nature microbiology

Analysis

https://doi.org/10.1038/s41564-025-02134-0

# Taxonomic expansion and reorganization of *Flaviviridae*

Received: 22 January 2025

Accepted: 27 August 2025

**Peter Simmonds**[1,13,15], **Anamarija Butković**[2,14], **Joe Grove**[3,14], **Richard Mayne**[1,14], **Jonathon C. O. Mifsud**[4,14], **Martin Beer**[5], **Jens Bukh**[6], **J. Felix Drexler**[7], **Amit Kapoor**[8], **Volker Lohmann**[9], **Donald B. Smith**[1], **Jack T. Stapleton**[10], **Nikos Vasilakis**[11] & **Jens H. Kuhn**[12,15]

*Flaviviridae* is a family of non-segmented positive-sense RNA viruses that includes major pathogens such as hepatitis C virus, dengue viruses and yellow fever virus. Recent large-scale metagenomic surveys have identified many RNA viruses related to members of this family, such as orthoflaviviruses and pestiviruses. These viruses diverge by having different genome lengths and configurations, and host range. Here we performed an analysis of RNA-directed RNA polymerase (RdRP) hallmark gene sequences of flaviviruses and 'flavi-like' viruses. We uncovered four divergent clades and multiple lineages that are congruent with phylogenies of their helicase genes, protein profile hidden Markov model profiles, and evolutionary relationships based on predicted RdRP protein structures. These results support their classification into three families (*Flaviviridae*, *Pestiviridae* and *Hepaciviridae*) and 12 genera in the established order *Amarillovirales*, with groupings correlating with genome properties and host range. This taxonomy provides a framework for future evolutionary studies on this important viral family.

*Flaviviridae* is a family of positive-sense RNA viruses that incorporates several major human and veterinary pathogens, including a wide range of often highly virulent arthropod-borne viruses, including yellow fever virus (YFV) and dengue viruses (original genus *Flavivirus*)[1]. The family also includes the *Pestivirus* genus comprising a number of systemic pathogens of cows, pigs and sheep, and was subsequently expanded to incorporate a third genus, *Hepacivirus*, for hepatitis C virus (HCV)[2] and its relatives in non-human primates, bats and rodents[3]. In 2012, a fourth genus, *Pegivirus*, was added for a range of apparently non-pathogenic RNA viruses infecting humans (human pegivirus [HPgV]) and a broad range of non-human primates, bats and other mammals[4], and one avian (goose) host[5]. Since then, there have been only minor changes to flavivirus classification, essentially limited to the expansion of the number of species assigned to the *Pestivirus, Pegivirus* and *Hepacivirus* genera[6] and the renaming of the genus *Flavivirus* to *Orthoflavivirus*[6].

Current members of the *Flaviviridae* have consistent genome organizations (positive-sense, non-segmented, linear, non-3′-poly-adenylated RNA 9.0–13 kb in length), protein expression strategies (synthesis of a single polyprotein with a conserved organization that is cleaved into structural proteins located at the N terminus and non-structural proteins at the C terminus) (Fig. 1a), and a primarily mammalian host range (in case of orthoflaviviruses, also arthropods)[1]. In addition to the gene encoding the RNA-directed RNA polymerase (RdRP), flaviviruses are homologous in their superfamily 2 helicase and serine protease domain sequences[2].

However, members of the *Flaviviridae* family also differ considerably, including possession of structurally distinct capsid proteins (and the apparent lack of a capsid in pegiviruses), packaging mechanisms, polyprotein translation strategy that may be 5′-cap-dependent (orthoflaviviruses) or driven through an internal ribosomal entry site (IRES) in other flaviviruses (Fig. 1a), and two disparate fusion glycoprotein systems (Fig. 1a). Indeed, other than the helicase and RdRP genes, no non-structural protein gene sequences are homologous across viruses of all four genera[6].

The more recent application of high-throughput sequencing technologies to detect and genetically characterize viruses in a much wider range of potential hosts has transformed our understanding of the diversity and abundance of viruses of this family[7]. Analyses

---

A full list of affiliations appears at the end of the paper. e-mail: peter.simmonds@utu.fi; jenshkuhn@comcast.net

particularly of arthropods and more recently fish and other aquatic life have created a case for a major expansion in viruses assigned to the established genera and revealed a diversity of genome organizations. For example, Jīngmén tick virus (JMTV)[8] and related 'jingmenviruses' have multipartite genomes with four or more separate segments[9–12]. The majority of flavi-like viruses have furthermore been discovered outside the primarily mammalian and vector host range of classified flaviviruses. A comprehensive list of currently described flavi-like viruses and their hosts is provided in Extended Data Table 1.

These discoveries beg the question of how to classify these viruses and how to organize the current order *Amarillovirales* to best reflect their evolutionary relationships[7,13]. For example, by phylogenetic analysis of RdRP domain sequences, JMTV and other segmented flavi-like viruses typically group closer to members of the *Orthoflavivirus* genus than to flaviviruses of other genera, but JMTV's putative assignment to this genus conflicts with what had been previously considered family-defining characteristics across the *Flaviviridae* family, that is, a non-segmented genome. The genomic diversity of flavi-like viruses and expanded host ranges that now encompasses multiple kingdoms of eukaryotes profoundly questions how a virus family such as the *Flaviviridae* can be best defined[7].

The future development of a robust taxonomy framework for flaviviruses, and indeed other virus families with similarly expanded genetic and genome organizational complexities, clearly requires a radically different set of criteria than those that have guided their current classification.

In this suggested update to the classification of flaviviruses, we have assigned primacy to the RdRP (hallmark) gene phylogeny predicated on the previously established principles for a genomics-based taxonomy of viruses[14], and that assignments should be based on the most evolutionarily conserved gene within virus groups[15–17]. At present, most viruses are assigned into seven realms that are divided on the basis of their separate origins and possession of shared orthologous gene(s)[18–20]. Almost all RNA viruses, including flaviviruses, have been assigned to the realm *Riboviria* on the basis of possession of evolutionarily related RdRPs that are distinct from those of cellular polymerases, substantiating their likely single origin. As the hallmark gene for ribovirians, its phylogenetic relationships therefore serve to best delineate the evolutionary history of RNA viruses. Recasting virus taxonomy through evolutionary relationships of hallmark genes[14] has furthermore enabled the development of a hierarchical higher-rank classification of RNA viruses, in which the family *Flaviviridae* has been assigned as a (single) member of the order *Amarillovirales* in the class *Flasuviricetes*, phylum *Kitrinoviricota*, and kingdom *Orthornavirae*.

This approach also acknowledges that the deeper evolution of all viruses may be punctuated by major genome reorganizations associated with changes in host range or ecologies. Examples of modular evolution in flaviviruses include multiple IRES exchanges within and between flaviviruses and viruses of the family *Picornaviridae*[21,22], and acquisition of glycoproteins associated with commitment to vertebrate hosts[23]. Collectively, these create different evolutionary histories and phylogenetic relationships between different genome regions. While these evolutionary events can be represented in the form of reticulated trees or networks, the strictly hierarchical classification of viruses required by the ICTV (and indeed of other biological taxonomies) requires primacy to be assigned to hallmark genes in the creation of coherent taxonomy frameworks.

Accordingly, we present reclassification of flaviviruses and flavi-like viruses grounded in varied analyses of RdRP hallmark gene relationships. This provides a robust, well-supported and coherent taxonomic framework that can serve as a scaffold for future expansion of this group of viruses. Moreover, our multimodal approach may provide a template for re-examining other viral taxonomies, as we seek to organize the huge viral diversity revealed by metagenomics.

## Results

To establish a new taxonomic framework for classified flaviviruses and unclassified flavi-like viruses, we generated a dataset of representative sequences from currently classified members of the family *Flaviviridae* (viruses of 52, 19, 14 and 11 species in four genera: *Orthoflavivirus*, *Pestivirus*, *Hepacivirus* and *Pegivirus*, respectively[6]) and a comprehensive sample of coding-complete sequences of currently unclassified flavi-like viruses[23] (listed in Supplementary Table 5). This dataset includes representatives from groups of segmented flavi-like viruses, divergent 'pesti-like' viruses, many with extended genomes, many additional 'hepaci-like' and 'pegi-like' viruses primarily found in marine vertebrates, plant-infecting 'koshoviruses' and a selection of flavi-like viruses recovered from environmental samples, such as diatom colony-associated virus (DCAV). Genome regions encoding the enzymatic domains of the RdRP and helicase genes were extracted for alignment and analysis.

### RdRP phylogeny splits flavi-/flavi-like viruses into 4 clades

Our initial analysis focused on translated flavivirus and flavi-like'RdRP domain sequences. Using the distantly related viruses of family *Tombusviridae* as an outgroup, the RdRP domain phylogeny of our sequence set, calculated with IQ-TREE[24], resulted in four main bootstrap-supported clades (I–IV), each comprising a number of bootstrap-supported lineages (labelled as Ia–l, IIm–s, IIIt–w anticlockwise around the tree) (Fig. 1b). Viruses of the four currently established flavivirus genera were distributed in three of the four clades, with hepaciviruses and pegiviruses clustering together in and largely defining clade III. In this phylogeny, jingmenviruses, 'tamanaviruses' and numerous 'insect-specific flaviviruses' (ISFs) cluster with orthoflaviviruses in clade I. The viruses that were previously loosely defined as 'large-genome flaviruses' (LGFs) join pestiviruses in clade II and include diverse viruses distributed across multiple lineages (for example, Bólè tick virus, a potentially tick-vectored pathogen of mammals)[25] in clade IIq, and plant-infecting koshoviruses[10,26,27] in clade IIs (Fig. 1b and Supplementary Table 1).

To investigate the robustness of the clade and lineage groupings, the IQ-TREE-based maximum-likelihood (ML) phylogeny generated for Fig. 1b was compared with trees generated through a temporal reconstruction using the Bayesian evolutionary analysis by sampling trees cross-platform programme (BEAST; https://github.com/beast-dev/beast-mcmc) and, in parallel, the protein distance-based unweighted pair group method with arithmetic mean (UPGMA) phylogeny method as implemented in MEGA7.0 (ref. 20) (Extended Data Fig. 1). BEAST and UPGMA results reproduced clades I–IV with bootstrap support comparable to the original ML analysis (Fig. 1b and Extended Data Fig. 1a). However, in the time-rooted BEAST tree, *Tombusviridae* became an inlier (Extended Data Fig. 1c), in marked contrast to its clearly defined outgroup positions in the IQ-TREE and UPGMA trees. The grouping of individual sequences within lineages a–w defined by IQ-TREE analysis was almost entirely reproduced in the BEAST and UPGMA trees, except for eight sequences that did not group with their lineages in the UPGMA tree (Extended Data Fig. 1b). Despite these, the clade and lineage assignments were relatively robust to different evolutionary reconstruction methods.

These RdRP phylogenies provide an excellent framework for the taxonomic reorganization of the *Flaviviridae*. Nonetheless, we recognized that phylogenetic inference over these very large genetic distances is challenging and therefore sought to corroborate (or, indeed, refute) these apparent taxonomic groupings with alternative and complementary approaches.

### Congruence of RdRP and helicase domain phylogenetic trees

There is evidence for modular exchange of structural or assessor proteins among flaviviruses of the four established current genera (for instance, acquisition of glycoprotein-encoding genes)[23], but it is less

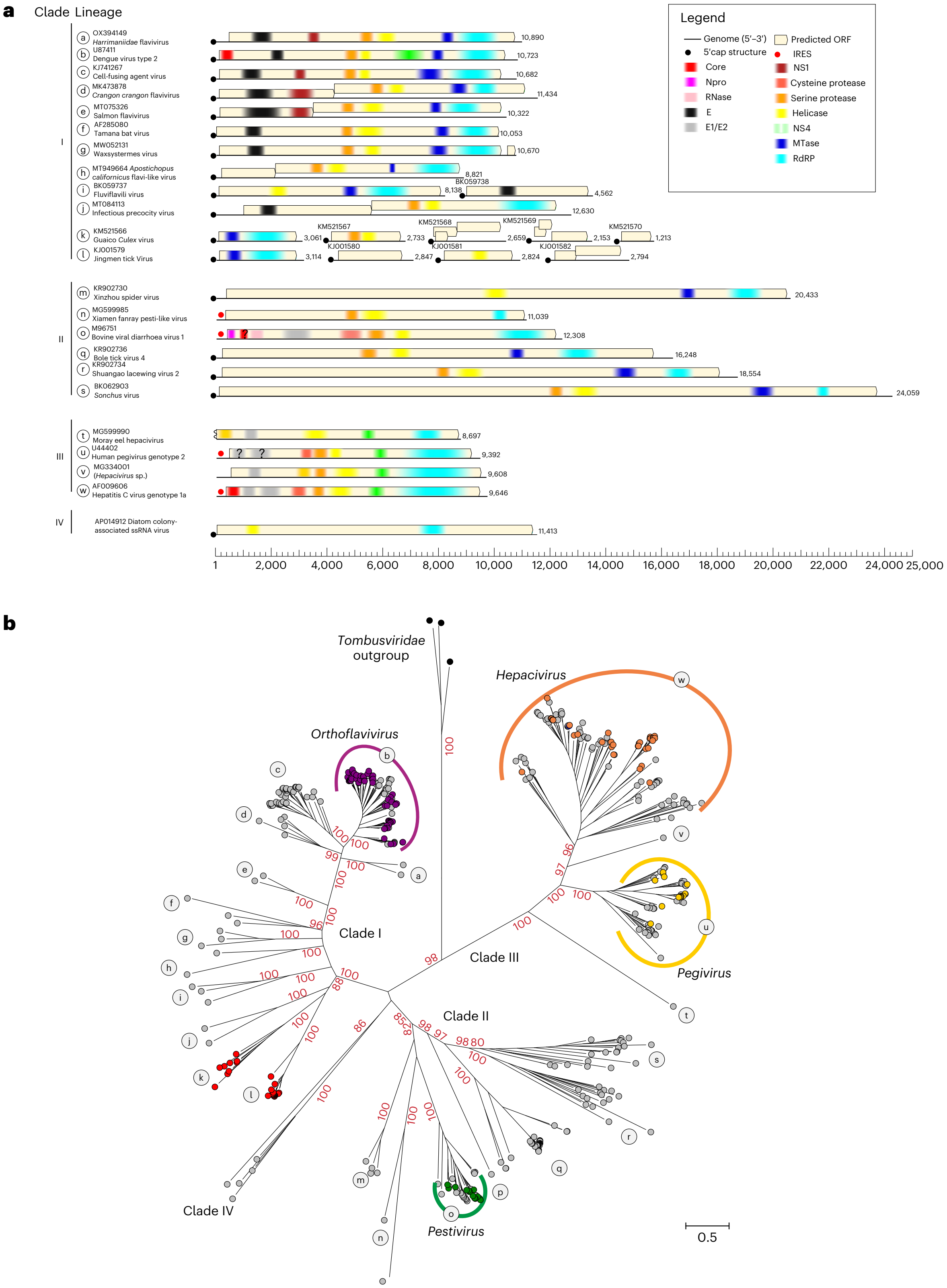


**Fig. 1 | Genome organizational and genetic diversity of flaviviruses and flavi-like viruses. a**, Organization of example genomes in each lineage of flaviviruses and flavi-like viruses. Genome diagrams for the example viruses listed in Table 1 drawn to scale (bottom scale bar) and main functional domains identified by InterProScan browser v.103 (https://www.ebi.ac.uk/interpro/search/sequence/). **b**, RdRP amino-acid sequence phylogenies differentiate flaviviruses and flavi-like viruses into four highly supported main clades: I, II, III and IV, and lineages (a–w) within these clades. Maximum-likelihood tree of aligned flavivirus and flavi-like RdRP domain amino-acid sequences, estimated using the LG + F + R10 model using IQ-TREE[24] with tombusvirus sequences as an outgroup. Bootstrap support values for the main branches of the tree are shown in red if ≥70%. Already classified flaviviruses are shown in colour-filled circles. The main clades are numbered I–IV and lineages are labelled with lowercase letters. The alignment and component sequences within each clade are provided in a fully annotated tree (Supplementary Data 1 and 2 and Supplementary Table 5).

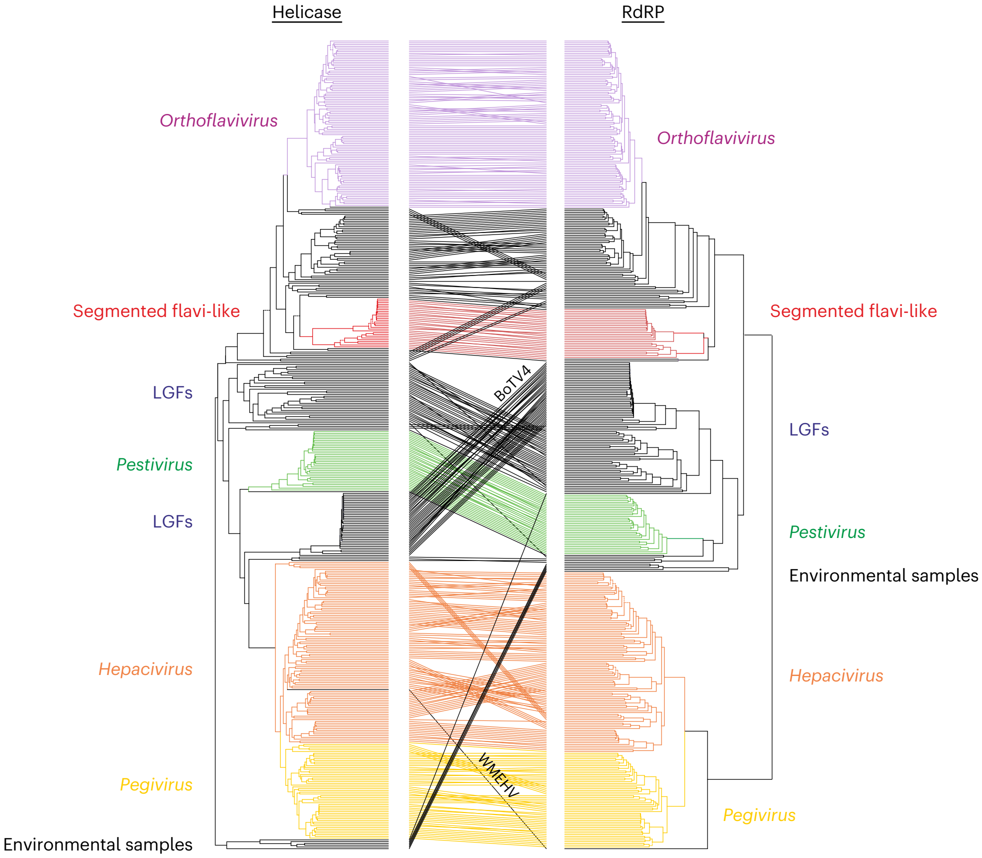


**Fig. 2 | Helicase domain (NS3) amino-acid sequence phylogeny supports partitioning of flaviviruses and flavi-like viruses into four main clades.** Tanglegrams of flavivirus and flavi-like virus helicase and RdRP domain sequences constructed from ML phylogenetic trees generated by IQ-TREE[24] trees, with established flavivirus genera coloured as in Fig. 1b, and segmented flavi-like viruses (lineages k and l in red). A copy of the figure with the branches individually labelled is provided as Supplementary Fig. 1. The alignment of helicase sequences is provided in Supplementary Data 3.

clear whether sequences encoding the core replication module of these viruses (including serine protease, helicase and RdRP) evolve as a unit or are similarly subject to genome exchange and rearrangements. Compared to the serine protease, the helicase domain was clearly identified in all the flavi- and flavi-like viruses, which made it a good candidate to complement our RdRP analyses. We thus deduced flavivirus and flavi-like helicase domain amino-acid sequences, aligned them, performed IQ-TREE[24] analysis and compared the result to the RdRP IQ-TREE tree using a tanglegram (Fig. 2). Phylogenetic groupings of viruses of the four current genera were highly concordant, with only minor differences in branching order within genera. The positions of flavi-like viruses grouping with orthoflaviviruses were generally concordant among regions, but with some exceptions. For instance, there was a change in the topology of the deeper branches underlying the LGF, pestivirus and hepaci-/pegivirus groupings, creating paraphyletic groups not observed in the RdRP tree. This result is not necessarily surprising because the helicase domain is shorter and more divergent than the RdRP domain, and hence probably reflects lower resolution of relationships rather than indicating genome reorganization between the two domains. Other minor exceptions included Wēnlǐng moray eel hepacivirus (WMEHV) moving from an outlier position in clade III (lineage IIIt) into *Hepacivirus* in the helicase tree, a finding not incompatible with potential sequence errors in the deposited RdRP region sequence. Bólè tick virus 4 (BoTV4) fell within the IIq lineage in the RdRP domain tree, but as an outlier to pestiviruses in the helicase domain tree. Finally, flavi-like viruses from environmental samples (clade IV in the RdRP tree) were located within the pestivirus–LGF branches in the helicase domain tree.

## RdRP protein structure relationships match sequence phylogenies

The function of any given protein is primarily a feature of its three-dimensional structure. Consequently, protein structure is fundamentally more conserved than the underlying protein sequence. The advent of accurate protein structure prediction through machine learning (for example, AlphaFold2) is enabling surveys of protein form and function at enormous scales[28,29] and has driven the development of new high-throughput structure comparisons tools (for example, Foldseek[30]) that enable structure-guided inference of deep evolutionary relationships[30,31].

In a recent investigation of glycoproteins, we systematically applied protein structure prediction to viruses of *Flaviviridae*[23], generating thousands of structures spanning the complete polyproteins of all viruses represented in the RdRP phylogeny (Supplementary Table 5). Drawing on this dataset, we generated complete NS5/NS5B RdRP domain structure predictions for each virus using ColabFold[32]. After filtering for prediction confidence and length, we analysed 400 RdRP structures using FoldTree[31] to produce a structure-guided tree (Fig. 3a) based on the local distance difference test (lddt) structural similarity metric (examples of comparisons are shown in Fig. 3b). As a control, experimentally determined RdRP structures from orthoflavi-, jingmen-, pesti- and hepaciviruses, taken from the protein databank, were included in the analysis. This structure-based approach does not explicitly consider protein sequence and, therefore, represents an independent recapitulation of the sequence-based phylogeny (Fig. 1b).

The topology of the structure-based tree (Fig. 3a) was remarkably similar to that of the sequence-based tree (Fig. 1b) and supports the

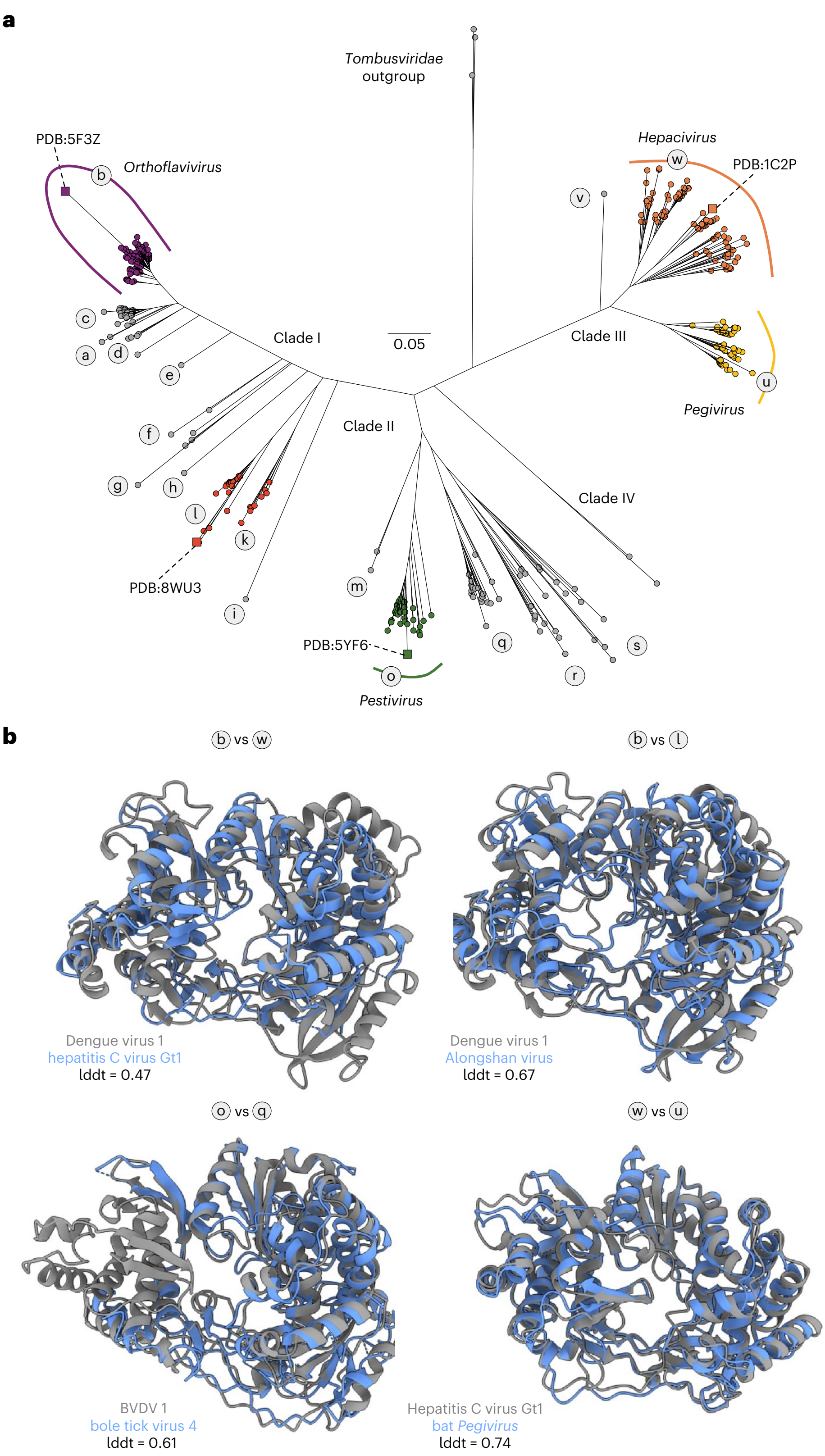


**Fig. 3 | RdRP domain (NS5/NS5B) structural comparison supports partitioning of flaviviruses and flavi-like viruses into four main clades. a**, Structure-based tree of 400 flavivirus and flavi-like virus' RdRP domains, derived from an lddt distance matrix (calculated using FoldTree[31], powered by Foldseek[30]; scale bar indicates lddt distance, which is approximate to the inverse of the pairwise lddt score). Main clades and lineages are labelled as in Fig. 1b. Experimentally determined structure references were included in the analyses and are shown as square tips with labels: PDB: 5F3Z, dengue virus type 3; PDB: 8WU3 Ālóngshān virus; PDB: 5YF6 classical swine fever virus; PDB: 1C2P, hepatitis C virus genotype 1b. **b**, Examples of aligned RdRP domain structures, colour-coded as stated in the label; lineage identifiers (for example, b versus w) indicate the position of the compared structures on the tree. lddt values represent structural similarity, with values of 1 being perfectly aligned identical structures. BVDV, bovine viral diarrhoea virus.

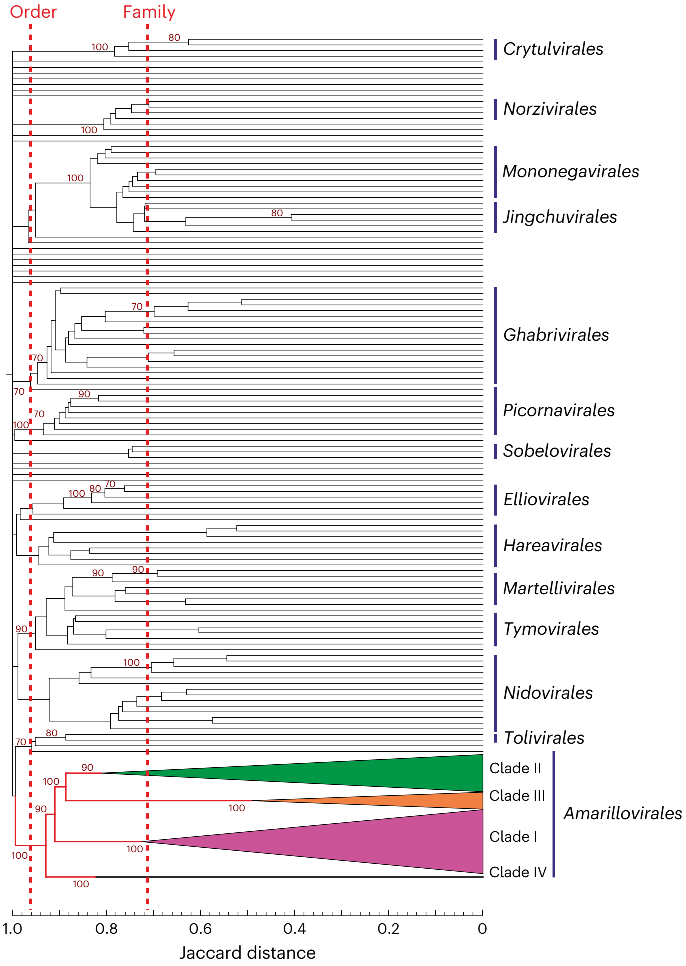


**Fig. 4 | Alignment-free hidden Markov model homology analysis supports partitioning of flaviviruses and flavi-like viruses into one order and four family rank clades.** GRAViTy[35] Jaccard distances calculated for classified flaviviruses and flavi-like viruses and a representative member of each established RNA virus family in ribovirian kingdom Orthornavirae ($n = 135$), showing approximate demarcation thresholds for orders and families (dashed vertical red lines). Bootstrap support values (10 iterations) ≥70% are shown in red. A copy of the figure with the branches individually labelled is provided as Supplementary Fig. 2.

same four major clades. Moreover, most of the lineages represented in the sequence phylogeny were consistent in their position and composition. The only exceptions are lineage Ia, which in the structure-based tree formed a basal branch from lineage Ic, and lineages Ii and IIIv which moved subtly in relation to their neighbouring clades. Clades Ij, IIn and IIIt were lost from the structure-based tree due to filtering of lower-confidence structural predictions. Major clade IV for flavi-like viruses from environmental samples also shifted position slightly, branching between clade II and the *Tombusviridae* outgroup in the structure-based tree. However, this divergent and basal taxon will probably remain difficult to place by any method without further discovery of similar viruses. Reassuringly, experimentally determined structure references were placed within the expected clades (for example, PDB: 1C2P HCV genotype 1b NS5B within the hepacivirus clade); this suggests that the predicted structures faithfully recapitulate the structural diversity of RdRP.

Example structural alignments of clade representatives corroborate their distribution on the tree (Fig. 3b). Dengue virus type 1 and HCV genotype 1 were at opposite ends of the tree (clade I and III) and their respective RdRPs align with a relatively low lddt score (0.47), whereas pairings within clade I, II and III gave higher scores: 0.67, 0.61 and 0.74, respectively (note that an lddt score of 1.0 represents perfect alignment of identical structures).

ColabFold is an AlphaFold2-based approach and, therefore, infers protein structure using co-evolutionary patterns extracted from multiple sequence alignments (MSAs). Consequently, it may be argued that our structure-only phylogeny (Fig. 3a) is not truly independent of sequence similarity. Therefore, we repeated our analysis using a protein language model (pLM)-based structure prediction method (Chai-1 run using ESM-2 embeddings[33,34]), which infers structure purely from primary sequence and without MSAs. This method results in lower prediction confidence (as measured by the pLDDT metric; Supplementary Fig. 2a); despite this, the resultant structures are in excellent agreement with reference structures (Supplementary Fig. 2b). Generating a structure-based tree using pLM-based predictions preserves all the major clades and groupings found in the ColabFold tree, with consistent placement of experimentally determined structure references (Supplementary Fig. 2c). However, lower prediction confidence appears to increase branch lengths and creates some topological ambiguity (for example, subclade Ia being nested within Ib). Nonetheless, inferring evolutionary relatedness through structure-only analysis (using both MSA- and pLM-based structure prediction methods) corroborates sequence-based approaches and is highly supportive of the organization of flaviviruses and flavi-like viruses into four main clades.

## Alignment-free hidden Markov model homology analysis

Genome Relationships Applied to Virus Taxonomy (GRAViTy) is a non-supervised, alignment-free method to assess the relatedness of virus genome sequences through calculation of protein profile hidden Markov model (PPHMM) homologies and through metrics of genome organization such as the order and orientation of genes[35]. We performed GRAViTy analysis using our flavivirus and flavi-like virus dataset (Supplementary Data 4) for phylogeny and RdRP structure comparisons; genomes of segmented jingmenviruses were concatenated in order of segment length from short to long. Results were remarkably concordant with those determined by RdRP and helicase domain phylogenies, with bootstrap-supported segregation of the same sequences into three main clades I–III (Extended Data Fig. 2). Flavi-like viruses from environmental samples formed an outlier position as clade IV.

To depict the wider inter-relationships of classified flaviviruses and flavi-like viruses, we compared their sequences with sequences representing viruses of all other established ribovirian families. Flaviviruses and flavi-like viruses were monophyletic in the GRAViTy dendrogram, supporting their assignment to a common higher taxonomic rank, the established order *Amarillovirales* (Fig. 4). The Jaccard distances calculated by GRAViTy do not provide a precise quantitative estimate of evolutionary distance or thresholds for taxonomic assignments. However, ribovirian families differ from each other in the distance range of 0.7–0.85, whereas viruses within families typically are associated with Jaccard distances of ~0.9 (Fig. 4). Although very general, the distances between clades I, II, III and IV were within the range of between-family distances elsewhere in the dendrogram, whereas their combined grouping occurs at the level of order assignments for other virus families assigned to orders (for example, *Nidovirales* and *Picornavirales*).

## Taxonomic associations with genome properties and host range

Within clade I, lineage Ib comprises the currently classified members of the *Orthoflavirus* genus, as well as a high number of currently unclassified primarily ISFs. Lineage Ic is similarly populated by ISFs, including the current unclassified cell-fusing agent virus (CFAV), consistent with a previous suggested assignment to a new genus within the *Flaviviridae*[36]. The similarly divergent Tamana bat virus (TABV) falls in the very diverse lineage If, which also includes flavi-like viruses infecting lumpfish (*Cyclopterus lumpus*) and pygmy squid (*Xipholeptos notoides*). The

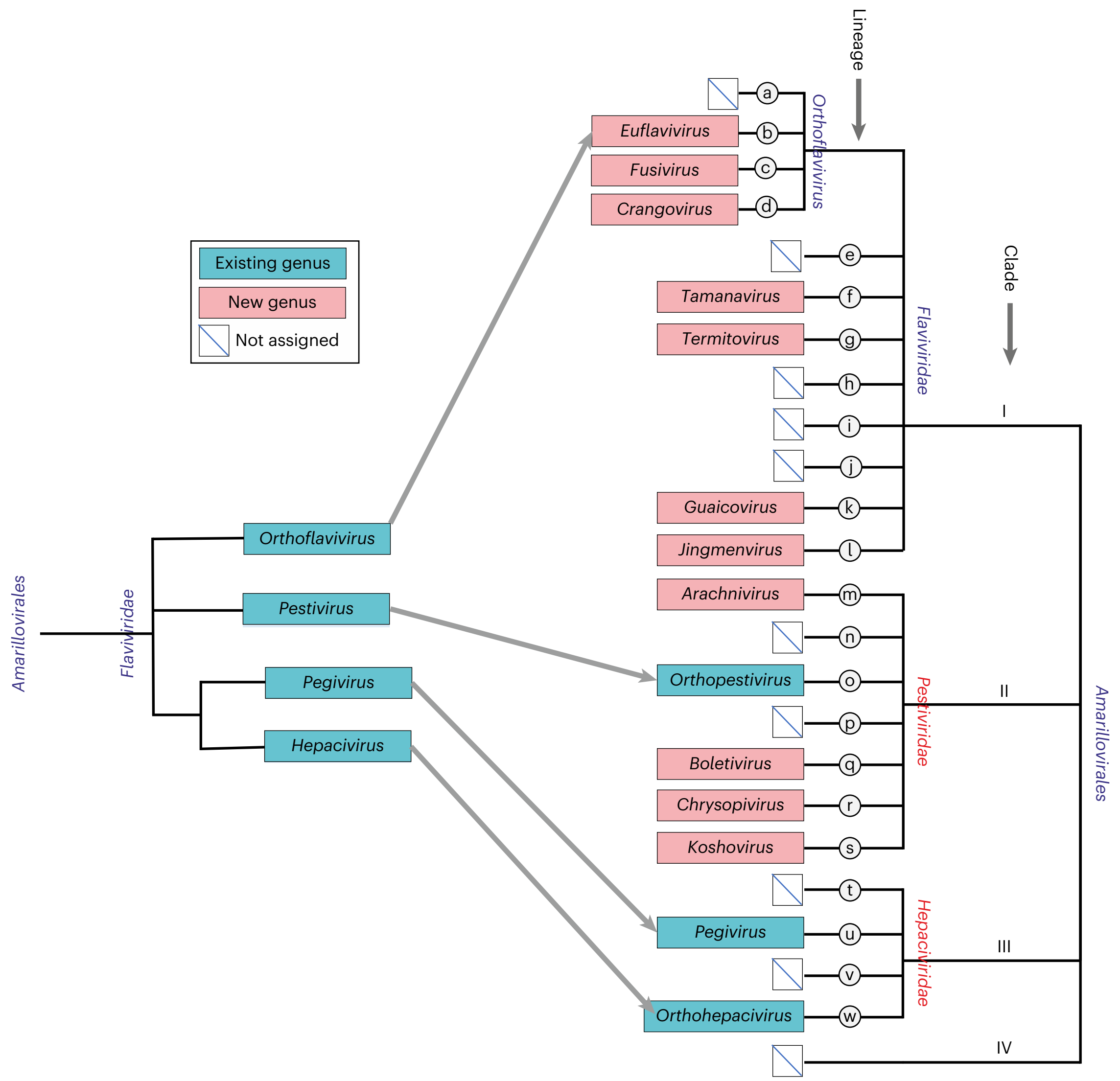


**Fig. 5 | Diagrammatic summary of suggested changes to flavivirus taxonomy.** Relationships between genus and higher-rank taxa in the current classification of *Flaviviridae* (left side, blue boxes) with those of the suggested expanded classification of flaviviruses (right side); suggested new taxa are shown in red.

segmented JMTV and Guaico *Culex* virus (GCuV) and their relatives have been assigned to lineages Ik and Il, adjacent to lineage Ij consisting of the non-segmented infectious precocity virus. Extending the host range of this clade were Cnidaria flavivirus and Harrimaniidae flavivirus infecting basal metazoans such as jellies and acorn worms[37] in lineage Ia. Despite the evident lineage (and host) diversity, clade I is clearly monophyletic with a bootstrap-supported long branch separating members from other clades.

Clade II similarly subdivides into several bootstrap-supported lineages IIm–IIs, with currently classified pestiviruses clustering exclusively in lineage IIo along with currently unassigned more divergent viruses exclusively infecting vertebrates, as does lineage IIn (in fish). Viruses assigned to other lineages within clade II are arthropod hosted, with the striking exception of the plant-infecting Apis flavivirus, carrot flavi-like virus 1, Gentian Kobu-sho-associated virus, Coptis virus 1 and Sonchus virus 1 (SONV1) that form a subgroup, termed koshoviruses, within lineage IIs[26,27,38]. Arthropod and plant-infecting members of clade II, such as Bólè tick virus 4 (BoTV4) and SONV1, frequently showed substantially longer genomes than the vertebrate-infecting members of lineages IIn and IIo (ranges 11,038–11,450 and 11,555–15,154 nucleotides [nt], respectively) compared with lineage IIm (20,432–22,622 nt), IIq (13,599–18,696 nt), IIr (14,731–26,314 nt) and IIs (18,749–27,708 nt).

Members of clade III were all non-segmented, possessed similar genome lengths (8,294–12,290 nt) and formed a well-defined separate grouping from other flaviviruses. In contrast to the wide host range of clade I and II, all clade III members had presumed or demonstrated vertebrate hosts, spanning a wide range of mammals, birds, reptiles, and bony and cartilaginous fish. Two lineages, IIIu and IIIw, comprised currently classified members of the *Pegivirus* and *Hepacivirus* genera along with a range of more divergent viruses with a greater host range beyond mammals and birds.

### A suggested reclassification of flavi-/flavi-like viruses

The reproducible phylogenetic relationships between flaviviruses and flavi-like viruses in the RdRP domain sequences using different tree construction methods (Fig. 1b and Extended Data Fig. 1) were recapitulated by protein structure-guided analysis (Fig. 3). The RdRP gene is evidently co-evolving with the helicase gene (Fig. 2), suggesting that the essential replicase of flaviviruses and flavi-like viruses traces a single coherent evolutionary history. This is also reflected in the relationships identified by an alignment-free method for analysing whole genome sequences (Extended Data Fig. 2). Therefore, we are confident that the foundational RdRP phylogeny in Fig. 1a provides a robust framework for a genomics-based reclassification of flaviviruses.

**Table 1 | Listing of clades and lineages of flavivirus and flavi-like viruses and suggested revised taxonomy**

| Clade | (Suggested) family[a] | Lineage | n[b] | (Suggested) genus[a,c] | Suggested subgenus | INSDC accession number | Example virus |
|---|---|---|---|---|---|---|---|
| I | *Flaviviridae* | ⓐ | 2 | *Orthoflavivirus* | | OX394137 | Cnidaria flavivirus (CnFV) |
| | | ⓑ | 95 | | ***Euflavivirus*** | U87411 | Dengue virus type 2 (DENV-2) |
| | | ⓒ | 39 | | ***Fusivirus*** | KJ741267 | Cell-fusing agent virus (CFAV) |
| | | ⓓ | 5 | | ***Crangovirus*** | MK473878 | Crangon crangon flavivirus (CCFV) |
| | | ⓔ | 2 | | | MT075326 | Salmon flavivirus (SaFV) |
| | | ⓕ | 3 | ***Tamanavirus*** | | AF285080 | Tamana bat virus (TABV) |
| | | ⓖ | 3 | ***Termitovirus*** | | MW052131 | waxsystermes virus (WXTV) |
| | | ⓗ | 2 | | | OX394159 | Chowder Bay tunicate associated flavi-like virus (CbtuFV) |
| | | ⓘ | 3 | | | BK059737, BK059738 | Fluviflavili virus |
| | | ⓙ | 2 | | | MT084113 | Infectious precocity virus (IPV) |
| | | ⓚ | 9 | ***Guaicovirus*** | | KM521566–KM521570 | Guaico Culex virus (GCuV) |
| | | ⓛ | 19 | ***Jingmenvirus*** | | KJ001579–KJ001582 | Jīngmén tick virus (JMTV) |
| II | ***Pestiviridae*** | ⓜ | 3 | ***Arachnivirus*** | | KR902730 | Xīnzhōu spider virus (XSV) |
| | | ⓝ | 2 | | | MG599985 | Xiàmén fanray pesti-like virus (XFPV) |
| | | ⓞ | 34 | ***Orthopestivirus*** | | M96751 | Bovine viral diarrhoea virus 1 (BVDV1) |
| | | ⓟ | 1 | | | SRR7976360 | |
| | | ⓠ | 39 | ***Boletivirus*** | | KR902736 | Bólè tick virus 4 (BoTV4) |
| | | ⓡ | 15 | ***Chrysopivirus*** | | KR902734 | Shuāngào lacewing virus 2 (SgLWV2) |
| | | ⓢ | 20 | ***Koshovirus*** | | BK062903 | Sonchus virus 1 (SONV1) |
| III | ***Hepaciviridae*** | ⓣ | 1 | | | MG599990 | Wēnlǐng moray eel hepacivirus |
| | | ⓤ | 54 | *Pegivirus* | | U44402 | Human pegivirus (HPgV) genotype 2 |
| | | ⓥ | 1 | | | MG334001 | (Hepacivirus sp.) |
| | | ⓦ | 101 | ***Orthohepacivirus*** | | AF009606 | Hepatitis C virus (HCV) genotype 1a |
| IV | (unclassified) | | | | | AP014912 | Diatom colony-associated ssRNA virus (DCAV) |

[a] Suggested new and renamed taxa are printed in bold font. [b] Available distinct coding-complete genome sequences; examples representing species range within the larger groups. [c] A genus was suggested when the lineage was represented by at least three or more examples (excluding SRR library-annotated sequences) and with viruses that consistently grouped together by different phylogenetic methods and showed some commonality in genome organization.

These analyses based on clade and lineage relationships provide a framework that maps clades and lineages onto families and genera. However, there are quite variable branch lengths within lineages (Fig. 1b) and differences in thresholds that split lineages in BEAST and UPGMA trees across clades I, II and III (Extended Data Fig. 1). These observations indicate that a purely cladistic classification may not conform to specific sequence divergence thresholds that are often used elsewhere in virus taxonomy. Indeed, formal comparison of mean amino-acid sequence identities between and within lineages in the three clades showed considerable variation (Extended Data Fig. 3), particularly in clade I for which sequence identities across lineages Ia, Ib, Ic and Id were all much greater than within-group values of other lineages (notably If, Ih, Ii and Ij, and lineage n in clade II). For classification purposes, it might be considered that the four lineages 1a, 1b, 1c and 1d were assigned as a group equivalent to those formed in other lineages; this would create a higher threshold in the UPGMA and BEAST trees (blue dotted line in Extended Data Fig. 1b,c) that is more consistent with between-lineage thresholds in clades II and III.

With this caveat, we can make some tentative suggestions for the reclassification of flaviviruses that accommodates the large number of additional flavi-like viruses described since the last classification of the family[1] (Table 1). The following findings were considered in the design of a new flavivirus taxonomy:

A. Flavivirus and flavi-like viruses form a monophyletic group in comparison with all other members of the *Riboviria*, enabling their assignment to a single taxonomic rank. Although not precise, divergence among flaviviruses at this rank was comparable to that of members of virus orders in GRAViTy[35] analysis (Fig. 4). We therefore posit that all flavivirus and flavi-like viruses can be assigned to the established order *Amarillovirales*.
B. The level of sequence divergence among members of clades I–IV was comparable to interfamily distances of other RNA viruses on GRAViTy[35] analysis. We therefore suggest the assignment of members of lineages I, II and III into three new families. We retain the name *Flaviviridae* for lineage I and suggest the names '*Pestiviridae*' and '*Hepaciviridae*' for lineages II and III, respectively, reflecting the historical key virus members of these families.
C. While a separate, bootstrap-supported lineage IV was consistently observed, its members are highly divergent genetically and derive from environmental samples. Where hosts are suspected, these are extremely diverse and require further verification. We therefore do not suggest creating a family for lineage IV at this time.
D. All three suggested families comprise several bootstrap-supported clades of sequences, often with distinct genome organizations, lengths and host tropisms. We suggest that the clade assignments can be used for genus demarcation in specific circumstances:
a. Clades Ib, IIo, IIIu and IIIw comprise currently classified flaviviruses and should be assigned in the revised taxonomy. We suggest that the current genus names *Orthoflavivirus*, *Pegivirus* and *Hepacivirus* be retained, with *Pestivirus* modified to '*Orthopestivirus*'

(because of the suggested *Pestiviridae* family name), and *Hepacivirus* modified to '*Orthohepacivirus*' (because of the suggested *Hepaciviridae* family name). As much as possible, this maintains continuity with established nomenclature while acknowledging the greatly expanded diversity of viruses assigned to each.

b. Criteria for designating further genera include:

i. Bootstrap-supported groupings consistent with multiple methods
ii. A minimum of three members so that the extent of grouping can be assessed, and
iii. Priority should be given to clades including previously described and well-characterized viruses, such as CFAV and TABV.

c. We further note that the group comprising lineages Ia, Ib, Ic and Id within clade I show greater comparability with genetic divergence thresholds between genera elsewhere. The name *Orthoflavivirus* could then be applied to this group (with an expanded membership), while its component lineages could be assigned as the subgenera '*Euflavivirus*' (true flaviviruses), '*Crangovirus*' and '*Fusivirus*'. Retaining the name *Orthoflavivirus* for the genus avoids the need to rename the 49 currently assigned species within the current genus.
d. On this basis, we suggest the creation and naming of a total of 8 additional genera and four subgenera in addition to three genera that remain as previously assigned (Table 1).

A diagrammatic summary of the suggested changes is provided in Fig. 5.

## Discussion

The RdRP domain is considered the 'hallmark gene' for the classification of RNA viruses, primarily because comparison of RdRP sequences provided evidence for their monophyletic origin separate from all known cellular polymerases[18,39]. Using the RdRP phylogeny as the reference point, the evolutionary history of flaviviruses and flavi-like viruses has evidently been punctuated by multiple genome reorganizations, expansions, exchange of structural sequence modules and genome segmentation[7,13,23]. The continuous discovery of flavi-like viruses in highly divergent hosts means that it is no longer possible to use the criteria of genome length, coding strategy and host range that had been used for taxonomic placement of currently classified flaviviruses.

The requirement of all biological taxonomies for a hierarchical classification necessitates selection of a common marker gene present in all clades. Thus, the confirmed grouping of already classified flaviviruses and flavi-like viruses into established order *Amarillovirales* and their suggested assignment to 3 families and 12 genera (Table 1) is based on the phylogeny of RdRP. Other genome regions exhibit much more horizontal gene transfer, resulting in distinct evolutionary histories that may even originate from cellular life (for example, E$^{rns}$, expressed by pestiviruses and LGFs, is of bacterial origin[23]); clearly, organization based on these genomic regions would confound taxonomic classification.

We observed a primary division into four clades, all of which are bootstrap supported by each of the methods we used. The established genus assignments were replicated in these analyses, but there is a much closer relationship between hepaciviruses and pegiviruses than between the other currently established genera. Indeed, hepaciviruses and pegiviruses cluster in one main clade (III), LGFs cluster with pestiviruses in another (II), whereas jingmenviruses, orthoflaviviruses and tamanaviruses cluster in another (I). There is no pre-defined level or evolutionary distance range in the RdRP phylogeny (or hallmark genes elsewhere in the ICTV taxonomy) that dictates family rank assignments. However, a threshold Jaccard distance level of 0.8–0.85 drawn through the amarilloviral grouping reproduces family rank assignments elsewhere in realm *Riboviria* (Fig. 4).

Consequently, splitting the current family *Flaviviridae* into three families is consistent with degrees of relatedness between and within other classified RdRP-encoding RNA virus families. This split results in the removal of genera *Hepacivirus*, *Pegivirus* and *Pestivirus* from family *Flaviviridae*, which would be restricted to genus *Orthoflavivirus* and novel genera for flavi-like viruses.

The suggested family *Pestiviridae* would absorb genus *Pestivirus* but also include the multiple and highly divergent clades of LGFs. Mammalian pestiviruses possess a type IV IRES[40,41] and encode major envelope proteins that are structurally homologous to those of hepaciviruses and pegiviruses[23]. However, unclassified pesti-like viruses of spiders (lineage IIm) may use cap-dependent translation and encode envelope proteins structurally unrelated to those of the originally assigned mammalian pestiviruses[23]. LGFs are far more diverse genetically in the RdRP sequence than viruses in the other suggested families. However, lowering the family rank assignment threshold would create up to five or more new families of pesti-like viruses, a step we consider inappropriate at least until these viruses are better characterized.

The remainder of flaviviruses and flavi-like viruses group in a deep and phylogenetically well-defined third family, *Hepaciviridae*. The depth of grouping provides no support for the formation of separate families for hepaciviruses and pegiviruses. The addition of recently described hepaci-like or pegi-like viruses infecting fish has greatly expanded the genetic diversity of both groups and blurs the originally clear distinction between them. Generally, however, the apparent absence of a capsid-encoding sequence in pegiviruses, indicative of a likely radically different virion structure (or conceivably undetected segmentation of the pegivirus genome), differentiates pegiviruses from hepaciviruses at least for now.

These analyses provide the basis for assignments of new genera within each of the three new families (Table 1), where they are supported into genetically well-defined groups and have some commonality in phenotypic properties, such as host range, and genome organization. The suggested additional eight genera and three subgenera, represent obvious candidates for classification by these criteria. However, additional assignments may be made in the future pending collection of further characterized flavi-like viruses in ongoing clinical, veterinary and entomology screens and in metagenomic data from the widening range of invertebrates.

Overall, we believe we have made a robust case for an evolutionarily based reclassification of flaviviruses using an approach that puts primacy on genetic relationships of the hallmark RdRP gene. The seeming propensity of flaviviruses and potentially other RNA viruses to undergo radical changes in genome organization (for example, segmentation, changes to translation mechanisms and exchanges of structural gene modules) indeed reinforces the need to base classification and inference on evolutionary histories of the most stable elements within the genome. The use of protein structure relationships of RdRP and potentially other replication-associated enzymes provides an exciting new approach to determine deeper evolutionary histories of RNA viruses beyond the ranks of family and order analysed in the current study.

## Methods

### Genome sequences

A comprehensive set of coding-complete genome sequences representing the 97 currently established flavivirus species[6], supplemented with flavi-like viruses analysed previously[23], was used as the basic set for analysis and putative taxon assignments (listed in Supplementary Table 5). Genome regions encoding the RdRP and helicase domains were extracted for amino-acid sequence deduction, alignment and analysis.

### Sequence alignment and phylogenetic analysis

Alignments were created from highly conserved genome regions of each virus encoding the RdRP. This region was extracted from each sequence by aligning polyprotein sequence subsets according to their taxonomy and using both pre-existing and newly generated NS5b annotations obtained from InterProScan (v.5.56-89.0) using SuperFamily (v.1.75), CDD (v.3.18) and Pfams (v.34.0) databases as a guide. Ambiguously aligned regions, including large insertions present in the RdRPs of some large-genome flaviviruses, were removed using trimAl (v.1.4.1)[42] with multiple methods and conservation thresholds as previously described[23]. The final alignment was obtained using MUSCLE (v.5.1) with a trimAl consensus and gap threshold value of 5 and 0.9, respectively (final alignment provided in Supplementary Data 1). The phylogeny of sequences in the final alignment was reconstructed using the ML-based IQ-TREE programme (v.1.6.12)[24] with an empirically determined optimal model, Lascuel + F + 10 rate categories (LG + F + R10) selected on the basis of the minimum Bayesian information criteria (BIC) score[43]. Robustness of branching was estimated by bootstrap resampling (1,000 replicates)[44]. An unrooted tree with bootstrap support values shown for the main clades and lineages was plotted using MEGA7.0 (ref. 45) (Supplementary Data 2).

The RdRP domain amino-acid sequence dataset was analysed in parallel by the distance-based UPGMA phylogeny method as implemented in MEGA7.0 (ref. 45), using Jones–Taylor–Thornton (JTT) matrix protein distances and 100 bootstrap replicates.

A temporal reconstruction of amarilloviral evolution was performed with BEAST v.10.05 (https://github.com/beast-dev/beast-mcmc) using dated sequences based on sample date (or International Nucleotide Sequence Database Collaboration [INSDC] submission date if this information was not annotated; $n = 143$), with uniform rate, constant population size and BLOSUM protein distances as priors.

A maximum-likelihood phylogenetic tree of alignment helicase-encoding sequences containing the 13 conserved motifs (Q-VI) specific for superfamily 2 (Supplementary Data 3) was similarly generated using IQ-TREE[24], with the (lowest BIC[43]) substitution model. The alignment was obtained with MAFFT[46], the presence of the 13 conserved motifs visually confirmed, and the alignment trimmed using TrimAl[42] with the gappyout option. RdRP and helicase domain trees were compared using Tanglegram[47,48]. Potyvirus, poxvirus and DEAH helicase family sequences were used to root the tree.

### GRAViTy

Genome relatedness of flaviviruses and flavi-like viruses to all currently classified ribovirians was determined using the GRAViTy version 2 implementation[35] of the original algorithm[49] (run parameters provided in Supplementary Data 4): analysis was performed in a single step, using the new classification function. Default parameters were used, except for initial translated open reading frame sequence clustering inflation (6.0), and use of the translated open reading frame alignment method (G-INS-I) and PPHMM similarity cut-off hitscore. Taxonomic assignments were bootstrapped with 10 iterations, using the sumtrees method.

### RdRP structure comparisons

We started with a previous dataset of *Flaviviridae* protein structure predictions (https://zenodo.org/records/11092288)[23]. This covers all viruses examined in the current study, with their respective polyprotein sequences broken into sequential 300-residue blocks (overlapping by 100 residues) for protein structure prediction with ColabFold and ESMFold[29,32,34]. The fragmented nature of these structures was insufficient for accurate structural alignment, therefore we queried the dataset against experimental RdRP structure references using Foldseek[30]: orthoflavivirus PDB: 5F3Z, pestivirus PDB: 5YF6 and hepacivirus PDB: 1C2P (refs. 50–52), enabling us to extract continuous RdRP domain sequences for all viruses. Structures were predicted for these RdRP sequences using ColabFold (v.1.5.5)[32], taking the highest confidence model from five predictions. These structures were filtered for average predicted local distance difference test (pLDDT) prediction confidence using a semi-arbitrary cut-off of ≥80%. We also discarded structures below 400 residues in length, reasoning that heavily truncated structures may be misinformative. This yielded a final structure dataset for this study of 400 flavivirus and flavi-like virus RdRP structures. For MSA-free structure prediction, this final set of 400 RdRPs were also analysed with Chai-1 (v.0.5.2)[33] using protein language model ESM-2 embeddings to guide inference (this is the default prediction option for Chai-1)[33,34]. All structure prediction was performed locally, using an NVIDIA A800 40 Gb GPU.

Predicted RdRP structures and experimentally determined structure references (Orthoflavi- PDB: 5F3Z, Jingmen- PDB: 8WU3 (ref. 53), Pesti- PDB: 5YF6 and Hepaci- PDB: 1C2P) were analysed using FoldTree[31] to produce a structure-guided tree based on the local distance difference test structural similarity metric (lddt; not to be confused with pLDDT), which provides a measure of structural similarity while accommodating for some structural flexibility[54]. In short, FoldTree[31] performs an all-vs-all structure comparison, driven by Foldseek[30], to derive pairwise lddt values that are used to calculate a distance matrix and derive a neighbour-joining tree via QuickTree[55]. The tree was visualized and prepared for publication using iTOL[56]. Example structural models in Fig. 3b were aligned for visualization using flexible FATCAT[57], and models were viewed and prepared for publication using UCSF ChimeraX[58].

## Materials availability

No biological materials were used in the study.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

Databases, sequence alignments and raw sequence distance data are provided in Supplementary Information. All RdRP predicted structures and resultant structure-based trees can be found in GitHub at https://github.com/GroveLab/Flavi_RdRp_Structures_Simmonds_2025 (ref. 59).

## Code availability

All code used in the analysis is freely available from the sources cited in the manuscript. Correspondence about the analysis should be addressed to P.S. or J.C.O.M. (RdRP sequence analysis), J.G. (RdRP structure analysis), R.M. (GRAViTy analysis) or A.B. (helicase analysis).

## References

1. Simmonds, P. et al. ICTV virus taxonomy profile: *Flaviviridae*. *J. Gen. Virol.* **98**, 2–3 (2017).
2. Choo, Q.-L. et al. Genetic organization and diversity of the hepatitis C virus. *Proc. Natl Acad. Sci. USA* **88**, 2451–2455 (1991).
3. Scheel, T. K. H., Simmonds, P. & Kapoor, A. Surveying the global virome: identification and characterization of HCV-related animal hepaciviruses. *Antivir. Res.* **115**, 83–93 (2015).
4. Stapleton, J. T., Foung, S., Muerhoff, A. S., Bukh, J. & Simmonds, P. The GB viruses: a review and proposed classification of GBV-A, GBV-C (HGV), and GBV-D in genus *Pegivirus* within the family *Flaviviridae*. *J. Gen. Virol.* **92**, 233–246 (2011).
5. Wu, Z. et al. The first nonmammalian pegivirus demonstrates efficient in vitro replication and high lymphotropism. *J. Virol.* **94**, e01150-20 (2020).
6. Postler, T. S. et al. Renaming of the genus *Flavivirus* to *Orthoflavivirus* and extension of binomial species names within the family *Flaviviridae*. *Arch. Virol.* **168**, 224 (2023).
7. Shi, M. et al. Divergent viruses discovered in arthropods and vertebrates revise the evolutionary history of the *Flaviviridae* and related viruses. *J. Virol.* **90**, 659–669 (2016).

8. Qin, X.-C. et al. A tick-borne segmented RNA virus contains genome segments derived from unsegmented viral ancestors. *Proc. Natl Acad. Sci. USA* **111**, 6744–6749 (2014).
9. Ladner, J. T. et al. A multicomponent animal virus isolated from mosquitoes. *Cell Host Microbe* **20**, 357–367 (2016).
10. Zhang, S. et al. Conserved untranslated regions of multipartite viruses: natural markers of novel viral genomic components and tags of viral evolution. *Virus Evol.* **10**, veae004 (2024).
11. Paraskevopoulou, S. et al. Viromics of extant insect orders unveil the evolution of the flavi-like superfamily. *Virus Evol.* **7**, veab030 (2021).
12. Colmant, A. M. G., Charrel, R. N. & Coutard, B. Jingmenviruses: ubiquitous, understudied, segmented flavi-like viruses. *Front. Microbiol.* **13**, 997058 (2022).
13. Bamford, C. G. G., de Souza, W. M., Parry, R. & Gifford, R. J. Comparative analysis of genome-encoded viral sequences reveals the evolutionary history of flavivirids (family *Flaviviridae*). *Virus Evol.* **8**, veac085 (2022).
14. Simmonds, P. et al. Four principles to establish a universal virus taxonomy. *PLoS Biol.* **21**, e3001922 (2023).
15. Brüssow, H. The not so universal tree of life or the place of viruses in the living world. *Phil. Trans. R. Soc. Lond. B Biol. Sci.* **364**, 2263–2274 (2009).
16. Krupovic, M., Dolja, V. V. & Koonin, E. V. Origin of viruses: primordial replicators recruiting capsids from hosts. *Nat. Rev. Microbiol.* **17**, 449–458 (2019).
17. Nasir, A., Romero-Severson, E. & Claverie, J.-M. Investigating the concept and origin of viruses. *Trends Microbiol.* **28**, 959–967 (2020).
18. Koonin, E. V. et al. Global organization and proposed megataxonomy of the virus world. *Microbiol. Mol. Biol. Rev.* **84**, e00061-19 (2020).
19. Kuhn, J. H. et al. Classify viruses—the gain is worth the pain. *Nature* **566**, 318–320 (2019).
20. Varsani, A. et al. Summary of taxonomy changes ratified by the International Committee on Taxonomy of Viruses (ICTV) from the Animal DNA Viruses and Retroviruses Subcommittee, 2025. *J. Gen. Virol.* **106**, 002113 (2025).
21. Arhab, Y., Bulakhov, A. G., Pestova, T. V. & Hellen, C. U. T. Dissemination of internal ribosomal entry sites (IRES) between viruses by horizontal gene transfer. *Viruses* **12**, 612 (2020).
22. Belsham, G. J. Divergent picornavirus IRES elements. *Virus Res.* **139**, 183–192 (2009).
23. Mifsud, J. C. O. et al. Mapping glycoprotein structure reveals *Flaviviridae* evolutionary history. *Nature* **633**, 695–703 (2024).
24. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
25. Kartashov, M. Y. et al. Novel flavi-like virus in ixodid ticks and patients in Russia. *Ticks Tick Borne Dis.* **14**, 102101 (2023).
26. Debat, H. & Bejerman, N. Two novel flavi-like viruses shed light on the plant-infecting koshoviruses. *Arch. Virol.* **168**, 184 (2023).
27. Schönegger, D., Marais, A., Faure, C. & Candresse, T. A new flavi-like virus identified in populations of wild carrots. *Arch. Virol.* **167**, 2407–2409 (2022).
28. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
29. Varadi, M. et al. AlphaFold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res.* **52**, D368–D375 (2024).
30. van Kempen, M. et al. Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* **42**, 243–246 (2024).
31. Moi, D. et al. Structural phylogenetics unravels the evolutionary diversification of communication systems in gram-positive bacteria and their viruses. *Nat. Struct. Mol. Biol.* https://doi.org/10.1038/s41594-025-01649-8 (2025).
32. Mirdita, M. et al. ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).
33. Chai Discovery Team et al. Chai-1: decoding the molecular interactions of life. Preprint at *bioRxiv* https://doi.org/10.1101/2024.10.10.615955 (2024).
34. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
35. Mayne, R., Aiewsakun, P., Turner, D., Adriaenssens, E. M. & Simmonds, P. GRAViTy-V2: a grounded viral taxonomy application. *NAR Genom. Bioinform.* **6**, lqae183 (2024).
36. Marin, M. S., Zanotto, P. M. D. A., Gritsun, T. S. & Gould, E. A. Phylogeny of TYU, SRE, and CFA virus: different evolutionary rates in the genus *Flavivirus*. *Virology* **206**, 1133–1139 (1995).
37. Mifsud, J. C. O. et al. Transcriptome mining extends the host range of the *Flaviviridae* to non-bilaterians. *Virus Evol.* **9**, veac124 (2023).
38. Kobayashi, K. et al. Gentian Kobu-sho-associated virus: a tentative, novel double-stranded RNA virus that is relevant to gentian Kobu-sho syndrome. *J. Gen. Plant Pathol.* **79**, 56–63 (2013).
39. Wolf, Y. I. et al. Origins and evolution of the global RNA virome. *mBio* **9**, e02329-18 (2018).
40. Fletcher, S. P. & Jackson, R. J. Pestivirus internal ribosome entry site (IRES) structure and function: elements in the 5′ untranslated region important for IRES function. *J. Virol.* **76**, 5024–5033 (2002).
41. Reusken, C. B. E. M., Dalebout, T. J., Eerligh, P., Bredenbeek, P. J. & Spaan, W. J. M. Analysis of hepatitis C virus/classical swine fever virus chimeric 5′NTRs: sequences within the hepatitis C virus IRES are required for viral RNA replication. *J. Gen. Virol.* **84**, 1761–1769 (2003).
42. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
43. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
44. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
45. Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for bigger datasets. *Mol. Biol. Evol.* **33**, 1870–1874 (2016).
46. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
47. Scornavacca, C., Zickmann, F. & Huson, D. H. Tanglegrams for rooted phylogenetic trees and networks. *Bioinformatics* **27**, i248–i256 (2011).
48. Galili, T. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* **31**, 3718–3720 (2015).
49. Aiewsakun, P. & Simmonds, P. The genomic underpinnings of eukaryotic virus taxonomy: creating a sequence-based framework for family-level virus classification. *Microbiome* **6**, 38 (2018).
50. Lesburg, C. A. et al. Crystal structure of the RNA-dependent RNA polymerase from hepatitis C virus reveals a fully encircled active site. *Nat. Struct. Biol.* **6**, 937–943 (1999).
51. Liu, W., Shi, X. & Gong, P. A unique intra-molecular fidelity-modulating mechanism identified in a viral RNA-dependent RNA polymerase. *Nucleic Acids Res.* **46**, 10840–10854 (2018).
52. Noble, C. G. et al. A conserved pocket in the dengue virus polymerase identified through fragment-based screening. *J. Biol. Chem.* **291**, 8541–8548 (2016).

53. Liu, Z. et al. Crystal structures of RNA-dependent RNA polymerases from Jingmen tick virus and Alongshan virus. *hLife* **2**, 18–31 (2024).
54. Mariani, V., Biasini, M., Barbato, A. & Schwede, T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* **29**, 2722–2728 (2013).
55. Howe, K., Bateman, A. & Durbin, R. QuickTree: building huge neighbour-joining trees of protein sequences. *Bioinformatics* **18**, 1546–1547 (2002).
56. Letunic, I. & Bork, P. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Res.* **52**, W78–W82 (2024).
57. Bittrich, S., Segura, J., Duarte, J. M., Burley, S. K. & Rose, Y. RCSB protein data bank: exploring protein 3D similarities via comprehensive structural alignments. *Bioinformatics* **40**, btae370 (2024).
58. Meng, E. C. et al. UCSF ChimeraX: tools for structure building and analysis. *Protein Sci.* **32**, e4792 (2023).
59. GroveLab. Flavivirus RdRP Structures - Simmonds et al. 2025. *GitHub* https://github.com/GroveLab/Flavi_RdRp_Structures_Simmonds_2025 (2025).

## Acknowledgements

We thank A. Crane for critically editing the manuscript. This work was supported in part through a Laulima Government Solutions, LLC, prime contract with the National Institute of Allergy and Infectious Diseases (Contract No. HHSN272201800013C). J.H.K. performed this work as an employee of Tunnell Government Services (TGS), a subcontractor of Laulima Government Solutions, LLC, under Contract No. HHSN272201800013C. N.V. acknowledges partial support from the Centers for Research in Emerging Infectious Diseases (CREID) Coordinating Research on Emerging Arboviral Threats Encompassing the NEOtropics (CREATE-NEO) U01AI151807 grant by the National Institutes of Health (NIH). A.B. was supported by a postdoctoral fellowship from Foundation pour la Recherche Mèdicale (grant number SPF202110014092). J.G. was supported by a Wellcome Trust/Royal Society Sir Henry Dale Fellowship (107653/Z/15/Z) and MRC-University of Glasgow Centre for Virus Research core support from the Medical Research Council (MC_UU_00034/1). J.T.S. was supported by Veterans Administration Merit Review BX000207 and VA SEQCure Network grants. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the US Department of Health and Human Services or of the institutions and companies affiliated with the authors. Mention of trade names, commercial products or organizations does not imply endorsement by the US Government.

## Author contributions

P.S. and J.H.K., in correspondence with other members of the ICTV *Flaviviridae* Study Group (M.B., J.B., J.F.D., A.K., V.L., J.T.S., D.B.S. and N.V.), conceived the study. P.S., A.B., J.G., R.M, J.C.O.M. and J.H.K. conceptualized the experimental section. P.S., A.B., J.G., R.M., D.B.S. and J.C.O.M. performed analyses. All authors wrote/revised the manuscript and P.S. and J.H.K. supervised the work. All authors read and approved the final manuscript.

## Competing interests

All authors declare no competing interests.

## Additional information

**Extended data** is available for this paper at https://doi.org/10.1038/s41564-025-02134-0.

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41564-025-02134-0.

**Correspondence and requests for materials** should be addressed to Peter Simmonds or Jens H. Kuhn.

**Peer review information** *Nature Microbiology* thanks Patrick Dolan, Alexander Ploss and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at www.nature.com/reprints.

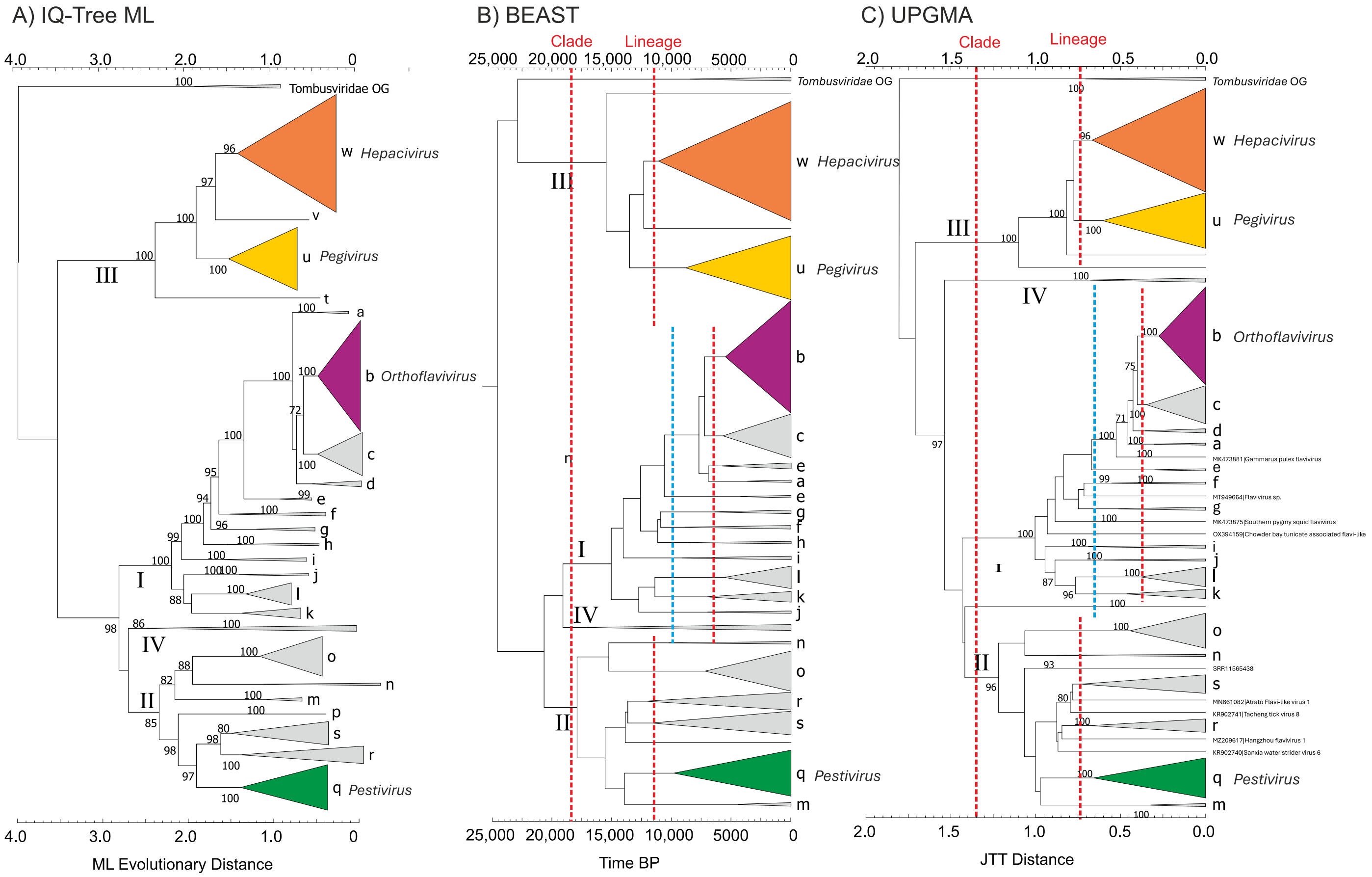
---

[1]Nuffield Department of Medicine, University of Oxford, Oxford, UK. [2]Archaeal Virology Unit, Institut Pasteur, Université Paris Cité, CNRS UMR6047, Paris, France. [3]MRC-University of Glasgow Centre for Virus Research, Glasgow, UK. [4]Sydney Institute for Infectious Diseases, School of Medical Sciences, The University of Sydney, Sydney, New South Wales, Australia. [5]Institute of Diagnostic Virology, Friedrich-Loeffler-Institut, Federal Research Institute for Animal Health, Greifswald-Insel Riems, Germany. [6]Copenhagen Hepatitis C Program (CO-HEP), Department of Infectious Diseases, Copenhagen University Hospital, Hvidovre, and Department of Immunology and Microbiology, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [7]Charité-Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Institute of Virology, Berlin, Germany. [8]Center for Vaccines and Immunity, The Research Institute at Nationwide Children's Hospital, Columbus, OH, USA. [9]Department of Infectious Diseases, Molecular Virology, Heidelberg University, Heidelberg, Germany. [10]Departments of Internal Medicine, Microbiology and Immunology, University of Iowa and Iowa City VA Healthcare, Iowa City, IA, USA. [11]Department of Pathology and Center for Vector-Borne and Zoonotic Diseases, University of Texas Medical Branch, Galveston, TX, USA. [12]Integrated Research Facility at Fort Detrick, Division of Clinical Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Fort Detrick, Frederick, MD, USA. [13]Present address: Institute of Biomedicine, University of Turku, Turku, Finland. [14]These authors contributed equally: Anamarija Butković, Joe Grove, Richard Mayne, Jonathon C. O. Mifsud. [15]These authors jointly supervised this work: Peter Simmonds, Jens H. Kuhn. ✉e-mail: peter.simmonds@utu.fi; jenshkuhn@comcast.net

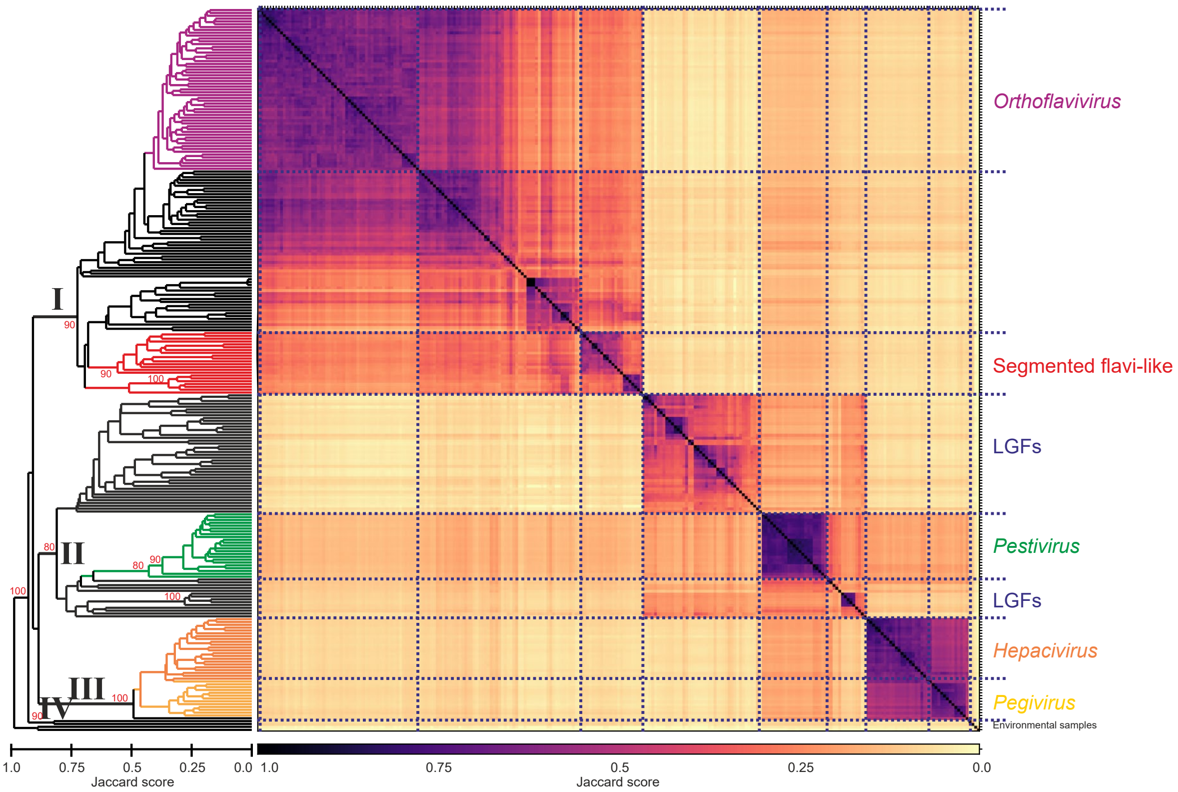
Extended Data Table 1 | Listing of hosts where 'flavi-like' viruses have been described

| Host | Reference |
|---|---|
| Primates, other mammals and birds | 1 - PMID:28786787, 2 - PMID:27692039, 3 - PMID:25545071, 4 - PMID:27741408, 5 - PMID:24511820 |
| Arboviruses (dual host) | 6 - PMID:27569558, 7 - PMID:38361819, 8 - PMID:34026271, 9 - PMID:36299728 |
| Arthropods | 8 - PMID:34026271, 10 - PMID:27609921, 11 - PMID:26491167, 12 - PMID:36529011, 13 - PMID:39018195, 14 - PMID:33050289 |
| Amphibians (frogs) | 15 - PMID:38059479, 16 - PMID:36694816 |
| Reptiles | 17 - PMID:32434883 |
| Cartilaginous and bony fish | 11 - PMID:26491167, 16 - PMID:36694816, 17 - PMID:32434883, 18 - PMID:30397510, 19 - PMID:29147783, 20 - PMID:34017611 |
| Poriferans (sponges) | 13 - PMID:39018195 |
| Cnidarians (jellies) | 16 - PMID:36694816 |
| Mollusks (squid) | 21 - PMID:31068424 |
| Diplurans | 8 - PMID:34026271 |
| Crustaceans | 16 - PMID:36694816, 21 - PMID:31068424, 22 - PMID:34100644 |
| Nematodes | 23 - PMID:24643877 |
| Platyhelminths | 24 - PMID:35536058, 25 - PMID:35902778 |
| Echinoderms (sea cucumbers) | 26 - PMID:32972018 |
| Hemichordates (acorn worms) | 16 - PMID:36694816 |
| Stramenopiles (diatoms and oomycotes) | 27 - PMID:26877136, 28 - PMID:38361818 |
| Angiosperm plants | 7 - PMID:38361819, 29 - PMID:37338667, 30 - PMID:35962825 |

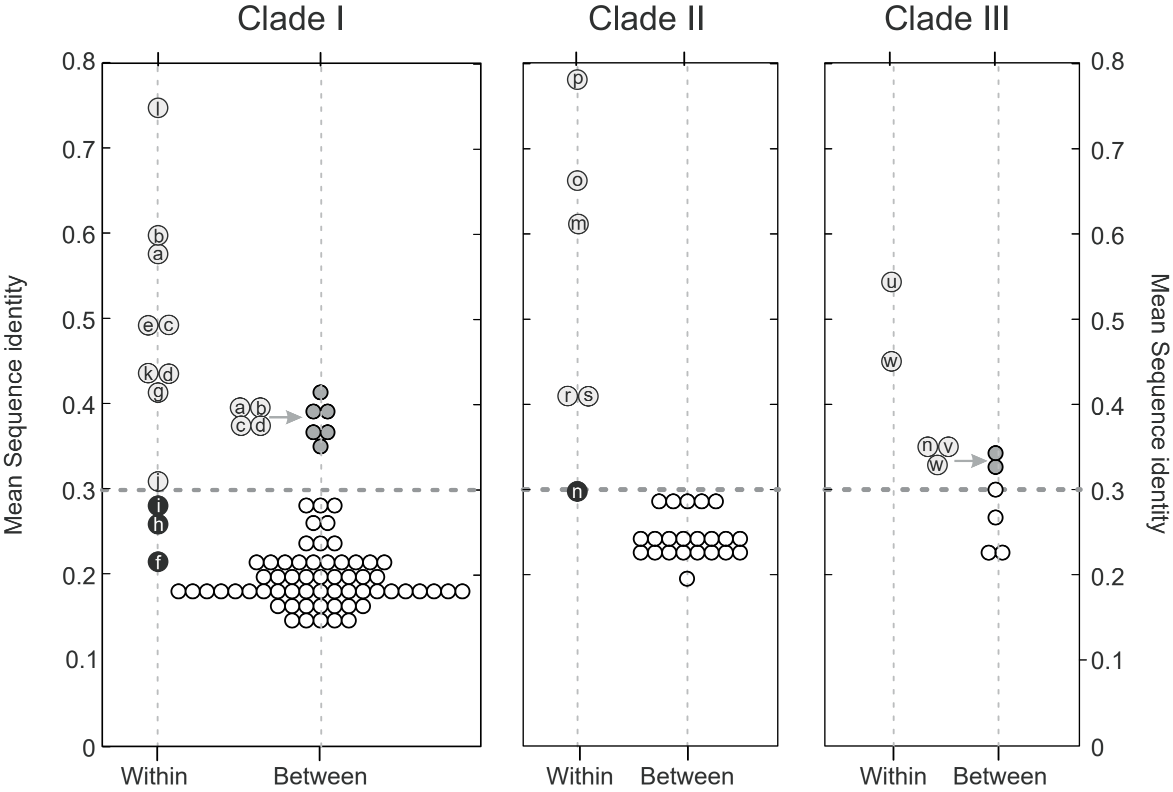
https://doi.org/10.1038/s41564-025-02134-0



**Extended Data Fig. 1 | Comparison of phylogenetic trees of RNA-directed RNA polymerase domain (NS5/NS5B) produced by different tree-building methods.** Phylogenetic trees constructed by likelihood (**a**, **b**) and distance-based (**c**) methods using flavivirus and 'flavi-like' RNA-directed RNA polymerase (RdRP) domain amino-acid sequences. Clades (I–IV) and lineages (a–w) labelled in each tree are based on those in Fig. 1b. Tentative threshold levels of divergence separating clades and lineages are shown as red dotted lines in BEAST and UPGMA trees. An alternative threshold corresponding to the assignment of lineages Ia-Id to a common lineage is shown in a blue dotted line. Abbreviations: BP, before present; BEAST, Bayesian evolutionary analysis by sampling trees cross-platform program; JTT, Jones-Taylor-Thornton matrix; ML, maximum likelihood; UPGMA, unweighted pair group method with arithmetic mean.



**Extended Data Fig. 2 | Heatmap and dendrogram depicting relationships among classified flaviviruses and 'flavi-like' viruses.** Clades I–IV identified in the RdRP phylogeny (Fig. 1b) were added to equivalent branches in dendrogram. Bootstrap support values (10 iterations) for deeper branches are shown in red if ≥70%. A copy of the figure with the branches individually labelled is provided as Fig. S3.



**Extended Data Fig. 3 | Mean pairwise amino acid sequence identities between lineages in clades I–III.** Mean pairwise sequence identities of RdRP domain amino acid sequences between and within lineages of clades I–III. Dotted line indicates an approximate threshold dividing within- and between-lineage distances; between-lineage comparisons above threshold shaded in grey; within-lineage distances below the inter-lineage threshold shown in black.

nature portfolio

Corresponding author(s): Jens H Kuhn

Last updated by author(s): 2025/08/08

# Reporting Summary

## Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

| n/a | Confirmed | |
|---|---|---|
| ☒ | ☐ | The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☒ | ☐ | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☒ | ☐ | The statistical test(s) used AND whether they are one- or two-sided<br>*Only common tests should be described solely by name; describe more complex techniques in the Methods section.* |
| ☒ | ☐ | A description of all covariates tested |
| ☒ | ☐ | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☒ | ☐ | A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☒ | ☐ | For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted<br>*Give P values as exact values whenever suitable.* |
| ☒ | ☐ | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒ | ☐ | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☒ | ☐ | Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated |

*Our web collection on statistics for biologists contains articles on many of the points above.*

## Software and code

Policy information about availability of computer code

Data collection: NCBI GenBank, ICTV Virus Metadata Resource

| | |
|---|---|
| Data analysis | GRAViTy version 2.0<br>IQ-Tree version 1.6.12<br>BEAST version 10.05<br>MAFFT<br>MEGA version 7.0<br>Tanglegram<br>ColabFold v. 1.5.5<br>ESMFold<br>FoldTree<br>Chai-1v0.5.2<br>TrimAl version 1.4.1MUSCLE version 5.1<br>iTOL<br>FATCAT<br>UCSF ChimeraX<br>InterProScan version 5.56-89.0<br>Superfamily version 1.75<br>CCOD version 3.18<br>Pfams version 34.0 |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

## Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Databases, sequence alignments and raw sequence distance data provided in Supplementary Information

## Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

| | |
|---|---|
| Reporting on sex and gender | N/A |
| Reporting on race, ethnicity, or other socially relevant groupings | N/A |
| Population characteristics | N/A |
| Recruitment | N/A |
| Ethics oversight | N/A |

Note that full information on the approval of the study protocol must also be provided in the manuscript.

# Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

# Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

| | |
|---|---|
| Study description | Published data only used |
| Research sample | Availabe nucleotide sequences from classified flaviviruses and from flavi-;like viruses avaialbe on NCBI GenBank database |

| | |
|---|---|
| Sampling strategy | N/A |
| Data collection | N/A |
| Timing and spatial scale | N/A |
| Data exclusions | Only coding complete genome sequences analysed |
| Reproducibility | N/A |
| Randomization | N/A |
| Blinding | N/A |

Did the study involve field work? ☐ Yes ☒ No

# Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

## Materials & experimental systems

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ Antibodies |
| ☒ | ☐ Eukaryotic cell lines |
| ☒ | ☐ Palaeontology and archaeology |
| ☒ | ☐ Animals and other organisms |
| ☒ | ☐ Clinical data |
| ☒ | ☐ Dual use research of concern |
| ☒ | ☐ Plants |

## Methods

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ ChIP-seq |
| ☒ | ☐ Flow cytometry |
| ☒ | ☐ MRI-based neuroimaging |

## Plants

| | |
|---|---|
| Seed stocks | N/A |
| Novel plant genotypes | N/A |
| Authentication | N/A |