBLNs) in human, mouse and squirrel genomes (Honda and Tomonaga, 2016). A recombinant EBLN from the ground squirrel genome was expressed in human oligodendroglialoma cells, where it co-localized with bornaviral sites of replication in the nucleus and inhibited the activity of infecting Bornavirus polymerase (Fujino et al., 2014). This dominant negative effect was assumed to be dependent on the relatively high-degree of amino-acid sequence identity (77%) between the protein expressed from the integrated complete viral ORF from the ground squirrel genome and the nucleocapsid protein of the Bornavirus (Fujino et al., 2014). However, the minimum degree of sequence-identity between the NIRVS-expressed protein and that of cognate viruses essential to exert competitive or dominant interaction was not established.

A transcript of an EBLN of the human genome (i.e. hsEBLN-1) has been proposed to function as a long non-coding RNA and control the expression of the neighbouring COMMD3 gene. COMMD3 encodes a protein that inhibits the NFκB immune pathway (Burstein et al., 2005; Sofuku et al., 2015). As such, hsEBLN-1 is considered an epigenetic immune regulator (Sofuku et al., 2015; Honda and Tomonaga, 2016). This type of function has not been observed for insect NIRVS, likely because, in addition to differences between the insect and vertebrate antiviral immune pathways, insect NIRVS are predominantly found in repetitive regions of the genome and correspond to partial viral ORFs.

The first evidence of a biological role for an insect NIRVS came when a ≈ 420 bp sequence from the Israeli acute paralysis virus (IAPV) was found integrated into the genome of bees (Apis mellifera), resulting in their resistance to subsequent infection with IAPV (Maori et al., 2007). The molecular mechanisms underlying this phenomenon were not
elucidated however (Maori et al., 2007; see Chapter 3 for details of similar observations in Drosophila).

**Endogenous viral elements and the piRNA pathway in mosquitoes**

A possible biological function for NIRVS in mosquitoes may be to enhance antiviral defence (Meisen et al., 2016b; Blair and Olson, 2017; Olson and Bonizzoni, 2017). An antiviral function was suggested by two observations. First, detection of RNA virus genome-derived piRNAs was shown in both uninfected and infected mosquitoes and mosquito cells (Scott et al., 2010; Brackney et al., 2010; Arensburger et al., 2011; Morazzoni et al., 2012; Vodovar et al., 2012; Schnettler et al., 2013; Miesen et al., 2016a,b). Second, integration of most NIRVS occurred in or near retroelements in piRNA clusters of *Ae. aegypti* and *Ae. albopictus* genomes. These observations not only suggest a mechanism by which non-retroviral RNA genomes were converted to cDNA and inserted into host genomes, but also that NIRVS function similarly to piRNA clusters, the genomic loci that express transposon-derived piRNAs implicated in defending host genome integrity by restricting transposon activity.

The small-interfering RNA (siRNA) pathway of the RNA interference (RNAi) system is the major antiviral defence system in acute arboviral infections of *Aedes* spp. mosquitoes (Sánchez-Vargas et al., 2009; Scott et al., 2010; Blair, 2011). In mosquitoes, the siRNA pathway is a cell-autonomous, cytoplasmic system in which the dsRNA-specific RNase III family member Dicer 2 (Dcr2) recognizes viral dsRNA produced during replication. Dcr2 binds the dsRNA and cleaves it into 21-nt siRNA duplexes with perfect base-pairing and 2-nt overhangs at the 3’ ends. In complex with the dsRNA-binding protein R2D2, Dcr2 loads the siRNA duplex onto Argonaute 2 (Ago2), an endonuclease that is part of the multi-component RNA-induced silencing complex (RISC). Ago2 cleaves and releases one strand of the siRNA duplex, retaining the second strand as a guide to hybridize to a complementary coding sequence in viral genome/mRNA, which it then cleaves in the centre of the complementary region. Transiently knocking down expression of Dcr2, R2D2, or Ago2 results in greatly increased arbovirus replication, indicating their central role in siRNA-mediated antiviral defence (Keene et al., 2004; Sánchez-Vargas et al., 2009). In either *Ae. aegypti* mosquitoes or *Ae. aegypti* embryo-derived Aag2 cell cultures, infection with the flavivirus dengue (DENV) results in production of DENV RNA-derived 21 nt siRNAs (Scott et al., 2010). However, in *Ae. albopictus*-derived C6/36 cells, which have a single-nucleotide deletion in their Dcr2 gene (Scott et al., 2010), arbovirus infection drives production of virus RNA-derived 27-nt piRNAs (Scott et al., 2010, Brackney et al., 2010). The more robust replication of DENV in exo-siRNA-deficient mosquito cells emphasizes the importance of this pathway in antiviral defence in *Aedes* spp., but production of piRNAs suggests that this pathway might have a secondary role in anti-arbovirus defence.

In *Drosophila*, which has served as a model for the mosquito RNAi system, production of mature piRNAs is carried out by the three members of the Piwi subfamily of the Argonaute protein family, Piwi, Aubergine (Aub) and Ago3, and is Dcr-independent (Brennecke, 2007, Gunawardane, 2007). These proteins are expressed in germline cells of the fly and their major function is to protect the genome by silencing transposable element (TE) activity. The substrates for production of primary piRNAs are nuclear transcripts from piRNA clusters in the genome that contain defective TE sequences. The transcripts are processed in the nucleus to produce mature, antisense Piwi- or Aub-associated 24–27 nt piRNAs, known as piRNA-induced silencing complexes (piRISC), that hybridize to and cleave complementary sense-strand RNA in the cytoplasm. This ‘slicer’ action of the piRISC results in positive-sense, Ago3-bound piRNA that can initiate a ‘ping-pong’ amplification loop (Siomi, 2011).

The Piwi protein subfamily has undergone expansion to eight proteins, Ago3 and Piwi1–7, in *Ae. aegypti* (Campbell et al., 2008), suggesting additional functional roles of these proteins. Also, the piRNA pathway is active in somatic as well as germline tissues of *Ae. aegypti* (Morazzani et al., 2012). Further studies are needed to determine which of the eight *Aedes* Piwi-family proteins may be involved in binding, processing and cleavage of...
the piRNA precursor. Determining whether a particular Piwi-family protein affects virus replication could depend on the mosquito tissue or cultured cells analysed, the time post-infection at which analysis is done, and the infecting arbovirus (flavivirus or alphavirus) (Vodovar et al., 2012; Morazzani et al., 2012; Schnettler et al., 2013; Miesen et al., 2015, 2016a; Goic et al., 2016). In addition to piRNAs derived from NIRVS in Aedes spp., newly generated piRNAs derived from the infecting arbovirus genome occur during acute infections of Aedes spp. cells and mosquitoes by flaviviruses, alphaviruses, and bunyaviruses (Scott, 2010; Morazzani, 2012; Léger, 2013; Schnettler, 2013; Miesen, 2015, 2016a,b). However, their role in antiviral defence is uncertain and the template for transcription of virus genome-derived piRNAs (vpiRNAs) and protein(s) involved in their biogenesis are not fully understood.

Several studies have shown that Aedes mosquitoes and mosquito cells are capable of synthesizing short segments of cDNA (vDNA) from alphavirus, flavivirus, or bunyavirus genome templates, using endogenous reverse transcriptase (RT), likely through recombination between viral RNA and a retrotransposon transcript during reverse transcription (Vodovar et al., 2012; Nag et al., 2016; Nag and Kramer, 2017). By analogy to the proviral DNA form of retrovirus genomes, this hybrid DNA is termed ‘proviral’. Proviral DNA is hypothesized to enter the nucleus where it is transcribed into primary piRNA precursors from vDNA episomes or after integration into a piRNA cluster in the genome using the retroelement integrase. Although robust production of virus genome-derived piRNAs has been observed in alphavirus-infected mosquitoes (Morazzani et al., 2012), no alphavirus genome-related NIRVS have been reported in mosquitoes. The evolutionary conservation of NIRVS and their variation in different populations of Aedes spp. provide support for their acquisition during acute infection. NIRVS acquisition and any long-term antiviral effects for a particular arbovirus would depend on infection of mosquito germ-line tissue and integration into the vector genome of NIRVS derived from the infecting viral RNA. Recent bioinformatics analyses have detected NIRVS derived from Dengue virus-1 in the genome of a distinct Ae. albopictus population from China; this may signify a recent NIRVS acquisition (Chen et al., 2015).

Furthermore, germ-line infections and vertical transmission of dengue viruses in Ae. aegypti probably occur more frequently than previously thought (Sánchez-Vargas et al., 2018).

**Concluding remarks and future directions**

The first experimental evidence that non-retroviral RNA viruses can integrate genome sequences into host DNA genomes surfaced in the 1970s (Zhdanov, 1975) but this evidence was ignored until the development of powerful and inexpensive next-generation sequencing technology that enabled the sequencing of many host genomes. Analyses of genome data are revealing an increasing number of viral sequence integrations in different organisms including insects. Although current data are mostly descriptive, preliminary observations are already transforming our perception of virus–host co-evolution, allowing for the generation of hypotheses about connections between virus evolution, insect ecology, insect genomic structure and insect antiviral immunity.

Understanding the significance of the widespread occurrence of viral integrations within insect genomes will require analysis of the genomes of more species (Fig. 2.1). It would be beneficial for the search for – and characterization of – viral integrations to be standard aims during the annotation of insect genomes. This would help answer whether integration of viral sequences is a common phenomenon within insects, or if it depends on the structure of the host genome (e.g. on the content of LTR transposons). It will also be important to verify which viral species/genera can efficiently integrate, especially as non-retroviral RNA viruses have been proposed as delivery tools for gene therapy (Mogler and Kamrud, 2015). This type of medical treatment is based on the transient presence of non-retroviral RNA viruses in host cells. For safety, it is essential to verify whether integrations into host genomes occur because of the properties of the host genome or of the virus. The use of long-read technologies to refine the genome assembly of species where NIRVS have been characterized will help resolve repetitive regions and unravel the landscape of NIRVS, including their numbers and their positioning within the genome (Whitfield et al., 2017; Matthews et al., 2018). Pertinent to this aim, a new,
A- Discovery

- Widespread occurrence of NIRVS in insect genomes
- Polymorphism and dynamics of NIRVS in natural populations
- Genomic context of NIRVS integrations
- Recent acquisitions of NIRVS in natural vector populations following outbreaks

B- Biological Function

- Domestication of NIRVS in gene exons
- ISV infections of cell lines as models for NIRVS biology and virus persistence
- Interplay of NIRVS and TE control elements associated with piRNA clusters

C- Exploitation

- Modification of the NIRVS landscape or engineering design of single NIRVS for vector control
- NIRVS as molecular markers for population genetics
- NIRVS and the biogenesis/function of piRNA clusters

Figure 2.1 Current knowledge gaps and areas for future research on endogenous viral elements from non-retroviral RNA viruses (i.e. non-retroviral integrated RNA virus sequences or NIRVS) in insect genomes. (A) NIRVS Discovery, a better understanding of the widespread occurrence of NIRVS in insect genomes, the timing of integrations and the genomic context in which integrations occur will lead to understanding of the process of endogenization and its impact on insect genomes. (B) Biological Function, indirect evidence on the distribution of NIRVS in the genomes of the mosquitoes Ae. aegypti and Ae. albopictus and their viral origin point to an interplay with TE control elements associated with piRNA clusters and mainly ISVs, which could be experimentally addressed in cell-line based systems. Cases of NIRVS domestication could also have occurred given the finding of several NIRVS as complete ORFs or part of annotated exons. (C) Exploitation, depending on the differential distribution of NIRVS in insect populations, NIRVS could be co-opted as novel molecular markers to depict population genetic structure or trace recent invasions. Additionally, given that in Ae. aegypti and Ae. albopictus NIRVS are statistically significantly enriched in piRNA clusters, unravelling the process of endogenization could offer insights on the biogenesis and functions of piRNA clusters in non-model organisms. Finally, NIRVS may be harnessed for vector control by modifying the landscape of NIRVS within the vector genome or using NIRVS sites as anchors to drive expression of effector molecules. In this regard, studies need to determine whether NIRVS have a direct role in anti-viral immunity, how the level of sequence-similarity between NIRVS and incoming viruses can attain antiviral activity, and if it is possible to genetically modify mosquitoes by introducing novel NIRVS into the host genome.

more complete and accurate annotation of the Ae. aegypti genome, designated AeagL5, has been released to resolve additional NIRVS and piRNA cluster sites (Matthews et al., 2018). This information could enable testable predictions regarding the mechanisms and entities involved in the integration process (e.g. by confirming NIRVS arrangements with respect to specific TE lineages). At the same time, elucidating the sites of integrations could help design experiments to test for NIRVS functions. If integrations are numerous, if NIRVS are duplicated and located primarily within piRNA clusters, which are composed of TE fragments, using genetic engineering strategies that require unique anchor sequences to modify NIRVS (i.e. knock-out NIRVS or introduce novel viral sequences) may be difficult. Examining the genomic landscape and polymorphism of NIRVS in natural specimens will enable exploration of the pervasiveness and micro-evolution of integrations. Finally, the similarity of most NIRVS to the genomes of ISVs, most of which are easily cultivated in insect cells, point to the promise of using ISVs and insect hosts as ‘animal model’ systems to elucidate the endogenization process and study the impact of NIRVS on exogenous virus infections.

The major controller of acute arbovirus infection in the vector is the siRNA pathway of RNAi. We (and others) have observed that when the RNAi inhibitor B2 (from Flock House virus) is expressed
by the arbovirus Sindbis (genus *Alphavirus*), Sindbis virus infection in the vector is greatly increased and the fitness of the vector is significantly lowered (Myles et al., 2008; Cirimotich et al., 2009). We hypothesize that NIRVS act by expressing virus genome-derived piRNAs in the vector to target and control infections. Thus, NIRVS may represent a fortuitous co-adaptation between arboviruses and vectors (Blair and Olson, 2015; Olson and Blair, 2016; Palatini et al., 2017). NIRVS probably do not prevent or ablate infections but help maintain virus persistence in the vectors.

We have developed RNAi- or ribozyme-based antiviral effector genes. Integration and expression of either of these effector genes in *Ae. aegypti* at the time of infection profoundly lowers or eliminates vector competence for several arboviruses (Franz et al., 2006; Mishra et al., 2016). However, these antiviral transgenes are integrated randomly in the genome using mariner or piggyBAC transposable elements, and their expression varies according to their insertion site. As we improve our understanding of NIRVS integration, expression, and biological functions, NIRVS could provide valuable insights into how we might optimally engineer antiviral transgenes into the vector genome to maintain antiviral gene cassettes for multiple generations, as observed for piRNA clusters. As NIRVS reveal naturally selected genetic engineering of the vector to control infection outcomes, NIRVS may provide the perfect road map for how to successfully engineer antiviral transgenes.

**References**


occasional event or something more? Curr. Opin. Insect Sci. 22, 45–53. https://doi.org/10.1016/j.cois.2017.05.010


Sánchez-Vargas, I., Scott, J.C., Poole-Smith, B.K., Franz, A.W., Barbosa-Solomieu, V., Wilusz, J., Olson, K.E., and Blair, C.D. (2009). Dengue virus type 2 infections of *Aedes aegypti* are modulated by the mosquito’s RNA interference pathway. PLOS Pathog. 5, e1000299. https://doi.org/10.1371/journal.ppat.1000299


