# Multiple origins of viral capsid proteins from cellular ancestors

Mart Krupovic[a,1] and Eugene V. Koonin[b,1]

[a]Institut Pasteur, Department of Microbiology, Unité Biologie Moléculaire du Gène chez les Extrêmophiles, 75015 Paris, France; and [b]National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD 20894

Contributed by Eugene V. Koonin, February 3, 2017 (sent for review December 21, 2016; reviewed by C. Martin Lawrence and Kenneth Stedman)

**Viruses are the most abundant biological entities on earth and show remarkable diversity of genome sequences, replication and expression strategies, and virion structures. Evolutionary genomics of viruses revealed many unexpected connections but the general scenario(s) for the evolution of the virosphere remains a matter of intense debate among proponents of the cellular regression, escaped genes, and primordial virus world hypotheses. A comprehensive sequence and structure analysis of major virion proteins indicates that they evolved on about 20 independent occasions, and in some of these cases likely ancestors are identifiable among the proteins of cellular organisms. Virus genomes typically consist of distinct structural and replication modules that recombine frequently and can have different evolutionary trajectories. The present analysis suggests that, although the replication modules of at least some classes of viruses might descend from primordial selfish genetic elements, bona fide viruses evolved on multiple, independent occasions throughout the course of evolution by the recruitment of diverse host proteins that became major virion components.**

virus evolution | capsid proteins | nucleocapsids | origin of viruses | primordial replicons

Viruses are the most abundant biological entities on our planet and have a profound impact on global ecology and the evolution of the biosphere (1–4), but their provenance remains a subject of debate and speculation. Three major alternative scenarios have been put forward to explain the origin of viruses (5). The virus-first hypothesis, also known as the primordial virus world hypothesis, regards viruses (or virus-like genetic elements) as intermediates between prebiotic chemical systems and cellular life and accordingly posits that virus-like entities originated in the precellular world. The regression hypothesis, in contrast, submits that viruses are degenerated cells that have succumbed to obligate intracellular parasitism and in the process shed many functional systems that are ubiquitous and essential in cellular life forms, in particular the translation apparatus. Finally, the escape hypothesis postulates that viruses evolved independently in different domains of life from cellular genes that embraced selfish replication and became infectious. The three scenarios are not mutually exclusive, because different groups of viruses potentially could have evolved via different routes. Over the years, all three scenarios have been revised and elaborated to different extents. For instance, the diversity of genome replication-expression strategies in viruses, contrasting the uniformity in cellular organisms, had been considered to be most compatible with the possibility that the virus world descends directly from a precellular stage of evolution (4, 6); the discovery of giant viruses infecting protists led to a revival of the regression hypothesis (7–9); and an updated version of the escape hypothesis states that the first viruses have escaped not from contemporary but rather from primordial cells, predating the last universal cellular ancestor (10). The three evolutionary scenarios imply different timelines for the origin of viruses but offer little insight into how the different components constituting viral genomes might have combined to give rise to modern viruses.

A typical virus genome encompasses two major functional modules, namely, determinants of virion formation and those of genome replication. Understanding the origin of any virus group is possible only if the provenances of both components are elucidated (11). Given that viral replication proteins often have no closely related homologs in known cellular organisms (6, 12), it has been suggested that many of these proteins evolved in the precellular world (4, 6) or in primordial, now extinct, cellular lineages (5, 10, 13). The ability to transfer the genetic information encased within capsids—the protective proteinaceous shells that comprise the cores of virus particles (virions)—is unique to bona fide viruses and distinguishes them from other types of selfish genetic elements such as plasmids and transposons (14). Thus, the origin of the first true viruses is inseparable from the emergence of viral capsids.

Viral capsid proteins (CPs) typically do not have obvious homologs among contemporary cellular proteins (6, 15), raising questions regarding their provenance and the circumstances under which they have evolved. One possibility is that genes encoding CPs have originated de novo within the genomes of nonviral selfish replicons by generic mechanisms, such as overprinting and diversification (16). Alternatively, these proteins could have first performed cellular functions, subsequently being recruited for virion formation. For instance, it has been proposed that virus-like particles could have served as gene-transfer agents in precellular communities of replicators (17). Another possibility is that virus-like particles evolved in the cellular context as micro- and nanocompartments, akin to prokaryotic carboxysomes and encapsulins, for the sequestration of various enzymes for specialized biochemical reactions (18, 19).

Studies on the origin of viral capsids are severely hampered by the high sequence divergence among these proteins. Nevertheless, numerous structural comparisons have uncovered unexpected similarities in the folds of CPs from viruses infecting hosts from different cellular domains, testifying to the antiquity of the CPs and

## Significance

The entire history of life is the story of virus–host coevolution. Therefore the origins and evolution of viruses are an essential component of this process. A signature feature of the virus state is the capsid, the proteinaceous shell that encases the viral genome. Although homologous capsid proteins are encoded by highly diverse viruses, there are at least 20 unrelated varieties of these proteins. We show here that many, if not all, capsid proteins evolved from ancestral proteins of cellular organisms on multiple, independent occasions. These findings reveal a stronger connection between the virosphere and cellular life forms than previously suspected.

Author contributions: M.K. and E.V.K. designed research; M.K. performed research; M.K. and E.V.K. analyzed data; and M.K. and E.V.K. wrote the paper.

Reviewers: C.M.L., Montana State University; and K.S., Portland State University.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

[1]To whom correspondence may be addressed. Email: koonin@ncbi.nlm.nih.gov or krupovic@pasteur.fr.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1621061114/-/DCSupplemental.

the evolutionary connections between the viruses that encode them (20–23). It also became apparent that the number of structural folds found in viral CPs is rather limited. For instance, viruses with dsDNA genomes from 20 families have been shown to possess CPs with only five distinct structural folds (23). Here, to investigate the extent of the diversity and potential origins of viral capsids and to gain further insight into virus origins and evolution, we performed a comparative analysis of the major structural proteins across the entire classified virosphere and made a focused effort to identify cellular homologs of these viral proteins.

## Results and Discussion

**A Comprehensive Census of Viral Capsid and Nucleocapsid Proteins.** Viruses display remarkable diversity in the complexity and organization of their virions. With few exceptions, nonenveloped virions are constructed from one major capsid protein (MCP), which determines virion assembly and architecture, and one or a few minor CPs. By contrast, enveloped virions often contain nucleocapsid (NC) proteins which form nucleoprotein complexes with the respective viral genomes, matrix proteins linking the nucleoprotein to the lipid membrane, and envelope proteins responsible for host recognition and membrane fusion. These proteins often constitute a considerable fraction of the virion mass, making it challenging to single out the major virion protein. Nevertheless, NC proteins of some enveloped viruses are homologous to the MCPs of nonenveloped viruses [e.g., nonenveloped tenuiviruses and enveloped phleboviruses (24)] and thus are considered to be functionally equivalent herein.

Analysis of the available sequences and structures of major CP and NC proteins encoded by representative members of 135 virus taxa (117 families and 18 unassigned genera; Table S1) (25, 26) allowed us to attribute structural folds to 76.3% of the known virus families and unassigned genera. The remaining taxa included viruses that do not form viral particles (3%) and viruses for which the fold of the major virion proteins is not known and could not be predicted from the sequence data (20.7%). The former group includes capsidless viruses of the families *Endornaviridae*, *Hypoviridae*, *Narnaviridae*, and *Amalgaviridae*, all of which appear to have evolved independently from different groups of full-fledged capsid-encoding RNA viruses (27–29). The latter category includes eight taxa of archaeal viruses with unique morphologies and genomes (30), pleomorphic bacterial viruses of the family Plasmaviridae, and 19 diverse taxa of eukaryotic viruses (Table S1). It should be noted that, with the current explosion of metagenomics studies, the number and diversity of newly recognized virus taxa will continue to rise (31). Although many of these viruses are expected to have previously observed CP/NC protein folds, novel architectural solutions doubtlessly will be discovered as well.

The 76.3% of viral taxa for which the fold of the major virion proteins was defined could be divided into 18 architectural classes (Fig. 1), also referred to as "structure-based viral lineages" (20, 23). These architectural classes were unevenly populated by viral taxa: Seven major architectural classes covered 64.4% of the known virosphere. Of the remaining 11 minor classes, seven contained folds unique to a single virus family, three folds were found in two families each, and the fold specific to the NC protein of members of the order Nidovirales was conserved in viruses from three families (Fig. 1).

Among viral taxa for which the fold of the major virion proteins could be defined, 73 taxa (71%) have icosahedral virions, and 29 taxa (29%) include viruses with helical (nucleo)capsids. By contrast, among viral taxa with unknown CP/NC protein folds, nearly half (13 taxa) contain viruses with helical nucleoprotein complexes, whereas those with icosahedral capsids belong to only six taxa (21%); the rest of these taxa include viruses with bacilliform, droplet-shaped, pleomorphic, spherical, bottle-shaped and spindle-shaped virions (Table S1). Thus, the CP structures of viruses with icosahedral capsids seem already to have been sampled to considerable depth, whereas viruses with helical (nucleo)capsids are understudied and might be found to have novel structural folds in the future.

Icosahedral capsids of characterized viruses are constructed from CPs with 10 remarkably diverse structural folds, which range

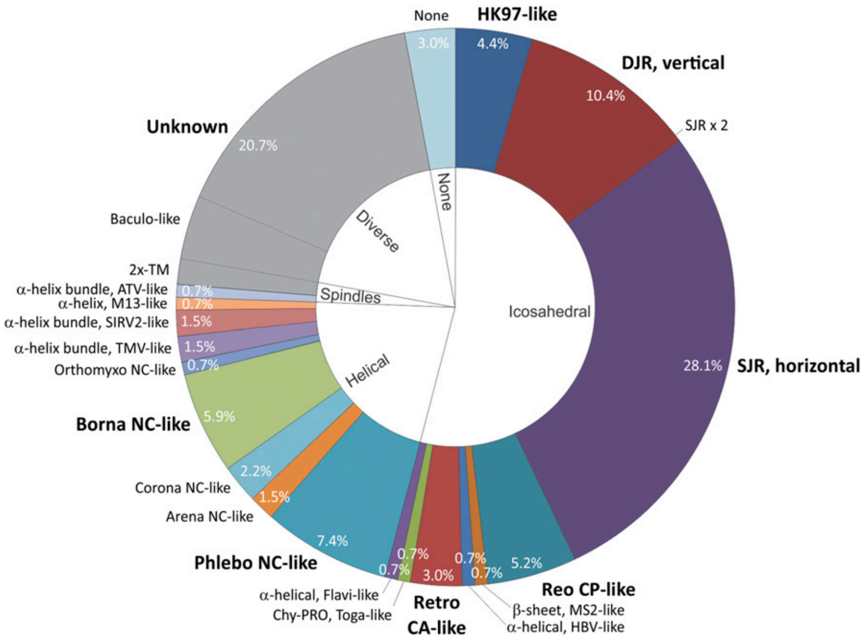

**Fig. 1.** Structural diversity of viral CP and NC proteins. The pie chart shows the distribution of architectural classes among 135 virus taxa (117 families and 18 unassigned genera; see Table S1). Arena, arenavirus; ATV, Acidianus two-tailed virus; Baculo-like, baculovirus-like viruses; Chy-PRO, chymotrypsin-like protease; Corona, coronavirus; CP, capsid protein; DJR, double jellyroll; Flavi, flavivirus; HBV, hepatitis B virus; NC, nucleocapsid protein; Orthomyxo, orthomyxovirus; Phlebo, phlebovirus; Reo, reovirus; Retro, retrovirus; SIRV2, *Sulfolobus islandicus* rod-shaped virus 2; SJR, single jellyroll; Toga, togavirus; TMV, tobacco mosaic virus; TM, transmembrane domain.

from exclusively α-helical to β-strand–based. The inherent ability of so many structurally unrelated proteins to assemble into icosahedral particles refutes the argument that the structural similarity between the MCPs of viruses infecting hosts in different domains of life is a result of convergent evolution, whereby the sheer geometry of the icosahedral shell constrains the evolution of a CP to a particular fold (32). Interestingly, the unrelated folds of the viral proteins that form helical (nucleo)capsids are largely α-helical. The reasons for such a bias are unclear, given that cellular helical filaments, such as certain bacterial pili, can be formed from β-strand–based proteins (33). Here we present a focused attempt to infer the likely evolutionary ancestry of the different classes of major virion proteins including CP, NC, and matrix proteins.

**Origins of Viral Structural Proteins.** Origins of viral (nucleo)capsids is one of the key unanswered questions in virus evolution. To understand the provenance of major proteins constituting viral particles, we performed systematic comparisons of viral proteins with the global database of protein sequences and structures.

***Jellyroll fold.*** The single jellyroll (SJR) is the most prevalent fold among viral CPs, representing ~28% of the CPs in the analyzed set of virus taxa (Fig. 1). High-resolution CP structures are available for viruses from 23 of the 38 taxa with SJR CPs (Table S1). Searches against the Protein Data Bank (PDB) database using the DALI server (34) seeded with representative CP structures resulted in multiple matches to cellular proteins containing the SJR-fold domains. Indeed, SJR proteins are widespread in organisms from all three cellular domains and are functionally diverse. However, most of those with the highest similarity to viral CPs can be classified into four major groups (Fig. 2). One of the most common functions of SJR domains in cellular proteins is carbohydrate recognition and binding. Accordingly, SJR domains are often appended to various carbohydrate-active enzymes, such as glycoside hydrolases (35). For example, a search seeded with the CP of satellite panicum mosaic virus (PDB ID code: 1STM) retrieved the carbohydrate-binding module from *Ruminococcus flavefaciens* (PDB ID code: 4D3L) with a highly significant DALI Z score of 7.9 (Fig. 2) despite the lack of appreciable sequence similarity. Similar results were obtained when searches were initiated with CPs from other virus families. Another family of cellular SJR proteins includes the P domain found in archaeal, bacterial, and eukaryotic subtilisin-like proteases, in which this domain is thought to assist in protein stabilization (36, 37). The P domain of *Saccharomyces cerevisiae* protease Kex2 (PDB ID code: 1R64) was retrieved with the CP of tobacco streak virus (Bromoviridae) (PDB ID code: 4Y6T) with a Z score of 5.8 (Fig. 2). The third family of viral CP-like SJR proteins includes nucleoplasmins and nucleophosmins (Fig. 2), molecular chaperones that bind to core histones and promote nucleosome assembly in eukaryotes (38). The core SJR domain of nucleoplasmins/nucleophosmins forms stable pentameric and decameric complexes that serve as platforms for binding histone octamers (39, 40). Hits to nucleoplasmins/nucleophosmins were obtained with different CPs, including the MCP of dsDNA bacteriophage P23-77 (PDB ID code: 3ZMO) of the family Sphaerolipoviridae (hit to PDB ID code: 2P1B; Z score, 5.2).

The fourth broad group of cellular proteins with CP-like SJR domains comprises cytokines of the TNF superfamily. TNF-like ligands and their corresponding receptors play pivotal roles in mammalian cell host-defense processes, inflammation, apoptosis, autoimmunity, and organogenesis (41). The biologically active form of TNF-like proteins is a trimer. Remarkably, however, soluble tumour necrosis factor- and Apo-L-related leucocyte-expressed ligand-1 (sTALL-1), a member of the TNF superfamily, has been shown to form 60-subunit (20 trimers) virus-like particles (42) that superficially resemble 60-subunit T = 1 virions (12 pentamers) of satellite tobacco necrosis virus (STNV) (Fig. 2 *A* and *B*). Furthermore, sTALL-1 is identified as a structural homolog of STNV CP with a DALI Z score of 7.7. Importantly, TNF-like proteins are not exclusive to eukaryotes but also are prevalent in bacteria. Perhaps most notable among these is the Bacillus collagen-like protein of anthracis (BclA) protein found in the outermost surface layer of *Bacillus anthracis* spores (43). A DALI search with the CP of cowpea mosaic virus (family Secoviridae; PDB ID code: 1NY7) retrieved BclA (PDB ID code: 3AB0) with a Z score of 5.4.

More divergent SJR domains are found in functionally diverse proteins of the Cupin superfamily (44). Although the conserved structural core in this superfamily consists of six β-strands, some members, such as oxygenases and Jumonji C (JmjC) domain-containing histone demethylases, contain eight antiparallel β-strands (45, 46). The Cupin superfamily includes bacterial transcription factors related to the arabinose operon regulator, AraC, in which the N-terminal SJR Cupin domain (Fig. 2*B*) is responsible for arabinose binding and dimerization and is fused to the C-terminal helix-turn-helix (HTH) DNA-binding domain (44). Searches seeded with AraC from *Escherichia coli* (PDB ID code: 2ARC) resulted in a match to the CP of San Miguel sea lion virus (family Caliciviridae) (PDB ID code: 2GH8) with a Z score of 2.7.

The ubiquity and functional diversity of cellular SJR proteins testifies to their antiquity. Indeed, it is highly probable that cellular proteins with the SJR fold had experienced substantial diversification before the emergence of the last universal cellular ancestor. Given the structural similarity between the cellular and viral SJR proteins and that some of these cellular proteins, such as the TNF superfamily, are capable of forming assemblies resembling virus-like particles (Fig. 2*B*), it is likely that the ancestor of the viral SJR CP evolved through recruitment of a cellular SJR protein. The original function of this protein could have involved recognition of carbohydrates. A protein with such a property would be immediately beneficial to the virus because, in addition to providing a protective shell for the genome, it could ensure specific binding of the viral particle to the host cell. It is noteworthy that many contemporary viruses bind directly to various glycan receptors on the surface of their hosts via the SJR CPs (47). The alternative possibility, that viral CPs gave rise to cellular SJR proteins, appears less likely, given the wide taxonomic distribution and functional diversity of SJR proteins in all three domains of cellular life, in sharp contrast to the scarcity of prokaryotic viruses with SJR CPs.

Transformation of a cellular protein into a bona fide CP would necessitate specific recognition and encapsidation of the viral genome. This function typically is performed by terminal extensions appended to the SJR core. For instance, some ssRNA and ssDNA viruses (e.g., tombusviruses and circoviruses, respectively) have largely unstructured, positively charged N-terminal domains that interact with the nucleic acids (48, 49).

Clustering of the SJR proteins with DALI, based on a pairwise comparison of the Z scores, suggests that the CPs from the majority of RNA viruses and eukaryotic ssDNA viruses form a monophyletic group (Fig. 2*C* and Dataset S1). Notably, circoviral CPs are nested among RNA viruses, as is consistent with the previously proposed scenario in which the CP genes of some eukaryotic ssDNA viruses have been horizontally acquired from ssRNA viruses (50–55). The compact CPs of bromoviruses (Fig. 2*A*) cluster with the P domain of Kex2-like subtilisin proteases, separately from other CPs (Fig. 2*C* and Dataset S1), whereas the CPs of bacterial microviruses (ssDNA genomes) appear to be more closely similar to the TNF-like proteins than to other viral CPs. The most divergent among viral SJR proteins, embellished with extended loops, are CPs of parvoviruses, polyomaviruses, and papillomaviruses. The CPs from the two latter virus groups form a clade separate from other viral CPs (Fig. 2*C* and Dataset S1). However, because of the high divergence of these proteins, their affinities are difficult to ascertain. Concurrently, among the cellular SJR proteins, cupins show the least similarity to other cellular and viral SJR proteins.

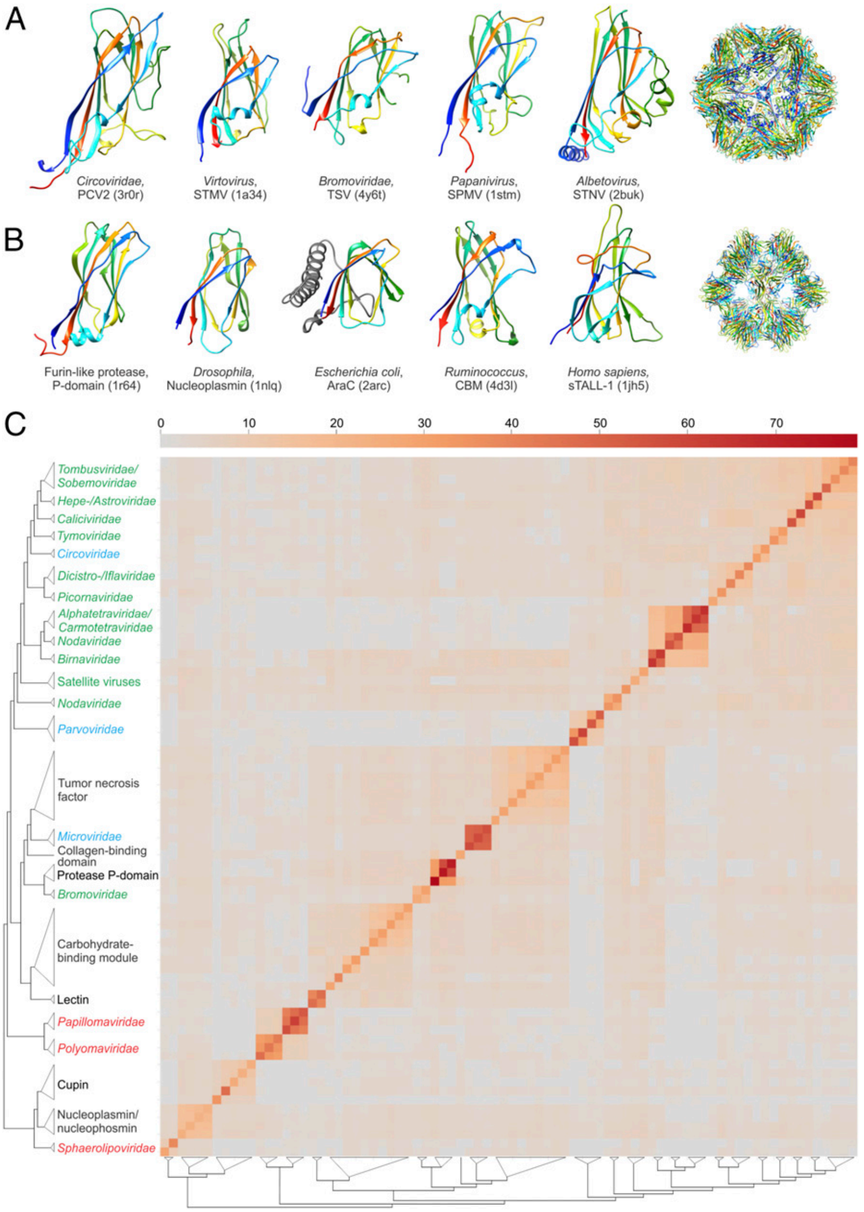

**Fig. 2.** Viral and cellular SJR proteins. (*A*) A selection of viral SJR CP structures. The rightmost structure corresponds to the virion of STNV. (*B*) A selection of cellular SJR protein structures. The rightmost structure corresponds to the 60-subunit virion-like assembly of the human sTALL-1 protein. All structures are colored using the rainbow scheme from blue (N terminus) to red (C terminus). The linker region leading to the DNA-binding domain in AraC is shown in gray. (*C*) Relationships between cellular and viral SJR proteins. The matrix and cluster dendrograms are based on the pairwise Z score comparisons calculated using DALI. For the complete matrix, see Dataset S1. The color scale indicates the corresponding Z scores. RNA viruses are shown in green, ssDNA viruses in blue, and dsDNA viruses in red. All compared structures are indicated with the corresponding PDB identifiers. CBM, carbohydrate-binding module; NP, nucleoplasmin/nucleophosmin; PCV2, porcine circovirus 2; PRO-P, P domain of subtilisin-like proteases; SPMV, satellite panicum mosaic virus; STMV, satellite tobacco mosaic virus; TSV, tobacco streak virus.

These observations suggest that viral SJR CPs could have evolved from bona fide cellular proteins, possibly on several independent occasions. However, we may never be able to pinpoint the exact family(ies) of cellular SJR proteins at the origin of the viral CPs with confidence because of the high evolution rates in viral genomes.

***Double jellyroll CPs.*** The second largest architectural class includes viruses with the double jellyroll (DJR) MCP, which consists of two consecutive jellyroll domains and is found in ~10% of the virus taxa (Fig. 1). Unlike SJR CPs, the DJR β-strands are oriented vertically with respect to the capsid surface (20, 21, 56). The DJR MCPs are exclusive to dsDNA viruses that are classified into 13 taxa and infect hosts from all three cellular domains (Table S1). This architectural class includes members of the bacterial virus families Corticoviridae and Tectiviridae, archaeal viruses of the family Turriviridae, and eukaryotic viruses of the families Adenoviridae and Lavidaviridae as well as the proposed order Megavirales that includes most of the large and giant eukaryotic viruses. Viruses with DJR MCPs are evolutionarily linked to a large group of unclassified eukaryotic endogenous viruses/transposons called "Polintoviruses/Polintons" (mavericks) that also encode a typical DJR protein, although the formation of virions remains to be demonstrated (57–59).

A straightforward evolutionary scenario proposes that the ancestral DJR MCP derives from a SJR CP via gene duplication (21, 60). Bacterial and archaeal viruses of the Sphaerolipoviridae family (61) display a potentially archaic virion architecture that might have given rise to the one observed in the DJR MCP viruses (62, 63). All sphaerolipoviruses encode two MCPs, each with the SJR fold, that form homo- and heterodimers involved in the formation of the icosahedral capsid with vertical orientation of the β-strands similar to the orientation in DJR MCPs. Consistent with the proposed SJR CP gene-duplication event in the evolution of the DJR ancestor, the two sphaerolipoviral CPs are most similar to each other among known protein structures (Fig. 2*C* and Dataset S1). Furthermore, sphaerolipoviruses and DJR viruses share genome-packaging ATPases of the A32-like family (named after the respective protein of vaccinia virus) (64) that thus far have not been found in viruses with other MCP types. Based on these common characteristics, we include sphaerolipoviruses in the architectural class of viruses encoding the DJR MCPs (Fig. 1). Notably, the two SJR CPs of sphaerolipoviruses cluster with nucleoplasmins/nucleophosmins, separately from other viral SJR CPs (Fig. 2*C* and Dataset S1), suggesting an independent origin from a cellular SJR ancestor.

***HK97-like MCP fold.*** The second major structural fold found in MCPs of dsDNA viruses is exemplified by and named after the gp5 protein of bacteriophage HK97 (65). This fold is characteristic of the MCPs of bacterial and archaeal members of the order Caudovirales (families *Myoviridae*, *Siphoviridae*, and *Podoviridae*) (66), one of the most abundant, widespread, and diverse groups of viruses on the planet (2, 3, 67). The HK97-fold is also found in the floor domain of the MCP of herpesviruses (order Herpesvirales) (68). In addition to homologous MCPs, herpesviruses and tailed prokaryotic dsDNA viruses share closely similar mechanisms of virion assembly, maturation, and genome packaging, indicating that at least the morphogenetic modules of the two groups evolved from a common ancestor (23, 56, 68, 69). Outside the virosphere, the HK97-like fold is found only in encapsulins, a class of bacterial and archaeal nanocompartments that encapsulate a variety of cargo proteins related to oxidative stress response, including ferritin-like proteins and DyP-type peroxidases (19). High-resolution structures are available for three encapsulins (Fig. S1), which, similar to viruses, assemble into icosahedral T = 1 or T = 3 cages (70–72). Structural comparison of the available cellular and viral proteins with the HK97-like fold showed that bacterial and archaeal encapsulins form a tight, apparently monophyletic cluster, whereas viral MCPs are more divergent (Fig. S1). This observation, along with the ubiquity of tailed dsDNA viruses, as opposed to the more narrow spread of encapsulins, might be interpreted as an indication of the viral origin of encapsulins via domestication of the HK97-like MCP. Notably, however, encapsulins are also encoded in certain groups of archaea, namely the phylum Crenarchaeota (73), which are not known to be parasitized by members of the Caudovirales. Thus, given our limited knowledge about the structural diversity and taxonomic distribution of encapsulins, the exact evolutionary relationship between encapsulins and MCPs remains unresolved.

***Chymotrypsin-like protease fold.*** Viruses of the genus *Alphavirus* (family *Togaviridae*) present another example of a CP that is evolutionarily related to cellular proteins. It has been noticed previously that alphavirus core (C) protein shares sequence similarity with chymotrypsin-like serine proteases (74, 75). The C protein consists of two domains: the largely unstructured, positively charged N-terminal region responsible for RNA binding and the C-terminal protease domain, which forms the icosahedral capsid shell located under the glycoprotein-containing envelope. In addition to its structural role in capsid formation, the protein acts as a protease and cleaves off the C protein from the polyprotein precursor. Following cleavage, the C terminus of the C protein inhibits its own protease activity (76). Structural studies have unequivocally shown that the C protein adopts the chymotrypsin-like fold (76). Strikingly, the closest homologs of alphaviral C protein (PDB ID code: 1WYK) are encoded by members of the family Flaviviridae (77, 78). The protease NS3 of Hepatitis C virus (PDB ID code: 1RGQ) is recovered with a DALI Z score of 10.8, and the protease HtrA from humans (PDB ID code: 3NWU) follows with a Z score of 10.6 (Fig. S2). The NS3 protease does not play a structural role in virion formation of flaviviruses but rather is responsible for the proteolytic processing of the polyprotein at several sites to produce mature viral proteins (77). In addition to the related proteases, flaviviruses and alphaviruses encode homologous class II envelope glycoproteins, which form the icosahedral shells around the membrane (79). However, unlike alphaviruses, flaviviruses do not form internal icosahedral capsids, and their C protein has a unique α-helical fold (80). The parsimonious scenario for the origin of the alphaviral C protein includes partial refunctionalization of the viral nonstructural protease, such as the flavivirus NS3, which itself evolved from the HtrA-like cellular protease. The key adaptation in this process was the addition of a positively charged N-terminal region to the protease domain, enabling the protein to bind viral RNA. Another notable difference between C protein and NS3 (and cellular proteases) is the absence of a conserved C-terminal α-helix in the C protein (Fig. S2). Remarkably, it has been recently demonstrated that the C protein of alphaviruses is not essential for virion formation (81). Deletion of the C gene results in the production of infectious, pleomorphic membrane vesicles decorated with viral glycoproteins and carrying the viral genome. This observation implies that the C protein is a relatively recent elaboration in alphaviral virions that is not central for virus propagation.

***Helical nucleocapsids.*** Strikingly, in the course of evolution, endonucleases appear to have been recruited to function as viral nucleocapsid proteins. The racket-shaped NC protein of nairoviruses (proposed family Nairoviridae) contains head and stalk domains (Fig. S3). It has been shown that the head domain of the Crimean-Congo hemorrhagic fever virus (CCHFV) nucleocapsid has a metal-dependent DNA-specific endonuclease activity (82). However, the protein does not display any recognizable similarity to known cellular nucleases. The NC protein of arenaviruses, such as Lassa mammarenavirus (LASV), which contains a head domain similar to that of nairoviruses (Fig. S3), instead displays a dsRNA-specific 3′–5′ exonuclease activity (83, 84). In the latter case, the activity is conferred not by the head domain but by the dedicated C-terminal domain homologous to various exonucleases of the DEDDh superfamily (named after four invariant acidic residues, DEDD, in the active site) (Fig. S3). It seems highly probable that the arenaviral nucleocapsid evolved from a nairoviral-like ancestor by acquiring the host-derived exonuclease domain.

In a case coming from a completely different part of the virosphere, evolution of a viral nucleocapsid from a nuclease has been also recently demonstrated for the enveloped filamentous archaeal virus, Thermoproteus tenax virus 1 (TTV1). Sequence analysis suggests that one of the two major NC proteins of TTV1 is a truncated and inactivated derivative of the CRISPR-associated nuclease Cas4, a component of adaptive CRISPR-Cas immune systems (85). Thus, it appears that during virus evolution cellular proteins involved in nucleic acid metabolism, nucleases in particular, have been recruited to function as structural components of the virion on several independent occasions.

***Retroviral Gag polyprotein.*** In all retroviruses, the structural polyprotein, group-specific antigen (Gag), is proteolytically processed into matrix (MA), capsid (CA), and NC proteins (Fig. 3*A*), but some viruses contain additional domains, such as p6 in HIV-1 (86). The MA is typically myristoylated at the N terminus and is required for Gag transport and subsequent binding to the cytoplasmic membrane; the CA and NC are both required for Gag multimerization and for the formation of immature spherical particles (87). Several high-resolution structures are available for all three major Gag domains from various retroviruses. Analysis of the retroviral MA structures reveals an α-helical fold that is

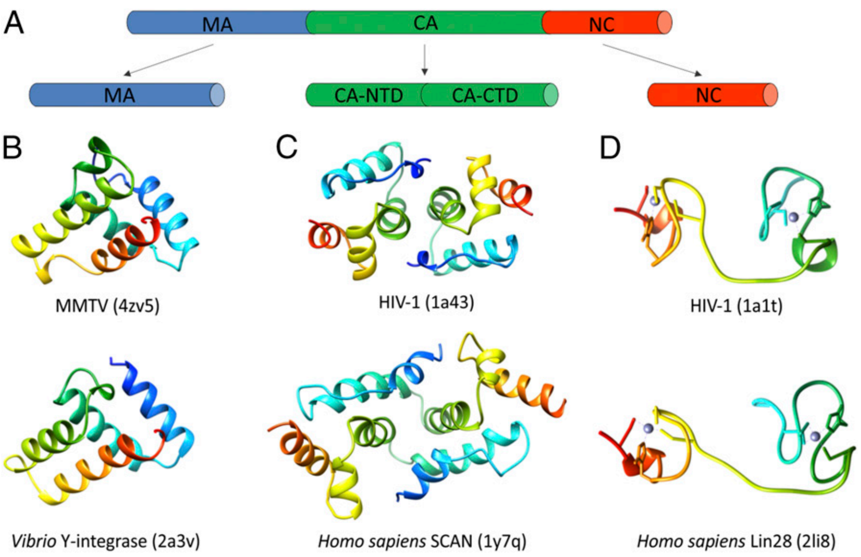
A
MA
CA
NC
MA
CA-NTD
CA-CTD
NC
B
C
D
MMTV (4zv5)
HIV-1 (1a43)
HIV-1 (1a1t)
*Vibrio* Y-integrase (2a3v)
*Homo sapiens* SCAN (1y7q)
*Homo sapiens* Lin28 (2li8)

**Fig. 3.** Cellular homologs of the retroviral proteins constituting the Gag polyprotein. (*A*) Proteolytic processing of the retroviral Gag polyprotein into MA, CA, and NC proteins. (*B*) Structural comparison of the matrix protein (*Upper*) of mouse mammary tumor virus (MMTV) with the N-terminal DNA-binding domain (*Lower*) of the tyrosine recombinase of *V. cholerae*. (*C*) Structural comparison of a dimer of the CA C-terminal domain (CA-CTD) (*Upper*) of HIV-1 with the dimer of the human SCAN domain protein (*Lower*). (*D*) Structural comparison of the NC protein (*Upper*) of HIV-1 with the human pluripotency factor Lin28 (*Lower*). All structures are colored using the rainbow scheme from blue (N terminus) to red (C terminus), and the corresponding PDB identifiers are shown.

remarkably similar to that of the N-terminal HTH DNA-binding domain found in various integrases of the tyrosine recombinase superfamily. A DALI search seeded with the MA of mouse mammary tumor virus (MMTV) (PDB ID code: 4ZV5) resulted in a significant hit (Z score, 4.8) to the HTH domain from the integron integrase of *Vibrio cholerae* (PDB ID code: 2A3V) (Fig. 3*B*). Notably, upon virus entry into the host cell and following reverse transcription, HIV-1 MA becomes a component of the preintegration complex and binds to dsDNA (88). Accordingly, the retroviral matrix displays not only structural but also functional similarity to the DNA-binding HTH domains and in all likelihood was exapted from this source.

Within the virosphere, the N-terminal domain of the CA protein appears to be unique to reverse-transcribing viruses. By contrast, a domain homologous to the C-terminal domain of the CA protein is commonly found in vertebrate transcription factors and is known as the "SCAN domain" (PF02023), a protein-interaction module that mediates self-association or selective association with other proteins (89). The SCAN domain is always accompanied by multiple C2H2 zinc fingers and/or Krüppel-associated box (KRAB) domains, none of which are of retroviral origin (90). The crystal structure of the SCAN dimer from the human ZNF174 protein indicates that this protein is indeed a domain-swapped homolog of the C-terminal domain of the retroviral CA protein (Fig. 3*C*) (91). It was previously concluded that known SCAN domains have been recruited from retrotransposons at or near the root of the tetrapod animal branch (89, 90). However, given the generic utility of the SCAN domain for

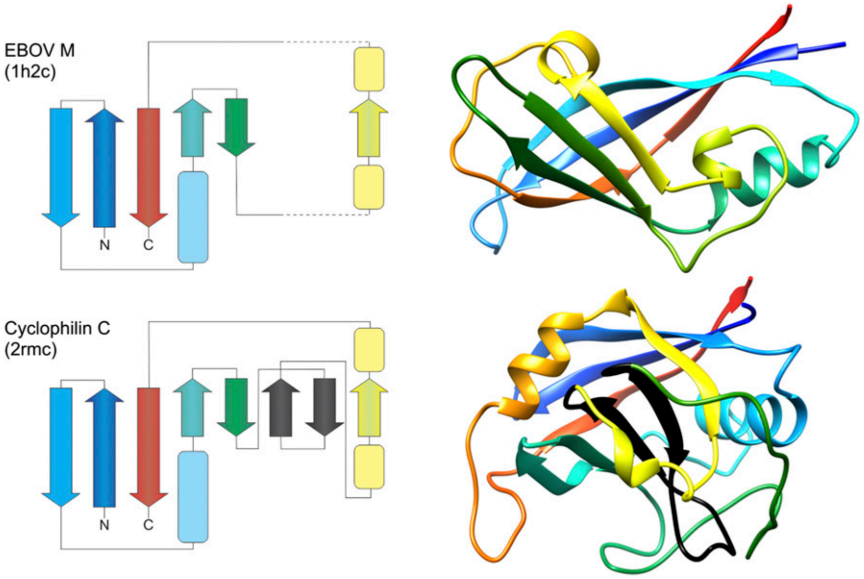

**Fig. 4.** Structural comparison of the Ebola virus MA protein with CypC. Topology diagrams (*Left*) and structural models (*Right*) are colored using the rainbow scheme from blue (N terminus) to red (C terminus), and the corresponding PDB identifiers are shown. The β-hairpin insert in CypC is colored black.

protein dimerization in both viruses and hosts (for functions unrelated to virion formation), the exact provenance of the ancestral SCAN-like dimerization domain remains uncertain.

The NC protein contains one or two CCHC Zn-knuckle motifs and binds the viral genome (87). An HHpred analysis of the NC protein sequence from HIV-1 showed that it is closely related to other Zn-knuckle domain proteins, most notably pluripotency factor Lin28 (probability = 99.3%) and Air2p, a substrate recognition component of a polyA RNA polymerase (probability = 99.2%). Comparison of the HIV-1 NC protein (PDB ID code: 1A1T) and human Lin28 (PDB ID code: 2LI8) further underscores the close structural similarity between the two proteins (Fig. 3*D*). Thus, at least two of the three major building blocks of retroviral virions are likely to have evolved from cellular proteins.

***MA protein of arenaviruses.*** The matrix protein, Z, of arenaviruses performs multiple functions, one of which is to bridge the viral surface glycoprotein, the viral ribonucleoprotein, and the host cell budding machinery (92). Similar to the retroviral MA, the N terminus of Z protein is myristoylated, facilitating its membrane anchoring and intracellular targeting, self-assembly, and interaction with other viral proteins. Structural studies have shown that LASV Z protein contains a typical Zn-binding RING domain (93), and an HHpred search retrieves with high probability (99.1%) E3 ubiquitin ligases and other cellular RING domains proteins as close homologs of the LASV Z (Fig. S4). Pairwise structural comparison of the LASV Z protein with the RING domain of human E3 ubiquitin ligase (PDB ID code: 4V3L) using DALI returned a Z score of 4.3. Notably, BLASTP searches seeded with the LASV Z protein yielded multiple significant hits to cellular multidomain proteins. For instance, a protein from plants (XP_018837173) was retrieved with the E value of 8e-05 and showed 37% identity to the LASV Z. Considering that among viral matrix proteins the RING domain is restricted to arenaviruses but otherwise is widespread in eukaryotic proteins, it is highly probable that a cellular RING domain protein has been exapted as the matrix protein in the ancestor of arenaviruses, following or concomitant with its diversification from bunyaviruses.

***Matrix proteins of mononegaviruses.*** Like many other enveloped viruses, members of the order Mononegavirales encode matrix proteins that direct virion assembly and budding. Structures of the matrix proteins are available for mononegaviruses of the families *Filoviridae*, *Bornaviridae*, *Paramyxoviridae*, *Pneumoviridae*, and *Rhabdoviridae*. The rhabdovirus matrix protein has a unique fold and is unrelated to the matrix proteins of other mononegaviruses (94). The latter are homologous to each other and consist of one (bornaviruses) or two (filoviruses, paramyxoviruses, and pneumoviruses) domains with similar β-sandwich folds, suggesting gene duplication during evolution (95, 96). Analysis of the mononegaviral (except for rhabdoviral) matrix protein structures uncovered unexpected similarity to cyclophilins. Cyclophilins are ubiquitous cellular proteins that possess peptidyl-prolyl-isomerase activity and participate in protein folding; these proteins also are receptors for the immunosuppressive drug cyclosporin A, which gave them their name (97). Fig. 4 shows a comparison between the N-terminal domain of the Ebola virus (EBOV) matrix protein and cyclophilin C (CypC). The match between the EBOV matrix protein and CypC was obtained with the low but significant Z score of 2.3. It should be noted that in the same search seeded with the EBOV matrix protein homologs from other mononegaviruses were obtained with similarly low Z scores. For instance, matrix proteins from human respiratory syncytial virus (family *Pneumoviridae*) (PDB ID code: 2VQP) and Borna disease virus (*Bornaviridae*) (PDB ID code: 3F1J) were matched to the EBOV matrix protein with the Z scores of 3.5. Nevertheless, visual inspection of the matrix and CypC proteins further confirmed the validity of the DALI matches. The main difference between the EBOV matrix protein and CypC is the presence of an additional β-hairpin in the structure of the latter protein (Fig. 4). Matches to cyclophilins were also obtained when DALI searches were seeded with structures of matrix proteins from other mononega viruses (Z scores of 2.5–3.1). Notably, cyclophilins are known to play an important role in viral infections (98). In particular, cyclophilin A (CypA) is incorporated into virions by binding to capsids or nucleocapsids of many unrelated viruses, including HIV-1 (*Retroviridae*), vesicular stomatitis virus (*Rhabdoviridae*), vaccinia virus (*Poxviridae*), and severe acute respiratory syndrome coronavirus (SARS-CoV; *Coronaviridae*) (98). All members of the Mononegavirales, including rhabdoviruses, encode homologous NC proteins (99), the unrelated matrix proteins notwithstanding. Binding of CypA to the nucleocapsid of rhabdoviruses (100) resembles the interaction between the nucleocapsid and the matrix protein of other mononegaviruses (101). Thus, the matrix protein-encoding gene of mononegaviruses likely evolved from a cyclophilin gene acquired from the host. The alternative possibility, i.e., that matrix protein of mononegaviruses is at the origin of cyclophilins, is hardly possible, given the ubiquity of the latter in cellular organisms and its scarcity among viruses. The cellular cyclophilin that gave rise to the matrix protein might have interacted with the viral nucleocapsid, as in the case of rhabdoviruses. Given that, in RNA-dependent RNA polymerase (RdRp)-based phylogenies, rhabdoviruses do not occupy the basal

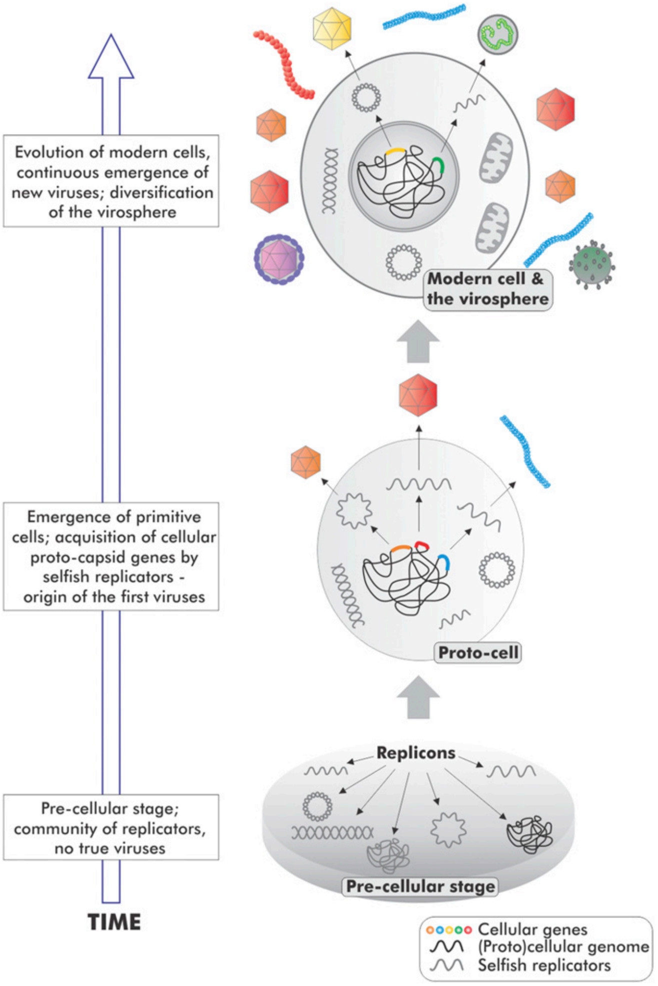
Evolution of modern cells, continuous emergence of new viruses; diversification of the virosphere
Modern cell & the virosphere
Emergence of primitive cells; acquisition of cellular proto-capsid genes by selfish replicators - origin of the first viruses
Proto-cell
Replicons
Pre-cellular stage; community of replicators, no true viruses
Pre-cellular stage
TIME
Cellular genes
(Proto)cellular genome
Selfish replicators

**Fig. 5.** A scenario for the origin of viruses from selfish replicators upon acquisition of capsid protein genes from cellular life forms at different stages of evolution.

position within the order Mononegavirales (102), it is likely that the ancestral cyclophilin-like matrix protein-encoding gene was replaced in the rhabdovirus ancestor by a nonhomologous gene with similar properties. It would be interesting to test whether any of the matrix proteins of mononegaviruses retained the peptidyl-prolyl-isomerase activity typical of cyclophilins.

Given the structural similarities between other cellular proteins and viral CPs, a scenario emerges in which bona fide viruses evolved on multiple, independent occasions by recruiting diverse host proteins that became major virion components (Fig. 5).

## Concluding Remarks

The findings on the apparent independent recruitment of diverse proteins from cellular organisms for the role of CPs and other major virion proteins compel us to adjust our concept of the virus world (4, 6). The grand scenario for virus evolution becomes a hybrid between the virus-first and escape hypotheses (Fig. 5). Given the lack of close cellular homologs for the hallmark virus proteins involved in genome replication, the diversity of viral genomic strategies, and the general considerations on the early stages in the evolution of replicating genomes that implicate ensembles of small, partially autonomous, virus-like genetic elements (103), the origin of viral replicative modules seems likely to hark all the way back to the precellular era. At that stage, some of these primordial replicators coalesced and gave rise to the first cellular genomes, whereas others became genetic parasites. Conceivably, however, such parasites gave rise to true viruses only after the emergence of cells. Viruses emerged through the recruitment of cellular carbohydrate- or nucleic acid-binding proteins as CPs and other major virion proteins. Given the simple, symmetrical, thermodynamically favored structures of the widespread capsids, such as icosahedra or helices, the structural requirements for such exaptation might not have been prohibitive. Indeed, this view is compatible with the scenario of multiple recruitment events occurring throughout the course of evolution of life. Some of the CPs were coopted at the earliest stages of cellular evolution, as is likely to have been the case for the SJR and DJR folds. Other structural proteins were likely adapted at the root of particular cellular domains, as is likely the case of retroviral Gag, given the wide spread and diversity of retroviruses in Eukarya, in sharp contrast to their absence in bacteria and archaea (11). Finally, in all likelihood, some virion components have evolved rather recently, e.g., protease recruitment for the alphavirus capsids or the RING domain and cyclophilin exaptation as the matrix proteins of arenaviruses and mononegaviruses, respectively. Notably, virus-like structures also appear to have evolved in the cellular context, e.g., bacterial microcompartments, large icosahedral organelles that, unlike encapsulins, are built from proteins with no identifiable homologs in the viral world (18). The evolution of virions certainly is not a one-way street. Once multiple capsids evolved, viral structural proteins were recruited for cellular functions on multiple occasions. A well-known example is the exaptation of retroviral envelope proteins for the role of mammalian placental receptors, syncytins (104). The history of virions can be considered the ultimate manifestation of virus–host coevolution.

## Materials and Methods

**Sequence and Structural Data.** To analyze the diversity of folds in the major virion proteins, structural and sequence information for representative proteins was collected from all currently recognized viral families and unassigned genera (25). The list of approved virus taxa was downloaded from the International Committee on the Taxonomy of Viruses (ICTV) website (https://talk.ictvonline.org/files/master-species-lists/). In accordance with the recently proposed taxonomy, which has been approved by the Executive Committee of the ICTV (105), different genera within the family Bunyaviridae were considered as separate families. In total, the analyzed dataset covered 135 virus taxa (117 families and 18 unassigned genera). The genera *Dinodnavirus* and *Rhizidiovirus* were excluded from the analysis because of the complete lack of available sequence or structural information. The protein structures and sequences were downloaded from the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank (www.rcsb.org) and the National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/protein/), respectively.

**Sequence and Structure Analysis.** Structure-based searches were performed using the DALI server (34, 106). Structural similarities between cellular and viral proteins were evaluated based on the DALI Z score, which is a measure of the quality of the structural alignment. Z scores above 2, i.e., two SDs above expected, are usually considered significant (107). The relevance of the matches was evaluated further by visual inspection of structural alignments between the cellular and viral proteins. Structural homologs were additionally searched for using the TopSearch server (https://topsearch.services.came.sbg.ac.at/). Structural similarity matrices from all-against-all structure comparisons as well as corresponding dendrograms were obtained using the latest release of the DALI server (34). Structures were aligned using the MatchMaker algorithm implemented in University of California, San Francisco (UCSF) Chimera (108) and were visualized using the same software. Sequence-similarity searches were performed using PSI-BLAST (109) against the nonredundant protein sequence database at the NCBI. For distant sequence similarity detection, homologous sequences of viral proteins were aligned using MUSCLE (110), and the resulting multiple sequence alignments (or individual sequences) were used as seeds in profile-against-profile searches using HHpred (111).

1. Danovaro R, et al. (2016) Virus-mediated archaeal hecatomb in the deep seafloor. *Sci Adv* 2(10):e1600492.
2. Chow CE, Suttle CA (2015) Biogeography of viruses in the sea. *Annu Rev Virol* 2(1): 41–66.
3. Cobián Güemes AG, et al. (2016) Viruses as winners in the game of life. *Annu Rev Virol* 3(1):197–214.
4. Koonin EV, Dolja VV (2013) A virocentric perspective on the evolution of life. *Curr Opin Virol* 3(5):546–557.
5. Forterre P (2006) The origin of viruses and their possible roles in major evolutionary transitions. *Virus Res* 117(1):5–16.
6. Koonin EV, Senkevich TG, Dolja VV (2006) The ancient virus world and evolution of cells. *Biol Direct* 1:29.
7. Abergel C, Legendre M, Claverie JM (2015) The rapidly expanding universe of giant viruses: Mimivirus, Pandoravirus, Pithovirus and Mollivirus. *FEMS Microbiol Rev* 39(6): 779–796.
8. Nasir A, Caetano-Anollés G (2015) A phylogenomic data-driven exploration of viral origins and evolution. *Sci Adv* 1(8):e1500527.
9. Koonin EV, Krupovic M, Yutin N (2015) Evolution of double-stranded DNA viruses of eukaryotes: From bacteriophages to transposons to giant viruses. *Ann N Y Acad Sci* 1341:10–24.
10. Forterre P, Krupovic M (2012) The origin of virions and virocells: The Escape hypothesis revisited. *Viruses: Essential Agents of Life*, ed Witzany G (Springer Science+ Business Media, Dordrecht), pp 43–60.
11. Koonin EV, Dolja VV, Krupovic M (2015) Origins and evolution of viruses of eukaryotes: The ultimate modularity. *Virology* 479–480:2–25.
12. Kazlauskas D, Krupovic M, Venclovas Č (2016) The logic of DNA replication in double-stranded DNA viruses: Insights from global analysis of viral genomes. *Nucleic Acids Res* 44(10):4551–4564.
13. Forterre P (2006) Three RNA cells for ribosomal lineages and three DNA viruses to replicate their genomes: A hypothesis for the origin of cellular domain. *Proc Natl Acad Sci USA* 103(10):3669–3674.
14. Krupovic M, Bamford DH (2009) Does the evolution of viral polymerases reflect the origin and evolution of viruses? *Nat Rev Microbiol* 7(3):250–, author reply 250.
15. Cheng S, Brooks CL, 3rd (2013) Viral capsid proteins are segregated in structural fold space. *PLOS Comput Biol* 9(2):e1002905.
16. Sabath N, Wagner A, Karlin D (2012) Evolution of viral proteins originated de novo by overprinting. *Mol Biol Evol* 29(12):3767–3780.
17. Jalasvuori M, Mattila S, Hoikkala V (2015) Chasing the origin of viruses: Capsid-forming genes as a life-saving preadaptation within a community of early replicators. *PLoS One* 10(5):e0126094.
18. Bobik TA, Lehman BP, Yeates TO (2015) Bacterial microcompartments: Widespread prokaryotic organelles for isolation and optimization of metabolic pathways. *Mol Microbiol* 98(2):193–207.
19. Giessen TW (2016) Encapsulins: Microbial nanocompartments with applications in biomedicine, nanobiotechnology and materials science. *Curr Opin Chem Biol* 34: 1–10.
20. Abrescia NG, Bamford DH, Grimes JM, Stuart DI (2012) Structure unifies the viral universe. *Annu Rev Biochem* 81:795–822.
21. Krupovic M, Bamford DH (2008) Virus evolution: How far does the double beta-barrel viral lineage extend? *Nat Rev Microbiol* 6(12):941–948.
22. Rossmann MG, Johnson JE (1989) Icosahedral RNA virus structure. *Annu Rev Biochem* 58:533–573.
23. Krupovic M, Bamford DH (2011) Double-stranded DNA viruses: 20 families and only five different architectural principles for virion assembly. *Curr Opin Virol* 1(2):118–124.
24. Kormelink R, Garcia ML, Goodin M, Sasaya T, Haenni AL (2011) Negative-strand RNA viruses: The plant-infecting counterparts. *Virus Res* 162(1-2):184–202.

25. Adams MJ, et al. (2016) Ratification vote on taxonomic proposals to the International Committee on Taxonomy of Viruses (2016). *Arch Virol* 161(10):2921–2949.
26. King AMQ, Adams MJ, Carstens EB, Lefkowitz EJ (2011) *Virus Taxonomy. Ninth Report of the International Committee on Taxonomy of Viruses* (Elsevier Academic, London).
27. Koonin EV, Dolja VV (2014) Virus world as an evolutionary network of viruses and capsidless selfish elements. *Microbiol Mol Biol Rev* 78(2):278–303.
28. Krupovic M, Dolja VV, Koonin EV (2015) Plant viruses of the Amalgaviridae family evolved via recombination between viruses with double-stranded and negative-strand RNA genomes. *Biol Direct* 10:12.
29. Sabanadzovic S, Valverde RA, Brown JK, Martin RR, Tzanetakis IE (2009) Southern tomato virus: The link between the families Totiviridae and Partitiviridae. *Virus Res* 140(1-2):130–137.
30. Iranzo J, Koonin EV, Prangishvili D, Krupovic M (2016) Bipartite network analysis of the archaeal virosphere: Evolutionary connections between viruses and capsid-less mobile elements. *J Virol* 90(24):11043–11055.
31. Simmonds P, et al. (2017) Consensus statement: Virus taxonomy in the age of metagenomics. *Nat Rev Microbiol* 15(3):161–168.
32. Moreira D, López-García P (2009) Ten reasons to exclude viruses from the tree of life. *Nat Rev Microbiol* 7(4):306–311.
33. Hospenthal MK, et al. (2016) Structure of a chaperone-usher pilus reveals the molecular basis of rod uncoiling. *Cell* 164(1-2):269–278.
34. Holm L, Laakso LM (2016) Dali server update. *Nucleic Acids Res* 44(W1):W351–W355.
35. Boraston AB, Bolam DN, Gilbert HJ, Davies GJ (2004) Carbohydrate-binding modules: Fine-tuning polysaccharide recognition. *Biochem J* 382(Pt 3):769–781.
36. Holyoak T, Kettner CA, Petsko GA, Fuller RS, Ringe D (2004) Structural basis for differences in substrate selectivity in Kex2 and furin protein convertases. *Biochemistry* 43(9):2412–2421.
37. Foophow T, et al. (2010) Crystal structure of a subtilisin homologue, Tk-SP, from Thermococcus kodakaraensis: Requirement of a C-terminal beta-jelly roll domain for hyperstability. *J Mol Biol* 400(4):865–877.
38. Prado A, Ramos I, Frehlick LJ, Muga A, Ausió J (2004) Nucleoplasmin: A nuclear chaperone. *Biochem Cell Biol* 82(4):437–445.
39. Namboodiri VM, Dutta S, Akey IV, Head JF, Akey CW (2003) The crystal structure of Drosophila NLP-core provides insight into pentamer formation and histone binding. *Structure* 11(2):175–186.
40. Eitoku M, Sato L, Senda T, Horikoshi M (2008) Histone chaperones: 30 years from isolation to elucidation of the mechanisms of nucleosome assembly and disassembly. *Cell Mol Life Sci* 65(3):414–444.
41. Locksley RM, Killeen N, Lenardo MJ (2001) The TNF and TNF receptor superfamilies: Integrating mammalian biology. *Cell* 104(4):487–501.
42. Liu Y, et al. (2002) Crystal structure of sTALL-1 reveals a virus-like assembly of TNF family ligands. *Cell* 108(3):383–394.
43. Thompson BM, Stewart GC (2008) Targeting of the BclA and BclB proteins to the Bacillus anthracis spore surface. *Mol Microbiol* 70(2):421–434.
44. Dunwell JM, Purvis A, Khuri S (2004) Cupins: The most functionally diverse protein superfamily? *Phytochemistry* 65(1):7–17.
45. Aik W, McDonough MA, Thalhammer A, Chowdhury R, Schofield CJ (2012) Role of the jelly-roll fold in substrate binding by 2-oxoglutarate oxygenases. *Curr Opin Struct Biol* 22(6):691–700.
46. Klose RJ, Zhang Y (2007) Regulation of histone methylation by demethylimination and demethylation. *Nat Rev Mol Cell Biol* 8(4):307–318.
47. Neu U, Bauer J, Stehle T (2011) Viruses and sialic acids: Rules of engagement. *Curr Opin Struct Biol* 21(5):610–618.
48. Sarker S, et al. (2016) Structural insights into the assembly and regulation of distinct viral capsid complexes. *Nat Commun* 7:13014.
49. Olson AJ, Bricogne G, Harrison SC (1983) Structure of tomato busy stunt virus IV. The virus particle at 2.9 A resolution. *J Mol Biol* 171(1):61–93.
50. Krupovic M (2012) Recombination between RNA viruses and plasmids might have played a central role in the origin and evolution of small DNA viruses. *BioEssays* 34(10):867–870.
51. Krupovic M (2013) Networks of evolutionary interactions underlying the polyphyletic origin of ssDNA viruses. *Curr Opin Virol* 3(5):578–586.
52. Diemer GS, Stedman KM (2012) A novel virus genome discovered in an extreme environment suggests recombination between unrelated groups of RNA and DNA viruses. *Biol Direct* 7:13.
53. Krupovic M, Ravantti JJ, Bamford DH (2009) Geminiviruses: A tale of a plasmid becoming a virus. *BMC Evol Biol* 9:112.
54. Roux S, et al. (2013) Chimeric viruses blur the borders between the major groups of eukaryotic single-stranded DNA viruses. *Nat Commun* 4:2700.
55. Kazlauskas D, et al. (2017) Evolutionary history of ssDNA bacilladnaviruses features horizontal acquisition of the capsid gene from ssRNA nodaviruses. *Virology* 504:114–121.
56. Bamford DH, Grimes JM, Stuart DI (2005) What does structure tell us about virus evolution? *Curr Opin Struct Biol* 15(6):655–663.
57. Krupovic M, Koonin EV (2015) Polintons: A hotbed of eukaryotic virus, transposon and plasmid evolution. *Nat Rev Microbiol* 13(2):105–115.
58. Iranzo J, Krupovic M, Koonin EV (2016) The double-stranded DNA virosphere as a modular hierarchical network of gene sharing. *MBio* 7(4):e00978-16.
59. Krupovic M, Bamford DH, Koonin EV (2014) Conservation of major and minor jelly-roll capsid proteins in Polinton (Maverick) transposons suggests that they are bona fide viruses. *Biol Direct* 9:6.
60. Abrescia NG, et al. (2008) Insights into virus evolution and membrane biogenesis from the structure of the marine lipid-containing bacteriophage PM2. *Mol Cell* 31(5):749–761.
61. Pawlowski A, Rissanen I, Bamford JK, Krupovic M, Jalasvuori M (2014) Gammasphaerolipovirus, a newly proposed bacteriophage genus, unifies viruses of halophilic archaea and thermophilic bacteria within the novel family Sphaerolipoviridae. *Arch Virol* 159(6):1541–1554.
62. Gil-Carton D, et al. (2015) Insight into the assembly of viruses with vertical single $\beta$-barrel major capsid proteins. *Structure* 23(10):1866–1877.
63. Rissanen I, et al. (2013) Bacteriophage P23-77 capsid protein structures reveal the archetype of an ancient branch from a major virus lineage. *Structure* 21(5):718–726.
64. Strömsten NJ, Bamford DH, Bamford JK (2005) In vitro DNA packaging of PRD1: A common mechanism for internal-membrane viruses. *J Mol Biol* 348(3):617–629.
65. Wikoff WR, et al. (2000) Topologically linked protein rings in the bacteriophage HK97 capsid. *Science* 289(5487):2129–2133.
66. Suhanovsky MM, Teschke CM (2015) Nature's favorite building block: Deciphering folding and capsid assembly of proteins with the HK97-fold. *Virology* 479-480:487–497.
67. Krupovic M, Prangishvili D, Hendrix RW, Bamford DH (2011) Genomics of bacterial and archaeal viruses: Dynamics within the prokaryotic virosphere. *Microbiol Mol Biol Rev* 75(4):610–635.
68. Baker ML, Jiang W, Rixon FJ, Chiu W (2005) Common ancestry of herpesviruses and tailed DNA bacteriophages. *J Virol* 79(23):14967–14970.
69. Brown JC, Newcomb WW (2011) Herpesvirus capsid assembly: Insights from structural analysis. *Curr Opin Virol* 1(2):142–149.
70. Akita F, et al. (2007) The crystal structure of a virus-like particle from the hyperthermophilic archaeon Pyrococcus furiosus provides insight into the evolution of viruses. *J Mol Biol* 368(5):1469–1483.
71. McHugh CA, et al. (2014) A virus capsid-like nanocompartment that stores iron and protects bacteria from oxidative stress. *EMBO J* 33(17):1896–1911.
72. Sutter M, et al. (2008) Structural basis of enzyme encapsulation into a bacterial nanocompartment. *Nat Struct Mol Biol* 15(9):939–947.
73. Heinemann J, et al. (2011) Fossil record of an archaeal HK97-like provirus. *Virology* 417(2):362–368.
74. Gorbalenya AE, Donchenko AP, Koonin EV, Blinov VM (1989) N-terminal domains of putative helicases of flavi- and pestiviruses may be serine proteases. *Nucleic Acids Res* 17(10):3889–3897.
75. Hahn CS, Strauss JH (1990) Site-directed mutagenesis of the proposed catalytic amino acids of the Sindbis virus capsid protein autoprotease. *J Virol* 64(6):3069–3073.
76. Choi HK, Lu G, Lee S, Wengler G, Rossmann MG (1997) Structure of Semliki Forest virus core protein. *Proteins* 27(3):345–359.
77. Kim JL, et al. (1996) Crystal structure of the hepatitis C virus NS3 protease domain complexed with a synthetic NS4A cofactor peptide. *Cell* 87(2):343–355.
78. Love RA, et al. (1996) The crystal structure of hepatitis C virus NS3 proteinase reveals a trypsin-like fold and a structural zinc binding site. *Cell* 87(2):331–342.
79. Vaney MC, Rey FA (2011) Class II enveloped viruses. *Cell Microbiol* 13(10):1451–1459.
80. Dokland T, et al. (2004) West Nile virus core protein; tetramer structure and ribbon formation. *Structure* 12(7):1157–1163.
81. Ruiz-Guillen M, et al. (2016) Capsid-deficient alphaviruses generate propagative infectious microvesicles at the plasma membrane. *Cell Mol Life Sci* 73(20):3897–3916.
82. Guo Y, et al. (2012) Crimean-Congo hemorrhagic fever virus nucleoprotein reveals endonuclease activity in bunyaviruses. *Proc Natl Acad Sci USA* 109(13):5046–5051.
83. Hastie KM, King LB, Zandonatti MA, Saphire EO (2012) Structural basis for the dsRNA specificity of the Lassa virus NP exonuclease. *PLoS One* 7(8):e44211.
84. 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 USA* 108(6):2396–2401.
85. Krupovic M, Cvirkaite-Krupovic V, Prangishvili D, Koonin EV (2015) Evolution of an archaeal virus nucleocapsid protein from the CRISPR-associated Cas4 nuclease. *Biol Direct* 10:65.
86. Mattei S, Schur FK, Briggs JA (2016) Retrovirus maturation-an extraordinary structural transformation. *Curr Opin Virol* 18:27–35.
87. Bell NM, Lever AM (2013) HIV Gag polyprotein: Processing and early viral particle assembly. *Trends Microbiol* 21(3):136–144.
88. Cai M, Huang Y, Craigie R, Clore GM (2010) Structural basis of the association of HIV-1 matrix protein with DNA. *PLoS One* 5(12):e15675.
89. Collins T, Sander TL (2013) The superfamily of SCAN domain containing zinc finger transcription factors. *Madame Curie Bioscience Database [Internet]*. Available at https://www.ncbi.nlm.nih.gov/books/NBK6264/. Accessed February 24, 2017.
90. Kaneko-Ishino T, Ishino F (2012) The role of genes domesticated from LTR retrotransposons and retroviruses in mammals. *Front Microbiol* 3:262.
91. Ivanov D, Stone JR, Maki JL, Collins T, Wagner G (2005) Mammalian SCAN domain dimer is a domain-swapped homolog of the HIV capsid C-terminal domain. *Mol Cell* 17(1):137–143.
92. Fehling SK, Lennartz F, Strecker T (2012) Multifunctional nature of the arenavirus RING finger protein Z. *Viruses* 4(11):2973–3011.
93. Hastie KM, et al. (2016) Crystal structure of the oligomeric form of Lassa virus matrix protein Z. *J Virol* 90(9):4556–4562.
94. Graham SC, et al. (2008) Rhabdovirus matrix protein structures reveal a novel mode of self-association. *PLoS Pathog* 4(12):e1000251.
95. Battisti AJ, et al. (2012) Structure and assembly of a paramyxovirus matrix protein. *Proc Natl Acad Sci USA* 109(35):13996–14000.
96. Leyrat C, Renner M, Harlos K, Huiskonen JT, Grimes JM (2014) Structure and self-assembly of the calcium binding matrix protein of human metapneumovirus. *Structure* 22(1):136–148.
97. Stamnes MA, Rutherford SL, Zuker CS (1992) Cyclophilins: A new family of proteins involved in intracellular folding. *Trends Cell Biol* 2(9):272–276.
98. Zhou D, Mei Q, Li J, He H (2012) Cyclophilin A and viral infections. *Biochem Biophys Res Commun* 424(4):647–650.
99. Sun Y, Guo Y, Lou Z (2012) A versatile building block: The structures and functions of negative-sense single-stranded RNA virus nucleocapsid proteins. *Protein Cell* 3(12):893–902.

100. Bose S, Mathur M, Bates P, Joshi N, Banerjee AK (2003) Requirement for cyclophilin A for the replication of vesicular stomatitis virus New Jersey serotype. *J Gen Virol* 84(Pt 7): 1687–1699.

101. Liljeroos L, Huiskonen JT, Ora A, Susi P, Butcher SJ (2011) Electron cryotomography of measles virus reveals how matrix protein coats the ribonucleocapsid within intact virions. *Proc Natl Acad Sci USA* 108(44):18085–18090.

102. Li CX, et al. (2015) Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. *eLife* 4:4.

103. Koonin EV, Martin W (2005) On the origin of genomes and cells within inorganic compartments. *Trends Genet* 21(12):647–654.

104. Cornelis G, et al. (2015) Retroviral envelope gene captures and syncytin exaptation for placentation in marsupials. *Proc Natl Acad Sci USA* 112(5):E487–E496.

105. Briese T, et al. (2016) ICTV taxonomic proposal 2016.030a-vM.A.v6.Bunyavirales. Create the order *Bunyavirales*, including eight new families, and one renamed family. Available at https://talk.ictvonline.org/files/proposals/taxonomy_proposals_plant1/m/plant04/6471#. Accessed February 24, 2017.

106. Holm L, Rosenstrom P (2010) Dali server: Conservation mapping in 3D. *Nucleic Acids Res* 38(Web Server issue):W545–W549.

107. Holm L, Sander C (1995) Dali: A network tool for protein structure comparison. *Trends Biochem Sci* 20(11):478–480.

108. Pettersen EF, et al. (2004) UCSF Chimera–a visualization system for exploratory research and analysis. *J Comput Chem* 25(13):1605–1612.

109. Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25(17):3389–3402.

110. Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32(5):1792–1797.

111. Söding J (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics* 21(7):951–960.