Comparative analysis of genome-encoded viral sequences reveals the evolutionary history of the *Flaviviridae*.

Connor G. G. Bamford[1]*, William Marciel de Souza[2]*, Rhys Parry[3] and Robert J. Gifford[4]

**Equal contributions*

[1] Wellcome-Wolfson Institute for Experimental Medicine, School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast, Belfast, UK

[2] World Reference Center for Emerging Viruses and Arboviruses and Department of Microbiology and Immunology, University of Texas Medical Branch, Galveston, TX, USA

[3] School of Chemistry and Molecular Biosciences, University of Queensland, Brisbane, QLD 4072, Australia

[4] MRC-University of Glasgow Centre for Virus Research, Glasgow, UK

Corresponding author:
Robert J. Gifford,
MRC-University of Glasgow Centre for Virus Research,
464 Bearsden Rd, Bearsden, Glasgow G61 1QH
E-mail: robert.gifford@glasgow.ac.uk

Keywords: Evolution | Virus | Flavivirus | Paleovirology | Genomics |

Abstract

The flaviviruses (family *Flaviviridae*) are a group of positive-strand RNA viruses, many of which pose serious risks to human health on a global scale. Here we use flavivirus-derived DNA sequences, identified in animal genomes, to reconstruct the long-term evolutionary history of the family *Flaviviridae*. We demonstrate that the family is at least 100 million years old and show that this timing can be integrated with dates inferred from co-phylogenetic analysis and paleontological records to produce a cohesive overview of flavivirus evolution in which the main subgroups originate early in animal evolution and broadly co-diverge with animal phyla. In addition, we show that the arthropod-borne 'classical' flaviviruses first evolved from tick-specific viruses, and only later adapted to become insect-borne. Our findings demonstrate that the biological properties of flaviviruses have been acquired over many millions of years, implying that broad-scale comparative analysis will reveal fundamental insights into flavivirus biology. Furthermore, we implement a novel approach to computational genomic studies of viruses that can support these efforts by enabling more efficient utilization of evolution-related domain knowledge in virus research.

Introduction

Flaviviruses are a major cause of disease in humans and domestic animals. Arthropod-borne flaviviruses such as Zika virus (ZIKV), dengue virus (DENV), and yellow fever virus (YFV) are the causative agents of large-scale outbreaks that result in millions of human infections every year, while the bloodborne hepatitis C virus (HCV) causes chronic liver disease throughout the world [1, 2]. Currently, four flavivirus genera are recognised: *Pegivirus, Pestivirus*, *Hepacivirus* and *Flavivirus* [3]. These genera contain viruses with monopartite genomes ~10 kilobases (Kb) in length and encoding one or more large polyproteins that are co- and post-translationally cleaved to generate mature virus proteins. The structural proteins of the virion - capsid (C), premembrane (prM) and envelope (E) - are encoded toward the 5' end of the genome, while genes encoding non-structural (NS) proteins are located further downstream [4].

A diverse variety of novel 'flavivirus-like' virus species have been described over recent years, reflecting progress in the use of metagenomic approaches to broadly survey the global virome [5-9]. Many of these newly identified viruses – particularly those identified in invertebrate species - have yet to be incorporated into official taxonomy. They exhibit a much greater range of variation in genome structure than is found among representatives of officially recognised flavivirus genera, with

genome sizes ranging up to 20Kb [10]. Furthermore, one novel group – the Jingmenviruses (JMVs) – comprises viruses with genomes that are segmented rather than monopartite [11].

To prevent the spread of pathogenic viruses, it is helpful to understand their evolutionary history in as much detail as possible because this can provide crucial insights into virus biology and host-virus relationships. The flaviviruses are a taxonomically diverse group that have been extensively examined using comparative approaches, revealing uniquely clear correlations between phylogenetic relationships and ecological characteristics [12-15]. The current, rapid accumulation of genome sequence data from novel flavivirus species presents unparalleled opportunities to build on these comparative studies and derive further insights into flavivirus biology.

Presently, however, there are two major obstacles to efficient use of flavivirus genome data. Firstly, comparative analysis of virus genomes typically entails the construction of complex data sets comprised of sequences, multiple sequence alignments (MSAs) and phylogenies linked to other diverse kinds of data. In theory, these data sets can be re-used across a wide range of analysis contents while also being collaboratively developed and refined by multiple contributors. This kind of collaborative approach would likely accelerate progress in understanding viruses and expedite the development of systems that utilise virus genome data to monitor and control virus spread. Unfortunately, however, it remains challenging to implement in practice [16].

Secondly – knowledge of the appropriate evolutionary timescale is critically lacking [17, 18]. So far, most studies of flavivirus evolution have proposed relatively short timelines in which individual flavivirus genera and sub-groups emerge within the last 10-100 thousand years [14]. However, most of these studies were based on viruses sampled from a relatively restricted range of animal hosts. By contrast, more recent studies utilising metagenomic techniques to sample flavivirus diversity across a broad range of animal species have prompted suggestions of a much longer timeline extending over hundreds of millions of years [6]. Evidence for a longer timeline has also come in the form of *endogenous viral elements* (EVEs) - virus-derived sequences found within metazoan genomes - which often have demonstrably ancient origins [19]. EVEs are thought to originate when infection of germline cells leads to virus DNA becoming incorporated into chromosomal DNA so that integrated viral genes were not only inherited as host alleles, but also persisted in the gene pool over many generations until they were genetically fixed (i.e., reached a frequency of 100% in the species gene pool). Genome comparisons – combined with species divergence dates obtained from paleontological data - show that many fixed EVE loci

originated many millions of years ago. EVEs derived from flaviviruses - referred to here as *endogenous flaviviral elements* (EFVs) - have been identified in some species, including insects [19-21] and crustaceans [7], but appear to be relatively uncommon compared to EVEs derived from other RNA virus families [22]. Partly reflecting this scarcity, robust calibrations of the flavivirus timeline are lacking.

In this study we used endogenous viral element (EVE) sequences to investigate the long-term evolutionary history of the *Flaviviridae* and calibrate their evolutionary timeline. We show that EFV sequences have only rarely been fixed in the animal germline, but those that have represent a broad diversity of flavivirus groups. Moreover, incorporating these sequences into a comprehensive phylogenetic analysis of the *Flaviviridae* enables calibration of the timeline of flavivirus evolution and revises current understanding of flavivirus evolution. To facilitate the downstream use of the evolution-related domain knowledge generated in our study, we published our data and analyses in the form of a cross-platform, interactive database.

RESULTS

Creation of open resources for reproducible genomic analysis of flaviviruses

To facilitate greater reproducibility and reusability in comparative genomic analyses we previously developed a software framework called GLUE ("Genes Linked by Underlying Evolution") that supports the development of re-usable sequence-based resources or 'projects' [23]. Here, we used the GLUE framework to create Flavivirus-GLUE [24], an openly accessible online resource for comparative analysis of flavivirus genomes (**Fig. S1a-b**). To support comparative genomic analyses of the family *Flaviviridae*, the Flavivirus-GLUE project includes: (i) a set of 237 reference genome sequences each representing a distinct flavivirus species and linked to isolate-associated data (**Table S1**); (ii) a standardized set of 81 flavivirus genome features (**Table S2**); (iii) genome annotations specifying the coordinates of genome features within selected 'master' reference genome sequences (**Table S3**); (iv) a set of MSAs constructed to represent distinct taxonomic levels within the family *Flaviviridae* – together these MSAs comprehensively represent inferred sequence homologies between reference genome sequences included in the project (**Table 1**).

Loading the Flavivirus-GLUE project into the GLUE 'engine' generates a relational database that not only contains the data items associated with our analysis, but also represents the semantic links between them (**Fig. S1a**). Standardised, reproducible comparative genomic analyses can be implemented by using GLUE's command layer to coordinate interactions between the database and bioinformatics software tools. To illustrate this, we implemented a simple script that uses

information collated in Flavivirus-GLUE to derive nucleotide, dinucleotide and amino acid composition metrics for all flavivirus species, broken down by genome feature (**Table S4**).

To fully investigate phylogenetic relationships using all available information it is often necessary to construct several separate MSAs each representing a distinct taxonomic grouping. The difficulties encountered in maintaining these MSAs in sync with one another while avoiding irreversible data loss are a key factor underlying the low levels of reproducibility and re-use in comparative genomic analyses [16], particularly those that examine more distant evolutionary relationships [25]. To address this, GLUE allows MSAs to be hierarchically linked via a 'constrained alignment tree' data structure [23]. We applied this approach in Flavivirus-GLUE - for each taxonomic level represented in the project, one reference genome was selected as the constraining 'master' reference that defines the genomic coordinate space (**Table 1**). Within this MSA hierarchy, each MSA is linked to its child and/or parent MSAs via our chosen set of references (**Fig. S1c**). This creates in effect a single, unified MSA that can be used to implement comparisons across a range of taxonomic levels while also making use of the maximum amount of information available at each level. This approach also simplifies the incorporation of novel taxa into the alignment tree, since new sequences only need be aligned to the most closely related reference genome to be aligned with all other flavivirus taxa included in the Flavivirus-GLUE project. MSAs representing internal nodes (see **Table 1**) contain only master reference sequences but can be recursively populated with all taxa contained in child alignments when exported from GLUE for analysis [23]. The use of constrained MSAs also has the effect of standardising the genomic coordinate space to the constraining reference sequence selected for each MSA, thus enabling standardised genome comparisons across the entire *Flaviviridae* family.

Flavivirus-GLUE can be installed on all commonly-used computing platforms, and is also fully containerised via Docker [26]. Hosting of the resource in an openly accessible online version control system (GitHub) provides a platform for managing its ongoing development, following practices established in the software industry (**Fig. S1c**).

Mapping the genomic fossil record of the *Flaviviridae*

To identify EFVs, we performed systematic *in silico* screening of whole genome sequence (WGS) data representing 1075 animal species and identified 374 EFV loci in 36 species (**Table 2**, **Table S5**). EFVs were identified in a diverse range of species, including fish, arthropods, jellyfish and mollusks. Overall, however,

screening revealed that most animal lineages have not incorporated flavivirus-derived DNA into their germline - of 1075 species screened in our analysis only 36 were found to harbour EFVs. We used phylogenetic approaches to examine the evolutionary relationships between EFVs and contemporary flaviviruses. Most EFVs could be confidently identified as members of recognised taxonomic groups within the family *Flaviviridae*, but some appear to represent divergent lineages (**Table 2**). We constructed MSAs of groups of EFV sequences putatively derived from a single germline incorporation event (i.e., orthologs and duplicated elements) and generated representative EFV consensus sequences from these (**Fig. 1**, **Fig. S2**). EFV consensus sequences, MSAs, and all EFV-associated metadata were incorporated into Flavivirus-GLUE. We extended the Flavivirus-GLUE database schema to capture information specific to EFV loci (e.g., species in which they occur, locations within contigs/chromosomes, and orientation) (**Fig. S1**).

All major flavivirus lineages are represented in the genomic 'fossil record'

We used flavivirus and EFV sequences to investigate the evolutionary relationships between EFVs and contemporary flaviviruses. Bootstrapped, maximum likelihood phylogenies were reconstructed from MSAs representing a range of taxonomic levels within the *Flaviviridae* and including both virus and EFV sequences [24]. In line with official taxonomy and previously published studies [3, 15] phylogenetic reconstructions split the *Flaviviridae* into two major lineages - "pegi-hepaci" and "flavi-pesti" – each of which contains several well-supported subgroups [9] (**Fig. 2a-e, Fig S3**). However, a divergent EFV recovered from the genome of a priapulid worm (*Priapulus caudatus*) may represent a third sub-lineage (**Fig. 2b**).

EFVs derived from a broad range of flavivirus lineages and subgroups are represented in the genomic 'fossil record'. Within the "pegi-hepaci" lineage the *Pegivirus* and *Hepacivirus* genera occupy a derived position [8] while an EFV identified in the genome of a marine mollusc - the Eastern emerald elysia (*Elysia chlorotica*) groups basally (**Fig. 2a**). The "flavi-pesti" lineage comprises robustly supported "pesti-like" and "flavi-like" clades, each of which has robustly supported internal structure (**Fig. 2b**). Our approach to phylogenetic reconstruction, which entails reconstructing separate phylogenies for distinct taxonomic levels (see **Table 1**), supports a clean division of "flavi-like" viruses into three monophyletic clades (**Fig. 2e**) corresponding to genus *Flavivirus*, the Jingmenviruses, and a clade of viruses related to 'Tamana bat virus' (TABV) which we refer to here as genus *Tamanavirus*, in line with a previous proposal [27].

TABV was isolated from insectivorous bats in 1973, and has no known vector species. It is clearly distinct from all known members of the genus *Flavivirus* not only in its phylogenetic placement, but also in other signature genomic traits such as nucleotide composition [27]. Despite this, it has often been seen as a 'primitive' member of the Flavivirus genus. However, we find that TABV is split from genus *Flavivirus* by an EFV in the genome of the freshwater jellyfish *Craspadecusta sowerbii*. This observation, combined with the extensive diversity and broad host range of TABV-like viruses [7, 9, 28, 29], suggests that these groups might have entirely distinct origins in the distant evolutionary past.

Among 'Jingmenvirus-like' EFVs, we identified EFVs that group robustly within the diversity of JMVs as well as a more distantly related, 'Jingmenvirus-like' lineage – here labelled X2 - that is comprised entirely of EFVs. Notably, X2-derived EFVs were identified in an arthropod species (the copepod *Eurytemora affinis*) as well as in multiple ray-finned fish (clade Actinopterygii) (**Table 1**). The copepod element exhibited duplications and rearrangements suggesting it might fall within a piRNA cluster.

The 'pesti-like' lineage contains three robustly supported subclades corresponding to the genus *Pestivirus* (extended to include basal, pestivirus-like viruses recently identified in lower vertebrates [6]) and a clade of invertebrate-associated viruses that have been collectively referred to as 'large genome flaviviruses' (LGFs) [9]. Among LGFs we observed a robustly supported division into two groups, here labelled 'PL1' and 'PL2', one of which (PL2) contains EFVs in addition to viruses (**Fig. 2c**).

Germline integration of flaviviruses is relatively rare

While EFVs were only identified in a small proportion of the animal species we screened, they occur at relatively high copy number in the germline of some insect groups, including genus *Aedes* (mosquitoes) and superfamily *Apoidea* (bees and sphecoid wasps). Interestingly, EFVs in *Aedes* mosquitoes - have been shown to produce PIWI-interacting RNAs (piRNAs) in the antisense orientation, indicating they represent a system of heritable antiviral immunity analogous to that of CRISPR-Cas [30]. But while it seems likely that high EFV copy number in *Aedes* species is related to a role in antiviral immunity [31], it is not clear that it this reflects the existence of a dynamic system of immunity, as has been suggested previously. A misleading impression may have been created by the presence of dengue-derived DNA sequences in genomes derived from *Aedes* cell lines – these sequences may be cell culture artefacts. Furthermore, the remaining EFV sequences in Aedes mosquitoes –

the presence of which has been largely confirmed by investigations of wild populations [32] - all derive from the classical insect specific flavivirus (CISF) group. These EFVs are often <1% divergent from one another in overlapping regions and are often arranged as tandemly repeating sections (**Table S5**). This suggests that longer cDNA sequences may been fixed in the germline initially before subsequently being mobilised (presumably by retroelement-mediated transposition events). Nonetheless, the presence of some divergent EFVs within both *Aedes* genomes suggests that multiple distinct genome invasions have probably occurred in these species.

We identified numerous EFVs derived from LGFs in the insect superfamily *Apoidea* (bees and sphecoid wasps). Copy number was dramatically inflated in certain species, such as the spurred ceratina (*Ceratina calcarata*) which contains at least 86 distinct EFV sequences in its published genome sequence (**Table 1**). It was not possible to determine the precise number of distinct insertions, since many of the matches identified in our screen occurred on relatively short contigs (**Table S5**). The high copy number of PL2-derived EFVs in superfamily *Apoidea* suggests that amplification of an EFV in association with retroelement activity has occurred, but as with the multicopy EFVs in *Aedes* genomes, multiple invasions involving related viruses cannot be conclusively ruled out.

Major flavivirus lineages originated in the distant evolutionary past

We used a variety of approaches to calibrate the evolutionary timeline of the *Flaviviridae*. Calibrations based on orthology were obtained for jingmenvirus-derived sequences found in *Chironomus* midges, X2-derived EVEs in fish, and a PL2-derived EFV in superfamily Apoidea (**Table 2**). Orthologous, PL2-derived EFVs were identified in two species – the spurred ceratina (*Ceratina calcarata*) and the green orchid bee (*Euglossa dilemma*) - that diverged >100 million years ago (Mya) [33] (**Fig. S4**). We also derived age estimates ranging between 3-62 Mya for pairs of putatively duplicated EVE sequences based on the assumption a neutral molecular clock (**Table 2**). These calibrations, combined with the identification of EFVs in basal animal lineages such as cnidarians and priapulids, suggested to us that the *Flaviviridae* might have truly primordial origins extending back to the origins of the metazoan ~800-500 Mya. Consistent with this, we identified several clades in which (i) the branching relationships broadly matched those of host phyla, and (ii) divergence times among animal lineages are significantly correlated with branch lengths (**Fig. S6**). These included (i) the divergence between invertebrate and vertebrate lineages in the Hepaci-Pegi lineage; (ii) shark/mammal and

invertebrate/vertebrate splits in the Pesti-like lineage and (iii) the cnidarian-arthropod, chelicerata-hexapoda, and crustacean-insect splits in the *Flavivirus* genus.

Arthropod-vectored flaviviruses emerged from an arachnid source

The long history of association between flaviviruses and their hosts implied by our investigation is consistent with a scenario in which the vector-borne classical flaviviruses (CFVs) emerge in association with the evolution of hematophagy (blood-feeding) in arthropods. The vertebrate blood vascular system evolved >400 Mya and was established in its basic form in endothermic vertebrates (birds and mammals) by ~200 Mya [34], and once a system for the transport of nutrients had evolved, parasitism of this system would have inevitably followed. Hematophagous behaviour is thought to have evolved independently in arthropods on >20 occasions [35], and no doubt provides an excellent opportunity for viruses to acquire new hosts and mechanisms of transmission. Although no EVEs derived from arthropod-borne CFVs were identified in our screen, our phylogenetic reconstructions show that the 'classical insect-specific flaviviruses' (cISFs) are robustly separated from the CFVs by viruses identified in crustaceans, sea spiders (Chelicerata), and ticks (Acari) (**Fig 3a**). In addition, trees rooted on TABV or the cnidarian EFVs show that the most basal CFV lineages are either tick-associated (Mpulungu and the tick-borne CFVs) or potentially tick-associated viruses with no known vector [27]. We identified several synapomorphies in the NS5 protein where an ancestral residue is conserved between ancestral and tick-borne flaviviruses but mutated in insect-vectored flaviviruses (**Fig 3a**, **Fig. S7**). A simple, parsimonious interpretation of these observations is that they reflect the emergence of tick-vectored viruses from tick-specific viruses, followed by the emergence of insect-vectored viruses from tick-vectored ancestors on at least one occasion (**Fig 3b**).

DISCUSSION

The flaviviruses are a diverse and highly studied group of viruses that include an unusually large number of taxonomically recognised species, many of which are associated with disease [15]. As a genetically and ecologically diverse group, they offer unique possibilities for researchers interested in applying comparative approaches to viruses. Because flaviviruses have occupied an important position within virus research for over a century - YFV being the first human virus identified - it is likely that their taxonomic richness has historically reflected sampling effort, at least to some extent, and is biased toward sampling of certain taxa (e.g., potential vector or reservoir species) [36]. However, with dramatic advances in DNA

sequencing technology it is now possible to investigate flavivirus distribution and diversity much more broadly, building on previous comparative investigations [12-15].

In this report we addressed two important challenges to more effective use of flavivirus sequence data in comparative studies. Firstly, we implemented our investigation within a computational framework that supports re-use of complex data sets and enables the reproduction of comparative genomic analyses. Data are not only openly available but are also stored in a cross-platform database that represents the semantic links between data items so that they are 'poised' for further use in computational analyses.

Secondly, we used comparative approaches to investigate the long-term evolutionary history of flaviviruses, making extensive use of the EVE 'fossil record' to calibrate their evolutionary timeline. We identified EFVs >100 My old, demonstrating that the evolution of the *Flaviviridae* spans geological eras. Furthermore, we show that calibrations obtained from EVEs can be integrated with dates based on co-phyletic analysis to produce a cohesive overview of flavivirus evolution in which the major flavivirus lineages originated early in animal evolution and broadly co-diverged with animal phyla. Integration of phylogenetic with the various date calibrations obtained via comparative genomic and co-phyletic analysis allows us to reconstruct a broad overview of *Flaviviridae* evolution (**Fig. 3c**).

The extended evolutionary timeline implied by our research raises interesting questions about ways in which animal evolution may have impacted flavivirus evolution, since it encompasses the development of entire organ systems (e.g., the liver and vascular system), and spans the evolution of fundamental changes in animal physiology, such as the emergence of endothermic vertebrates. The identification of flavivirus-derived EVEs in animals lacking a circulatory system (e.g. Cnidarians and Priapulids) is consistent with a primordial origin in animals, and suggests that virus transmission via exosomes - as has been reported for TBFVs [37] - might represent the most archaic form of cell-to-cell transmission for flaviviruses.

The phylogenetic relationships within genus *Flavivirus*, combined with the extended evolutionary timeline implied by our study, are consistent with a parsimonious evolutionary scenario in which vector borne CFVs originate in arachnids and later acquire the capacity to infect and be transmitted by insects (**Fig. 3a**). This model of CFV evolution has some appealing properties. Firstly, tick feeding entails a relatively long exposure and available evidence indicates that ticks transmit infection to one another via host blood when multiple individuals feed in proximity [38]. To the extent that persisting longer in vertebrate blood is advantageous for increasing transmission among tick populations, it could eventually select for viruses

capable of limited replication in vertebrate cells. These viruses could then ultimately evolve into viruses capable of causing sustained viremia, which would in turn create opportunities for biting insects such as mosquitoes to acquire a vector role as they become exposed to tick-vectored viruses via feeding on viremic vertebrate hosts (**Fig. 3c**). Interestingly, this 'ticks first' hypothesis is consistent with previous comparative genomics studies, which found evidence that the 3' UTR – which contains numerous RNA structural elements involved in mediating the viral life cycle [39] - evolved via duplications of a ~200 nucleotide region within the NS5 ORF [40]. These 'long repeated sequences' are best preserved in the TBFV lineage - with only remnants being present in the more derived CFV groups [41].

It should be acknowledged that it is still possible to conceive of ways in which certain flavivirus genera or sub-lineages might have originated within shorter time frames. While our analysis reveals clear evidence that the Flaviviridae emerged millions of years ago, our co-phyletic calibrations rely on relatively strong assumptions regarding codivergence the rooting of virus trees, and for now remain speculative. Overall, however, our findings are consistent with a wide range of other data indicating that features of virus genomes can remain conserved over immense expanses of time [42-47]. This suggests that comparative analyses of flavivirus genomes – combined with other forms of empirical investigation – will yield key insights into the structural and functional basis of flavivirus replication, and the biology underlying important adaptations such as the capacity to replicate in both arthropod and vertebrate hosts.

Our analysis shows that flavivirus-derived DNA has rarely been fixed in the animal germline - only 36 of 1075 species genomes screened in our analysis contained any EFVs. However, in species or species groups where they do occur, EFVs are frequently multicopy, and some insect groups exhibit relatively high EFV copy number (e.g., genus *Aedes*, superfamily *Apoidea*) – the high level of similarity between distinct EFV inserts in mosquito genomes is consistent with recent mobilization of EFV DNA. The LGF-derived EFVs in *Apoidea* show a much higher level of diversity, but these sequences appear to derive from an ancient germline incorporation event, thus diversity could potentially have arisen among retroposed copies within the germline. We have conservatively assumed that the presence of multiple EFV loci generally reflects intragenomic amplification rather than multiple independent germline incorporation events. However, because multiple germline incorporation events by closely related viruses cannot be ruled out, it is possible that the number is higher.

Resources that allow more effective usage of genomic data in efforts to control the spread of viral diseases are a public health priority. There are a wide variety of ways in which the data collated in Flavivirus-GLUE can facilitate efforts to monitor and control flavivirus-associated disease. On the host side, the mapping of EFVs and their alignment to virus genomes can support investigations of piRNA-mediated defences. On the virus side, Flavivirus-GLUE can support broad-based comparative analyses as well as providing a foundation for the development of sequence-based resources that are more specifically focussed on individual flavivirus species (e.g., see HCV-GLUE [23, 48]). The novel computational approaches described here are scalable (e.g., see CoV-GLUE [49]) and allow researchers working in distinct areas of flavivirus research - but utilizing related data and domain knowledge - to benefit from one another's work.

MATERIALS & METHODS

Genome screening in silico

Systematic *in silico* genome screening was performed using the database-integrated genome screening (DIGS) tool [50] – a PERL-based screening framework within which the basic local alignment search tool (BLAST) program suite [51] is used to perform similarity searches while the MySQL relational database management system (Community Server 8.0.26) is used to record their output. WGS data were obtained from the National Center for Biotechnology Information (NCBI) genome database [52] - we obtained all animal genomes available as of March 2020 (see [24]). Flavivirus reference genomes (**Table S1**) and coding feature annotations (**Table S2**) collated in Flavivirus-GLUE [24] were used to derive polypeptide probes for tBLASTn-based screening in the DIGS framework. For virus genomes lacking detailed annotations, we created polypeptide probes based on arbitrary fragments of the major polyprotein. Via screening of WGS assemblies using the DIGS tool we generated a non-redundant database of putative EFV loci (see [24]). We used DIGS to investigate these loci and categorise them into: (i) putatively novel EFV elements; (ii) orthologs of previously characterised EFVs (e.g., copies containing large indels); (iii) non-viral sequences that cross-matched to flavivirus probes (e.g., retrotransposons).

Construction of sequence-based resources for comparative genomic analysis

We used the GLUE software framework [23] to create Flavivirus-GLUE [24], an openly accessible online resource for comparative analysis of flavivirus genomes.

A library of flavivirus representative genome sequences was collated from GenBank via reference to the International Committee for Taxonomy of Viruses website (see [24]). GenBank sequence entries in XML format were imported into the Flavivirus-GLUE project using an appropriately configured version of GLUE's GenBank importer module. We extracted isolate-specific information (e.g., date and location of isolation, isolate name, host species) from XML files as well standard GenBank fields (e.g., submission date). Additional/missing data was loaded from tabular files using GLUE's TextFilePopulator module. Via reference to previous studies [5, 8, 9, 15, 29] we assigned all flavivirus sequences included in Flavivirus-GLUE to a taxonomic group and defined a standard set of genome features for flaviviruses. The coordinates of genome features (where known) within all master references were recorded within the Flavivirus-GLUE database. These reference sequences and annotations were used in combination with a codon-aware, BLAST-based sequence aligner implemented in GLUE [23, 51] to generate constrained MSAs (i.e., MSAs in which the coordinate space is constrained to a selected 'master' reference) for each taxonomic level within the flavivirus family (**Table 1**). To handle genome and gene-coverage-related issues we used GLUE to generated genome feature coverage data for member sequence and used this information to condition the way in which taxa were selected into MSA partitions [53].

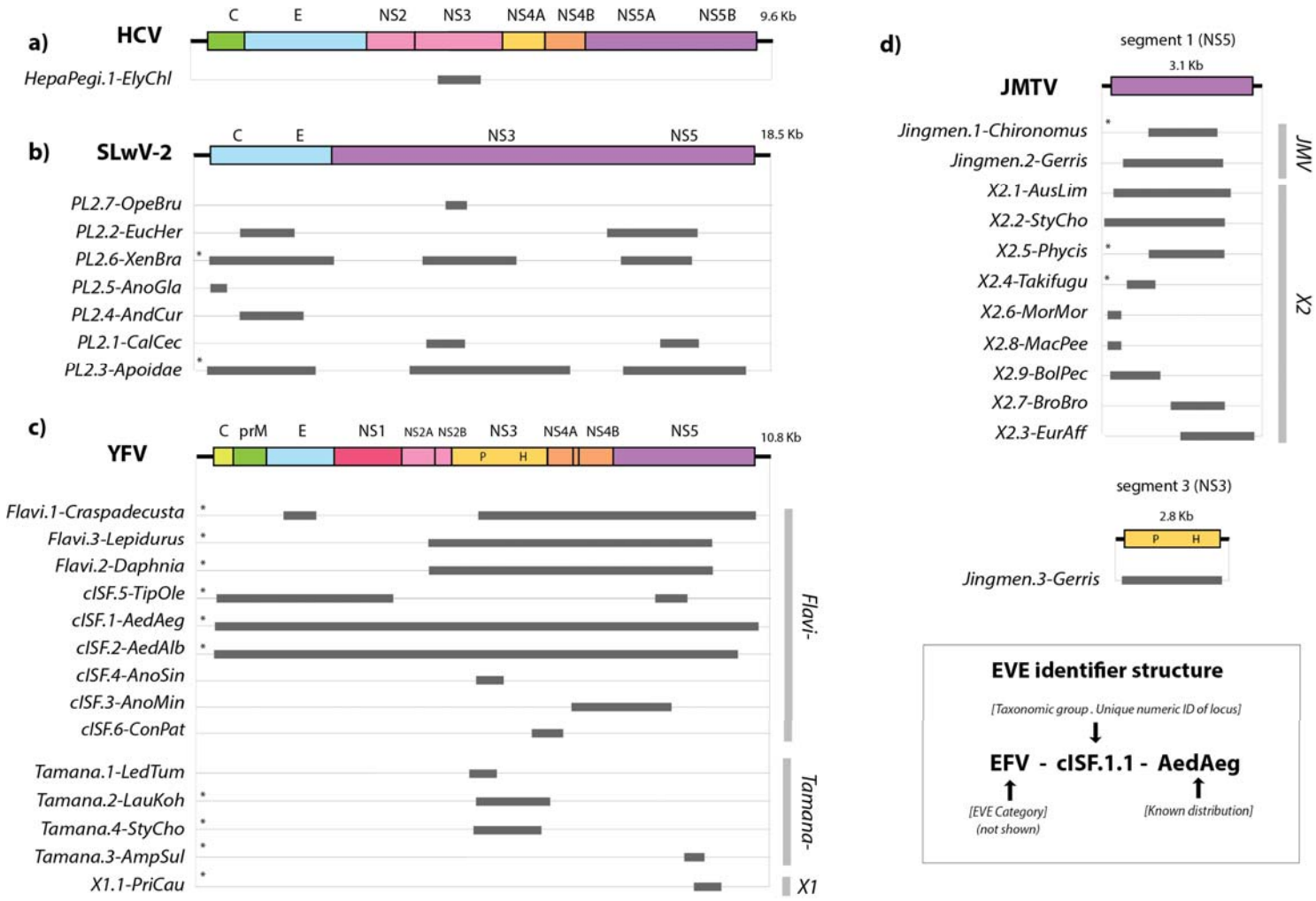
Phylogenetic and genomic analysis

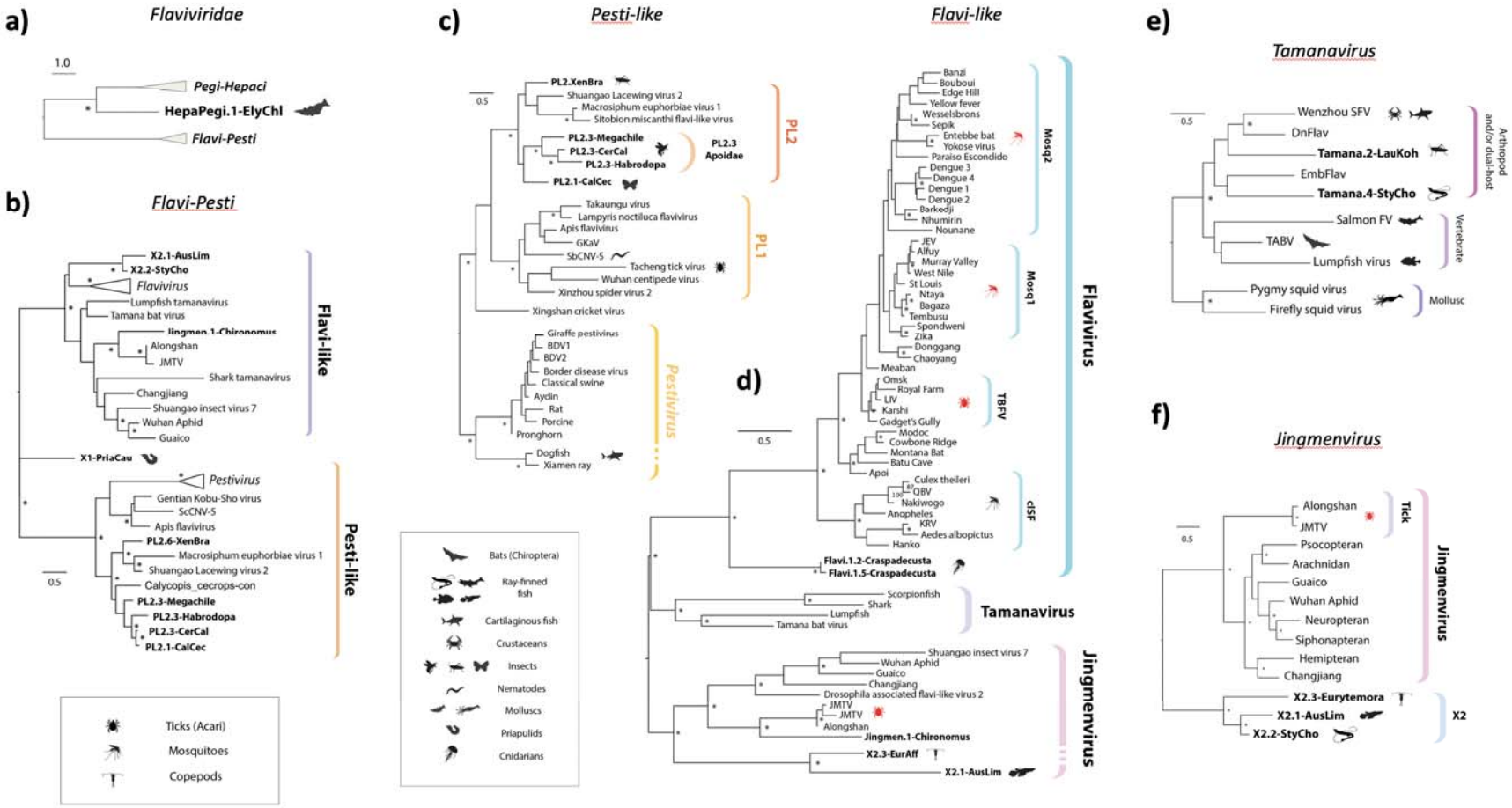
Flavivirus-GLUE was used to implement an automated process for reconstructing midpoint-rooted, bootstrapped phylogenies from MSA partitions representing each level within the constrained MSA tree. Phylogenies were reconstructed using the maximum likelihood approach implemented in RAxML (version 8.2.12) [53]. Protein substitution models were selected via hierarchical maximum likelihood ratio test using the PROTAUTOGAMMA option in RAxML. JalView [54] (version 2.11.1.4) and Se-Al (version 2.0a11) were used to inspect MSAs.

Composition analysis

To infer the location coordinates of genome features among all flavivirus species in Flavivirus-GLUE (i.e., not only sequences in which feature annotations had been explicitly defined - **Table S3**), constrained MSAs were used as input for GLUE's 'inherit features' command [23]. This function infers feature locations in all member sequences of an MSA based on their positions within the constraining reference genome. GLUE's native JavaScript-based scripting layer was used to

implement a process for calculating composition within each genome feature annotated within each reference sequence.


a)
HCV
C
E
NS2
NS3
NS4A
NS4B
NS5A
NS5B
9.6 Kb
HepaPegi.1-ElyChl
b)
SLwV-2
C
E
NS3
NS5
18.5 Kb
PL2.7-OpeBru
PL2.2-EucHer
PL2.6-XenBra
PL2.5-AnoGla
PL2.4-AndCur
PL2.1-CalCec
PL2.3-Apoidae
c)
YFV
C
prM
E
NS1
NS2A
NS2B
NS3
NS4A
NS4B
NS5
10.8 Kb
P
H
Flavi.1-Craspadecusta
Flavi.3-Lepidurus
Flavi.2-Daphnia
cISF.5-TipOle
cISF.1-AedAeg
cISF.2-AedAlb
cISF.4-AnoSin
cISF.3-AnoMin
cISF.6-ConPat
Flavi-
Tamana.1-LedTum
Tamana.2-LauKoh
Tamana.4-StyCho
Tamana.3-AmpSul
Tamana-
X1.1-PriCau
X1
d)
segment 1 (NS5)
3.1 Kb
JMTV
Jingmen.1-Chironomus
Jingmen.2-Gerris
JMV
X2.1-AusLim
X2.2-StyCho
X2.5-Phycis
X2.4-Takifugu
X2.6-MorMor
X2.8-MacPee
X2.9-BolPec
X2.7-BroBro
X2.3-EurAff
X2
segment 3 (NS3)
2.8 Kb
P
H
Jingmen.3-Gerris
EVE identifier structure
[Taxonomic group . Unique numeric ID of locus]
EFV - cISF.1.1 - AedAeg
[EVE Category]
(not shown)
[Known distribution]


a)
Flaviviridae
1.0
Pegi-Hepaci
HepaPegi.1-ElyChl
Flavi-Pesti
b)
Flavi-Pesti
X2.1-AusLim
X2.2-StyCho
Flavivirus
Lumpfish tamanavirus
Tamana bat virus
Jingmen.1-Chironomus
Alongshan
JMTV
Shark tamanavirus
Changjiang
Shuangao insect virus 7
Wuhan Aphid
Guaico
Flavi-like
X1-PriaCau
Pestivirus
Gentian Kobu-Sho virus
ScCNV-5
Apis flavivirus
PL2.6-XenBra
Macrosiphum euphorbiae virus 1
Shuangao Lacewing virus 2
Calycopis_cecrops-con
PL2.3-Megachile
PL2.3-Habrodopa
PL2.3-CerCal
PL2.1-CalCec
Pesti-like
0.5
Ticks (Acari)
Mosquitoes
Copepods
c)
Pesti-like
PL2.XenBra
Shuangao Lacewing virus 2
Macrosiphum euphorbiae virus 1
Sitobion miscanthi flavi-like virus
PL2.3-Megachile
PL2.3-CerCal
PL2.3-Habrodopa
PL2.3 Apoidae
PL2.1-CalCec
PL2
Takaungu virus
Lampyris noctiluca flavivirus
Apis flavivirus
GKaV
SbCNV-5
Tacheng tick virus
Wuhan centipede virus
Xinzhou spider virus 2
PL1
Xingshan cricket virus
Giraffe pestivirus
BDV1
BDV2
Border disease virus
Classical swine
Aydin
Rat
Porcine
Pronghorn
Dogfish
Xiamen ray
Pestivirus
0.5
Bats (Chiroptera)
Ray-finned fish
Cartilaginous fish
Crustaceans
Insects
Nematodes
Molluscs
Priapulids
Cnidarians
d)
Flavi-like
Banzi
Bouboui
Edge Hill
Yellow fever
Wesselsbrons
Sepik
Entebbe bat
Yokose virus
Paraiso Escondido
Dengue 3
Dengue 4
Dengue 1
Dengue 2
Barkedji
Nhumirin
Nounane
Mosq2
JEV
Alfuy
Murray Valley
West Nile
St Louis
Ntaya
Bagaza
Tembusu
Spondweni
Zika
Donggang
Chaoyang
Meaban
Mosq1
Omsk
Royal Farm
LIV
Karshi
Gadget's Gully
TBFV
Modoc
Cowbone Ridge
Montana Bat
Batu Cave
Apoi
Culex theileri
QBV
Nakiwogo
Anopheles
KRV
Aedes albopictus
Hanko
cISF
Flavivirus
Flavi.1.2-Craspadecusta
Flavi.1.5-Craspadecusta
Scorpionfish
Shark
Lumpfish
Tamana bat virus
Tamanavirus
Shuangao insect virus 7
Wuhan Aphid
Guaico
Changjiang
Drosophila associated flavi-like virus 2
JMTV
JMTV
Alongshan
Jingmen.1-Chironomus
X2.3-EurAff
X2.1-AusLim
Jingmenvirus
0.5
e)
Tamanavirus
Wenzhou SFV
DnFlav
Tamana.2-LauKoh
EmbFlav
Tamana.4-StyCho
Arthropod and/or dual-host
Salmon FV
TABV
Lumpfish virus
Vertebrate
Pygmy squid virus
Firefly squid virus
Mollusc
0.5
f)
Jingmenvirus
Alongshan
JMTV
Tick
Psocopteran
Arachnidan
Guaico
Wuhan Aphid
Neuropteran
Siphonapteran
Hemipteran
Changjiang
Jingmenvirus
X2.3-Eurytemora
X2.1-AusLim
X2.2-StyCho
X2
0.5

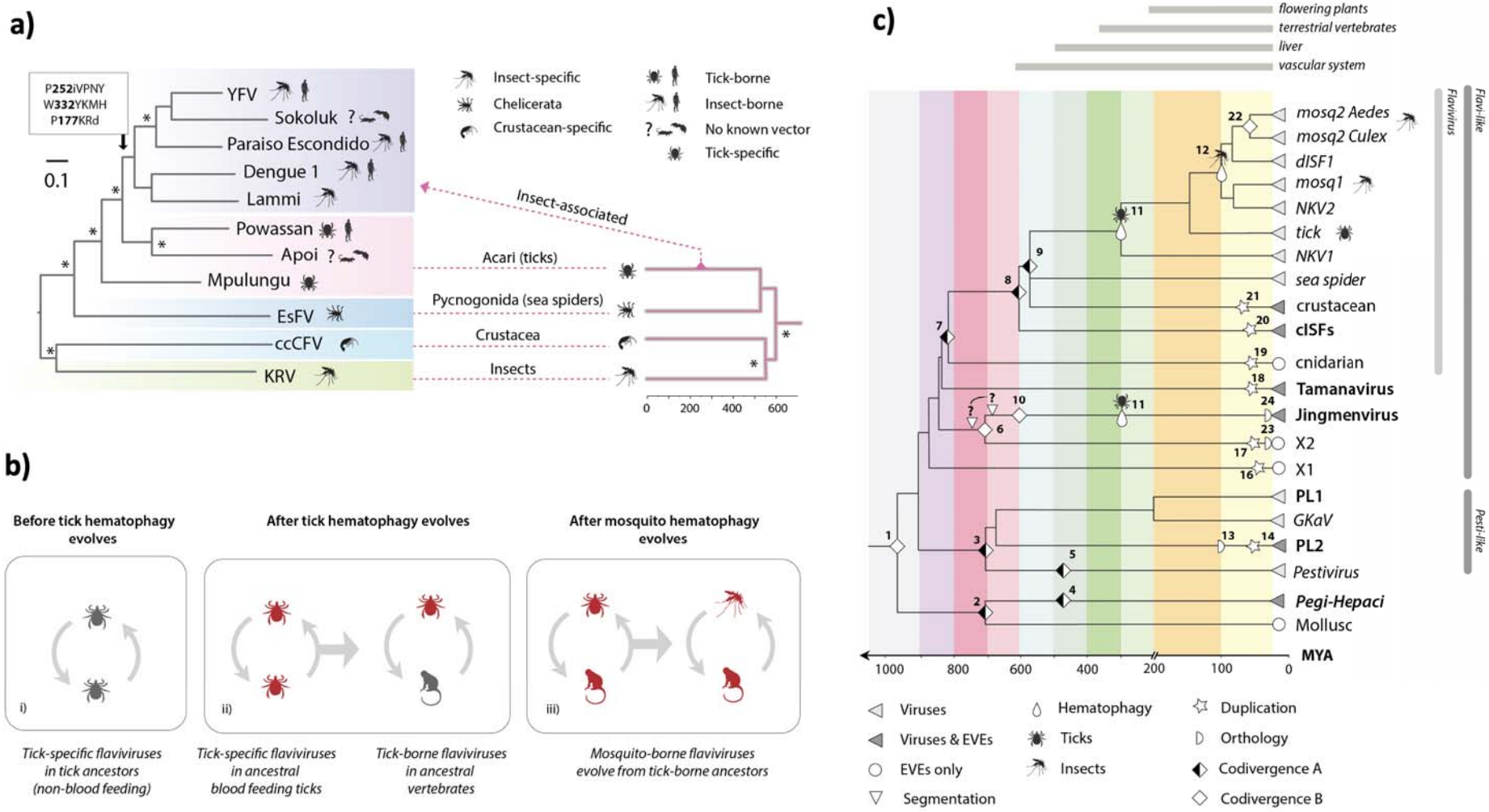

a)
P252iVPNY
W332YKMH
P177KRd
0.1
YFV
Sokoluk ?
Paraiso Escondido
Dengue 1
Lammi
Powassan
Apoi ?
Mpulungu
EsFV
ccCFV
KRV
Insect-specific
Chelicerata
Crustacean-specific
Tick-borne
Insect-borne
No known vector
Tick-specific
Insect-associated
Acari (ticks)
Pycnogonida (sea spiders)
Crustacea
Insects
0
200
400
600
b)
Before tick hematophagy evolves
After tick hematophagy evolves
After mosquito hematophagy evolves
i)
ii)
iii)
Tick-specific flaviviruses in tick ancestors (non-blood feeding)
Tick-specific flaviviruses in ancestral blood feeding ticks
Tick-borne flaviviruses in ancestral vertebrates
Mosquito-borne flaviviruses evolve from tick-borne ancestors
c)
mammals
flowering plants
terrestrial vertebrates
liver
vascular system
Flavi-like
Flavivirus
Pesti-like
mosq2 Aedes
mosq2 Culex
dISF1
mosq1
NKV2
tick
NKV1
sea spider
crustacean
cISFs
cnidarian
Tamanavirus
Jingmenvirus
X2
X1
PL1
GKaV
PL2
Pestivirus
Pegi-Hepaci
Mollusc
MYA
1000
800
600
400
200
100
0
Viruses
Viruses & EVEs
EVEs only
Segmentation
Hematophagy
Ticks
Insects
Duplication
Orthology
Codivergence A
Codivergence B

Figure 1. Genome structures of endogenous flaviviral elements.

Schematic diagrams showing the genomic regions represented by endogenous flaviviral element (EFV) sequences. **(a)** Hepaci-Pegivirus-like elements are shown relative to hepatitis C virus (HCV); **(b)** elements derived from the 'Pestivirus-like 2' (PL2) lineage of viruses shown relative to Shuangao lacewing virus 2 (SLwV-2); **(c)** Elements derived from the *Flavivirus* genus, the proposed '*Tamanavirus*' genus, and the X1 lineage shown relative to Dengue virus type 1 (DENV-1); **(d)** Jingmenvirus-derived EFVs shown relative to Jingmen tick virus (JMTV). Homologous regions represented by EFV sequences are shown as grey horizontal bars. Bars to the right show taxonomic groups. EFV identifiers (IDs) are shown to the left. EFV IDs were constructed as indicated in the key, following a convention established for endogenous retroviruses [55]. IDs have three components- the first is the classifier 'EFV' and can be dropped when it is implied by context. The second comprises two subcomponents separated by a period; (i) the name of the taxonomic group of viruses from which the EFV is thought to derive; (ii) a numeric ID that uniquely identifies the integration locus. The third component describes the known taxonomic distribution of orthologous copies of the element among host species. For the Jingmenviruses, which have segmented genomes, we used classifiers that specify the gene from which the EVE is derived, in line with conventions established for EVEs derived from segmented viruses [45]– endogenous jingmenvirus helicase (EJH) and endogenous jingmenvirus polymerase (EJP). **Abbreviations**: Kb = kilobase. X1= unclassified flavivirus-like lineage X1. X2 = unclassified flavivirus-like lineage X2. C = Capsid, prM = Pre-Membrane, E = Envelope, NS = Non-Structural protein, P = Protease, H = Helicase. HP = Hepacivirus/Pegivirus-like virus. ***** Indicates consensus sequence.

Figure 2. Phylogenetic relationships of Flaviviruses and endogenous flaviviral elements (EFVs). Bootstrapped maximum likelihood phylogenies (1000 replicates), reconstructed for viruses and EFVs across a range of taxonomic levels, as follows: **(a)** All *Flaviviridae*, showing placement of Hepacivirus-related EFV (122 amino acid (aa) residues in multiple sequence alignment (MSA) spanning conserved regions in NS3, substitution model=RTREV); **(b)** Flavi-Pesti lineage (MSA spanning 99 aa residues in NS5, substitution model= BLOSUM62); **(c)** Pesti-like lineage (104 aa residues in NS5, substitution model=RTREV); **(d)** Flavi-like lineage (MSA spanning 179 aa residues in NS5, substitution model=LG likelihood); **(e)** Tentative genus 'Tamanavirus' (MSA spanning 292 aa residues in NS5, substitution model=BLOSUM62); **(f)** Jingmenviruses and related lineages (MSA spanning 727 residues in NS5, substitution model=LG likelihood). EFV names are shown in bold. Only EFVs that had sufficient coverage (>50% of total MSA length) were included in the analysis. Triangular terminal branches indicate collapsed clades containing multiple taxa. Asterisks indicate bootstrap support ≥70% (1000 replicates). The scale bar indicates evolutionary distance in substitutions per site. Brackets to the right indicate genera and sub-lineages. All trees are midpoint rooted for display purposes. **Abbreviations**: cISF = classical insect-specific flaviviruses. NKV = no known vector. Jingmen = Jingmen tick virus =. PL = Pestivirus-like; MBFV2 = Mosquito borne.

Figure 3. Timeline of flavivirus evolution

(a) Evolution of arthropod-borne flaviviruses. The phylogeny shown on the left was constructed from an alignment spanning 804 residues of the precursor polyprotein (substitution model=LG likelihood). The phylogeny on the right is a time-calibrated phylogeny of arthropod hosts/vectors of classical flaviviruses, obtained via TimeTree [33]. The figure shows the hypothesised relationships between the host and virus trees, with virus phylogeny broadly following host phylogeny at higher taxonomic levels but diverging dramatically from this pattern when tick-borne viruses of vertebrates emerge.

(b) Proposed model for the origin and evolution of the vector-borne flaviviruses. Three stages are shown: (i) an ancestral group of non-vectored tick viruses is present; (ii) tick hematophagy provides prolonged exposure to vertebrate blood allowing tick-specific flaviviruses to eventually acquire the ability to replicate in vertebrate cells and become tick-borne flaviviruses (TBFV); (iii) the evolution of TBFV gives rise to the occurrence of viraemic vertebrate hosts, thereby exposing mosquitoes and other hematophagous insects such that they might acquire a role as vectors.

(c) An integrated overview of flavivirus evolution. A time-calibrated phylogeny, summarising the phylogenetic relationships recovered in our analysis is shown. The key indicates the different types of time calibrations incorporated into the figure**.** Numbers adjacent to key symbols correspond to numbered rows in **Table 3** ‘Codivergence type A’ = codivergence supported by co-phylogeny. Codivergence type B = potential codivergence-based calibrations without supporting evidence. Vertical lines to right of the tree show taxonomic groups.

Table 1. Endogenous flaviviral elements detected in this study

Sequence ID[a]	#Species[b]	#Sequences[c]	lORF[d]	Host species[e]	Common name
Flavivirus					
cISF.1-AedAeg*	1	92	564	*Aedes aegypti*	Yellow fever mosquito
cISF.2-AedAlb*	1	35	643	*Aedes albopictus*	Tiger mosquito
cISF.3-AnoMin*	1	1	97	*Anopheles minimus*	
cISF.4-AnoSin*	1	1	138	*Anopheles sinensis*	
cISF.5-TipOle	1	4	226	*Tipula oleracea*	Marsh crane fly
cISF.6-ConPat	1	1	38	*Condylostylus patibulatus*	Long-legged fly
Flavi.1-CraSow	1	13	1182	*Craspedacusta sowerbyi*	Freshwater jellyfish
Flavi.2-DapMag*	1	2	573	*Daphnia magna*	Water flea
Flavi.3-LepArc*	1	3	281	*Lepidurus arcticus*	Tadpole shrimp
Tamanavirus					
Tamana.1-LedTum	1	1	55	*Lednia tumana*	Meltwater stonefly
Tamana.2-LauKoh	1	1	56	*Laupala kohalensis*	Hawaiian cricket
Tamana.3-AmpSul	1	2	44	*Amphinemura sulcicollis*	Spring stonefly
Tamana.4-StyCho	1	1	873	*Stylephorus chordatus*	Tube-eye
Jingmenvirus					
EJP-Jingmen.1-Chironomus	2	2	248	*Chironomus sp.*	
EJP-Jingmen.2-*Gerris*	1	1	45	*Gerris buenoi*	Water strider
EJH-Jingmen.1-*Gerris*	1	1	51	"	"
X1					
X1.1-PriCau	1	2	347	*Priapulus caudatus*	Penis worm
X2					
X2.1-AusLim	1	2	214	*Austrofundulus limnaeus*	Mangrove killifish
X2.2-StyCho	1	3	873	*Stylephorus chordatus*	Tube-eye
X2.3-EurAff	1	2	190	*Eurytemora affinis*	Copepod
X2.4-Takifugu	3	5	103	*Takifugu*	Pufferfish
X2.5-Phycis	2	2	129	*Phycis*	Phycid hakes
X2.6-MorMor	1	1	36	*Mora moro*	Common mora
X2.7-BroBro	1	1	56	*Brosme brosme*	Cusk
X2.8-MacPee	1	1	51	*Maccullochella peelii*	Murray cod
X2.9-BolPec	1	1	155	*Boleophthalmus pectinirostris*	Blue spotted mudhopper
Pesti-like					
PL2.1-CalCec	1	5	78	*Calycopis cecrops*	Red-banded hairstreak
PL2.2-EucHer	1	2	187	*Eucheros histo*	Stink bug
PL2.3-Apoidae	11	239	2469	Family Apoidae	Sweat bees
PL2.4-AndCur	2	6	961	Andricus	Cynipid gall wasp
PL2.5-AnoGla	1	1	26	*Anoplophora glabripennis*	Long-horned beetle
PL2.6-XenBra	1	16	2232	*Xenocatantops brachycerus*	Grasshopper
PL2.7-OpeBru	1	2	38	*Operophtera brumata*	Winter moth
Pegi-Hepaci					
HepaPegi.1-ElyChl	1	1	60	*Elysia chlorotica*	Eastern emerald elysia

Footnote: [a] EFVs have been assigned standard IDs based on conventions established for endogenous retroviruses, wherein information about virus taxonomy and locus orthology are incorporated into the ID itself (see Methods). Asterisks indicate EFVs loci or lineages that have been reported previously. For jingmenviruses the classifier component of the ID specifies the gene it is derived from following conventions established for EFVs derived from mRNAs and/or segmented viruses (EJP=endogenous jingmenvirus polymerase; EJH=endogenous jingmenvirus helicase). [b] Number of species in which EVE locus was identified. [c] Number of sequences (i.e., distinct insertions) derived from this EVE that were identified via *in silico* screening. [e] Host species or species groups. Abbreviations: L-ORF=Longest open reading frame; PL2=Pesti-like 2; cISF=classical insect-specific flaviviruses.

Table 2. Comprehensive mapping of flavivirus homology via hierarchically linked multiple sequence alignments

# [a]	Taxonomic scope	Name	Parent [b]	Children [c]	Constraining reference [d]	Genome Coverage [e]	Virus count	EVE count
Family								
1	Flaviviridae-like	Flaviviridae	none	2	YFV	~6%	2	
Major lineage								
2	Hepaci/Pegi-like viruses	HepaPegi	Flaviviridae	2	BVDV1	~60%	2	1
3	Flavi/Pesti-like	FlaviPesti	Flaviviridae	2	YFV	~80%	2	1
4	Flavi-like	FlaviLike	FlaviPesti	2	YFV	~35%	2	
5	Pesti-like	PestiLike	FlaviPesti	3	BVDV1	~60%	3	
Minor lineage								
6	Flavi-Tamana-like	FlaviTamana	FlaviLike	2	YFV	~93%		
Genus								
7	Flavivirus genus	Flavivirus	FlaviTamana	11	YFV		8	3
8	Hepacivirus genus*	Hepacivirus	HepaPegi	0	HCV	~90%	14	
9	Pegivirus genus*	Pegivirus	HepaPegi	0	HPgV2	~60%	9	
10	Insect pesti-like*	PL2	PestiLike	0	SLV2	~73%	6	7
11	SCNV+arthropod*	PL1	PestiLike	0	SCNV5	~72%	8	
12	Pestivirus genus*	Pestivirus	PestiLike	0	BVDV1	~88%	11	
13	X2-derived EVEs	X2	FlaviLike	0		NS5		9
14	Jingmenvirus seg 1	Jingmen_1	FlaviLike	0	JMTV Seg1	Segment 1	7	2
14	Jingmenvirus seg 3	Jingmen_3	FlaviLike	0	JMTV Seg3	Segment 3	4	1
15	Tamanavirus genus*	Tamanavirus	FlaviTamana	0	TABV	100%	6	5
Subgenus								
16	Crustacean	crustacean	Flavivirus	0		100%Δ-genome	2	
17	cISF	cISF	Flavivirus	0	KRV	100%	14	6
18	NKV1	NKV2	Flavivirus	0	APOIV	100%	6	
19	Tick	tick	Flavivirus	0	POWV	~91%	15	
20	dISF	dISF	Flavivirus	0	LAMV	~95%	9	
21	Mosquito1	mosquito-1	Flavivirus	0	DEN1	~97%	35	
22	NKV2	NKV2	Flavivirus	0	SOKV	100%	3	
23	Mosquito2	mosquito-2	Flavivirus	0	YFV	~93%	14	
	Totals						**159**	**35**

Footnote: [a] Numbers correspond to labelled nodes in **Fig. S1d**. [b] The parent multiple sequence alignment (MSA) in the hierarchy. [c] Children of the MSA in the hierarchy. [d] Reference sequence that constrains the genomic co-ordinate coordinate space in the MSA. [e] Percentage of the constraining reference genome spanned by the MSA. Phylogenies constructed for each of these alignment partitions are available in Flavivirus-GLUE [24].

Table 3. Dates and age estimates used to calibrate flavivirus evolution.

#	Virus lineage	Host lineage(s)	MYA	Low	High
Codivergence A (minimum age)					
1	*Flaviviridae*	Animalia	952	757	1147
6	Jingmen-X2	Insecta-Crustacea	794	678	916
22	Mosquito 2 subgenus	Culex-Aedes	40	22	52
10	Jingmenvirus	Arachnida-Insecta	601	568	642
Codivergence B (minimum age)					
2	Hepaci-Pegi lineage	Mollusc-vertebrate	794	678	916
3	Pesti-like viruses	Invertebrates-vertebrates	794	678	916
4	Pegi-Hepaci	Chondrichthyes-Mammalia	465	450	497
5	Pestiviruses	Chondrichthyes-Mammalia	465	450	497
7	Flavivirus	Cnidaria-Arthropoda	824	652	973
8	Arthropod flaviviruses	Crustacea-Arachnida	601	568	642
9	Chelicerata flaviviruses	Pycnogonida-Arachnida	553	476	653
Orthology (minimum age)					
13	PL2	Family Apoidea	102	71	148
24	Jingmenvirus	Genus *Chironomus*	*n/k*	*n/k*	*n/k*
23	X2	Genus *Phycis*	16.9	12.1	37.1
Duplicates /Molecular clock (minimum age)					
20	cISF	Tipula		16	40
21	Crustacean FV	Daphnia 2		26	62
19	Cnidarian FV	Craspedacusta 1		17	40
16	X1	Priapulus (NS5)		12	29
17	X2	Eurytemora (NS5)		20	47
18	Tamanavirus	Austrofundulus 1 (NS5)		19	43
14	PL2	Operophtera		17	40
Origin hypothesis (maximum age)					
11	All vectored flaviviruses	Tick hematophagy	300	90	400
12	Mosquito-vectored flaviviruses	Mosquito hematophagy		79	100

Footnote: Node numbers correspond to those shown in **Fig. 3a**. Hematophagy estimates obtained from [35]. Culex-Aedes divergence date obtained from [56]. All other divergence data estimates were obtained from TimeTree [33].

Acknowledgements

RJG was funded by the Medical Research Council of the United Kingdom (MC_UU_12014/12). We thank Joseph Hughes, Alain Kohl, Spyros Lytras, Emilie Pondeville, Charles Rice, David Robertson, Greg Towers, and Sam J. Wilson for critical reading of the manuscript.

References

1. Manns, M.P., et al., *Hepatitis C virus infection.* Nature Reviews Disease Primers, 2017. **3**(1): p. 17006.
2. Pierson, T.C. and M.S. Diamond, *The continued threat of emerging flaviviruses.* Nat Microbiol, 2020. **5**(6): p. 796-812.
3. Simmonds, P., et al., *ICTV Virus Taxonomy Profile: Flaviviridae.* J Gen Virol, 2017. **98**(1): p. 2-3.
4. Chambers, T.J., et al., *Flavivirus genome organization, expression, and replication.* Annu Rev Microbiol, 1990. **44**: p. 649-88.
5. Shi, M., et al., *Divergent Viruses Discovered in Arthropods and Vertebrates Revise the Evolutionary History of the Flaviviridae and Related Viruses.* J Virol, 2016. **90**(2): p. 659-69.
6. Shi, M., et al., *The evolutionary history of vertebrate RNA viruses.* Nature, 2018. **556**(7700): p. 197-202.
7. Parry, R. and S. Asgari, *Discovery of novel crustacean and cephalopod flaviviruses: insights into evolution and circulation of flaviviruses between marine invertebrate and vertebrate hosts.* J Virol, 2019.
8. Porter, A.F., et al., *Novel hepaci- and pegi-like viruses in native Australian wildlife and non-human primates.* Virus Evol, 2020. **6**(2): p. veaa064.
9. Paraskevopoulou, S., et al., *Viromics of extant insect orders unveil the evolution of the flavi-like superfamily.* Virus Evol, 2021. **7**(1): p. veab030.
10. Kobayashi, K., et al., *Gentian Kobu-sho-associated virus: a tentative, novel double-stranded RNA virus that is relevant to gentian Kobu-sho syndrome.* Journal of General Plant Pathology, 2012. **79**: p. 56-63.
11. Qin, X.C., et al., *A tick-borne segmented RNA virus contains genome segments derived from unsegmented viral ancestors.* Proc Natl Acad Sci U S A, 2014. **111**(18): p. 6744-9.
12. Zanotto, P.M., et al., *Population dynamics of flaviviruses revealed by molecular phylogenies.* Proc Natl Acad Sci U S A, 1996. **93**(2): p. 548-53.
13. Cook, S. and E.C. Holmes, *A multigene analysis of the phylogenetic relationships among the flaviviruses (Family: Flaviviridae) and the evolution of vector transmission.* Arch Virol, 2006. **151**(2): p. 309-25.
14. Gould, E.A., et al., *Origins, evolution, and vector/host coadaptations within the genus Flavivirus.* Adv Virus Res, 2003. **59**: p. 277-314.
15. Moureau, G., et al., *New insights into flavivirus evolution, taxonomy and biogeographic history, extended by analysis of canonical and alternative coding sequences.* PLoS One, 2015. **10**(2): p. e0117849.
16. Grüning, B., et al., *Practical Computational Reproducibility in the Life Sciences.* Cell Syst, 2018. **6**(6): p. 631-635.
17. Duchene, S., E.C. Holmes, and S.Y. Ho, *Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates.* Proc Biol Sci, 2014. **281**(1786).
18. Aiewsakun, P. and A. Katzourakis, *Time-Dependent Rate Phenomenon in Viruses.* J Virol, 2016. **90**(16): p. 7184-95.
19. Katzourakis, A. and R.J. Gifford, *Endogenous viral elements in animal genomes.* PLoS Genet, 2010. **6**(11): p. e1001191.

20. Lequime, S. and L. Lambrechts, *Discovery of flavivirus-derived endogenous viral elements in Anopheles mosquito genomes supports the existence of Anopheles-associated insect-specific flaviviruses.* Virus Evol, 2017. **3**(1): p. vew035.
21. Whitfield, Z.J., et al., *The Diversity, Structure, and Function of Heritable Adaptive Immunity Sequences in the Aedes aegypti Genome.* Curr Biol, 2017. **27**(22): p. 3511-3519.e7.
22. Blair, C.D., K.E. Olson, and M. Bonizzoni, *The Widespread Occurrence and Potential Biological Roles of Endogenous Viral Elements in Insect Genomes.* Curr Issues Mol Biol, 2020. **34**: p. 13-30.
23. Singer, J.B., et al., *GLUE: a flexible software system for virus sequence data.* BMC Bioinformatics, 2018. **19**(1): p. 532.
24. Gifford, R.J. *Flavivirus-GLUE*. 2021; Available from: https://giffordlabcvr.github.io/Flavivirus-GLUE.
25. Holmes, E.C. and S. Duchêne, *Can Sequence Phylogenies Safely Infer the Origin of the Global Virome?* mBio, 2019. **10**(2).
26. Merkel, D., *Docker: lightweight linux containers for consistent development and deployment.* Linux Journal, 2014. **239**(2).
27. Blitvich, B.J. and A.E. Firth, *A Review of Flaviviruses that Have No Known Arthropod Vector.* Viruses, 2017. **9**(6).
28. Skoge, R.H., et al., *New virus of the family Flaviviridae detected in lumpfish (Cyclopterus lumpus).* Arch Virol, 2018. **163**(3): p. 679-685.
29. Geoghegan, J.L., et al., *Hidden diversity and evolution of viruses in market fish.* Virus Evol, 2018. **4**(2): p. vey031.
30. Ophinni, Y., et al., *piRNA-Guided CRISPR-like Immunity in Eukaryotes.* Trends Immunol, 2019. **40**(11): p. 998-1010.
31. Suzuki, Y., et al., *Non-retroviral Endogenous Viral Element Limits Cognate Virus Replication in Aedes aegypti Ovaries.* Curr Biol, 2020. **30**(18): p. 3495-3506.e6.
32. Suzuki, Y., et al., *Uncovering the Repertoire of Endogenous Flaviviral Elements in Aedes Mosquito Genomes.* J Virol, 2017. **91**(15).
33. Kumar, S., et al., *TimeTree: A Resource for Timelines, Timetrees, and Divergence Times.* Mol Biol Evol, 2017. **34**(7): p. 1812-1819.
34. Monahan-Earley, R., A.M. Dvorak, and W.C. Aird, *Evolutionary origins of the blood vascular system and endothelium.* J Thromb Haemost, 2013. **11 Suppl 1**(Suppl 1): p. 46-66.
35. Mans, B.J., *Evolution of vertebrate hemostatic and inflammatory control mechanisms in blood-feeding arthropods.* J Innate Immun, 2011. **3**(1): p. 41-51.
36. Gibb, R., et al., *Mammal virus diversity estimates are unstable due to accelerating discovery effort.* Biol Lett, 2022. **18**(1): p. 20210427.
37. Zhou, W., et al., *Exosomes serve as novel modes of tick-borne flavivirus transmission from arthropod to human cells and facilitates dissemination of viral RNA and proteins to the vertebrate neuronal cells.* PLoS Pathog, 2018. **14**(1): p. e1006764.
38. Hermance, M.E. and S. Thangamani, *Tick□Virus□Host Interactions at the Cutaneous Interface: The Nidus of Flavivirus Transmission.* Viruses, 2018. **10**(7).
39. Ochsenreiter, R., I.L. Hofacker, and M.T. Wolfinger, *Functional RNA Structures in the 3'UTR of Tick-Borne, Insect-Specific and No-Known-Vector Flaviviruses.* Viruses, 2019. **11**(3).
40. Gritsun, T.S. and E.A. Gould, *Direct repeats in the flavivirus 3' untranslated region; a strategy for survival in the environment?* Virology, 2007. **358**(2): p. 258-65.

41. Gritsun, D.J., et al., *Molecular archaeology of Flaviviridae untranslated regions: duplicated RNA structures in the replication enhancer of flaviviruses and pestiviruses emerged via convergent evolution.* PLoS One, 2014. **9**(3): p. e92056.
42. Gifford, R.J., *Viral evolution in deep time: lentiviruses and mammals.* Trends Genet, 2012. **28**(2): p. 89-100.
43. Zhang, Y.Z., et al., *The diversity, evolution and origins of vertebrate RNA viruses.* Curr Opin Virol, 2018. **31**: p. 9-16.
44. Dolja, V.V. and E.V. Koonin, *Metagenomics reshapes the concepts of RNA virus evolution by revealing extensive horizontal virus transfer.* Virus Res, 2018. **244**: p. 36-52.
45. Kawasaki, J., et al., *100-My history of bornavirus infections hidden in vertebrate genomes.* Proc Natl Acad Sci U S A, 2021. **118**(20).
46. Ghafari, M., et al., *A mechanistic evolutionary model explains the time-dependent pattern of substitution rates in viruses.* Curr Biol, 2021. **31**(21): p. 4689-4696.e5.
47. Li, Y., et al., *Discovery of Flaviviridae-derived endogenous viral elements in shrew genomes provide novel insights into Pestivirus ancient history.* bioRxiv, 2022: p. 2022.02.11.480044.
48. Gifford, R.J. and J. Singer. *HCV-GLUE*. 2021; Available from: http://hcv-glue.cvr.gla.ac.uk/.
49. Singer, J., et al., *CoV-GLUE: A Web Application for Tracking SARS-CoV-2 Genomic Variation.* Preprints, 2020.
50. Zhu, H., et al., *Database-integrated genome screening (DIGS): exploring genomes heuristically using sequence similarity search tools and a relational database.* bioRxiv, 2018.
51. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nuc. Acids Res., 1997. **25**: p. 3389-3402.
52. Kitts, P.A., et al., *Assembly: a resource for assembled genomes at NCBI.* Nucleic Acids Res, 2016. **44**(D1): p. D73-80.
53. Stamatakis, A., *RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.* Bioinformatics, 2014. **30**(9): p. 1312-3.
54. Waterhouse, A.M., et al., *Jalview Version 2--a multiple sequence alignment editor and analysis workbench.* Bioinformatics, 2009. **25**(9): p. 1189-91.
55. Gifford, R.J., et al., *Nomenclature for endogenous retrovirus (ERV) loci.* Retrovirology, 2018. **15**(1): p. 59.
56. Sieglaff, D.H., et al., *Comparative genomics allows the discovery of cis-regulatory elements in mosquitoes.* Proc Natl Acad Sci U S A, 2009. **106**(9): p. 3053-8.