Comparative analysis of genome-encoded viral sequences reveals the evolutionary history of *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 calibrate the timeline of flavivirus evolution using flavivirus-derived DNA sequences identified in animal genomes. 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 of evolution, implying that broad-scale comparative analysis can reveal fundamental insights into flavivirus biology. 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.

Significance

Understanding how pathogenic viruses evolved can provide vital insights into their biology. In this study we use genomic data to show that flaviviruses – a group of viruses that includes important pathogens such as Dengue virus and hepatitis C virus - arose through an extended history of evolutionary interaction with host and vector species. Our findings show that comparative studies can productively utilise genomic data to reveal insights into flavivirus biology, which will help to facilitate the development of more effective antiviral treatments and strategies.

Introduction

The flaviviruses (family *Flaviviridae*) are a diverse group of enveloped, positive-strand RNA viruses, many of which cause 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) is a major cause of chronic liver disease [1, 2].

Currently, four flavivirus genera are recognised: *Pegivirus, Pestivirus*, *Hepacivirus* and *Flavivirus* [3]. These genera contain viruses that have 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 virome. Most of these newly identified viruses have yet to be incorporated into official taxonomy [5-9]. 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], and 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 including mechanisms of pathogenesis and likelihood of zoonosis. However, knowledge of the appropriate evolutionary timescale is critical. 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 [12]. However, 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]. Unfortunately, due to time-dependent effects on viral evolutionary rates [13, 14], it is not possible to discriminate between these starkly contrasting possibilities using genome data sampled from contemporary flaviviruses.

'Endogenous viral elements' (EVEs) are virus-derived DNA sequences found in the genomes of metazoan species [15]. EVEs derived from RNA viruses are thought to originate when infection of germline cells leads to the generation of viral cDNA and its subsequent integration into chromosomal DNA (presumably mediated by retroelement activity) [16]. The integrated viral genes can be inherited as host alleles and sometimes persist in the gene pool long enough to become genetically fixed (i.e. reach a frequency of 100% in the species gene pool). These fixed EVEs often derive from viruses that circulated millions of years ago and thereby provide a rare glimpse into the distant history of virus evolution. EVEs derived from flaviviruses, which we refer to here as 'endogenous flaviviral elements' (EFVs), have been identified in some arthropod species including insects [15, 17, 18] and crustaceans [7]. Interestingly, some of these – most notably those found in the genomes of *Aedes* mosquitoes - have been shown to produce PIWI-interacting RNAs (piRNAs) in the antisense orientation, suggesting they might provide the basis for a form of heritable antiviral immunity analogous to that of CRISPR-Cas [16, 18-20]. Here, we use comparative approaches to investigate the long-term evolutionary history of flaviviruses, making extensive use of the EVE 'fossil record'. We show that although EFV sequences have only rarely been fixed in the animal germline, 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 that can be used to reproduce our analyses and facilitates re-use of the associated data items.

RESULTS

Creation of open resources for reproducible genomic analysis of flaviviruses

Comparative genomic analyses generally entail the construction of complex data sets comprising molecular sequence data linked to other kinds of information. Usually these include genome feature annotations and multiple sequence alignments (MSAs) as well as other diverse kinds of data. We used GLUE - an open, platform-independent software toolkit for storage and interpretation of sequence data [21] - to develop an openly accessible online resource for comparative genomic analysis of flaviviruses (Flavivirus-GLUE) that preserves the 'state' of our data so that our analyses can be precisely and widely replicated (**Fig. S1a-b**). Furthermore, hosting of the Flavivirus-GLUE project in an openly accessible online version control system (GitHub) provides a platform for ongoing development of this resource by multiple collaborators, following practices established in the software industry (**Fig. S1c**).

The Flavivirus-GLUE project incorporates all of the data items typically used in comparative sequence analysis of flaviviruses, including: (i) a set of reference sequences representing all known flavivirus species (**Table S1**); (ii) sequence and isolate-specific information (e.g. host species, vector species) in tabular form; (iii) a standardized set of flavivirus genome features and their coordinates within selected reference genome sequences; (iv) a set of MSAs incorporating all sequences in the project. Loading the project into the GLUE 'engine' generates a relational database that not only contains the data items associated with our analysis, but also represents the complex semantic links between them (**Fig. S1a**). The Flavivirus-

GLUE project can be installed on all commonly-used computing platforms and is also fully containerised via Docker [22]. Reproducible comparative analysis can be implemented within the Flavivirus-GLUE project using GLUE's command layer, which can coordinate interactions between the relational database and bioinformatics software tools (**Fig. S1b**). These interactions can leverage 'structured query language' (SQL) - the standard user and application program interface of relational databases – to provide a high degree of flexibility and power over the way in which data are retrieved.

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

We incorporated EFVs into our Flavivirus-GLUE project so that the unique information contained in these sequences could be utilised in comparative analysis. 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 1**, **Table S1**). We constructed multiple sequence alignments (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**). We used MSAs containing EFV sequences as input for maximum likelihood phylogenetic analysis to reconstruct the evolutionary relationships between EFVs and contemporary flaviviruses (**Fig. 2, Fig S3**). EFVs were assigned unique identifiers that (i) capture information about the taxonomic group of viruses they are derived from, (ii) uniquely identify the germline incorporation event and genomic locus associated with each sequence, and (iii) indicate the distribution of related sequences across host species (**Fig. 1**). EFV locus and consensus sequences, MSAs, and all associated metadata were incorporated into Flavivirus-GLUE. We extended the database schema of Flavivirus-GLUE to capture the semantic

relationships between these data items, so that they can be interactively accessed via the underlying database.

Phylogenetic reconstructions show that the *Flaviviridae* are clearly split into two major lineages - “pegi-hepaci” and “flavi-pesti” – each of which contains several well-supported subgroups [9]. EFVs derived from each of these major flavivirus lineages and subgroups are represented in the genomic ‘fossil record’. A single EFV that derives from the “pegi-hepaci” lineage was identified in the genome of a marine mollusc - the Eastern emerald elysia (*Elysia chlorotica*), demonstrating that the host range of these viruses extends to invertebrates (**Fig 2a**). However, all other EFVs identified in our screen grouped with the “flavi-pesti” lineage, and these included representatives of diverse “flavi-like” and “pesti-like” groups (**Fig 2b-g, Fig S3**). In addition, an EFV recovered from the genome of a priapulid worm (*Priapulus caudatus*) appears to represent a novel subgroup, here labelled ‘X1’, within the “flavi-like” lineage (**Fig 2b**).

Fixation of flavivirus-derived elements in the animal germline is rare

Of the 1075 species screened in our analysis only 36 were found to harbour EFVs in their genomes, but those that did often contained multiple distinct EFV loci. We estimate these EFVs to represent at least 25 distinct germline colonisation events. The low number of events relative to sequences reflects our decision to conservatively assume that the presence of multiple EFV loci in individual species or taxonomic groups reflects intragenomic amplification rather than multiple independent germline incorporation events. The number of events could potentially be much higher if distinct germline incorporation events by closely related viruses have occurred during the evolution of these species. Notably, however, where multiple EFV sequences occur, they generally comprise contiguous and/or overlapping fragments of larger genome segments and in many cases, the overlapping segments are nearly identical at nucleotide level (**Fig. S2)**. This suggests

that longer cDNA sequences may been fixed in the germline initially before subsequently being duplicated and/or fragmented.

The largest set of EFV sequences putatively derived from a single germline incorporation event is found in the insect superfamily Apoidae (bees and sphecoid wasps). We identified >200 copies in total, and copy number is dramatically inflated in certain species, such as the spurred ceratina (*Ceratina calcarata*) which has at least 86 distinct copies (**Table 1**). The identification of orthologous elements demonstrates these elements derive from a germline incorporation event that occurred >100 Mya (**Fig. S4**). Furthermore, their occurrence across a wide range of related species suggests that germline incorporation predates deep divergences within the Apoidae, and high copy number reflects subsequent intragenomic amplification. Notably, while most members of this set of elements, labelled EFV-PL.2-Apoidae, are highly degraded, some elements contain relatively intact ORFs, with the longest intact coding regions being identified in *C. calcarata* - the species with the highest overall copy number (**Table 1**).

Comprehensive phylogenetic analysis revises flavivirus phylogeny

We performed a comprehensive phylogenetic analysis of the *Flaviviridae*, using all available information, including endogenous flaviviral elements (EFVs). We used GLUE's 'constrained alignment tree data structure to reconstruct bootstrapped, maximum likelihood phylogenies across a range of taxonomic levels within the *Flaviviridae* (**Table 2, Fig. 2, Fig. S5**). Our analysis revealed new information about phylogenetic structure within each major lineage. Within the "pegi-hepaci" lineage the *Pegivirus* and *Hepacivirus* genera occupy a derived position [8], viruses identified in sharks group in an intermediate position, and the invertebrate lineage represented by EFV-PegiHepaci-*Elysia* groups basally (**Fig. 2a**). The "flavi-pesti" lineage is more structured – it comprises robustly supported "pesti-like" and "flavi-like" clades, each of which has a clear and well-supported internal structure (**Fig. 2b**).

In contrast to previous reports [9] we find robust support for a clean division of "flavi-like" viruses into three discrete, monophyletic clades (**Fig 2c**). Significantly, this reflects a separation between the flaviviruses (genus *Flavivirus*) and viruses related to 'Tamana bat virus' (TABV), an unusual flavivirus that is often thought of a basal member of genus *Flavivirus* [23]. However, we identified a *Flavivirus*-derived EFV in the genome of a basal animal - the freshwater jellyfish *Craspadecusta sowerbii* – that splits TABV from the *Flavivirus* genus. This observation, combined with the extensive diversity and broad host of TABV-like viruses, indicates that they represent an entirely distinct subgroup of 'flavi-like' viruses, which we refer to here as genus *Tamanavirus*, in line with a previous proposal [24] (**Fig. 2e**).

Within the flavivirus genus derived lineages include arthropod-vectored viruses ('arboviruses'), with the tick-borne and mosquito-borne flaviviruses emerging as distinct, robustly supported clades (**Fig. 2e, Fig S3b**). Additional, basal lineages are present containing viruses with no known vector, and viruses that are restricted to arthropod hosts [7, 9, 25]. In most phylogenies NKV flaviviruses cluster basal to all arthropod/vertebrate flaviviruses, but an alternative topology places these NKV flaviviruses more closely with the tick-borne flaviviruses, as reported by other authors [23, 26].

The remaining "flavi-like" viruses cluster together with jingmenviruses in a robustly supported clade (**Fig. 2f**). In addition to identifying EFVs that group robustly within the diversity of JMVs, we also identified a more distantly related, 'Jingmenvirus-like' lineage – here labelled X2 - that is comprised entirely of EFVs and forms a sister clade to the main JMV lineage (**Fig. 2c**). Notably, X2-derived EFVs were identified in an arthropod species (the calanoid copepod *Eurytemora affinis*) as well as in a diverse range of ray-finned fish (clade Actinopterygii) (**Table 1**). We did not identify any X2-derived insertions that spanned outside of NS5 gene, therefore the event that generated the segmented structure of the JMVs could potentially

predate their split from the X2 lineage, so that X2 viruses also have segmented genomes.

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 two invertebrate-associated clades (here labelled 'PL1' and 'PL2') that have been collectively referred to as 'large genome flaviviruses' (LGFs) [9]. The PL2 subgroup contains both viruses and EFVs (**Fig. 2g**). The EFV sequences identified in this group are among the largest we identified and included many overlapping fragments that could be assembled into larger genome fragments. However, the longest consensus sequences we could assemble were ~20% shorter than the genome of the most closely related contemporary virus (**Fig. 1b**).

All 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 the *Chironomus* midges, for several X2-derived EVEs in fish, and for a PL2-derived EVE in superfamily Apoidea (**Table 2**). The PL2-derived EFV was the oldest - we identified orthologous loci in two species – the spurred ceratina (*Ceratina calcarata*) and the green orchid bee (*Euglossa dilemma*) - that diverged >100 million years ago (Mya) [27] (**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**).

Despite their demonstrably ancient origins, PL2-derived EFV sequences show a relatively close evolutionary relationship to contemporary LGFs (**Fig 2f**). This observation, together 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.

We integrated the main findings of our comprehensive phylogenetic analysis (**Fig. 2**) with the various date calibrations obtained via comparative genomic and co-phyletic analysis to reconstruct a broad overview of *Flaviviridae* evolution (**Fig. 3a**) (**Table 3**). The extended evolutionary timeline implied by our analysis is consistent with a scenario in which vector-borne flaviviruses emerge broadly in parallel with the evolution of hematophagy (blood-feeding) in ticks and insects. We therefore incorporated additional maximum age calibrations based on the assumption that blood-feeding arthropods vectors evolved prior to the emergence of vectored viruses. Similarly, we infer that Gentian Kobu-sho-associated virus (GKaV) - a putative double-stranded RNA virus of plants [10] that groups within the PL1 lineage (**Fig 2c**) - emerged from an arthropod source via plant feeding (as has been proposed for other enveloped virus groups found in plants [28]), and that both this event and the subsequent transition from single-stranded to double-stranded RNA postdate the evolution and expansion of flowering plants (Angiospermae).

Arthropod-vectored flaviviruses emerged from an arachnid source

Phylogenetic reconstructions show that the ‘classical insect-specific flaviviruses’ (cISFs) are robustly separated from the ‘classical vector-borne flaviviruses’ (CFVs) by viruses identified in crustaceans, sea spiders (Chelicerata), and ticks (Acari) (**Fig 3b**). In addition, trees rooted on TABV or the cnidarian EFVs show that the most basal CFV lineages are either tick-associated (Mpulungu and the

TBFVs) or potentially tick-associated [23]. In addition, several synapomorphies were identified in the NS5 protein where an ancestral residue is conserved between ancestral and tick-borne flaviviruses, but mutated in insect-vectored flaviviruses (**Fig 3b**, **Fig. S7**). The most parsimonious interpretation of these observations is that they reflect firstly 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. This model would imply that members of the flavivirus genus for which no known vector has been identified [23] are in reality transmitted by an arthropod vector, or else evolved from viruses that were.

DISCUSSION

In this study we used EFV sequences to investigate the long-term evolutionary history of the *Flaviviridae* and calibrate their evolutionary timeline. We identified EFVs derived from a wide range of flavivirus groups in a diverse range of host species. We reveal the first evidence that hepacivirus-like viruses, which have so far only been identified in vertebrates, have counterparts in invertebrate species. We also show that Jingmenvirus-like elements are present in fish, providing first unambiguous evidence for the existence of EFVs in vertebrate genomes. Hepacivirus-related EVEs have previously been reported in rabbits [29], but these sequences are very short (i.e. 20-30 nucleotides) making their origins difficult to investigate further.

We show that flavivirus-like viruses have rarely been fixed in the animal germline - only 36 of 1075 species genomes screened in our analysis contained any EFVs (**Table 1**). Our screen suggests they most often occur in arthropod genomes, but less commonly than EFVs derived from other RNA virus groups [20]. But while EFVs are rare overall, they occur at relatively high copy number in the germline of some animal species and groups (e.g., *Aedes* mosquitoes, superfamily Apoidea). High EFV copy number in *Aedes* species is likely related to role that these

sequences play in antiviral immunity [19]. However, it remains unclear the extent to which it reflects intragenomic duplication of EFVs following rare germline incorporation events, versus numerous distinct incorporations as would be expected for a dynamic system of heritable antiviral immunity [16]. The extremely high copy number of PL2-derived EFVs in superfamily Apoidea could potentially reflect amplification in association with retroelement activity, as has been reported for other kinds of EVE [30, 31], but multiple invasions involving related viruses cannot be ruled out.

The EFVs identified in our study include some remarkably ancient examples >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 (**Fig 3a**). The extended evolutionary timeline implied by our research raises interesting questions about the fundamental way 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 ecological shifts such as the emergence of endothermic vertebrates and flowering plants. The identification of flavivirus-derived EVEs in animals lacking a circulatory system (e.g. Cnidarians and Priapulids) is also consistent with a primordial origin in animals, and suggests that virus transmission via exosomes - as has been reported for TBFVs [32] - might represent the most archaic form of cell-to-cell transmission for RNA viruses.

The vertebrate blood vascular system evolved >400 Mya and was established in its basic form in endothermic vertebrates (birds and mammals) by ~200 Mya [33]. Once systems for the transport of nutrients had evolved, parasitism of these systems inevitably followed. Accordingly, hematophagous behaviour is thought to have evolved independently in arthropods on >20 occasions [34]. The time-calibrated

phylogeny obtained in our study suggests this also led to independent emergence of arthropod-borne flaviviruses that infect vertebrate species. Besides the various groups of arthropod-borne classical flaviviruses (CFVs), it is now clear that some JMVs are arthropod-borne viruses that infect mammals, including humans [35], and recent studies suggest that dual-host relationships involving vertebrates and arthropods may also exist among tamanaviruses [7]. Furthermore, we find that the JMV-related X2 lineage is represented by EVEs in both fish and copepod genomes, suggesting it may contain fish viruses that are (or were) transmitted by hematophagous copepods.

We show that phylogenetic relationships within the *Flavivirus* genus are consistent with a scenario under which the ancestors of the arthropod-borne classical flaviviruses (CFVs) circulated exclusively within ancestral arthropods and expanded into vertebrates in association with the evolution of hematophagy in specific arthropod groups (**Fig. 2**, **Fig. 3a**). No EVEs derived from the arthropod-borne lineage of classical flaviviruses were identified in our screen and estimates of when hematophagous traits evolved in ticks and mosquitoes vary widely [34]. Thus, we cannot draw definitive conclusions about the timing of their emergence [36]. However, the phylogenetic relationships within genus *Flavivirus* provide compelling evidence that CFVs originated in arachnids and latterly acquired the capacity to infect and be transmitted by insects (**Fig. 3b**). 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 [37]. 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 [38] - evolved via duplications of a ~200 nucleotide region within the NS5 ORF [39]. These 'long repeated sequences' are best preserved in the TBFV lineage - with only remnants being present in the more derived CFV groups [40].

Our findings are consistent with a growing understanding that, despite extremely high background mutation rates and an inherent capacity for rapid adaptation, the overall structural and functional architecture of virus genomes often remains fundamentally conserved over immense expanses of time [30, 41-44]}. This implies that comparative approaches can be applied to reveal the structural and functional basis of key adaptations, such as the capacity to replicate in both arthropod and vertebrate hosts. In this study we introduced a template for computational genomics studies of viruses that focusses on facilitating the reproduction of comparative analyses and re-use of the complex datasets that underpin them (e.g. MSAs) (**Fig. S1**). This approach can not only scale to accommodate greatly increased quantities of virus species and sequences, but also introduces new levels of reproducibility and re-usability so that researchers working in different areas of flavivirus genomics - but utilizing related data and domain knowledge - can benefit from one another's work.

MATERIALS & METHODS

Sequences and data

A library of flavivirus representative genome sequences (**Table S1**) was collated from GenBank via reference to the International Committee for Taxonomy of Viruses website. Whole genome sequence data of animal species were obtained from the National Center for Biotechnology Information genome database [45]. We obtained all animal genomes available as of March 2020.

Genome screening in silico

WGS data of animal species were obtained from the National Center for Biotechnology Information) genome database [45]. We obtained all animal genomes available as of March 2020. We used the database-integrated genome screening (DIGS) tool [46] to derive a non-redundant database of EFV loci within published WGS assemblies (**Table S1**). The DIGS tool uses the basic local alignment search tool (BLAST) program suite [47] to perform similarity searches and the MySQL relational database management system (Community Server 8.0.26) to coordinate screening and record output. A user-defined reference sequence library provides both a source of 'probes' for searching WGS data and a means of classifying DNA sequences recovered by screening. We used Flavivirus-GLUE to export a representative library of polypeptide sequences for the *Flaviviridae*, incorporating all previously characterised flavivirus species and EVEs, as input for DIGS (**Fig S8**). Sequences that could produce false positive matches – e.g. the host-derived sequence insertions found in the reference genome sequence of bovine viral diarrhoea virus 1 [48] – were excluded from the probe set. We extended the screening database schema to incorporate tables representing the taxonomic classifications of viruses, EFVs and host species, and used database queries constructed using SQL to filter matches based on their similarity reference sequences and their taxonomic distribution across hosts. Using this approach, we

categorised sequences 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). For genome sequences that were obtained from cell lines we discarded two sequences that disclosed <3% nucleotide divergence from contemporary Dengue strains, as these could potentially represent integrations that occurred in cell culture.

EFVs were assigned standard identifiers following a convention established for endogenous retroviruses [49] - the first component is the classifier 'EFV' (which may be dropped when it is implied by context). The second component comprises two subcomponents separated by a period; (i) the name of the taxonomic group of viruses from which the EFV is thought to derive (the lowest level taxonomic group into which the element can be confidently placed by phylogenetic analysis is used); (ii) a numeric ID that uniquely identifies the integration event from which the insertion derives within the taxonomic lineage it is assigned to. 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 mRNAs and from segmented viruses [43]– endogenous jingmenvirus helicase (EJH) and endogenous jingmenvirus polymerase (EJP).

Phylogenetic analysis

A process for reconstructing evolutionary relationships across the entire *Flaviviridae* was implemented using GLUE. We used a data structure called a 'constrained MSA tree' to coordinate genomic analyses across the large phylogenetic distances found in flaviviruses **Fig. S7d**. This approach addresses the issue that MSAs constructed at higher taxonomic levels (e.g., above genus-level in the *Flaviviridae*) typically contain far fewer columns than those constructed at low

taxonomic levels (e.g., subgenus), meaning that to fully investigate phylogenetic relationships using all of the available information it is often necessary to construct several separate MSAs each representing a distinct taxonomic grouping (e.g. see **Table 2**). 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 [50], particularly those that examine more distant evolutionary relationships genomic [51]. To address these issues the 'constrained MSA tree' data structure represents the hierarchical links between MSAs constructed at distinct taxonomic levels, creating in effect a single, unified MSA that can be used to reconstruct both shallow and deep evolutionary relationships while making use of the maximum amount of available information at each level (**Table 2, Fig. S1d**). This approach also has the effect of standardising the genomic coordinate space to the constraining reference sequences selected for each MSA without imposing any limitations on which references are used (e.g., laboratory strains versus wild-type references), since additional, alternative constraining references can be incorporated, and contingencies such as insertions relative to the constraining reference are dealt with in a standardised way [21].

MSAs partitions derived from the constrained MSA tree were used as input for phylogenetic reconstructions. Nucleotide and protein phylogenies were reconstructed using maximum likelihood (ML) as implemented in RAxML (version 8.2.12) [52]. To handle coverage-related issues we generated gene coverage data prior to phylogenetic analysis and used this information to condition the way in which taxa are selected into MSA partitions (**Supplementary**). Protein substitution models were selected via hierarchical maximum likelihood ratio test using the PROTAUTOGAMMA option in RAxML.

Genomic analysis

Open reading frames (ORFs) of EVEs were inferred by manual comparison to virus genomes. We used a BLAST-based, codon-aware aligner implemented in GLUE [21, 47] to generate in-frame alignments of virus and EFV virus coding domains, and to assemble a consensus genome of sea spider flavivirus from published transcripts [53]. JalView [54] (version 2.11.1.4) and Se-Al (version 2.0a11)

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 [49]. 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) 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.

(b) 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 [27]. 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.

(c) 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 viremic vertebrate hosts, thereby exposing mosquitoes and other hematophagous insects such that they might acquire a role as vectors.

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*	Bluespotted 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.

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
15	Gentian Kobu-Sho virus	Evolution of angiosperms		190	315

Footnote: Node numbers correspond to those shown in **Fig. 3a**. Hematophagy estimates obtained from [34]. Culex-Aedes divergence date obtained from [55]. Angiosperm emergence taken from [56]. All other divergence data estimates were obtained from TimeTree [27].

Acknowledgements

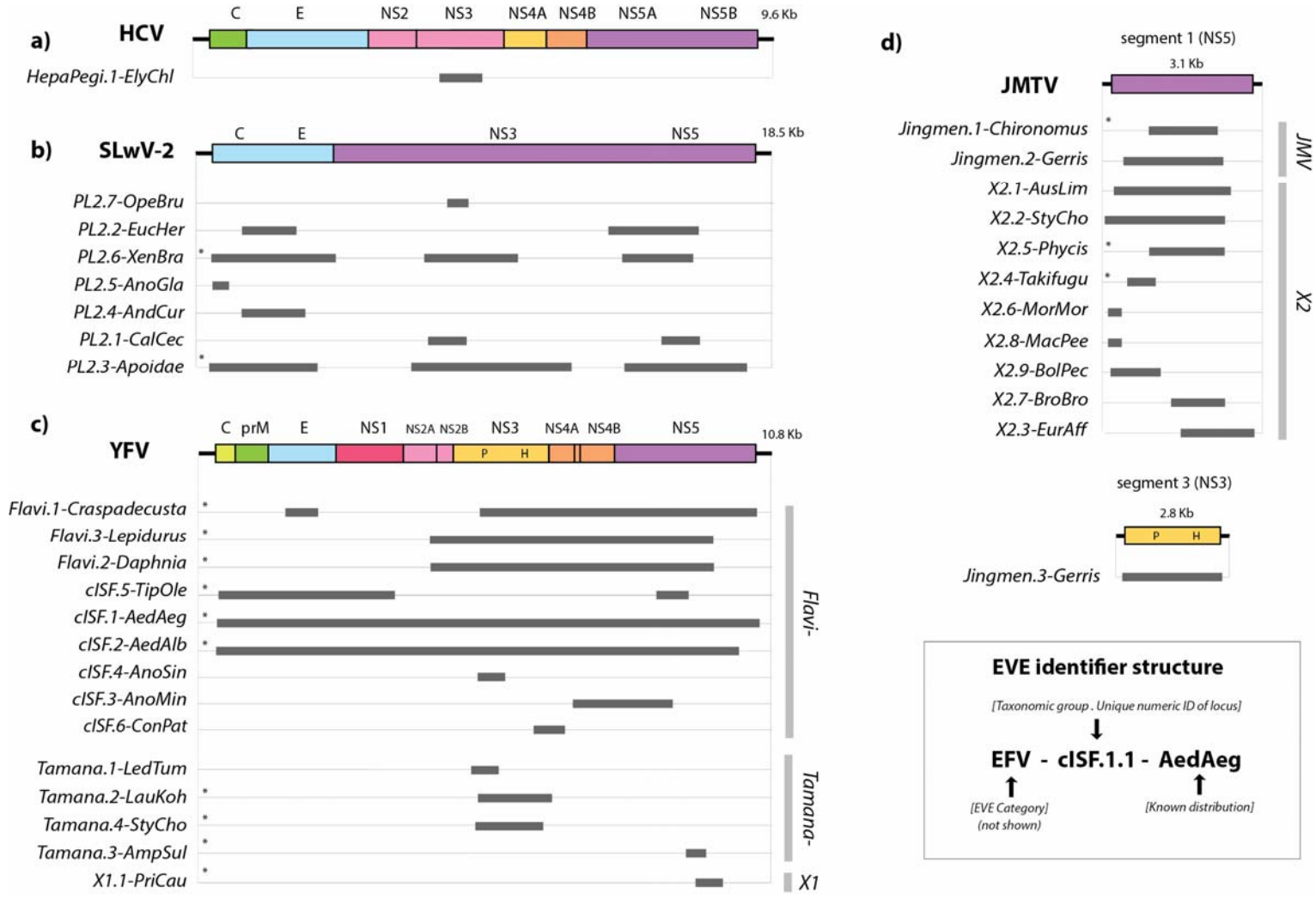
RJG was funded by the Medical Research Council of the United Kingdom (MC_UU_12014/12). We thank Joseph Hughes, Alain Kohl, Spyros Lytras, Emilie Pondeville, Charles Rice, David Robertson, Greg Towers, and Sam J. Wilson for critical reading of the manuscript.

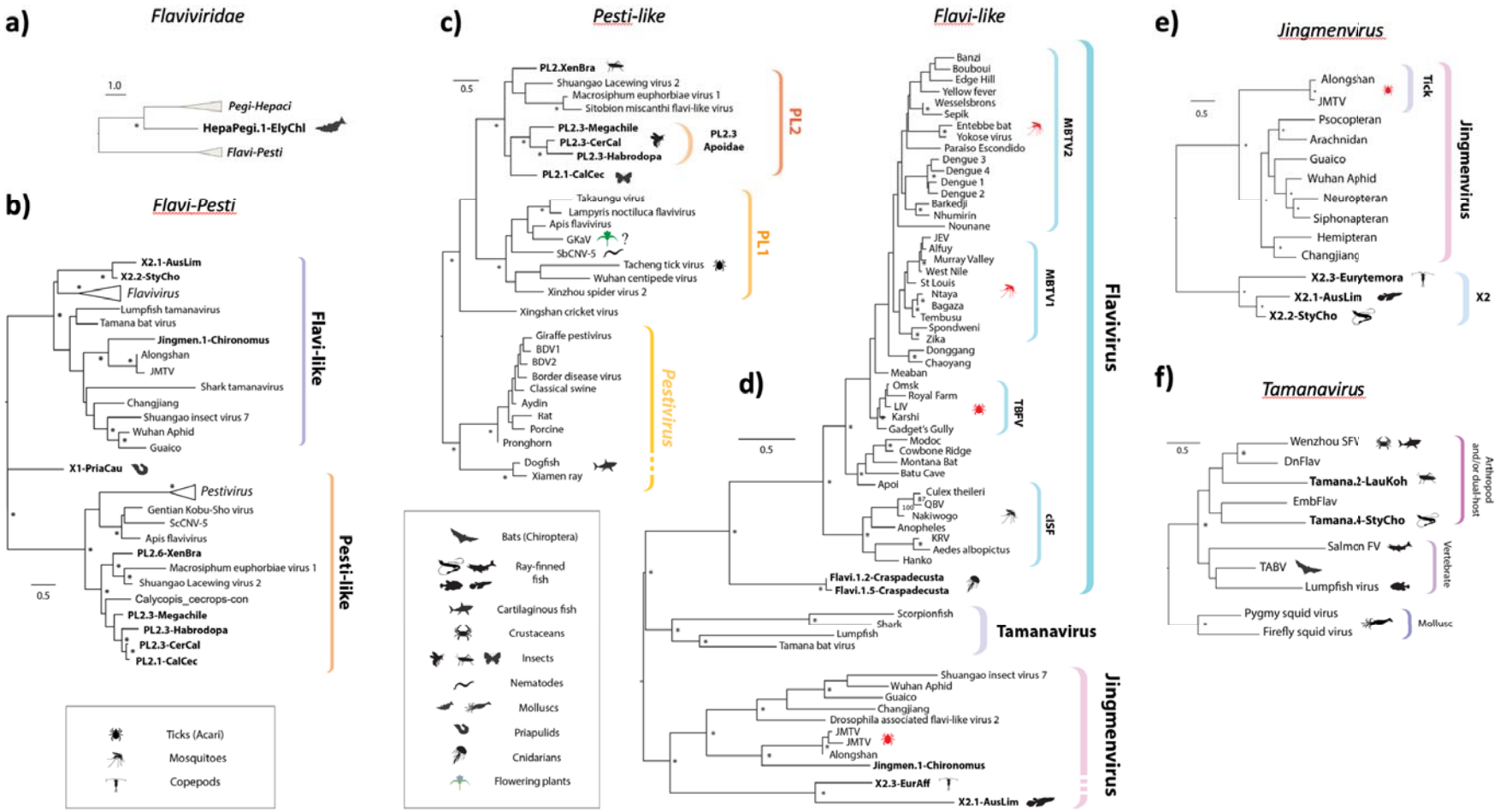
References

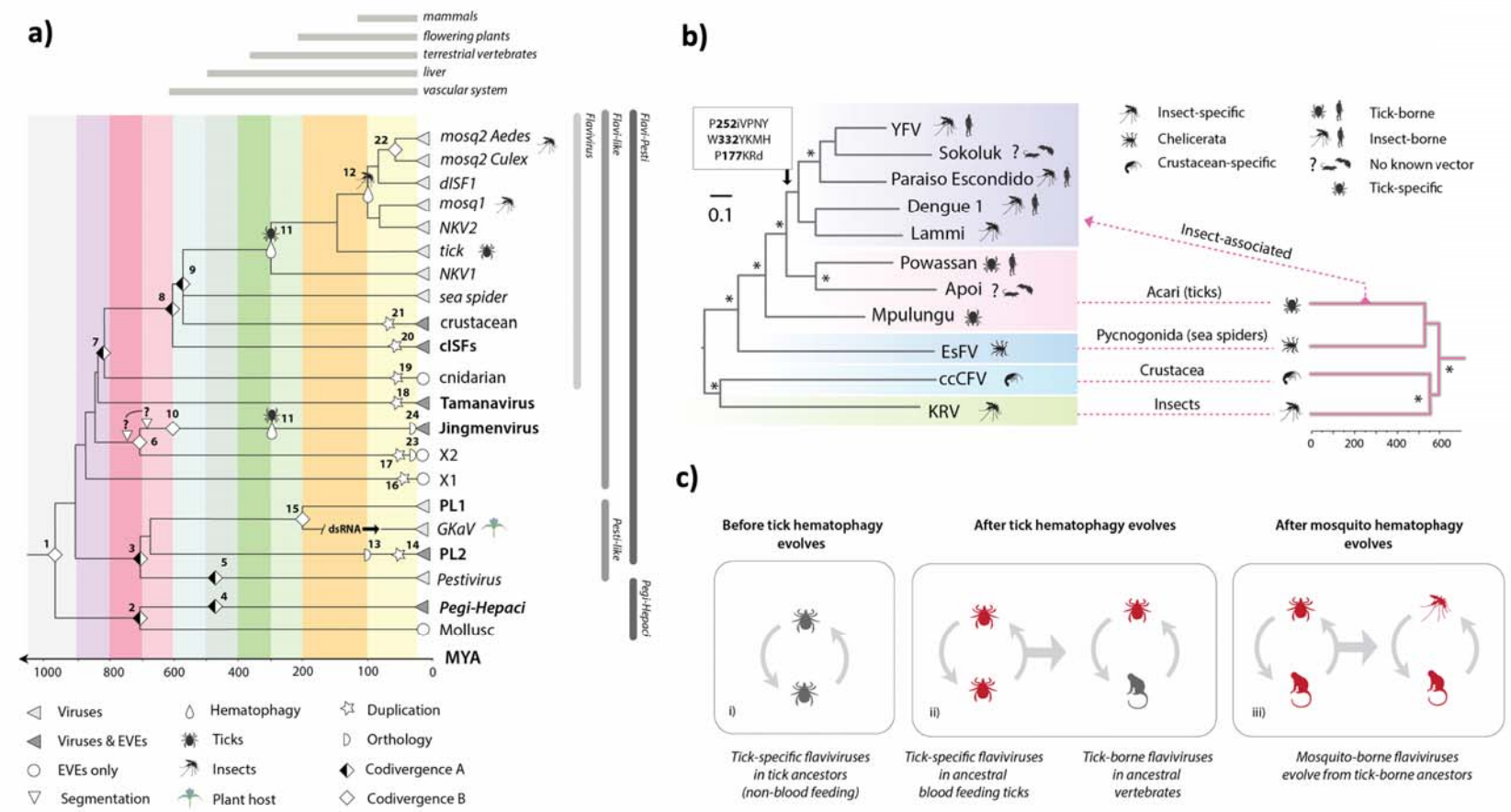
1. Manns, M.P., et al., *Hepatitis C virus infection.* Nature Reviews Disease Primers, 2017. **3**(1): p. 17006.
2. Pierson, T.C. and M.S. Diamond, *The continued threat of emerging flaviviruses.* Nat Microbiol, 2020. **5**(6): p. 796-812.
3. Simmonds, P., et al., *ICTV Virus Taxonomy Profile: Flaviviridae.* 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. Gould, E.A., et al., *Origins, evolution, and vector/host coadaptations within the genus Flavivirus.* Adv Virus Res, 2003. **59**: p. 277-314.
13. 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).
14. Aiewsakun, P. and A. Katzourakis, *Time-Dependent Rate Phenomenon in Viruses.* J Virol, 2016. **90**(16): p. 7184-95.
15. Katzourakis, A. and R.J. Gifford, *Endogenous viral elements in animal genomes.* PLoS Genet, 2010. **6**(11): p. e1001191.
16. Ophinni, Y., et al., *piRNA-Guided CRISPR-like Immunity in Eukaryotes.* Trends Immunol, 2019. **40**(11): p. 998-1010.
17. 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.
18. 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.

19. 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.
20. 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.
21. Singer, J.B., et al., *GLUE: a flexible software system for virus sequence data.* BMC Bioinformatics, 2018. **19**(1): p. 532.
22. Merkel, D., *Docker: lightweight linux containers for consistent development and deployment.* Linux Journal, 2014. **239**(2).
23. Blitvich, B.J. and A.E. Firth, *A Review of Flaviviruses that Have No Known Arthropod Vector.* Viruses, 2017. **9**(6).
24. de Lamballerie, X., et al., *Genome sequence analysis of Tamana bat virus and its relationship with the genus Flavivirus.* J Gen Virol, 2002. **83**(Pt 10): p. 2443-54.
25. Blitvich, B.J. and A.E. Firth, *Insect-specific flaviviruses: a systematic review of their discovery, host range, mode of transmission, superinfection exclusion potential and genomic organization.* Viruses, 2015. **7**(4): p. 1927-59.
26. Temmam, S., et al., *Characterization of Viral Communities of Biting Midges and Identification of Novel Thogotovirus Species and Rhabdovirus Genus.* Viruses, 2016. **8**(3): p. 77.
27. Kumar, S., et al., *TimeTree: A Resource for Timelines, Timetrees, and Divergence Times.* Mol Biol Evol, 2017. **34**(7): p. 1812-1819.
28. Buchmann, J.P. and E.C. Holmes, *Cell Walls and the Convergent Evolution of the Viral Envelope.* Microbiol Mol Biol Rev, 2015. **79**(4): p. 403-18.
29. Silva, E., et al., *Endogenous hepatitis C virus homolog fragments in European rabbit and hare genomes replicate in cell culture.* PLoS One, 2012. **7**(11): p. e49820.
30. Lytras, S., G. Arriagada, and R.J. Gifford, *Ancient evolution of hepadnaviral paleoviruses and their impact on host genomes.* Virus Evol, 2021. **7**(1): p. veab012.
31. Inoue, Y., et al., *Complete fusion of a transposon and herpesvirus created the Teratorn mobile element in medaka fish.* Nat Commun, 2017. **8**(1): p. 551.
32. 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.
33. 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.
34. Mans, B.J., *Evolution of vertebrate hemostatic and inflammatory control mechanisms in blood-feeding arthropods.* J Innate Immun, 2011. **3**(1): p. 41-51.
35. Wang, Z.D., et al., *A New Segmented Virus Associated with Human Febrile Illness in China.* N Engl J Med, 2019. **380**(22): p. 2116-2125.
36. Mans, B.J., A.I. Louw, and A.W.H. Neitz, *Evolution of Hematophagy in Ticks: Common Origins for Blood Coagulation and Platelet Aggregation Inhibitors from Soft Ticks of the Genus Ornithodoros.* Molecular Biology and Evolution, 2002. **19**(10): p. 1695-1705.
37. Hermance, M.E. and S. Thangamani, *Tick□Virus□Host Interactions at the Cutaneous Interface: The Nidus of Flavivirus Transmission.* Viruses, 2018. **10**(7).
38. 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).

39. 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.
40. 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.
41. Gifford, R.J., *Viral evolution in deep time: lentiviruses and mammals.* Trends Genet, 2012. **28**(2): p. 89-100.
42. Zhang, Y.Z., et al., *The diversity, evolution and origins of vertebrate RNA viruses.* Curr Opin Virol, 2018. **31**: p. 9-16.
43. Kawasaki, J., et al., *100-My history of bornavirus infections hidden in vertebrate genomes.* Proc Natl Acad Sci U S A, 2021. **118**(20).
44. Barreat, J.G.N. and A. Katzourakis, *Paleovirology of the DNA viruses of eukaryotes.* Trends Microbiol, 2021.
45. Kitts, P.A., et al., *Assembly: a resource for assembled genomes at NCBI.* Nucleic Acids Res, 2016. **44**(D1): p. D73-80.
46. Zhu, H., et al., *Database-integrated genome screening (DIGS): exploring genomes heuristically using sequence similarity search tools and a relational database.* bioRxiv, 2018.
47. 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.
48. Rinck, G., et al., *A cellular J-domain protein modulates polyprotein processing and cytopathogenicity of a pestivirus.* J Virol, 2001. **75**(19): p. 9470-82.
49. Gifford, R.J., et al., *Nomenclature for endogenous retrovirus (ERV) loci.* Retrovirology, 2018. **15**(1): p. 59.
50. Grüning, B., et al., *Practical Computational Reproducibility in the Life Sciences.* Cell Syst, 2018. **6**(6): p. 631-635.
51. Holmes, E.C. and S. Duchêne, *Can Sequence Phylogenies Safely Infer the Origin of the Global Virome?* mBio, 2019. **10**(2).
52. Stamatakis, A., *RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.* Bioinformatics, 2014. **30**(9): p. 1312-3.
53. Conway, M.J., *Identification of a Flavivirus Sequence in a Marine Arthropod.* PLoS One, 2015. **10**(12): p. e0146037.
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. 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.
56. Qin, L., et al., *Insights into angiosperm evolution, floral development and chemical biosynthesis from the Aristolochia fimbriata genome.* Nat Plants, 2021.


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
P
H
10.8 Kb
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
Tamana.1-LedTum
Tamana.2-LauKoh
Tamana.4-StyCho
Tamana.3-AmpSul
X1.1-PriCau
Flavi-
Tamana-
X1
d)
segment 1 (NS5)
3.1 Kb
JMTV
Jingmen.1-Chironomus
Jingmen.2-Gerris
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
JMV
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
0.5
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
Bats (Chiroptera)
Ray-finned fish
Cartilaginous fish
Crustaceans
Insects
Nematodes
Molluscs
Priapulids
Cnidarians
Flowering plants
d)
Flavi-like
0.5
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
MBTV2
JEV
Alfuy
Murray Valley
West Nile
St Louis
Ntaya
Bagaza
Tembusu
Spondweni
Zika
Donggang
Chaoyang
MBTV1
Meaban
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
e)
Jingmenvirus
0.5
Alongshan
JMTV
Tick
Psocopteran
Arachnidan
Guaico
Wuhan Aphid
Neuropteran
Siphonapteran
Hemipteran
Changjiang
Jingmenvirus
X2.3-Eurytemora
X2.1-AusLim
X2.2-StyCho
X2
f)
Tamanavirus
0.5
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


a)
mammals
flowering plants
terrestrial vertebrates
liver
vascular system
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
dsRNA
Flavivirus
Flavi-like
Flavi-Pesti
Pesti-like
Pegi-Hepaci
MYA
1000
800
600
400
200
100
0
Viruses
Viruses & EVEs
EVEs only
Segmentation
Hematophagy
Ticks
Insects
Plant host
Duplication
Orthology
Codivergence A
Codivergence B
b)
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
c)
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