The protein structurome of Orthornavirae and its dark matter

Pascal Mutz, Antonio Pedro Camargo, Harutyun Sahakyan, Uri Neri, Anamarija Butkovic, Yuri I Wolf, Mart Krupovic, Valerian V Dolja, Eugene V Koonin

AMERICAN SOCIETY FOR MICROBIOLOGY | mBio

Virology | Research Article

The protein structurome of Orthornavirae and its dark matter

Pascal Mutz,[1] Antonio Pedro Camargo,[2] Harutyun Sahakyan,[1] Uri Neri,[2] Anamarija Butkovic,[3] Yuri I. Wolf,[1] Mart Krupovic,[3] Valerian V. Dolja,[4] Eugene V. Koonin[1]

AUTHOR AFFILIATIONS See affiliation list on p. 22.

ABSTRACT Metatranscriptomics is uncovering more and more diverse families of viruses with RNA genomes comprising the viral kingdom Orthornavirae in the realm Riboviria. Thorough protein annotation and comparison are essential to get insights into the functions of viral proteins and virus evolution. In addition to sequence- and hmm profile-based methods, protein structure comparison adds a powerful tool to uncover protein functions and relationships. We constructed an Orthornavirae "structurome" consisting of already annotated as well as unannotated ("dark matter") proteins and domains encoded in viral genomes. We used protein structure modeling and similarity searches to illuminate the remaining dark matter in hundreds of thousands of orthornavirus genomes. The vast majority of the dark matter domains showed either "generic" folds, such as single α-helices, or no high confidence structure predictions. Nevertheless, a variety of lineage-specific globular domains that were new either to orthornaviruses in general or to particular virus families were identified within the proteomic dark matter of orthornaviruses, including several predicted nucleic acid-binding domains and nucleases. In addition, we identified a case of exaptation of a cellular nucleoside monophosphate kinase as an RNA-binding protein in several virus families. Notwithstanding the continuing discovery of numerous orthornaviruses, it appears that all the protein domains conserved in large groups of viruses have already been identified. The rest of the viral proteome seems to be dominated by poorly structured domains including intrinsically disordered ones that likely mediate specific virus-host interactions.

IMPORTANCE Advanced methods for protein structure prediction, such as AlphaFold2, greatly expand our capability to identify protein domains and infer their likely functions and evolutionary relationships. This is particularly pertinent for proteins encoded by viruses that are known to evolve rapidly and as a result often cannot be adequately characterized by analysis of the protein sequences. We performed an exhaustive structure prediction and comparative analysis for uncharacterized proteins and domains ("dark matter") encoded by viruses with RNA genomes. The results show the dark matter of RNA virus proteome consists mostly of disordered and all-α-helical domains that cannot be readily assigned a specific function and that likely mediate various interactions between viral proteins and between viral and host proteins. The great majority of globular proteins and domains of RNA viruses are already known although we identified several unexpected domains represented in individual viral families.

KEYWORDS RNA virus, Orthornaviria, proteome, protein structure prediction, novel protein domains

Viruses are the most abundant biological entities on earth infecting all life forms. In the recently adopted comprehensive taxonomy, all viruses have been divided into six realms one of which, *Riboviria*, consists of an enormous variety of viruses with RNA genomes that encode homologous replication enzymes, RNA-directed RNA polymerase (RdRp), in the kingdom *Orthornavirae*, or reverse transcriptase, in the kingdom

Editor Michael S. Diamond, Washington University in St Louis School of Medicine, St. Louis, Missouri, USA

Address correspondence to Pascal Mutz, Pascal.Mutz@nih.gov, Valerian V. Dolja, valerian.dolja@oregonstate.edu, or Eugene V. Koonin, koonin@ncbi.nlm.nih.gov.

The authors declare no conflict of interest.

Received 17 October 2024
Accepted 28 October 2024
Published 23 December 2024

Pararnavirae. In the last few years, metatranscriptomics along with targeted approaches have been uncovering diverse orthornavirus families at an increasing pace (1–4). Orthornaviruses have small genomes, mostly, between 3 and 20 kilobases (kb), with the upper bound of 35–64 kb in nidoviruses (5–7) and up to 40 kb in flavi-like viruses (8), and accordingly, encode limited repertoires of protein domains. Thorough annotation and comparison of viral proteins provide ample insights into protein functions, virus evolution, and in some cases, virus-host association. The RdRp is the only protein that is conserved in all orthornaviruses (9) and thus serves as the primary query for the discovery of orthornaviruses in metatranscriptomes (1, 3, 10). Other proteins conserved in large groups of orthornaviruses include helicases and proteases of different families, mRNA capping enzymes, and capsid proteins of different types. Several other protein domains are conserved across more narrow ranges of orthornaviruses, often associated with specific host organisms. Such domains include the movement proteins (11) and AlkB family oxygenases (12) found in a variety of plant viruses, ADP-ribose-binding Macro domains encoded by several families of animal viruses (13), lysozymes encoded by a variety of RNA bacteriophages (1, 14), and more. All these protein domains are conserved at the sequence level so that annotation using protein family profiles can delineate their core regions in viral genomes with high confidence. However, even exhaustive (viral) protein profile generation and searches leave many unannotated proteins and large protein regions in newly discovered orthornaviruses. The question thus remains how these unannotated portions of viral proteins are related to each other and what are their structures and potential functions.

Protein structures are in general far more strongly conserved than sequences, and therefore, structural comparisons have the potential to illuminate the "dark matter" of viral proteomes as compellingly demonstrated for proteomes of cellular life forms, in particular, the human proteome (15, 16). With the advent of high-accuracy protein structure prediction methods, such as AI-based AlphaFold2 and RoseTTAFold, or using transformer protein language models such as ESMfold, comprehensive protein structure prediction and analysis have become realistic (17–19). Large-scale databases of protein structure models are already available (e.g., EBI: https://alphafold.ebi.ac.uk/ [20] and ESM Atlas: https://esmatlas.com/ [19]), covering either single species proteomes or metatranscriptomes such as ESM atlas. However, no such databases were available for viral proteins at the time this research was being conducted, in part, due to the difficulty of processing polyproteins that are encoded by many viruses, in particular, members of *Orthornavirae*. Only recently, Nomburg and colleagues presented predicted structures for proteins of eukaryotic viruses (21) and Kim and colleagues modeled structures for viral representatives of uniref30 clusters (22).

We have previously demonstrated the utility of protein structure prediction using AF2, followed by comparison to structure databases, for predicting functions of uncharacterized proteins or protein domains of DNA viruses, revealing, in particular, multiple cases of exaptation of host enzymes (23, 24). We were then motivated to explore in depth the proteomic "dark matter" of orthornaviruses using a similar approach. To this end, we used a previously published data set of (predicted) orthornavirus genomes spanning over 300,000 viral contigs from nearly 500 virus families discovered in metatranscriptomics (1). Of note, nearly 400 of the analyzed virus families are operational and not formally ratified. We pre-processed the predicted viral proteins to isolate unannotated domains and hypothetical open reading frames (ORFs), modeled them in addition to well-annotated domains using AF2, and performed structure comparisons among all viral proteins to compose a virus protein "structurome." These structural models were then compared to structures of cellular proteins represented in the PDB. We found that the vast majority of unannotated regions of orthornaviral proteins and smaller unannotated ORFs were either predicted to form a "generic" fold, such as a single α-helix, or could not be modeled with high confidence, suggesting a non-globular structure. Nevertheless, several globular domains, mostly, represented in one or more narrow viral lineages and not previously detected either in any

orthornaviruses or at least in a given viral family were predicted. Taken together, the results of this analysis indicate that the widespread globular domains comprising the proteome of orthornaviruses are largely known, whereas newly identified proteins and domains are lineage specific, are in many cases non-globular, and are likely to be involved in interactions between viral and host proteins.

RESULTS

Annotated and unannotated domains and proteins of orthornaviruses

To compile a comprehensive set of orthornavirus proteins, we used a recently published data set (hereafter environmental metatranscriptome RNA virus [EMRV] set) of predicted orthornavirus genomes spanning more than 370,000 viral contigs from nearly 500 operationally defined virus families of which 98 had been approved by the ICTV at the time of publication (2022) (1). All proteins and domains annotated in this study were extracted, yielding a set of 647,383 protein sequences. Then, evolutionarily conserved but unannotated (putative) proteins and domains that are conserved in groups of viruses were identified (see Fig. S1 in the supplemental material for a schematic and Materials and Methods for details). In brief, ORFs from start to stop were predicted in all six frames and matched to the published annotations (1). ORFs without annotation and with only partial annotation were processed to retrieve the unannotated sequence stretches. In the case of polyproteins and multidomain proteins, unannotated and partially annotated proteins of at least 200 amino acids (aa), in which a continuous stretch of at least 60 aa was unannotated, hereinafter conserved unannotated domains (CUDs), were extracted. Unannotated ORFs between 60 and 199 aa (n = 1,362,871) were not sliced further and kept for downstream analysis (hereinafter "ORFans"). To identify evolutionarily conserved sequences that are likely to be expressed and functional, ORFans and CUDs were clustered by sequence similarity (see Materials and Methods). All clusters that included proteins from different genomes which are represented by at least three leaves in the RdRp-based phylogenetic trees or at least one CUD were retained for further analysis, resulting in 6,117 clusters of ORFans representing 100,694 sequences and 13,085 CUDs representing 31,247 sequences (Fig. S2A and B).

Next, we assessed which known protein domains were missed during profile-based annotation and might be present within the CUD and ORFan set. To this end, we downloaded a curated set of viral reference genomes from the ICTV (ICTV exemplars, https://ictv.global/vmr, VMR_MSL38_v2) matching the 98 virus families recognized in the EMRV set and extracted the annotated domains and proteins (n = 32,648) from the GenBank files (hereafter exemplar domains). Exemplar domains were clustered and compared against the same viral profile databases used for EMRV annotation ran using hhsearch (25). Clusters with at least one highly confident hit (probability ≥95%) were harmonized and assigned accordingly. As a result, 64% of ICTV clusters were assigned confidently, spanning 83% of all exemplar domains. Across virus families, about 70% of all clusters associated with a given family could be confidently assigned, increasing to 80% if considering only clusters spanning domains on the RdRp-encoding segment, those recovered by Neri et al. (1) and included in the EMRV data set (Fig. S3A). The clusters without confident assignment were dominated by functionally uncharacterized domains (Fig. S3B). Thus, some known viral proteins and domains are not identified with high probability by the used viral protein profiles and might be present among the CUDs and ORFans. These under-annotated domains were identified as part of the refinement of the orthornavirus proteome as described below.

The pan-proteome of *Orthornavirae*

Combining the annotated domains and proteins with CUDs and ORFans, we constructed a pan-proteome for each orthornavirus family. All domains and ORFs (annotated or not) were clustered by sequence similarity and domains were either labeled by their functional tag or as CUD or ORFan. Similar procedure was performed for the ICTV

exemplar set, and the resulting virus family pan-proteomes were compared (Fig. 1A). For the 98 virus families represented in both the ICTV exemplars and the EMRV set, comparable numbers of functions (profile assignments) per virus family were detected, confirming that robust domain annotation was obtained for the EMRV set. Functional assignment per virus family across all 498 families was lower because about 25% of the families included viruses for which only the RdRp was detected, that is, most likely, viruses with segmented genomes (Fig. 1B).

A substantial majority of the domains annotated by a particular profile were found in a single virus family in both the EMRV set (85%) and the ICTV exemplar set (89%). Thus, these domains represent virus family- or even genus- or species-specific functions (Fig. 1C and E). The RdRp was the only protein represented in all virus families. Very few other domains and functions were found to be broadly distributed across virus families. These widespread functions are different types of capsids (40% of families in the EMRV set vs 76% in the ICTV exemplar set), mRNA capping enzymes (35% vs 47%), helicases (33% vs 45%), and proteases (22% vs 30%) (Fig. 1D and F). The consistently lower abundance of these domains across virus families within the EMRV set compared to the ICTV exemplars is probably due to the absence of non-RdRp-encoding segments of multipartite genomes in the EMRV set.

The structurome of *Orthornavirae*

To predict the structures and functions of the CUDs and ORFans, we generated structural models using AlphaFold2 (17). Only a fraction of structures (about 34% of CUDs and 10% of ORFans) could be predicted with high confidence (mean plddt score ≥70), indicating the possibility of larger intrinsically disordered stretches (Fig. S4A and 5A). To address this possibility, we predicted intrinsically disordered structures in CUDs and ORFans. Although we found no significant correlation with the plddt score of the structural models, proteins predicted to have disordered regions have mostly a plddt score below 70 (Fig. S4B and 5B). To account for the uncertainty of correct folds among low plddt structures, we clustered all CUDs and ORFans independently with Foldseek (0.8 coverage) and kept only clusters in which at least one member had a mean plddt score of 70 or higher, resulting in 412 ORFan and 1,594 CUD representatives.

These representatives were compared to PDB using Dali (26). Secondary structure prediction with Psique (27) indicated an all-α-helical fold for about 69% of CUDs and nearly 76% of ORFans (Fig. 2). Furthermore, for 53% of the CUDs and 35% of the ORFans, folds distinct from the simplest ones, such as a single α-helix or helix-turn-helix (HTH), were predicted. Those were considered as CUD and ORFan of interest (COI and OOI hereafter, respectively). About 37% and 13% of the ORFans and CUDs with a predicted simple fold were predicted to contain at least one transmembrane domain compared to about 8% and 9% of the more structured ORFans and CUD, respectively, indicating a substantial enrichment of small transmembrane proteins among ORFans with a simple fold. Most of the COIs and OOIs were confined to a single virus family (Fig. S6). Foldseek clusters spanning representatives of multiple families represent mainly capsid proteins (see below). Furthermore, COIs were checked with a "neighborhood" approach to determine whether a given COI was confidently annotated in a related genome from the same virus family, pointing to its function (Fig. 3A and B). In brief, COIs were searched against the annotated domains using psi-blast to obtain a provisional annotation of the COI. Next, related proteins from neighboring genomes with or without COI were aligned and the respective annotations were mapped to the alignment. Whenever annotations from neighboring proteins overlapped with the COI, the annotation was compared with the psi-blast hit. If consistent, the COI was considered a "refinement" and was not investigated further. Whenever there was no overlap with an annotated domain, a conflicting or mixed result, or no psi-blast hit, the COI was kept for manual inspection of the Dali results. Of the 852 COIs with confidently predicted structures, 62 were from genomes not phylogenetically assigned, 553 represented "refinements," many as part of the RdRp.

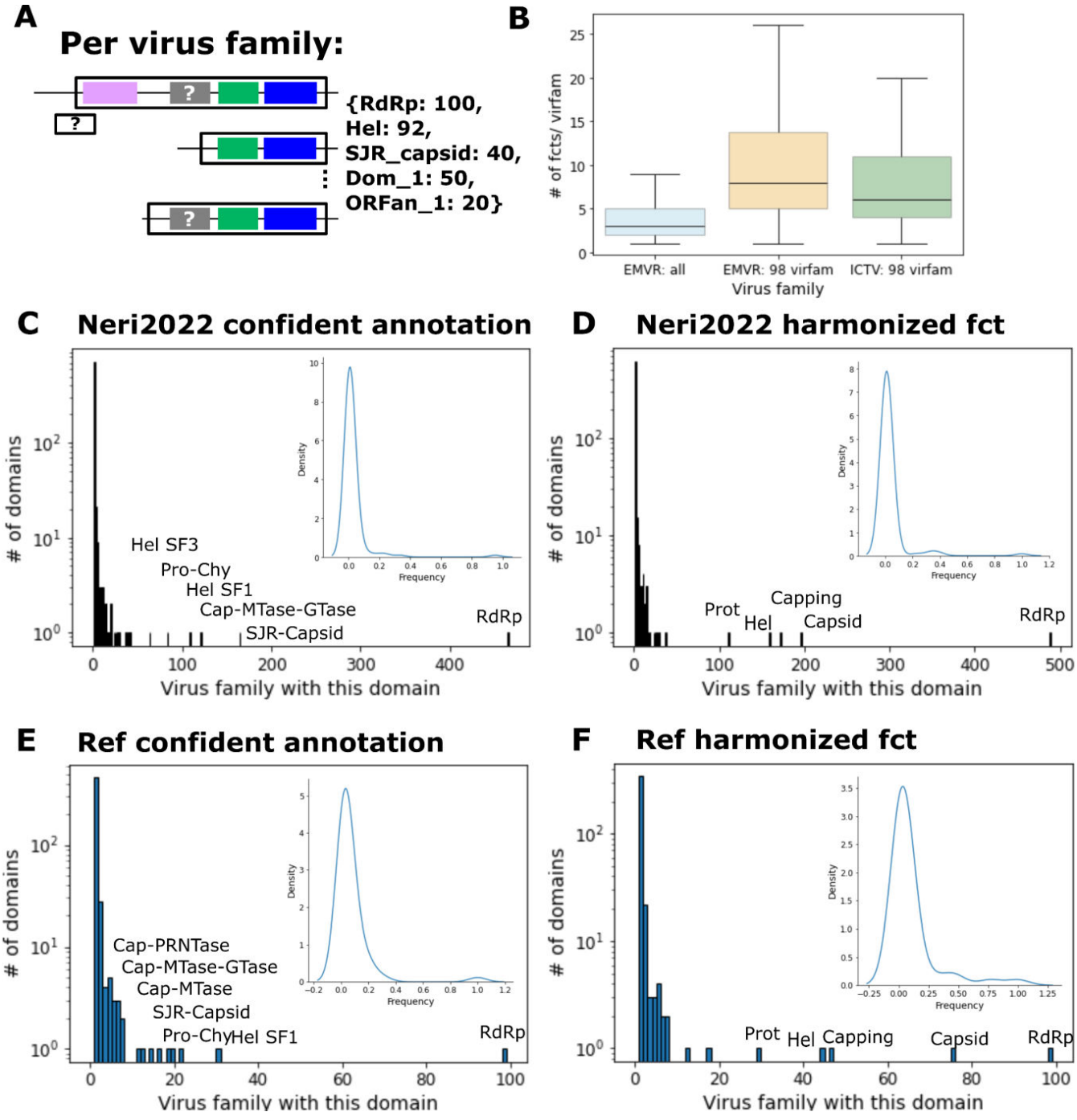

FIG 1 Frequencies of EMRV profile annotated vs ICTV exemplar-annotated proteins across virus families (A) Schematic of a prototype pangenome per virus family with 100 contigs which are either incomplete or complete but all contain the RdRp (blue), and some also contain helicase (Hel, green), a single jelly-roll capsid (SJR_capsid, pink), an unannotated domain (Dom_1, CUD, gray) or an ORFan (ORFan_1, white). (B) A number of unique annotated domains per virus family ("# of fcts/virfam") for all virus families in the EMRV set ("EMRV: all"), of the 98 named virus families with a corresponding family in the ICTV exemplar set ("EMRV: 98 virfam") and of the ICTV exemplar set ("ICTV: 98 virfam"). (C) A number of unique annotations based on profile comparison per virus family across all 498 families. (D) Same as panel C but with harmonized functions (e.g., combining all Helicase-related labels as "'Hel"). (E) Number of unique annotations within the ICTV exemplar virus families based on nvpc profile db comparison. (F) Harmonized functions (e.g., "capsid" represents the functional tags "nucleoprotein," "SJR capsid," "core," and others assigned to capsid and nucleocapsid proteins) across the ICTV exemplar virus families as in panel E together with proteins which are annotated in GenBank but not in nvpc. (C–F) Inset shows frequencies for all functional domains that are present in at least two families.

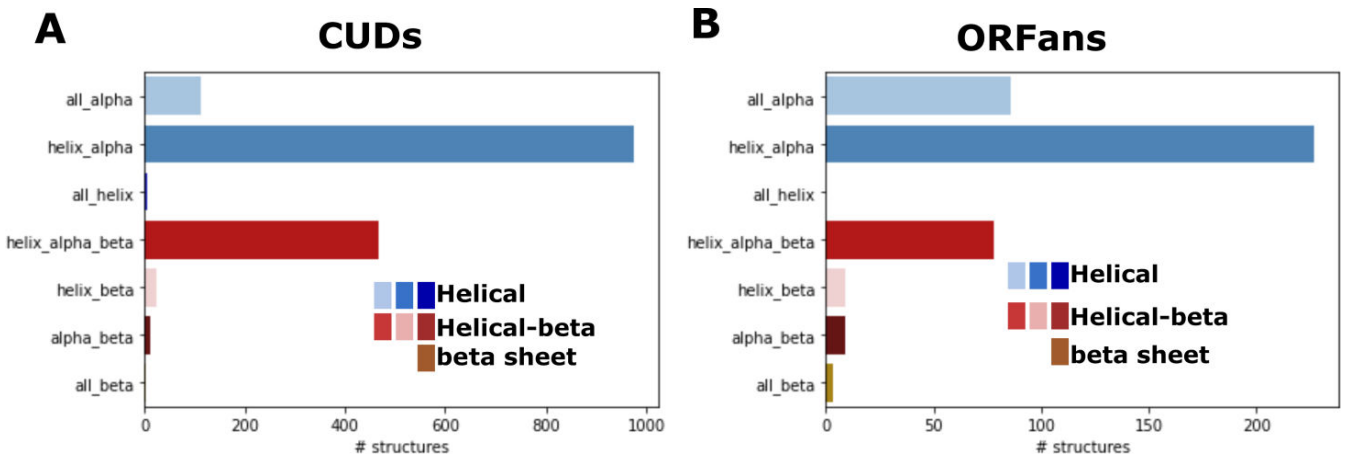

FIG 2 Secondary structure assignments for CUDs and ORFans Psique-based secondary structure assignments are shown for all CUDs (A) and ORFans (B) with a mean plddt ≥70. α-helical types in the blue color range, β-strand and α-helical in the red color range, all-β in brown, and other in gray.

To incorporate the predicted structures of COIs and OOI into the context of known structures across all *Orthornavirae* families, we constructed a structurome of *Orthornavirae* by modeling one representative structure per functional tag per virus family from the pangenome (e.g., one representative RdRp structure per virus family, 2,421 annotated representatives in total). The ICTV exemplar domains lacking strong profile annotation were modeled too (3,000 representatives). This modeling resulted in 6,419 predicted structures which were pre-clustered with Foldseek (0.8 coverage, 4,022 clusters). The great majority of the structures, 88%, remained singletons (Fig. S7A). In 27 cases, OOIs were found in a Foldseek cluster together with annotated domains and/or ICTV exemplar domains, and the same was found for 13 COIs. Then, 4,022 cluster representatives were analyzed by an all-vs-all Dali comparison, finalizing the *Orthornavirae* structurome. Structures within the structurome were clustered based on the Dali all-vs-all z-scores in an iterative procedure in which individual structures were added to a cluster as long as the mean z-score to each other structure already present was above a given z-score. The threshold was set at a z-score of 7 or higher ("z7 clusters") to avoid over-clustering. This procedure resulted in 333 z7 clusters of which 59% contained two structures, whereas the largest cluster consisted of 43 structures (Fig. S7B). These 333 z7 clusters represented 1,201 (30%) of all representative structures in the Dali all-vs-all run. The structurome was visualized as a structure-similarity network in which each structure is a node connected to other nodes via edges weighted by the pairwise z-score. Figure S8 shows a subset of the network in which only structures present in a z7 cluster are shown or which have at least one connection to a structure within a z7 cluster with a z-score of 7 or higher (about 36% of all structures in the Dali all-vs-all run). As expected, we detected z7 clusters of structures coming from well known, functionally characterized domains such as RdRp, helicases, single jelly-roll (SJR) capsid proteins, and proteases. Some functional labels were distributed across several z7 clusters. For example, RdRp structures that dominate the structurome (489 structures initially) contributed to 4 z7 clusters with one harboring about 88% of the representative RdRp structures (Fig. S7B).

About 28% (n = 92) of z7 clusters consisted of OOIs and COIs only, that is, represented unique shared folds. Notably, nearly all α-helical domains and proteins (annotated or not) were placed in eight larger z7 clusters (10 representatives or more) in which they are connected by z-scores of 7 or higher but these connections apparently reflect generic structural similarity rather than shared distinct folds (hubs labeled "Helical" in Fig. S8).

OOIs and COIs were inspected semi-manually by considering Dali results, secondary structures, and structural relationships within the structurome. All-helical COIs and OOIs mostly did not show conclusive Dali hits, with moderate structural similarity between small α-helical modules detected across apparently unrelated proteins. Inspection of

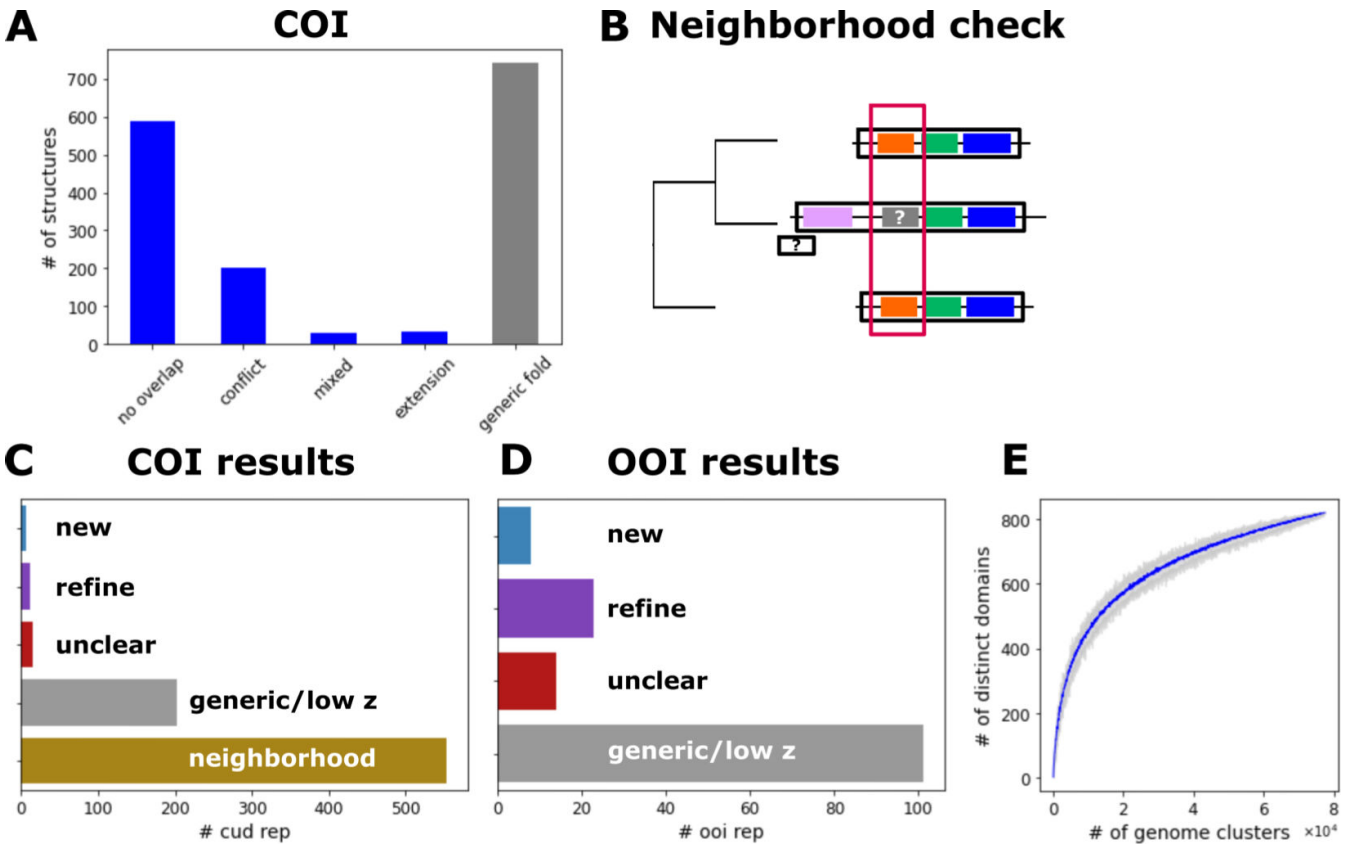

FIG 3 Overview of domains and ORFans of interest. (A) A number of representative COI (conserved unannotated domain of interest) structures binned as follows: (i) no overlap with present annotation in genome; (ii) conflict: there is a present annotation that slightly overlaps with the provisional CUD annotation; (iii) mixed: members of a CUD cluster had substantially different provisional PSI-BLAST annotations; (iv) extension: the provisional psi-blast annotation of a CUD extended the annotation of an existing profile-based annotated domain; (v) generic fold: based on Dali results, the fold is a single helix, HTH, a beta-hairpin or disordered. Categories i–iii were analyzed further. (B) Schematic of the neighborhood analysis. Homologous multidomain proteins or polyproteins of neighboring genomes were aligned, and protein annotations were mapped on the alignment. If a putative COI region overlapped a confident annotation, it was considered annotated. (C) Number of COIs that were considered annotated as a result of the neighborhood analysis (bottom bar) and results of the semi-manual examination of COIs. New: a COI representative with a predicted structure not reported previously for the given virus family. Refine: Dali results pointed to a refinement of the annotation as the structure/function was already reported for other members in this virus family. Unclear: a high-confidence model was obtained for a COI but Dali hits were inconclusive (mainly, alpha/beta domains). Generic/ low z: the structure is too generic to produce meaningful Dali hits (e.g., an alpha helix with a small beta-hairpin) or the Dali z-score was not significant (below 4). (D) Results of the semi-manual check of OOIs. Binning is as in **C**. (E) Rarefaction curve of distinct domains as a function of the number of sampled genome clusters (leaves). The blue line represents a mean of 30 bootstraps and gray area shows the range of unique domains at each sampling step (step size: 50 genome clusters).

COIs and OOIs predicted to fold into globular structures either revealed a fold and function not yet reported for a given virus family ("new"), or a fold not yet reported for a given virus family for which no clear function could be assigned due to significant but too variable Dali hits ("unclear"), or indicated a refinement of the current profile-based annotation of a given virus family ("refined"). Altogether, we classified the predicted structures for OOIs and COIs, respectively, as follows: 8 and 7 new; 14 and 16 unclear; and 23 and 12 refined (Fig. 3C and D).

Saturation of the rarefaction curve of distinct domains sampled randomly across *Orthornavirae* genome clusters suggests that the substantial majority of widely distributed *Orthornavirae* domains are already known (Fig. 3E). Conversely, new viral domains are expected to be found in single virus families which is in line with the findings of this study. Having both the structure-based and profile-based annotations at hand, we aimed to identify the core set of unique domains per virus family shared by at least 50% of genomes (see Materials and Methods for details) and compare it to

the overall distribution of domains per virus family. For most orthornavirus families, we identified a small core, often consisting of the RdRp alone, and a "shell" of domains with intermediate frequencies (Fig. S19). This distribution of domain frequencies could reflect true variation across a virus family, for example, on the genus level but also the presence of incomplete genomes in the data set, in particular, those with multiple segments, and a lack of complete domain annotation.

Discovery of new domains in the orthornaviral structurome

"Unexpected" OOIs and COIs identified in a particular virus family potentially could be either truly novel, that is, not reported so far in any *Orthornavirae*, or new to the given virus family. There were no unequivocal examples of the former case although we identified two OOIs from the putative family *f.0145* (o.0036, c.0025, *Kitrinoviricota*) with a predicted β-barrel fold. Best but not consistent Dali hits were two bacterial β-barrel fold proteins (top z-scores ~ 5, PilZ domain and HCP3, a paralog of type VI secretion system effector, Hcp1, pfam PF05638, for the two OOIs, respectively) and no structural similarity was observed to known virus proteins in the *Orthornavirae* structurome or, to our knowledge, any other viruses (Fig. S9). Apart from these two β-barrel domains, we identified several other domains known to be encoded by *Orthornavirae* members but here found in unexpected virus families.

Nucleic-acid-binding domains

Several COI and OOI products with different folds seem to be involved in nucleic acid binding. One of these nucleic acid-binding domains is the phytoreovirus core-P7 dsRNA-binding domain (P7-dsRBD) that is known to be encoded by members of *Sedoreoviridae (28*) and here was found in proteins of various virus families as annotated by profile comparison: *Chrysoviridae* (70/81 leaves covered), *Endornaviridae* (30/222), *Megabirnaviridae* (3/12), *f.0281.base-Megabirna* (7/29), *f.0285* (31/94), *Flaviviridae* (4/360), with z-scores of 8–11 to each other (see pangenome in the Supplementary Material on zenodo) for all virus families with this domain).

Besides a refined census of P7-dsRBDs, for example, in *Cystoviridae*, we identified an OOI with a similar fold in *Picobirnaviridae* (in genomes of 14/1376 leaves scattered across the tree; Fig. 4). Structure comparison against PDB revealed prominent similarity to nucleotide monophosphate (NMP) kinases, such as adenylate kinase (z-score ~8). Thus, proteins with a core-P7-like RBD fold likely originated from cellular NMP kinases and might have spread horizontally among diverse viruses. Representative viral protein structures with P7-dsRBD from different virus families including *Sedoreoviridae*, *Chrysoviridae,* and *Picobirnaviridae* were aligned with homologous cellular kinases using FoldMason (29), and a phylogenetic tree was constructed using IQTree2 (30). Most of the viral P7-dsRBDs formed a distinct clade separate from the cellular kinases (Fig. 4C) which is compatible with functional divergence after exaptation and subsequent horizontal spread. Exceptions are the P7-dsRBDs of *Picobirnaviridae* which clustered within the cellular kinase clades, suggestive of independent exaptation events (Fig. 4C). While the Walker A motif is intact in viral core-P7-like dsRBD fold proteins, the Walker B motif is missing, indicating loss of kinase activity (Fig. S10). The P7-dsRBD domains are found either as stand-alone proteins or are incorporated into viral polyproteins (Fig. S11). This domain was identified in three families from the order *Ghabrivirales* (*Chrysoviridae*, *f.0296*, and *f.0285*) indicative of an old acquisition but, in contrast, seems to have been more recently acquired at least twice by *Picobirnaviridae* members (Fig. 4B and C). Other members of *Picobirnaviridae*, predicted to infect bacterial hosts, have been shown to encode a putative lysozyme (completely unrelated to dsRBD or kinases) in the same genomic location (1). Furthermore, other members of *Picobirnaviridae* encode a capsid protein 5′ of the RdRp ORF, demonstrating the flexibility of *Picobirnaviridae* to capture diverse ORFs 5′ of the RdRp ORF. The exact functions of the P7-dsRBD domains in different virus families remain unclear.

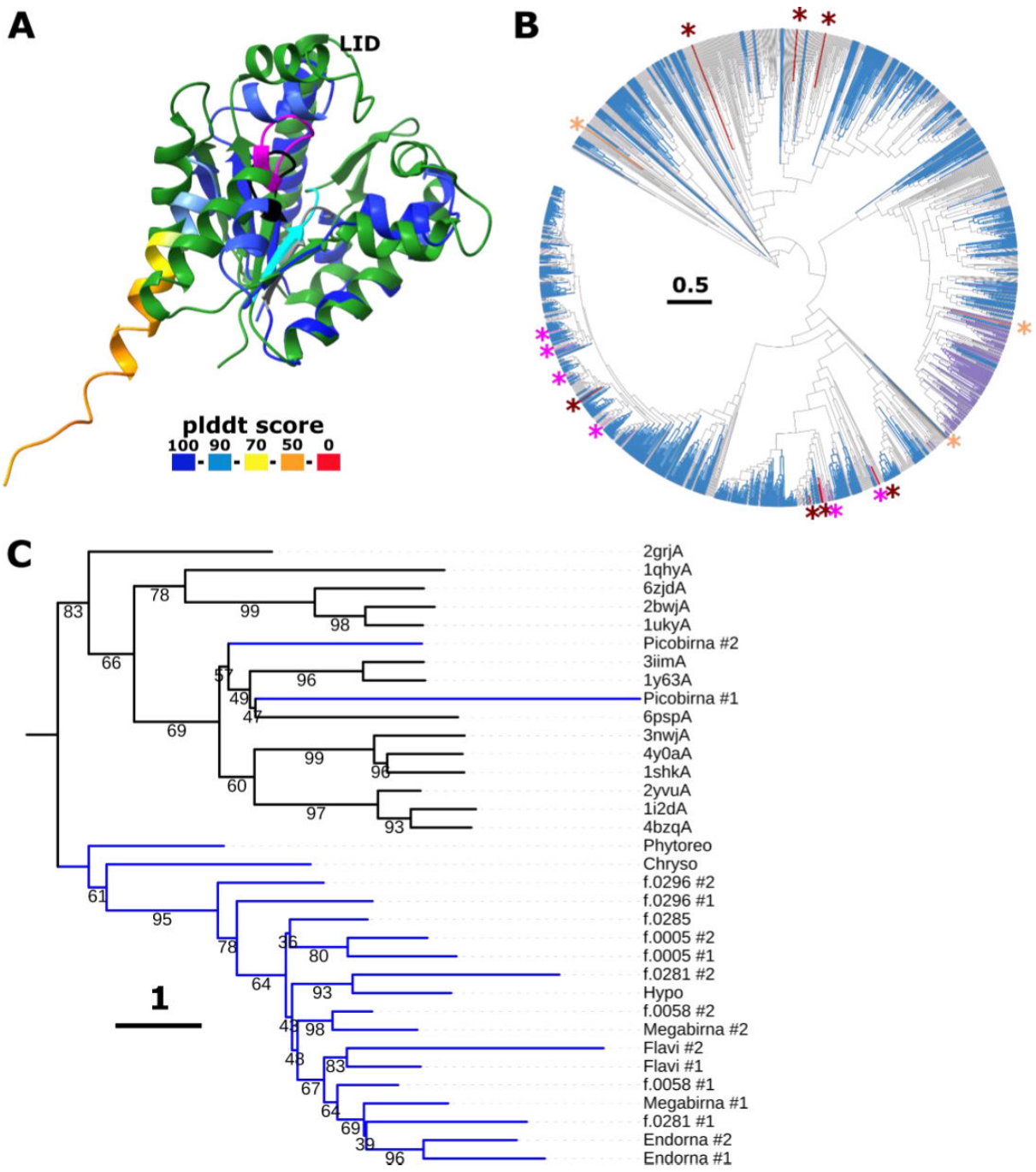

FIG 4 Phytoreovirus core-P7 dsRNA-binding domain with a kinase fold in orthornaviruses. (A) Superposition of *Picobirnaviridae* 5′ ORFan (colored by plddt score, Walker A motif is shown in black, position of degraded Walker B motif is shown in gray) with adenylate kinase from *Methanococcus igneus* (6psp, green, Walker A motif shown in magenta, Walker B motif is shown in cyan, z-score 9.1). (B) Phylogenetic distribution of contigs encoding P7-dsRBD (red branches and asterisks), lysozyme (orange and asterisk), and capsid protein (purple) 5′ of the RdRp within *Picobirnaviridae*. Blue color indicates contigs containing less than 180 nt in front of the RdRp ORF (likely incomplete). (C) Phylogenetic tree based on a structure-guided alignment of viral P7-dsRBD domains found in different virus families by structure comparison (*Picobirnaviridae*) or profile comparison (EMRV set) with structurally similar kinases (z-scores 6–11; order as in tree: Dephospho-CoA kinase from *Thermotoga maritim*a [2grjA]; chloramphenicol phosphotransferase from *Streptomyces venezuelae* [1qhyA]; adenylate kinase 3 from *Homo sapiens* [6zjdA]; adenylate kinase 5 from *Homo sapiens* [2bwjA]; uridylate kinase from *Saccharomyces cerevisiae* [1ukyA]; atypical mammalian nuclear adenylate kinase hCINAP from *Homo sapiens* [3iimA]; probable kinase from *Leishmania major* Friedlin [1y63A]; adenylate kinase from *Methanococcus igneus* [6pspA]; shikimate kinase from *Arabidopsis thaliana* [3nwjA]; shikimate kinase from *Acinetobacter baumannii* [4y0aA]; shikimate kinase from *Erwinia chrysanthemi* [1shkA]; APE1195 from *Aeropyrum pernix* K1 [2yvuA]; ATP sulfurylase from *Penicillium chrysogenum* [1i2dA]; and APS kinase CysC from *Mycobacterium tuberculosis* [4bzqA]). Branches of viral P7-dsRBD are colored in blue, and those of cellular kinases are colored in black.

Notably, all experimentally characterized viruses with a P7-dsRBD have dsRNA genomes, suggesting that this domain is specifically involved in capsid-associated dsRNA transactions. To our knowledge, there are no experimental studies for viruses with ssRNA genomes, such as flaviviruses, on the role of the observed P7-dsRBD. Nevertheless, given that all orthornaviruses produce replicative dsRNA intermediates in the host cell, it seems likely that these domains are involved in transcription and/or RNA packaging as reported for the phytoreovirus P7 (31), but additional or alternative roles, for example, in the suppression of the host RNAi system, cannot be ruled out.

An OOI containing a common RBD fold (32) consisting of four β-strands and two α-helices (obviously, unrelated to the kinase-derived RBD discussed above) was identified in members of the *Hepeviridae* family (Fig. 5). This OOI is located 3′ of the non-structural polyprotein and capsid protein ORFs. It is unclear whether this OOI actually binds RNA because the top Dali hits are the RBD that have lost the RNA-binding capacity and are instead involved in protein-protein interactions, such as RBD2 of the *Arabidopsis* protein HYL1 (33).

Another nucleic acid-binding domain is a dsRBD that is found in many virus families such as *Nodaviridae*, *Astroviridae,* and *Reoviridae* and functions as a viral suppressor of host RNAi defense (see virus family pangenomes). We additionally identified related dsRBD as a COI domain in *f.0092*, basal to *Permutotetraviridae* (Fig. S12). This domain was detected in genomes represented by nearly all leaves of this family (16/17).

Yet another fold implicated in nucleic acid binding is a winged helix-turn-helix (wHTH) domain found in family *f.0008* basal of *Polycipiviridae,* the only virus so far identified in *f.0008* is Lothians earthworm picorna-like virus 1 (34) (Fig. 6). The wHTH domain apparently was acquired recently by *f.0008* members as it was only found in a distinct, distal clade (Fig. 6C). In addition to the wHTH domain, the same and additional *f.0008* family members were shown to encode an SJR-fold protein (Fig. 6B through D). Given the genome architecture of *f.0008* members with an annotated larger capsid ORF 5′ of the newly identified SJR protein, and given that it clusters with other plant movement proteins (MP) in the structurome, it is highly likely that this is a 30K superfamily MP that evolved by duplication of an SJR capsid protein followed by neofunctionalization (11). Thus, plants are likely the hosts for at least this clade of *f.0008* members.

A galactose-binding domain

Independently of the putative MP and wHTH, other members of the family *f.0008* were found to encode a galactose-binding domain (Fig. S13; Dali z-score of 15.5 against a bacterial carbohydrate-binding domain). The general genome organization of *f.0008* members differs between those encoding the wHTH protein and MP, and those encoding the carbohydrate-binding domain. The former viruses encode a 5′ non-structural polyprotein followed by the capsid protein ORF whereas the latter ones encode a single polyprotein with the capsid protein domain at the N-terminus followed by the galactose-binding domain. The galactose-binding domain is specific for a long branch within the *f.0008* family and is likely to be involved in host-specific interactions (35).

Endonuclease domains

With well over 4,000 members, *Marnaviridae* is currently the largest family in the *Pisuviricota* phylum (1). The few characterized viruses of this family infect diverse marine protists (36). We found that a number of marnaviruses encode an unexpected domain that is located at the C-terminus of the RdRp and shows significant structural similarity to NucS-like endonucleases (restriction endonuclease fold) (Fig. 7; Dali z score 8.2). Endonucleases of different folds are also encoded by several groups of orthornaviruses, such as influenza viruses, where this enzyme of the PD-(D/E)XK nuclease superfamily cleaves host mRNAs, snatching the cap for viral RNA synthesis (37), and nidoviruses, in which NendoU (nidoviral uridylate-specific endoribonuclease) (38) is involved in viral replication and evasion of host innate immunity (39, 40). The *Marnaviridae* endonuclease domain (MED) represents only the catalytic C-terminal domain of NucS-like

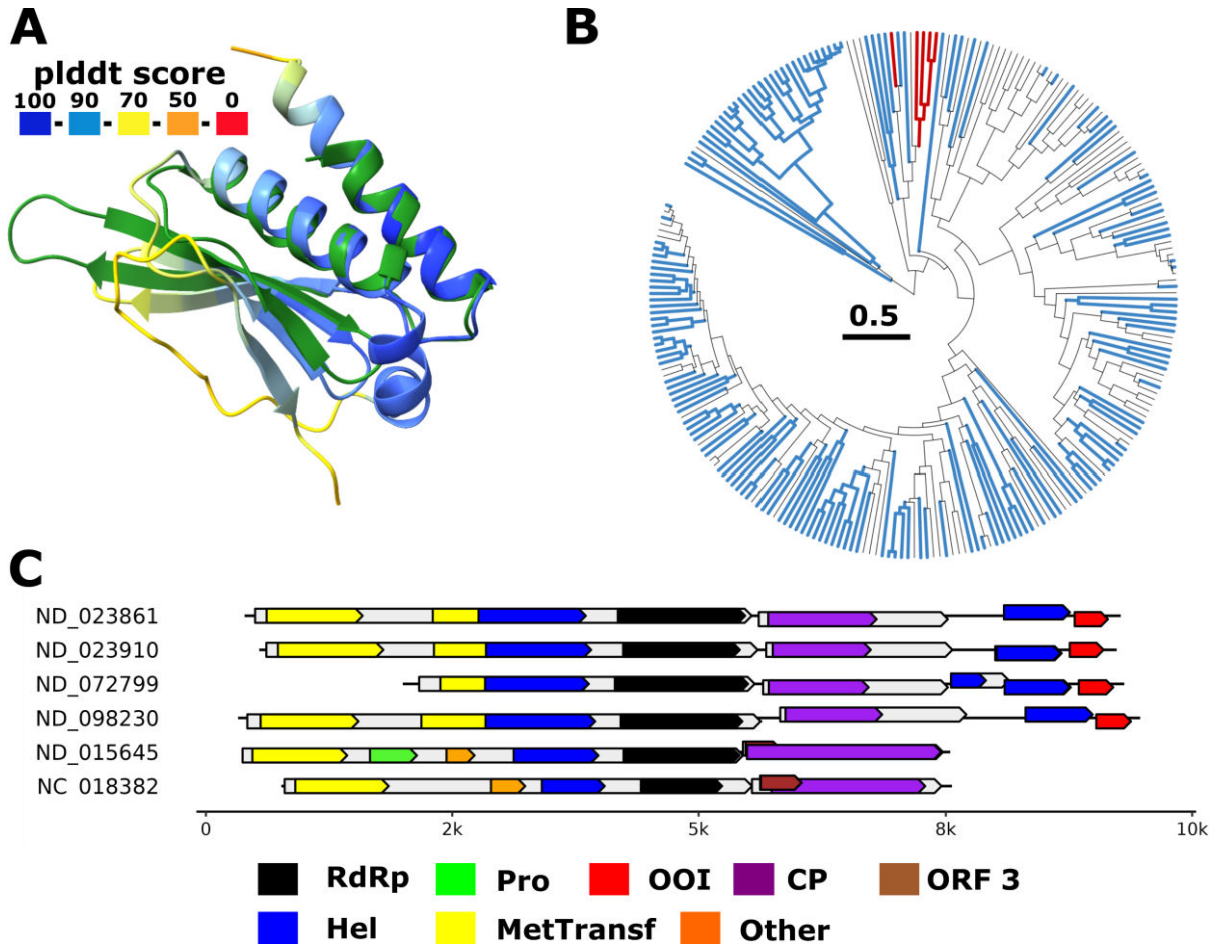

FIG 5 Inactive RNA-binding domain fold in *Hepeviridae*. (A) Superposition of likely inactive RNA-binding domain (RBD)-fold found in *Hepeviridae* (colored by plddt score) with RBD 2 from *A. thaliana* protein HYL1 (pdb 3adj, green, z-score 7.5). (B) Phylogenetic tree of *Hepeviridae* RdRp. Branches containing the OOI with RBD-fold are colored in red. Branches with no coding capacity after the capsid are colored in blue. (C) Genome maps for representative *Hepeviridae* members encoding (or not) for the OOI. Annotations are based on profile analysis (1) and GenBank annotation (NC 018382). Protein domains: RdRp, RNA-directed RNA polymerase, Hel, helicase, CP, capsid protein, Pro, protease, ORF3, *Hepeviridae* ORF 3, Other: additional *Hepeviridae* domains, OOI: ORFan of interest.

endonucleases ("DEK" motif) (41) but lacks the N-terminal dimerization domain (e.g., reference 41). The catalytic residues are conserved in MED, suggesting it is an active endonuclease (Fig. 7A and B). Typically, endonucleases with this fold cleave double-stranded or single-stranded DNA (41–43) suggesting that MED could target host DNA in marnavirus-infected cells. It remains unclear whether MED is proteolytically cleaved off the *Marnaviridae* polyprotein and functions as a distinct protein or remains fused to the RdRp domain. Contigs encoding MED are scattered across the RdRp tree of *Marnaviridae* suggesting spread via HGT and/or multiple losses of the MED domain (Fig. 7D).

In addition, we identified an exonuclease in the ~450 members-strong, distinct viral family *f.0181* in which the majority of viruses are likely hosted by protists that use alternative genetic codes (1). This family could not be assigned to known orders or classes in the phylum *Kitrinoviricota*. Profile annotation indicated the presence of the RdRp in these contigs but left a ~1,500 amino acid residues-long unannotated N-terminal region and a ~300 residues-long unannotated C-terminal region in the polyprotein. Dali search revealed a DEDD superfamily exonuclease domain located at the very N-terminus (Fig. S14; z-score: 10.6, e.g., RNase AS, a polyadenylate-specific exoribonuclease of *Mycobacterium tuberculosis*, which belongs to the DEDDh family within the DEDD superfamily). In the orthornavirus structurome, it showed structural similarities with coronavirus ExoN (z-score 8.6) which is involved in proofreading during RNA replication

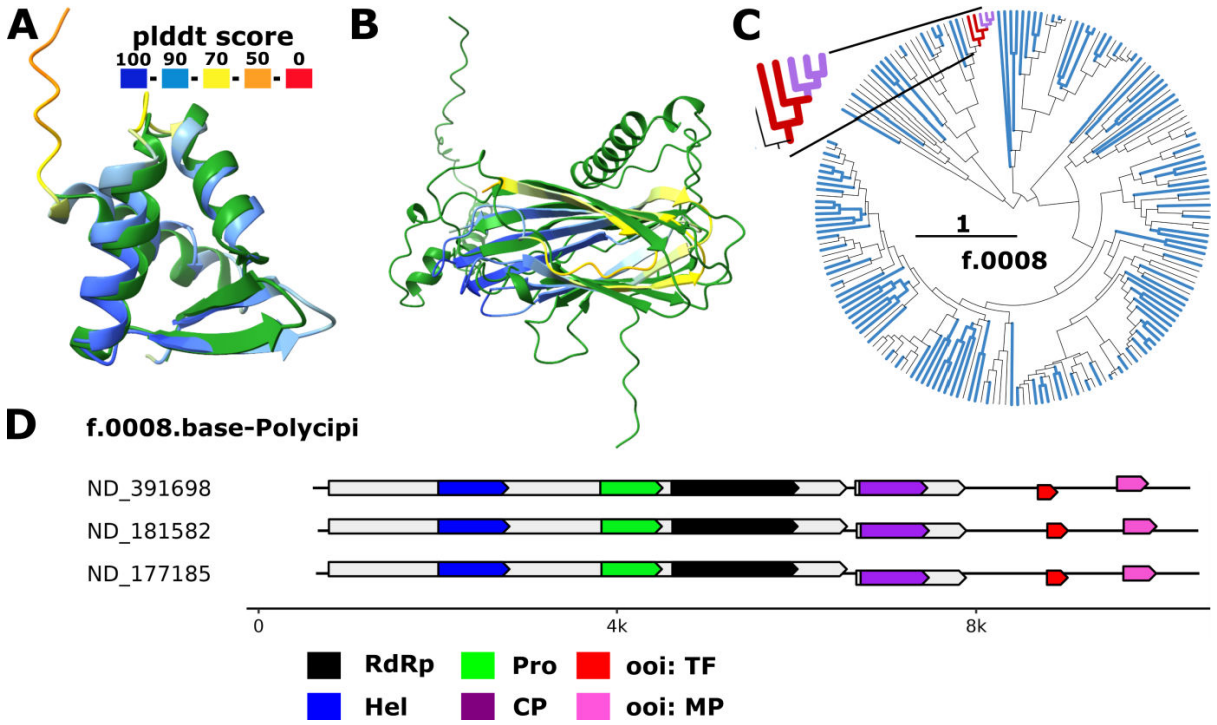

FIG 6 Winged helix-turn-helix domain and movement protein ORFans in a novel viral family. (A) Superposition of wHTH domain identified in *f.0008.base-Polycipi* (plddt score colored) with mouse HOP2 DNA-binding wHTH domain (2mh2, green, z-score: 11.1). (B) Superposition of predicted movement protein encoded by *f.0008* members (plddt colored) with an annotated movement protein domain from *Betaflexiviridae* (AlphaFold2 modeled, green, MP_30K, z-score: 7). (C) Phylogenetic distribution of contigs encoding only the predicted movement protein (pink) or both movement protein and wHTH (red). Blue branches indicate contigs with less than 180 nt after the capsid encoding ORF which are likely incomplete. (D) Representative genome maps for members of *f.0008* carrying the respective ORFs of interest (OOI). Protein domains: RdRp, RNA-directed RNA polymerase; Hel, helicase; CP, capsid protein; Pro, protease; OOI, ORF of interest (wHTH [red] or MP [pink]).

in some of these viruses possessing the largest known RNA genomes (44, 45), and the C-terminal nuclease domain of the nucleocapsid protein of mammarenaviruses (z-score 7.1) implicated in immune invasion (46–48). Given that the maximum genome size of *f.0181* members is only up to 7500 nt, it is unlikely that this exonuclease is involved in proofreading. Of note, the DEDDh catalytic site is modified to DEEEh in the *f.0185* exonuclease domain, a variation that, to our knowledge, has not been reported previously. Moreover, structurome comparison revealed a viral methyltransferase domain within the last third of the large N-terminal region (z-score 10–11). This domain is involved in virus RNA capping during replication, a function that is widespread in diverse members of this phylum (49). The C-terminal region was identified as an SJR capsid protein by Dali search and neighborhood analysis. With the identification of these three functional domains, the *f.0181* genome maps became less inscrutable.

A hydrolase domain

The *Solemoviridae* family (*Pisuviricota*) includes four genera of plant viruses (50) and roughly 2,000 newly discovered members found primarily in aquatic, soil, and invertebrate metatranscriptomes (1). A fraction of solemoviruses encode an OOI with a typical α/β hydrolase fold (Fig. 8). The top Dali hits for this OOI include different types of hydrolases (e.g., deacetylases and cutinases with z-scores of 6–7); hence, no clear target can be predicted. A phylogenetic distribution suggests the acquisition of this OOI in a larger clade of unclassified *Solemoviridae*, although it is not found across all genomes in this clade. Of note, we cannot detect related folds in other orthornaviruses indicating a unique acquisition of this fold or major fold-remodeling post-exaptation by

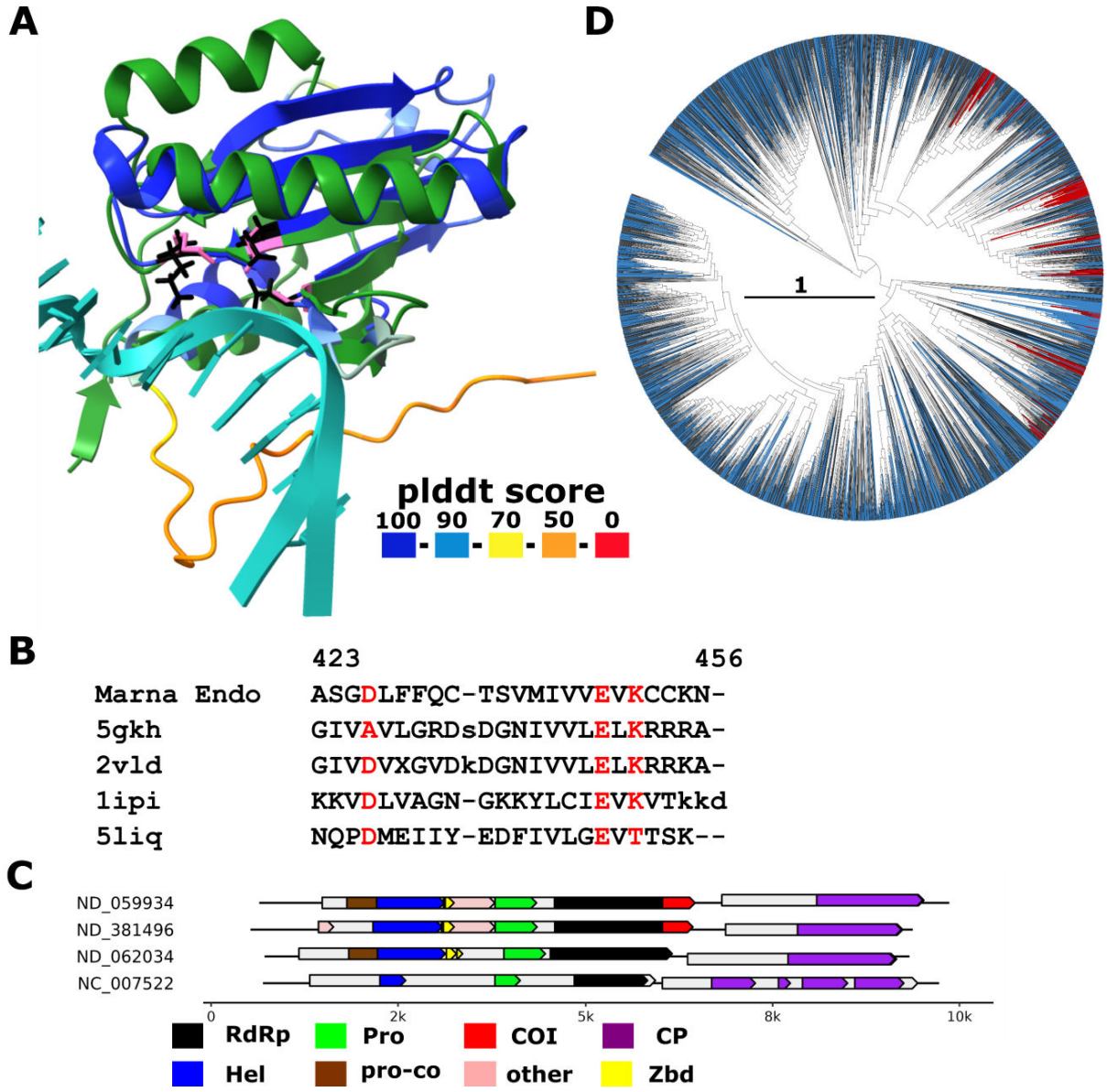

FIG 7 Endonuclease domain in *Marnaviridae*. (A) Superposition of representative *Marnaviridae* endonuclease domains (plddt score colored) with the C-terminal domain of endonuclease EndoMS (pdb 5gkh, aa 127-end, protein: green; DNA: light sea green; z-score 8.2). Catalytic residues of the nuclease are highlighted in pink for EndoMS (K181, E179, and D156A from left to right) and in black for *Marnaviridae* endonuclease (K69, E67, and D54). D156A is experimentally mutated in 5gkh to obtain the structure with uncleaved DNA. (B) Structure-guided alignment between *Marnaviridae* endonuclease and the four Dali top hits: endonuclease EndoMS (5gkh, Archaea, *Thermococcus kodakarensis* KOD1, z-score 8.2), NucS (2vld, Archaea, *Pyrococcus abyssi*, z-score 7.3), Holiday junction resolvase Hjc (1ipi, Archaea, *Pyrococcus furiosus*, z-score 6.7) and nicking endonuclease Nt.BspD6I (5liq, Bacteria, *Bacillus sp.*, z-score 6.1). (C) Genome maps of representative (nearly) complete *Marnaviridae* members from reference 1 and ICTV exemplar (NC_007522) which either contain (top two) or lack (bottom two) the endonuclease domain. Annotations are based on profile analysis (1) and GenBank annotation (NC_007522). Protein domains: RdRp, RNA-directed RNA polymerase, Hel, helicase, CP, capsid protein, Pro, protease, Pro-Co: protease cofactor_calici-como32k-like, Zbd, Zn-binding domain, COI: unannotated domain of interest (*Marnaviridae* endonuclease), other: other unannotated domain. (D) Phylogenetic tree of *Marnaviridae* RdRps (1); clades in which each leaf represents at least one contig that encodes an endonuclease are shown in red. Blue branches indicate contigs with less than 60 aa left unannotated C-terminal of the RdRp domain in the polyprotein.

Solemoviridae members. Given that none of the known plant solemoviruses encodes this hydrolase domain and that many solemoviruses were identified in plant-less aquatic environments, it seems likely that the host range of this virus family will be extended to additional host phyla.

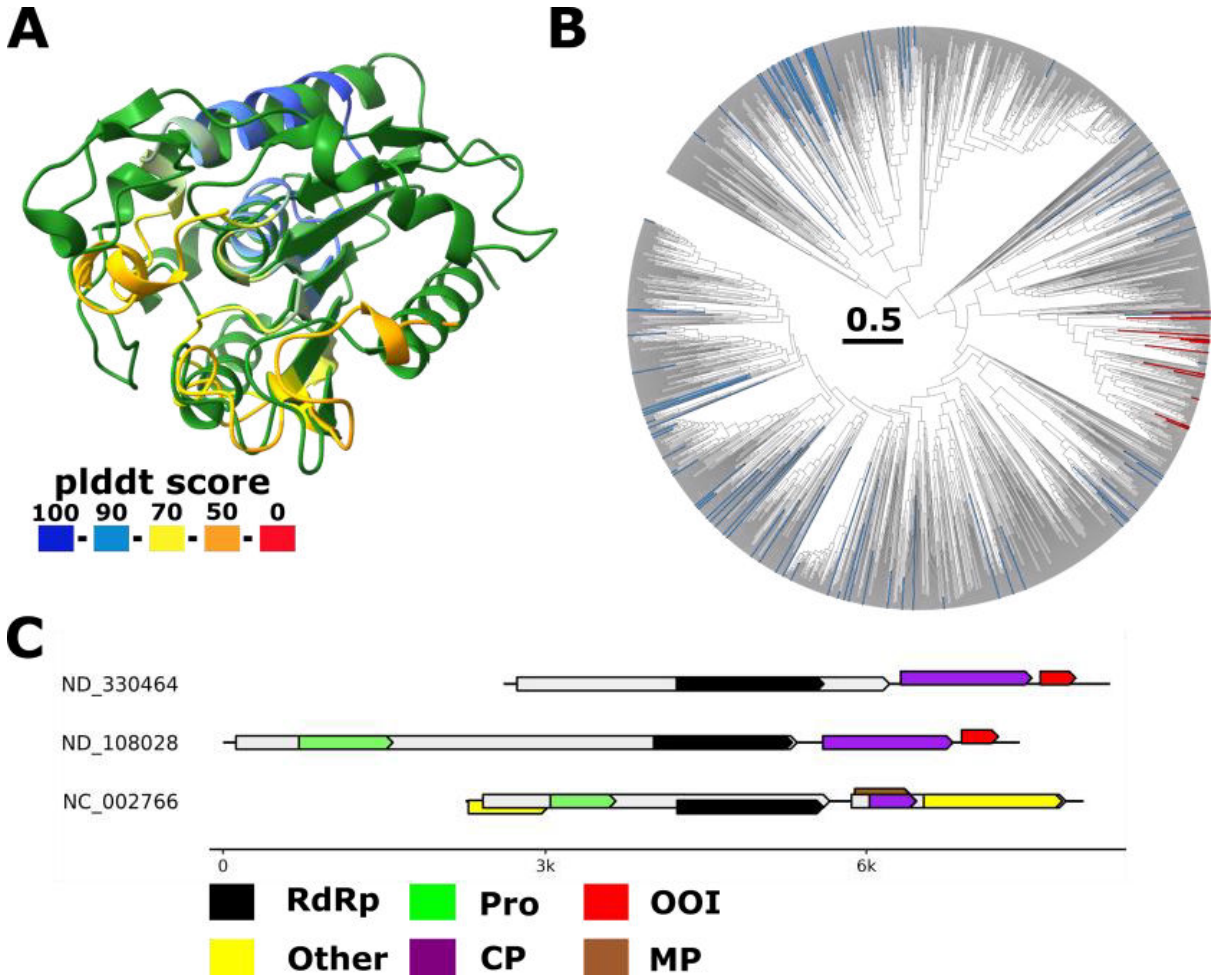

FIG 8 Hydrolase fold in *Solemoviridae*. (A) Superposition of putative hydrolase identified in *Solemoviridae* (colored by plddt score) with *Arabidopsis thaliana* SOBER1 deacetylase (pdb 6avw, green, z-score 7.0). (B) Phylogenetic tree in which leaves representing members encoding the putative hydrolase are colored red and leaves representing genomes with no coding capacity at the 3′ end for the putative hydrolase are colored blue (likely incomplete genomes). (C) Representative genome maps of *Solemoviridae* members. Annotations are on profile analysis (1) and GenBank annotation (NC_002766; start of RdRp encoding ORF at proposed frameshift leading to protease-RdRp polyprotein). Protein domains: RdRp, RNA-directed RNA polymerase, CP, capsid protein, Pro, protease, OOI, ORFan of interest, other: *Solemoviridae*-specific proteins p0 and p5.

Uncharacterized α/β-domains

In several cases, high-quality models were obtained for a COI, but there were no significant functionally characterized hits in structure comparisons. A case in point is the C-terminal domain of the *Secoviridae* (*Pisuviricota* [51]) polyprotein located immediately downstream of the RdRp (Fig. S15). Dali search showed mainly uncharacterized proteins, the closest being the N-terminal α/β-domain of TolB (z-score of 6.8) without a known function. No significant hits (z-score ≥ 5) were found for this COI in the orthornavirus structurome. The COI was only found in members of one *Secoviridae* genus, *Nepovirus*. Nepoviruses infect plants and are transmitted by beetles, aphids, whiteflies, leafhoppers, or nematodes (52). There are no indications that this C-terminal domain is cleaved from the RdRp, so further research is needed to decipher the role of this COI during nepovirus replication and host specificity.

Similarly, an OOI encoded between the matrix and the glycoprotein in *Rhabdoviridae* (*Negarnaviricota* [53]) members was predicted to adopt an α/β fold with 4 β-strands and 2 α-helices but without conclusive Dali hits (Fig. S16). Orthornavirus structurome comparison showed similarity to another hypothetical protein of *Rhabdoviridae*

(OA33_gp5 from NC_025382.1, a *Betapaprhavirus* member; encoded between G an L; z-score: 6.1). The function of this OOI remains unclear as is the case for other additional smaller *Rhabdoviridae* proteins that are encoded, for example, by members of the genus *Ephemerovirus* (NC 028241; Fig. S16C) (54).

A papain-like protease

A viral OTU (vOTU) domain, a papain-like-fold thiol protease (55), was identified in several families basal to *Deltaflexiviridae* (*Tymovirales*, *Kitrinoviricota*) family of fungal viruses (*f.0208, f.0210,* and *f.0212*; Fig. S17). Previously, a vOTU protease was identified in some members of *Tymovirales* (55), and subsequently, in *Deltaflexiviridae* (1), but not in the newly discovered basal provisional families. The present structure comparison confidently extended the spread of the vOTU protease to these families. The catalytic Cys-His dyad is conserved in all these viral proteins indicating that they are active proteases (Fig. S17A). In plant and animal viruses, vOTU domains function as deubiquitinylases implicated in immune evasion (56), and a similar role can be envisioned for the vOTUs of *Deltaflexiviridae* and their basal relatives.

An SJR domain

In the same vein, we identified the SJR capsid protein in the virus family *f.0198* of the same order, *Tymovirales* (Fig. S18). Similar capsid proteins were also found for the families *f.0194, f.0218,* and *f.0217*. Based on their overall relationship to other members of *Tymovirales*, identification of capsid proteins in these families should have been expected, but profile comparison failed to detect these proteins due to sequence divergence.

ORFan refinements

Apart from the discovery of new domains, ORFan refinements were mainly achieved for capsid proteins, for example, for nine distinct OOI representatives in *Leviviricetes*, previously unannotated capsid proteins homologous to the typical levivirus capsid proteins were identified across various families such as *Steitz-*, *Atkins-*, *Blume-*, *Solspi-*, *Fiers-*, and *Duinviridae* but also for related putative new families in which no capsid has been so far reported (e.g., *f.0361.base-Solspi* and *f.0367.base-Atkins*). Other refinements included *Rhabdoviridae* matrix protein, phytoreovirus core-P7 dsRNA-like-binding domain in *Cystoviridae,* methyltransferases, E proteins or RNAi suppressor proteins in *Arteriviridae*, *Flaviviridae*, and *Dicistroviridae* and *Mymonaviridae,* respectively.

DISCUSSION

Reliable modeling of protein structure is now available through various methods and techniques, such as AlphaFold2 and 3, RosettaFold, ESMFold, and more, and is widely applied for large-scale prediction of protein structures and functions. These analyses produced a variety of structural databases (19, 20) and many studies applying structure prediction for the exploration of evolutionary relationships among proteins and functional annotation of various genomes (19, 23, 51, 57, 58). To our knowledge, however, there were no large-scale databases available for virus-encoded proteins until very recently. For example, as of June 10th, 2024, the EBI AlphaFold repository dismissed viral proteins until computational polyprotein processing would improve (https://alphafold.ebi.ac.uk/ [20]). Only when this paper was in preparation, Kim et al. published a Foldseek database for viral representatives of UniRef30 clusters (22) and Nomburg et al. presented a study on predicted virus protein structures (21). Here, we sought to start closing this gap by modeling the structures of the proteins from 498 orthornavirus families using AlphaFold2. The majority of unannotated domains and putative ORFs yielded low-quality models which could be expected given that unannotated ORFs and protein regions are a highly heterogeneous set that includes variable sequences, linkers, and intrinsically disordered protein-protein interaction interfaces. Furthermore,

some of the smaller ORFans could have been acquired recently or evolved *de novo*, are virus specific, and are unlikely to have a large footprint in the databases used for model prediction. This would lead to alignments comprising low sequence diversity and therefore less confident structure predictions. Nevertheless, we identified a large set of high-quality models for both CUDs and ORFans many of which could be linked to structures from proteins with known functions.

The spread of almost all of the predicted new structures and functions was limited to a single virus family, and typically, only to a subset of its members. Thus, the general conclusion from this work is that the current catalog of widespread protein domains of orthornaviruses is effectively complete, even for novel groups known only from metatranscriptome mining. The dark matter of the orthornavirus pan-proteome consists mostly of all-α-helical and intrinsically disordered domains and proteins that are not readily amenable to structure modeling and comparison. This seems to be independent of the method chosen for structure modeling as we observed a preponderance of similar, simple folds when predicting ORFan structures using ESMFold (19), a large language model which does not rely on protein alignments (ESMFold predictions of ORFans can be found together with the supplemental material). This conclusion on the near saturation of the orthornavirus domain repertoire contrasts the ever-expanding diversity of the RdRps that shows no signs of saturation (1–4, 59). This difference reflects the strong constraints on the size of RNA genomes that limit the potential for new gene capture, in sharp contrast to viruses with large DNA genomes.

The general conclusion on the limited domain repertoire of orthornaviruses notwithstanding, the globular domains that we identified here do show certain trends. Specifically, some of these are predicted nucleic acid-binding domains and nucleases that could be involved in viral interference with host-specific immune systems. Typically, the ORFans with predicted new structures and functions are not found in a single viral clade but rather are scattered over the evolutionary trees of the respective viral families. This pattern is likely to reflect the dynamic evolution of these relatively recently captured genes involving multiple HGT events as well as gene loss. Generally, uncharacterized regions of known viral (poly)proteins in which globular domains were predicted span broader taxonomic ranges of viruses than ORFans. This difference seems plausible because the acquisition or *de novo* emergence of a new small ORF, while maintaining a replication-competent virus, appears to be more likely than the insertion of a new domain inside a large virus protein.

Somewhat serendipitously, during this structural analysis of the uncharacterized proteome of orthornaviruses, we identified a notable case of exaptation of a host enzyme, NMP kinase, for dsRNA-binding function. This protein is widely spread across diverse orthornaviral families suggestive of an important role(s) in virion morphogenesis but, possibly, also in suppression of host immunity. Exaptation of enzymes for structural roles seems to be a common theme in the evolution of large viruses with dsDNA genomes, such as poxviruses (60), but is less frequent in orthornaviruses (61) and other viruses with small genomes. More generally, exaptation of the host- and virus-encoded proteins is a leading trend in the evolution of viruses (62).

The expansion of the RNA virosphere via metatranscriptome mining is ongoing at an accelerating pace, and novel viruses can be confidently expected to emerge for many years to come, at least, at the levels of family and below. Characterization of functional domains, including the identification of novel ones specific to a few or individual families, can be performed by extensive profile comparison (1) and complemented by protein structure comparison. A recent expansive prediction of RNA viruses in metatranscriptomes using artificial intelligence approaches provides ample material for the discovery of such domains that are narrowly distributed but inform our understanding of virus biology (63). Thus, the pangenome and structurome of orthornaviruses constructed in this work, along with advancing tools for protein structure modeling and comparison, should be helpful to researchers investigating the structural and functional diversity of the RNA virosphere.

Conclusions

The current analysis of the orthornavirus pan-proteome and its dark matter using structure prediction and comparison methods suggests that all broadly conserved proteins and domains encoded by viruses of this kingdom are already known. The proteomic dark matter seems to consist largely of disordered and all-α-helical proteins that cannot be readily assigned a specific function. It appears likely that these domains mediate various interactions between viral proteins and between viral and host proteins but further experimental study is required to uncover their function. Nevertheless, we also identified a substantial number of globular domains that have not been reported previously. These domains are primarily encoded by ORFans, which are present only in narrow groups of orthornaviruses or are scattered across several such groups. With the accelerating discovery of orthornaviruses in metatranscriptomes, protein structure modeling, and analysis is now the approach of choice for the characterization of lineage-specific viral proteins and domains. The orthornavirus structurome constructed in this work can be expected to facilitate such studies.

MATERIALS AND METHODS

ORF and domain identification

Virus contigs and protein annotations were retrieved from the data deposited by Neri et al. (1) (DOI:10.5281/zenodo.7368133), named EMRV set (environmental metatranscriptome RNA viruses) hereafter. ORF boundaries were identified by running emboss getorf (minimal number of nucleotides: 150, stop-stop to allow for the identification of incomplete ORFs in incomplete genomes) with the standard genetic code or the code indicated in reference 1. Putative start codons were identified within the extracted ORFs and assumed to be ORF starts unless the ORF was located at the very 5′ end of the contig (likely incomplete assembly) or overlapped with an annotation from reference 1 (see below, possibility of programmed frame-shift). Annotations from reference 1 were mapped back to the ORFs. An ORF (and its protein sequence) was called fully annotated if it contained no continuous unannotated stretches of 60 aa or more. Otherwise, the annotated part was extracted as a domain, retrieving in a total of 647,383 annotated ORFs and domains.

To identify conserved unannotated domains (CUD) and ORFs, a pipeline was run as follows: At the first step, all proteins larger than 200 aa with partial or no annotation (297,411 proteins, with at least one stretch of at least 60 aa unannotated) were clustered and aligned using the snakemake pipeline (64) (Suppl. file "protein-clustering-diamond-mcl_full.smk" ran with snakemake --config input = proteins.faa precluster_min_seq_id = 0.95 diamond_min_seq_id = 0.0 min_aln_cov = 0.9 mcl_inflation = 2.5 j 16 s protein-clustering-diamond-mcl_full.smk). In detail, sequences were pre-clustered with mmseqs2 (mmseqs easy-linclust --kmer-per-seq 100 c 1.0 --cluster-mode 2 --cov-mode 1 --min-seq-id 0.95) (65) and representatives were searched against each other using diamond (diamond blastp -e 1e-3 --very-sensitive --id 0.0) (66). Pairs with at least 90% coverage were identified and clustered with mcxload (mcxload -abc {input} --stream-mirror --stream-neg-log10 -stream-tf 'ceil(300)' -o {output[0]} -write-tab {output[1]}) (67) and mcl (mcl {input[0]} -use-tab {input[1]} -o {output} -te {threads} -I 2.5) (67). Protein clusters including five or more sequences were kept to focus on domains from abundant proteins. Clusters were aligned with mafft (68) and written to a file with seqkit (69) (mafft --quiet --anysymbol --thread 4 --auto {input} | seqkit seq -w 0 > {output}), resulting in 2,410 alignments spanning 31,041 sequences. These alignments were used to conduct a first iteration of hhblits (70) to identify known protein domains. First, fasta alignments were converted to a3m alignments with hhconsensus (-M 50) (70) and searched against the following databases: pdb70 (71), pfama (72), scope70 (73, 74), ECOD (ECOD_F70_20220613) (75), and nvpc (1) (hhblits -cpu 1 -norealign -n 1 p 0.9 -z 0 -Z 5000 -b 0 -B 5000 -i {input} -o {output} -d {pdb70} -d {pfama} -d {scope70}

-d {ECOD} -d {nvpc}). Unannotated stretches of at least 60 aa in the alignment were extracted if not overlapping with an annotation (90% probability) by more than 10% (Suppl. File get_uncovered_coord_first_round.py and snakemake files snakemake_interdomain_p1.sh and snakefile_interdomain_p1.smk). The extracted stretches of unannotated alignments were run with hhblits against the same databases with the same parameters and processed as in the first round to retrieve the final unannotated alignment stretches. This procedure resulted in 2,831 unannotated alignment stretches spanning 34,710 domains (31,247 larger than 60 aa which were kept as conserved unannotated domains [CUDs]).

To identify abundant unannotated ORFans between 60 and 200 aa in size, only ORFs between start and stop codons were considered (no programmed frameshifts or incomplete ORFs at the 5′ end of the genome lacking the start codon included). In the EMRV set, 1,363,871 such unannotated ORFs were detected including those located on either the forward or the reverse strand and those nested with other ORFs. ORFans were clustered together with CUDs using mmseqs2 (-min-seq-ident 0.4 and -c 0.85). Only clusters with at least one CUD or at least 5 ORFans were considered, resulting in 59,815 clusters spanning 655,787 ORFans and 13,099 clusters consisting of CUDs including 325 clusters that included 1052 ORFans together with CUDs. Representatives from these clusters were kept for protein structure prediction.

Processing of ICTV exemplars

To obtain a reference set of functional domains represented in each virus family, reference virus genomes were downloaded from ICTV exemplars for all *Orthornavirae* (https://ictv.global/vmr, VMR_MSL38_v2, 1 December 2023). GenBank files of viral genome were retrieved from NCBI for all ICTV-approved virus families in the EMRV set (98 virus families). Proteins and domains were extracted as annotated within the genome GenBank file. To account for a more fine-grained annotation of individual proteins, GenBank files for individual proteins were retrieved whenever the protein in the genome GenBank file was flagged as polyprotein or was at least 500 amino acids long. Again, individual domains were isolated and mapped to the corresponding virus family (32,648 domains in total). To compare the GenBank annotation with our profile-based annotation, isolated ICTV exemplar domains were run against the viral profile database from reference 1 (nvpc db) using hhsearch with default settings. Hits with 95% or higher probability were harmonized by taking the most prevalent function. All domains were then clustered using mmseqs2 (min-seq-id 0.4, coverage 0.85; 9245 clusters), and profile hits were harmonized across all members of a cluster. The most frequent function was then assigned to all members of a cluster. Clusters that lacked at least one highly significant profile hit were inspected for GenBank annotation of the cluster members (4,697 sequences across 3,163 clusters). Those GenBank annotations were harmonized by keyword (e.g., "hypothetical," "unknown," and "putative protein" were all denoted "hypothetical") and the most frequent label was assigned to the cluster. The clusters were then assigned to the respective virus families and profile annotation coverage per virus family was calculated (Fig. S2). This procedure was performed for clusters harboring either all domains or domains encoded on the RdRp encoding segment for viruses with segmented genomes.

Building pangenomes of orthornavirus families

Annotated domains from the EMRV set were mapped to the ORFs and assigned to their respective virus families. Nucleotide and amino acid positions were orientated such that the RdRp encoding ORF is on the forward strand on each genome. Next, CUDs were mapped to the respective genomes. As CUDs were retrieved by a slightly different method than the annotation in reference 1, 10 aa overlaps between CUDs and annotated domains were permitted.

Protein structure prediction, analysis, comparison and visualization

Protein structures were modeled using AlphaFold2 (version 2.3.1, default parameters) for all proteins and domains of orthornaviruses (annotated, unannotated, and ICTV exemplars) except ORFans (all Prodigal annotated viral proteins from reference 1 were added to the uniref90 database to improve alignments) on the high-performance biowulf cluster at NIH. ORFans were modeled using ESM-fold (version 2.0.0) because of the prohibitive computational cost of modeling such a large number of sequences with AlphaFold2. Model quality was assessed by the plddt score. Protein structures were clustered with Foldseek (-c 0.8) (76). Protein structures of selected representatives of CUD and ORFan clusters with a plddt score of 70 or higher were searched against a local version of pdb70 using Dali (26). Protein secondary structure was assessed by psique (27) for all CUDs and ORFan structures and based on the Dali search files for representative structures ran against PDB (an individual helix was called if at least 10 consecutive amino acids were assigned as helix and a beta-strand was called if at least three consecutive amino acids were assigned as beta-strand member). Protein structures were visualized with Chimera X (version 1.3) (77).

The orthornavirus "structurome"

To obtain the orthornavirus "structurome," one representative protein of each functional tag per virus family was taken from the *Riboviria* pangenome and ICTV exemplar domains lacking confident annotation, the structures of these representatives were predicted as described above and combined with representative structures of OOIs and COIs. All 7,996 structures were pre-clustered with Foldseek (-c 0.8) and 5,528 representatives were run all-vs-all using Dali (default parameters). Clusters were identified using an iterative process in which, first, all closest related pairs across the all-vs-all matrix were identified (based on their z-score, minimal z-score here 7, "z7 cluster") as preliminary clusters. Then, the next closest related pairs were identified. Whenever the members of a pair belonged to different clusters, such clusters were merged if the mean inter-cluster z-score was 7 or higher. In case only one member was part of a cluster, the second one was added if the mean z-score to all members in a cluster was 7 or higher. In case none of the members of a pair belonged to a cluster, a new preliminary cluster was created. The iterations stopped when no new members could be added, and no clusters could be merged.

A structure similarity network was constructed based on Dali z-scores of the all-vs-all comparison to visualize the structure relationships. All structures present in a z7 cluster were considered. Structures outside of such clusters with a z-score of 7 or higher to any structure within a cluster were also included. Each structure was a node and each edge linking two nodes was weighted by the z-score. Node and edge tables were loaded into Gephi (78) (version 0.10.1) and arranged by running the openOrd algorithm with default parameters (stages and their percentages: liquid [25%], expansion [25%], cooldown [25%], crunch [10%], and simmer [15%]) and other parameters as follows: Edge cut: 0.8, Nin threads: 5, Num iterations: 750, fixed time: 0.2, random seed: 2644300876718762555), followed by a round noverlap (speed: 3, ratio: 1.2, margin:2).

Provisional annotation of CUDs and ORFans

All initially extracted 31,247 CUDs were run against a blast database of all annotated domains from the EMRV set as well as all ICTV exemplar domains using PSI-BLAST (79) (default parameters, eval 0.0001). Annotations were mapped on the 13,085 CUD mmseqs clusters and the cluster representative provisionally labeled if at least 50% of the members produced a hit; the hits were then harmonized. If no harmonization was possible, the hit with the lowest e-value was taken. Through this procedure, 37% of the 13,085 CUD mmseqs representatives were annotated.

ORFans and CUDs were then clustered together with ICTV exemplar domains using mmseqs2 (min-seq-ident 0.4, coverage 0.85) and clusters were aligned with mafft (68)

(default parameters). Aligned clusters were searched against various databases (ECOD [75], scope70 [73, 74], pfam [72], ncbi CD [80], nvpc [1], virusDB2020 [81]) using hhblits (70) (default parameters). Clusters that produced hits with a probability of 90% or higher were provisionally annotated. The ICTV exemplar domain annotation was considered for a cluster whenever such a domain was present and no hhblits hit with a probability of 90% or higher was obtained. About 1% of the clusters were provisionally annotated.

Identification of CUDs of interest

All 13,000 CUD models were filtered by their mean plddt score (70 or higher) and secondary structure. To be kept, a CUD had to encompass at least three helices or one helix and one beta-strand or more than two beta strands based on the Dali secondary structure assignment, where at least 10 consecutive "helix" assigned amino acids counted as a helix and at least three consecutive "beta-strand" assigned amino acids counted as a beta-strand. This procedure excludes single helices, helix-turn-helix folds, beta-hairpins, and disordered proteins, resulting in 830 CUDs of interest (COI). Then, COIs were compared to the N- and C-terminal annotated domains in the same genome to see whether the COI might be an extension of the existing annotation. To determine whether a COI was annotated already in a related protein, genomic neighborhood analysis was performed: similar proteins with and without the COI annotation were aligned using mafft (68) (default parameters), and annotated domains and COI positions of each included protein were mapped to the alignment. If the COI region overlapped by 50% or more with any annotated domain of another protein in the alignment, the provisional COI PSI-BLAST annotation (if any, see above) was compared with that of the profile-annotated domain. If there was no conflict, the COI was called "resolved by neighborhood." In case of a conflict, the Dali structure comparison results for the given COI were manually inspected. If there was no overlap by at least 50%, the COI was kept for manual inspection. Only COIs with a top Dali z-score of 4 or higher were inspected. In addition, all-α-helical proteins were only inspected if the top z-score was above 10 given that all-α-helical proteins tend to produce inconclusive results because helices in many different types of proteins are hit. None of the all-α-helical COIs with a top z-score above 10 produced a conclusive hit. Furthermore, randomly checked all-α-helical COIs with lower top z-scores could not be assigned to specific folds. The remaining COI structures with a Dali top z-score of 4 or higher were inspected with respect to (i) the respective Dali pdb hits, (ii) their position in the genomes of a virus family, (iii) their relationship to other structures in the structurome; and (iv) their provisional hhblits annotation, if any.

Identification of ORFans of interest

The modeled ORFans were clustered with Foldseek (-c 0.8) and inspected by their plddt score. ORFans present in a cluster with a plddt score of 70 or higher were further analyzed (5,944/59,185 ORFan structures). Structure representatives were run against a local version of pdb70 using Dali. As for CUDs, structures with "simple" folds were excluded (see above), leaving about 1,600 ORFans of interest (OOIs). As for COIs, OOIs were manually inspected whenever they were not all-α-helical. The remaining COI structures with a Dali top z-score of 6 or higher were inspected with respect to (i) the respective Dali pdb hits, (ii) their position in the genomes of a virus family, (iii) their relationship to other structures in the structurome, and (iv) their provisional hhblits annotation, if any.

Detection of intrinsically disordered regions

The fraction of intrinsically disordered regions in proteins was analyzed using a local version of MobiDB-lite (version 3.10.0) (82).

Phylogenetic distribution of OOIs and COIs

For each virus family, representative OOIs or COIs of the same function were aligned with their Foldseek and mmseqs2 cluster members using muscle (83) (version 5.1, default parameters). Aligned sequences were searched using PSI-BLAST (84) (version 2.15) (psiblast -in_msa ip -threshold 9 db db -evalue 0.01 -out op -num_threads 8 -max_target_seqs 50000 -outfmt 6) against all extracted ORFs (annotated or not) from the EMRV set. All regions covered were extracted, realigned together with the initial sequences, and ran again against all ORFs using PSI-BLAST with the same parameters. The final sequence sets were mapped back to the respective virus family phylogeny obtained from reference 1. Each leaf covering at least one genome containing the respective OOI or COI was called positive. The leaves in each tree represent clusters of RdRp core sequences at 95% sequence identity (1). Trees were visualized using iTol (v6) (85).

Visualization of viral genomes

Representative viral genome maps were generated with R using the libraries ggplot2 and gggenomes (https://github.com/thackl/gggenomes).

Structure-guided alignment of core-P7 RBD domains and cellular kinases and tree construction

Representative structures of viral core-P7 RBD domains and related cellular kinases were aligned using the FoldMason web server (29). The alignment was trimmed to remove columns with more than 35% gaps, and a phylogenetic tree was built using IQtree2 with the implemented modelfinder (-m MFP [86]) and ultrafast bootstrapping (-B 10000 [87, 88]).

Sampling of unique domains across *Orthornavirae*

Increasing numbers of clusters of similar genomes (based on 90% RdRp aa identity, each cluster corresponding to a leaf in the virus family phylogeny) were sampled randomly across all *Orthornavirae* (step size: 50 clusters; till all clusters were sampled) and the number of distinct domains was extracted. This procedure was bootstrapped 30 times and the mean, minimal, and maximal number of unique domains were plotted as a function of the number of sampled clusters.

Transmembrane domain prediction

TMHMM 2.0 (89, 90) was used to predict transmembrane domains in CUDs and ORFans.

Defining the "core" set of unique domains per virus family

Distribution of associated contig lengths was obtained for each virus family; contigs of length of at least 2/3rd of the 75th percentile we operationally classified as "near full-length." For families with at least 10 near full-length contigs, the full set of identified structural domains was identified along with their frequencies. All domains were ranked by their frequencies; for domains, present in at least 50% of the near full-length contigs, the "frequency gap" (difference between its frequency and that of the next-ranked domain) was calculated; the domain with the highest frequency gap was defined as the last core domain.

ACKNOWLEDGMENTS

P.M., H.S., Y.I.W., and E.V.K. are supported by the Intramural Research Program of the National Institutes of Health (National Library of Medicine). This work utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov). V.V.D.

research was supported in part by an appointment to the National Center for Biotechnology Information Scientific Visitors Program administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and the National Institute of Health. ORISE is managed by ORAU under DOE contract number DE-SC0014664. The work conducted by the US Department of Energy Joint Genome Institute (https://ror.org/04xm1d337) and the National Energy Research Scientific Computing Center (https://ror.org/05v3mvq14) is supported by the US Department of Energy Office of Science user facilities, operated under contract no. DE-AC02-05CH11231.

All opinions expressed in this paper are the author's and do not necessarily reflect the policies and views of NIH, NCBI, DOE, or ORAU/ORISE. A.B. is supported by a postdoctoral fellowship from the Foundation pour la Recherche Mèdicale (grant number SPF202110014092).

P.M. and E.V.K. conceptualized the project; P.M., A.P.C., H.S., A.B., U.N., and Y.I.W. developed the methodology and collected the data; P.M. performed the research; P.M., M.K., V.V.D., and E.V.K. analyzed the data; P.M. and E.V.K. wrote the manuscript which was read, edited, and approved by all authors.

AUTHOR AFFILIATIONS

[1]Division of Intramural Research, Computational Biology Branch, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA
[2]Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, California, USA
[3]Institut Pasteur, Université Paris Cité, CNRS UMR6047, Archaeal Virology Unit, Paris, France
[4]Department of Botany and Plant Pathology, Oregon State University, Corvallis, Oregon, USA

AUTHOR ORCIDs

Pascal Mutz http://orcid.org/0000-0002-0430-2095
Uri Neri http://orcid.org/0000-0003-0894-2484
Anamarija Butkovic http://orcid.org/0000-0002-1435-0912
Yuri I. Wolf http://orcid.org/0000-0002-0247-8708
Mart Krupovic http://orcid.org/0000-0001-5486-0098
Valerian V. Dolja http://orcid.org/0000-0002-8148-4670
Eugene V. Koonin http://orcid.org/0000-0003-3943-8299

AUTHOR CONTRIBUTIONS

Pascal Mutz, Conceptualization, Formal analysis, Investigation, Writing – original draft, Writing – review and editing | Antonio Pedro Camargo, Data curation, Formal analysis | Harutyun Sahakyan, Formal analysis, Investigation | Uri Neri, Data curation | Anamarija Butkovic, Formal analysis | Yuri I. Wolf, Conceptualization, Formal analysis, Investigation, Methodology | Mart Krupovic, Formal analysis, Writing – review and editing | Valerian V. Dolja, Conceptualization, Formal analysis, Writing – original draft | Eugene V. Koonin, Conceptualization, Formal analysis, Funding acquisition, Investigation, Project administration, Supervision, Writing – original draft, Writing – review and editing

DIRECT CONTRIBUTION

This article is a direct contribution from Valerian V. Dolja, a Fellow of the American Academy of Microbiology, who arranged for and secured reviews by Andrew Lang, Memorial University of Newfoundland Department of Biology, and Andrew Firth, University of Cambridge.

DATA AVAILABILITY

This work is based on the analysis of genomes publicly available in GenBank. All other data generated by this analysis are contained in the supplemental material or are publicly available at zenodo (https://doi.org/10.5281/zenodo.13770614).

ADDITIONAL FILES

The following material is available online.

Supplemental Material

Supplemental figures (mBio03200-24-s0001.pdf). Figures S1 to S19.

REFERENCES

1. Neri U, Wolf YI, Roux S, Camargo AP, Lee B, Kazlauskas D, Chen IM, Ivanova N, Zeigler Allen L, Paez-Espino D, et al. 2022. Expansion of the global RNA virome reveals diverse clades of bacteriophages. Cell 185:4023–4037. https://doi.org/10.1016/j.cell.2022.08.023
2. Zayed AA, Wainaina JM, Dominguez-Huerta G, Pelletier E, Guo J, Mohssen M, Tian F, Pratama AA, Bolduc B, Zablocki O, et al. 2022. Cryptic and abundant marine viruses at the evolutionary origins of earth's RNA virome. Science 376:156–162. https://doi.org/10.1126/science.abm5847
3. Edgar RC, Taylor B, Lin V, Altman T, Barbera P, Meleshko D, Lohr D, Novakovsky G, Buchfink B, Al-Shayeb B, Banfield JF, de la Peña M, Korobeynikov A, Chikhi R, Babaian A. 2022. Petabase-scale sequence alignment catalyses viral discovery. Nature New Biol 602:142–147. https://doi.org/10.1038/s41586-021-04332-2
4. Lauber C, Seitz S. 2022. Opportunities and challenges of data-driven virus discovery. Biomolecules 12:1073. https://doi.org/10.3390/biom12081073
5. Bukhari K, Mulley G, Gulyaeva AA, Zhao L, Shu G, Jiang J, Neuman BW. 2018. Description and initial characterization of metatranscriptomic nidovirus-like genomes from the proposed new family abyssoviridae, and from a sister group to the coronavirinae, the proposed genus alphaletovirus. Virology (Auckl) 524:160–171. https://doi.org/10.1016/j.virol.2018.08.010
6. Saberi A, Gulyaeva AA, Brubacher JL, Newmark PA, Gorbalenya AE. 2018. A planarian nidovirus expands the limits of RNA genome size. PLoS Pathog 14:e1007314. https://doi.org/10.1371/journal.ppat.1007314
7. Neuman BW, Smart A, Vaas J, Bartenschlager R, Seitz S, Gorbalenya AE, Caliskan N, Lauber C. 2024. RNA genome expansion up to 64 kb in nidoviruses is host constrained and associated with new modes of replicase expression. bioRxiv. https://doi.org/10.1101/2024.07.07.602380
8. Petrone ME, Grove J, Mélade J, Mifsud JCO, Parry RH, Marzinelli EM, Holmes EC. 2024. A ~40-kb flavi-like virus does not encode a known error-correcting mechanism. Proc Natl Acad Sci U S A 121:e2403805121. https://doi.org/10.1073/pnas.2403805121
9. Koonin EV, Dolja VV, Krupovic M, Varsani A, Wolf YI, Yutin N, Zerbini FM, Kuhn JH. 2020. Global organization and proposed megataxonomy of the virus world. Microbiol Mol Biol Rev 84. https://doi.org/10.1128/MMBR.00061-19
10. Lauber C, Zhang X, Vaas J, Klingler F, Mutz P, Dubin A, Pietschmann T, Roth O, Neuman BW, Gorbalenya AE, Bartenschlager R, Seitz S. 2024. Deep mining of the sequence read archive reveals major genetic innovations in coronaviruses and other nidoviruses of aquatic vertebrates. PLoS Pathog 20:e1012163. https://doi.org/10.1371/journal.ppat.1012163
11. Butkovic A, Dolja VV, Koonin EV, Krupovic M. 2023. Plant virus movement proteins originated from jelly-roll capsid proteins. PLoS Biol 21:e3002157. https://doi.org/10.1371/journal.pbio.3002157
12. van den Born E, Omelchenko MV, Bekkelund A, Leihne V, Koonin EV, Dolja VV, Falnes PØ. 2008. Viral AlkB proteins repair RNA damage by oxidative demethylation. Nucleic Acids Res 36:5451–5461. https://doi.org/10.1093/nar/gkn519
13. Egloff MP, Malet H, Putics A, Heinonen M, Dutartre H, Frangeul A, Gruez A, Campanacci V, Cambillau C, Ziebuhr J, Ahola T, Canard B. 2006. Structural and functional basis for ADP-ribose and poly(ADP-ribose) binding by viral macro domains. J Virol 80:8493–8502. https://doi.org/10.1128/JVI.00713-06
14. Gan T, Wang D. 2023. Picobirnaviruses encode proteins that are functional bacterial lysins. Proc Natl Acad Sci U S A 120:e2309647120. https://doi.org/10.1073/pnas.2309647120
15. Binder JL, Berendzen J, Stevens AO, He Y, Wang J, Dokholyan NV, Oprea TI. 2022. AlphaFold illuminates half of the dark human proteins. Curr Opin Struct Biol 74:102372. https://doi.org/10.1016/j.sbi.2022.102372
16. Porta-Pardo E, Ruiz-Serra V, Valentini S, Valencia A. 2022. The structural coverage of the human proteome before and after AlphaFold. PLoS Comput Biol 18:e1009818. https://doi.org/10.1371/journal.pcbi.1009818
17. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. 2021. Highly accurate protein structure prediction with AlphaFold. Nature New Biol 596:583–589. https://doi.org/10.1038/s41586-021-03819-2
18. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD, et al. 2021. Accurate prediction of protein structures and interactions using a three-track neural network. Science 373:871–876. https://doi.org/10.1126/science.abj8754
19. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, Dos Santos Costa A, Fazel-Zarandi M, Sercu T, Candido S, Rives A. 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 379:1123–1130. https://doi.org/10.1126/science.ade2574
20. Varadi M, Bertoni D, Magana P, Paramval U, Pidruchna I, Radhakrishnan M, Tsenkov M, Nair S, Mirdita M, Yeo J, Kovalevskiy O, Tunyasuvunakool K, Laydon A, Žídek A, Tomlinson H, Hariharan D, Abrahamson J, Green T, Jumper J, Birney E, Steinegger M, Hassabis D, Velankar S. 2024. AlphaFold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. Nucleic Acids Res 52:D368–D375. https://doi.org/10.1093/nar/gkad1011
21. Nomburg J, Doherty EE, Price N, Bellieny-Rabelo D, Zhu YK, Doudna JA. 2024. Birth of protein folds and functions in the virome. Nature New Biol 633:710–717. https://doi.org/10.1038/s41586-024-07809-y
22. Kim RS, Karin EL, Steinegger M. 2024. BFVD - a large repository of predicted viral protein structures. bioRxiv. https://doi.org/10.1101/2024.09.08.611582
23. Mutz P, Resch W, Faure G, Senkevich TG, Koonin EV, Moss B. 2023. Exaptation of inactivated host enzymes for structural roles in orthopoxviruses and novel folds of virus proteins revealed by protein structure modeling. MBio 14:e0040823. https://doi.org/10.1128/mbio.00408-23
24. Krupovic M, Kuhn JH, Fischer MG, Koonin EV. 2024. Natural history of eukaryotic DNA viruses with double jelly-roll major capsid proteins. Proc Natl Acad Sci U S A 121:e2405771121. https://doi.org/10.1073/pnas.2405771121
25. Söding J. 2005. Protein homology detection by HMM-HMM comparison. Bioinformatics 21:951–960. https://doi.org/10.1093/bioinformatics/bti125
26. Holm L. 2020. Using dali for protein structure comparison. Methods Mol Biol 2112:29–42. https://doi.org/10.1007/978-1-0716-0270-6_3
27. Adasme-Carreño F, Caballero J, Ireta J. 2021. PSIQUE: protein secondary structure identification on the basis of quaternions and electronic

structure calculations. J Chem Inf Model 61:1789–1800. https://doi.org/10.1021/acs.jcim.0c01343
28. Nakashima K, Kakutani T, Minobe Y. 1990. Sequence analysis and product assignment of segment 7 of the rice dwarf virus genome. J Gen Virol 71 (Pt 3):725–729. https://doi.org/10.1099/0022-1317-71-3-725
29. Gilchrist CLM, Mirdita M, Steinegger M. 2024. Multiple protein structure alignment at scale with FoldMason. bioRxiv. https://doi.org/10.1101/2024.08.01.606130:2024.08.01.606130
30. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol 37:1530–1534. https://doi.org/10.1093/molbev/msaa015
31. Zhong BX, Shen YW, Omura T. 2005. RNA-binding domain of the key structural protein P7 for the Rice dwarf virus particle assembly. Acta Biochim Biophys Sin (Shanghai) 37:55–60. https://doi.org/10.1093/abbs/37.1.55
32. Doyle M, Jantsch MF. 2002. New and old roles of the double-stranded RNA-binding domain. J Struct Biol 140:147–153. https://doi.org/10.1016/s1047-8477(02)00544-0
33. Yang SW, Chen HY, Yang J, Machida S, Chua NH, Yuan YA. 2010. Structure of *Arabidopsis* HYPONASTIC LEAVES1 and its molecular implications for miRNA processing. Structure 18:594–605. https://doi.org/10.1016/j.str.2010.02.006
34. Waldron FM, Stone GN, Obbard DJ. 2018. Metagenomic sequencing suggests a diversity of RNA interference-like responses to viruses across multicellular eukaryotes. PLoS Genet 14:e1007533. https://doi.org/10.1371/journal.pgen.1007533
35. Bell CL, Gurda BL, Van Vliet K, Agbandje-McKenna M, Wilson JM. 2012. Identification of the galactose binding domain of the adeno-associated virus serotype 9 capsid. J Virol 86:7326–7333. https://doi.org/10.1128/JVI.00448-12
36. Lang AS, Vlok M, Culley AI, Suttle CA, Takao Y, Tomaru Y, Ictv Report C. 2021. ICTV virus taxonomy profile: marnaviridae 2021. J Gen Virol 102. https://doi.org/10.1099/jgv.0.001633
37. Dias A, Bouvier D, Crépin T, McCarthy AA, Hart DJ, Baudin F, Cusack S, Ruigrok RWH. 2009. The cap-snatching endonuclease of influenza virus polymerase resides in the PA subunit. Nature New Biol 458:914–918. https://doi.org/10.1038/nature07745
38. Ivanov KA, Hertzig T, Rozanov M, Bayer S, Thiel V, Gorbalenya AE, Ziebuhr J. 2004. Major genetic marker of nidoviruses encodes a replicative endoribonuclease. Proc Natl Acad Sci U S A 101:12694–12699. https://doi.org/10.1073/pnas.0403127101
39. Posthuma CC, Nedialkova DD, Zevenhoven-Dobbe JC, Blokhuis JH, Gorbalenya AE, Snijder EJ. 2006. Site-directed mutagenesis of the nidovirus replicative endoribonuclease NendoU exerts pleiotropic effects on the arterivirus life cycle. J Virol 80:1653–1661. https://doi.org/10.1128/JVI.80.4.1653-1661.2006
40. Yang L, He J, Wang R, Zhang X, Lin S, Ma Z, Zhang Y. 2019. Nonstructural protein 11 of porcine reproductive and respiratory syndrome virus induces STAT2 degradation to inhibit interferon signaling. J Virol 93:e01352-19. https://doi.org/10.1128/JVI.01352-19
41. Nakae S, Hijikata A, Tsuji T, Yonezawa K, Kouyama KI, Mayanagi K, Ishino S, Ishino Y, Shirai T. 2016. Structure of the endoMS-DNA complex as mismatch restriction endonuclease. Structure 24:1960–1971. https://doi.org/10.1016/j.str.2016.09.005
42. Ishino S, Nishi Y, Oda S, Uemori T, Sagara T, Takatsu N, Yamagami T, Shirai T, Ishino Y. 2016. Identification of a mismatch-specific endonuclease in hyperthermophilic archaea. Nucleic Acids Res 44:2977–2986. https://doi.org/10.1093/nar/gkw153
43. Ren B, Kühn J, Meslet-Cladiere L, Briffotaux J, Norais C, Lavigne R, Flament D, Ladenstein R, Myllykallio H. 2009. Structure and function of a novel endonuclease acting on branched DNA substrates. EMBO J 28:2479–2489. https://doi.org/10.1038/emboj.2009.192
44. Robson F, Khan KS, Le TK, Paris C, Demirbag S, Barfuss P, Rocchi P, Ng WL. 2020. Coronavirus RNA proofreading: molecular basis and therapeutic targeting. Mol Cell 79:710–727. https://doi.org/10.1016/j.molcel.2020.07.027
45. Minskaia E, Hertzig T, Gorbalenya AE, Campanacci V, Cambillau C, Canard B, Ziebuhr J. 2006. Discovery of an RNA virus 3'->5' exoribonuclease that is critically involved in coronavirus RNA synthesis. Proc Natl Acad Sci U S A 103:5108–5113. https://doi.org/10.1073/pnas.0508200103
46. Qi X, Lan S, Wang W, Schelde LM, Dong H, Wallat GD, Ly H, Liang Y, Dong C. 2010. Cap binding and immune evasion revealed by lassa nucleoprotein structure. Nature New Biol 468:779–783. https://doi.org/10.1038/nature09605
47. Hastie KM, Kimberlin CR, Zandonatti MA, MacRae IJ, Saphire EO. 2011. Structure of the lassa virus nucleoprotein reveals a dsRNA-specific 3' to 5' exonuclease activity essential for immune suppression. Proc Natl Acad Sci U S A 108:2396–2401. https://doi.org/10.1073/pnas.1016404108
48. Jiang X, Huang Q, Wang W, Dong H, Ly H, Liang Y, Dong C. 2013. Structures of arenaviral nucleoproteins with triphosphate dsRNA reveal a unique mechanism of immune suppression. J Biol Chem 288:16949–16959. https://doi.org/10.1074/jbc.M112.420521
49. Mushegian A. 2022. Methyltransferases of riboviria. Biomolecules 12:1247. https://doi.org/10.3390/biom12091247
50. Sõmera M, Fargette D, Hébrard E, Sarmiento C, ICTV Report Consortium. 2021. ICTV virus taxonomy profile: solemoviridae 2021. J Gen Virol 102. https://doi.org/10.1099/jgv.0.001707
51. Segelke BW. 2022. Functional annotation from structural homology. Methods Mol Biol 2349:215–257. https://doi.org/10.1007/978-1-0716-1585-0_11
52. Ictv Report C. 2023. Family: secoviridae, on ICTV. Available from: https://ictv.global/report/chapter/secoviridae/secoviridae
53. Walker Peter J., Freitas-Astúa J, Bejerman N, Blasdell KR, Breyta R, Dietzgen RG, Fooks AR, Kondo H, Kurath G, Kuzmin IV, Ramos-González PL, Shi M, Stone DM, Tesh RB, Tordo N, Vasilakis N, Whitfield AE, ICTV Report Consortium. 2022. ICTV virus taxonomy profile: rhabdoviridae 2022. J Gen Virol 103. https://doi.org/10.1099/jgv.0.001689
54. Walker PJ, Firth C, Widen SG, Blasdell KR, Guzman H, Wood TG, Paradkar PN, Holmes EC, Tesh RB, Vasilakis N. 2015. Evolution of genome size and complexity in the rhabdoviridae. PLoS Pathog 11:e1004664. https://doi.org/10.1371/journal.ppat.1004664
55. Bailey-Elkin BA, van Kasteren PB, Snijder EJ, Kikkert M, Mark BL. 2014. Viral OTU deubiquitinases: a structural and functional comparison. PLoS Pathog 10:e1003894. https://doi.org/10.1371/journal.ppat.1003894
56. Bailey-Elkin BA, Knaap RCM, Kikkert M, Mark BL. 2017. Structure and function of viral deubiquitinating enzymes. J Mol Biol 429:3441–3470. https://doi.org/10.1016/j.jmb.2017.06.010
57. Borujeni PM, Salavati R. 2024. Functional domain annotation by structural similarity. NAR Genom Bioinform 6:lqae005. https://doi.org/10.1093/nargab/lqae005
58. Ruperti F, Papadopoulos N, Musser JM, Mirdita M, Steinegger M, Arendt D. 2023. Cross-phyla protein annotation by structural prediction and alignment. Genome Biol 24:113. https://doi.org/10.1186/s13059-023-02942-9
59. Urayama SI, Fukudome A, Hirai M, Okumura T, Nishimura Y, Takaki Y, Kurosawa N, Koonin EV, Krupovic M, Nunoura T. 2024. Double-stranded RNA sequencing reveals distinct riboviruses associated with thermoacidophilic bacteria from hot springs in Japan. Nat Microbiol 9:514–523. https://doi.org/10.1038/s41564-023-01579-5
60. Elde NC, Child SJ, Eickbush MT, Kitzman JO, Rogers KS, Shendure J, Geballe AP, Malik HS. 2012. Poxviruses deploy genomic accordions to adapt rapidly against host antiviral defenses. Cell 150:831–841. https://doi.org/10.1016/j.cell.2012.05.049
61. Krupovic M, Koonin EV. 2017. Multiple origins of viral capsid proteins from cellular ancestors. Proc Natl Acad Sci U S A 114:E2401–E2410. https://doi.org/10.1073/pnas.1621061114
62. Koonin EV, Dolja VV, Krupovic M. 2022. The logic of virus evolution. Cell Host & Microbe 30:917–929. https://doi.org/10.1016/j.chom.2022.06.008
63. Hou X, He Y, Fang P, Mei S-Q, Xu Z, Wu W-C, Tian J-H, Zhang S, Zeng Z-Y, Gou Q-Y, Xin G-Y, Le S-J, Xia Y-Y, Zhou Y-L, Hui F-M, Pan Y-F, Eden J-S, Yang Z-H, Han C, Shu Y-L, Guo D, Li J, Holmes EC, Li Z-R, Shi M. 2024. Using artificial intelligence to document the hidden RNA virosphere. Cell, October. https://doi.org/10.1016/j.cell.2024.09.027
64. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, Forster J, Lee S, Twardziok SO, Kanitz A, Wilm A, Holtgrewe M, Rahmann S, Nahnsen S, Köster J. 2021. Sustainable data analysis with snakemake. F1000Res 10:33. https://doi.org/10.12688/f1000research.29032.2

65. Steinegger M, Söding J. 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol 35:1026–1028. https://doi.org/10.1038/nbt.3988
66. Buchfink B, Reuter K, Drost HG. 2021. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods 18:366–368. https://doi.org/10.1038/s41592-021-01101-x
67. Van Dongen S. 2008. Graph clustering via a discrete uncoupling process. SIAM J Matrix Anal & Appl 30:121–141. https://doi.org/10.1137/040608635
68. Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform. Nucleic Acids Res 30:3059–3066. https://doi.org/10.1093/nar/gkf436
69. Shen W, Le S, Li Y, Hu F. 2016. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS One 11:e0163962. https://doi.org/10.1371/journal.pone.0163962
70. Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J. 2019. HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics 20:473. https://doi.org/10.1186/s12859-019-3019-7
71. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The protein data bank. Nucleic Acids Res 28:235–242. https://doi.org/10.1093/nar/28.1.235
72. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, Finn RD, Bateman A. 2021. Pfam: the protein families database in 2021. Nucleic Acids Res 49:D412–D419. https://doi.org/10.1093/nar/gkaa913
73. Fox NK, Brenner SE, Chandonia JM. 2014. SCOPe: structural classification of proteins--extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res 42:D304–9. https://doi.org/10.1093/nar/gkt1240
74. Chandonia JM, Guan L, Lin S, Yu C, Fox NK, Brenner SE. 2022. SCOPe: improvements to the structural classification of proteins - extended database to facilitate variant interpretation and machine learning. Nucleic Acids Res 50:D553–D559. https://doi.org/10.1093/nar/gkab1054
75. Cheng H, Schaeffer RD, Liao Y, Kinch LN, Pei J, Shi S, Kim BH, Grishin NV. 2014. ECOD: an evolutionary classification of protein domains. PLoS Comput Biol 10:e1003926. https://doi.org/10.1371/journal.pcbi.1003926
76. van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, Söding J, Steinegger M. 2022. Foldseek: fast and accurate protein structure search. bioRxiv. https://doi.org/10.1101/2022.02.07.479398
77. Pettersen EF, Goddard TD, Huang CC, Meng EC, Couch GS, Croll TI, Morris JH, Ferrin TE. 2021. UCSF ChimeraX: structure visualization for researchers, educators, and developers. Protein Sci 30:70–82. https://doi.org/10.1002/pro.3943
78. Bastian M, Heymann S, Jacomy M. 2009. Gephi: an open source software for exploring and manipulating networks. ICWSM 3:361–362. https://doi.org/10.1609/icwsm.v3i1.13937
79. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402. https://doi.org/10.1093/nar/25.17.3389
80. Wang J, Chitsaz F, Derbyshire MK, Gonzales NR, Gwadz M, Lu S, Marchler GH, Song JS, Thanki N, Yamashita RA, Yang M, Zhang D, Zheng C, Lanczycki CJ, Marchler-Bauer A. 2023. The conserved domain database in 2023. Nucleic Acids Res 51:D384–D388. https://doi.org/10.1093/nar/gkac1096
81. Wolf YI, Silas S, Wang Y, Wu S, Bocek M, Kazlauskas D, Krupovic M, Fire A, Dolja VV, Koonin EV. 2020. Doubling of the known set of RNA viruses by metagenomic analysis of an aquatic virome. Nat Microbiol 5:1262–1270. https://doi.org/10.1038/s41564-020-0755-4
82. Necci M, Piovesan D, Dosztányi Z, Tosatto SCE. 2017. MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins. Bioinformatics 33:1402–1404. https://doi.org/10.1093/bioinformatics/btx015
83. Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113. https://doi.org/10.1186/1471-2105-5-113
84. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421. https://doi.org/10.1186/1471-2105-10-421
85. Letunic I, Bork P. 2024. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. Nucleic Acids Res 52:W78–W82. https://doi.org/10.1093/nar/gkae268
86. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods 14:587–589. https://doi.org/10.1038/nmeth.4285
87. Minh BQ, Nguyen MAT, von Haeseler A. 2013. Ultrafast approximation for phylogenetic bootstrap. Mol Biol Evol 30:1188–1195. https://doi.org/10.1093/molbev/mst024
88. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: improving the ultrafast bootstrap approximation. Mol Biol Evol 35:518–522. https://doi.org/10.1093/molbev/msx281
89. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305:567–580. https://doi.org/10.1006/jmbi.2000.4315
90. Sonnhammer EL, von Heijne G, Krogh A. 1998. A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol 6:175–182.