Unique structural homologs

Aromatic acid detoxification

Protein K4NVH5 from Ascovirus was structurally homologous to Phenolic acid decarboxylase (PAD, Table S5), a class of enzyme that decarboxylates phenolic compounds to their corresponding p-vinyl derivatives via a non-oxidative mechanism (1). Although PADs have been identified in bacteria, amoeba, protozoa and algae, this appears to be the first report from a virus. Ascoviruses multiply within the larval tissues of Lepidoptera, to which they are eventually fatal (2). Lepidopteran larvae are voracious herbivores, encountering an array of plant phenolic compounds generated for anti-herbivore defense. These compounds act by covalent inactivation of larval digestive enzymes and covalent reaction with larval gut tissue (3, 4). They may be detoxified by enzymes secreted into the gut lumen by the larva or the gut microbiota (5, 6).

Ascovirus-filled vesicles accumulate in the larval gut lumen before spreading throughout the larval body (7). However, the products of some detoxification systems of the larval host, such as quinones produced by the prophenoloxidases, remain toxic to viruses. Baculovirus infectivity, for example, is significantly lowered by the binding of quinones to viral occlusion bodies (8). Ascovirus protein K4NVH5 may serve to detoxify phenolics in the host's diet while diverting them away from quinone-producing pathways that could remain toxic to the virus.

K4NVH5 had a second structural homolog, Burkholdia beta lacatamase, whose structural homology to various aromatic acid decarboxylases (PADs, Ferulic acid decarboxylases and p-Coumaric acid decarboxylase) is apparent in the RCSB database. The relationship between beta lactam ring-opening and aromatic acid decarboxylation is unclear. They both involve hydrolysis at a carbonyl bond, though for beta lactamase this is an N-C bond in a 4-membered ring, while for PAD it is a C-C bond and the carbonyl is part of a terminal carboxyl group. Their catalytic mechanisms may not be related.

Ribosomal protein

The uncharacterized Emiliania huxleyi virus accession Q4A2G2 showed structural homology to Deinococcus 50S ribosomal protein L19 exceeding the 80% probability threshold (Table S5) as well as to the equivalent protein from other prokaryotes, Archeae, and the *S. cerevisiae* mitoribosome (data not shown). Eukaryotic cytoplasmic ribosomes possess no homolog to this protein. The large protein sequence family encompassing L19 includes a homolog in the red algae chloroplast (Table 5) and, like red algae, the coccolithophore host of Emiliania huxleyi virus is photosynthetic. Emiliania huxleyi virus may therefore modulate host photosynthesis via its chloroplast ribosome. A number of other ribosomal proteins have recently been found in viruses (mainly phage (9)), though to our knowledge this is the first report of a prokaryotic L19 homolog or of any prokaryotic-type ribosomal protein in genome of a eukaryotic virus. Q4A2G2 is conserved in all Emiliania huxleyi virus genomes sequenced to date.

Endocytosis

Ranavirus protein Q6GZV8 is structurally homologous to the 'V-shaped' domain of the human modular protein PDCD6IP/ALIX. PDCD6IP/ALIX functions in the ESCRT pathway for intralumenal endosomal vesicle formation, at the abscission stage of cytokinesis (10). It is also involved in the abscission and budding of enveloped (lenti)viruses via hijack of the cellular ESCRT machinery (in which a short peptide motif in lentivirus GAG protein interacts with ALIX V-shaped domain (11, 12)). Tiger frog virus (TFV), a member of the *Ranavirus* genus, uses the ESCRT pathway during virus budding, recruiting ALIX and other proteins that bind to the ESCRT protein complex to mediate its release from the host cell (13). We speculate that Q6GZV8 may be involved in this pathway.

Pandoravirus protein A0A0B5J0R1 was structurally homologous to an SHD1 domain ("SLA1 homology domain 1"). SHD1 domains in yeast protein sla1p act as adaptors during endocytocis: In clathrin-coated vesicles sla1p binds actin while its SHD1 binds cargo proteins containing an $NPFX_{(1,2)}D$ endocytic targeting signal. This signal is found in plasma membrane proteins destined for rapid endocytic internalization (14-17). Instead of sorting to the lysosomes for complete degradation, however, $NPFX_{(1,2)}D$-containing proteins are recycled back to the

plasma membrane (17). In Pandoravirus, protein A0A0B5J0R1 might be acting as a decoy to block the endocytotic destruction of viral membrane proteins.

Septum formation

Entomopox alpha protein W6JIY4 showed structural homology to *Pyrococcus* protein SepF (Table S5). SepF is involved in the binary fission of gram positive bacteria during septum formation between vestigial daughter cells (18-20). SepF stimulates the bundling of protofilaments of the tubulin-like GTPase FtsZ protein, thereby stimulating formation of the contractile FtsZ "Z ring" that marks the physical site of division of the mother cell followed by ring constriction and fission (21). The C-terminal portion of SepF contains the FtsZ binding site and is sufficient to promote FtsZ ring formation, while the N-terminal portion contains a transmembrane domain that presumably anchors the Z ring to the dividing bacterial membrane, allowing the membrane to be pulled inwards during contraction. Entomopox W6JIY4 and *Pyrococcus* SepF are comparable in length (109 vs 131 aa, respectively) and the structural homology region covers the C-terminal FtsZ-binding portion of SepF (Table S5). A prokaryotic tubulin-binding type domain therefore seems to have been co-opted in Entomopox alpha due to its potential for binding the host cell cytoskeleton. Like bacterial SepF, Entomopox alpha W6JIY4 has an N-terminal transmembrane region. A number of steps in virion morphogenesis may involve the cytoskeleton-driven constriction or remodeling of membranes, an obvious candidate being virus budding and abscission as promoted by cellular ESCRT complexes in many viruses (22). Speculatively, in Entomopox alpha, this role may have adopted a prokaryotic-type constriction/vesicularization mechanism. However, to have such a fundamental role in the virus lifecycle, the protein would likely be conserved among the Entomopoxviruses at least, yet no W6JIY4 orthologs or SepF structural homologs were found in the two other Entomopoxvirus type-species analyzed here (Table 1; data not shown) and no proteins with significant sequence similarity to W6JIY4 were identified in BLASTP searches. This lack of conservation may be more typical of a viral defense-type protein. Whatever its role, the structural homology detected here, only just exceeding our 80% threshold, may indicate some divergence during its adaptation to a eukaryotic virus.

Gasdermin

Vaccinia protein A47 (P26673) showed structural homology to the C-terminal domain of Gasdermin A3. This is a conserved auto-inhibitory domain found in various Gasdermins. Caspase-directed Gasdermin cleavage at a linker region connecting the N- and C-terminal domains unmasks the N-terminal domain from auto-inhibition, allowing it to undergo a conformational change that promotes oligomerization leading to the formation of membrane-spanning pores (23), pyroptotic cell death, and cytokine release (24-26). A47 may be a molecular decoy for the activated (unmasked) Gasdermin N-terminal domain, thereby suppressing pyroptosis. Protein A47 is expressed early during Vaccinia infection and contains unusually high numbers of CD8+ T cell epitopes able to prime T cells *in vivo* (27).

Proteasomal degradation

Megavirus protein K7YHS8 was structurally homologous to the N-terminal beta-grasp fold domain of Arabidopsis NPL4-like protein 1. This fold is found in diverse protein families (28) including the compact globular ubiquitin-like (UbL) domain found in ubiquitin and other proteins. UbL-containing proteins bind substrates destined for degradation and also bind subunits of the proteasome, and thus regulate protein turnover (29). The beta-grasp fold of NPL4-like protein 1 is likely also a UbL domain (28). NPL4 interacts with the N-terminal domain of the AAA ATPase VCP/p97 (30) which has diverse functions in the cell mostly centered around ubiquitin-dependent processes (31): For example, it facilitates the degradation of monoubiquitylated, polyubiquitylated, and non-degradative ubiquitin chain-containing proteins. It also extracts proteins from membranes and other cellular structures for degradation (or activation in the case of transcription factors precursors). It seems possible that Megavirus protein K7YHS8 may regulate the degradation of Megavirus proteins during infection.

Also in Megavirus, protein K7Z7B4 was structurally homologous to a variety of AAA domain containing proteins, with K7Z7B4 residues 135 – 304 showing AAA domain alignment. A small subset of AAA domain containing proteins, namely the AAA domain-containing proteasome regulatory subunits, showed more extensive homology to K7Z7B4. The top homolog, overall, was 26S proteasome regulatory subunit 7 protein

(Table S5 - also known as PSMC2/RPT1/MSS1) (32). This appears to be the first finding of an MSS1 homolog in a non-eukaryote. The 26S proteasome comprises a barrel-shaped, proteolytic 20S core with a 19S regulatory "lid" at one or both ends. 19S serves to unfold ubiquitinated target proteins and to translocate them into the 20S proteolytic chamber (33). 19S contains at least 18 subunits including a hexameric ring of six distinct AAA ATPases - one of which is our top structural homolog, subunit 7. Subunit 7 appeared unique in both the degree and extent of homology with K7Z7B4, being the only protein showing structural homology to the N-terminal side of the AAA ATPase domain (K7Z7B4 residues 72 – 133), a region of unknown function. Other proteasome regulatory subunits show homology within this region (residues 82 – 133) K7Z7B4, which has the Pfam designation "Proteasomal ATPase OB C-terminal domain" (PF16450). However, since these other regulatory subunits lacked RCSB structural data immediately N-terminal to residue 82, it was unclear whether subunit 7's slightly more extensive homology was real or illusory. Interestingly, K7Z7B4 lacks the "AAA+lid" domain present in these proteasome regulatory subunits. K7Z7B4 may serve to modulate the target specificity of the proteasome. In this regard, it would seem to be a reasonable partner for K7YHS8, above. Nonetheless, some proteasomal regulatory subunits (eg. subunit 6A/PSMC3) are multifunctional, with roles in transcriptional tumor suppression and binding to HIV TAT protein (34, 35). This appears to be the first finding of a proteasomal subunit in a virus of any kind.

Hypoxic response

Chlorellavirus protein Q98541, which is currently annotated as an integral membrane protein, was found to be a structural homolog of human HIG1 domain family member 1B, an integral membrane protein induced by hypoxia (36) whose functions remain poorly understood. This may be the first identification of a HIG1 domain family member in a virus.

Antimicrobial peptides

Cystine knot proteins: Cystine knots are highly stable structural motifs comprising four beta sheets crosslinked by three disulfide bridges (37). One class of cystine knot proteins, the inhibitor cystine knot ("Knottin") class, exhibits toxic, insecticidal or anti-microbial activity (38, 39). The 217 residue Chloriridovirus protein Q197F5 and the 281 residue Mollivirus protein A0A0M5KJJ9 contained regions homologous to the Knottin motif. Specifically, Q197F5 (residues 120 - 158) was homologous to the antimicrobial and antifungal peptides Alo-3 (Harlequin beetle, Table S5) and antimicrobial peptide 1 (Pokeweed, not shown) along with various conotoxins and other toxin peptides (not shown). A0A0M5KJJ9 contains three adjacent regions of structural homology (between residues 100 and 273) to the pharmacologically inert 32 residue peptide "Asteropsin G" from the marine sponge Asteropus. While these regions of A0A0M5KJJ9 matched the general requirements for cystine knots, they did not match the highly specific requirements for Knottin, perhaps consistent with the apparently non-toxic character of Asteropsin G (40). We are therefore circumspect about whether A0A0M5KJJ9 has actual knottin character. Cystine knots have been identified in many plants and animals, but have not, to our knowledge, been reported in a virus.

Entomopox beta protein R4ZER6 was structurally homologous to Defensin-like protein 1 from horse chestnut, which is a knottin-fold protein. Defensins more generally are arthropod and insect peptides active against Gram-positive bacteria (41, 42). They are found in many species, including Lepidoptera, the insect host of beta Entomopoxviruses (43).

Toxin-antitoxin systems

Marseillevirus contained three proteins comparable in size and structure to the 110-residue *E. coli* multidrug resistance-conferring membrane protein EmrE. EmrE belongs to a family of small multidrug resistance (SMR) transporters driving the efflux of aromatic cationic drugs from the cytoplasm via a drug/H^+ antiport mechanism (44). EmrE's transport substrates have few common structural features (44). (45). More than 200 SMR genes have been identified in bacteria (plus a few archaea) including bacterial strains with multiple paralogs (45). All share a critical conserved glutamate (Glu-14) also present in one of the Marseillevirus proteins (D2XAC8; residue 15). Their occurrence on plasmids or their proximity in the bacterial chromosome to insertion elements (e.g. EmrE is encoded within the DLP12 cryptic lambdoid prophage region of the *E. coli* chromosome) suggests a strategy for gene spread via horizontal gene transfer. EmrE homologs were previously found in two

Yellowstone Lake phycodnavirus metagenomes (46). Additional Marseillevirus orthologs of the three Marseillevirus proteins can be found by BLASTP (data not shown). These maybe the first identifications of members of this SMR protein family in eukaroytes or their viruses, and their roles are not obvious, though paralogous bacterial transporters show substrate complementarity (45). Drug resistance proteins may occur in NCLDV to promote virus persistence via symbiosis, immunity or addiction (47) or due to their amoebal hosts residing in complex aqueous and phagocytic environments.

Marseillevirus protein D2XAS7 showed structural homology to the short *E. Coli* antitoxin GhoS - an endoribonuclease that targets a specific site in a specific *E. coli* mRNA – namely that for toxin GhoT (48). GhoT functions by damaging the *E. coli* inner membrane via the formation of transient transmembrane pores (48, 49). Due to the nature of its fold (*E. coli* GhoS shows structural homology to the short CRISPR-associated sequence-specific endoribonuclease CAS2 (48)) we speculate that D2XAS7 acts as an endoribonuclease during Marseillevirus infection, though evidence it has antitoxin function therein is lacking.

Glycosylation and oligosaccharide degradation

Chlorellavirus protein Q84630 was identified as a structural homolog of bacterial membrane endo-alpha-mannosidase, an enzyme required for cell wall biosynthesis (50, 51). The latter enzyme is a structural prototype for glycan trimming enzymes of the endoplasmic reticulum (51). Cellular mannosidases function early during the diversification and maturation of protein-attached glycans in the ER and Golgi. Viral surface and secreted proteins are glycosylated (52), and the hijacking of N-glycan synthesis can occur in viral and other diseases (51). Mannosidases are also implicated in ER-associated protein degradation (53). Speculatively, Q84630 may serve to redirect the host protein glycosylation machinery to the production of an antigenically distinct pattern of viral protein glycosylation.

Mimivirus protein E3VYK8 was structurally homologous to two enzymes with roles in cleaving oligosaccharides at the glycosidic bond, namely the crystallized N-terminal region of beta galactosidase from Bacteroides, and beta-1,4-mannooligosaccharide phosphorylase. Entomopox beta proteins R4ZE02 and R4ZER5, showed structural homology to the N-terminal alpha-helical domain of streptococcal Hyaluronate lyase (54), a secreted enzyme that promotes bacterial tissue invasion by degrading the glycosaminoglycans found in extracellular matrix (55). The N-terminal alpha-helical domain in polysaccharide lyase family 8 (PL8) enzymes possesses the catalytic site and contributes one side of a structural cleft that binds substrate (54). Glycosaminoglycans are found in the insect midgut (56, 57), and the degradative activities of R4ZE02 and R4ZER5 may facilitate the host spread of Entomopox beta. Alternatively, in entomopoxvirus this domain may have lost catalytic activity - providing, instead, a viral attachment protein acting in comparable fashion to the glycosaminoglycan-binding chordopoxvirus attachment proteins (58-60).

Structural domains

ssDNA binding domains: Emiliania Huxleyi Virus protein Q4A2A1 showed three apparent single-stranded DNA binding domains covering three distinct types of OB fold (Table S5). It may have a role in virus genome replication and/or the maintenance of virus genome telomeres.

Coiled-coil domain: Two overlapping regions of the 447 residue Lymphocystivirus protein Q677M6 (residues 298 - 376 and 346 – 425) showed structural homology to a 121-residue core domain of the 155 residue *Bacillus* lipoprotein GerD. GerD is located in the inner membrane of the bacterial spore and functions in its rapid response to external germinants (61). The 121-residue core peptide forms an alpha helical homotrimer in solution and crystallizes into a neatly twisted superhelical rope (62) that may nucleate the clustering of spore inner membrane proteins. The corresponding triple-helical region in Lymphocystivirus could play any number of roles in virus biology. Vaccinia attachment protein A27, for example, forms a triple coiled-coiled homotrimer (63-65).

Ars operon repressor: A 69 residue region of the 290 residue Faustovirus protein A0A0H3TLY8 is structurally homologous to the 120 residue protein ArsD, a plasmid-encoded trans-acting repressor of the bacterial arsenical resistance (‘ars’) operon (arsRDABC). ArsD represses the operon to basal levels in the absence of trivalent/pentavalent arsenite or antimony metalloids (66) by binding a 24 nt segment of the ars promoter (66)

and is released from DNA by arsenite binding. ArsD also sequesters toxic intracellular metalloids (67) and shuttles them to the ATPase component of the arsenical pump (ArsA, encoded within the arsRDABC operon) for reduction and expulsion (68). It seems unlikely that Faustovirus A0A0H3TLY8 has any metal binding role since none of the metal-binding cys of ArsD (69, 70) are conserved (A0A0H3TLY8 is entirely cys-free). However, this fold may have been co-opted for its DNA binding properties or some other role.

References

1. **Cavin JF, Dartois V, Divies C.** 1998. Gene cloning, transcriptional analysis, purification, and characterization of phenolic acid decarboxylase from Bacillus subtilis. Appl Environ Microbiol **64:**1466-1471.
2. **Govindarajan R, Federici BA.** 1990. Ascovirus infectivity and effects of infection on the growth and development of noctuid larvae. J Invertebr Pathol **56:**291-299.
3. **Bi JL, Felton GW.** 1995. Foliar oxidative stress and insect herbivory: Primary compounds, secondary metabolites, and reactive oxygen species as components of induced resistance. J Chem Ecol **21:**1511-1530.
4. **Bhonwong A, Stout MJ, Attajarusit J, Tantasawat P.** 2009. Defensive role of tomato polyphenol oxidases against cotton bollworm (Helicoverpa armigera) and beet armyworm (Spodoptera exigua). J Chem Ecol **35:**28-38.
5. **Xia X, Gurr GM, Vasseur L, Zheng D, Zhong H, Qin B, Lin J, Wang Y, Song F, Li Y, Lin H, You M.** 2017. Metagenomic Sequencing of Diamondback Moth Gut Microbiome Unveils Key Holobiont Adaptations for Herbivory. Front Microbiol **8:**663.
6. **Wu K, Zhang J, Zhang Q, Zhu S, Shao Q, Clark KD, Liu Y, Ling E.** 2015. Plant phenolics are detoxified by prophenoloxidase in the insect gut. Sci Rep **5:**16823.
7. **Bigot Y, Rabouille A, Doury G, Sizaret PY, Delbost F, Hamelin MH, Periquet G.** 1997. Biological and molecular features of the relationships between Diadromus pulchellus ascovirus, a parasitoid hymenopteran wasp (Diadromus pulchellus) and its lepidopteran host, Acrolepiopsis assectella. J Gen Virol **78 (Pt 5):**1149-1163.
8. **Felton GW, Duffey SS.** 1990. Inactivation of baculovirus by quinones formed in insect-damaged plant tissues. J Chem Ecol **16:**1221-1236.
9. **Mizuno CM, Guyomar C, Roux S, Lavigne R, Rodriguez-Valera F, Sullivan MB, Gillet R, Forterre P, Krupovic M.** 2019. Numerous cultivated and uncultivated viruses encode ribosomal proteins. Nat Commun **10:**752.
10. **Hurley JH, Emr SD.** 2006. The ESCRT complexes: structure and mechanism of a membrane-trafficking network. Annu Rev Biophys Biomol Struct **35:**277-298.
11. **Martin-Serrano J, Yarovoy A, Perez-Caballero D, Bieniasz PD.** 2003. Divergent retroviral late-budding domains recruit vacuolar protein sorting factors by using alternative adaptor proteins. Proc Natl Acad Sci U S A **100:**12414-12419.
12. **Fisher RD, Chung HY, Zhai Q, Robinson H, Sundquist WI, Hill CP.** 2007. Structural and biochemical studies of ALIX/AIP1 and its role in retrovirus budding. Cell **128:**841-852.
13. **Mi S, Qin XW, Lin YF, He J, Chen NN, Liu C, Weng SP, He JG, Guo CJ.** 2016. Budding of Tiger Frog Virus (an Iridovirus) from HepG2 Cells via Three Ways Recruits the ESCRT Pathway. Sci Rep **6:**26581.
14. **Howard JP, Hutton JL, Olson JM, Payne GS.** 2002. Sla1p serves as the targeting signal recognition factor for NPFX(1,2)D-mediated endocytosis. J Cell Biol **157:**315-326.
15. **Tan PK, Howard JP, Payne GS.** 1996. The sequence NPFXD defines a new class of endocytosis signal in Saccharomyces cerevisiae. J Cell Biol **135:**1789-1800.
16. **Mahadev RK, Di Pietro SM, Olson JM, Piao HL, Payne GS, Overduin M.** 2007. Structure of Sla1p homology domain 1 and interaction with the NPFxD endocytic internalization motif. Embo Journal **26:**1963-1971.
17. **Piao HL, Machado IM, Payne GS.** 2007. NPFXD-mediated endocytosis is required for polarity and function of a yeast cell wall stress sensor. Molecular Biology of the Cell **18:**57-65.
18. **Hamoen LW, Meile JC, de Jong W, Noirot P, Errington J.** 2006. SepF, a novel FtsZ-interacting protein required for a late step in cell division. Molecular Microbiology **59:**989-999.
19. **Duman R, Ishikawa S, Celik I, Strahl H, Ogasawara N, Troc P, Lowe J, Hamoen LW.** 2013. Structural and genetic analyses reveal the protein SepF as a new membrane anchor for the Z ring. Proc Natl Acad Sci U S A **110:**E4601-4610.
20. **Szwedziak P, Wang Q, Freund SM, Lowe J.** 2012. FtsA forms actin-like protofilaments. Embo Journal **31:**2249-2260.
21. **Jekely G.** 2014. Origin and evolution of the self-organizing cytoskeleton in the network of eukaryotic organelles. Cold Spring Harb Perspect Biol **6:**a016030.
22. **Chen BJ, Lamb RA.** 2008. Mechanisms for enveloped virus budding: can some viruses do without an ESCRT? Virology **372:**221-232.
23. **Ruan J, Xia S, Liu X, Lieberman J, Wu H.** 2018. Cryo-EM structure of the gasdermin A3 membrane pore. Nature **557:**62-67.
24. **Lin PH, Lin HY, Kuo CC, Yang LT.** 2015. N-terminal functional domain of Gasdermin A3 regulates mitochondrial homeostasis via mitochondrial targeting. Journal of Biomedical Science **22:**44.

25. **Ding J, Wang K, Liu W, She Y, Sun Q, Shi J, Sun H, Wang DC, Shao F.** 2016. Pore-forming activity and structural autoinhibition of the gasdermin family. Nature **535:**111-116.
26. **Evavold CL, Ruan J, Tan Y, Xia S, Wu H, Kagan JC.** 2018. The Pore-Forming Protein Gasdermin D Regulates Interleukin-1 Secretion from Living Macrophages. Immunity **48:**35-44 e36.
27. **Yuen TJ, Flesch IE, Hollett NA, Dobson BM, Russell TA, Fahrer AM, Tscharke DC.** 2010. Analysis of A47, an immunoprevalent protein of vaccinia virus, leads to a reevaluation of the total antiviral CD8+ T cell response. J Virol **84:**10220-10229.
28. **Burroughs AM, Balaji S, Iyer LM, Aravind L.** 2007. Small but versatile: the extraordinary functional and structural diversity of the beta-grasp fold. Biol Direct **2:**18.
29. **Su V, Lau AF.** 2009. Ubiquitin-like and ubiquitin-associated domain proteins: significance in proteasomal degradation. Cell Mol Life Sci **66:**2819-2833.
30. **Isaacson RL, Pye VE, Simpson P, Meyer HH, Zhang X, Freemont PS, Matthews S.** 2007. Detailed structural insights into the p97-Npl4-Ufd1 interface. J Biol Chem **282:**21361-21369.
31. **Meyer H, Weihl CC.** 2014. The VCP/p97 system at a glance: connecting cellular function to disease pathogenesis. Journal of Cell Science **127:**3877-3883.
32. **Dubiel W, Ferrell K, Rechsteiner M.** 1993. Peptide sequencing identifies MSS1, a modulator of HIV Tat-mediated transactivation, as subunit 7 of the 26 S protease. Febs Letters **323:**276-278.
33. **Schweitzer A, Aufderheide A, Rudack T, Beck F, Pfeifer G, Plitzko JM, Sakata E, Schulten K, Forster F, Baumeister W.** 2016. Structure of the human 26S proteasome at a resolution of 3.9 A. Proc Natl Acad Sci U S A **113:**7816-7821.
34. **Sepe M, Festa L, Tolino F, Bellucci L, Sisto L, Alfano D, Ragno P, Calabro V, de Franciscis V, La Mantia G, Pollice A.** 2011. A regulatory mechanism involving TBP-1/Tat-Binding Protein 1 and Akt/PKB in the control of cell proliferation. PLoS ONE **6:**e22800.
35. **Shibuya H, Irie K, Ninomiya-Tsuji J, Goebl M, Taniguchi T, Matsumoto K.** 1992. New human gene encoding a positive modulator of HIV Tat-mediated transactivation. Nature **357:**700-702.
36. **Denko N, Schindler C, Koong A, Laderoute K, Green C, Giaccia A.** 2000. Epigenetic regulation of gene expression in cervical cancer cells by the tumor microenvironment. Clin Cancer Res **6:**480-487.
37. **Colgrave ML, Craik DJ.** 2004. Thermal, chemical, and enzymatic stability of the cyclotide kalata B1: the importance of the cyclic cystine knot. Biochemistry **43:**5965-5975.
38. **Park HG, Kyung SS, Lee KS, Kim BY, Choi YS, Yoon HJ, Kwon HW, Je YH, Jin BR.** 2014. Dual function of a bee (Apis cerana) inhibitor cysteine knot peptide that acts as an antifungal peptide and insecticidal venom toxin. Dev Comp Immunol **47:**247-253.
39. **Kolmar H.** 2009. Biological diversity and therapeutic potential of natural and engineered cystine knot miniproteins. Current Opinion in Pharmacology **9:**608-614.
40. **Su M, Li H, Wang H, Kim E, Kim HS, Kim EH, Lee J, Jung JH.** 2016. Stable and biocompatible cystine knot peptides from the marine sponge Asteropus sp. Bioorg Med Chem **24:**2979-2987.
41. **Lambert J, Keppi E, Dimarcq JL, Wicker C, Reichhart JM, Dunbar B, Lepage P, Van Dorsselaer A, Hoffmann J, Fothergill J, et al.** 1989. Insect immunity: isolation from immune blood of the dipteran Phormia terranovae of two insect antibacterial peptides with sequence homology to rabbit lung macrophage bactericidal peptides. Proc Natl Acad Sci U S A **86:**262-266.
42. **Bulet P, Cociancich S, Dimarcq JL, Lambert J, Reichhart JM, Hoffmann D, Hetru C, Hoffmann JA.** 1991. Insect immunity. Isolation from a coleopteran insect of a novel inducible antibacterial peptide and of new members of the insect defensin family. J Biol Chem **266:**24520-24525.
43. **Yi HY, Chowdhury M, Huang YD, Yu XQ.** 2014. Insect antimicrobial peptides and their applications. Appl Microbiol Biotechnol **98:**5807-5822.
44. **Rotem D, Schuldiner S.** 2004. EmrE, a multidrug transporter from Escherichia coli, transports monovalent and divalent substrates with the same stoichiometry. J Biol Chem **279:**48787-48793.
45. **Schuldiner S.** 2009. EmrE, a model for studying evolution and mechanism of ion-coupled transporters. Biochimica Et Biophysica Acta **1794:**748-762.
46. **Zhang W, Zhou J, Liu T, Yu Y, Pan Y, Yan S, Wang Y.** 2015. Four novel algal virus genomes discovered from Yellowstone Lake metagenomes. Sci Rep **5:**15131.
47. **Villarreal LP.** 2016. Persistent virus and addiction modules: an engine of symbiosis. Curr Opin Microbiol **31:**70-79.

48. **Wang X, Lord DM, Cheng HY, Osbourne DO, Hong SH, Sanchez-Torres V, Quiroga C, Zheng K, Herrmann T, Peti W, Benedik MJ, Page R, Wood TK.** 2012. A new type V toxin-antitoxin system where mRNA for toxin GhoT is cleaved by antitoxin GhoS. Nat Chem Biol **8:**855-861.
49. **Kim JS, Schantz AB, Song S, Kumar M, Wood TK.** 2018. GhoT of the GhoT/GhoS toxin/antitoxin system damages lipid membranes by forming transient pores. Biochem Biophys Res Commun **497:**467-472.
50. **Lubas WA, Spiro RG.** 1987. Golgi endo-alpha-D-mannosidase from rat liver, a novel N-linked carbohydrate unit processing enzyme. J Biol Chem **262:**3775-3781.
51. **Thompson AJ, Williams RJ, Hakki Z, Alonzi DS, Wennekes T, Gloster TM, Songsrirote K, Thomas-Oates JE, Wrodnigg TM, Spreitz J, Stutz AE, Butters TD, Williams SJ, Davies GJ.** 2012. Structural and mechanistic insight into N-glycan processing by endo-alpha-mannosidase. Proc Natl Acad Sci U S A **109:**781-786.
52. **Watanabe Y, Bowden TA, Wilson IA, Crispin M.** 2019. Exploitation of glycosylation in enveloped virus pathobiology. Biochim Biophys Acta Gen Subj **1863:**1480-1497.
53. **Pan S, Cheng X, Sifers RN.** 2013. Golgi-situated endoplasmic reticulum alpha-1, 2-mannosidase contributes to the retrieval of ERAD substrates through a direct interaction with gamma-COP. Molecular Biology of the Cell **24:**1111-1121.
54. **Li S, Jedrzejas MJ.** 2001. Hyaluronan binding and degradation by Streptococcus agalactiae hyaluronate lyase. J Biol Chem **276:**41407-41416.
55. **Hynes WL, Walton SL.** 2000. Hyaluronidases of Gram-positive bacteria. FEMS Microbiol Lett **183:**201-207.
56. **Dinglasan RR, Alaganan A, Ghosh AK, Saito A, van Kuppevelt TH, Jacobs-Lorena M.** 2007. Plasmodium falciparum ookinetes require mosquito midgut chondroitin sulfate proteoglycans for cell invasion. Proc Natl Acad Sci U S A **104:**15882-15887.
57. **Mathias DK, Pastrana-Mena R, Ranucci E, Tao D, Ferruti P, Ortega C, Staples GO, Zaia J, Takashima E, Tsuboi T, Borg NA, Verotta L, Dinglasan RR.** 2013. A small molecule glycosaminoglycan mimetic blocks Plasmodium invasion of the mosquito midgut. PLoS Pathog **9:**e1003757.
58. **Chung CS, Hsiao JC, Chang YS, Chang W.** 1998. A27L protein mediates vaccinia virus interaction with cell surface heparan sulfate. J Virol **72:**1577-1585.
59. **Hsiao JC, Chung CS, Chang W.** 1999. Vaccinia virus envelope D8L protein binds to cell surface chondroitin sulfate and mediates the adsorption of intracellular mature virions to cells. J Virol **73:**8750-8761.
60. **Lin CL, Chung CS, Heine HG, Chang W.** 2000. Vaccinia virus envelope H3L protein binds to cell surface heparan sulfate and is important for intracellular mature virion morphogenesis and virus infection in vitro and in vivo. J Virol **74:**3353-3365.
61. **Pelczar PL, Igarashi T, Setlow B, Setlow P.** 2007. Role of GerD in germination of Bacillus subtilis spores. Journal of Bacteriology **189:**1090-1098.
62. **Li Y, Jin K, Ghosh S, Devarakonda P, Carlson K, Davis A, Stewart KA, Cammett E, Pelczar Rossi P, Setlow B, Lu M, Setlow P, Hao B.** 2014. Structural and functional analysis of the GerD spore germination protein of Bacillus species. J Mol Biol **426:**1995-2008.
63. **Rodriguez JF, Paez E, Esteban M.** 1987. A 14,000-Mr envelope protein of vaccinia virus is involved in cell fusion and forms covalently linked trimers. J Virol **61:**395-404.
64. **Lai C, Gong S, Esteban M.** 1990. Structural and functional properties of the 14-kDaenvelope protein of vaccinia virus synthesized in Escherichia coli. Journal of Biological Chemistry **265:**22174-22180.
65. **Vazquez MI, Rivas G, Cregut D, Serrano L, Esteban M.** 1998. The vaccinia virus 14-kilodalton (A27L) fusion protein forms a triple coiled-coil structure and interacts with the 21-kilodalton (A17L) virus membrane protein through a C-terminal alpha-helix. J Virol **72:**10126-10137.
66. **Chen Y, Rosen BP.** 1997. Metalloregulatory properties of the ArsD repressor. J Biol Chem **272:**14257-14262.
67. **Li S, Rosen BP, Borges-Walmsley MI, Walmsley AR.** 2002. Evidence for cooperativity between the four binding sites of dimeric ArsD, an As(III)-responsive transcriptional regulator. J Biol Chem **277:**25992-26002.
68. **Lin YF, Walmsley AR, Rosen BP.** 2006. An arsenic metallochaperone for an arsenic detoxification pump. Proc Natl Acad Sci U S A **103:**15617-15622.
69. **Lin YF, Yang J, Rosen BP.** 2007. ArsD: an As(III) metallochaperone for the ArsAB As(III)-translocating ATPase. J Bioenerg Biomembr **39:**453-458.
70. **Lin YF, Yang J, Rosen BP.** 2007. ArsD residues Cys12, Cys13, and Cys18 form an As(III)-binding site required for arsenic metallochaperone activity. J Biol Chem **282:**16783-16791.

Figure S1. Substantial query/target overlap in the overwhelming majority of matches: Overview of the range of match-types encountered.

Figure S1, example #1 (VERY COMMON)

Query protein (single domain):
Homology across full length of both query and target

Figure S1, example #2 (VERY COMMON)

Query protein (multidomain):
Full length of query matches full length of target

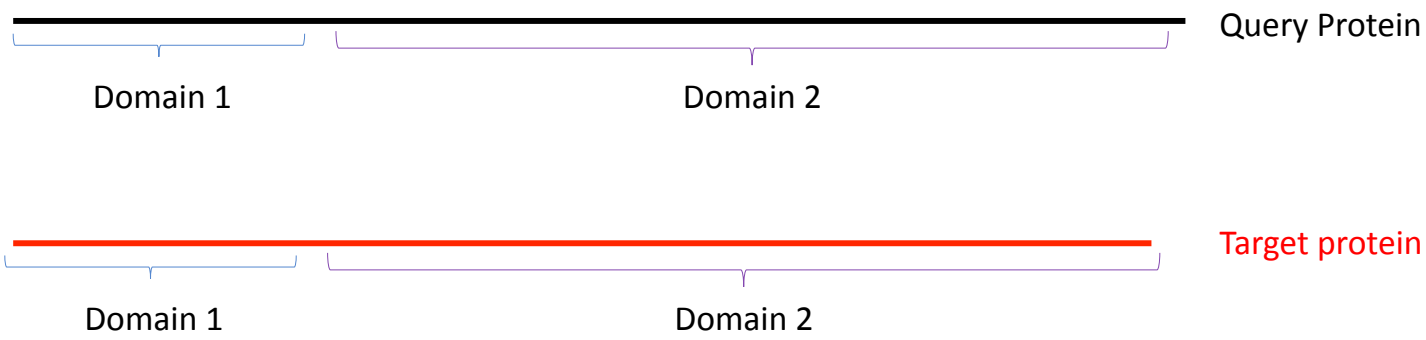


Figure S1, example #3 (COMMON)

Query protein (multidomain):
Query matches two crystal structures from same target, but covering different domains

Query protein

Target protein – domain #1 – crystal structure #1
Target protein – domain #2 – crystal structure #2
Target protein = residues not present in crystal structure

Figure S1, example #4 (LESS COMMON)

Query protein (any):
Query has homology to one target, plus an additional large area of the query (> 100 amino acids) with no homologs

Figure S1, example #5 (UNCOMMON)

Query protein (multidomain):
Different regions of the query have homology to domains from crystal structures of different target proteins

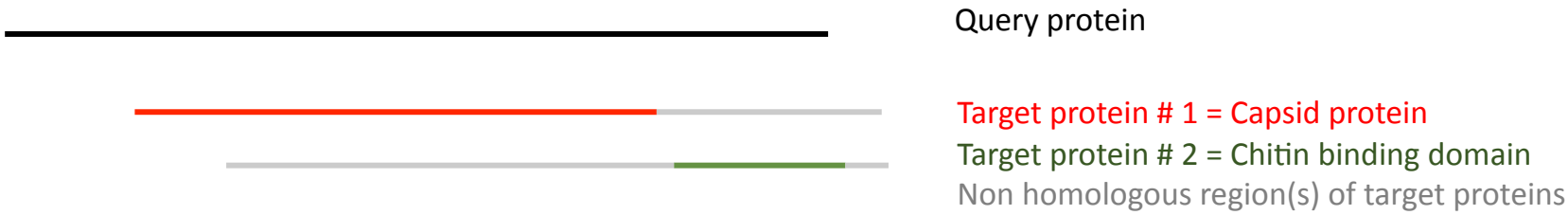


Figure S1, example #6 (UNCOMMON)

Query protein (any):
Repeating and overlapping homology regions in query, to repeat-rich targets e.g. Ankyrins, Collagen or Myosin-like proteins

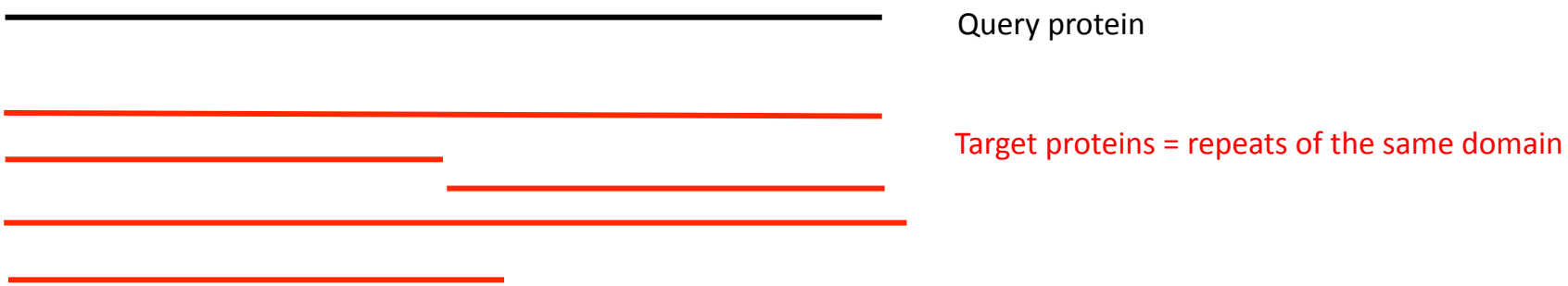
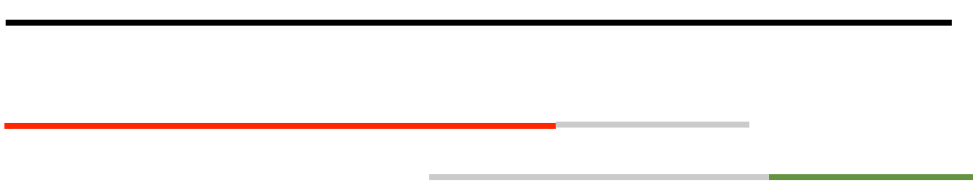


Figure S1, example #7 (VERY UNCOMMON)

Query protein (any):
Target protein has no Pfam or Superfamily – annotated, here, without Pfam

Figure S1, example #8 (VERY UNCOMMON)

Query protein (any):
One major domain of target protein, plus one transmembrane domain



Query protein

Target protein #1 = Major domain
Target protein #2 = helical transmembrane domain
Non-homologous region(s) of target proteins

Figure S1, example #9 (VERY UNCOMMON)

Query protein (any):
Helical transmembrane domain, only, identified in query

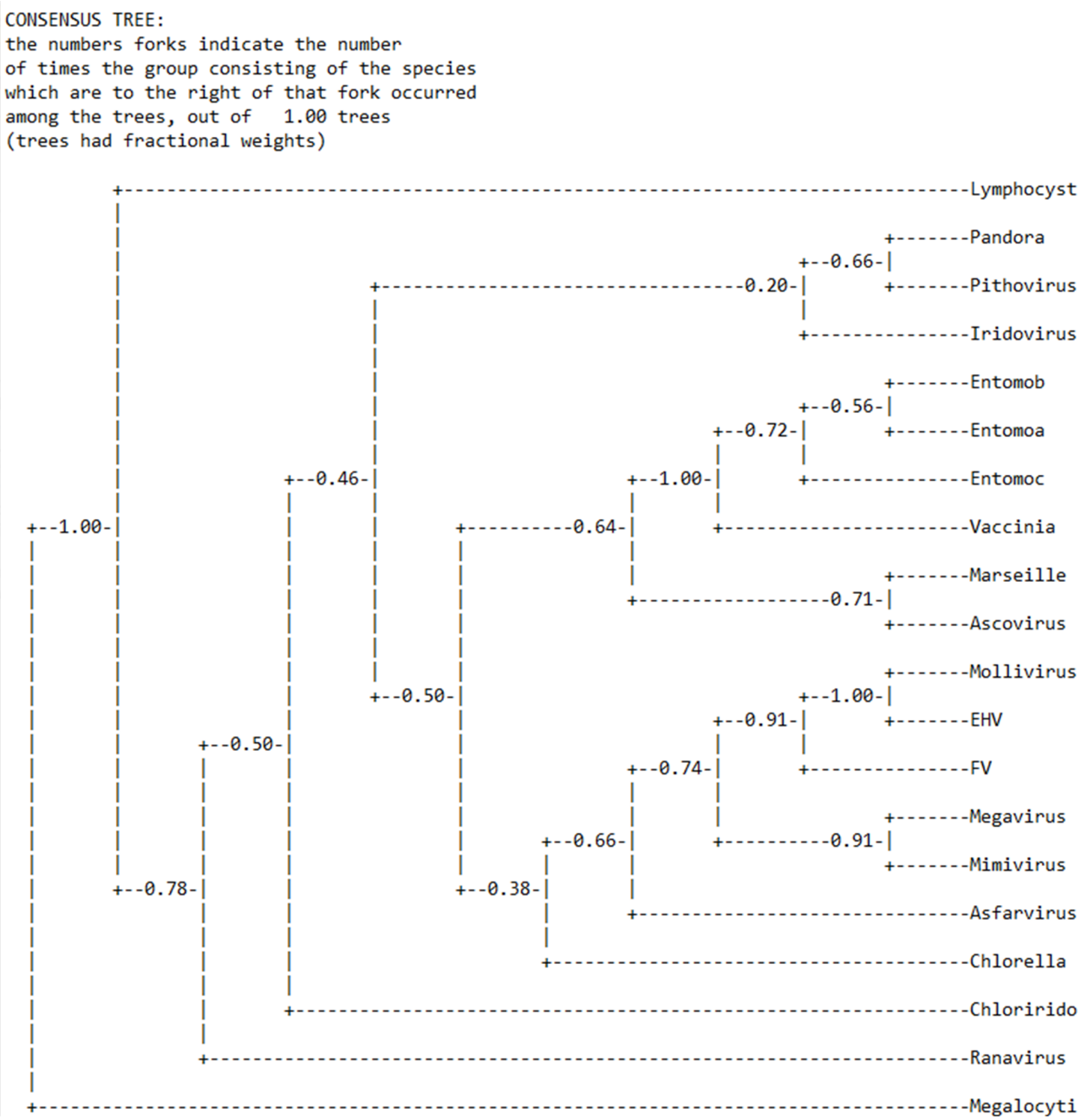


Figure S3. Phylogenetic tree inferred from a binary trait matrix of DNA-dependent RNA polymerase subunit/transcription factor presence/absence, generated using 'Dolpenny' (ref. 95 in the main manuscript) and 'Consense' from the PHYLIP package as described in "MATERIALS AND METHODS" in the main text. The tree shown is an "extended majority rule consensus" from the top 1000 individual trees.

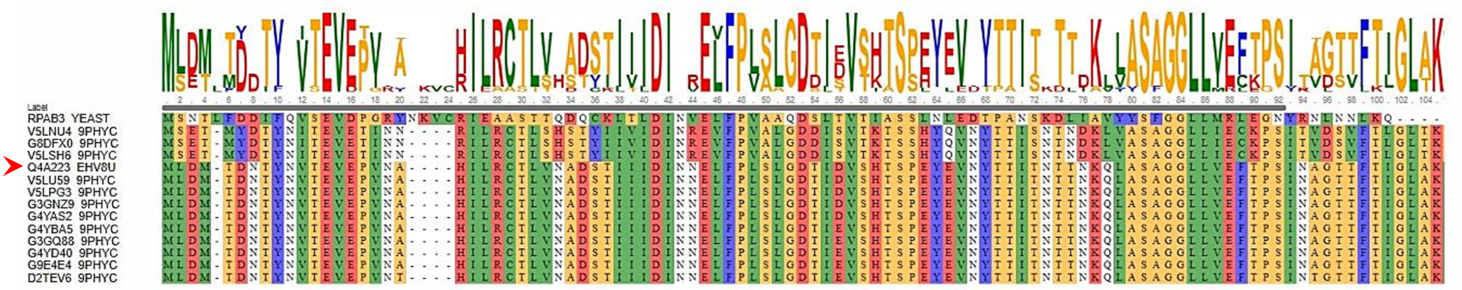


Figure S4. Multiple sequence alignment (ClustalW) of a cluster of RNA polymerase subunit RPB8 homologs encoded by Emiliania-Huxleyi viruses found using Q4A223_EHV8U (arrowed) as a BLASTP query. BLASTP e-values ranged from 10^{34} to 10^{73}. The alignment includes yeast RPB8 (top) from which amino acids 72 - 107 were removed (since they were absent from all viral sequences). Colors show similarity in amino acid chemical properties, and the consensus sequence is shown above the MSA. EhV orthologs fell into two apparent phylogenetic groups: V5LSH6 (EhV156), G8DFX0 (EhV-202), V5LNU4 (EhV-18) and Q4A223 (EhV-86), V5LU59 (EhV-164), V5LPG3 (EhV-145), G3GNZ9 (EhV-84), G4YAS2 (EhV-88), G4YBA5 (EhV-207), G3GQ88 (EhV-203), G4YD40 (EhV-208), G9E4E4 (EhV-201), D2TEV6 (EhV-99B1). These two groups showed around 71% amino acid similarity to one another, while yeast showed around 35.7% similarity with group 1. EhV orthologs were first aligned against each other by alignment order, then aligned against yeast RPB8 with fixed input order.

ETF1

```
# WEBSEQUENCE Length: 637
# WEBSEQUENCE Number of predicted TMHs:  1
# WEBSEQUENCE Exp number of AAs in TMHs: 29.1362699999999999
# WEBSEQUENCE Exp number, first 60 AAs:  19.57801
# WEBSEQUENCE Total prob of N-in:        0.64081
# WEBSEQUENCE POSSIBLE N-term signal sequence
WEBSEQUENCE     TMHMM2.0        inside       1     40
WEBSEQUENCE     TMHMM2.0        TMhelix     41     63
WEBSEQUENCE     TMHMM2.0        outside     64    637
```

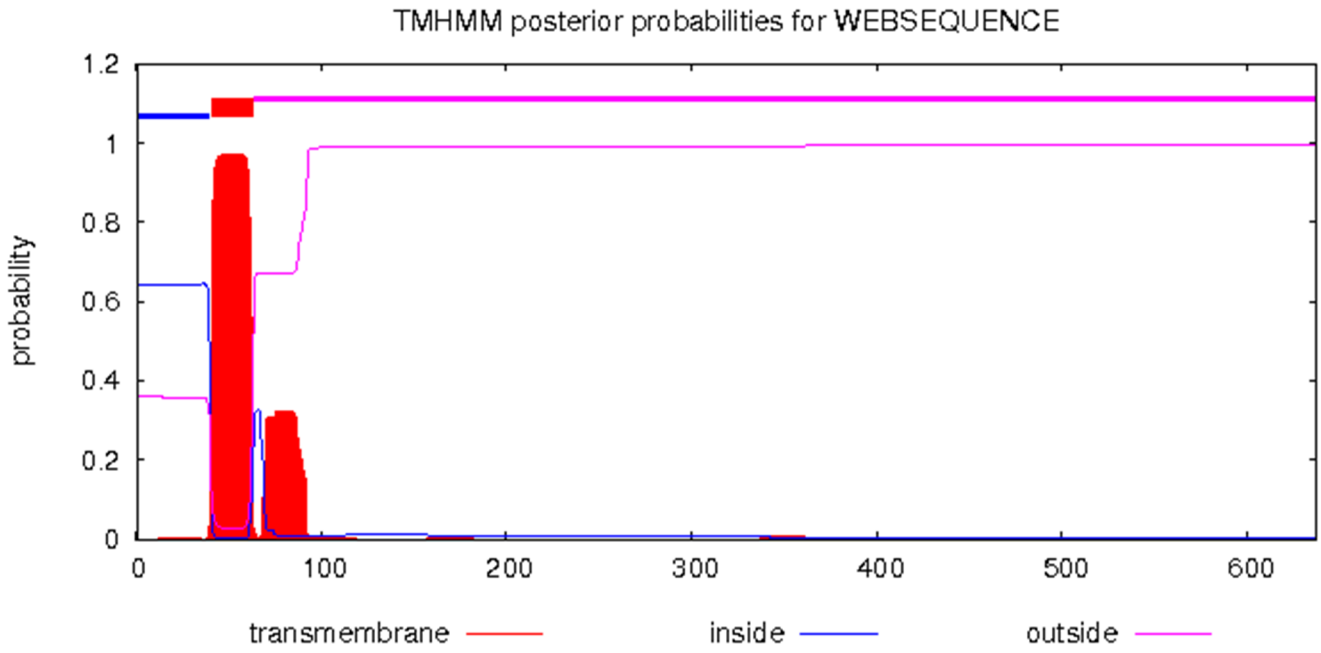


Figure S5. Strongly-predicted N-terminal TM domain/membrane anchor at positions 41 – 63 of protein ETF1. SignalP 5.0 identified no secretory signal peptide in this region. None of the other transcriptosome proteins had a strongly predicted membrane anchor.

Table S4. New/expanded trans-NCLDV protein families. For each row, structural homology combined with annotation associated with the structural h
Column 4: UniProt "Protein" field. Columns 5, 6: Pfam(s) covering, or overlapping with, the structural homology region (from the structural homolog's
NCLDV that was previously annotated as in column 6. Column 10: For a homology region covering the entire structural homolog, this is from UniProt'
PF07690 and PF00854, the Pfam hits to Mollivirus (previously) and Pandoravirus (here), respectively.

NCLDV query accession(s)	Structural homolog accession(s)	Structural homolog organism	Structural homolog name	Pfam(s) overlapping the homology region	Pfam descriptor(s)
D2XAQ6 (Marseillevirus) O55739 (Iridovirus)	Q9P0M2 (5jj2_A)	Human	A-kinase anchor protein 7 isoform gamma	PF10469	AKAP7 2'5' RNA ligase-like domain
Q84547 (Chlorella virus)	A0A0B8QKV5 (1zuj_A) P0CB53 (2cf7_A)	Lactococcus Streptococcus	DNA protection during starvation protein	PF00210	Ferritin-like domain
K7Z8N4 (Megavirus) A0A0G2Y127 (Mimivirus)	Q4WZ11 (3w0e_A)	Aspergillus	Elastase inhibitor AFUEI	PF11720	Peptidase inhibitor I78 family
K7YW37 (Megavirus) A0A0G2Y3W1 (Mimivirus)	W2SRJ3 (4uet_A)	Nematode	Fatty acid retinoid binding protein	PF05823	Nematode fatty acid retinoid binding protein (Gp-FAR-1)
A0A0B5JB34, A0A0B5JCF5 (Pandoravirus)	A0A0M3KKZ1 (4w6v_A)	Yersinia	di-/tripeptide transporter	CL0015	Major Facilitator Superfamily
K7Z7J6, K7Z8Q9, K7YFQ8 (Megavirus) A0A0G2Y657, A0A0G2Y9M2, F8V6J0 (Mimivirus)	Q8MQJ9 (4zlr_B)	Drosophila	Brain tumor protein	PF01436	NHL repeat
K7YFR4 (Megavirus) A0A0G2YBR1 (Mimivirus)	Q6Y7T6 (4csh_C)	Staphylococcus	Phage K_071	PF05257	CHAP domain
A0A0M5KAF0 (Mollivirus) A0A0B5JDI3 (Pandoravirus)	Q9S508 (3ff0_A) Q9GSQ9 (1n1i_A)	Pseudomonas Plasmodium	Phenazine biosynthesis protein B2 Merozoite surface protein 1	PF03284 PF12946 PF12947	Phenazine biosynthesis protein A/B MSP1 EGF domain 1 EGF domain
Q98542,Q98543 (Chlorella virus)	P39825 (3d9y_A) Q95VF7 (1acf_A)	Yeast Acanthamoeba	Profilin	PF00235	Profilin
Q4A2I6 (Emiliania-Huxleyi virus) D2XAY5 (Marseillevirus)	Q0SDB1 (4u5r_A) P70994 (2opa_A)	Rhodococcus Bacillus	Tautomerase_3 domain-containing protein 2-hydroxymuconate tautomerase	PF01361	4-Oxalocrotonate Tautomerase
K7YAA9, K7YUX8, K7YXF3, K7Z9E5 (Megavirus) A0A0G2YCB3, A0A0G2YCB5, A0A0G2Y4L1 (Mimivirus)	E7FCY1 (4BXR_A) Q9VI72 (4MPZ_A)	Zebrafish Drosophila	Centromere protein J Spindle assembly abnormal 4	PF07202	T-complex protein 10 C-terminus
K7Z767 (Megavirus) E3VYL3 (Mimivirus)	P0A8H8 (1LV3_A)	E. coli	DNA gyrase inhibitor YacG	PF03884	DNA gyrase inhibitor YacG
W6JPK9, W6JIZ4 (Entomopox alpha) R4ZDQ0, R4ZES4 (Entomopox beta) Q9YVZ3, Q9YW15 (Entomopox unclass.)	P06437 (2gum_A, 5fz2_A	Herpesvirus	Envelope glycoprotein B	PF17416 PF17417 PF00606	Herpesvirus Glycoprotein B Herpesvirus Glycoprotein B PH-like domain Herpesvirus Glycoprotein B ectodomain
A0A0M5KAC8 (Mollivirus) A0A0B5J3T1 (Pandoravirus)	B6JPK4 (3ub6_A)	Helicobacter	Methyl-accepting chemotaxis transmembrane sensory protein (MCP-like protein)	PF17200	Single Cache domain 2

Table S5. Structural homologies found uniquely in individual NCLDV. Almost all had a prior annotation of "Uncharacterized" (column 2). Columns are
Column 9: Probability values >99.8% are shown to two decimal places. Column 10: All functional annotation are sourced. Each row represents a disti
highest scoring member of a family of equivalent proteins. The four exceptions are: Rows 1 and 16 (K4NVH5 and E3VYK8): Two distinct families of s
multiple homologs with similar match probability; Row 9 Q4A223: Albeit 2f3i_A matched with marginally higher probability, 4ayb_G's homology regior

NCLDV query accession(s)	NCLDV query annotation	Structural homolog accession(s)	Structural homolog organism	Structural homolog name	Structural homology region	o ho
K4NVH5 (Ascovirus)	Uncharacterized protein	O07006 (2p8g_A) A4JJY8 (5hal_A)	Bacillus subtilis Burkholderia vietnamiensis	Phenolic acid decarboxylase PadC Uncharacterized protein]	5-106 (91%) 3-105 (92%)	
Q4A2G2 (Emiliania huxleyi virus)	Uncharacterized protein	Q9RWB4 (5dm6_M)	Deinococcus radiodurans	50S ribosomal protein L19	4-56 (85%)	
Q6GZV8 (Iridovirus/Ranavirus)	Uncharacterized protein 017L	Q8WUM4 (2r03_A)	Human	Programmed cell death 6-interacting protein	141-406 (53%)	
W6JIY4 (Entomopox alpha)	Uncharacterized protein	I6V3Q6 (3zig_A)	Pyrococcus furiosus	Uncharacterized protein	37-104 (61%)	
P26673 (Vaccinia)	Protein A47	Q5Y4Y6 (5b5r_A)	Mus musculus	Gasdermin-A3	61-245 (73%)	
K7YHS8 (Megavirus)	Uncharacterized protein	Q9LYC2 (1wf9_A)	Arabidopsis thaliana	NPL4-like protein 1	7-82 (90%)	
K7Z7B4 (Megavirus)	Uncharacterized protein	P35998 (5l4g_H)	Human	26S proteasome regulatory subunit 7	72-304 (76%)	
Q98541 (Chlorella virus)	Uncharacterized protein	Q9P298 (2lon_A)	Human	HIG1 domain family member 1B	3-66 (83%)	
Q4A223 (Emiliania huxleyi virus)	Uncharacterized protein	P52434 (2f3i_A) B8YB59 (4ayb_G)	Human Saccharolobus shibatae B12	DNA-directed RNA polymerases I, II, and III subunit RPABC3 RNA polymerase subunit 8	5-57 (46%) 7-100 (83%)	
Q197F5 (Iridoviridae/Chloriridovirus)	Uncharacterized protein 005L	P83653 (1q3j_A)	Acrocinus longimanus (Harlequin beetle)	Anti-microbial peptide Alo-3	120-158 (18%)	
A0A0M5KJJ9 (Mollivirus)	Uncharacterized protein	A0A1A9T940 (2n3p_A)	Asteropus (marine sponge)	Asteropsin_G	103-131 (10%) 167-198 (11%) 241-268 (10%)	
R4ZER6 (Entomopox beta)	Uncharacterized protein	Q7M1F3 (1bk8_A)	Aesculus hippocastanum (Horse-chestnut tree)	Defensin-like protein 1	44-78 (43%)	
D2XAM0 (Marseillevirus) D2XAC8 (Marseillevirus) D2XAC9 (Marseillevirus)	Uncharacterized protein Small membrane protein Small membrane protein	P23895 (2i68_A)	Escherichia coli	Multidrug transporter EmrE	65-102 (35%) 5-105 (94%) 13-110 (87%)	
D2XAS7 (Marseillevirus)	Uncharacterized protein	P0AF61 (2llz_A)	Escherichia coli	Endoribonuclease antitoxin GhoS	15-93 (70%)	
Q84630 (Chlorella virus)	Uncharacterized protein	Q8A109 (4acy_A)	Bacteroides thetaiotaomicron	Endo-alpha-mannosidase	173-437 (61%)	
E3VYK8 (Mimivirus)	Uncharacterized protein R118	Q8A921 (5muj_A) D9ZDQ9 (4udg_F)	Bacteroides thetaiotaomicron Uncultured organism	Beta galactosidase Uncharacterized protein	51-352 (85%) 27-352 (91%)	
R4ZE02 (Entomopox beta) R4ZER5 (Entomopox beta)	Uncharacterized protein Uncharacterized protein	Q53591 (1f1s_A)	Streptococcus agalactiae	Hyaluronate lyase	72-290 (34%) 87-303 (33%)	
Q4A2A1 (Emiliania huxleyi virus)	Uncharacterized protein	Q8Q045 (2kbn_A) P27694 (1jmc_A) O13988 (1qzg_A)	Methanosarcina mazei Human Schizosaccharomyces pombe	Conserved protein Replication protein A 70 kDa DNA-binding subunit Protection of telomeres protein 1	16-89 (16%) 17-210 (43%) 264-331 (15%)	P
Q677M6 (Iridovirus/Lymphocystivirus)	Uncharacterized protein	Q5L3Q1 (4o8w_A)	Geobacillus kaustophilus	Spore germination protein	298-425 (28%)	
A0A0H3TLY8 (Faustovirus)	Uncharacterized protein	P46003 (3mwh_A)	Escherichia coli	Arsenical resistance operon trans-acting repressor ArsD	26-94 (23%)	
A0A0B5J0R1 (Pandoravirus)	Uncharacterized protein	P32790 (2hbp_A)	Saccharomyces cerevisiae	Actin cytoskeleton-regulatory complex protein SLA1	54-93 (11%)	

Table S5 Reference

1. **Su M, Li H, Wang H, Kim E, Kim HS, Kim EH, Lee J, Jung JH.** 2016. Stable and biocompatible cystine knot peptides from the marine sponge Asteropus sp. Bi