# UNIVERSITA' DEGLI STUDI DI PAVIA

FACOLTA' DI INGEGNERIA
DIPARTIMENTO DI INGEGNERIA INDUSTRIALE E DELL'INFORMAZIONE



DOTTORATO DI RICERCA IN TECNOLOGIE PER LA SALUTE, BIOINGEGNERIA E BIOINFORMATICA
XXXIII CICLO - 2020


# COMPUTATIONAL METHODS AND TOOLS FOR INVESTIGATING MOSQUITO-ARBOVIRUS CO-EVOLUTION


PhD Thesis by

**ELISA PISCHEDDA**



**Advisor:**
**Prof. Mariangela Bonizzoni**

**PhD Program Chair:**
**Prof. Silvana Quaglini**

# Acknowledgments

I desire to thank Professor Mariangela Bonizzoni, for supporting me in these years of PhD, for all the professional growth and opportunities that she has allowed me to experience and for teaching me the true meaning of passion for research.

Thanks also to all the members of the laboratory that I have had the pleasure to meet over the years, thanks to them it was possible to carry out even the most difficult job.

# Abstract

L'interazione tra i virus e i loro ospiti avviene a molti livelli. I virus hanno un impatto sulle strutture genetiche delle popolazioni ospiti attraverso la deriva genetica e/o la selezione naturale se l'infezione virale provoca malattie e mortalità. L'esempio più notevole di questo fenomeno è la pandemia influenzale del 1918, che causò circa 50 milioni di morti in tutto il mondo. Un altro meccanismo attraverso il quale i virus influenzano il loro ospite è l'integrazione del genoma. L'integrazione degli acidi nucleici virali può avvenire sia nelle cellule somatiche che in quelle germinali. Le integrazioni somatiche sono associate all'instabilità del genoma, che può progredire nella carcinogenesi. Inoltre, l'integrazione virale nei genomi dell'ospite è un modo per eludere la risposta immunitaria dell'ospite e favorire la persistenza dell'infezione. Se le integrazioni avvengono nelle cellule germinali, possono essere trasmesse verticalmente. La persistenza e l'esito di queste integrazioni nelle popolazioni ospiti dipendono dai loro effetti sulla fitness dell'ospite. Se deleterie, le integrazioni vengono perse. In alternativa, le sequenze virali possono essere adottate funzionalmente dall'ospite ed esercitare funzioni benefiche. Per esempio, il prodotto del gene di restrizione del retrovirus murino, Fv1, protegge i topi contro l'infezione con il virus della leucemia murina (MLV) e altri retrovirus. Fv1 è derivato dal gene gag di un antico retrovirus endogeno lontano dal MLV. Una terza possibilità si verifica se le integrazioni sono in posizioni cromosomiche che non sono trascritte o mancano di funzioni di regolazione. In questo caso, le integrazioni virali possono persistere e subire riarrangiamenti, iniziando con l'accumulo di mutazioni. Le integrazioni virali dei virus del DNA e dei retrovirus sono un fenomeno comune, come dimostra l'abbondanza di sequenze virali identificate nei genomi di vari organismi sequenziati finora (es. il codice genetico da retrovirus costituisce circa l'8% del genoma umano). Poiché i virus a RNA non retrovirali mancano della codifica per la trascrittasi inversa e il macchinario di integrazione necessario per il trasferimento di successo al DNA genomi, il loro potenziale di integrazione è stato considerato minimo. Tuttavia, il numero di studi che mostrano integrazioni genomiche in cellule ospiti sia somatiche che germinali da virus a RNA non retrovirali, compresi virus a RNA a singolo filamento (positivo e negativo) e a doppio filamento, è in aumento. I virus a RNA comprendono arbovirus (arthropod-borne viruses) e virus specifici degli insetti (ISVs). Sequenze derivanti da NRVs sono state trovate integrate all'interno del genoma di eucarioti tra cui la zanzara della febbre gialla, *Aedes aegypti*, e la zanzara tigre asiatica, *Aedes albopictus*. Queste due specie sono i principali vettori per diversi arbovirus

epidemiologicamente rilevanti tra cui i virus di Dengue (DENVs), Zika e Chikungunya (CHIKV).

In entrambe le specie di *Aedes*, le integrazioni virali chiamate Non-retroviral Integrated RNA Virus Sequences (NIRVSs) o non-retroviral Endogenous Viral Elements (nrEVEs), si trovano in regioni del genoma che contengono anche sequenze di elementi trasponibili (TE), spesso all'interno di piRNA clusters e producono piRNA. I cluster di piRNA sono stati studiati soprattutto in *D. melanogaster* dove è stato dimostrato che sono composti da sequenze frammentate da TE. I piRNA sono prodotti da questi frammenti e limitano il movimento dei TE in base alla complementarità della sequenza.

La similarità osservata tra la modalità di organizzazione dei TE e degli nrEVEs suggerisce che gli nrEVEs si comportino come i TE all'interno dei piRNA clusters e che quindi anche loro possano essere marcatori di infezioni passate e avere un ruolo nell'immunità antivirale contro virus affini. Un corollario a questa ipotesi è che la distribuzione degli nrEVEs dipenda dall'esposizione virale delle zanzare e che quindi sia diversa tra campioni naturali di popolazioni geografiche diverse. Tuttavia, il meccanismo che permette un'integrazione virale e il ruolo degli nrEVEs è ancora poco chiaro.

L'assenza di farmaci specifici per arbovirus e la limitata presenza di vaccini, stimolano la ricerca di nuove strategie di controllo degli organismi vettori. Un'idea è di manipolare geneticamente i vettori in modo che diventino incapaci di permettere l'infezione, la replicazione e la trasmissione dell'agente patogeno.

In particolare, proprio perché gli nrEVEs possono contribuire alla capacità del vettore di trasmettere arbovirus, recenti studi stanno promuovendo l'utilizzo degli nrEVEs ingegnerizzati come nuova strategia di controllo dei vettori.

Negli ultimi anni, i ricercatori nel settore della biologia dei vettori hanno promosso il sequenziamento e l'assemblaggio di genomi di riferimento delle due specie di *Aedes* che fossero ben risolti anche nelle regioni ripetute del genoma. Questi sforzi certamente aprono nuove frontiere di ricerca per lo studio del polimorfismo delle integrazioni virali e per identificare integrazioni virali in campioni naturali. Inoltre, sono necessari strumenti per confrontare i risultati delle integrazioni virali tra diverse versioni dei genomi di riferimento.

Su queste basi, in questa tesi ho sviluppato approcci bioinformatici per annotare automaticamente le integrazioni virali in un genoma di riferimento, studiarne la distribuzione in campioni raccolti in campo e scoprire nuovi nrEVEs (integrazioni virali assenti nel genoma di riferimento), che possono essere specifici dell'esposizione virale delle zanzare di campo.

L'annotazione delle integrazioni virali nelle due specie di *Aedes* è stata resa disponibile per la comunità scientifica per studi futuri nel sito www.nreves.com, un database navigabile online di elementi virali endogeni provenienti da virus a RNA non retrovirale.

Inoltre, alla luce del sempre più largo utilizzo delle tecnologie semantiche anche nel settore delle scienze della vita, ho creato un'ontologia di nrEVEs. L'ontologia è stata modellata riutilizzando vocaboli di ontologie disponibili in BioPortal e creando *ex novo* solo i termini non trovati. L'ontologia delle

integrazioni virali è un primo passo verso l'utilizzo dei dati riguardanti gli nrEVEs in applicazioni automatiche e interoperabili.

Sia i metodi bioinformatici che l'ontologia presentata in questa tesi sono estensibili ad altri organismi ponendo le basi per studi futuri, in particolare nella comprensione del complesso sistema dell'integrazione virale nelle specie vettori.

# Abstract

Interaction between viruses and their hosts occur at many levels. Viruses impact the genetic structures of host populations through genetic drift and/or natural selection if viral infection results in disease and mortality. The most remarkable example of this phenomenon is the influenza pandemic of 1918, which caused about 50 million deaths worldwide.

Another mechanism through which viruses affect their host is genome integration. Integrations of viral nucleic acids can occur both in somatic and germline cells. Somatic integrations are associated with genome instability, which can progress into carcinogenesis. Additionally, viral integration into host genomes is a way to elude the host immunity response and favor the persistence of the infection. If integrations occur in germline cells, they can be vertically transmitted. The persistence and the outcome of these integrations in the host populations depend on their effects on the host fitness. If deleterious, integrations are lost. Alternatively, viral sequences can be functionally adopted by the host and exert beneficial functions. For instance, the product of the murine retrovirus restriction gene, Fv1, protects mice against infection with murine leukemia virus (MLV) and other retroviruses. Fv1 is derived from the gag gene of an ancient endogenous retrovirus distantly related to MLV. A third possibility occurs if integrations are in chromosomal locations that are not transcribed or lack regulatory functions. In this case, viral integrations may persist and undergo rearrangements, starting with the accumulation of mutations.

Viral integrations of DNA viruses and retroviruses is a common phenomenon as shown by the abundance of viral-related sequences identified in the genomes of various organisms sequenced so far (i.e., genetic code from retroviruses constitute around 8% of human genome). Because non-retroviral RNA viruses lack the coding for reverse transcriptase and the integration machinery needed for successful transfer to DNA genomes, their integration potentials were considered minimal. However, the number of studies showing genome integrations into both somatic and germline host cells from non-retroviral RNA viruses, including single-stranded (positive and negative) and double-stranded RNA viruses, is increasing.

RNA viruses comprise arboviruses (arthropod-borne viruses) and insect-specific viruses (ISVs). Sequences from NRVs have been detected integrated into the genomes of eukaryotes including the Yellow fever mosquito, *Aedes aegypti*, and the Asian tiger mosquito, *Aedes albopictus*. These two species are primary vectors for many epidemiologically relevant arboviruses including Dengue viruses (DENVs), Zika virus and Chikungunya virus (CHIKV).

In both *Aedes* spp., viral integrations, called Non-retroviral Integrated RNA Virus Sequences (NIRVSs) or non-retroviral Endogenous Viral Elements (nrEVEs) are embedded next to transposable element (TE) sequences, enriched in piRNA clusters, and produce piRNAs. piRNA clusters have been studied mostly in *D. melanogaster* where they were shown to be composed of fragmented sequences from TEs. piRNAs are produced from these fragments that limit TE movement based on sequence complementarity.

The similarities observed between the way TEs and nrEVEs are organized in piRNA clusters suggests that nrEVEs behave like TEs of piRNA clusters, thus they could be markers of past infections and have a role in antiviral immunity against cognate viruses. A corollary of this hypothesis is that nrEVEs landscape depends on mosquito viral exposure, thus it should be different across wild mosquito populations. The mechanism through which integrations occur and the nrEVE biological role are still poorly understood.

The absence of arbovirus-specific drugs and limited vaccines for arboviral diseases, stimulate the research of novel vector control strategies. One idea is to genetically manipulate the vectors so that they become unable to support pathogen infection, replication and transmission.

On this basis, recent studies are promoting genetically engineered nrEVEs to test whether they could reduce vector competence, thus be employed in novel vector replacement strategies.

In the last few years, researchers in vector biology have made extensive efforts in sequencing and assembling the genomes of *Aedes* spp. mosquitoes and produce assemblies that would be well-resolved also in repeated regions of the genome. These efforts certainly opened up new research possibilities to study the polymorphism of viral integrations and identify viral integrations from wild-collected samples. Additionally, tools to compare results on viral integrations across subsequent versions of genome assembly are needed.

On this basis, in this thesis I developed bioinformatics approaches to automatically annotate viral integrations in a reference genome assembly, study nrEVEs landscape in wild samples and discover new nrEVEs (i.e., viral integrations absent in the reference genome), which should be depend on viral exposure of wild samples.

The annotation of viral integrations in *Aedes* spp. has been made available for the community for future studies at www.nreves.com, an online browsable database.

Additionally, given the increasingly application of semantic technologies in life science, I created the semantic ontology for nrEVEs. The ontology was modeled re-using vocabularies of available ontologies in BioPortal and creating *ex novo* vocabularies only for missing terms. The ontology of viral integrations is a first step to make nrEVEs data available for automatic and interoperable applications.

Both the bioinformatics methods and the ontology presented in this thesis are extensible to nrEVEs of other organisms laying the foundations for future studies.

# Contents

# Abbreviation list

**BAM**: Binary Alignment Map
**BED**: Browser Extensible Data
**CDS**: coding sequence
**CFAV**: Cell Fusing Agent Virus
**CHIKV**: Chikungunya virus
**CSV**: Comma Separated Value
**DENV**: Dengue Virus
**evalue**: expected value
**EVE**: Endogenous Viral Element
**FALDO**: Feature Annotation Location Description Ontology
**F-NIRVS**: nrEVEs with similarities to Flavivirus
**FG**: Fast evolving Gene
**GBOL**: Genome Biology Ontology Language
**GpY**: Generations per Year
**INDEL**: INsertion-DELetion
**ISV**: Insect Specific Virus
**KRV**: Kamiti River Virus
**LOD**: Linked Open Data
**LoP**: level of Polymorphism
**LOV**: Linked Open Vocabulary
**LT**: Lateral Transfer
**LTR**: Long Terminal Repeat
**MR**: Mutation Rate
**mya**: million year ago
**NGS**: Next Generation Sequencing
**N-G:** gene including a nrEVEs in its coding sequence or untranslated region
**N protein**: Nucleocapsid protein
**NIRVS**: Non-retroviral Integrated RNA Virus Sequences
**NR**: Non-Redundant
**nrEVE**: non-retroviral Endogenous Viral Element
**NRV**: Non-retroviral RNA Viruses
**ORF**: Open Reading Frame
**OWL**: Web Ontology Language
**piRISC**: piRNA-Induced Silencing Complex
**piRNA**: PIWI-interacting RNA
**R-G**: genes of the RNA interference pathway
**RefSeq**: Reference Sequence
**RdRP**: RNA dependent RNA Polymerase
**R-NIRVS**: nrEVEs with similarities to Rhabdovirus
**RNAi**: RNA interference
**RO**: Relation Ontology

**RSA**: Reference Sequence Annotation
**SAM**: Sequence Alignment Map
**SG**: Slow evolving Gene
**SIO**: Semanticscience Integrated Ontology
**SNP**: Single Nucleotide Polymorphism
**SO**: Sequence types and features Ontology
**spp**: species
**SSM:** Single Sample Mosquitoes
**STR**: Short Tandem Repeat
**TE**: Transposable Element
**UP**: Uniprot KB
**UTR**: UnTranslated Region
**VCF**: Variant Calling Format
**vDNA**: viral DNA
**VSV**: Vesicular Stomatitis Virus
**WGS**: Whole Genome Sequencing
**ZIKV**: Zika virus

# Chapter 1

# Introduction

## 1.1. Arboviral diseases

Arboviral diseases are infections caused by viruses transmitted to humans through arthropod vectors, such as mosquitoes and ticks, and thus collectively called arthropod-borne viruses or arboviruses.

Dengue (DENVs), Zika (ZIKV), Yellow Fever (YFV) and Chikungunya (CHIKV) viruses are among the most prevalent arboviruses worldwide. The World Health Organization (WHO) estimated that the incidence of dengue has increased 30-fold in the past five decades and half of the world's population lives in countries where Dengue is endemic [1]. Arboviral diseases are also emerging in temperate areas of the world, such as Europe. For instance, cases of Dengue and Chikungunya have been reported in southern France and Croatia since 2010 and Italy suffered from Chikungunya outbreaks in 2007 and 2017. There are no arbovirus-specific drugs and vaccines are limited, thus prevention of arboviral diseases lays on vector control.

Historical methods of vector control such as the use of insecticides and environmental control are facing challenges due to the wide spread of insecticide resistance throughout natural mosquito populations and the complexity of breeding site elimination in modern urban environments. Innovative genetics-based strategies are emerging as promising complement to historical mosquito control methods. One idea is to genetically manipulate vectors so that they become unable to support pathogen infection, replication or transmission [2].

## 1.2. Arboviruses and Insect-Specific Viruses

Apart from few exceptions, all arboviruses are Non-Retroviral RNA viruses (NRVs) [3], [4]. Once acquired by the vector through the blood

feeding on an infected host, viral particles replicate and disseminate throughout the vector body until they reach the salivary glands. Once in the salivary glands, viral particles can be transmitted to a new host during a second blood-feeding (**Figure 1**) [5]. During the early stages of infection, viral titer increases, and mosquitoes mount a complex immunity response. Later a balance between viral replication and the mosquito immune system is reached resulting in the establishment of a non-pathogenic persistent infection which ensures mosquitoes a life-long viral transmission capacity [6]–[10].

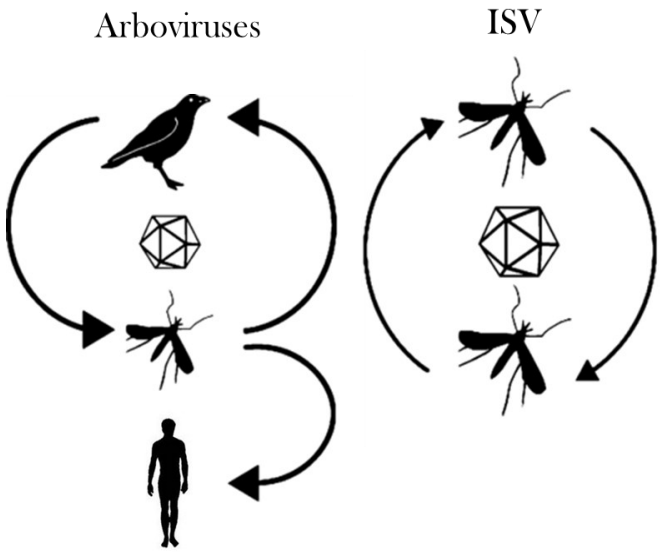


**Figure 1.** NRVs replication cycle [5].

Arboviruses of relevant importance for human and animal health belong to six main taxonomic families: *Togaviridae*, *Flaviviridae*, *Rhabdoviridae*, *Bunyaviridae*, *Reoviridae*, *Oxomyxoviridae* [3], [11]. Within the same phylogenetic families including arboviruses, there are viruses unable to replicate in vertebrate cells and have been called Insect Specific Viruses (ISVs). ISVs are not pathogenic to humans because they naturally infect and replicate exclusively in insect cells [12].

ISVs are studied in relation with arboviruses because they both infect arboviral vectors and thus interact and may compete within the insect body. The presence of ISVs at the root of many phylogenies of viral families including arboviruses support the hypothesis that arboviruses emerged from ISVs that expanded their host range to include vertebrate cells [4], [12]–[14]. Additionally, ISVs may alter host vectoral capacity by the upregulation of host antiviral immune responses or through superinfection exclusion [12], [14]. On this basis, ISVs have been proposed as novel biological control agents against arboviruses.

## 1.3. Arboviral vectors

The main arboviral vectors worldwide are the yellow fever mosquito *Aedes aegypti* and the Asian tiger mosquito *Aedes albopictus*. *Aedes aegypti* was described for the first time in 1894 by Frederick A. Askew Skuse [15]. Aedes albopictus was first described in 1757 by Linnaeus F. Hasselqvist [16].

The prominent role of *Aedes* spp. mosquitoes as arboviral vectors is also related to their invasive capacity.

*Aedes aegypti* originated in the sub-Saharan Africa and is currently found throughout tropical and subtropical regions of the world [17], [18] (**Figure 2**). In Europe, *Ae. aegypti* was recorded throughout the first half of the 20$^{th}$ century and in 2004 it re-emerged in limited areas [19]. The success of *Ae. aegypti* depends primarily on its domestication [19]. In particular, the ability to live indoors and in association to humans, which provides new habitats that can be exploited as oviposition sites [17].

*Aedes albopictus* is native of south east Asia and since the late 18$^{th}$ century it spread to islands of the Indian and Pacific Oceans [20]–[22] (**Figure 2**). In the last few decades, *Ae. albopictus* moved globally and become established in all continents with the exception of Antarctica [23]. *Aedes albopictus* was first reported in Europe in 1979 in Albania and is currently established permanently in different countries around the Mediterranean, including Italy which is considered the most heavily infested country in Europe [24], [25].

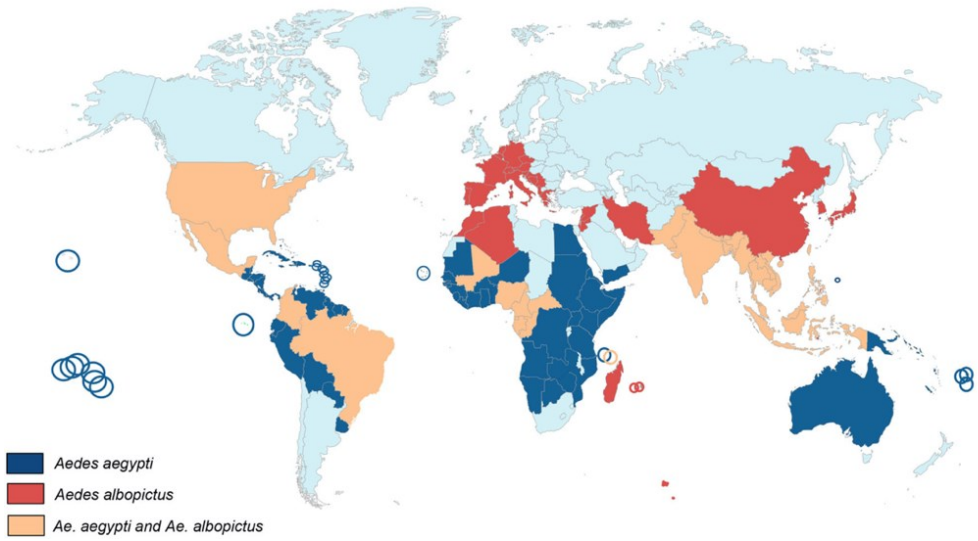


**Figure 2.** Global distribution of *Aedes* spp. [23].

The establishment of *Ae. albopictus* in temperate regions of the world is supported by the ability of some populations to enter photoperiodic diapause. Photoperiod diapause implies developmental arrest at the stage of eggs in presence of a reduced period of daylight, as occurring in fall-winter months,

and a resumption of metabolic activity and growth when daylight elongates as in spring. Photoperiodic diapause allows *Ae. albopictus* to overcome winters [26], [27]. Given its great ecological plasticity, *Ae. albopictus* was recognized as one of the top 100 invasive species in the world [28] and it is believed that will be able to spread also to the northern European countries in the future following climate change [29]–[31].

## 1.4. Viral integrations

Viral integrations are sequences from viruses that integrate into host genomes. Viral integrations can occur in somatic or germline cells. Somatic integrations of viruses have been linked to persistent viral infection and genotoxic effects, including various types of cancer in humans. Viral sequences that integrate into germline cells can be transmitted vertically, be maintained in host genomes and be co-opted for host functions. Viral integrations are referred to as Endogenous Viral Elements (EVEs) [32], [33]. EVEs have long been known, especially integrations from retroviruses in mammalian and *Drosophila melanogaster* genomes. Modern genomic sequencing analyses showed that non-model organisms may also harbor EVEs, which derive not only from DNA viruses and retroviruses, but also from NRVs. EVEs coming from NRVs, are called non-retroviral Endogenous Viral Elements (nrEVEs) or Non-retroviral Integrated RNA Virus Sequences (NIRVS) [4], [32], [34], [35]. The mechanisms behind the integration event and the frequency of this event are still poorly understood.

Viral integrations can be annotated in a reference genome assembly and are defined by the genomic coordinates delimiting the viral sequence. Whole genome sequencing (WGS) reads of a sample having the viral integration are all mapped in the reference genome and overlap both the host and viral regions with a homogeneous coverage (**Figure 3A**). Hereafter we refer to this category of viral integrations as ‘reference viral integrations’ or ‘reference nrEVEs’.

Wild-collected samples or samples collected under *ad hoc* experimental conditions, such as after a viral infection, may also harbor in their genome viral integrations that are not found in the reference genome assembly. These viral integrations can be identified through the analysis of Next Generation Sequencing (NGS) data. Paired-end WGS reads of a sample having the viral integration show a mate mapping to the host and one mate mapping to the virus suggesting the insertion of a viral integration in a specific position of the reference genome (**Figure 3B**). Mate reads mapping both to the viral portion appear unmapped from the alignment of the WGS data against the reference genome of the host. Hereafter we refer to this category of viral integrations as ‘novel viral integrations’ or ‘novel nrEVEs’.

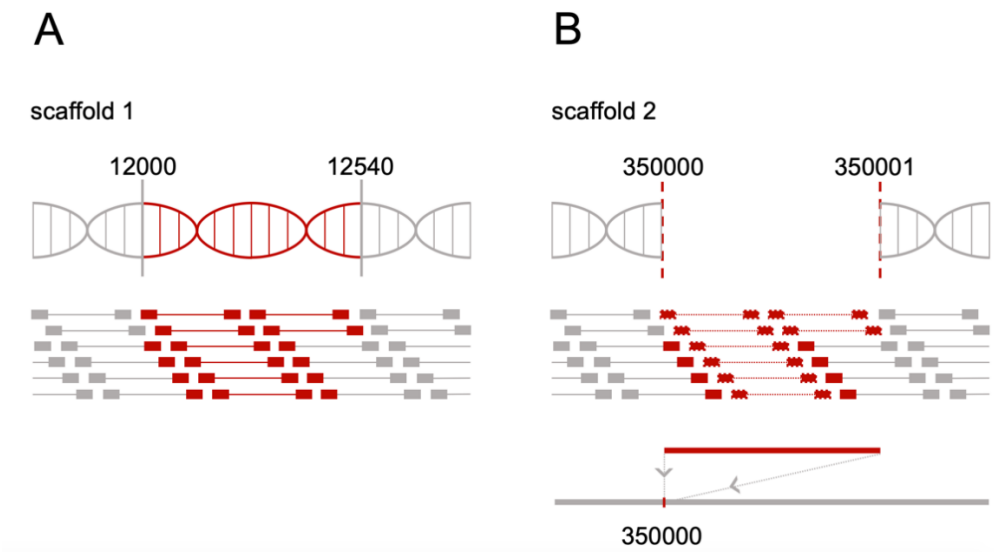


**Figure 3.** Schematic representation of viral integrations. **A)** A reference viral integration is a viral integration annotated in the reference genome assembly. It has specific coordinates delimiting the viral portion. Sample reads aligned in the reference genome overlap both the host and viral regions with a homogeneous coverage. **B)** A novel viral integration is a viral integration absent in the reference genome. It appears as an insertion in a specific position of the reference genome. Novel viral integrations are detectable through the alignment of reads from whole genome sequencing data from wild-collected samples or samples collected under special conditions (i.e., after a viral infection). Dashed mate reads mapping both to the viral portion, appear unmapped from the alignment of the WGS data.

### 1.4.1. Tools to detect EVEs in a reference genome assembly

In a reference genome assembly, bioinformatics-based identification of viral integrations is usually based on the comparison between the reference genome of the studied organism and a database of sequences from selected viruses. Most popular tools for the alignment of the viral sequences against the reference genome, and vice versa, are Blast+, either blastx or tblastn [36]. Alignment cut-off values of between $10^{-3}$ to $10^{-6}$ are generally considered [37]–[39]. After the identification of positive blast hits, the putative viral sequences are usually filtered by looking at additional criteria (i.e., presence of viral genome open reading frames or minimum size restrictions) to reduce false positive results resulting from low complexity reads with short tandem repeats (STR) or homopolymers.

The number of EVEs that will be characterized is influenced by the number of viral genomes considered along with the stringency of the criteria used to define an insect sequence as a viral integration (i.e., cut off blast expected value [evalue]; reverse blast; analyses of low complexity sequences; length of sequences; inclusion of newly discovered and unclassified viruses). Examples of EVEs identification include searches using a single virus families [13], [40], plant viruses transmitted by insects

[41], all known RNA viruses, including those most-recently identified (i.e., Whitfield et al., 2017; Palatini et al., 2017) [34], [38] or a mixture of DNA and RNA viruses [39], [42]. Modern genomic sequencing projects have detected numerous examples of viral integrations in organisms as different as the mouse and squirrel genomes, hematophagous and non-hematophagous insects, ticks, ants and other arthropods [33], [39], [43]–[45]. In these non-model organisms, viral integrations tend to occur in repetitive DNA, mostly in association with TE sequences, and have been proposed to constitute a novel form of heritable adaptive immunity elements [33], [34], [38].

### 1.4.2. Tools to detect novel nrEVEs

The majority of the bioinformatic tools to identify novel viral integrations have been developed in the context of cancer genetics [46], [47]. These tools can be categorized in three classes based on the strategy used to analyze whole genome sequencing (WGS) data [47]; in all cases WGS data consists of paired-end reads as obtained from next-generation sequencing (NGS) strategies such as Illumina.

- Host-Virus strategy:
    - raw reads are aligned to the host genome;
    - unmapped reads are extracted;
    - unmapped reads are aligned to a viral genome to recognize integration events.
- Virus-Host strategy:
    - raw reads are aligned to a viral genome;
    - partially mapped reads are extracted to detect viral integrations.
- Host and Virus strategy:
    - raw reads are aligned to a hybrid reference genome including both the host and the viral sequence.

One of the above-mentioned strategies or a combination of them are used by several tools such as VirusSeq, VirusFinder, ViralFusionSeq, VERSE, HIVID, SummonChimera, Vy-PER, Virus-Clip, BATVI, HGT-ID, ViFi, VirTect [48]–[59]. Each of these computational methods is differentially versatile in terms of data input format (i.e., RNA-seq or DNA-seq data, reference viral databases or customization opportunities), performances and computational requirements, but all have in common being geared towards the human genome thus, a well-annotated genome. Additionally, some of the above-mentioned tools depend on a large number of external programs making their installation cumbersome. Most of these tools do not have stringent quality control procedures and this might increase the possibility of false-positives detection [47].

Novel EVEs are usually studied in relation to human health, for instance by looking at integrations of possibly cancerogenic viruses such as HPV and

HBV or for studying viral latency for instance for HIV. In non-model organisms, in which EVEs often occur in repetitive DNA and in association with TE sequences, the identification of novel viral integrations is often impaired by the complexity and/or fragmentation of the reference genome assembly. Thus, rarely studies of viral integrations in non-model organisms have gone beyond their characterization from annotated reference genome sequences because of the lack of computational methods suited to handle repeated and fragmented genomes.

### 1.4.3. Viral integrations in insects

Viral integrations from NRV were first discovered in the genomes of *Ae. aegypti* and *Ae. albopictus* mosquitoes in 2004 as a fortuitus results from a PCR reaction with Flavivirus-specific primers on mosquito genomic DNA [60].

With the onset of NGS technology and the improvement of the bioinformatic techniques, nrEVEs were annotated in the *Aedes* spp. reference genome assemblies and nrEVEs were identified in the genomes of other insect species [61].

A peculiarity of nrEVEs of insect genomes is that they are often flanked by TEs, in particular, by long terminal repeats (LTRs) retroelements. In 2015, DNA fragments of CHIKV were detected in mosquitoes after CHIKV infection [62]. These viral DNA (vDNA) fragments were flanked by TEs [62]. vDNA fragments are produced by *D. melanogaster* and *Aedes* spp. mosquitoes after infection with various arboviruses [62]–[65]. These findings support the hypothesis that nrEVEs derive from vDNA fragments and are embedded within TEs. In *D. melanogaster*, vDNA fragments are produced by the DExD/X helicase domain of Dicer-2 from defective viral particles [63].

### 1.4.4. Biological relevance of nrEVEs in arboviral vectors

In 2017, nrEVEs deriving primarily from flaviviruses and rhabdoviruses were characterized from the genome of Culicidae mosquitoes, including the arboviral vectors *Ae. aegypti*, *Ae. albopictus* and *Cx. quinquefasciatus* and protozoan vectors such as *Anopheles* spp. [40].

nrEVEs were found to be ten-fold more abundant in the *Ae. aegypti* and *Ae. albopictus* genomes than in any other tested mosquito genomes [13], [34], [42] (**Figure 4**). Moreover, in both *Aedes* spp. nrEVEs were statistically significantly enriched in PIWI-interacting RNA (piRNA) clusters and produced piRNAs [34].

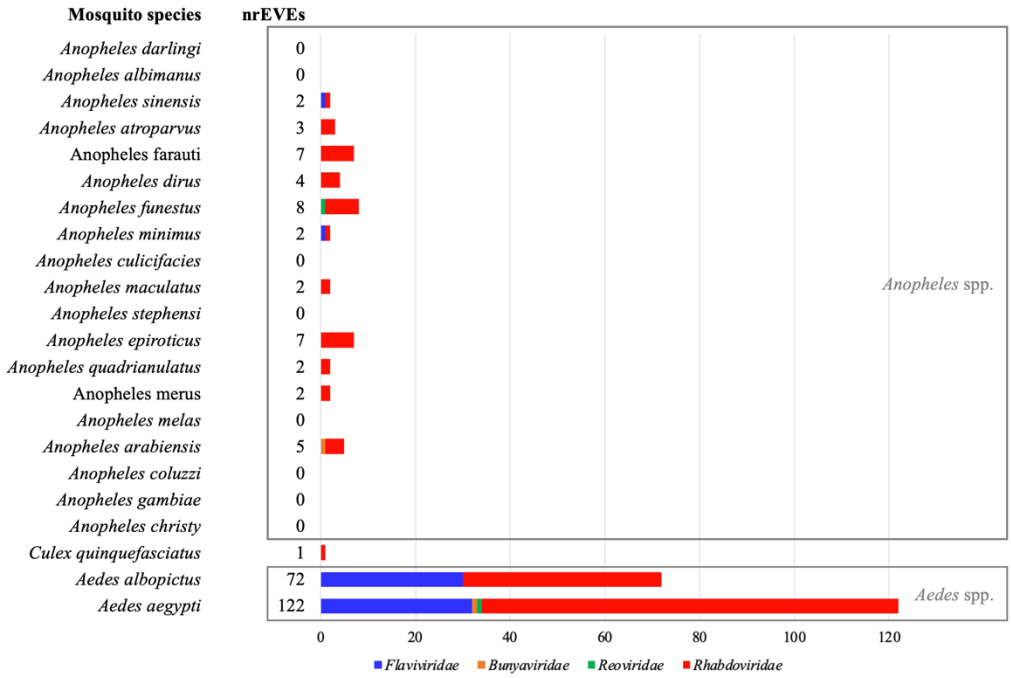


**Figure 4.** Distribution of nrEVEs several viral families. nrEVEs are 10-fold more abundant in *Aedes* spp. [34].

The piRNA pathway was identified only a decade ago and it has been largely studied in *D. melanogaster*, where it is of key importance for silencing TEs in germline tissues [66]. Briefly, in *D. melanogaster*, the synthesis of piRNAs is performed by proteins of the Piwi subfamily of the Argonaute protein family, namely Piwi, Aubergine (Aub) and Ago-3. In the nucleus, long piRNA precursors are produced from piRNA clusters. Precursors are processed in the nucleus to generate mature antisense piRNAs of 25-30 nucleotides which are moved to the cytoplasm where they form the piRNA-induced silencing complex (piRISC). Together with Ago-3, the Aub-piRNA complex serves as a trigger to start the "ping-pong" amplification pathway in the cytoplasm, leading to the formation of secondary piRNAs [67], [68] (**Figure 5**). piRNA clusters contain sequences of previously acquired TEs. Thus, piRNA clusters constitute an archive of past TEs invasions. Altering the composition of TE fragments within the *D. melanogaster flamenco* locus resulted in modification of the regulatory properties against cognate TEs [69].

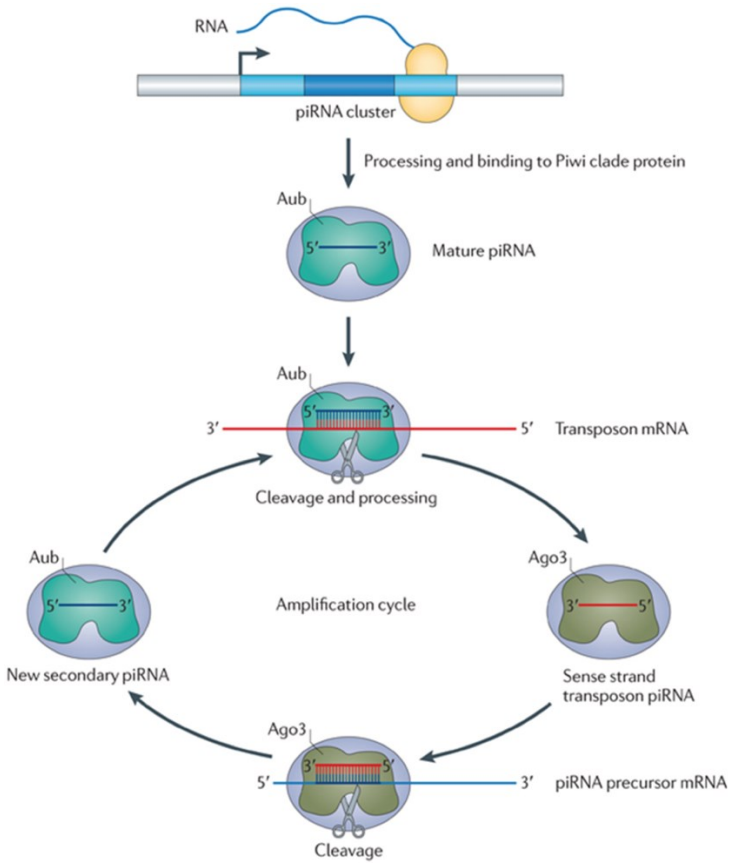


**Figure 5.** Schematic representation of the piRNA pathway [67].

Recently, it was shown that, in addition to its canonical function in preserving genome integrity, the piRNA pathway has antiviral activity in *Aedes* spp. mosquitoes [70]. Observations supporting a prominent role of the piRNA pathway in mosquito antiviral immunity include: 1. virus-derived piRNAs are produced upon viral infections of *Aedes* mosquitoes and they exhibit the typical hallmarks of the ping-pong amplification; 2. knockdown of PIWI expression results in an increase of virus replication; 3. PIWI proteins and piRNAs expression is not limited to germline cells in mosquitoes, but also occurs in the soma where arboviruses replicate; 4. the PIWI family of proteins is expanded in *Aedes* spp. mosquitoes in comparison to *D. melanogaster*, suggesting functional specialization [71].

The physical contiguity between TE and nrEVEs, along the production of piRNAs from nrEVEs support the hypothesis that nrEVEs behave analogously of TE fragments within the piRNA pathway. Corollaries of this hypothesis are that: 1. viral integrations are not only ancient events, but an ongoing process related to mosquito viral exposure [35]; 2. the landscape of nrEVEs differs in geography populations, 3. nrEVEs may confer protection from subsequent infections with cognate viruses and affect vector competence [72]–[75]; 4. biologically-relevant nrEVEs should be selectively retained in host genomes.

## 1.5. Databases in life science

The advent of high-throughput technologies has revolutionized the way to study biological systems. High-throughput technologies allowed researchers to systematically study the genomes of any organism (Genomics), the set of its RNA molecules (Transcriptomics and non-coding RNAs), and the set of its proteins including their structures and functions (Proteomics) [76]. These heterogeneous data sources continuously generate a huge amount of different types of data. These "big" data are stored and shared in databases, which are formulated and organized depending on the nature of the data [77].

Currently relational databases are the default and classical solution in the biomedical domain because of their simplicity in installation and usage [78]. However, conventional data resources and repositories do not naturally express semantics in their models. As a consequence, they are limited in their addition and cannot be investigate easily in relation with other data to extract more complex semantic information (network relationships) [78].

Core databases such as UniProt [79] and Ensembl [80], include data that needs to be structured defining an appropriate storage format and metadata to give semantics meaning to the data. Metadata could be included into the format, but usually they are provided externally in the form of annotations, for example in the form of ontology. Modern ontology languages provide enough technics to both the annotation and the description of biological objects, including the structural and semantical description of objects. This makes biological data available for interoperability applications [81].

Given the possibility of the semantic integration in Life Sciences, major bioinformatics institutions such as UniProt [79] embrace semantic ontologies and, more generally, all Linked Data technologies as a common platform for biological data integration. For the development of advanced use-case scenarios for the end users and their applications it is necessary to develop interlinked distributed datasets [82].

An example of the power of interlink data based on increasing knowledge is the Linked Open Data (LOD) Cloud. Currently LOD is the largest network of datasets published and interconnected (the version of May 2020 includes 1260 datasets and 16187 links) [83] (**Figure 6**). The majority of the data included in LOD derive from the government, linguistic, publication and life science environment. Every LOD data guarantees that it can be freely accessible by any human or machine with access to the Internet. Further, interconnected data could be investigated simultaneously to extract new information compared to just separated datasets.

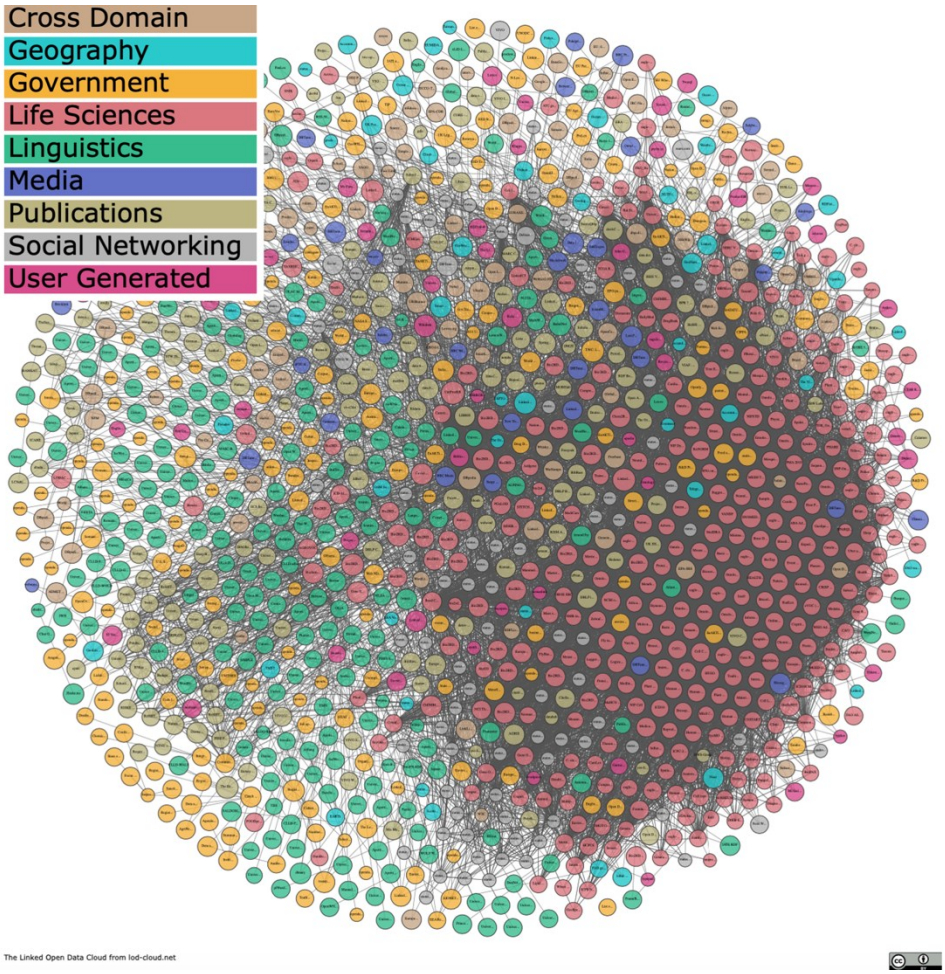


**Figure 6.** LOD Cloud Diagram in October 2020 [83].

Data on viral integrations have never been organized in a storage structure sharable with the research community so far. The possibility of organizing this type of data in an online repository from which data could be downloaded in bioinformatic standard formats conform to the current principles of open science [84].

# Chapter 2

# Thesis objectives and outlines

Viruses can impact the genetic structures of host populations not only if viral infection results in disease and mortality, but also by transferring parts of their genetic material into host genomes [32]. *Aedes albopictus* is an invasive mosquito able to transmit several public health relevant arboviruses. Previous work by members of the Bonizzoni laboratory demonstrated the presence of hundreds of nrEVEs in the genomes of *Ae. aegypti* and *Ae. albopictus* mosquitoes [34]. In *Aedes* spp. genomes, nrEVEs appear to be in close association with TEs, they are enriched in piRNA clusters and produce piRNAs. The physical contiguity of nrEVEs with TEs and their production of piRNAs suggests that nrEVEs may act as TEs of piRNA clusters, thus possibly constitute novel heritable antiviral effectors [75].

Understanding the widespread distribution of nrEVEs and their evolution could help deciphering their biological relevance and unravel mechanisms of integrations.

On this basis, the goals of my thesis are to develop: 1. computational methods and tools to characterize nrEVEs and study their evolution; 2. a database of nrEVEs to share results with the research community. This last goal is based on the paradigm of open science: the idea that sharing of materials, data and information is the foundation of a solid and reproducible scientific finding that can help speed scientific discoveries.

To reach these goals, I designed four focused aims, which correspond to dedicated chapters.

**Aim 1. Annotation of viral integrations in reference genome assemblies.**

I contributed to the development of an automatic short pipeline to avoid manual filtering after the selection of putative reference viral integrations. I will present results of the application of my pipeline to the genomes of *Ae. albopictus* in Chapter 3.

The pipeline that I describe can be extended to other organisms, thus allowing to understand whether the phenomenon of host genome integrations from NRVs occurs in all viral lineages, or it is limited to specific host-viral combinations.

**Aim 2. Reference nrEVEs polymorphism.**

I contributed with the development of a bioinformatic pipeline named 'Structural Variant Definition' (SVD) that allows to test nrEVEs polymorphism. nrEVEs polymorphism is analyzed in terms of 1. their presence/absence in WGS data from samples collected in the wild or under hypothesis-driven experimental conditions and 2. their sequence. I will present results of the application of my pipeline in Chapter 4.

**Aim 3. Detection of novel viral integrations.**

I contributed with the development of a bioinformatic pipeline named 'Virus in Repeats' (ViR). ViR was designed to solve intrasample variability and ameliorate predictions of viral integration sites when dealing with reference genomes full of repetitive DNA, duplications or suffering from assembly fragmentation. I will present results of the application of my pipeline in Chapter 5. The biological significance of nrEVEs in mosquito genomes is strongly dependent on their distribution in different geographic populations, which have been exposed to different circulating viruses. This aim will allow to identify novel nrEVEs using WGS data from wild-collected mosquitoes and test their frequency.

**Aim 4. Database of viral integrations.**

This aim will allow to make data on nrEVEs easily accessible and understandable by any researchers to facilitate their work and to provide a complete picture of the entities cooperating in this complex model. Because the discovery that nonretroviral RNA viruses can integrated into host genome is relatively recent and information on their widespread are limited, but growing, we have the unique opportunity to start building a reference nrEVEs database while nrEVEs are being characterized. Additionally, I propose the first ontology to describe viral integrations according to the linked data principles to increase the interoperability and the usage of these data in automatic application.

I will present results of the creation of the database in Chapter 6.

Overall conclusion of the thesis will be presented in Chapter 7.

# Chapter 3

# Annotation of viral integrations

## 3.1. *Aedes albopictus* reference genome assemblies

When I started my PhD program, the genome sequence of the Foshan strain of *Ae. albopictus* had just become available [40]. This sequence was assembled into the AaloF1 assembly starting from WGS data from a single pupa. AaloF1 is currently hosted in Vectorbase, the bioinformatic resource for invertebrate vectors of human pathogens [40], [85]. AaloF1 is highly fragmented biasing the annotation of nrEVEs. Thus, in the Bonizzoni laboratory we built an international consortium that re-sequenced the genome of *Ae. albopictus* mosquitoes of the Foshan strain using PacBio long-sequencing reads technologies and Hi-C. The new genome sequence and its assembly, which we called AalbF2, was published in 2020 [86] and is currently available on NCBI with the Reference Sequence (RefSeq) accession number GCF_006496715.1 [86].

The genome length of *Ae. albopictus* ranges between 1.190 Gb and 1.275 Gb as estimated by a cytofluorimetric approach [86]. This value is different from both the genome lengths of AaloF1 and AalbF2, due to duplications and miss assembly most probably related to the high abundance of repetitive DNA in this species [86]. AaloF1 is composed of 154782 scaffolds for a total genomic length of 1.9 Gb. AalbF2 is composed of 2197 scaffolds for a total genomic length of 2.5 Gb. AalbF2 shows a N50 length of 55.7 Mb, which represent a continuity increase of 2 orders of magnitude from AaloF1[86].

Both AaloF1 and AalbF2 were used to annotate viral integrations. Annotation of nrEVEs in AaloF1 preceded my joining the Bonizzoni's lab [34].

## 3.2. EVE annotation pipeline

I annotated nrEVEs in the newest *Ae. albopictus* reference genome (GCF_006496715.1) [86] using the EVE_finder pipeline published by Whitfield et al., 2017 [38] and a viral database composed of 1563 viral species [86]. The database includes all complete amino acids sequences of single strand (ss)RNA, double strand (ds)RNA and unclassified RNA viruses with a tropism for arthropods; viral sequences were downloaded from NCBI Viral Genomes Browser [87] in November 2019.

Briefly, the EVE_finder pipeline starts with the alignment of the reference genome against the viral protein database using blastx [36]. The result of the alignment is then converted in the Browser Extensible Data (BED) format using a custom script and sorted using the BEDtools 'sort' function [88]. Then, the BEDtools 'merge' function [88] selects the largest coordinates of the overlapped regions and a second custom script attributes the best result for the merged region. Finally, BEDtools 'getfasta' [88] is used to extract the sequences of the selected EVEs from the reference genome based on the merged coordinates. The result of the pipeline is a list of genomic coordinates where putative viral integrations have been identified.

### 3.2.1. Refinement of the EVEs annotation

I reasoned that the output FASTA file of the putative viral integrations detected by the EVE-finder pipeline should be further filtered to reduce the chance of false positives and avoid eventually untraceable errors due to manual curation. First filtering step was a reverse blastx [36] against the protein databases of RefSeq and Non-Redundant (NR) from NCBI as download in November 2019. The evalue threshold for this analysis was set at $10^{-6}$. Then, I refined the annotation using a custom bash and python-based pipeline. Starting from the output file of the reverse blastx and the BED file of the putative viral integrations detected by the EVE_finder pipeline the best protein match for each candidate was selected. The blastx output file is designed to show a specific column order including the protein accession number and taxid (**Figure 7A**).

Taxon-ids from the blastx file are then divided in two categories 'Viral' and 'Non-Viral' using the Virus-Host Classifier [89]. 'Non-Viral' entries are parsed with a set of regular expressions to discard matches with eukaryotic genes and uncharacterized or low-quality proteins (**Figure 7B**). evalue and Subject ID both for the best viral match and the best non-viral match are reported in the output file along with the number of total and viral hits. This output result format allows users to easily select integrations based on their custom selection criteria.

Finally, viral taxon-ids are parsed with the Taxonkit tool [90] to extract the corresponding viral family and order.

A

```
blastx -query tophits.fasta \
-db NR/RefSeq_protein \
-evalue 1e-06 \
-outfmt '6 qseqid qstart qend salltitles saccver evalue qframe pident qcovs sstart send slen staxid' \
-out TopHits.blastx
```

B

```
if re.search(r'.*AGAP.*-PA.*', k_hit_db) is None and \
    re.search(r'.*AAEL.*-P.*', k_hit_db) is None and \
    re.search(r'.*ncharacterized.*', k_hit_db) is None and \
    re.search(r'.*PREDICTED.*', k_hit_db) is None and \
    re.search(r'.*hypothetical.*', k_hit_db) is None and \
    re.search(r'.*CLUMA_CG.*', k_hit_db) is None and \
    re.search(r'.*baculoviral.*', k_hit_db) is None and \
    re.search(r'.*LOW QUALITY PROTEIN:.*', k_hit_db) is None and \
    re.search(r'.*unnamed protein product.*', k_hit_db) is None:
    selected_nonviral_evalues.append(v_hit_db.evalue)
    selected_nonviral_elem.append(k_hit_db)
```

**Figure 7.** EVE annotation pipeline sections. **A)** reverse blastx command line to search the best match in the NR or RefSeq protein database for each top hit sequence. The output file reports the format shown in the value of the parameter outfmt. **B)** Portion of the python script including the regular expressions used to filter out eukaryotic and inaccurate matches.

Additional information regarding piRNA clusters and TEs were included in the output file to describe the genomic context in which viral integrations occur. piRNA clusters and annotated coding sequences (CDSs) regions [86], [91] were intersected with the coordinates of the reference nrEVEs using the BEDtools 'intersect' function [88]. A database of *Ae. albopictus* TE sequences was obtained from Jake Tu (Virginia Tech) and used to align with the 1000 bp - boundary sequences of the reference nrEVEs using blastn [36] and restricting results with an evalue of $10^{-5}$. Class, Order and Superfamily were assigned to TEs and included in the output file of nrEVE annotation [92]–[97].

## 3.3. nrEVEs in the reference genome of *Aedes albopictus*

Using the EVE_finder pipeline followed by the previously-described filtering steps, I annotated a total of 456 nrEVEs with similarity to viruses from nine viral families in AalbF2 (**Figure 8**) [86].

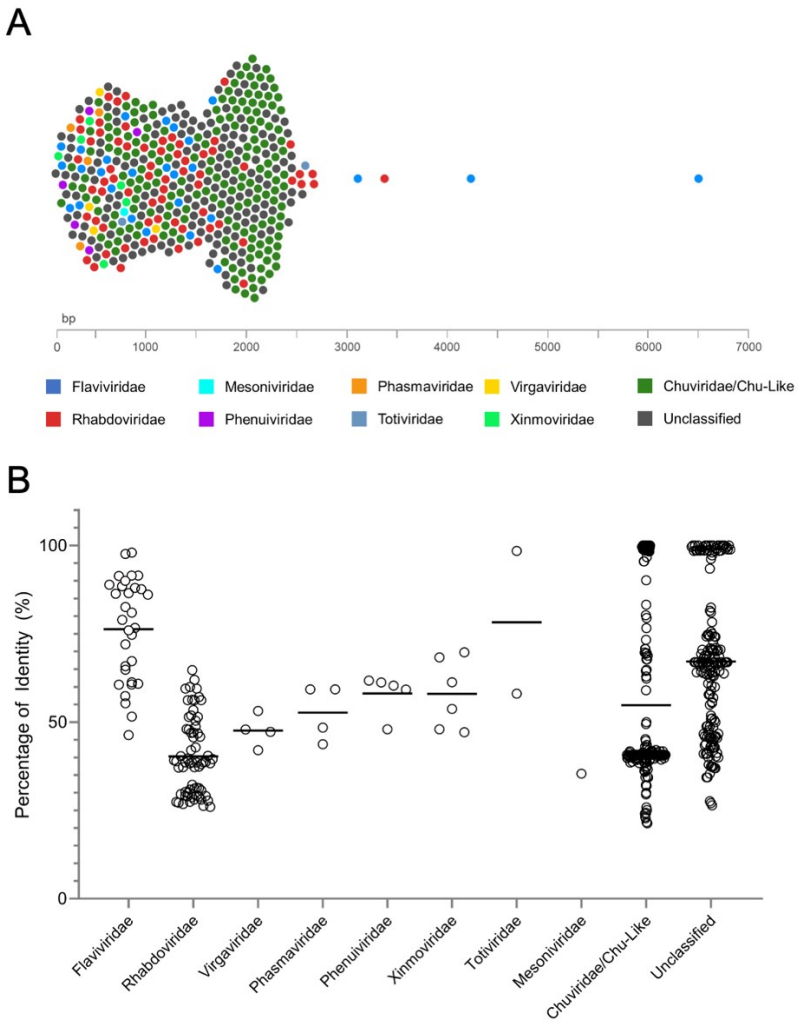


**Figure 8.** nrEVEs identified in the *Ae. albopictus* genome (AalbF2). **A)** Bee swarm plot showing nrEVEs (dots) distributed according to their length and color-coded based on their viral origin. **B)** Box plots representing the amino acid identity of each nrEVE and its best hit retrieved by blastx searches against NR database grouped by viral family. The average is shown by a line [86].

nrEVEs range in length from 131 to 6593 nt, with an average of 1289 nt. The majority of them have similarities to known ISVs of the Flavivirus and Rhabdovirus genera (**Figure 8A**). Considering average values of amino acidic percentage identities between nrEVEs and their best viral hits identified, nrEVEs from the Flaviviridae family are more similar to currently circulating viruses than nrEVEs with similarity to Rhabdoviruses (**Figure 8B**). nrEVEs from both Flaviviridae and Rhabdoviridae families often appear concatenated to each other, generating clusters of rearranged or duplicated sequences which are in tight association with TEs, primarily Gypsy and Pao LTRs (**Figure 9A**). This association appears to be driven by the enrichment of LTR retrotransposons into piRNA clusters (**Figure 9B**) [86].

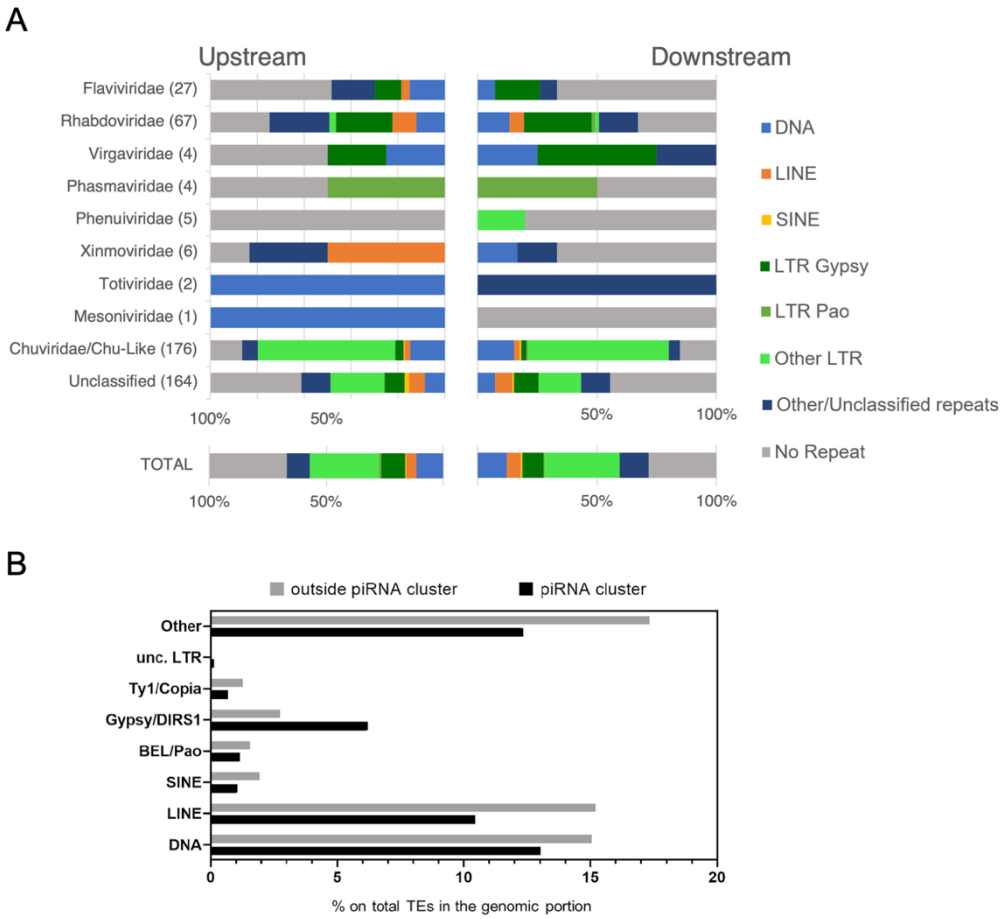


**Figure 9.** Distribution of TEs in boundary regions of nrEVEs and in piRNA clusters. **A)** Bar plots showing transposable elements identified upstream and downstream each reference nrEVE. Viral integrations are classified based on their viral origin [86]; **B)** Bar plots showing the percentage on the total genome content in and out piRNA clusters for each TE category.

## 3.4. Correspondence between nrEVEs annotated in AaloF1 and in AalbF2

To provide correspondence between the nrEVEs annotation in AaloF1 and AalbF2, I aligned the sequence of the 72 nrEVEs identified in AaloF1 against the sequence of the 456 nrEVEs identified in AalbF2 with blastn [36].

The difference in the number of nrEVEs detected in AaloF1 and AalbF2 is due to both the different annotation method adopted and the different viral database (**Figure 10**). The annotation of the nrEVEs in AaloF1 was described by Palatini et al., in 2017 [34]. Briefly, the reference genome was screened using tblastx [36] with a viral database including 424 NRVs and 1 DNA virus (African swine fever virus). Hits of at least 100 bp and high identity (evalue <0.0001) with viruses were selected and for each nrEVE locus the hit with the highest score was retained. To reduce the chance of false positives a series of filtering steps was implemented. Filtering steps included a reverse

tblastx [36] against all nucleotide sequences in the NCBI database [87], a search for Open Reading Frame (ORF) sequencing encompassing viral proteins based on NCBI ORF finder [98] and a functional annotation based on Argot2 [99].

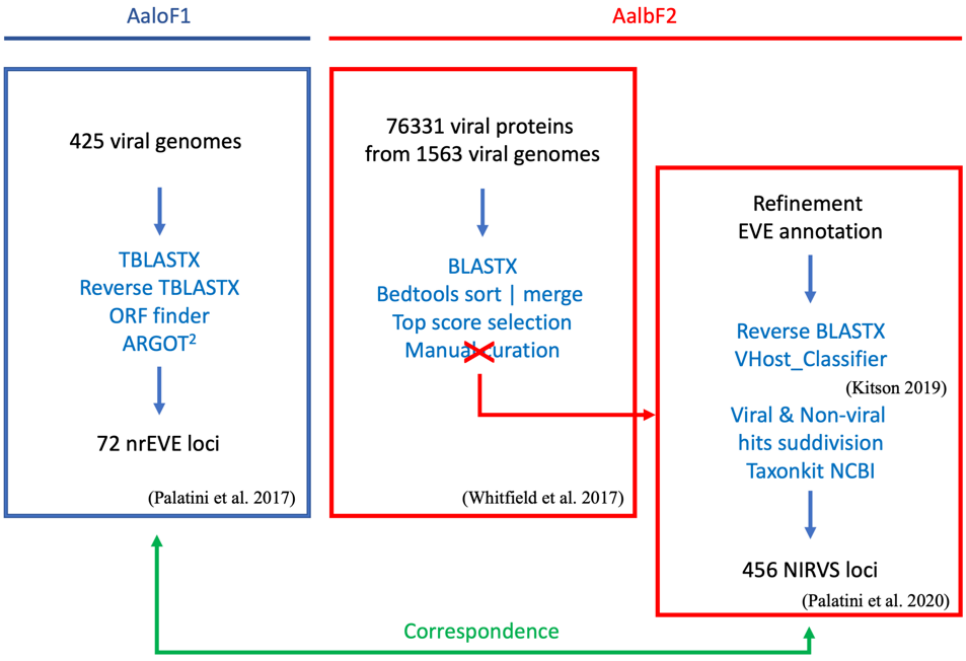


**Figure 10.** Comparison of the pipelines used to identify nrEVEs in AaloF1 [34] and AalbF2. The correspondence of nrEVEs annotated in AaloF1 and AalbF2 was studied to allow researchers to translate results from one annotation to the other.

The 72 nrEVEs annotated in AaloF1 were divided into two groups: 30 nrEVEs have similarity to viruses of the Flaviviridae family (F-NIRVS) and 42 nrEVEs have similarity to viruses of the Rhabdoviridae family (R-NIRVS). Considering blastn hits with nucleotide identity higher than 90% and alignment length higher than 100 bp, 16 out of the 30 F-NIRVS and 26 out of the 42 R-NIRVS showed to correspond to nrEVEs identified in AalbF2 for their entire length. 7 out of the 30 F-NIRVS and 5 out of the 42 R-NIRVS showed to correspond to nrEVEs identified in AalbF2 for a portion of their length. No matches were identified for 7 F-NIRVS and 11 R-NIRVS. For 14 F-NIRVS and 20 R-NIRVS more than one match was identified in the list of nrEVEs from AalbF2, probably as a result of the high sequence similarity among nrEVEs of the same viral family, as previously described [34]. Finally, in three cases for F-NIRVS and one case in R-NIRVS, multiple viral integrations in AaloF1 have match with the same nrEVE in AalbF2. This is a further demonstration of the improvement of the new reference genome of *Ae. albopictus* compared to the first on in terms of reduction of fragmentation and duplication (**Appendix 1)**.

# Chapter 4

# Reference nrEVEs polymorphism

I developed the 'Structural Variants Definition' (SVD) pipeline to study the polymorphism of reference nrEVEs both in terms of their presence/absence and their sequence using WGS data from wild-collected mosquitoes. Results of my work allowed to investigate an unexplored component of the mosquito repeatome and were published in 2019 in Frontiers in Genetics (10.3389/fgene.2019.00093).

## 4.1. Structural Variants Definition Pipeline

The overall scheme of the SVD pipeline is shown in **Figure 11**. The pipeline can be applied to WGS data from one mosquito, or alternatively to WGS data from multiple samples, hereafter called population.

All input requirements of the pipeline are shown in **Appendix 2** and are settable in one command line. Input requirements include the sample file, the pipeline directory, the output directory, the reference files (i.e., the genome of the studied organisms in FASTA and the coordinates of the loci of interest in BED), the position of the required tools and input parameters of the tools. The sample file includes the list of the alignment files of the WGS raw data in the reference genome (Binary Alignment Map file, BAM file).

The majority of the tools used in the pipeline are named 'Variant Callers' because they aim to find genomic variants: Single Nucleotide Polymorphisms (SNPs) and INsertions or DELetions (INDELs). The parameters of the variant callers allow to detect variants with specific quality features such as a minimum base and mapping quality and a minimum reads coverage. Because not all the callers allow to set in input the same parameters, I included some filter parameters to be used to homogenize the identified variants in terms of common features. Finally, some output parameters can be set by the user to include in the output file named 'AllData' a specific set of results.

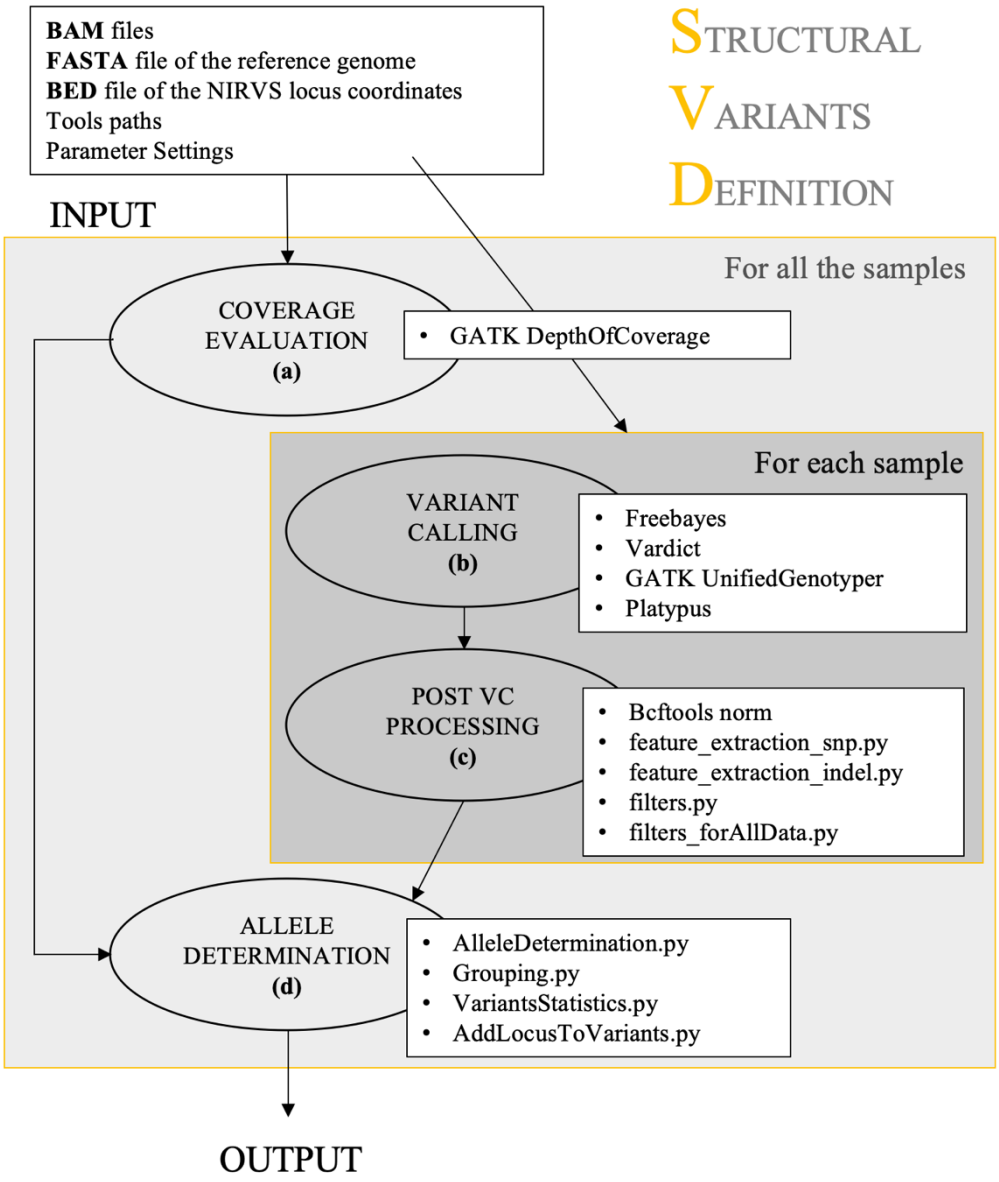


**Figure 11.** Scheme of the SVD Pipeline. The pipeline investigates presence of nrEVEs and their sequence polymorphism in WGS data through four subsequent steps: the coverage evaluation, the variant calling, the post variant calling processing and the allele determination [100].

The pipeline is divided into four main steps (**Figure 11**). In case of a population, the first and the last step are evaluated considering all samples at the same time, while the intermediate steps are evaluated separately for each sample.

In the first step, the 'DepthOfCoverage' function of the GATK tool [101] is used to evaluate the coverage of the regions of interest limiting to reads with Phred mapping quality greater than the threshold set by the user.

In the second step, four different Variant Callers are used to identify SNPs and INDELs within the regions of interest. The chosen Variant Caller are GATK UnifiedGenotyper [101], Freebayes [102], Platypus [103], and

Vardict [104], are used to identify SNPs and INDELs within the regions of interest. The search of SNPs and INDELS by different variant callers allows to increase the pool of variants and reduce the number of false positive.

In the third step, the 'Bcftools norm' function [105] is used to split multi-allelic variants calls into multiple records and simultaneously, to achieve the left alignment of the INDEL variants (**Figure 12**).

```
X       100639593       .       CAAA    CAA,C   47      QD
BRF=0.04;FR=0.5030,0.4968;HP=18;HapScore=4;MGOF=8;MMLQ=26;MQ=59.3;
GT:GL:GOF:GQ:NR:NV      1/2:-1,-1,-1:8:22:23,23:8,8

X       100639593       .       CA      C       47      QD
BRF=0.04;FR=0.503,0.4968;HP=18;HapScore=4;MGOF=8;MMLQ=26;MQ=59.3;
GT:GL:GOF:GQ:NR:NV      1/0:-1,-1,-1:8:22:23:8
X       100639593       .       CAAA    C       47      QD
BRF=0.04;FR=0.503,0.4968;HP=18;HapScore=4;MGOF=8;MMLQ=26;MQ=59.3;
GT:GL:GOF:GQ:NR:NV      0/1:-1,-1,-1:8:22:23:8
```

**Figure 12.** Bcftools example. In the top of the figure, one record contains in the alternate section two variants separate by comma (CAA, C). In the bottom of the image, this record is separated in two different lines and for the first variant the left alignment is implemented.

The resulting Variant Calling Format (VCF) files from each Variant Caller are then manipulated be able to compare results. Despite the VCF file is a standard format for genomic variants calling, sometimes several variant callers indicate the same parameter with different names (i.e., the number of reads supporting the variant in Freebayes is named Alternate Observation [AO], in Vardict and GATK is named Allelic Depth [AD], in Platypus is named Number of Variant reads [NV]). The feature 'extraction step' solves this problem creating a list of homogenized features. This step is implemented separately for SNPs and INDELs (**Appendix 3**).

The following features are considered:

- $GT$: the genotype of the variant identified by the caller. Variants showing genotype 0/0 are filter out as they are false positive variants (they are in homozygosity with the reference). Genotypes 0/1 and 1/0, were rewrite as 1 as they imply heterozygosity with respect to the reference and the genotype 1/1 was rewritten as 2 as it implies variant in homozygosity.

$$GT = \begin{cases} 1 & variant\ in\ heterozygosity \\ 2 & variant\ in\ homozygosity \end{cases}$$

- $AO$: Alternate Observation is the number of reads supporting the variant allele.
- $RO$: Reference Observation is the number of reads supporting the reference allele.

- $AO_f$: AO forward is the number of forward reads (5'-3') supporting the variant allele.
- $AO_r$: AO reverse is the number of reverse reads (3'-5') supporting the variant allele.
- $RO_f$: RO forward is the number of forward reads (5'-3') supporting the reference allele.
- $RO_r$: RO reverse is the number of reverse reads (3'-5') supporting the reference allele.
- $DP$: depth of coverage is the total number of reads supporting a nucleotide.
- $DP_{norm}$: depth of coverage of the variant position normalized with respect to the mean number of reads coverage given in input.
- $DP_f$: depth of coverage forward is the total number of forward reads (5'-3') supporting a nucleotide.
- $DP_r$: depth of coverage reverse is the total number of reverse reads (3'-5') supporting a nucleotide.
- $AF$: allele frequency is the ratio between the number of reads supporting the variant allele and the total number of reads supporting the nucleotide in which the variant occurs.

$$AF = \frac{AO}{DP}$$

- $StrandBias$: index of the fraction of reads supporting the variant in the two directories forward (5'-3') and reverse (3'-5').

In case of total absence of reads in one direction some alignment error could occur, thus, these variants must be filtered out. The value of the strand bias, showed in the feature extraction output file, is evaluated as Fisher score, after implementing the Fisher Exact Test. The resulting value ranges from 0 to 1. When the value tends to 0 there is not strand bias; when the value tends to 1 there is presence of strand bias. Strand bias is evaluated with $RO_f$, $RO_r$, $AO_f$ and $AO_r$. For Freebayes, Vardict and Platypus the value is evaluated as:

$$StrandBias = \begin{cases} Fisher\ Score = 1 - P_{value} & if\ \min\frac{(DP_r, DP_f)}{DP} \geq 0 \\ Null & if\ \min\frac{(DP_r, DP_f)}{DP} < 0 \end{cases}$$

The GATK VCF file does not contain $RO_f$, $RO_r$, $AO_f$ and $AO_r$ values, but it includes PhredFS that is the Phred scale Pvalue of the Fisher exact test. To make results comparable among callers, $PhredFS$ was converted with the following formula:

$$StrBiasFS = 1 - pow\left(10, -\frac{PhredFS}{10}\right)$$

- $MQ0$ and $MQ0F$ are the number and the frequency of the reads with mapping quality 0, respectively. Available only for GATK.
- $BQRankSum$: the Z score WilcoxonRankSumTest evaluates if there is statistical difference between the base quality of the reads supporting the reference and the base quality of the reads supporting the variant allele. The variant may be an artifact if there is bias in the base quality distribution.
- $MQRankSum$: the Z score WilcoxonRankSumTest evaluates if there is statistical difference between the mapping quality of the reads supporting the reference and the mapping quality of the reads supporting the variant allele.
- $QB$: base quality is the mean of the base quality of the reads supporting the alternate variant. Available only for Freebayes e Vardict.
- $Call$: if the flag is 1, it means that the caller found the variant. It is available for each of the four Variant Callers implemented.

Mean and median of the Allele frequency, the Depth of coverage and the Strand Bias are evaluated.

Not all the callers allow the usage of all the parameters setting in input: the minimum coverage parameter is available just for Freebayes; the minimum alternate fraction is available just for Freebayes and Vardict; the minimum number of alternative observation and the minimum mapping quality parameters are available just for Freebayes and Platypus; the minimum base quality parameter is available for all the tools except for Vardict.

To compensate the missing application of filters by some variant callers, I applied a further filter step. In particular, I exclude variants with allele fraction (AF) less than the minimum AF or with depth of coverage (DP) less than the minimum DP required by the user.

The final step of the pipeline produces in output:

- the list of nrEVEs detected in each sample;
- for each nrEVE, the list of SNPs and INDELs detected in each sample;
- for each nrEVE, the list of different alleles found in the samples;
- a summary table including the previous information.

For each sample in analysis, the summary table shows:

- the number of variants found in each locus;
- the number of variants found in heterozygosity in each locus. If the number of callers that call the variants in heterozygosity is equal or

higher than the number of callers that call the variant in homozygosity than the genotype is considered heterozygous.

$$Variant\ Genotype = \begin{cases} Homoz, & if\ Heteroz_{calls} < Homoz_{calls} \\ Heteroz, & if\ Heteroz_{calls} \geq Homoz_{calls} \end{cases}$$

If there is one variant in heterozygosity, the allele genotype is called in heterozygosity;

- the length of the allele;
- the Level of Polymorphism (LoP) defined as follow:

$$LoP = \frac{N_{SNPs} + N_{INDELs}}{Allele_{length}}$$

- the structure of the allele defined as follow:

$$start{:}\, stop\Delta startDel_1{:}\, stopDel_1 \ldots \mathrm{I}startIns_1{:}\, stopIns_1$$

$start$ and $stop$ are the most external position of the allele with at least the coverage required in input. Symbols $\Delta, \mathrm{I}$ indicate the presence of a deletion or an insertion of at least the INDEL length required in input, respectively. $stop\Delta startDel_x{:}\, stopDel_x$ and $startIns_y{:}\, stopIns_y$ are the coordinates for the start and the stop of the deletion $x$ and the insertion $y$. $start_{del}$, $stop_{del}$ and $start_{ins}$ refer to the position of the variant according to the reference genome. Instead, $stop_{ins}$ is an artificial value evaluated as $start\_ins + length(insertion)$.

## 4.2. Implementation of the SVD pipeline to WGS data from *Aedes albopictus*

WGS data from 16 mosquitoes of the Foshan strain were used for the analyses of nrEVEs polymorphism. Hereafter, I will call these samples Single Sample Mosquitoes or SSM.

DNA was extracted from each mosquito using the QIAGEN Blood and Tissue kit (Qiagen, Hilden, Germany) and it was sent to the "Polo D'Innovazione Genomica, Genetica e Biologia" (Siena, Italy) for DNA quality control, Illumina library preparation and WGS. Paired-end sequencing was done with Illumina HiSeq 2500 with an approximate 20x coverage.

WGS data were aligned to AaloF1 with 'bwa mem' with default settings [106]. The BAM files were analyzed with the Picard tool with the following functions and default settings: AddOrReplaceReadGroups, CleanSam, SortSam and MarkDuplicates [107]. This operation allows to add a single read group to all the reads in the BAM file, to soft-clip beyond-end-of-

reference alignments, setting mapping quality to 0 for unmapped reads, sort by coordinates the reads and identify duplicates.

Resulting BAM files and the BED file of the 72 reference nrEVEs of AaloF1 were used as input for the SVD pipeline. We set the parameter of the pipeline based on the depth of coverage of our WGS data and after established the minimum requirements to consider a locus and a variant present in a sample. nrEVE presence in a sample was established by imposing a minimum coverage of 5 reads with at least 20 Phred mapping quality and a minimum of 30 consecutive nucleotides with that depth of coverage. The following settings were used for the variant calling step of the pipeline: 20 as minimum Phred base and mapping quality, at least 8 reads as depth of coverage, at least 0.1 as allele fraction and 2 as minimum allele count.

The ratio between the number of R-NIRVS detected in a sample and the total R-NIRVSs annotated in the reference genome was used to estimate R-NIRVSs prevalence. The same calculation was done for F-NIRVS.

The hypergeometric test was applied to test whether the group of nrEVEs identified in each SSM was enriched in: 1. F- or R-NIRVSs; 2. any viral ORFs; 3. nrEVEs shorter or longer than 500 bp; 4. nrEVEs mapping in exons, piRNA clusters or intergenic regions.

Eleven nrEVEs (i.e., AlbFlavi19, AlbFlavi31, AlbFlavi32, AlbFlavi33, AlbFlavi38, AlbFlavi39, AlbFlavi40, AlbRha43, AlbRha79, AlbRha80, AlbRha95) were absent in all 16 SSMs. A total of 20 nrEVEs were found in all SSMs, with a statistical enrichment for R-NIRVS (Hypergeometric test, $p = 0.022$) and nrEVEs mapping in gene exons (Hypergeometric test, $p = 0.006$) (**Figure 13A**). These 20 nrEVEs constitute the core of *Ae. albopictus* nrEVEs and included R-NIRVSs identified within the coding sequence of genes (i.e., AlbRha12, AlbRha15, AlbRha28, AlbRha52, AlbRha85 and AlbRha9) and piRNA clusters (i.e., AlbRha14 and AlbRha36). Conversely, F-NIRVS were variably distributed among SSMs. Of note is AlbFlavi4, a 512bp sequence with similarity to the capsid gene of *Aedes flavivirus* [34]. AlbFlavi4 is annotated within the second exon of AALF003313 and is also included in piRNA cluster 95 [108]. AlbFlavi4 produces vepi4730383, a piRNA that is upregulated upon dengue infection [109]. In SSMs and *Ae. albopictus* geographic samples, variants were identified for AALF003313, only one of which includes AlbFlavi4 (**Figure 13B, C**).

Overall, mean base pairs (bp) occupied by F-NIRVSs and R-NIRVSs are 12095 and 19293 bp, respectively (**Figure 13D**).

Taken together, these results demonstrate that, with an average genome occupancy of 31389 bp, nrEVEs represent quantitatively a limited fraction of the mosquito repeatome. However, the enrichment of nrEVEs in piRNA clusters [34] and the fact that the pattern of nrEVEs is variable in host genomes support the hypothesis that nrEVEs are a dynamic component of the repeatome.

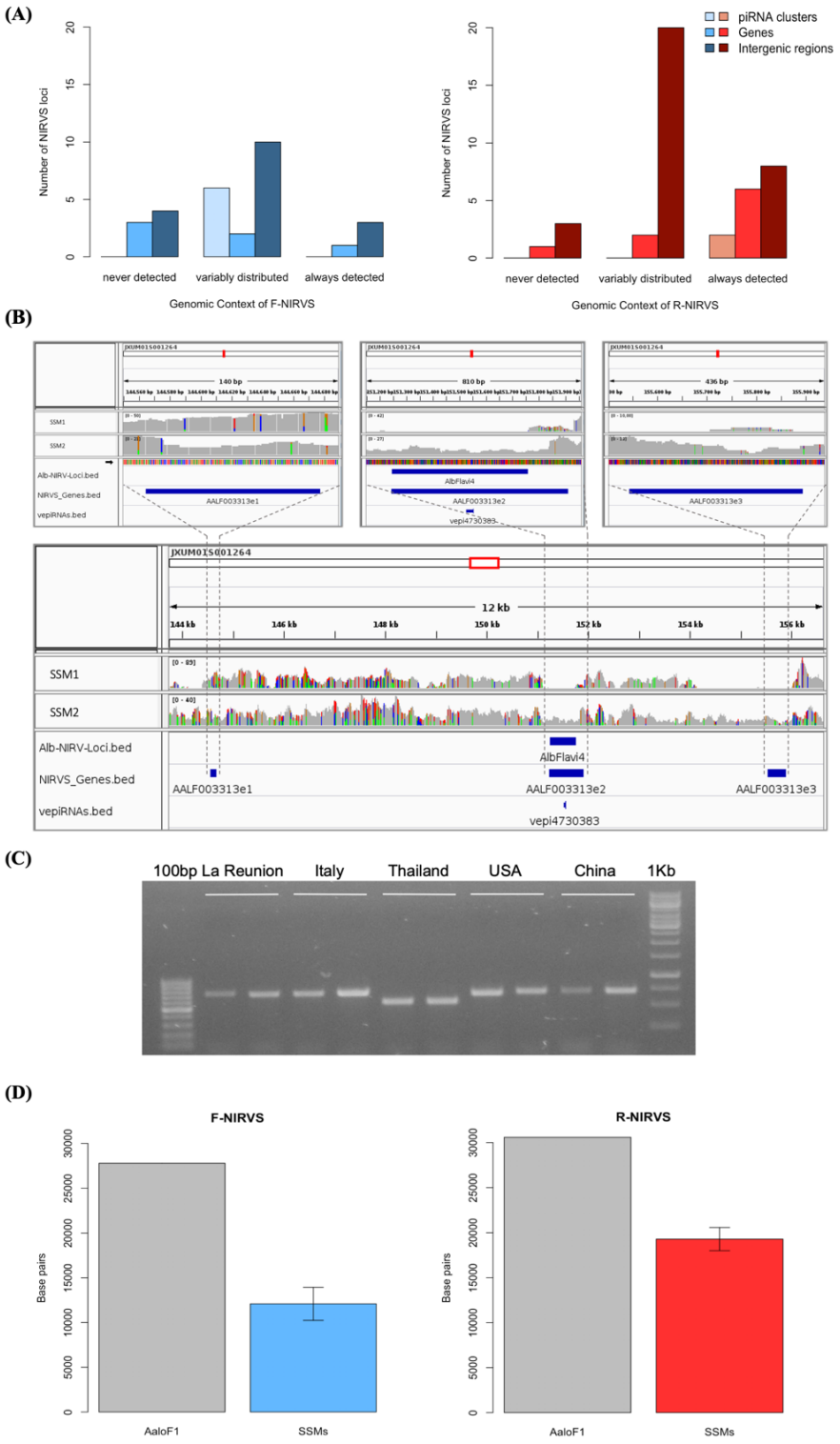


**Figure 13.** nrEVEs are variably distributed in the genome of 16 mosquitoes. **(A)** Number of F-NIRVS and R-NIRVS loci mapping within genes, piRNA clusters or intergenic regions, classified on the basis of read-coverage across SSMs. **(B)** Integrated Genomics Viewer (IGV) screen shot showing read coverage at AALF003313 in two mosquitoes, SSM1 and SSM2. Positions of the three exons annotated in AALF003313; the positions of AlbFlavi4 and vepi4730383 are indicated by blue bars. **(C)** PCR amplification of

AALF003313 exon2 in ten *Ae. albopictus* geographic samples. **(D)** Given their variable presence across the 16 tested mosquitoes, F-NIRVS and R-NIRVS loci occupancy (in base pairs) is about half of that expected based on the annotated sequences in AaloF1. F-NIRVS are in blue, R-NIRVS are in red.

## 4.3. Level of sequence polymorphism of reference nrEVEs

I implemented the SVD pipeline to nrEVEs and sets of selected genes using the same settings to compare their LoPs. LoP was evaluated as the ratio between the number of total mutations (SNPs and INDELs) found in the locus and its length.

Selected genes included: 1. genes of the RNA interference (RNAi) pathway, for which intraspecific rapid evolution has been observed in *Ae. aegypti* [110], 2. sets of slow and fast evolving genes as described below; 3. a total of 13 genes, which included nrEVEs in their CDS or untranslated regions (UTRs) [34]; these genes are called nrEVEs genes (N-Gs).

Genes of the RNAi pathway (R-Gs) included Ago1 (AALF020776), piwi6 (AALF016369), piwi1 and 3 (AALF005499, AALF005498), and Ago2 (AALF006056). Slow and fast evolving genes were identified through the following analysis. Orthologous genes across 27 insect species within the Nematocera sub-order, including *Ae. aegypti,* but not *Ae. albopictus*, were identified in OrthoDB v9.1 [111]. Levels of sequence divergence were computed for each orthologous group as the average of interspecies amino acid identified. These values were then normalized to the average identity of all interspecies best-reciprocal-hits, computed from pairwise Smith-Waterman alignments of protein sequences. These levels of sequence divergency were plotted and the genes of *Ae. aegypti* showing the 0.1% levels at each tail of the distribution (n = 14, number comparable to genes with nrEVEs in their CDS or UTRs, see below) were selected as representative of the conserved (left tail) and variable (right tail) gene sets, respectively. Orthologs of these genes were identified in AaloF1 and their single-copy status verified.

Conserved genes were also called slow evolving genes (SGs). In *Ae. albopictus*, SGs included genes with hypothetical protein transporter or vesicle-mediated transport activity (i.e., AALF003606, AALF014156, AALF014287; AALF014448; AALF004102), structural activity (AALF005886, annotated as tubulin alpha chain), signal transducer activity (AALF026109), protein and DNA binding activity (AALF027761, AALF028431), SUMO transferase activity (AALF020750), the homothorax homeobox encoding gene AALF019476, the tropomyosin invertebrate gene (AALF0082224), the Protein yippee-like (AALF018378) and autophagy (AALF018476). The variable genes were called fast evolving genes (FGs). In *Ae. albopictus*, FGs include genes with unknown functions (AALF004733, AALF009493, AALF009839, AALF012271, AALF026991,

AALF014993, AALF017064, AALF018679), proteolysis functions (AALF010748) a gene associated with transcriptional (AALF022019), DNA-binding (AALF019413, AALF024551), structural (AALF028390) and proteolytic (AALF010877) activities.

The median LoP of SGs within mosquitoes of the Foshan strain is 0.0071, a value higher than that observed across 63.3% of the detected nrEVEs. In particular, 18 out 23 detected F-NIRVS (78%) and 20 out 38 detected R-NIRVS (52%) have LoPs lower than the median LoP of the SGs (**Figure 14**).

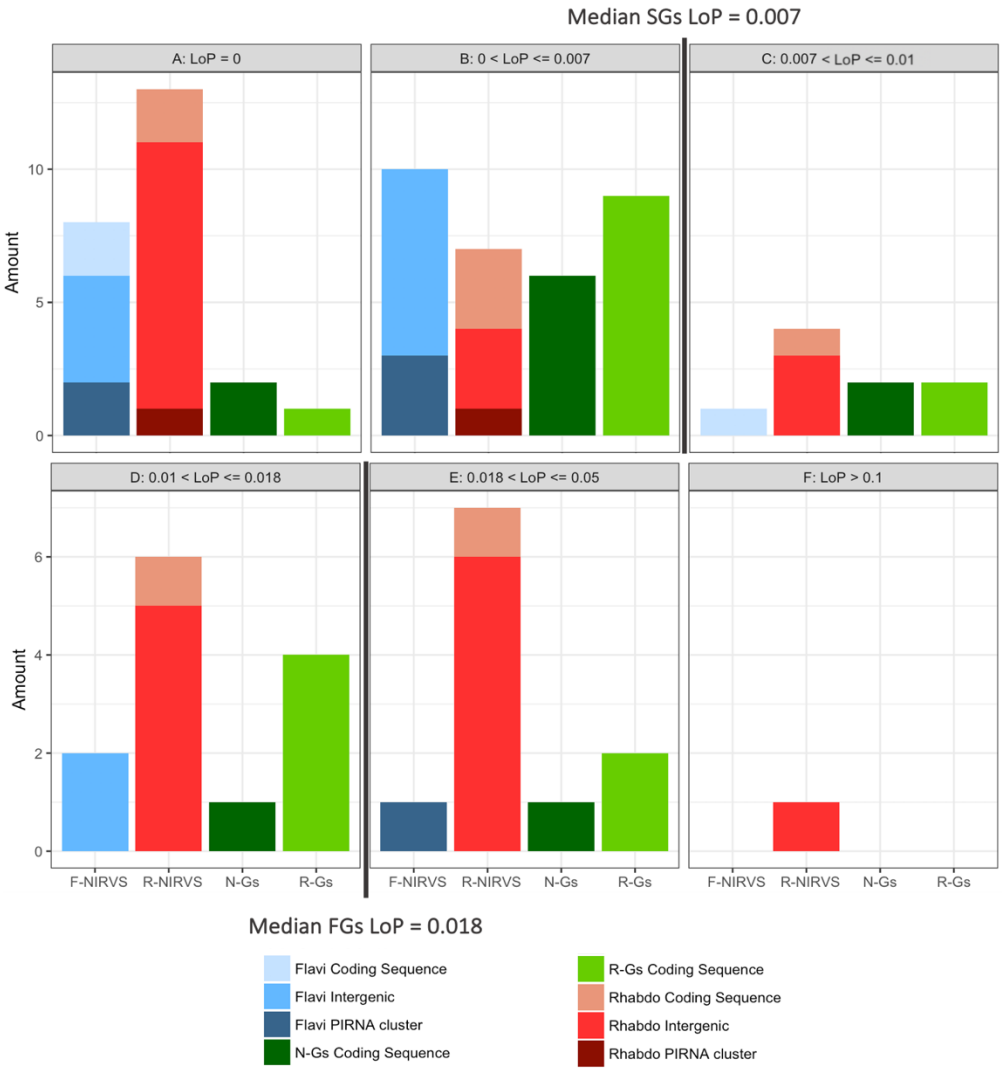


**Figure 14.** Distribution of nrEVEs, N-Gs and R-Gs based on their polymorphism levels (LoP). Grey lines are median LoP values of SGs and FGs. F-NIRVS are blue, R-NIRVS are red, N-Gs are dark green, R-Gs are light green. Within F-NIRVS and R-NIRVS groups, shades of colors are used to highlight nrEVEs mapping in exons of annotated genes, piRNA clusters or intergenic regions. A-F refer to different LoP classes: A) LoP is zero; B) LoP between 0 and 0.007; C) LoP between 0.007 and 0.01; D) LoP between 0.01 and 0.018; E) LoP between 0.018 and 0.05; F) LoP higher than 0.1.

I used the Kolmogorov-Smirnov test, from the 'stat' package of R studio to test the significance of the difference in the LoP distributions of nrEVEs, R-Gs, N-Gs and FGs with respect to that of SG LoP, [112]. SG LoP was the median of the LoPs of the tested SGs. The threshold of significance of 0.05 was adjusted with the Bonferroni correction and loci were separated according to the adjusted significance of the test. Results of ratio between the LoP of each locus and the median LoP of SGs (fold change [FC]) that were different from 0 were visualized in **Figure 15**.

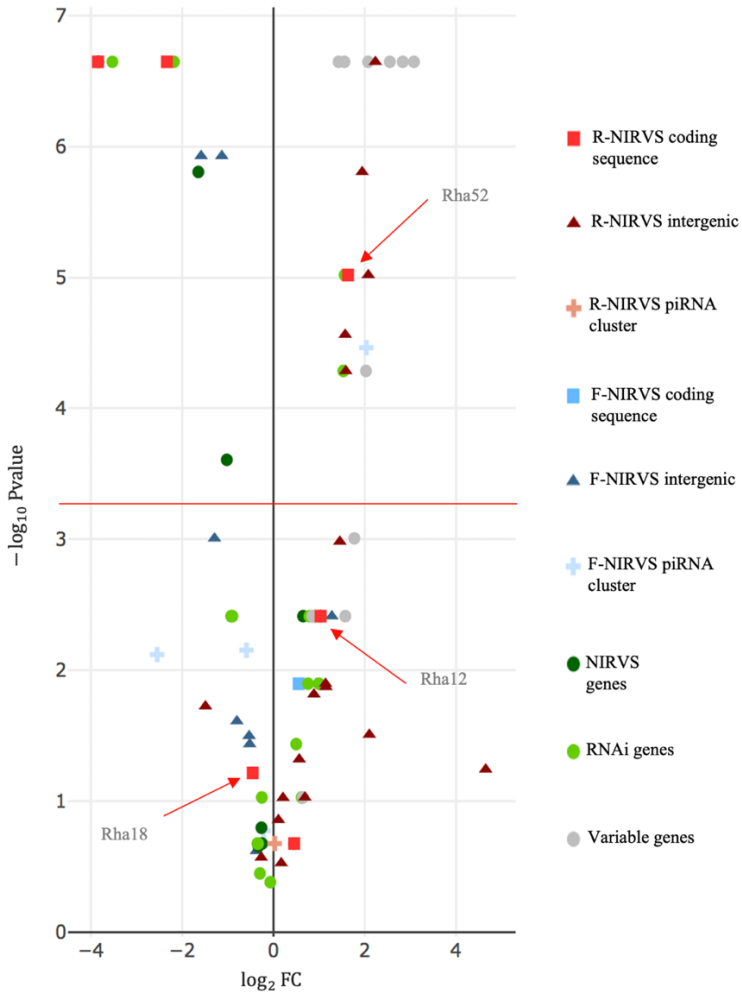


**Figure 15.** nrEVE sequence polymorphism. Volcano plot showing LoP comparison between SGs and nrEVEs, N-Gs [34], R-Gs and FGs. Entities with LoPs statistically different than that of SGs are above the red line [adjusted significance of the test (–log10 0.05/99 = 3.32)]. Entities on the left side of the panel (log2FC < 0) have smaller LoPs than SGs. The opposite for entities on right side of the panel (log2FC > 0).

Eleven out of fourteen FGs were more variable than SGs, with seven appearing also statistically more polymorphic than SGs. This result further supports our selection of SGs and FGs. R-Gs are heterogeneously

polymorphic. Ago1 (AALF020776) and piwi6 (AALF016369) are statistically more polymorphic than SGs. The opposite result was obtained for piwi1 and 3 (AALF005499, AALF005498), and Ago2 (AALF006056) (**Figure 15**).

nrEVEs identified within piRNA clusters [108] are all less polymorphic than SGs, with the exception of AlbFlavi12-17 that has a median LoP value of 0.0258. This large LoP may be due to the fact that AlbFlavi12-17 is composed of four small viral sequences nested one next to the other [34]. Unlike nrEVEs in piRNA clusters, nrEVEs spanning gene exons are more heterogeneous in their LoPs. Three (i.e., AlbFlavi34, AlbRha12, and AlbRha52) have LoP values higher than those of SGs, while others (i.e., AlbFlavi24, AlbRha28, AlbRha85) are less polymorphic than SGs. AlbFlavi24, AlbFlavi34, AlbRha12, and AlbRha28 are annotated as the only exons of AALF023281, AALF005432, AALF025780, AALF000478, respectively.

## 4.4. Testing the expression of nrEVEs annotated in coding sequences

The observed LoP of AALF020122, which contains AlbRha52, AALF025780, in which AlbRha12 was annotated and AALF005432, in which AlbFlavi34 was annotated, is analogous to that of rapidly evolving genes, suggesting co-option for immunity functions [32].

Because domestication of exogenous sequences is a multi-step process, including persistence, immobilization and stable expression of the newly acquired sequences besides rapid evolution [113], we analyzed the distribution and expression pattern of these genes. Expression analyses were extended to all other N-Gs (AALF025779 with a unique exon containing AlbRha9, AALF000476 with a unique exon corresponding to AlbRha15, AALF000477, and AALF004130 in which the unique exons are contained within AlbRha18 and AlbRha85, respectively) that were identified in all tested SSMs, but have LoP levels comparable to or lower than those of SGs (**Figure 15**).

AlbFlavi34 had been previously studied and showed to be expressed in pupae and adult males more than in larvae [34]. N-Gs form two groups of paralogs, with similarity to the *Rhabdovirus* RNA-dependent RNA polymerase (RdRP) and the nucleocapsid-encoding gene (N protein), respectively (**Table 1**). As shown in **Figure 16**, apart from AALF00477, all other genes are expressed throughout *Ae. albopictus* development with a similar profile, but at different levels. None of the genes showed sex-biased expression or tissue-specific expression in the ovaries; on the contrary highest expression was observed in sugar- and blood-fed females.

**Table 1**. Characteristics of genes with nrEVEs in their coding sequence.

| Gene ID | nrEVE | Viral ORF | PfamID | Median LoP |
|---|---|---|---|---|
| AALF000476 | AlbRha15 | *Rhabdovirus* N protein | PF00945 | 0.0086 |
| AALF000477 | AlbRha18 | *Rhabdovirus* N protein | PF00945 | 0.0052 |
| AALF000478 | AlbRha28 | *Rhabdovirus* N protein | PF00945 | 0.0004 |
| AALF025780 | AlbRha12 | *Rhabdovirus* N protein | PF00945 | 0.0129 |
| AALF025779 | AlbRha9 | *Rhabdovirus* N protein | PF00945 | 0.0031 |
| AALF004130 | AlbRha85 | *Rhabdovirus* RdRP | PF00946 | 0.0020 |
| AALF020122 | AlbRha52 | *Rhabdovirus* RdRP | PF00946 | 0.0196 |

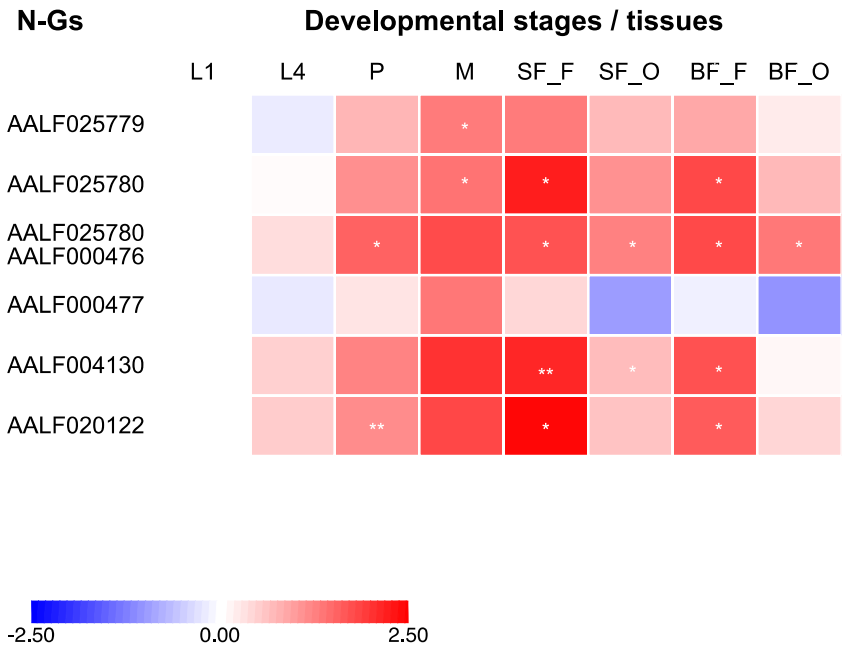


**Figure 16.** Expression of genes containing nrEVEs. Heatmap of the expression profiles of N-Gs across developmental stages and body tissues. L1-L4: 1st-4th instar larvae; P: pupae; M: male whole body; SF_F/BF_F: sugar/blood fed female whole body; SF_O/BF_O: ovaries from sugar/blood fed females. Color key expresses the log10-fold change relative to larva 1st instar (calibrator). Asterisks indicate significant differences in transcript abundances (Unpaired two-tailed t-tests, ∗P < 0.05, ∗∗P < 0.01).

## 4.5. Estimates of nrEVEs integration time

The higher prevalence of R-NIRVS with respect to F-NIRVS suggests R-NIRVS are older integrations (**Figure 17**).

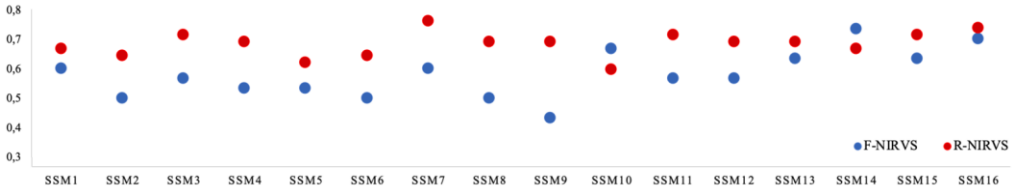


**Figure 17.** Prevalence of R-NIRVS (red) and F-NIRVS (blue) in each SSMs. Prevalence was calculated in each sample as the ration between detected nrEVEs and reference nrEVEs for both R- and F-NIRVS.

To verify this hypothesis, the presence and the polymorphism of a selected group of nrEVEs was evaluated in native populations from China and Thailand, old populations from La Reunion Island and invasive populations from United States and Italy by PCR with nrEVE-specific primers (**Appendix 4**) [114]. At least five amplification products per population per locus were sequenced with Sanger (see Material and Methods of Pischedda et al., 2019 [100]). This part of the analyses was conducted by biologists working in the Bonizzoni's Lab.

Seven F-NIRVS (AlbFlavi2, AlbFlavi4, AlbFlavi8-41, AlbFlavi10, AlbFlavi36, AlbFlavi1, and AlbFlavi12-17) and six R-NIRVS (AlbRha1, AlbRha7, AlbRha14, AlbRha36, AlbRha52, AlbRha85) were selected based on their unique occurrence in different regions of the mosquito genome and their similarity to various viral ORFs. AlbRha52 and AlbRha85 are annotated as unique exons of AALF020122 and AALF004130, respectively.

nrEVEs sequences from geographic samples were aligned in Ugene, version 1.26.1 [115] with MAFFT [116]. Default parameters with five iterative refinements were applied for the alignment. Alignments were manually curated to verify frameshifts, truncations, deletions, and insertions. All positions including gaps were filtered out from the analysis. The following formula was used to estimate the time of integration in years assuming that all mutations are neutral:

$$Mean\ Mutation_{seq} = \frac{Tot\ Mutation}{N_{seq} * Length_{seq}}$$

$$Age\ in\ Years_{seq} = \frac{Mean\ Mutation_{seq}}{MR * GpY}$$

Mutation rates (MR) were assumed to be comparable to those of *D. melanogaster* genes, in range of 3.5-8.4 × $10^{-09}$ [117], [118]. A range of 4-

17 number of Generations per Year (GpY) was tested considering mosquitoes of temperate or tropical environments [114].

Under these conditions, R-NIRVS integrated between 36 thousand and 2.7 million years ago (mya) and F-NIRVS between 7.4 thousand and 2.4 mya (**Figure 18**).

This large window supports the conclusion that integration of viral sequence is a dynamic process occurring occasionally at different times. As shown in **Figure 18**, estimates of integration times varied greatly depending on the genomic context of nrEVEs. nrEVEs annotated within gene exons appear statistically more recent than nrEVEs of piRNA clusters (ANOVA, ***P < 0.001). Besides reflecting a different integration time, this result is consistent with the hypothesis that integrations within exons are under rapid evolution, a hallmark of domestication [32].

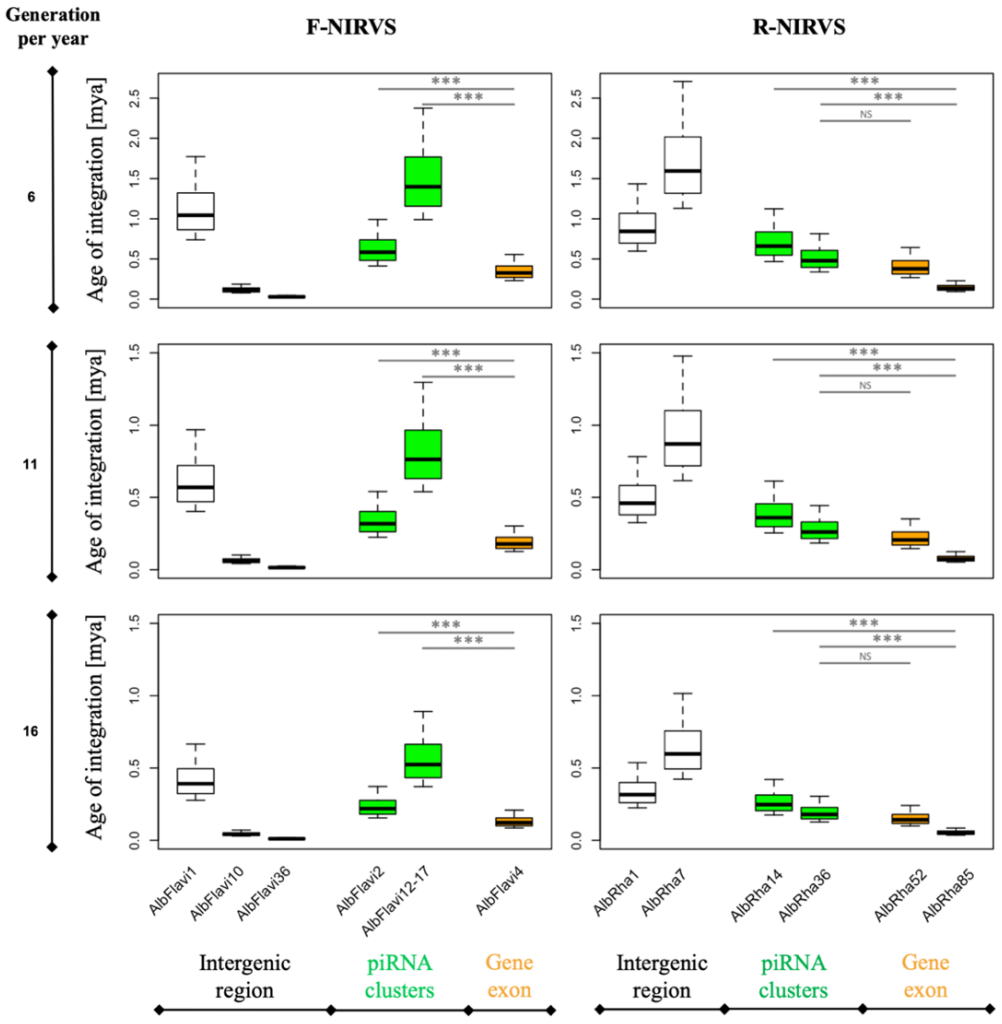


**Figure 18.** nrEVEs integration time. Boxplots showing the integration times for the nrEVEs whose variability was studied across five geographic populations. Estimates are based on the *D. melanogaster* mutation rate, i.e.,

3.5–8.4 × $10^{-9}$ per site per generation [117], [118], and a range of generations per year between 6 and 16 to include mosquitoes from temperate and tropical regions [114]. nrEVEs of piRNA clusters are statistically older than nrEVEs mapping in gene exons (ANOVA, $^{***}P < 0.001$).

## 4.6. Concluding remarks on the analysis of nrEVEs polymorphism in *Aedes albopictus*

Here I used both bioinformatic and molecular approaches to study the polymorphism of nrEVEs in SSMs of *Ae. albopictus* from the Foshan strain and from different geographical areas.

The 16 Foshan mosquitoes derive from a strain obtain by the collection of samples in the Chinese city of Foshan in the early 1980 and maintained in laboratory without viral exposure [40]. Their analysis highlighted a variable landscape of nrEVEs with a core set of nrEVEs from *Rhabdoviruses* and nrEVEs mapping in CDSs. This result demonstrates that viral integrations are a dynamic component of the repeatome and not all viral integrations are dispensable genomic elements.

nrEVEs were compared to FGs and SGs in mosquitoes. I found that nrEVEs identified within piRNA clusters were less polymorphic than SGs. Despite piRNAs have an incredible sequence diversity, selection constraints on sequences within piRNA clusters have been previously identified in both flies and mice [119].

To start to investigate the widespread occurrence of nrEVEs in natural population we selected a set of 13 viral integrations representative of both R- and F-NIRVS and mapping within piRNA clusters, intergenic regions and gene exons. Based on the invasion history of *Ae. albopictus* we included samples from the native home range of the species, China, Thailand and La Reunion island and samples from newly colonized areas such as Italy and United States.

We confirmed the variable landscape of nrEVEs across geographic populations and higher of R-NIRVS with respect to F-NIRVS. This result should be interpreted with caution. In fact, in newly invaded areas the mosquito populations present lack of isolation by distance and appear genetically mixed [114], [120], [121]. Thus, the variability of nrEVEs can be only partially explained by the occurrence of frequent bottlenecks followed by interbreeding. However, the enrichment for R-NIRVS, the variable distribution of nrEVEs within piRNA clusters and their heterogenous polymorphism indicate that evolutionary forces other than genetic drift and gene flow have played a role in the distribution of nrEVEs and suggests a multifaceted impact of nrEVEs on mosquito physiology.

Further, we noticed that R-NIRVS appeared older integrations than F-NIRVS. This last aspect is particularly intriguing because Mononegavirales, including *Rhabdoviruses*, are considered evolutionary more recent than

*Flaviviridae* [122]. This discovery opens to hypothesis: 1. because *Rhabdoviruses* have been shown to frequently transfer horizontally among host species based on their ecological and geographic proximity [123], thus the wide geographic distribution range of *Rhabdoviruses* may favor their integrations into mosquito genomes; 2. the frequent horizontal transfers of *Rhabdoviruses* could determine the emergence of generalist protection mechanisms, of which integrations could be part of.

Overall, my results emphasize the complexity of the composition and structure of the mosquito repeatome and provide an objective strategy to identify viral integrations that most probably affect mosquito biology.

# Chapter 5

# Detection of novel viral integration

I developed the ViR pipeline to ameliorate detection of novel viral integrations in organisms with a repetitive and/or fragmented genome sequence. The pipeline was made public to the research community through GitHub (https://github.com/epischedda/ViR).

## 5.1. Challenges in the detection of novel viral integrations

The transfer of genetic material between separate evolutionary lineages is a recognized event that occurs not only among prokaryotes, but also between viruses and eukaryotic cells [124]. Somatic integrations of different viral species, among the best of known of which are the human papilloma virus, hepatitis B and C viruses and the Epstein-Barr virus, have been linked to genotoxic effects possibly progressing into cancer [46]. Consequently, several pipelines have been developed to identify viral sequences integrated into the human genome using whole-genome sequencing (WGS) data [47].

Recent genomics and metagenomics analyses have shown that viruses also integrate into the genome of non-model organisms (i.e., arthropods, fish, plants, vertebrates) [39], [42], [43], [61], [125], [126]. However, in non-model organisms, studies of viral integrations have rarely gone beyond their annotation in reference genome assemblies. Additionally, in non-model organisms, we lack a thorough understanding of the widespread occurrence of nrEVEs and their biological relevance, apart from sporadic cases which nevertheless point to significant roles of nrEVEs in immunity and regulation of expression [75]. The absence of bioinformatic tools able to account for intrasample variability (i.e., repetitive DNA, duplications and/or assembly fragmentation) when mapping WGS data to a reference genome is hindering our ability to detect integration sites different than those already annotated in a reference genome assembly, thus to understand the widespread

occurrence and polymorphism of EVEs in host genomes and testing hypothesis on their biological function using WGS data collected under hypotheses-driven experimental conditions.

To ameliorate this issue, I developed ViR. As such ViR is not a new tool to detect viral integrations, rather it works downstream of any tool for identification of viral integrations based on paired-end reads to solve intrasample variability and ameliorate predictions of integration sites.

## 5.2. The Vy-PER pipeline

Among the pipelines available to identify viral integrations using WGS data, I chose Vy-PER for the accuracy of its results, low computational requirements and the possibility to test more viral genomes simultaneously [51]. Vy-PER workflow is shown in **Figure 19**.

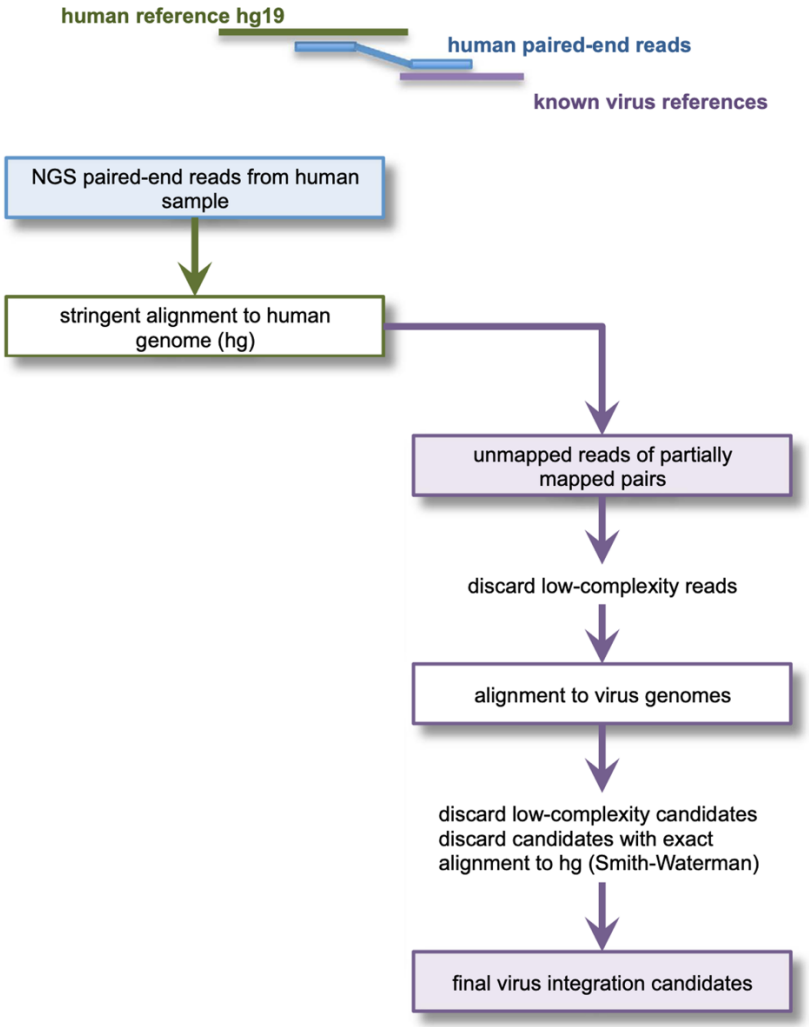


**Figure 19.** Vy-PER pipeline for the viral integration detection [51].

Briefly, the paired-end reads are first aligned to the reference genome using the BWA tool with '*aln + sampe'* functions [127]. The resulting Sequence Alignment Map (SAM) file is converted into sorted BAM format using SAMtools [128]. Chimeric read pairs in which one read maps to the reference genome and the other one does not, are extracted using the 'view' function of SAMtools [128]. The read that maps to the reference genome is referred to as host read. Among the selected unmapped reads, low-complexity reads are discarded using Phobos [129]. The *vyper_sam2fas_se* script is used to remove reads in which the main non-STR region is shorter than 30 bp (default and used threshold). The remaining unmapped reads are aligned with BLAT [130] to a user-defined viral genome database. Only the top 3 virus candidates per integration site are retained. Since a partial length mapping to a viral genome may be the result of a STR or of a homopolymer, low-complexity reads are discarded in a fourth step. Finally, the *vyper_final_filtering* custom script refines the output including the final virus integration candidates. Vy-PER output includes:

- an ideogram plot in PDF format giving a summary of candidate loci and virus types (only for human reference genome);
- a table of the top 10 virus candidates;
- a table of the clusters (genomic windows, number of candidates, virus name and NCBI ID);
- a table of phiX174 chimeras per chromosome/scaffolds/contigs;
- a detailed table of unfiltered virus candidates;
- FASTA files for each virus candidate for optional manual alignment/checking.

Vy-PER was built within the framework of human cancer genetics, thus it works perfectly with the well-assembled and annotated human genome.

With the collaboration of Engenome srl, a spin-off of the University of Pavia, the Vy-PER pipeline was written with Cosmos [131], a python library for the parallelization of the processes in the Linux environment.

## 5.3. The ViR pipeline

ViR works downstream of Vy-PER and any other EVE prediction tool that uses paired-end reads. ViR is composed of four scripts, which work in two modules (**Figure 20**). The first module includes three scripts, ViR_RefineCandidates, ViR_SolveDispersion and Vir_AlignToGroup, which work together to overcome the dispersion of reads due to intrasample variability. The second module includes one script, ViR_LTFinder, designed to test for a lateral transfer (LT) events of non-host sequences which have none or limited sequence similarity to sequences of the host.

Default input parameters of each script are presented in brackets. Their values were selected based on the coverage of the WGS data available in

laboratory in order to distinguish novel nrEVEs from already annotated nrEVEs. The user can modify parameter values based on its own data.

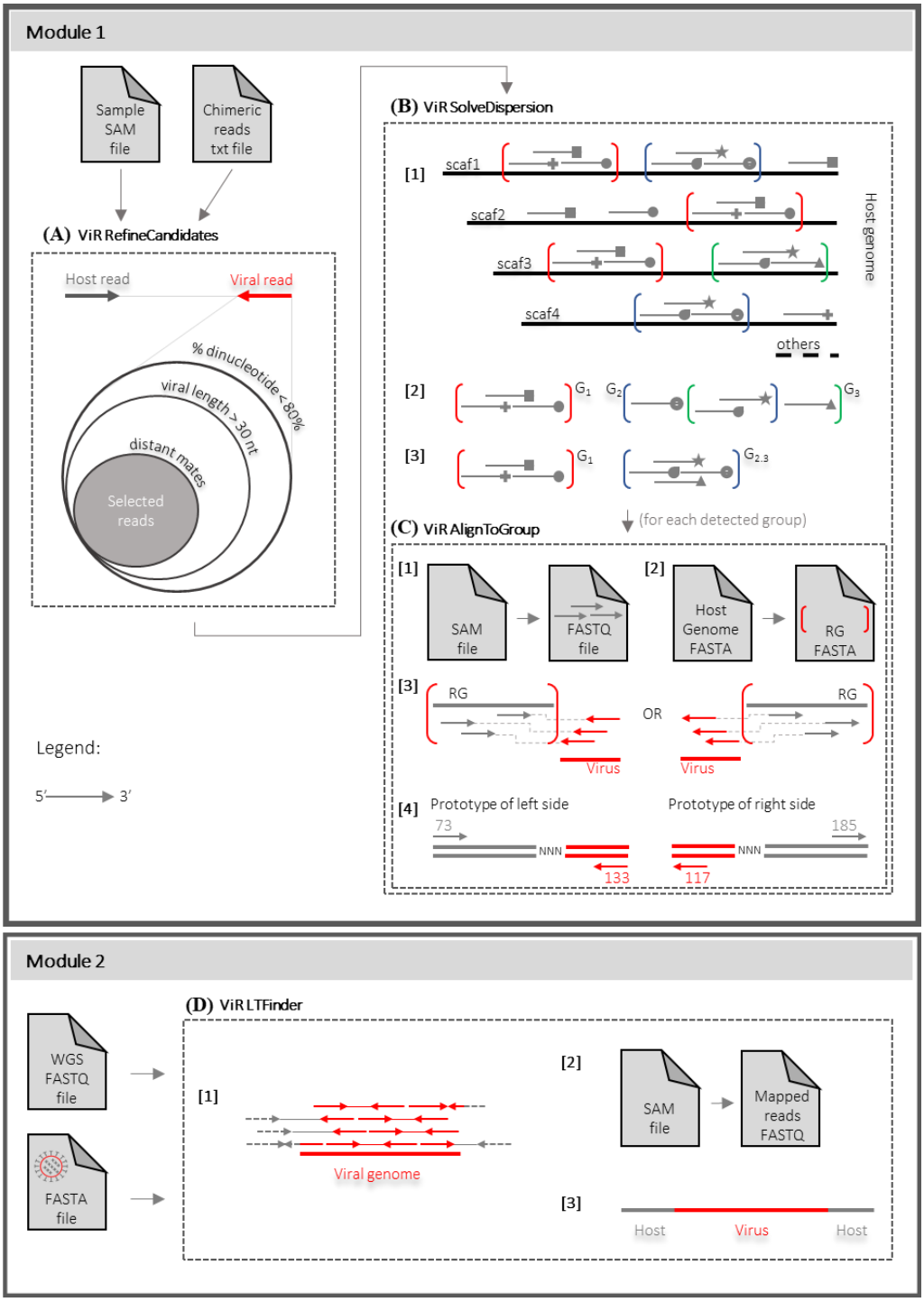


**Figure 20.** Overview of ViR.

The *ViR_RefineCandidates* script selects from a list of chimeric reads the best candidate pairs supporting a viral integration by filtering reads based on their sequence complexity, expressed as percentage of dinucleotides (default

< 80%), imposing a minimum length recognized as viral (default 30 nucleotides) and removing mates that can align within a defined window in the reference genome (default 10000). For this last filter, host and viral reads are aligned with blastn [132] (default evalue 1e-15) on the reference genome independently. The coordinates of the alignments are converted in the BED format and finally host and viral reads coordinates are compared with the function "closest" of BEDtools [88] package (**Figure 20A**).

The *ViR_SolveDispersion* script solves the dispersion of host reads by grouping together reads that map to regions of the genome with the same sequence (**Figure 20B**). These reads are called "read groups"; regions of the genome to which these reads can equivalently map because they contain the same repetitive element, or it is a sequence that has been erroneously assembled into different contigs or scaffolds, are called "equivalent regions". The script acquires as inputs a file listing all samples to analyze and the output directory of ViR_RefineCandidates. Reads mapping within equivalent regions are grouped together using the function "merge" from BEDtools [88] (default maximum merge distance 1000 nt) (**Figure 20B**, step 1). Then, identified read groups are compared in a pairwise mode in an iterative process in which read groups sharing more than a user-defined percentage of reads are collapsed in one (default is 80%) (**Figure 20B**, step 2). This procedure allows to identify the best candidate anchor genomic region of a candidate viral integration site (**Figure 20B**, step 3).

The *ViR_AlignToGroup* script predicts the right and left sides of the integration site by realigning reads supporting each candidate viral integration against the sequence of the equivalent region. First, for each candidate, this script extracts the host reads with their viral pair from the SAM file of the reads analyzed by ViR_RefineCandidates using the command-line utility "grep" (https://www.gnu.org/software/grep/manual/grep.html); the SAM file is converted into a BAM file using the function "view" of SAMtools [128]; the BAM file is converted into a FASTQ format using the function "bamtofastq" from BEDtools [88] (**Figure 20C**, step 1). Then, the script obtains the sequence of the equivalent region in fasta using the BEDtools function "getfasta" [88] (**Figure 20C**, step 2). Reads from step 1 are re-aligned to the sequence of the equivalent region using "bwa mem" with default parameters [106]. By taking advantage of the flags of alignment of each read of all chimeric pairs and eventual soft clipped reads, the left and right sides of the integration point can be predicted using Trinity (**Figure 20C**, step 3). Even if no assemblies are created, flags of alignment are used to predict the direction of the integration sites (https://broadinstitute.github.io/picard/explain-flags.html) (**Figure 20C**, step 4).

The *ViR_LTFinder* script is designed to test for an integration from non-host sequences which have a user-defined percentage of similarity to host sequences. WGS reads are mapped to a selected non-host sequence (i.e., an entire genome or selected portions) using BWA tool +"mem" function with default parameters [106]. Mapped reads are extracted using the function

"view" of SAMtools [128] (**Figure 20D**, step 1). The aligned reads are converted into FASTQ format using the function "bamtofastq" from BEDtools [88] (**Figure 20D**, step 2) and used for *de-novo* local assembly using Trinity [133] (**Figure 20D**, step 3). A consensus sequence is built if any instances of LT are identified. Output of ViR_LTFinder include files for visualization of the aligned reads using the Integrated Genomics Viewer (IGV) tool [134].

## 5.4. Evaluation of ViR performance using *in silico* dataset

The ViR pipeline was implemented with simulated dataset to evaluate its performances. Both single samples and pool sequencing data were used. The performances of module 1 and module 2 were tested separately considering:

- different sequencing coverage depths (5, 15, 30, 45, 60);
- different integration sizes (300 bp, 600 bp and 900 bp);
- different integration events in unique genomic loci (UL) or repeated genomic regions (repeated 10 and 100 times);
- for pools, different pool sizes were also analyzed (pools of 10, 30 and 50 individuals).

**Table 2.** Confusion matrix structure.

| | **Absent Integration (Negative)** | **Present Integration (Positive)** |
|---|---|---|
| **Predicted Absent Integration (Negative)** | TN | FN |
| **Predicted Present Integration (Positive)** | FP | TP |

ViR performance was computed based on the confusion matrix in **Table 2**. The confusion matrix shows in column the real values that need to be classified and in rows the predicted values by a method of classification.

The values of the matrix include: TN (True Negative): the number of events correctly identified as not integration; FP (False Positive): the number of integration events erroneously identified; TP (True Positive): the number of integration events correctly identified; FN (False Negative): the number of integration events missed by the tool.

The following performance parameters were evaluated [135], [136]:

$$ACC^1: Accuracy = \frac{TP+TN}{TP+TN+FN+FP}$$

$$SENS^{2}: Sensitivity = \frac{TP}{TP+FN}$$

$$SPEC^{3}: Specificity = \frac{TN}{TN+FP}$$

$$PREC^{4}: Precision = \frac{TP}{TP+FP}$$

$$F1^{5}: F1 = \frac{2TP}{2TP+FP+FN}$$

$$\text{BACC}^{6}: Balanced\ Accuracy = \frac{Sensitivity + Specificity}{2}$$

$$MCC^{7}: Matthews\ Correlation\ Coefficient = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

**Table 3.** ViR module 1 performances in single samples.

| | | ACC[1] | SENS[2] | SPEC[3] | PREC[4] | F1[5] | B ACC[6] | MCC[7] |
|---|---|---|---|---|---|---|---|---|
| | ALL | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| | Cov5 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| | Cov15 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| | Cov30 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| | Cov45 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| **Cov.** | Cov60 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| | INT0 | 100 | ND | 100 | ND | ND | ND | ND |
| | 300-INT0 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| | 600-INT0 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| **Int. size** | 900-INT0 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| | UL | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| | T10 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| **Int. site** | T100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |

**Table 4.** ViR module 2 performances in single samples.

| | | ACC[1] | SENS[2] | SPEC[3] | PREC[4] | F1[5] | B ACC[6] | MCC[7] |
|---|---|---|---|---|---|---|---|---|
| | ALL | 93,75 | 100 | 78,94 | 91,83 | 95,74 | 89,47 | 85,14 |
| | Cov5 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| | Cov15 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| | Cov30 | 85,71 | 100 | 60 | 81,81 | 90 | 80 | 70,06 |
| | Cov45 | 92,30 | 100 | 75 | 90 | 94,73 | 87,5 | 82,15 |
| **Cov.** | Cov60 | 92,30 | 100 | 75 | 90 | 94,73 | 87,5 | 82,15 |
| | INT0 | 100 | ND | 100 | ND | ND | ND | ND |
| | 300-INT0 | 93,75 | 100 | 88,23 | 88,23 | 93,75 | 94,11 | 88,23 |
| | 600-INT0 | 93,75 | 100 | 88,23 | 88,23 | 93,75 | 94,11 | 88,23 |
| **Int. size** | 900-INT0 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| | UL | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| | T10 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| **Int. site** | T100 | 83,33 | 100 | 55,55 | 78,94 | 88,23 | 77,77 | 66,22 |

Module 1 showed 100% sensitivity and 100% specificity in all cases with SSMs (**Table 3**). Across all tested conditions in SSMs, the performance of module 2 reached an overall accuracy and specificity of 93,75% and 78,94%, respectively. This result was driven by the situation in which the integration site occurred in a highly repeated (100 times) genomic sequence. This event affects the generation of *de novo* assemblies giving a ViR accuracy of 83,33% and a specificity of 55,55% (**Table 4**).

**Table 5.** ViR module 1 performances in pools.

| | | ACC[1] | SENS[2] | SPEC[3] | PREC[4] | F1[5] | B ACC[6] | MCC[7] |
|---|---|---|---|---|---|---|---|---|
| | ALL | 65,74 | 54,32 | 100 | 100 | 70,4 | 77,16 | 47,87 |
| | Cov30 | 55,55 | 40,74 | 100 | 100 | 57,89 | 70,37 | 38,29 |
| | Cov45 | 69,44 | 59,25 | 100 | 100 | 74,41 | 79,62 | 51,63 |
| **Cov.** | Cov60 | 72,22 | 62,96 | 100 | 100 | 77,27 | 81,48 | 54,61 |
| | INT0 | 100 | ND | 100 | ND | ND | ND | ND |
| | 300-INT0 | 64,81 | 29,62 | 100 | 100 | 45,71 | 64,81 | 41,70 |
| | 600-INT0 | 81,48 | 62,96 | 100 | 100 | 77,27 | 81,48 | 67,78 |
| **Int. size** | 900-INT0 | 85,18 | 70,37 | 100 | 100 | 82,60 | 85,18 | 73,67 |
| | POOL10 | 94,44 | 92,59 | 100 | 100 | 96,15 | 96,29 | 87,03 |
| | POOL30 | 63,88 | 51,85 | 100 | 100 | 68,29 | 75,92 | 46,05 |
| **pool size** | POOL50 | 38,88 | 18,51 | 100 | 100 | 31,25 | 59,25 | 23,18 |
| | UL | 80,55 | 74,07 | 100 | 100 | 85,10 | 87,03 | 64,54 |
| | Rep10 | 61,11 | 48,14 | 100 | 100 | 65 | 74,07 | 43,40 |
| **Int. site** | Rep100 | 55,55 | 40,74 | 100 | 100 | 57,89 | 70,37 | 38,29 |

**Table 6.** ViR module 2 performances in pools.

| | | ACC[1] | SENS[2] | SPEC[3] | PREC[4] | F1[5] | B ACC[6] | MCC[7] |
|---|---|---|---|---|---|---|---|---|
| | ALL | 53,33 | 33,33 | 100 | 100 | 50 | 66,66 | 36,11 |
| | Cov30 | 38,70 | 13,63 | 100 | 100 | 24 | 56,81 | 20,93 |
| | Cov45 | 58,06 | 40,90 | 100 | 100 | 58,06 | 70,45 | 40,90 |
| **Cov.** | Cov60 | 64,28 | 47,36 | 100 | 100 | 64,28 | 73,68 | 47,36 |
| | INT0 | 100 | ND | 100 | ND | ND | ND | ND |
| | INT300 | 61,70 | 10 | 100 | 100 | 18,18 | 55 | 24,49 |
| | INT600 | 78 | 52,17 | 100 | 100 | 68,57 | 76,08 | 60,88 |
| **Int. size** | INT900 | 72,34 | 35 | 100 | 100 | 51,85 | 67,5 | 48,60 |
| | POOL10 | 94,73 | 90 | 100 | 100 | 94,73 | 95 | 90 |
| | POOL30 | 57,14 | 42,30 | 100 | 100 | 59,45 | 71,15 | 39,83 |
| **pool size** | POOL50 | 27,77 | 3,70 | 100 | 100 | 7,14 | 51,85 | 9,75 |
| | UL | 50 | 26,31 | 100 | 100 | 41,66 | 63,15 | 32,08 |
| | Rep10 | 58,06 | 40,90 | 100 | 100 | 58,06 | 70,45 | 40,90 |
| **Int. site** | Rep100 | 51,61 | 31,81 | 100 | 100 | 48,27 | 65,90 | 34,54 |

When using pools, the sensitivity and the accuracy greatly varied with: 1. the size of the pool, pools of 50 mosquitoes had accuracy and sensitivity < 50% in both modules; 2. the sequencing coverage (with a 30X coverage accuracy was 55% and 38% in module 1 and 2 respectively, but increased to 72%, in module 1, and 64%, in module 2, for a 60X coverage); 3. the length and the site of the integration event. Integrations shorter than 300 bp will not be able to be sensitively detected overall in case of a highly (>100 times) repeated genomic sequence (**Table 5**, **Table 6**).

## 5.5. Implementation of ViR with WGS data from *Aedes albopictus*

WGS data were generated from wild-collected *Ae. albopictus* samples (**Table 7**) and were used to test for the presence of new viral integrations using the Vy-PER [51] and ViR [137] pipelines.

Wild-collected samples derive from La Reunion Island (France), Crema (Italy), Chiang Mai (Thailand) and Tapachula (Mexico). All samples were collected as adults, shipped to the University of Pavia where DNA was extracted using the QIAGEN Blood and Tissue kit (Qiagen, Hilden, Germany). All the samples were sent to the "Polo D'Innovazione Genomica, Genetica e Biologia" (Siena, Italy) or to the Biodiversa srl company (Rovereto, Italy) for DNA quality control, Illumina library preparation and whole genome sequencing (WGS). Paired-end sequencing was done on either Illumina HiSeq 2500 or HiSeq4000.

Data are divided into single sample and pool sequencing (hereafter called Pool). A total of 22 and 24 mosquitoes from La Reunion Island and Tapachula, respectively, were processed as SSMs and sequenced at an approximate 30x coverage (**Table 7**). Their library preparation and sequencing were conducted by Verily (Google).

Three replicate pools of 40 mosquitoes were processed from La Reunion Island, Crema, Chiang Mai and Tapachula and sequenced with an approximate 30x coverage. I will refer to these samples as Pool30 (**Table 7**). Eight replicate pools of 30 mosquitoes from different localities of La Reunion Island (**Figure 21**) were further processed and sequenced at an approximate 60x coverage. I will refer to these samples as Pool60 (**Table 7**).

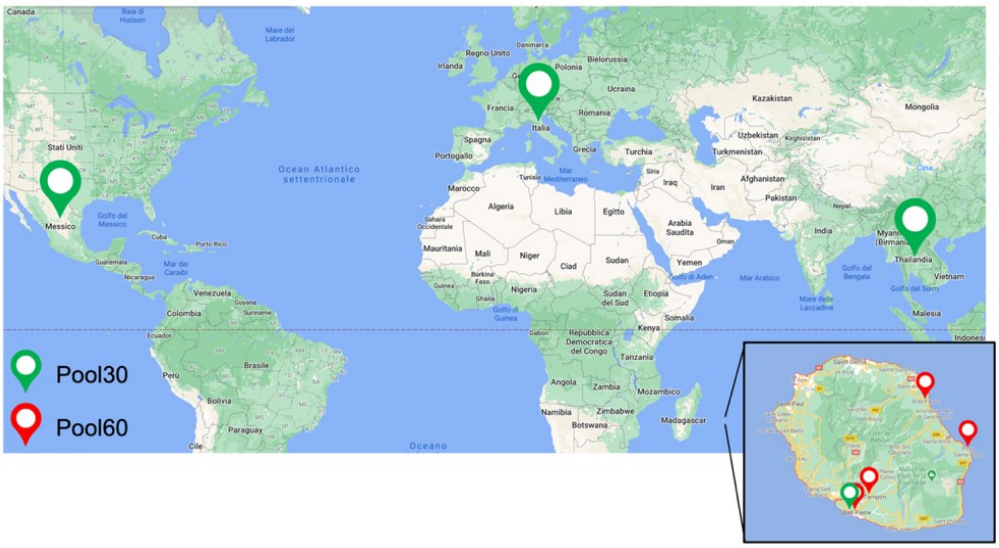


**Figure 21.** *Aedes albopictus* sampling sites. *Aedes albopictus* samples were collected in different geographical sites in the world with a specific focus in

La Reunion Island (France). Green pointers show Pool30 sampling sites; red pointers show Pool60 sampling sites.

**Table 7.** *Aedes albopictus* WGS data analyzed.

| Collection Site | Sample name | | Sequencing Strategy |
|---|---|---|---|
| La Reunion Island (France) | RosAnsR1 | RosAnsR2 | Pool 60 |
| | BraR1 | BraR2 | |
| | StpR1 | StpR2 | |
| | TamR1 | TamR2 | |
| | ReunionR1 | | Pool 30 |
| | ReunionR2 | | |
| | ReunionR3 | | |
| | Tam-1_LIN210A145 | Tam-14_LIN210A158 | SSM |
| | Tam-3_LIN210A147 | Tam-15_LIN210A159 | |
| | Tam-4_LIN210A148 | Tam-16_LIN210A160 | |
| | Tam-5_LIN210A149 | Tam-17_LIN210A161 | |
| | Tam-7_LIN210A151 | Tam-18_LIN210A162 | |
| | Tam-8_LIN210A152 | Tam-19_LIN210A163 | |
| | Tam-9_LIN210A153 | Tam-20_LIN210A164 | |
| | Tam-10_LIN210A154 | Tam-21_LIN210A165 | |
| | Tam-11_LIN210A155 | Tam-22_LIN210A166 | |
| | Tam-12_LIN210A156 | Tam-23_LIN210A167 | |
| | Tam-13_LIN210A157 | Tam-24_LIN210A168 | |
| Tapachula (Mexico) | TapachulaR1 | | Pool 30 |
| | TapachulaR2 | | |
| | TapachulaR3 | | |
| | JP-1_LIN210A121 | JP-13_LIN210A133 | SSM |
| | JP-2_LIN210A122 | JP-14_LIN210A134 | |
| | JP-3_LIN210A123 | JP-15_LIN210A135 | |
| | JP-4_LIN210A124 | JP-16_LIN210A136 | |
| | JP-5_LIN210A125 | JP-17_LIN210A137 | |
| | JP-6_LIN210A126 | JP-18_LIN210A138 | |
| | JP-7_LIN210A127 | JP-19_LIN210A139 | |
| | JP-8_LIN210A128 | JP-20_LIN210A140 | |
| | JP-9_LIN210A129 | JP-21_LIN210A141 | |
| | JP-10_LIN210A130 | JP-22_LIN210A142 | |
| | JP-11_LIN210A131 | JP-23_LIN210A143 | |
| | JP-12_LIN210A132 | JP-24_LIN210A144 | |
| Chiang Mai (Thailand) | ChiangMaiR1 | | Pool 30 |
| | ChiangMaiR2 | | |
| | ChiangMaiR3 | | |
| Crema (Italy) | CremaR1 | | Pool 30 |
| | CremaR2 | | |
| | CremaR3 | | |

WGS data were analyzed with Vy-PER with the latest version of the reference genome of *Ae. albopictus*, AalbF2 (RefSeq assembly accession: GCF_006496715.1) [86] and a viral genome database of Arboviruses and

ISV. The viral database, created in October 2019, is composed of 990 nucleotide sequences corresponding to 409 taxon-ids and is described in **Appendix 5**.

For each sample, chimeric reads from Vy-PER were evaluated with ViR_RefineCandidates.sh and ViR_SolveDispersion.sh [137] with default settings, as described in Chapter 5.3. Finally, in cases when the identified viral integration did not encompass both the left and right integration sites, I run ViR_LTFinder to find the whole integrated sequence. ViR_LTFinder was implemented using as reference the viral portion of the identified viral integration and the corresponding WGS data.

All the results obtained running ViR with the chimeric reads found by Vy-PER are presented in the **Appendix 6**. I found chimeric reads in each sample with the exception of JP-1, JP-11, TapachulaR2 and TapachulaR3. The minimum and the maximum number of chimeric reads detected in SSMs are 0-17 in La Reunion Island and 0-12 in Mexico; in Pool30 I detected between 6-23 chimeric reads in Italy, 9-25 in La Reunion Island, 0-18 in Mexico and 2-67 in Thailand; in Pool60, I detected between 22-44 chimeric reads. I called a viral integration when there was at least support from two chimeric read pairs. I also identified situations in which a chimeric pair was composed of good quality reads that mapped on the host genome and/or the viral genome in positions far apart from any other chimeric pair. I called these chimeric reads "Ungrouped". I found "Ungrouped" reads in all samples, with the exception of JP-15, Tam-4, Tam-9, Tam-18 and ChiangMaiR3.

Based on the maximum length of the DNA fragments obtained by the NGS sequencing step of all my data (10000 bp), I did not consider as novel a candidate viral integration occurring in a window of 10000 nucleotides from any reference nrEVE. Instead I assumed these chimeric reads identified polymorphisms of reference nrEVEs, which is frequent as I showed in the genome of Foshan mosquitoes [100].

I also discarded two putative novel viral integrations based on the dubious results from the viral reads. In one case, I observed a putative viral integration of a ~220 nucleotides with similarity to the Iridoviridae family (Accession number NC_023848.1). This sequence shows several alignments with the 28S ribosomal gene of *Aedes* species (score range: 239-319; query cover: 99%; expected value range: $6e^{-83}$-$4e^{-59}$; percentage identity range: 84.47-92.24%) and with analogous scores with the High Island Virus of the Reoviridae family (score range: 321; query cover: 99%; expected value range: $2e^{-83}$; percentage identity range: 92.24%). This result suggests that rather than being a viral integration, the identified sequence is a ribosomal gene. A second candidate viral integration was a ~170 nucleotides with similarity to an Orthobunyavirus of the Perybunyaviridae family (Accession number NC_018464.1). The alignment of the viral reads of this candidate viral integration against the NCBI NR database did not produce viral hits.

These two results highlight the importance to build the viral database with accurate viral sequences.

## 5.6. Novel nrEVEs of *Aedes Albopictus*

A total of 31 candidate viral integrations were identified across tested samples (**Appendix 6**), some of these were shared among samples resulting in a total of 13 novel viral integrations. The list these 13 novel nrEVEs is shown in **Table 8**, including the viral species to which they are similar and the samples in which they were identified. nrEVEnew-1, -3, -6, -7 were found only in samples from La Reunion Island; nrEVEnew-2 and nrEVEnew-8 were found in samples from La Reunion Island and Mexico; nrEVEnew-4 was found in samples from Italy, La Reunion Island and Mexico; nrEVEnew-5 was found in samples from La Reunion Island, Mexico and Thailand; nrEVEnew-9 and -10 were found only in samples from Thailand; nrEVEnew-11, -12, -13 were found only in samples from Mexico.

**Table 8**. list of the 13 novel nrEVEs identified. Virus refers to the viral species to which nrEVEs are derived from and sample name refers to the mosquito samples in which the candidate nrEVE was identified.

| Candidate name | Virus | Sample name |
|---|---|---|
| nrEVEnew-1 | KRV | RosAnsR2 |
| | | Tam-13_LIN210A157 |
| nrEVEnew-2 | CFAV | RosAnsR2 |
| | | StpR2 |
| | | JP-9_LIN210A129 |
| nrEVEnew-3 | KRV | BraR2 |
| | | RosAnsR1 |
| | | TamR2 |
| | | Tam-23_LIN210A167 |
| nrEVEnew-4 | KRV | CremaR3 |
| | | RosAnsR2 |
| | | StpR2 |
| | | TamR2 |
| | | ReunionR1 |
| | | ReunionR2 |
| | | Tam-12_LIN210A156 |
| | | Tam-17_LIN210A161 |
| | | Tam-19_LIN210A163 |
| | | JP-24_LIN210A144 |
| nrEVEnew-5 | KRV | StpR2 |
| | CFAV | TamR2 |
| | | JP-8_LIN210A128 |
| | | ChiangMaiR2 |
| | | ChiangMaiR3 |

| | | |
|---|---|---|
| nrEVEnew-6 | CFAV | TamR2<br>RosAnsR1 |
| nrEVEnew-7 | AeFV | RosAnsR2 |
| nrEVEnew-8 | CFAV | RosAnsR2<br>TapachulaR1 |
| nrEVEnew-9 | AeFV | ChiangMaiR1 |
| nrEVEnew-10 | KRV | ChiangMaiR1 |
| nrEVEnew-11 | KRV | JP-16_LIN210A136 |
| nrEVEnew-12 | AnFV | TapachulaR1 |
| nrEVEnew-13 | CFAV | JP-15_LIN210A135 |

All novel nrEVEs have similarities to ISVs of the Flaviviridae family. Specifically, nrEVEnew-1, -3, -4, -5, -10 and -11 include reads with similarity to the Kamiti River Virus (KRV) (NC_005064.1). nrEVEnew-2, -5, -6, -8 and -13 include reads from the Cell Fusing Agent Virus (CFAV) (M91671.1 and NC_001564.2). nrEVEnew-7 and -9 include reads from the *Aedes* Flavivirus (AeFV) and nrEVEnew-12 includes reads from the Anopheles Flavivirus (KX148547.1).

The viral portion of nrEVEnew-2 corresponds to two different CFAV isolates (Nucleotide Accession number M91671.1 and NC_001564.2) that have 96,5% nucleotide identity. The viral portion of nrEVEnew-5 correspond to a portion from both KRV (NC_005064.1) and CFAV (NC_001564.2). Overall KRV and CFAV have a sequence identity of 70%, which increases to 80% in the region to which nrEVEnew-5 maps.

nrEVEnew-3 maps within the refence nrEVE, Flavi12. However, the viral sequence in nrEVEnew-3 (1825 bp sequence corresponding to all non-structural proteins of KRV) is different than that of Flavi12 (99 bp are absent in Flavi12).

The ViR_LTFinder script was implemented for each of the 13 candidate novel nrEVEs to resolve the integration site.

By implementing ViR_LTFinder, I was able to solve both the left and right integrations sites for nrEVEnew-1, -2, -4, -5, -7 and -11; only the left integration site of nrEVEnew-3 and -13 and only the right integration site of nrEVEnew-8, -9. I was not able to identify the integration sites of nrEVEnew-10 and -12. Results of ViR_LTFinder showed that nrEVEnew-4 and -6 correspond respectively to the right and left regions of the same integration, hereafter called nrEVEnew-4/6.

Thanks to the contribution of other members of my laboratory it was possible to test molecularly by PCR the sequences obtained for nrEVEnews from -1 to -8. These predictions were confirmed, with the exception of

nrEVEnew-8 (**Figure 22**). nrEVE-new-8 is bioinformatically supported by three reads in RosAnsR1 and three reads in TapachulaR1; thus, we cannot exclude that the absence of the amplification could be due to its rarity as observed in *Aedes aegypti* [91].

The molecular validation of nrEVEnews from -9 to -13 is still ongoing.

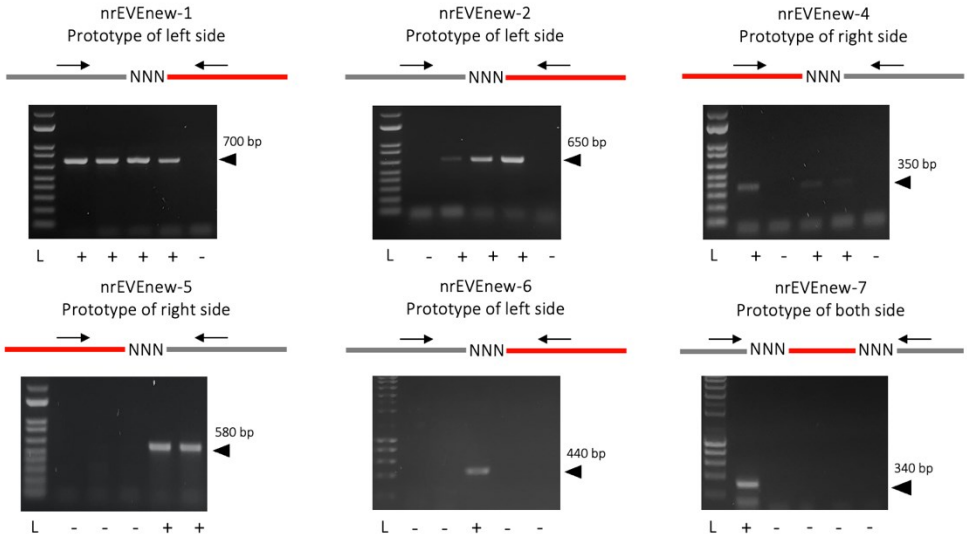


**Figure 22**. PCR molecular validation of the novel nrEVEs. For each viral integration is shown the scheme of the integration and the result of a PCR on mosquito genomic DNA with primers that were designed to check the left or right integration sites. '+' indicates the presence of the viral integration and '-' the absence [137].

The resolution of the left and right integrations sites showed that all novel nrEVEs are flanked by repeated sequence. In particular, for 5 out of the 12 novel nrEVEs the flanking regions recognized as transposable elements (TEs) but classified as 'Unknown TE' (**Figure 23**). For those sequences it is impossible to determine the precise insertion point of the integration in the reference genome.

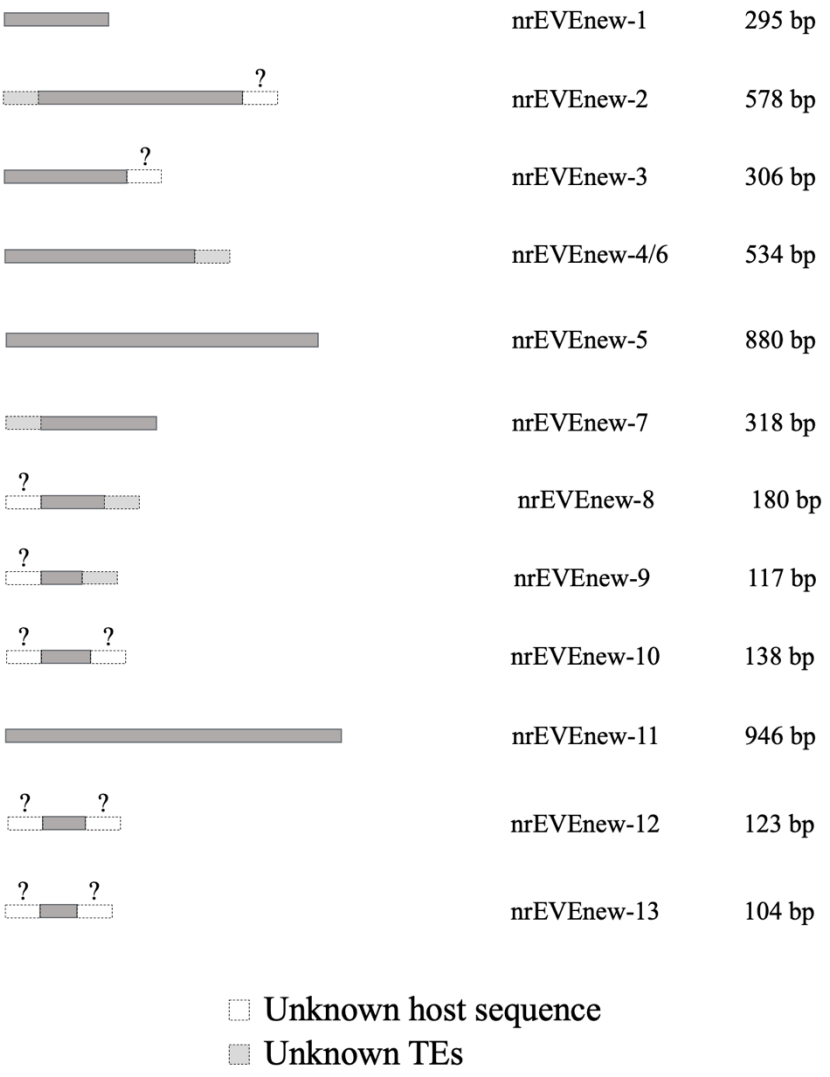


**Figure 23.** Schematic representation of the novel nrEVEs. All the novel nrEVEs are flanked by repeated sequences. In particular nrEVEnew-2, -4/6, -7, -8 and -9 are flanked by sequences recognized as TE of unknown origin.

## 5.7. Evaluation of ViR performance using WGS data

A subset of samples from La Reunion Island was used to estimate the ViR performance in terms of gain in solving the dispersion of the reads. Samples include: all the 22 single samples; RosAnsR2, StpR2 and TamR2 from the pool 30; ReunionR1, ReunionR2 and ReunionR3 from the pool 40 [137].

I used the concept of the 'Gain Index' parameter from the Information Theory [138] to assess the utility of ViR. Officially, the 'Gain Index' reflects the capacity of an attribute in segregating entities to different classes [138].

Instead, I used this index to evaluate the gain of enclosing in a single 'equivalent region' reads that had been originally assigned to different loci.

In particular, the attribute is the "equivalent region' identified by ViR_SolveDispersion, classes are the original read loci assigned and the entities are all the reads supporting candidate integrations (output of ViR_RefineCandidates).

I started from the concepts of Entropy $I$ and Residual Information $I_{res}$.

The Entropy $I$ is:

$$I = -\sum_{l} p(l) * \log_e p(l)$$

where $l$ is the locus ID, meaning the host genomic coordinates of the candidate integration identified before ViR and $p(l)$ is the relative frequency of the reads assigned to the locus ID $l$. Entropy is 0 when only one locus ID (i.e., one candidate integration site) is identified in the sample (i.e., WGS dataset). Entropy is > 0, when more than one locus ID is identified in the sample.

The residual information $I_{res}$ is defined by:

$$I_{res} = -\sum_{g} p(g) \sum_{l} p(l|g) * \log_e p(l|g)$$

where $g$ is the ID of the equivalent region identified by ViR_SolveDispersion, $p(g)$ is the relative frequency of the reads in the equivalent region $g$ and $p(l|g)$ is the relative frequency of reads assigned to the locus ID $l$ in the equivalent region $g$.

I evaluated the $Normalised\ Dispersion\ Gain$, which ranges between 0 and 1, as the ratio between the difference between I and $I_{res}$ and the value of the initial entropy:

$$Normalised\ Dispersion\ Gain = \frac{I - I_{res}}{I}$$

This operation allows to normalize results across samples, which were obtained using different experimental set ups (i.e., WGS from single or pools). The closer the value of $Normalised\ Dispersion\ Gain$ is to 0, the higher is the gain of ViR in solving the dispersion of reads.

To favor intuitive interpretation of results, we show Solve Dispersion Gain as:

$$Solve\ Dispersion\ Gain = 1 - Normalised\ Dispersion\ Gain$$

$$= 1 - \frac{I - I_{res}}{I} = \frac{I_{res}}{I}$$

Values of $Solve\ Dispersion\ Gain > 0$ are found when ViR was able to identify a unique equivalent region for at least two different reads previously assigned to two different loci ID. The higher the value of $Solve\ Dispersion\ Gain$, the higher is the performance of ViR.

I applied the $Solve\ Dispersion\ Gain$ considering as dataset the output reads from ViR_RefineCandidates, their initial loci assigned by Vy-PER and the reads groups assigned by ViR_SolveDispersion [137].

As an example, in Tampon-19 among the SSM, ViR solved the dispersion of seven reads identified by Vy-PER by grouping them into one group supporting nrEVEnew-4; two read remained ungrouped, resulting in a Solve Dispersion Gain value of 0,65 (**Figure 24A**). Ungrouped reads are chimeric reads identified by Vy-PER could not be grouped because they have alignments in the genome distant from the others. Even if they are not useful for the discovery of viral integrations, it is important to isolate them to avoid wasting time trying to interpret them, for this motif we included these ungrouped reads in the calculation of the Solve Dispersion Gain.

Considering all the tested, the median values of the solve dispersion gain were 0.5 in SSM, 0,42 in pool 60 and 0,48 in pool30 (**Figure 24B**). Dispersion gain values were not different between single vs pools or between pools 30 vs pools 60 samples, indicating the gain is not influenced by the sequencing strategy or depth of coverage.

**A**

AalbF2
single sample R19

| Locus ID | Read Group |
|---|---|
| L1 | Ungrouped |
| L2 | RG1 |
| L3 | RG1 |
| L3 | RG1 |
| L4 | RG1 |
| L5 | RG1 |
| L6 | Ungrouped |

**B**

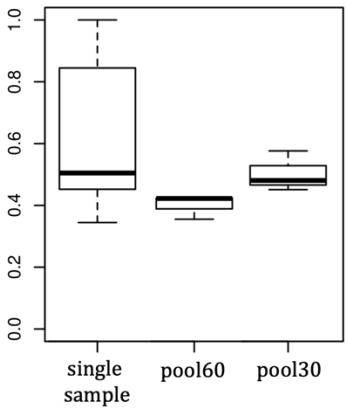


**Figure 24.** ViR performances. **A)** Solve dispersion gain data of Tampon-19. **B)** Solve dispersion gain evaluated in SSM, Pool30 and Pool60 samples from La Reunion Island [137].

# Chapter 6

# nrEVEs database

Guided by the paradigm of open science, that is the idea that sharing of materials, data and information is the foundation of a solid and reproducible scientific finding and can help speed scientific discoveries, I dedicated the last part of my PhD project to design and implement a database of nrEVEs. I was recently able to make this database public at www.nreves.com.

Nowadays, online webpage and a tabular format are the most popular ways to share biological data, but these methods do not allow data integration in automatic applications. In the last two decades, the addition of semantic annotation to tabular data has been introduced also in life science. The main advantage of semantic annotations is to describe the biological meaning of data in a computer-accessible manner, resulting in knowledge integration across several repositories and experimental data stores [139].

To follow this trend of innovation, I decided to build an ontology for nrEVEs based on already existing vocabularies. My work opens new biological and IT research fronts and places data on nrEVEs towards a logic of interoperability with other existing datasets.

## 6.1. Dataset description

Data to be included in the database are reference nrEVEs and novel nrEVEs not only from *Ae. albopictus,* which I described in **Chapter 3** (reference nrEVEs) and **Chapter 5** (novel nrEVEs) of thesis, but also from *Ae. aegypti*, as described in Crava et al., 2020.

Reference nrEVE sequences are associated to a viral protein by the annotation pipeline shown in **Chapter 3** [38] using separately proteins from the NCBI RefSeq and the NCBI NR databases.

For novel nrEVEs I was seldomly able to identify integration sites and thus, to provide the exact nrEVE mapping coordinates. In most cases I obtained a list of equivalent regions in which the integration could have

occurred. The viral protein related to novel nrEVEs is obtained by an online blastx against NCBI NR and NCBI RefSeq proteins databases.

I also included information on piwi RNA clusters, annotated CDS and TEs to describe the genomic context of where the integration occurred, when possible. TEs may be present in multiple copies in the reference genome, thus only TEs detected upstream or downstream nrEVEs are included in the dataset. TEs were described only through their Class, Order and Superfamily because of uncertainties in their classification/annotation in the *Aedes* spp. genomes.

All data were collected and organized in a Comma Separated Value (CSV) format file. In this data model, each nrEVE is represented as a tuple. A tuple is a row of the table in which the value of each column is the specific value of the attribute expressed by the column name [77].

*Aedes aegypti* and *Ae. albopictus* nrEVE datasets are associated to their reference paper.

## 6.2. nrEVEs web application

The above-described dataset was made available for the research community at www.nrEVEs.com. The graphic of the site was created using the HTML language [140] and React (node-v12.18.3 with npm v6.14.6) [141]. React is a JavaScript library for building user interfaces. The possibility to start from small and isolated pieces of code called "components" makes this library very flexible and efficient [141].

The user interface includes three web pages (Home, Browse and Download), which are browsable through the navigation bar. The navigation bar was created merging classes from the packages of react-bootstrap version 1.3.0 [142] and react-router-dom version 2.1.8 [143].

The "Home" page shows the release list of the database, the list of the publications to cite nrEVEs datasets and how to contact nrEVEs contributors. The database could be further expanded to include nrEVE data from other species leading to new releases (**Figure 25**).

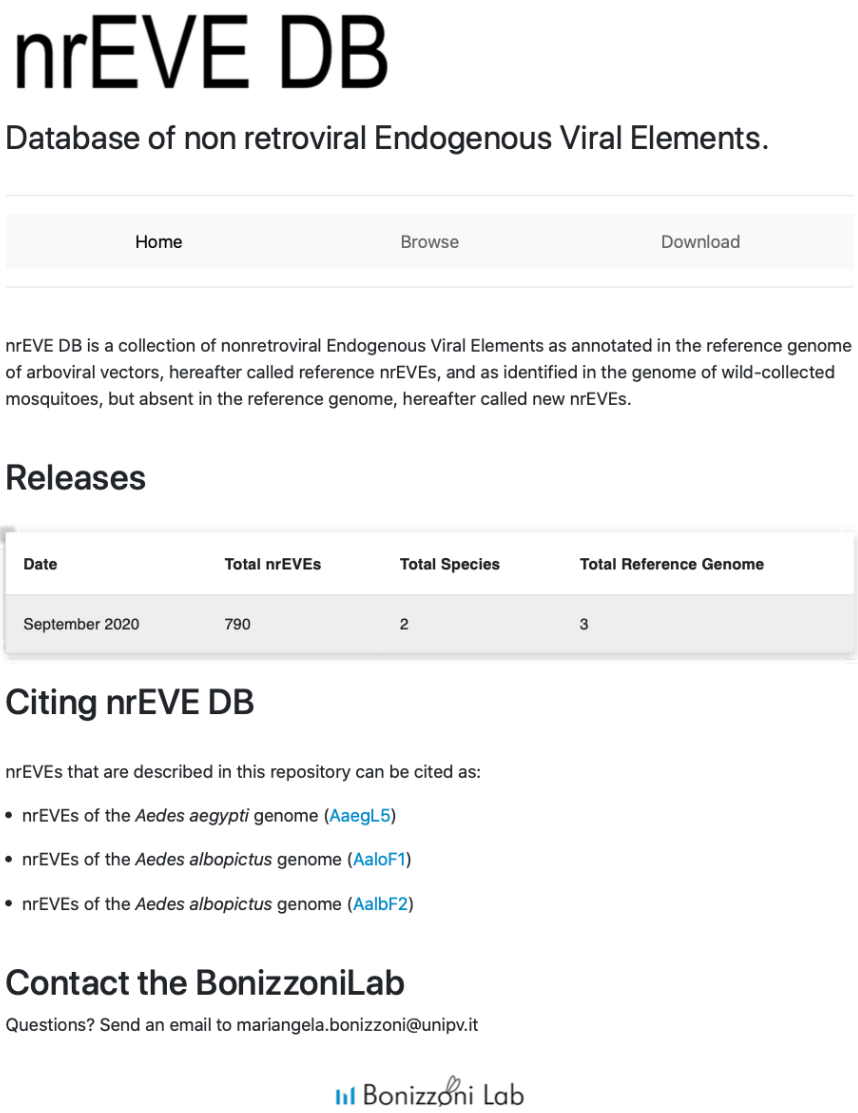


**Figure 25.** nrEVEs Database "Home" page.

The “Browse” page shows the information available for nrEVEs data. Reference and novel nrEVEs are shown in two separated tables, both created using the mui-datatables package version 3.4.0 [144]. The mui-datatable allow the user to customize the visualization of nrEVEs data with the specific settings. They include:

- the number of nrEVE instances per page;
- the option to scroll nrEVE instances;
- the option to sort columns;
- activate or deactivate columns visualization;
- filter nrEVEs based on a specific pattern of words;
- filter nrEVEs based on one or more specific columns category.

An example of the application of a mui-datatable in the nrEVEs database is shown in in **Figure 26**.

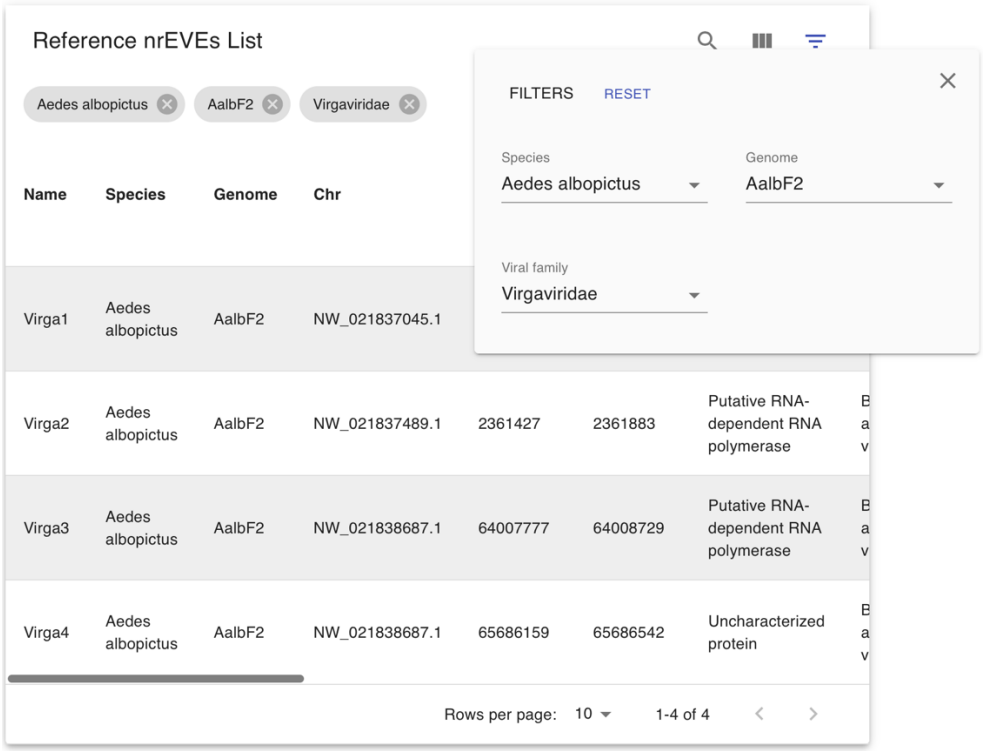


**Figure 26.** Example usage of nrEVEs Database browsable tables. Reference nrEVEs from the Virgaviridae family detected in the AalbF2 reference genome of *Ae. albopictus* are shown in the table. Filter could be set in the top right panel. Active filters are visible as removable tags under the table title. The number of rows per page are shown at the bottom bar.

Finally, in the "Download" page the user can download data in CSV format, in FASTA format, in BED format. Coordinates of piRNA clusters can be downloaded in BED format.

## 6.3. Ontology design

An ontology is a formal description of concepts in a certain domain of interest [145]. The creation of an ontology is done through the connection of different entities: classes, properties, restrictions and instances.

All concepts in a certain domain can be described by classes organized in taxonomies, connected with a "is a" relation. As an example, a branch of the Genome Biology Ontology Language (GBOL) ontology, regarding the description of the features related to a sequence, is shown in **Figure 27**. gbol:SequenceAnnotation and gbol:GenomicFeature classes are 'child' of the gbol:NAFeature class; gbol:GeneralFeature and gbol:NAFeature are 'child' of the Feature class.

Classes have intrinsic properties, named in Protégé Data Properties. For example gbol:SequenceAnnotation and gbol:GenomicFeature classes have as property a gbol:standardName which is a text (**Figure 27**).

Classes can be related among them through another type of properties, named in Protégé Object Properties. In our case, a sequence could have one or more features associated; thus, the gbol:Sequence is the domain of the Object Property gbol:feature and the class gbol:Feature is the range of the relation (**Figure 27**).

Restrictions set the type of data associated to a Data Property and the amount of connection that can be done with a property (both Data and Object Properties).

Instances are the specific individuals of a class. They inherit both the Data and the Object Properties of their class. For example, consider "Sequence_1" as an instance of the class gbol:Sequence, and "nrEVE" an instance of the class gbol:GenomicFeature. "Sequence_1" can be connected with "nrEVE" using the Object Property gbol:feature.

Finally, all the ontology entities can be annotated with metadata named in Protégé Annotation Property. For example, this kind of property regards the attribution of a version or a creator of an ontology entity.

Currently, according to the Semantic Web principles [146], all ontologies entities are commonly referred to using an Internationalized Resource Identifier (IRI), unique identifier for resources in the web [81].

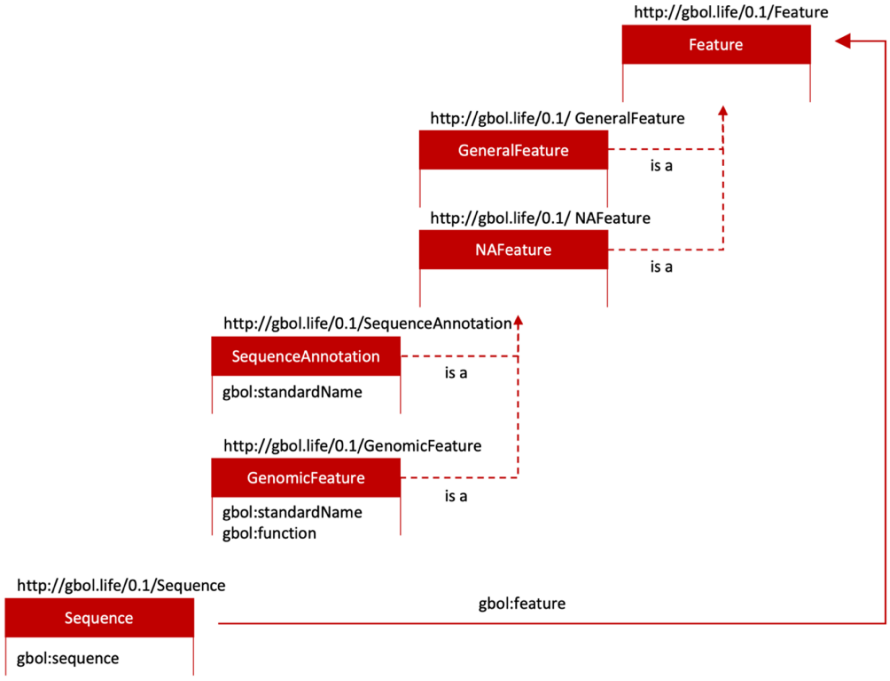


**Figure 27.** Example of the ontology entities. Entities from the GBOL ontologies [147] are shown. Both the gbol:SequenceAnnotation and gbol:GenomicFeature classes are connected to gbol:NAFeature with the relation "is a". The same relation is present between gbol:GeneralFeature and gbol:NAFeature with respect to the Feature class. Thus, all these 5 classes contribute to create a taxonomy of concepts. gbol:SequenceAnnotation and gbol:GenomicFeature classes have as Data Property gbol:standardName. The gbol:Sequence is the domain of the Object Property gbol:feature and the class gbol:Feature is the range of the relation.

nrEVEs have been recently identified as components of eukaryotic genomes, thus there is not a comprehensive ontology to describe nrEVE data. There are three approaches to overcome the creation of a new ontology [78]:

- To search for relevant classes and relations in already existing ontologies, compare them, find the most adaptable to nrEVEs and reuse them;
- To define new classes to describe nrEVEs and organize them in a newly designed ontology;
- To use a mixed strategy where vocabularies and relations from other ontologies are merged with newly defined classes to describe nrEVEs.

I chose a mixed strategy and started the ontology design applying a bottom-up strategy to take advantage of already existing ontologies and defining only missing terms. Reuse of existing knowledge model was a necessary step when a certain domain was already, totally or partially described. This step favors the standardization of the information avoiding inconsistency that may be generated from multiple representation of the same concept [145].

The majority of the terms to describe the nrEVEs dataset were identified in the BioPortal [148] and Linked Open Vocabulary (LOV) [149] platforms and they were extracted from the ontologies shown in **Table 9**.

**Table 9.** Source ontologies of reused terms.

| Acronym | Full name | Availability | Release date |
|---|---|---|---|
| FALDO | Feature Annotation Location Description Ontology | http://biohackathon.org/resource/faldo | 27/08/19 |
| RSA | Reference Sequence Annotation | http://rdf.biosemantics.org/ontologies/rsa | 31/01/14 |
| RO | Relation Ontology | http://purl.obolibrary.org/obo/ro/releases/2020-02-26/ro.owl | 26/02/20 |
| GBOL | Genome Biology Ontology Language | http://gbol.life/0.1/ | 05/09/17 |
| UP | Uniprot KB | http://purl.uniprot.org/core/ | October 2019 |
| SO | Sequence types and features ontology | http://purl.obolibrary.org/obo/so/2020-05-28/so.owl | 28/05/20 |
| SIO | Semanticscience Integrated Ontology | http://semanticscience.org/ontology/sio/v1.44/sio-release.owl | 19/04/20 |

The description of the coordinates of nrEVEs, piRNA clusters and CDS in the reference genome was made using terms available in the Feature Annotation Location Description Ontology (FALDO) [150] and in the Reference Sequence Annotation (RSA) [151]. Biologically, nrEVEs, piRNA clusters and CDS occupy a unique specific region of the chromosome/scaffold/contig of the reference genome. Thus, the region (faldo#Region) is delimited by specific start and end nucleotides, both described by the faldo#ExactPosition class. Each exact position belongs to a reference sequence (rsa#ReferenceSequence) using the Object Property faldo:reference. The relation between the reference sequence and the genome assembly (rsa#GenomeAssembly) is described through the Object Property ro:proper_part_of that is not currently available in the ontology [151]. As a consequence, this relation was substituted with the relation ro:part_of. Both these two last Object Properties derive from the Relation Ontology (RO) [152].

nrEVEs and piRNA clusters were represented by the same class from GBOL [147], gbol:Sequence. CDS were represented by gbol:CDS. Instead, TEs were described by the superclass SO_0000101 and its children from the Sequence types and features Ontology (SO) [153]. Reference nrEVEs and piRNA clusters instances from the gbol:Sequence and gbol:CDS classes show one related region in the reference genome. Novel nrEVEs could have more insertion points faldo#InBetweenPosition. In order to represent this concept I created two subclasses of gbol:Sequence named 'Annotated_sequence' and 'Not_annotated_sequence'. nrEVEs, piRNA clusters and CDS belong to 'Annotated_sequence'; novel nrEVEs belong to 'Not_annotated_sequence'.

Several GBOL classes have been adopted to describe the annotation of a sequence related to nrEVEs and piRNA clusters. Each gbol:Sequence instance has a gbol:feature to one of the two instances of gbol:GenomicFeature, "nrEVE" and "piRNA cluster" (thought to be categories defining the nature of a sequence, i.e., nrEVEs or piRNA clusters). Furthermore, the recognition of a sequence as nrEVEs derives from a classification process based on the comparison of the sequence with a database of proteins throughout the usage of blast tool. In the ontology this is expressed by the following relations: 1. gbol:Sequence gbol:feature gbol:SequenceAnnotation; 2. gbol:SequenceAnnotation gbol:db gbol:Database; 3. gbol:SequenceAnnotation gbol:deriveFrom gbol:Blast; 4. gbol:Blast sio:refersTo up:Protein.

sio:refersTo is an Object Property from the Semanticscience Integrated Ontology (SIO) [154].

The protein is represented by the class up:Protein from Uniprot [79]. This class is used in the ontology to represent two cases: a protein produced by a CDS annotated in the same region of a nrEVE and a viral protein related to a nrEVE through a relation of similarity express by a Blast result.

Two other classes were reused from the Uniprot schema. up:Taxon represents the organism from which the rsa:ReferenceAssembly derives.

up:Citation is used to add the reference of a sequence or a reference assembly.

As already described, nrEVEs were annotated in both the AloF1 and AlbF2 *Ae. albopictus* genome assemblies. To help comparing results from the two assemblies, a further annotation was added for nrEVEs which have similarity between the two assemblies. Briefly, each couple of nrEVEs with nucleotide identity % > 90% has a gbol:feature relation to the same instance of the class 'Similarity'. The score of their alignment is stored in an instance of gbol:Blast. With respect to first usage of the Blast instances, these set of gbol:Blast instances have no sio:refersTo relation with a up:Protein.

Once downloaded, the ontologies shown in **Table 9** were visualized and manipulated in the Protégé 5.5 software [155], open source editor ontology widely used in life science applications. Protégé is an ontology development environment available both online (Web Protégé) and in the downloadable version (Protégé desktop), and is supported by a rich online documentation [156]. Protégé support the standard ontology language Web Ontology Language (OWL) developed by the World Wide Web Consortium [157] [158].

Using Protégé, I applied a top-down approach to discard all the unnecessary entities and their asserted descendent entities from the downloaded ontologies shown in **Table 9**. Then, the resulting minimized ontologies were imported in a new ontology and merged.

Three classes were created *ex novo*. The 'Annotated_sequence' and 'Not_annotated_sequence' classes were created as subclasses of gbol:Sequence to distinguish sequences with a specific region in the reference genome and sequences absent in the reference genome. These last sequences appear as insertions occurring in specific points of the reference genome. The class 'Similarity' describes the similarity between nrEVEs annotated in an old reference genome respect to a new one. This connection helps researchers to connect results obtained in the past with a recent version of a reference genome.

All the terms of the ontology are shown in **Appendix 7**. The graphic representation of the final ontology is shown in **Figure 28**.

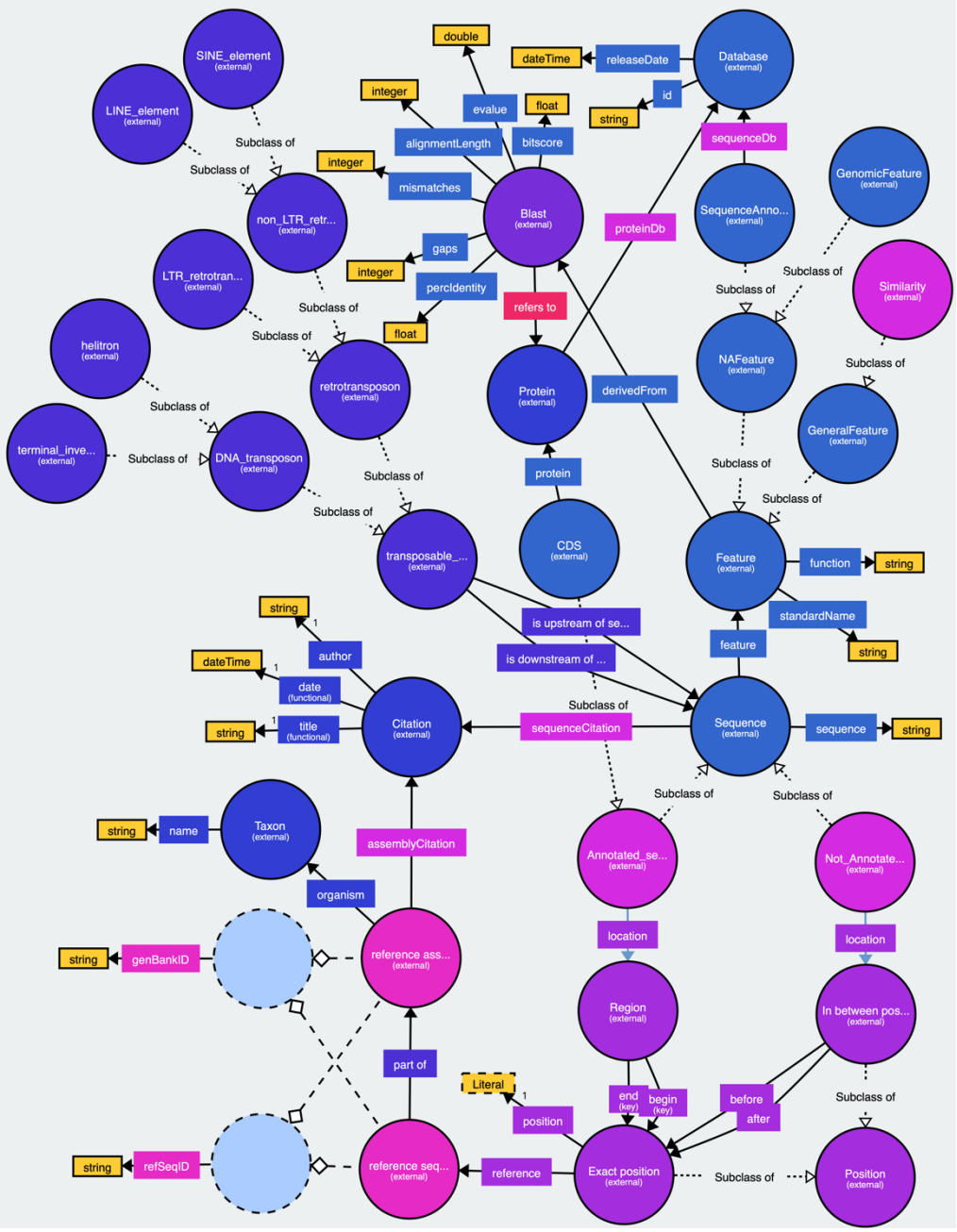


**Figure 28.** nrEVEs ontology represented through WebVOWL [159].

# Chapter 7

# Conclusion

Interaction between viruses and their hosts occurs at many levels. Viruses can impact the genetic structures of host populations if viral infection results in disease and mortality. One of the most remarkable examples of this phenomenon is the influenza pandemic of 1918, which caused about 50 million deaths worldwide [160]. Another mechanism through which viruses affect their host is through transfer and integration of their genetic material into host genomes. Integrations of viral sequences into host genomes is a well-recognized phenomenon for certain classes of viruses such as DNA viruses and retroviruses. For instance, several DNA viruses, as HBV and HPV, integrate into the genome of their host cells inducing genome instability, which can progress into carcinogenesis [46]. Integrations of the retrovirus HIV are a way to elude the host immunity response and favour infection persistence. The most common viruses that infect humans, and include epidemiologically relevant viruses such as the influenza, Ebola, Chikungunya, Dengue and West Nile viruses, are non-retroviral RNA viruses, meaning viruses with an RNA-based genome that lacks the machinery needed for integration into host genomes [122]. Because of this property, several species of non-retroviral RNA viruses are used in gene therapy applications as delivery systems for drugs and vaccines [161]. As a consequence, the recent finding of sequences from non-retroviral RNA viruses integrated into animal and plant genomes came as a shocking surprise to the scientific community, a discovery that the virologist Edward Holmes calls “one of the most remarkable observations in viral evolution of recent years” [162]. However, the widespread of this phenomenon, its mechanisms and the biological significance of nrEVEs in host genomes are still largely unknown [61].

*Aedes aegypti* and *Aedes albopictus* are the main worldwide vectors of arboviruses. Previous work by members of the Bonizzoni laboratory

demonstrated a significantly higher concentration of nrEVEs in the genomes of *Aedes* spp. mosquitoes compared Anophelinae mosquitoes [34].

Starting from these discoveries, during my PhD project I participated in the annotation of nrEVEs in the latest version of the reference genome of *Ae. albopictus*, AalbF2 [86]. I developed a short pipeline to avoid manual filtering after the application of the EVE_finder pipeline [38] and to organize data in a repeatable format (**Chapter 3**).

A total of 456 loci harboring sequences from nine viral families were identified in AalbF2. nrEVEs appear to be in close association with TEs, they are enriched in piRNA clusters and produce piRNAs. This result suggests the involvement of nrEVEs in the piRNA pathway, one of the main antiviral mechanisms of mosquito.

The application of the nrEVE annotation pipeline, including the filtering steps that I developed, to other organisms could allow to understand whether the phenomenon of host genome integrations from nonretroviral RNA viruses occurs in all viral lineages, or it is limited to specific host-viral combinations thus unbaling the choice of "safe" nonretroviral RNA virus species to be used in gene therapy applications. Furthermore, understanding the mechanisms of mosquito immunity could provide new tools to control arbovirus spread.

Starting from the coordinates of reference nrEVEs, I studied the polymorphism of nrEVEs to understand their widespread occurrence and evolution within mosquito genomes. To reach this goal, I developed the SVD bioinformatic pipeline as described in **Chapter 4**. I applied this pipeline to WGS data from 16 mosquitoes of the Foshan strain. Results of this work were published in the article titled 'Insights into an unexplored component of the mosquito repeatome: Distribution and variability of viral sequences integrated into the genome of the arboviral vector *Aedes albopictus*' of which I am the first author [100].

The biological significance of nrEVEs is strongly dependent on their distribution in geographical different mosquitoes, which are exposed to different circulating viruses. To identify novel nrEVEs in WGS data from wild-collected mosquitoes I developed the ViR bioinformatic pipeline, as described in **Chapter 5**. ViR starts with a set of chimeric reads, which support an undefined number of viral integrations, and organizes them in groups which support the same integration even if it occurs in a repeated region of the host genome. The pipeline was tested using WGS data from both single and pool samples. The pipeline is described in the article titled 'ViR: a tool to account for intrasample variability in the detection of viral integrations' in which I am the first author [137]. The article is currently under revision.

In the final part of the PhD, I build a database of reference and novel nrEVEs of *Aedes* spp. mosquitoes. The database is public at www.nreves.com.

Additionally, to increase the interoperability and the usage of these data in automatic application, I created the first ontology capable of a full description of the nrEVE dataset described in my thesis. According to the Semantic Web and Linked Data principles, I studied already standard and published ontologies from BioPortal and LOV to take advantage of existing vocabularies, including classes and properties. I included *ex novo* terms only for missing vocabularies. This ontology is a first step towards the integration of nrEVEs data with other sources.

Ontologies are considered a fundamental part of the biological and biomedical research because they are commonly used to explicit deliver the knowledge behind life sciences data. Currently, the amount of biological data continues to increase, data come from different disciplines and a large number of databases is emerging. Under this scenario, it is important not only to represent data, but also to easily investigate data.

The description of biological data in a computer-accessible manner helps researchers find new models and enable knowledge integration across several repositories and experimental data stores [167]. However, the comparison and the combination of different data is a complex interoperability task. It is difficult to understand if different data sources could coexist in the same structure and work together coherently and this requires a huge effort to study both the structure and the content of each involved datasets [163]. The Semantic Web technologies and the Linked Data principles are facilitating this process creating an interlinked web of knowledge that can be easily navigated and processed by software agents. For example, data mining applications can respond to complex queries combining many heterogeneous biological sources.

During my PhD program I also had the opportunity to apply my pipelines to WGS data from *Ae. aegypti*, besides *Ae. albopictus*. Briefly, I contributed to three different projects:

Project1: the SVD and ViR pipelines were applied to *Ae. aegypti* wild mosquitoes collected in Gabon, Ghana, Kenya, Mexico and America Samoa. This work was done in collaboration with Dr. Cristina Crava [91].

Project2: The SDV pipeline was applied to identify the pattern of reference nrEVE in Brazilian samples from Bebedouro and Botucatu. The ViR pipeline was applied to the same dataset to identify novel nrEVEs. This project was in collaboration with Prof. Jayme Souza Neto, the Sao Paulo State University Botucatu (Brazil). Because mosquitoes from Bebedouro and Botucatu have a different vector competence for Dengue viruses, this study wants to test whether the landscape of viral integrations contributed to this different vector competence.

Project3: I applied the ViR pipeline to WGS data from the *Ae. aegypti* Aag2 cell line after infection with different arboviruses to test for the formation of nrEVEs. This work was done in collaboration with Annabella Failloux with the Institute Pasteur in Paris (France). Novel nrEVEs were detected after infection with CHIKV, DENV and Cell Fusing Agent Virus (CFAV); while no nrEVEs were discovered in Aag2 cells after infecting with Vesicular Stomatitis Virus (VSV) and Kamiti River Virus (KRV).

To conclude, during my PhD project, I produced pipelines to facilitate the annotation, study and identification of nrEVE and I applied them to *Aedes* spp. mosquitoes. I also delivered the first database of nrEVEs. The pipelines include previously absent standard methods to study specific biological hypothesis. The formalization of these methods in automated pipelines allows their usage in massive analysis producing comparable results and saving time. All the pipelines are versatile and can be applied to both model and non-model organisms.

# Appendix

## nrEVEs correspondence between AaloF1 and AalbF2

Summary of the correspondence between nrEVEs in AaloF1 and AalbF2 reference genomes [86].

**Table 10.** Reference nrEVEs annotated in AaloF1 are divided in four categories: nrEVEs that have no match in AalbF2, nrEVEs annotated in AalbF2 for their entire length, nrEVEs annotated in AalbF2 only for a portion of their length; nrEVEs having more matches in AalbF2.

| **nrEVEs without correspondence in AalbF2** | | **nrEVEs corresponding with nrEVEs in AalbF2** | | **nrEVEs partially corresponding with nrEVEs in AalbF2** | | **nrEVEs with more matches in AalbF2** | |
|---|---|---|---|---|---|---|---|
| AlbFlavi18 | AlbRha10 | AlbFlavi1 | AlbRha1 | AlbFlavi25 | AlbRha58 | AlbFlavi1 | AlbRha12 |
| AlbFlavi19 | AlbRha3 | AlbFlavi10 | AlbRha11 | AlbFlavi3 | AlbRha62 | AlbFlavi10 | AlbRha15 |
| AlbFlavi20 | AlbRha38 | AlbFlavi 12_17 | AlbRha12 | AlbFlavi31 | AlbRha66 | AlbFlavi 12_17 | AlbRha2 |
| AlbFlavi28 | AlbRha41 | AlbFlavi2 | AlbRha14 | AlbFlavi32 | AlbRha85 | AlbFlavi22 | AlbRha28 |
| AlbFlavi38 | AlbRha42 | AlbFlavi22 | AlbRha15 | AlbFlavi33 | AlbRha87 | AlbFlavi23 | AlbRha32 |
| AlbFlavi39 | AlbRha44 | AlbFlavi23 | AlbRha18 | AlbFlavi34 | | AlbFlavi25 | AlbRha4 |
| AlbFlavi40 | AlbRha7 | AlbFlavi24 | AlbRha2 | AlbFlavi7 | | AlbFlavi26 | AlbRha48 |
| | AlbRha73 | AlbFlavi26 | AlbRha28 | | | AlbFlavi27 | AlbRha52 |
| | AlbRha74 | AlbFlavi27 | AlbRha32 | | | AlbFlavi34 | AlbRha62 |
| | AlbRha79 | AlbFlavi36 | AlbRha33 | | | AlbFlavi36 | AlbRha66 |
| | AlbRha80 | AlbFlavi37 | AlbRha36 | | | AlbFlavi37 | AlbRha71 |
| | | AlbFlavi4 | AlbRha4 | | | AlbFlavi41 | AlbRha83 |
| | | AlbFlavi41 | AlbRha43 | | | AlbFlavi42 | AlbRha84 |
| | | AlbFlavi42 | AlbRha45 | | | AlbFlavi8 | AlbRha85 |
| | | AlbFlavi6 | AlbRha48 | | | | AlbRha87 |
| | | AlbFlavi8 | AlbRha49 | | | | AlbRha9 |
| | | | AlbRha52 | | | | AlbRha92 |
| | | | AlbRha71 | | | | AlbRha94 |
| | | | AlbRha83 | | | | AlbRha95 |
| | | | AlbRha84 | | | | AlbRha96 |

| | | |
|---|---|---|
| | AlbRha88 | |
| | AlbRha9 | |
| | AlbRha92 | |
| | AlbRha94 | |
| | AlbRha95 | |
| | AlbRha96 | |

**Table 11.** Multiple viral integrations in AaloF1 have match with the same nrEVE in AalbF2.

| | **AaloF1** | **AalbF2** |
|---|---|---|
| **F-NIRVS** | AlbFlavi 2 | Flavi12 |
| | AlbFlavi3 | |
| | AlbFlavi24 | Flavi27 |
| | AlbFlavi31 | |
| | AlbFlavi32 | |
| | AlbFlavi33 | |
| | AlbFlavi6 | Flavi1 |
| | AlbFlavi7 | |
| **R-NIRVS** | AlbRha33 | Rhabdo4 |
| | AlbRha43 | |
| | AlbRha58 | |

# SVD: pipeline parameters

**Table 12**. Input parameters of the pipeline SVD. The table is divided into three columns: Section column, Parameter column and Description column. The Section column includes schematic categories to summarize input parameters. The Parameter column includes the parameters used in the pipeline. The Description column includes a description of the parameter.

| Section | Parameter | Description |
|---|---|---|
| Samples info | -c\|--configsample) | path of the configuration file including the list of samples. One per raw, each sample is represented by its id and its BAM path file. |
| Paths of the pipeline files and output directory | -i\|--filepath) | path of the pipeline directory (i.e., /AbsPathTo/StructuralVariantsDefinition/) |
| | -o\|--output) | path of the output directory |
| Path of the references | -b\|--bedfile) | path of the bed file including contig start stop name of the loci of interest |
| | -b_pl\|--bedfile_platypus) | path of the platypus bed file including contig start stop name of the loci of interest |
| | -f\|--fasta) | path of the fasta file of the reference genome |
| Path of the executable tools | -fbpath\|--freebayespath) | |
| | -gkpath\|--gatkpath) | |
| | -vdpath\|--vardictpath) | |
| | --fileR) | |
| | --filePl) | |
| | -plpath\|--platypuspath) | |
| | -btpath\|--bcftoolspath) | |
| Parameters of the variant callers | -th\|--threads) | number of threads |
| | -R\|--ram) | ram to use in variant calling for GATK (es. 3g) |
| | --MIN_MQ) | minimum phred mapping quality to call a variant |
| | --MIN_BQ) | minimum phred base quality to call a variant |

| | | |
|---|---|---|
| | --MIN_AF) | minimum allele frequency to call a variant |
| | --MIN_AO) | minimum allele observations to call a variant |
| | --MIN_COV) | minimum depth of coverage to call a variant |
| | --MAX_DEPTH) | maximum depth of coverage to call a variant |
| Parameters for features extraction | --DP_expected_mean) | mean of the read coverage in the sample |
| Parameters for filter file | -af\|--minallfreq) | minimum allele frequency to call a variant |
| | -dp\|--mindepth) | minimum depth of coverage to call a variant |
| Parameters to filter variants and define allele for the AllData file output | --AFallData) | minimum allele frequency to include variant in the output |
| | --MaxStrBias) | maximum strand bias to include variant in the output |
| | --MinLengthINDELallData) | length of the INDEL to include variant in the output |
| | --NumCallersallData) | number of callers that have to simultaneously call a variant for it to be accepted |
| | --minReadsAllDef) | minimum number of reads coverage to consider an allele present |
| | --minLengthAllele) | minimum number of consecutive nucleotides to define an allele |
| | --thresholdSimilarity) | maximum percentage of the annotated sequence length to consider two alleles equal |

# SVD: Variant callers features

**Table 13**. Features extracted from the VCF files of each Variant Caller. Feature values are directly drawn in the feature extraction output file if present in the VCF file; otherwise, features are evaluated with the drawn feature values when possible. The abbreviation of each feature is explained in the text.

| Features | Freebayes | Platypus | GATK | Vardict | Complete feature name |
|---|---|---|---|---|---|
| **GT** | Drawn | Drawn | Drawn | Drawn | Genotype |
| **AO** | Drawn | Drawn | Drawn | Drawn | Alternate observations |
| **RO** | Drawn | Evaluated | Drawn | Drawn | Reference observations |
| **AO_f** | Drawn | Drawn | . | Drawn | Alternate observations forward |
| **AO_r** | Drawn | Drawn | . | Drawn | Alternate observations reverse |
| **RO_f** | Drawn | Evaluated | . | Drawn | Reference observations forward |
| **RO_r** | Drawn | Evaluated | . | Drawn | Reference observations reverse |
| **DP** | Drawn | Drawn | Drawn | Drawn | Depth of coverage |
| **DP_f** | Evaluated | Drawn | . | Evaluated | DP reads forward |
| **DP_r** | Evaluated | Drawn | . | Evaluated | DP reads reverse |
| **AF** | Evaluated | Evaluated | Evaluated | Evaluated | Allele frequency |
| **StrandBias** | Evaluated | Evaluated | Evaluated | Evaluated | Strand bias |
| **MQ0F** | . | . | . | . | Fraction of reads with mapping quality 0 |
| **MQ0** | . | . | Evaluated | . | Number of reads with mapping quality 0 |

| | | | | | |
|---|---|---|---|---|---|
| **MQRankSum** | . | . | Evaluated | . | Zscore WilcoxonRankSum Test of Alt VS Ref mapping/base qualities |
| **BQRankSum** | . | . | Evaluated | . | |
| **BQ** | Evaluated | . | . | Drawn | Base quality |
| **Call** | 1/0 | 1/0 | 1/0 | 1/0 | If the caller recognizes a variant the call feature is set to 1 otherwise is set to 0 |

# nrEVE-specific primers

**Table 14.** List of primers used for population genetics. The respective amplicon size for each primer set is shown in brackets.

| F-NIRVS | R-NIRVS |
|---|---|
| **AlbFlavi12_17 (436 bp)**<br>5617-91F:<br>TTTCTACTGCCTCGCCATGA<br>5617CD-R:<br>GACGCATCCTAATTGTTCCGA | **AlbRha1 (1090 bp)**<br>AR1_Fext:<br>GGAGTTGCTGCCTCGGTC<br>AR1_Rext:<br>GCATTTCCTGGGCTCCTAAGT |
| **AlbFlavi1, AlbFlavi12_17 (233 bp-AlbFlavi1; 1262 bp-AlbFlavi12_17)**<br>5617-91F:<br>TTTCTACTGCCTCGCCATGA<br>5617-91R:<br>GAGTTGAATGGAGGAAGTCGTG | **AlbRha7 (862 bp)**<br>AR7_Fext:<br>CGAGAGAAGGTGGACTGGTT<br>AR7_Rext:<br>ACAGTTCGTCACGCCACTTA |
| **AlbFlavi10 (1583 bp)**<br>5171Fext:<br>CACCCACATCCGAAAGCTTC<br>5171Rex:<br>TTCCCGCGACCAGTATTCTT | **AlbRha14 (350 bp)**<br>AR14_Fext:<br>TAACTGTTCGCTAGTGGACTCG<br>AR14_Rext:<br>GCTTCAAACATTGCGCGTGA |
| **AlbFlavi2 (960 bp)**<br>157AF:<br>TCACAAACGCATGCTACACC<br>157AR:<br>TTCATTTGAGAGCAAGCGGG | **AlbRha36 (829 bp)**<br>AR36_Fext:<br>CAACAACCGCGAGAAGAAGC<br>AR36_Rext:<br>AATACCATTCCAGGGCGTCC |
| **AlbFlavi36 (1055 bp)**<br>14636EF:<br>AAGTTCGTGTTTTGGGTGCA<br>14636ER:<br>GATGCGCTCTCCTACTCACT | **AlbRha52 (968 bp)**<br>AR52_F4:<br>GAGAAGCCAATGACCCTGTGT<br>AR52_Rext:<br>GATTGACTGATGGACCAAGAACA |
| **AlbFlavi4 (690 bp)**<br>1256F:<br>AGGAGCGAAAAGTTCTTGGT<br>1256R:<br>TGATTCGACAGACCCGGAC | **AlbRha85 (638 bp)**<br>AR85_F2:<br>GACCCCTCTGTCCTGGATCA<br>AR85_Rext:<br>TCGAGCCCCATATTTTGAAGC |
| **AlbFlavi8, AlbFlavi41 (681 bp)**<br>4896-8815F:<br>CCGTGACGCTTGATGAGTTT<br>4896-8815Rext:<br>TGGTACTATCAACGGCATCTCT | |

# Viral genome of Arboviruses and ISVs

I developed a viral genome database including arboviruses as available in NCBI Viral Genome Browser (VGB) [87] and the ViruSITE database [164] by October,10th 2019.
A total of 3825 and 3809 Riboviria taxon-ids were selected from the NCBI Viral Genome Browser (VGB) [87] and the ViruSITE [164] databases, respectively. Among these, 3802 taxon-ids were shared by the two databases, while 23 and 7 taxon-ids were specific of NCBI VGB or of ViruSITE, respectively. I enriched the viral database with information regarding the host species using the Virus-Host Classifier developed by Kitson et al. in 2019 [89] and with the lineage of each taxon-id using the NCBI taxonomy toolkit [90].
In the next paragraphs, I will describe the viral species included in the database, keeping them separated based on whether they are arboviruses or ISVs.

To identify arboviruses, I followed the classification proposed by Go et al. in 2014 [165], which describes arboviruses in nine genus from six viral families (**Table 15**). Using this classification 309 arboviral taxon-ids were maintained in my viral database. At the moment, the only DNA virus considered as arbovirus is the African Swine Fever Virus [166]. Its reference genome was added to the database.

**Table 15**. Taxon-ids selected in the viral database based on the classification proposed by Go et al. 2014 [165].

| Viral Genome | Family | Genus | Num tax-ids |
|---|---|---|---|
| Single-stranded positive-sense RNA | Togaviridae | Alphavirus | 33 |
| | Flaviviridae | Flavivirus | 93 |
| Single-stranded negative-sense RNA | Bunyaviridae | Orthobunyavirus | 78 |
| | | Nairovirus | 0 |
| | | Phlebovirus | 29 |
| | | Tospovirus | 23 |
| | Rhabdoviridae | Vesiculovirus | 16 |
| | Orthomixoviridae | Thogotovirus | 3 |
| Double-stranded RNA | Reoviridae | Orbivirus | 32 |

| | |
|---|---|
| Coltivirus | 2 |
| Total | 309 |

Currently, a database of ISVs does not exist. Thus, to include as many ISVs as possible in my database, I searched the NCBI PubMed from 2015 to October 2019 using the keyword 'insect specific viruses'. I selected a total of 14 publications, which I investigated for ISVs (**Table 16**).

**Table 16**. Articles selected for the identification of ISVs.

| # | Reference | Title | Authors | Year |
|---|---|---|---|---|
| 1 | [165] | Zoonotic encephalitides caused by arboviruses: transmission and epidemiology of alphaviruses and flaviviruses | Go et al. | 2014 |
| 2 | [167] | Insect-specific flaviviruses: A systematic review of their discovery, host range, mode of transmission, superinfection exclusion potential and genomic organization | Blitvich & Firth | 2015 |
| 3 | [12] | Insect-specific virus discovery: Significance for the arbovirus community | Bolling et al. | 2015 |
| 4 | [168] | Arboviral screening of invasive *Aedes* species in northeastern Turkey: West Nile virus circulation and detection of insect-only viruses | Akiner et al. | 2016 |
| 5 | [169] | Discovery and characterisation of a new insect-specific bunyavirus from Culex mosquitoes captured in northern Australia | Hobson-Peters et al. | 2016 |
| 6 | [170] | Insect-specific flaviviruses, a worldwide widespread group of viruses only detected in insects | Calzolari et al. | 2016 |
| 7 | [171] | West African Anopheles gambiae mosquitoes harbor a taxonomically diverse virome including new insect-specific flaviviruses, mononegaviruses, and totiviruses | Fauver et al. | 2016 |
| 8 | [172] | Mosquito-specific and mosquito-borne viruses: evolution, infection, and host defense | Halbach et al. | 2017 |
| 9 | [173] | Genetic characterization, molecular epidemiology, and phylogenetic relationships of insect-specific viruses in the taxon Negevirus | Nunes et al. | 2017 |
| 10 | [174] | Characterization of three new insect-specific flaviviruses: Their relationship to the mosquito-borne flavivirus pathogens | Guzman et al. | 2018 |

| | | | | |
|---|---|---|---|---|
| 11 | [175] | Arboviral screening of invasive *Aedes* species in northeastern Turkey: West nile virus circulation and detection of insect-only viruses | Agboli et al. | 2019 |
| 12 | [176] | The discovery and global distribution of novel mosquito-associated viruses in the last decade (2007-2017) | Atoni et al. | 2019 |
| 13 | [177] | Cell-Fusing Agent Virus Reduces Arbovirus Dissemination in *Aedes aegypti* Mosquitoes In Vivo | Baidaliuk et al. | 2019 |
| 14 | [14] | Insect-specific virus evolution and potential effects on vector competence | Öhlund et al. | 2019 |

Taxon-ids of viruses for which only the nucleotide sequence ID was available were searched with the NCBI E-utility tool [178]. Few unresolved cases were searched manually in NCBI online. Briefly, among the manually checked 22 taxon-ids, 11 showed to be already included in the database with other names, while 4 were not included because they correspond to unverified sequences or partial cds.

The final database is composed of 440 taxon-ids of which 264 arboviruses and 176 ISVs. For all these taxon-ids, the accession number of the reference genome was extracted from NCBI assembly refseq database using the NCBI E-utility tool [178]. If the reference genome was not present in refseq, the taxon-id was searched in the genbank assembly database or used to extract all the accession numbers of nucleotide sequences present in the NCBI nucleotide database. The obtained sequences were filtered so that to include in the title the terms ('complete' or 'genomic'), ('sequence' or ‘genome’) and not the terms ('cds' and ‘UNVERIFIED').

After this filtering, 25 taxon-ids were excluded from the database.

199 of the remaining taxon-ids had more than one associated sequence. This could be due to the presence of various strain sequence IDs for the same segment or for the presence of more segments for the same virus or both. In case of multiple sequences, MAFFT [116] was used to identify the longest sequence among sequences with a percentage of identity higher than 90%.

The final database includes 990 nucleotide sequences corresponding to 409 taxon-ids (**Table 17**).

**Table 17**. Arboviruses and ISVs taxon ids.

| Viral name | Taxid | Viral name | Taxid |
|---|---|---|---|
| Groundnut bud necrosis virus | 198612 | Heartland virus | 1216928 |
| Murray Valley encephalitis virus | 11079 | Umatilla virus | 40060 |
| Japanese encephalitis virus | 11072 | Middelburg virus | 11023 |
| Venezuelan equine encephalitis virus | 11036 | Maraba virus | 1046251 |
| Dengue virus 2 | 11060 | Vesicular stomatitis Alagoas virus | 198833 |
| Dengue virus 3 | 11069 | Malpais Spring vesiculovirus | 1972570 |
| Dengue virus 1 | 11053 | Yug Bogdanovac vesiculovirus | 1972567 |
| Onyong-nyong virus | 2169701 | Perinet vesiculovirus | 1972569 |
| Ross River virus | 11029 | Jutiapa virus | 64299 |
| Sindbis virus | 11034 | Cacipacore virus | 64305 |
| Vesicular stomatitis Indiana virus strain 98COE | 11277 | Sokoluk virus | 64317 |
| Tick-borne encephalitis virus | 11084 | Bhanja virus | 1213620 |
| Barmah Forest virus | 11020 | Wad Medani virus | 40067 |
| Louping ill virus | 11086 | Chenuda virus | 40065 |
| Bunyamwera virus | 35304 | Chobar Gorge virus | 1679172 |
| Yellow fever virus | 11089 | Spanish goat encephalitis virus | 1691889 |
| Tomato spotted wilt tospovirus | 1933298 | Chrysanthemum stem necrosis virus | 83871 |
| Dengue virus 4 | 11070 | Paraiso Escondido virus | 1566298 |
| Semliki Forest virus | 11033 | Cocal virus | 50713 |
| Mayaro virus | 59301 | Potiskum virus | 64314 |
| Sleeping disease virus | 78540 | Saboya virus | 64284 |
| Peanut bud necrosis virus | 40687 | Spondweni virus | 64318 |
| Impatiens necrotic spot virus | 11612 | Adana virus | 1611877 |
| Modoc virus | 64300 | Edge Hill virus | 64296 |
| Rio Bravo virus | 64285 | Tospovirus kiwifruit/YXW/2014 | 1857323 |
| Apoi virus | 64280 | Huangpi Tick Virus 2 | 1608048 |
| Powassan virus | 11083 | Zucchini lethal chlorosis virus | 83872 |
| Langat virus | 11085 | New Mapoon virus | 499854 |
| Eyach virus | 62352 | Kokobera virus | 44024 |
| Watermelon silver mottle tospovirus | 1933300 | Bouboui virus | 64295 |
| Eastern equine encephalitis virus | 11021 | Uganda S virus | 64297 |
| Aura virus | 44158 | Jugra virus | 64309 |
| Western equine encephalitis virus | 11039 | Meaban virus | 35279 |
| Salmon pancreas disease virus | 84589 | Gadgets Gully virus | 64307 |
| Tamana bat virus | 161675 | Kadam virus | 64310 |
| La Crosse virus | 11577 | Saumarez Reef virus | 40012 |
| Montana myotis leukoencephalitis virus | 64312 | Pepper chlorotic spot virus | 1414655 |
| Chikungunya virus | 37124 | Orbivirus SX-2017a | 1955493 |
| Colorado tick fever virus | 46839 | Melon severe mosaic tospovirus | 485724 |
| Yokose virus | 64294 | Urucuri virus | 1926502 |
| Omsk hemorrhagic fever virus | 12542 | Ambe virus | 1926500 |
| Uukuniemi virus | 11591 | Anhanga virus | 904722 |
| Oropouche virus | 118655 | Munguba virus | 1048854 |

| | |
|---|---|
| Palyam virus | 40059 |
| St Croix River virus | 104581 |
| African horse sickness virus | 40050 |
| Bluetongue virus | 40051 |
| Sandfly fever Naples virus | 206160 |
| Thogoto thogotovirus | 11569 |
| Usutu virus | 64286 |
| Getah virus | 59300 |
| Saint Louis encephalitis virus | 11080 |
| Yunnan orbivirus | 306276 |
| Peruvian horse sickness virus | 356862 |
| Capsicum chlorosis virus | 163325 |
| Melon yellow spot virus | 89471 |
| Entebbe bat virus | 64283 |
| Sepik virus | 44026 |
| Aroa virus | 64303 |
| Akabane virus | 70566 |
| West Nile virus lineage 2 | 11082 |
| Tomato zonate spot virus | 460926 |
| Zika virus strain Natal RGN | 64320 |
| Kedougou virus | 64311 |
| Bagaza virus | 64290 |
| Highlands J virus | 11024 |
| Wesselsbron virus | 164416 |
| Epizootic hemorrhagic disease virus (serotype 1 / strain New Jersey) | 449133 |
| Fort Morgan virus | 48544 |
| Rift Valley fever virus | 11588 |
| Great Island virus | 204269 |
| Chandiru virus | 629725 |
| Sandfly fever Turkey virus | 688699 |
| Aguacate virus | 1006583 |
| Groundnut ringspot and Tomato chlorotic spot virus reassortant | 1027232 |
| Tembusu virus | 64293 |
| Ndumu virus | 59302 |
| Southern elephant seal virus | 1159195 |
| Whataroa virus | 48543 |
| Bebaru virus | 59305 |
| Bean necrotic mosaic virus | 1033976 |
| SFTS virus HB29 | 992212 |
| Aino virus | 11582 |
| Sathuperi orthobunyavirus | 159141 |
| Shamonda orthobunyavirus | 159150 |
| Simbu orthobunyavirus | 35306 |
| Ntaya virus | 64292 |
| Chandipura virus | 11272 |
| Isfahan virus | 290008 |
| Wallal virus | 40061 |
| Tapara virus | 1926501 |
| Uriurana virus | 1055750 |
| Phnom Penh bat virus | 64894 |
| Yaounde virus | 64319 |
| Dhori thogotovirus | 11318 |
| Oriboca virus | 192199 |
| Caraparu virus | 192196 |
| Bwamba orthobunyavirus | 35310 |
| Nyando virus | 35316 |
| Capim virus | 35312 |
| Kaeng Khoi virus | 307164 |
| Marituba virus | 292278 |
| Madrid virus | 348013 |
| Guaroa virus | 80941 |
| Morreton vesiculovirus | 1972565 |
| Wolkberg virus | 1867943 |
| Tomato chlorotic spot virus | 12851 |
| Kabuto mountain virus | 1851087 |
| Calla lily chlorotic spot virus | 309542 |
| Zerdali virus | 1764086 |
| Toros virus | 1764085 |
| Arrabida virus | 1457322 |
| Iris yellow spot virus | 60456 |
| Carajas virus | 239239 |
| Piry virus | 11274 |
| Radi vesiculovirus | 1972566 |
| Watermelon bud necrosis virus | 76052 |
| Corriparta virus | 40053 |
| Equine encephalosis virus | 201490 |
| Eubenangee virus | 40056 |
| Lebombo virus | 40057 |
| Orungo virus | 40058 |
| Warrego virus | 40062 |
| Cabassou virus | 60879 |
| Everglades virus | 2083198 |
| Mucambo virus | 60875 |
| Pixuna virus | 60876 |
| Rio Negro virus | 332097 |
| Tonate virus | 60877 |
| Alajuela virus | 1552846 |
| Lukuni virus | 1678227 |
| Anopheles B virus | 35308 |
| Batama virus | 611709 |
| Bimiti virus | 1678224 |
| Catu virus | 1678225 |
| Gamboa virus | 35313 |
| Guajara orthobunyavirus | 1933272 |

| | |
|---|---|
| Mobuck virus | 1408137 |
| Razdan virus | 1405807 |
| Changuinola virus | 40052 |
| American bat vesiculovirus TFFN-2013 | 1344113 |
| Tyuleniy virus | 40004 |
| Kama virus | 1456752 |
| Arumowot virus | 904698 |
| Madariaga virus | 1440170 |
| Cat Que virus | 1495866 |
| Vesicular stomatitis New Jersey virus | 11280 |
| Aedes camptorhynchus reo-like virus | 2010269 |
| Bellavista virus | 1856565 |
| Kyasanur Forest disease virus | 33743 |
| Royal Farm virus | 64288 |
| Kaisodi virus | 1564120 |
| Oz virus | 2137161 |
| Alstroemeria yellow spot virus | 2212644 |
| Rocio virus | 64315 |
| Ilheus virus | 59563 |
| Guangxi orbivirus | 2306813 |
| Thimiri orthobunyavirus | 1819305 |
| Umbre virus | 552554 |
| Banzi virus | 38837 |
| Bukalasa bat virus | 64281 |
| Carey Island virus | 64289 |
| Cowbone Ridge virus | 64298 |
| Dakar bat virus | 64282 |
| Israel turkey meningoencephalomyelitis virus | 64291 |
| Koutango virus | 44025 |
| Sal Vieja virus | 64301 |
| San Perlita virus | 64302 |
| Wongorr virus | 47465 |
| Frijoles virus VP-161A | 426788 |
| Groundnut yellow spot virus | 345030 |
| Trocara virus | 135246 |
| Una virus | 59304 |
| M'Poko virus | 442712 |
| Groundnut ringspot virus | 12675 |
| Mukawa virus | 1569922 |
| Tensaw virus | 273347 |
| Peaton virus | 159151 |
| Sabo virus | 159138 |
| Sango virus | 159152 |
| Jamestown Canyon virus | 35511 |
| Leanyer virus | 999729 |
| Anhembi virus | 273355 |
| Iaco virus | 273356 |
| Guama virus | 1678234 |
| Kairi virus | 80939 |
| Koongol virus | 35314 |
| Mosso das Pedras virus | 2083199 |
| Main Drain virus | 80938 |
| Tete orthobunyavirus | 35319 |
| Wyeomyia orthobunyavirus | 273350 |
| Punta Toro virus | 11587 |
| Polygonum ringspot tospovirus | 430606 |
| Jurona vesiculovirus | 1972568 |
| Culex rhabdo-like 1 | 2010277 |
| Culex mononega-like virus 2 | 2010272 |
| Nanay virus | 1254420 |
| Culex theileri flavivirus | 1244563 |
| Kampung Karu virus | 2045186 |
| Culex pipiens pallens densovirus | 465914 |
| Culex densovirus 0507JS11 | 642590 |
| Culex Y virus | 1230254 |
| Dezidougou virus | 1170421 |
| Santana virus | 1170427 |
| Mosquito X virus | 1237117 |
| Kamphang Phet virus | 1332247 |
| Culicine-associated Z virus | 1398940 |
| North Creek virus | 1406950 |
| Mosinovirus | 1545703 |
| Wuhan Mosquito Virus 3 | 1608128 |
| Wuhan Mosquito Virus 5 | 1608130 |
| Wuhan Mosquito Virus 7 | 1608132 |
| Jiangxia Mosquito Virus 1 | 1608051 |
| Jiangxia Mosquito Virus 2 | 1608052 |
| Xinzhou Mosquito Virus | 1608142 |
| Qingnian Mosquito Virus | 1608059 |
| Terena virus | 1795443 |
| Anopheles Cypovirus | 1769781 |
| Anopheles flavivirus variant2 | 1903341 |
| Anopheles totivirus | 1903415 |
| Bolahun virus variant 1 | 1903426 |
| Bolahun virus variant 2 | 1903560 |
| Gambie virus | 1903427 |
| Chaq virus-like1 | 1903431 |
| Chaq virus-like2 | 1903533 |
| Chaq virus-like3 | 1903534 |
| dsRNA virus-like 1 | 1903437 |
| dsRNA virus-like 2 | 1903535 |
| dsRNA virus-like 3 | 1903536 |
| dsRNA virus-like 4 | 1903537 |
| dsRNA virus-like 5 | 1903538 |

| | |
|---|---|
| Sororoca virus | 273354 |
| Cachoeira Porteira virus | 1138490 |
| Jatobal virus | 150058 |
| Batai virus | 80942 |
| Schmallenberg virus | 1133363 |
| Ilesha virus | 273341 |
| Ingwavuma virus | 159145 |
| Mermet virus | 159147 |
| Utinga virus | 159144 |
| Buttonwillow virus | 159140 |
| Pacui virus | 1538454 |
| Rio Preto da Eva virus | 1538455 |
| Guertu virus | 1763596 |
| Enseada virus | 1821545 |
| Fort Sherman virus | 273345 |
| Cache Valley virus | 80935 |
| Keystone virus | 35514 |
| Lumbo virus | 80940 |
| Melao virus | 35515 |
| San Angelo virus | 45767 |
| Serra do Navio virus | 45768 |
| Potosi virus | 273360 |
| Birao virus | 273358 |
| Bozo virus | 273349 |
| Laurel Lake virus | 2027354 |
| Zegla virus | 2303488 |
| Patois virus | 35318 |
| Tacaiuma orthobunyavirus | 611707 |
| Shuni orthobunyavirus | 159148 |
| Koyama Hill virus | 1435294 |
| Cholul virus | 1093160 |
| Salt ash virus | 1406136 |
| El Huayo virus | 1769592 |
| Anadyr virus | 1642852 |
| Parry's Lagoon virus | 1853202 |
| Calbertado virus | 537023 |
| Long Pine Key virus | 2045185 |
| La Tina virus | 2045187 |
| Marisma mosquito virus | 1105173 |
| CFLV NEI 1 | 11093 |
| Anopheles hinesorum orbivirus | 2026602 |
| Anopheles annulipes orbivirus | 2026603 |
| Skunk River virus | 2488682 |
| Cell fusing agent virus | 31658 |
| Kamiti River virus | 218849 |
| Culex flavivirus | 390844 |
| dsRNA virus-like 6 | 1903539 |
| dsRNA virus-like 7 | 1903540 |
| dsRNA virus-like 8 | 1903541 |
| dsRNA virus-like 9 | 1903542 |
| Endornavirus-like 1 | 1903433 |
| Endornavirus-like 2 | 1903543 |
| Endornavirus-like 3 | 1903544 |
| Mononegavirus-like 1 | 1903435 |
| Mononegavirus-like 2 | 1903545 |
| Mononegavirus-like 3 | 1903546 |
| Mononegavirus-like 4 | 1903547 |
| Mononegavirus-like 5 | 1903548 |
| Partitivirus-like 3 | 1903550 |
| ssRNA virus-like 1 | 1903553 |
| ssRNA virus-like 2 | 1903554 |
| ssRNA virus-like 3 | 1903555 |
| ssRNA virus-like 4 | 1903556 |
| ssRNA virus-like 5 | 1903557 |
| ssRNA virus-like 6 | 1903558 |
| Guaico Culex virus | 1665361 |
| Hubei sobemo-like virus 39 | 1923226 |
| Hubei noda-like virus 12 | 1922968 |
| Hubei partiti-like virus 22 | 1922978 |
| Wuhan Mosquito Virus 4 | 1608129 |
| Shuangao partiti-like virus 1 | 1923472 |
| Sinu virus | 1927799 |
| Panmunjeom flavivirus | 1928710 |
| Omono River virus | 753758 |
| Ohlsdorf virus | 2040592 |
| Bontang Baru virus | 2109378 |
| Culex phasma-like virus | 2010276 |
| Ngewotan virus | 1265748 |
| Aedes alboannulatus reo-like virus | 2010267 |
| Shuangao chryso-like virus 1 | 1923465 |
| Wilkie qin-like virus | 2010285 |
| Culex mononega-like virus 1 | 2010271 |
| Wuhan Mosquito Virus 6 | 1608131 |
| Culex Luteo-like virus | 2010270 |
| Hubei chryso-like virus 1 | 1922855 |
| Lobeira virus | 2027352 |
| Merida-like Turkey virus | 2053815 |
| Barkedji virus | 478577 |
| Houston virus | 1332246 |
| Sabethes flavivirus | 2491660 |
| Hubei reo-like virus 7 | 1923182 |
| Armigeres subalbatus virus SaX06-AK20 | 556524 |

| | |
|---|---|
| Quang Binh virus | 643132 |
| Stretch Lagoon orbivirus | 559180 |
| Aedes flavivirus | 390845 |
| Donggang virus | 985683 |
| Chaoyang virus | 631267 |
| Eilat virus | 1231903 |
| Mosquito flavivirus | 673515 |
| Brazoran virus | 1368616 |
| Murrumbidgee virus | 1406134 |
| Fengkai orbivirus | 1692107 |
| Tibet orbivirus | 1428763 |
| Parramatta River virus | 1708654 |
| Mercadeo virus | 1708574 |
| Hanko virus | 1125677 |
| Tai Forest alphavirus | 1930825 |
| Palm Creek virus | 1302179 |
| Nounane virus | 486494 |
| T'Ho virus | 577122 |
| Ochlerotatus caspius flavivirus | 1244565 |
| Culex Negev-like virus 3 | 2010275 |
| Aedes alboannulatus toti-like virus | 2010265 |
| African swine fever virus | 10497 |
| Culex originated Tymoviridae-like virus | 1236047 |
| Anopheline-associated C virus | 1398939 |
| Anopheles minimus iridovirus | 1465751 |
| Tanay virus | 1489714 |
| Mosquito Circovirus | 1611039 |
| Anopheles flavivirus variant1 | 1903340 |
| Zhejiang mosquito virus 3 | 1923779 |
| Big Cypress virus | 1955196 |
| Biratnagar virus | 1955197 |
| San Bernardo virus | 1955199 |
| Cordoba virus | 1955175 |
| Fort Crockett virus | 1955198 |
| Wilkie Partiti-like virus 2 | 2010284 |
| Wilkie narna-like virus 2 | 2010281 |
| Point Douro narna-like virus | 2010279 |
| Wilkie Partiti-like virus 1 | 2010283 |
| Culex Negev-like virus 2 | 2010274 |
| Culex Negev-like virus 1 | 2010273 |
| Leschenault Partiti-like virus | 2010278 |
| Wilkie narna-like virus 1 | 2010280 |
| Aedes camptorhynchus negev-like virus | 2010268 |

# VIR pipeline implementation

**Table 18**. ViR output. In each raw from left to right: sample name, number of reads obtained using ViR_RefineCandidate **(a)**, number of "Ungrouped" reads **(b)**, number of groups created by ViR_SolveDispersion **(c)**, number of groups supporting a reference nrEVE polymorphism **(d),** number of candidate viral integrations with dubious Viral ID **(e),** number of candidate novel nrEVEs (novel nrEVEs).

| Sample name | (a) | (b) | (c) | (d) | (e) | novel nrEVEs |
|---|---|---|---|---|---|---|
| RosAnsR1 | 41 | 6 | 9 | 7 | 1 | 1 |
| RosAnsR2 | 35 | 6 | 8 | 3 | 0 | 5 |
| BraR1 | 22 | 9 | 4 | 3 | 1 | 0 |
| BraR2 | 29 | 2 | 6 | 6 | 0 | 0 |
| StpR1 | 32 | 7 | 4 | 3 | 1 | 0 |
| StpR2 | 37 | 3 | 8 | 4 | 0 | 4 |
| TamR1 | 28 | 9 | 5 | 3 | 2 | 0 |
| TamR2 | 44 | 4 | 8 | 5 | 1 | 3 |
| ReunionR1 | 10 | 3 | 3 | 2 | 0 | 1 |
| ReunionR2 | 9 | 3 | 2 | 1 | 0 | 1 |
| ReunionR3 | 25 | 15 | 5 | 4 | 1 | 0 |
| Tam-1_LIN210A145 | 4 | 2 | 1 | 1 | 0 | 0 |
| Tam-10_LIN210A154 | 3 | 3 | 0 | 0 | 0 | 0 |
| Tam-11_LIN210A155 | 17 | 3 | 4 | 3 | 1 | 0 |
| Tam-12_LIN210A156 | 7 | 2 | 2 | 0 | 1 | 1 |
| Tam-13_LIN210A157 | 4 | 2 | 1 | 0 | 0 | 1 |
| Tam-14_LIN210A158 | 2 | 2 | 0 | 0 | 0 | 0 |
| Tam-15_LIN210A159 | 4 | 1 | 1 | 1 | 0 | 0 |
| Tam-16_LIN210A160 | 5 | 1 | 1 | 1 | 0 | 0 |
| Tam-17_LIN210A161 | 10 | 2 | 3 | 2 | 0 | 1 |
| Tam-18_LIN210A162 | 2 | 0 | 1 | 1 | 0 | 0 |
| Tam-19_LIN210A163 | 7 | 2 | 1 | 0 | 0 | 1 |
| Tam-20_LIN210A164 | 4 | 2 | 1 | 1 | 0 | 0 |
| Tam-21_LIN210A165 | 3 | 3 | 0 | 0 | 0 | 0 |
| Tam-22_LIN210A166 | 3 | 1 | 1 | 1 | 0 | 0 |
| Tam-23_LIN210A167 | 7 | 4 | 1 | 1 | 0 | 0 |
| Tam-24_LIN210A168 | 3 | 3 | 0 | 0 | 0 | 0 |
| Tam-3_LIN210A147 | 4 | 2 | 1 | 0 | 1 | 0 |
| Tam-4_LIN210A148 | 1 | 0 | 0 | 0 | 0 | 0 |
| Tam-5_LIN210A149 | 14 | 1 | 3 | 2 | 1 | 0 |
| Tam-7_LIN210A151 | 5 | 5 | 0 | 0 | 0 | 0 |
| Tam-8_LIN210A152 | 5 | 1 | 2 | 1 | 1 | 0 |
| Tam-9_LIN210A153 | 7 | 0 | 2 | 2 | 0 | 0 |
| TapachulaR1 | 18 | 11 | 3 | 1 | 0 | 2 |
| TapachulaR2 | 0 | 0 | 0 | 0 | 0 | 0 |
| TapachulaR3 | 0 | 0 | 0 | 0 | 0 | 0 |
| JP-1_LIN210A121 | 0 | 0 | 0 | 0 | 0 | 0 |

| | | | | | | |
|---|---|---|---|---|---|---|
| JP-2_LIN210A122 | 1 | 1 | 0 | 0 | 0 | 0 |
| JP-3_LIN210A123 | 3 | 3 | 0 | 0 | 0 | 0 |
| JP-4_LIN210A124 | 2 | 2 | 0 | 0 | 0 | 0 |
| JP-5_LIN210A125 | 4 | 4 | 0 | 0 | 0 | 0 |
| JP-6_LIN210A126 | 5 | 5 | 0 | 0 | 0 | 0 |
| JP-7_LIN210A127 | 4 | 1 | 1 | 1 | 0 | 0 |
| JP-8_LIN210A128 | 4 | 1 | 1 | 0 | 0 | 1 |
| JP-9_LIN210A129 | 9 | 3 | 3 | 1 | 1 | 1 |
| JP-10_LIN210A130 | 3 | 3 | 0 | 0 | 0 | 0 |
| JP-11_LIN210A131 | 0 | 0 | 0 | 0 | 0 | 0 |
| JP-12_LIN210A132 | 9 | 3 | 2 | 2 | 0 | 0 |
| JP-13_LIN210A133 | 3 | 1 | 1 | 1 | 0 | 0 |
| JP-14_LIN210A134 | 9 | 3 | 2 | 2 | 0 | 0 |
| JP-15_LIN210A135 | 2 | 0 | 1 | 0 | 0 | 1 |
| JP-16_LIN210A136 | 5 | 1 | 2 | 0 | 1 | 1 |
| JP-17_LIN210A137 | 9 | 2 | 2 | 1 | 1 | 0 |
| JP-18_LIN210A138 | 5 | 1 | 1 | 1 | 0 | 0 |
| JP-19_LIN210A139 | 10 | 2 | 2 | 1 | 1 | 0 |
| JP-20_LIN210A140 | 9 | 2 | 2 | 2 | 0 | 0 |
| JP-21_LIN210A141 | 5 | 3 | 1 | 1 | 0 | 0 |
| JP-22_LIN210A142 | 4 | 4 | 0 | 0 | 0 | 0 |
| JP-23_LIN210A143 | 12 | 1 | 3 | 2 | 1 | 0 |
| JP-24_LIN210A144 | 6 | 2 | 2 | 1 | 0 | 1 |
| ChiangMaiR1 | 67 | 15 | 5 | 3 | 0 | 2 |
| ChiangMaiR2 | 11 | 4 | 3 | 2 | 0 | 1 |
| ChiangMaiR3 | 2 | 0 | 1 | 0 | 0 | 1 |
| CremaR1 | 8 | 4 | 2 | 1 | 1 | 0 |
| CremaR2 | 6 | 1 | 1 | 1 | 0 | 0 |
| CremaR3 | 23 | 13 | 5 | 3 | 1 | 1 |

# Ontology vocabularies

The vocabularies used in the nrEVEs ontology are divided in classes (**Table 19**), object properties (**Table 20**) and data properties (**Table 21**). The terms are extracted by existent ontologies in BioPortal [148] and LOV[149] or created *ex novo*.

**Table 19.** Classes of the nrEVEs ontology.

| Source Ontology | Class | ID |
|---|---|---|
| GBOL | CDS | http://gbol.life/0.1/CDS |
| GBOL | Blast | http://semantics.systemsbiology.nl/sapp/0.1/Blast |
| GBOL | Database | http://gbol.life/0.1/Database |
| GBOL | Sequence | http://gbol.life/0.1/Sequence |
| GBOL | Feature | http://gbol.life/0.1/Feature |
| GBOL | GeneralFeature | http://gbol.life/0.1/ GeneralFeature |
| GBOL | NAFeature | http://gbol.life/0.1/ NAFeature |
| GBOL | SequenceAnnotation | http://gbol.life/0.1/SequenceAnnotation |
| GBOL | GenomicFeature | http://gbol.life/0.1/GenomicFeature |
| Uniprot KB | Citation | http://purl.uniprot.org/core/Citation |
| Uniprot KB | Taxon | http://purl.uniprot.org/core/Taxon |
| Uniprot KB | Protein | http://purl.uniprot.org/core/Protein |
| FALDO | Region | http://biohackathon.org/resource/faldo#Region |
| FALDO | InBetweenPosition | http://biohackathon.org/resource/faldo#InBetweenPosition |
| FALDO | ExactPosition | http://biohackathon.org/resource/faldo#ExactPosition |
| RSA | ReferenceSequence | http://rdf.biosemantics.org/ontologies/rsa#ReferenceSequence |
| RSA | GenomeAssembly | http://rdf.biosemantics.org/ontologies/rsa#GenomeAssembly |
| SO | transposable_element | http://purl.obolibrary.org/obo/SO_0000101 |
| SO | DNA_transposon | http://purl.obolibrary.org/obo/SO_0000182 |
| SO | helitron | http://purl.obolibrary.org/obo/SO_0000544 |
| SO | terminal_inverted_repeat_element | http://purl.obolibrary.org/obo/SO_0000208 |
| SO | retrotransposon | http://purl.obolibrary.org/obo/SO_0000180 |
| SO | LTR_retrotransposon | http://purl.obolibrary.org/obo/SO_0000186 |
| SO | non_LTR_retrotransposon | http://purl.obolibrary.org/obo/SO_0000189 |
| SO | LINE_element | http://purl.obolibrary.org/obo/SO_0000194 |
| SO | SINE_element | http://purl.obolibrary.org/obo/SO_0000206 |
| nrEVEs | Similarity | http://www.nrEVEdb.com/Similarity |
| nrEVEs | Annotated_sequence | http://www.nrEVEdb.com/Annotated_sequence |
| nrEVEs | Not_annotated_sequence | http://www.nrEVEdb.com/Not_annotated_sequence |

**Table 20.** Object properties of the nrEVEs ontology.

| Source Ontology | Property | ID |
|---|---|---|
| RO | part_of | http://purl.obolibrary.org/obo/BFO_0000050 |
| RO | is_upstream_of_sequence_of | http://purl.obolibrary.org/obo/RO_0002528 |
| RO | is_downstream_of_sequence_of | http://purl.obolibrary.org/obo/RO_0002529 |
| GBOL | protein | http://gbol.life/0.1/protein |
| GBOL | db | http://gbol.life/0.1/db |
| GBOL | derivedFrom | http://gbol.life/0.1/derivedFrom |
| GBOL | feature | http://gbol.life/0.1/feature |
| SIO | refersTo | http://semanticscience.org/resource/SIO_000628 |
| FALDO | location | http://biohackathon.org/resource/faldo#location |
| FALDO | begin | http://biohackathon.org/resource/faldo#begin |
| FALDO | end | http://biohackathon.org/resource/faldo#end |
| FALDO | before | http://biohackathon.org/resource/faldo#before |
| FALDO | after | http://biohackathon.org/resource/faldo#after |
| FALDO | reference | http://biohackathon.org/resource/faldo#reference |
| Uniprot KB | citation | http://purl.uniprot.org/core/citation |
| Uniprot KB | organism | http://purl.uniprot.org/core/organism |
| nrEVEs | proteinDb | http://www.nrEVEdb.com/proteinDb |
| nrEVEs | sequenceDb | http://www.nrEVEdb.com/sequenceDb |
| nrEVEs | assemblyCitation | http://www.nrEVEdb.com/assemblyCitation |
| nrEVEs | sequenceCitation | http://www.nrEVEdb.com/sequenceCitation |

**Table 21.** Data properties of the nrEVEs ontology.

| Source Ontology | Property | ID |
|---|---|---|
| GBOL | id | http://gbol.life/0.1/id |
| GBOL | releaseDate | http://gbol.life/0.1/releaseDate |
| GBOL | sequence | http://gbol.life/0.1/sequence |
| GBOL | standardName | http://gbol.life/0.1/standardName |
| GBOL | function | http://gbol.life/0.1/function |
| GBOL | alignmentLength | http://gbol.life/0.1/alignmentLength |
| GBOL | bitscore | http://gbol.life/0.1/bitscore |
| GBOL | evalue | http://gbol.life/0.1/evalue |
| GBOL | gaps | http://gbol.life/0.1/gaps |
| GBOL | mismatches | http://gbol.life/0.1/mismatches |
| GBOL | percIdentity | http://gbol.life/0.1/percIdentity |
| FALDO | position | http://biohackathon.org/resource/faldo#position |
| RSA | genBankID | http://rdf.biosemantics.org/ontologies/rsa#genBankID |
| RSA | refSeqID | http://rdf.biosemantics.org/ontologies/rsa#refSeqID |
| Uniprot KB | name | http://purl.uniprot.org/core/name |
| Uniprot KB | author | http://purl.uniprot.org/core/author |
| Uniprot KB | title | http://purl.uniprot.org/core/title |
| Uniprot KB | date | http://purl.uniprot.org/core/date |

# References


[1] World Health Organization, “WHO Report: Global Strategy for dengue prevention and control, 2012–2020,” 2012.

[2] W. R. Shaw and F. Catteruccia, “Vector biology meets disease control: using basic research to fight vector-borne diseases,” *Nat. Microbiol.*, vol. 4, no. 1, pp. 20–34, Jan. 2019, doi: 10.1038/s41564-018-0214-7.

[3] N. Vasilakis, A. Lambert, N. J. MacLachlan, and A. C. Brault, “Genomic Organization of Arboviral Families,” in *Arboviruses: Molecular Biology, Evolution and Control*, no. January, 2016, pp. 31–44.

[4] K. E. Olson and M. Bonizzoni, “Nonretroviral integrated RNA viruses in arthropod vectors: an occasional event or something more?,” *Curr. Opin. Insect Sci.*, vol. 22, no. 1, pp. 45–53, 2017, doi: 10.1016/j.cois.2017.05.010.

[5] R. A. Hall and J. Hobson-Peters, “Newly discovered mosquito viruses help control vector-borne viral diseases,” *Microbiol. Aust.*, vol. 39, no. 2, p. 72, 2018, doi: 10.1071/MA18020.

[6] L. Lambrechts and T. W. Scott, “Mode of transmission and the evolution of arbovirus virulence in mosquito vectors,” *Proc. R. Soc. B Biol. Sci.*, 2009, doi: 10.1098/rspb.2008.1709.

[7] W. S. Lee, J. A. Webster, E. T. Madzokere, E. B. Stephenson, and L. J. Herrero, “Mosquito antiviral defense mechanisms: A delicate balance between innate immunity and persistent viral infection,” *Parasites and Vectors*, 2019, doi: 10.1186/s13071-019-3433-8.

[8] J. S. Salas-Benito and M. De Nova-Ocampo, “Viral interference and persistence in mosquito-borne flaviviruses,” *Journal of Immunology Research*. 2015, doi: 10.1155/2015/873404.

[9] G. Cheng, Y. Liu, P. Wang, and X. Xiao, “Mosquito Defense Strategies against Viral Infection,” *Trends in Parasitology*. 2016, doi: 10.1016/j.pt.2015.09.009.

[10] J. Liu, L. Swevers, A. Kolliopoulou, and G. Smagghe, “Arboviruses and the Challenge to Establish Systemic and Persistent Infections in Competent Mosquito Vectors: The Interaction With the RNAi Mechanism,” *Front. Physiol.*, vol. 10, p. 890, 2019, doi: 10.3389/fphys.2019.00890.

[11] V. H. Peña-García, M. K. McCracken, and R. C. Christofferson, “Examining the Potential for South American Arboviruses to Spread Beyond the New World,” *Curr. Clin. Microbiol. Reports*, vol. 4, no. 4, pp. 208–217, Dec. 2017, doi: 10.1007/s40588-017-0076-4.

[12] B. G. Bolling, S. C. Weaver, R. B. Tesh, and N. Vasilakis, “Insect-

specific virus discovery: Significance for the arbovirus community," *Viruses*, vol. 7, no. 9, pp. 4911–4928, 2015, doi: 10.3390/v7092851.

[13] P. Fort *et al.*, "Fossil rhabdoviral sequences integrated into arthropod genomes: ontogeny, evolution, and potential functionality," *Mol. Biol. Evol.*, vol. 29, no. 1, pp. 381–390, 2012, doi: 10.1093/molbev/msr226.

[14] P. Öhlund, H. Lundén, and A.-L. Blomström, "Insect-specific virus evolution and potential effects on vector competence," *Virus Genes*, vol. 55, no. 2, pp. 127–137, Jan. 2019, doi: 10.1007/s11262-018-01629-9.

[15] F. A. A. Skuse, "The banded mosquito of Bengal," *Indian Museum Notes*, vol. 3 (for 189, p. 20, 1895.

[16] A. N. Clements and R. E. Harbach, "Controversies over the scientific name of the principal mosquito vector of yellow fever virus - expediency versus validity," *J. Vector Ecol.*, vol. 43, no. 1, pp. 1–14, Jun. 2018, doi: 10.1111/jvec.12277.

[17] S. C. Weaver and W. K. Reisen, "Present and future arboviral threats," *Antiviral Res.*, vol. 85, no. 2, pp. 328–345, Feb. 2010, doi: 10.1016/j.antiviral.2009.10.008.

[18] J. E. Brown *et al.*, "Human impacts have shaped historical and recent evolution in *Aedes aegypti*, the Dengue and Yellow Fever mosquitoe," *Evolution (N. Y).*, vol. 68, no. 2, pp. 514–525, Feb. 2014, doi: 10.1111/evo.12281.

[19] D. Rogers and S. Hay, "The climatic suitability for dengue transmission in continental Europe," Stockholm, 2012. doi: 10.2900/62095.

[20] P. F. Mattingly, "New Records and a New Species of the Subgenus Stegomyia (Diptera, Culicidae) from the Ethiopian Region," *Ann. Trop. Med. Parasitol.*, vol. 47, no. 3, pp. 294–298, Oct. 1953, doi: 10.1080/00034983.1953.11685571.

[21] D. Metselaar *et al.*, "An outbreak of type 2 dengue fever in the Seychelles, probably transmitted by *Aedes albopictus* (Skuse).," *Bull. World Health Organ.*, vol. 58, no. 6, pp. 937–43, 1980, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/6971192.

[22] S. A. Elliott, "*Aedes albopictus* in the Solomon and Santa Cruz Islands, South Pacific," *Trans. R. Soc. Trop. Med. Hyg.*, vol. 74, no. 6, pp. 747–748, Jan. 1980, doi: 10.1016/0035-9203(80)90192-3.

[23] V. Houé, M. Bonizzoni, and A.-B. Failloux, "Endogenous non-retroviral elements in genomes of *Aedes* mosquitoes and vector competence," *Emerg. Microbes Infect.*, vol. 8, no. 1, pp. 542–555, Jan. 2019, doi: 10.1080/22221751.2019.1599302.

[24] J. Adhami and P. Reiter, "Introduction and establishment of *Aedes (Stegomyia) albopictus* skuse (Diptera: Culicidae) in Albania.," *J. Am. Mosq. Control Assoc.*, vol. 14, no. 3, pp. 340–3, Sep. 1998, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/9813831.

[25] J. M. Medlock *et al.*, "An entomological review of invasive mosquitoes in Europe," *Bull. Entomol. Res.*, vol. 105, no. 6, pp. 637–663, Dec. 2015, doi: 10.1017/S0007485315000103.

[26] J. M. Medlock, D. Avenell, I. Barrass, and S. Leach, "Analysis of the potential for survival and seasonal activity of *Aedes albopictus* (Diptera: Culicidae) in the United Kingdom.," *J. Vector Ecol.*, vol. 31,

no. 2, pp. 292–304, Dec. 2006, doi: 10.3376/1081-1710(2006)31[292:aotpfs]2.0.co;2.

[27] S. M. Thomas, U. Obermayr, D. Fischer, J. Kreyling, and C. Beierkuhnlein, “Low-temperature threshold for egg survival of a post-diapause and non-diapause European aedine strain, *Aedes albopictus* (Diptera: Culicidae),” *Parasit. Vectors*, vol. 5, no. 1, p. 100, Dec. 2012, doi: 10.1186/1756-3305-5-100.

[28] Global Invasive Species Database, “100 of the World’s Worst Invasive Alien Species,” 2020. http://www.iucngisd.org/gisd/100_worst.php.

[29] C. Caminade *et al.*, “Suitability of European climate for the Asian tiger mosquito *Aedes albopictus*: recent trends and future scenarios,” *J. R. Soc. Interface*, vol. 9, no. 75, pp. 2708–2717, Oct. 2012, doi: 10.1098/rsif.2012.0138.

[30] D. Fischer, S. M. Thomas, M. Neteler, N. B. Tjaden, and C. Beierkuhnlein, “Climatic suitability of *Aedes albopictus* in Europe referring to climate change projections: comparison of mechanistic and correlative niche modelling approaches,” *Eurosurveillance*, vol. 19, no. 6, Feb. 2014, doi: 10.2807/1560-7917.ES2014.19.6.20696.

[31] J. C. Semenza and J. E. Suk, “Vector-borne diseases and climate change: a European perspective,” *FEMS Microbiol. Lett.*, vol. 365, no. 2, Jan. 2018, doi: 10.1093/femsle/fnx244.

[32] J. A. Frank and C. Feschotte, “Co-option of endogenous viral sequences for host cell function,” *Curr. Opin. Virol.*, vol. 25, pp. 81–89, 2017, doi: 10.1016/j.coviro.2017.07.021.

[33] T. Honda and K. Tomonaga, “Endogenous non-retroviral RNA virus elements evidence a novel type of antiviral immunity,” *Mob. Genet. Elements*, vol. 6, no. 3, pp. 1–6, 2016, doi: 10.1080/2159256X.2016.1165785.

[34] U. Palatini *et al.*, “Comparative genomics shows that viral integrations are abundant and express piRNAs in the arboviral vectors *Aedes aegypti* and *Aedes albopictus*,” *BMC Genomics*, vol. 18, no. 1, pp. 1–15, 2017, doi: 10.1186/s12864-017-3903-3.

[35] P. Aiewsakun and A. Katzourakis, “Endogenous viruses : Connecting recent and ancient viral evolution,” *Virology*, vol. 479–480, pp. 26–37, 2015, doi: 10.1016/j.virol.2015.02.011.

[36] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, “Basic local alignment search tool,” *J. Mol. Biol.*, vol. 215, no. 3, pp. 403–410, Oct. 1990, doi: 10.1016/S0022-2836(05)80360-2.

[37] H. Kondo, S. Chiba, K. Maruyama, I. B. Andika, and N. Suzuki, “A novel insect-infecting virga/nege-like virus group and its pervasive endogenization into insect genomes,” *Virus Res.*, vol. 262, pp. 37–47, 2017, doi: 10.1016/j.virusres.2017.11.020.

[38] Z. J. Whitfield *et al.*, “The diversity, structure, and function of heritable adaptive immunity sequences in the *Aedes aegypti* genome,” *Curr. Biol.*, vol. 27, no. 22, pp. 3511–3519, 2017, doi: 10.1016/j.cub.2017.09.067.

[39] A. M. ter Horst, J. C. Nigg, F. M. Dekker, and B. W. Falk, “Endogenous Viral Elements Are Widespread in Arthropod Genomes and Commonly Give Rise to PIWI-Interacting RNAs,” *J. Virol.*, vol.

93, no. 6, pp. e02124-18, Mar. 2019, doi: 10.1128/JVI.02124-18.
[40] X.-G. Chen *et al.*, "Genome sequence of the Asian Tiger mosquito, *Aedes albopictus*, reveals insights into its biology, genetics and evolution," *Proc. Natl. Acad. Sci.*, vol. 112, no. 44, pp. E5907–E5915, 2015, doi: 10.1073/pnas.1516410112.
[41] J. Cui and E. C. Holmes, "Endogenous RNA viruses of plants in insect genomes," *Virology*, vol. 427, no. 2, pp. 77–79, Jun. 2012, doi: 10.1016/j.virol.2012.02.014.
[42] A. Katzourakis and R. J. Gifford, "Endogenous viral elements in animal genomes," *PLoS Genet.*, vol. 6, no. 11, p. e1001191, 2010, doi: 10.1371/journal.pgen.1001191.
[43] A. G. Russo, A. G. Kelly, D. Enosi Tuipulotu, M. M. Tanaka, and P. A. White, "Novel insights into endogenous RNA viral elements in Ixodes scapularis and other arbovirus vector genomes," *Virus Evol.*, vol. 5, no. 1, pp. 1–18, 2019, doi: 10.1093/ve/vez010.
[44] P. J. Flynn and C. S. Moreau, "Assessing the Diversity of Endogenous Viruses Throughout Ant Genomes," *Front. Microbiol.*, vol. 10, no. May, 2019, doi: 10.3389/fmicb.2019.01139.
[45] C. D. Blair, K. E. Olson, and M. Bonizzoni, "The Widespread Occurrence and Potential Biological Roles of Endogenous Viral Elements in Insect Genomes," *Curr. Issues Mol. Biol.*, pp. 13–30, 2020, doi: 10.21775/cimb.034.013.
[46] Y. Chen, V. Williams, M. Filippova, V. Filippov, and P. Duerksen-Hughes, "Viral Carcinogenesis: Factors Inducing DNA Damage and Virus Integration," *Cancers (Basel).*, vol. 6, no. 4, pp. 2155–2186, Oct. 2014, doi: 10.3390/cancers6042155.
[47] X. Chen, J. Kost, and D. Li, "Comprehensive comparative analysis of methods and software for identifying viral integrations," *Brief. Bioinform.*, vol. 20, no. 6, pp. 2088–2097, Nov. 2019, doi: 10.1093/bib/bby070.
[48] D. W. Ho, K. M. Sze, and I. O. Ng, "Virus-Clip: a fast and memory-efficient viral integration site detection tool at single-base resolution with annotation capability," *Oncotarget*, vol. 6, no. 25, Aug. 2015, doi: 10.18632/oncotarget.4187.
[49] Y. Xia, Y. Liu, M. Deng, and R. Xi, "Detecting virus integration sites based on multiple related sequencing data by VirTect," *BMC Med. Genomics*, vol. 12, no. S1, p. 19, Jan. 2019, doi: 10.1186/s12920-018-0461-8.
[50] Q. Wang, P. Jia, and Z. Zhao, "VERSE: a novel approach to detect virus integration in host genomes through reference genome customization," *Genome Med.*, vol. 7, no. 1, p. 2, 2015, doi: 10.1186/s13073-015-0126-6.
[51] M. Forster *et al.*, "Vy-PER: eliminating false positive detection of virus integration events in next generation sequencing data," *Sci. Rep.*, vol. 5, no. 1, p. 11534, 2015, doi: 10.1038/srep11534.
[52] S. Baheti *et al.*, "HGT-ID: an efficient and sensitive workflow to detect human-viral insertion sites using next-generation sequencing data," *BMC Bioinformatics*, vol. 19, no. 1, p. 271, Dec. 2018, doi: 10.1186/s12859-018-2260-9.

[53] J.-W. Li, R. Wan, C.-S. Yu, N. N. Co, N. Wong, and T.-F. Chan, "ViralFusionSeq: accurately discover viral integration events and reconstruct fusion transcripts at single-base resolution," *Bioinformatics*, vol. 29, no. 5, pp. 649–651, Mar. 2013, doi: 10.1093/bioinformatics/btt011.

[54] J. P. Katz and J. M. Pipas, "SummonChimera infers integrated viral genomes with nucleotide precision from NGS data.," *BMC Bioinformatics*, vol. 15, no. 1, p. 348, 2014, doi: 10.1186/s12859-014-0348-4.

[55] C. Tennakoon and W. K. Sung, "BATVI: Fast, sensitive and accurate detection of virus integrations," *BMC Bioinformatics*, vol. 18, no. S3, p. 71, Mar. 2017, doi: 10.1186/s12859-017-1470-x.

[56] N. D. Nguyen, V. Deshpande, J. Luebeck, P. S. Mischel, and V. Bafna, "ViFi: accurate detection of viral integration and mRNA fusion reveals indiscriminate and unregulated transcription in proximal genomic regions in cervical cancer," *Nucleic Acids Res.*, vol. 46, no. 7, pp. 3309–3325, Apr. 2018, doi: 10.1093/nar/gky180.

[57] Q. Wang, P. Jia, and Z. Zhao, "VirusFinder: Software for Efficient and Accurate Detection of Viruses and Their Integration Sites in Host Genomes through Next Generation Sequencing Data," *PLoS One*, vol. 8, no. 5, p. e64465, May 2013, doi: 10.1371/journal.pone.0064465.

[58] Y. Chen, H. Yao, E. J. Thompson, N. M. Tannir, J. N. Weinstein, and X. Su, "VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue," *Bioinformatics*, vol. 29, no. 2, pp. 266–267, Jan. 2013, doi: 10.1093/bioinformatics/bts665.

[59] W. Li *et al.*, "HIVID: An efficient method to detect HBV integration using low coverage sequencing," *Genomics*, vol. 102, no. 4, pp. 338–344, Oct. 2013, doi: 10.1016/j.ygeno.2013.07.002.

[60] S. Crochu *et al.*, "Sequences of flavivirus-related RNA viruses persist in DNA form integrated in the genome of *Aedes spp.* mosquitoes," *J. Gen. Virol.*, vol. 85, no. 7, pp. 1971–1980, 2004, doi: 10.1099/vir.0.79850-0.

[61] C. D. Blair, K. E. Olson, and M. Bonizzoni, "The widespread occurrence and potentil biological roles of endognenous viral elements in insects genomes," in *Insect Molecular Virology*, June 2019., Bryony C. Bonning, Ed. Norfolk, UK: Caister Academic Press, 2019, pp. 13–30.

[62] B. Goic *et al.*, "Virus-derived DNA drives mosquito vector tolerance to arboviral infection," *Nat. Commun.*, vol. 7, pp. 1–10, 2016, doi: 10.1038/ncomms12410.

[63] E. Z. Poirier *et al.*, "Dicer-2-Dependent Generation of Viral DNA from Defective Genomes of RNA Viruses Modulates Antiviral Immunity in Insects," *Cell Host Microbe*, vol. 23, no. 3, pp. 353-365.e8, Mar. 2018, doi: 10.1016/j.chom.2018.02.001.

[64] D. K. Nag, M. Brecher, and L. D. Kramer, "DNA forms of arboviral RNA genomes are generated following infection in mosquito cell cultures," *Virology*, vol. 498, pp. 164–171, 2016, doi: 10.1016/j.virol.2016.08.022.

[65] D. K. Nag and L. D. Kramer, “Patchy DNA forms of the Zika virus RNA genome are generated following infection in mosquito cell cultures and in mosquitoes,” *J. Gen. Virol.*, vol. 98, no. 11, pp. 2731–2737, 2017, doi: 10.1099/jgv.0.000945.

[66] J. Brennecke *et al.*, “Discrete small RNA-Generating loci as master regulators of transposon activity in *Drosophila*,” *Cell*, vol. 128, no. 6, pp. 1089–1103, 2007, doi: 10.1016/j.cell.2007.01.043.

[67] H. L. Levin and J. V. Moran, “Dynamic interactions between transposable elements and their hosts,” *Nat. Rev. Genet.*, vol. 12, no. 9, pp. 615–627, 2011, doi: 10.1038/nrg3030.

[68] Y. W. Iwasaki, M. C. Siomi, and H. Siomi, “PIWI-Interacting RNA: Its Biogenesis and Functions,” *Annu. Rev. Biochem.*, vol. 84, pp. 405–433, 2015, doi: 10.1146/annurev-biochem-060614-034258.

[69] V. Zanni *et al.*, “Distribution, evolution, and diversity of retrotransposons at the flamenco locus reflect the regulatory properties of piRNA clusters,” *Proc. Natl. Acad. Sci.*, vol. 110, no. 49, pp. 19842–19847, Nov. 2013, doi: 10.1073/pnas.1313677110.

[70] P. Miesen, J. Joosten, and R. P. van Rij, “PIWIs Go Viral: Arbovirus-derived piRNAs in vector mosquitoes,” *PLoS Pathog.*, vol. 12, no. 12, pp. 1–17, 2016, doi: 10.1371/journal.ppat.1006017.

[71] M. Marconcini *et al.*, “Polymorphism analyses and protein modelling inform on functional specialization of Piwi clade genes in the arboviral vector *Aedes albopictus*,” *PLoS Negl. Trop. Dis.*, vol. 13, no. 12, p. e0007919, Dec. 2019, doi: 10.1371/journal.pntd.0007919.

[72] N. Tromas, M. P. Zwart, J. Forment, and S. F. Elena, “Shrinkage of Genome Size in a Plant RNA Virus upon Transfer of an Essential Viral Gene into the Host Genome,” *Genome Biol. Evol.*, vol. 6, no. 3, pp. 538–550, Mar. 2014, doi: 10.1093/gbe/evu036.

[73] E. Maori, E. Tanne, and I. Sela, “Reciprocal sequence exchange between non-retro viruses and hosts leading to the appearance of new host phenotypes,” *Virology*, vol. 362, no. 2, pp. 342–349, 2007, doi: 10.1016/j.virol.2006.11.038.

[74] K. Fujino, M. Horie, T. Honda, D. K. Merriman, and K. Tomonaga, “Inhibition of Borna disease virus replication by an endogenous bornavirus-like element in the ground squirrel genome,” *Proc. Natl. Acad. Sci.*, vol. 111, no. 36, pp. 13175–13180, 2014, doi: 10.1073/pnas.1407046111.

[75] Y. Suzuki *et al.*, “Non-retroviral Endogenous Viral Element Limits Cognate Virus Replication in Aedes aegypti Ovaries,” *Curr. Biol.*, vol. 30, no. 18, pp. 3495-3506.e6, Sep. 2020, doi: 10.1016/j.cub.2020.06.057.

[76] C. Chen, H. Huang, and C. H. Wu, “Protein Bioinformatics Databases and Resources,” 2017, pp. 3–39.

[77] N. H. N. D. de Silva, “Relational Databases and Biomedical Big Data,” 2017, pp. 69–81.

[78] P. Jezek and R. Moucek, “Semantic framework for mapping object-oriented model to semantic web languages,” *Front. Neuroinform.*, vol. 9, Feb. 2015, doi: 10.3389/fninf.2015.00003.

[79] The Uniprot Consortium, “UniProt: a worldwide hub of protein

knowledge," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D506–D515, Jan. 2019, doi: 10.1093/nar/gky1049.
[80] A. D. Yates *et al.*, "Ensembl 2020," *Nucleic Acids Res.*, Nov. 2019, doi: 10.1093/nar/gkz966.
[81] R. Hoehndorf, P. N. Schofield, and G. V Gkoutos, "The role of ontologies in biological and biomedical research: a functional perspective.," *Brief. Bioinform.*, vol. 16, no. 6, pp. 1069–80, Nov. 2015, doi: 10.1093/bib/bbv011.
[82] M. J. García Godoy, E. López-Camacho, I. Navas-Delgado, and J. F. Aldana-Montes, "Sharing and executing linked data queries in a collaborative environment," *Bioinformatics*, vol. 29, no. 13, pp. 1663–1670, Jul. 2013, doi: 10.1093/bioinformatics/btt192.
[83] "Linked Open Data Cloud." https://lod-cloud.net.
[84] *Open innovation, Open Science, open to the world. A vision for Europe*. Brussels: European Commission, Directorate-General for Research and Innovation., 2016.
[85] G. I. Giraldo-Calderón *et al.*, "VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases.," *Nucleic Acids Res.*, vol. 43, no. D1, pp. D707-13, Jan. 2015, doi: 10.1093/nar/gku1117.
[86] U. Palatini *et al.*, "Improved reference genome of the arboviral vector *Aedes albopictus*," *Genome Biol.*, vol. 21, no. 1, p. 215, Dec. 2020, doi: 10.1186/s13059-020-02141-w.
[87] J. R. Brister, D. Ako-adjei, Y. Bao, and O. Blinkova, "NCBI Viral Genomes Resource," *Nucleic Acids Res.*, vol. 43, no. D1, pp. D571–D577, Jan. 2015, doi: 10.1093/nar/gku1207.
[88] A. R. Quinlan and I. M. Hall, "BEDTools: A flexible suite of utilities for comparing genomic features," *Bioinformatics*, vol. 26, no. 6, pp. 841–842, 2010, doi: 10.1093/bioinformatics/btq033.
[89] E. Kitson and C. A. Suttle, "VHost-Classifier: virus-host classification using natural language processing," *Bioinformatics*, vol. 35, no. 19, pp. 3867–3869, Oct. 2019, doi: 10.1093/bioinformatics/btz151.
[90] W. Shen and J. Xiong, "TaxonKit: a cross-platform and efficient NCBI taxonomy toolkit," *bioRXiv*, 2019, [Online]. Available: https://www.biorxiv.org/content/10.1101/513523v1.
[91] C. Crava *et al.*, "Immunity to infections in arboviral vectors by integrated viral sequences: an evolutionary perspective," *bioRxiv*, p. 2020.04.02.022509, Jan. 2020, doi: 10.1101/2020.04.02.022509.
[92] L. I. Gilbert, *Insect Molecular Biology and Biochemistry*. Elsevier Science, 2012.
[93] W. Makałowski, V. Gotea, A. Pande, and I. Makałowska, "Transposable Elements: Classification, Identification, and Their Use As a Tool For Comparative Genomics," 2019, pp. 177–207.
[94] Z. Tu, "Molecular and Evolutionary Analysis of Two Divergent Subfamilies of a Novel Miniature Inverted Repeat Transposable Element in the Yellow Fever Mosquito, *Aedes aegypti*," *Mol. Biol. Evol.*, vol. 17, no. 9, pp. 1313–1325, Sep. 2000, doi: 10.1093/oxfordjournals.molbev.a026415.
[95] P. Arensburger *et al.*, "Phylogenetic and functional characterization of

the hAT transposon superfamily.," *Genetics*, vol. 188, no. 1, pp. 45–57, May 2011, doi: 10.1534/genetics.111.126813.
[96] H.-H. Zhang *et al.*, "Unexpected invasion of miniature inverted-repeat transposable elements in viral genomes," *Mob. DNA*, vol. 9, no. 1, p. 19, Dec. 2018, doi: 10.1186/s13100-018-0125-4.
[97] R. Hubley *et al.*, "The Dfam database of repetitive DNA families," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D81–D89, Jan. 2016, doi: 10.1093/nar/gkv1272.
[98] D. L. Wheeler, "Database resources of the National Center for Biotechnology," *Nucleic Acids Res.*, vol. 31, no. 1, pp. 28–33, Jan. 2003, doi: 10.1093/nar/gkg033.
[99] M. Falda *et al.*, "Argot2: a large scale function prediction tool relying on semantic similarity of weighted Gene Ontology terms," *BMC Bioinformatics*, vol. 13, no. S4, p. S14, Dec. 2012, doi: 10.1186/1471-2105-13-S4-S14.
[100] E. Pischedda *et al.*, "Insights into an unexplored component of the mosquito repeatome: Distribution and variability of viral sequences integrated into the genome of the arboviral vector *Aedes albopictus*," *Front. Genet.*, vol. 10, no. FEB, pp. 1–15, 2019, doi: 10.3389/fgene.2019.00093.
[101] A. Mckenna *et al.*, "The Genome Analysis Toolkit : A MapReduce framework for analyzing next-generation DNA sequencing data The Genome Analysis Toolkit : A MapReduce framework for analyzing next-generation DNA sequencing data," *Genome Res.*, vol. 20, no. 9, pp. 1297–1303, 2010, doi: 10.1101/gr.107524.110.programming.
[102] E. Garrison and G. Marth, "Haplotype-based variant detection from short-read sequencing," *arXiv*, p. 9, 2012, doi: arXiv:1207.3907 [q-bio.GN].
[103] A. Rimmer, H. Phan, I. Mathieson, Z. Iqbal, and S. R. F. Twigg, "Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications," *Nat. Genet.*, vol. 46, no. 8, pp. 912–918, 2014, doi: 10.1038/ng.3036.Integrating.
[104] Z. Lai *et al.*, "VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research," *Nucleic Acids Res.*, vol. 44, no. 11, pp. 1–11, 2016, doi: 10.1093/nar/gkw227.
[105] H. Li, "A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data," *Bioinformatics*, vol. 27, no. 21, pp. 2987–2993, Sep. 2011, doi: 10.1093/bioinformatics/btr509.
[106] H. Li, "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM," vol. 00, no. 00, pp. 1–3, 2013.
[107] "Picard." https://broadinstitute.github.io/picard/.
[108] P. Liu, Y. Dong, J. Gu, S. Puthiyakunnon, Y. Wu, and X. G. Chen, "Developmental piRNA profiles of the invasive vector mosquito *Aedes albopictus*," *Parasites and Vectors*, vol. 9, no. 1, pp. 1–15, 2016, doi: 10.1186/s13071-016-1815-8.
[109] Y. Wang, B. Jin, P. Liu, J. Li, X. Chen, and J. Gu, "PiRNA profiling of dengue virus type 2-infected Asian tiger mosquito and midgut tissues," *Viruses*, vol. 10, no. 4, pp. 1–20, 2018, doi: 10.3390/v10040213.

[110] S. A. Bernhardt, M. P. Simmons, K. E. Olson, B. J. Beaty, C. D. Blair, and W. C. Black, "Rapid intraspecific evolution of miRNA and siRNA Genes in the mosquito *Aedes aegypti*," *PLoS One*, vol. 7, no. 9, p. e44198, 2012, doi: 10.1371/journal.pone.0044198.
[111] E. M. Zdobnov *et al.*, "OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D744–D749, 2016, doi: 10.1093/nar/gkw1119.
[112] M. RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, "R Studio." 2015, [Online]. Available: https://www.rstudio.com.
[113] Z. Joly-Lopez and T. E. Bureau, "Exaptation of transposable element coding sequences," *Curr. Opin. Genet. Dev.*, vol. 49, pp. 34–42, 2018, doi: 10.1016/j.gde.2018.02.011.
[114] M. Manni *et al.*, "Genetic evidence for a worldwide chaotic dispersion pattern of the arbovirus vector, *Aedes albopictus*," *PLoS Negl. Trop. Dis.*, vol. 11, no. 1, p. e0005332, 2017, doi: 10.1371/journal.pntd.0005332.
[115] K. Okonechnikov *et al.*, "Unipro UGENE: a unified bioinformatics toolkit," *Bioinformatics*, vol. 28, no. 8, pp. 1166–1167, 2012, doi: 10.1093/bioinformatics/bts091.
[116] K. D. Yamada, K. Tomii, and K. Katoh, "Application of the MAFFT sequence alignment program to large data - Reexamination of the usefulness of chained guide trees," *Bioinformatics*, vol. 32, no. 21, pp. 3246–3251, 2016, doi: 10.1093/bioinformatics/btw412.
[117] C. Haag-Liautard *et al.*, "Direct estimation of per nucleotide and genomic deleterious mutation rates in *Drosophila*," *Nature*, vol. 445, no. 7123, pp. 82–85, 2007, doi: 10.1038/nature05388.
[118] P. D. Keightley, U. Trivedi, M. Thomson, F. Oliver, S. Kumar, and M. L. Blaxter, "Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines," *Genome Res.*, vol. 19, pp. 1195–1201, 2009, doi: 10.1101/gr.091231.109.more.
[119] G. W. Chirn *et al.*, "Conserved piRNA expression from a distinct set of piRNA cluster loci in Eutherian mammals," *PLoS Genet.*, vol. 11, no. 11, pp. 1–26, 2015, doi: 10.1371/journal.pgen.1005652.
[120] P. Kotsakiozi *et al.*, "Population genomics of the Asian tiger mosquito, *Aedes albopictus*: insights into the recent worldwide invasion," *Ecol. Evol.*, vol. 7, no. 23, pp. 10143–10157, Dec. 2017, doi: 10.1002/ece3.3514.
[121] A. J. Maynard *et al.*, "Tiger on the prowl: Invasion history and spatio-temporal genetic structure of the Asian tiger mosquito *Aedes albopictus* (Skuse 1894) in the Indo-Pacific," *PLoS Negl. Trop. Dis.*, vol. 11, no. 4, pp. 1–27, 2017, doi: 10.1371/journal.pntd.0005546.
[122] E. V. Koonin, V. V. Dolja, and M. Krupovic, "Origins and evolution of viruses of eukaryotes: the ultimate modularity," *Virology*, vol. 479–480, pp. 2–25, 2015, doi: 10.1016/j.virol.2015.02.039.
[123] J. L. Geoghegan, S. Duchêne, and E. C. Holmes, "Comparative analysis estimates the relative frequencies of co-divergence and cross-species transmission within viral families," *PLOS Pathog.*, vol. 13, no. 2, p.

e1006215, Feb. 2017, doi: 10.1371/journal.ppat.1006215.
[124] P. J. Keeling and J. D. Palmer, “Horizontal gene transfer in eukaryotic evolution,” *Nat. Rev. Genet.*, vol. 9, no. 8, pp. 605–618, Aug. 2008, doi: 10.1038/nrg2386.
[125] V. A. Belyi, A. J. Levine, and A. M. Skalka, “Unexpected Inheritance: Multiple Integrations of Ancient Bornavirus and Ebolavirus/Marburgvirus Sequences in Vertebrate Genomes,” *PLoS Pathog.*, vol. 6, no. 7, p. e1001030, Jul. 2010, doi: 10.1371/journal.ppat.1001030.
[126] M. Horie *et al.*, “Endogenous non-retroviral RNA virus elements in mammalian genomes,” *Nature*, vol. 463, no. 7277, pp. 84–87, Jan. 2010, doi: 10.1038/nature08695.
[127] H. Li and R. Durbin, “Fast and accurate short read alignment with Burrows–Wheeler transform,” *Mass Genomics*, vol. 25, no. 14, pp. 1754–1760, 2009, doi: 10.1093/bioinformatics/btp324.
[128] H. Li *et al.*, “The Sequence Alignment/Map format and SAMtools,” *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009, doi: 10.1093/bioinformatics/btp352.
[129] C. Mayer, “Phobos 3.3.11,” 2010. http://www.rub.de/ecoevo/cm/cm_phobos.htm.
[130] J. K. Kent, “BLAT—The BLAST-Like Alignment Tool,” *Genome Res.*, vol. 12, no. 4, pp. 656–664, 2002, doi: 10.1101/gr.229202.
[131] E. Gafni *et al.*, “COSMOS: Python library for massively parallel workflows,” *Bioinformatics*, vol. 30, no. 20, pp. 2956–2958, Oct. 2014, doi: 10.1093/bioinformatics/btu385.
[132] Z. Zhang, S. Schwartz, L. Wagner, and W. Miller, “A Greedy Algorithm for Aligning DNA Sequences,” *J. Comput. Biol.*, vol. 7, no. 1–2, pp. 203–214, Feb. 2000, doi: 10.1089/10665270050081478.
[133] M. G. Grabherr *et al.*, “Full-length transcriptome assembly from RNA-Seq data without a reference genome,” *Nat. Biotechnol.*, vol. 29, no. 7, pp. 644–652, Jul. 2011, doi: 10.1038/nbt.1883.
[134] J. T. Robinson *et al.*, “Integrative genomics viewer,” *Nat. Biotechnol.*, vol. 29, no. 1, pp. 24–26, Jan. 2011, doi: 10.1038/nbt.1754.
[135] A. Tharwat, “Classification assessment methods,” *Appl. Comput. Informatics*, vol. ahead-of-p, no. ahead-of-print, Aug. 2020, doi: 10.1016/j.aci.2018.08.003.
[136] D. Chicco and G. Jurman, “The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation,” *BMC Genomics*, vol. 21, no. 1, p. 6, Dec. 2020, doi: 10.1186/s12864-019-6413-7.
[137] E. Pischedda, C. Crava, M. Carlassara, L. Gasmi, and M. Bonizzoni, “ViR: a tool to account for intrasample variability in the detection of viral integrations,” *bioRxiv*, 2020, [Online]. Available: https://www.biorxiv.org/content/10.1101/2020.06.16.155119v1.
[138] C. E. Shannon, “A Mathematical Theory of Communication,” *Bell Syst. Tech. J.*, 1948, doi: 10.1002/j.1538-7305.1948.tb01338.x.
[139] M. L. Neal *et al.*, “Harmonizing semantic annotations for computational models in biology,” *Brief. Bioinform.*, vol. 20, no. 2, pp. 540–550, Mar. 2019, doi: 10.1093/bib/bby087.

[140] W3School, “HTML Language,” 2020. https://www.w3schools.com/html/html_intro.asp.
[141] R. Team, “ReactJs.org,” 2020. https://reactjs.org/tutorial/tutorial.html.
[142] “React-bootstrap.” 2020, [Online]. Available: https://www.npmjs.com/package/react-bootstrap.
[143] “React-router-dom.” 2020, [Online]. Available: https://www.npmjs.com/package/react-router-dom.
[144] “mui-datatables.” 2020, [Online]. Available: https://github.com/gregnb/mui-datatables#licence.
[145] E. G. Caldarola and A. M. Rinaldi, “A Multi-strategy Approach for Ontology Reuse Through Matching and Integration Techniques,” 2018, pp. 63–90.
[146] T. Berners-Lee, J. Hendler, and O. Lassila, “The Semantic Web,” *Sci. Am.*, vol. 284, no. 5, pp. 34–43, Nov. 2001, doi: 10.1038/scientificamerican0501-34.
[147] J. C. J. J. van Dam, J. J. Koehorst, J. O. Vik, V. A. P. P. Martins dos Santos, P. J. Schaap, and M. Suarez-Diez, “The Empusa code generator and its application to GBOL, an extendable ontology for genome annotation,” *Sci. data*, vol. 6, no. 1, p. 254, Dec. 2019, doi: 10.1038/s41597-019-0263-7.
[148] N. F. Noy *et al.*, “BioPortal: ontologies and integrated data resources at the click of a mouse.,” *Nucleic Acids Res.*, vol. 37, no. Web Server issue, pp. W170-3, Jul. 2009, doi: 10.1093/nar/gkp440.
[149] P.-Y. Vandenbussche, G. A. Atemezing, M. Poveda-Villalón, and B. Vatant, “Linked Open Vocabularies (LOV): A gateway to reusable semantic vocabularies on the Web,” *Semant. Web*, vol. 8, no. 3, pp. 437–452, Dec. 2016, doi: 10.3233/SW-160213.
[150] J. T. Bolleman *et al.*, “FALDO: a semantic standard for describing the location of nucleotide and protein feature annotation,” *J. Biomed. Semantics*, vol. 7, no. 1, p. 39, Dec. 2016, doi: 10.1186/s13326-016-0067-z.
[151] Z. Tatum *et al.*, “Preserving sequence annotations across reference sequences,” *J. Biomed. Semantics*, vol. 5, no. Suppl 1, p. S6, 2014, doi: 10.1186/2041-1480-5-S1-S6.
[152] C. J. Mungall, “Relation Ontology Homepage.” https://github.com/oborel/obo-relations/.
[153] K. Eilbeck *et al.*, “The Sequence Ontology: a tool for the unification of genome annotations.,” *Genome Biol.*, vol. 6, no. 5, p. R44, 2005, doi: 10.1186/gb-2005-6-5-r44.
[154] M. Dumontier *et al.*, “The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery,” *J. Biomed. Semantics*, vol. 5, no. 1, p. 14, 2014, doi: 10.1186/2041-1480-5-14.
[155] M. A. Musen, “The protégé project: A Look Back and a Look Forward,” *AI Matters*, vol. 1, no. 4, pp. 4–12, Jun. 2015, doi: 10.1145/2757001.2757003.
[156] “Protegè.” https://protege.stanford.edu/.
[157] “Web Ontology Language.” https://www.w3.org/TR/owl-features/.
[158] B. C. Grau, I. Horrocks, B. Motik, B. Parsia, P. Patel-Schneider, and U. Sattler, “OWL 2: The next step for OWL,” *J. Web Semant.*, vol. 6, no.

4, pp. 309–322, Nov. 2008, doi: 10.1016/j.websem.2008.05.001.
[159] S. Lohmann, V. Link, E. Marbach, and S. Negru, “WebVOWL: Web-based Visualization of Ontologies,” in *Knowledge Engineering and Knowledge Management*, 2015, pp. 154–158.
[160] J. K. Taubenberger and D. M. Morens, “1918 Influenza: the Mother of All Pandemics,” *Emerg. Infect. Dis.*, vol. 12, no. 1, pp. 15–22, Jan. 2006, doi: 10.3201/eid1201.050979.
[161] M. A. Mogler and K. I. Kamrud, “RNA-based viral vectors,” *Expert Rev. Vaccines*, vol. 14, no. 2, pp. 283–312, Feb. 2015, doi: 10.1586/14760584.2015.979798.
[162] E. C. Holmes, *The Evolution and Emergence of RNA Viruses*, vol. 16, no. 5. New York, NY, USA, 2010.
[163] C. Goble, S. Bechhofer, and K. Wolstencroft, “Semantic Web, Interoperability,” in *Encyclopedia of Systems Biology*, New York, NY: Springer New York, 2013, pp. 1921–1925.
[164] M. Stano, G. Beke, and L. Klucar, “viruSITE—integrated database for viral genomics,” *Database*, vol. 2016, p. baw162, Dec. 2016, doi: 10.1093/database/baw162.
[165] Y. Y. Go, U. B. R. Balasuriya, and C.-K. Lee, “Zoonotic encephalitides caused by arboviruses: transmission and epidemiology of alphaviruses and flaviviruses.,” *Clin. Exp. Vaccine Res.*, vol. 3, no. 1, pp. 58–77, Jan. 2014, doi: 10.7774/cevr.2014.3.1.58.
[166] S. C. Weaver and W. K. Reisen, “Present and Future Arboriral Threaths,” *Antiviral Res.*, vol. 85, no. 2, pp. 328–345, 2010, doi: 10.1016/j.antiviral.2009.10.008.
[167] B. Blitvich and A. Firth, “Insect-Specific Flaviviruses: A Systematic Review of Their Discovery, Host Range, Mode of Transmission, Superinfection Exclusion Potential and Genomic Organization,” *Viruses*, vol. 7, no. 4, pp. 1927–1959, Apr. 2015, doi: 10.3390/v7041927.
[168] M. M. Akiner, B. Demirci, G. Babuadze, V. Robert, and F. Schaffner, “Spread of the Invasive Mosquitoes *Aedes aegypti* and *Aedes albopictus* in the Black Sea Region Increases Risk of Chikungunya, Dengue, and Zika Outbreaks in Europe,” *PLoS Negl. Trop. Dis.*, vol. 10, no. 4, p. e0004664, Apr. 2016, doi: 10.1371/journal.pntd.0004664.
[169] J. Hobson-Peters *et al.*, “Discovery and characterisation of a new insect-specific bunyavirus from Culex mosquitoes captured in northern Australia,” *Virology*, vol. 489, pp. 269–281, Feb. 2016, doi: 10.1016/j.virol.2015.11.003.
[170] M. Calzolari, L. Zé-Zé, A. Vázquez, M. P. Sánchez Seco, F. Amaro, and M. Dottori, “Insect-specific flaviviruses, a worldwide widespread group of viruses only detected in insects,” *Infect. Genet. Evol.*, vol. 40, pp. 381–388, Jun. 2016, doi: 10.1016/j.meegid.2015.07.032.
[171] J. R. Fauver *et al.*, “West African Anopheles gambiae mosquitoes harbor a taxonomically diverse virome including new insect-specific flaviviruses, mononegaviruses, and totiviruses,” *Virology*, vol. 498, pp. 288–299, Nov. 2016, doi: 10.1016/j.virol.2016.07.031.
[172] R. Halbach, S. Junglen, and R. P. van Rij, “Mosquito-specific and mosquito-borne viruses: evolution, infection, and host defense,” *Curr.*

*Opin. Insect Sci.*, vol. 22, pp. 16–27, Aug. 2017, doi: 10.1016/j.cois.2017.05.004.

[173] M. R. T. Nunes *et al.*, "Genetic characterization, molecular epidemiology, and phylogenetic relationships of insect-specific viruses in the taxon Negevirus," *Virology*, vol. 504, pp. 152–167, Apr. 2017, doi: 10.1016/j.virol.2017.01.022.

[174] H. Guzman *et al.*, "Characterization of Three New Insect-Specific Flaviviruses: Their Relationship to the Mosquito-Borne Flavivirus Pathogens," *Am. J. Trop. Med. Hyg.*, vol. 98, no. 2, pp. 410–419, Feb. 2018, doi: 10.4269/ajtmh.17-0350.

[175] Agboli, Leggewie, Altinli, and Schnettler, "Mosquito-Specific Viruses-Transmission and Interaction," *Viruses*, vol. 11, no. 9, p. 873, Sep. 2019, doi: 10.3390/v11090873.

[176] E. Atoni *et al.*, "Metagenomic Virome Analysis of *Culex* Mosquitoes from Kenya and China," *Viruses*, vol. 10, no. 1, p. 30, Jan. 2018, doi: 10.3390/v10010030.

[177] A. Baidaliuk *et al.*, "Cell-Fusing Agent Virus Reduces Arbovirus Dissemination in *Aedes aegypti* Mosquitoes In Vivo," *J. Virol.*, vol. 93, no. 18, Jun. 2019, doi: 10.1128/JVI.00705-19.

[178] J. Kens, "Entrez Direct: E-utilities on the UNIX Command Line." Bethesda (MD): National Center for Biotechnology Information (US), 2013, [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK179288/.