The logic of virus evolution

Eugene V Koonin, Valerian V Dolja, Mart Krupovic

Cell Host Microbe. 2022; 30(7):917-929. doi: 10.1016/j.chom.2022.06.008.

The Logic of Virus Evolution

Eugene V. Koonin[1*#], Valerian V. Dolja[1,2] and Mart Krupovic[3*]

[1] National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD 20894, USA

[2]Department of Botany and Plant Pathology, Oregon State University, OR 97331, USA

[3]Institut Pasteur, Université Paris Cité, CNRS UMR6047, Archaeal Virology Unit, F-75015 Paris, France

* - Correspondence to koonin@ncbi.nlm.nih.gov or mart.krupovic@pasteur.fr

- Lead contact

Summary
Viruses are obligate intracellular parasites. Despite their dependence on host cells, viruses are evolutionarily autonomous, with their own genomes and evolutionary trajectories locked in arms races with the hosts. Here, we discuss a simple functional logic to explain virus macroevolution that appears to define the course of virus evolution. A small core of virus hallmark genes that are responsible for genome replication apparently descended from primordial replicators, whereas most virus genes, starting with those encoding capsid proteins, were subsequently acquired from hosts. The oldest of these acquisitions antedate the last universal cellular ancestor (LUCA). Host gene capture followed two major routes: convergent recruitment of genes with functions that directly benefit virus reproduction and exaptation when host proteins are repurposed for unique virus functions. These forms of host protein recruitment by viruses result in different levels of similarity between virus and host homologs, with the exapted ones often changing beyond easy recognition.

Keywords: capsid proteins; exaptation; origins of viruses; virus evolution; virus structure.

Introduction

Theoretical argument and empirical evidence converge to suggest that all cellular life forms, with the possible exception of some intracellular parasites, host multiple viruses (Koonin et al., 2020a; Forterre and Prangishvili, 2009). Viruses display enormous diversity that exceeds the diversity of their cellular hosts in at least two dimensions. First, whereas cellular life forms employ a uniform scheme of genetic information storage and expression, with double-stranded (ds) DNA genome transcribed into various RNAs, including mRNAs, that are translated into proteins, viruses effectively make use of all information transmission schemes that are possible with the two types of nucleic acids (Koonin et al., 2021b; Baltimore, 1971). Second, virus genes typically evolve much faster than their cellular homologs (Domingo et al., 2021; Aiewsakun and Katzourakis, 2016), enabling virus gene sequences to span comparatively larger regions of the sequence space. Indeed, sequence diversity of a single family of viruses can exceed the diversity of entire domains of cellular life (Mihara et al., 2018), rendering the recognition of deep evolutionary relationships and ancestry of virus genes particularly challenging.

The range of virus genome sizes spans about 3 orders of magnitude, from less than 2 kilobases (kb) in the smallest single-stranded (ss) RNA and DNA viruses to more than 2 megabases (Mb) in the giant pandoraviruses (Legendre et al., 2018). Thus, the virosphere encompasses both the simplest of the known protein-coding replicators and the giant viruses, which surpass some prokaryotes and unicellular eukaryotes in genome size and complexity.

The standard concept in virus evolution is that viruses are locked with cellular hosts in a perennial arms race (Koonin et al., 2020b). During this race, hosts evolve and diversify antivirus defense mechanisms, whereas viruses respond with evolving counter-defense systems that suppress the host defenses. Complementary to the arms race and intertwined with it is the extensive gene flow between viruses and their hosts. Viruses frequently acquire host genes and employ them for roles in virus reproduction and virus-host interaction, and conversely, hosts capture virus genes and repurpose them for defense and other cellular functions (Koonin et al., 2020b). Compared to horizontal gene transfer between organisms, the gene flow between viruses and hosts is facilitated by the key feature of virus reproduction, namely, that it occurs strictly within cells, putting replicating virus genomes in close contact with the genome and mRNAs of the host.

In this Perspective, we examine the capture of host genes by viruses and how this process shaped virus genomes through the nearly 4 billion years of virus-host coevolution. We provide support for the idea that gene gain, which produced the astonishing diversity of virus genomes, is governed by simple logic of adaptation of cellular proteins for functions that are essential or beneficial for virus reproduction. Such adaptation involves either direct aquisition of a cellular function or exaptation, whereby the function of the acquired protein changes, often accompanied by elimination of the original activity, although the protein structure is largely preserved. This logic of repurposing cellular proteins determines the trajectories of virus macroevolution including the emergence of new groups of viruses, in an interplay with the arms race.

Virus Hallmark Proteins (VHPs) and their Origins

Comparative analysis of virus proteomes led to the delineation of a small set of Virus Hallmark Proteins (VHP). These proteins are widely distributed among diverse viruses and play key roles in the replication of virus genomes ('replication-associated VHPs') and formation of virions, including genome packaging into capsids ('structural VHPs') (Koonin et al., 2020a). The observations on the provenance of these two functional classes of VHPs, those involved in virus genome replication and in virion formation, strongly suggest different routes of evolution and translate into a parsimonious scenario for the origin of viruses (Krupovic et al., 2019).

The replication VHPs, particularly, RNA-dependent RNA polymerase (RdRP) and reverse transcriptase (RT), but also, protein-primed DNA polymerase, rolling-circle replication endonuclease (RCRE) and superfamily 3 helicase (S3H), have only distant homologs among cellular proteins, suggestive of ancient origins, most likely, antedating the Last Universal Cellular Ancestor (LUCA) (Krupovic et al., 2019). Many evolutionary reconstructions performed using widely different computational approaches indicate that the LUCA was a population of organisms comparable in complexity with the extant prokaryotes (Krupovic et al., 2020a). In particular, multiple polymerases and helicases were apparently already represented in the LUCA. Furthermore, RdRP and RT are the types of replicative enzymes that are likely to have evolved during the transitions from the hypothetical primordial RNA world to the RNA-protein stage, and then, to the DNA-RNA-protein life that spawned LUCA and the subsequent evolution of cellular organisms.

Underneath all the diversity of the virus replication enzymes, there is striking unity: the structural core of all these proteins is the RNA Recognition Motif (RRM) domain which binds nucleic acids in a great variety of proteins (Krupovic et al., 2019). In addition to RNA-binding, this domain evolved the catalytic sites for nucleotide polymerization as well as nuclease activity in the virus and some cellular replication enzymes. The relative simplicity, ubiquity and enormous versatility of the RRM domain implies that this might have been one of the earliest protein domains to evolve, possibly, serving as a cofactor to ribozyme polymerases in the late RNA world. The diversification of the RRM domain, yielding RdRP, RT and other virus as well as cellular replicative proteins, probably occurred within a primordial pool of replicators that inhabited protocells, prior to the establishment of the modern-type replication of DNA genomes and, accordingly, prokaryote-like cells.

Structural VHPs: ancient exaptation

Virus capsid proteins (CP) are remarkably diverse (Krupovic and Koonin, 2017). Because the amino acid sequences of CPs typically evolve much faster than those of replication-associated VHPs, the principal approach employed to establish evolutionary relationship among CPs of different groups of viruses and infer their origins is protein structure

comparison. Altogether, about 20 major CP groups with distinct structures are currently known (Krupovic and Koonin, 2017). Importantly, however, the abundances of different capsid shapes and distinct capsid proteins differ widely, so that there are a few common forms and a number of rare and odd ones. The majority of virus capsids are icosahedral and a substantial minority are helical. Both icosahedral and helical capsids are thermodynamically favored during protein aggregation and these forms can be achieved using multiple, unrelated proteins. In particular, at least 10 unrelated protein folds were identified among the major CPs of icosahedral capsids (Krupovic and Koonin, 2017). Of these, only 3 protein families are spread widely enough among viruses to qualify as structural VHPs.

The single β barrel "jelly-roll" (SJR) structure is commonly found in virus CPs. The SJR CP is the major component of the icosahedral capsids of a great variety of positive-sense RNA viruses of the kingdom *Orthornavira* (realm *Riboviria*) and ssDNA viruses in the realm *Monodnaviria* (Figure 1). Distinct varieties of the SJR CP are also present among dsDNA viruses including the kingdom *Helvetiavirae* within the realm *Varidnaviria* and the families *Papillomaviridae, Polyomaviridae* within *Monodnaviria* (Krupovic and Koonin, 2017; Krupovic et al., 2022).

The SJR domain is also found in a broad variety of cellular proteins, most of which bind carbohydrates and some are carbohydrate-active enzymes (Krupovic and Koonin, 2017). Many of the cellular SJR proteins, in particular those of the tumor necrosis factor (TNF) family, form aggregates resembling capsid-like particles (Zhukovsky et al., 2004), and thus appear to be facile material for recruitment as virus CPs. The SJR CPs of eukaryotic RNA and ssDNA viruses and those of prokaryotic microviruses (*Monodnaviria*) and helvetiaviruses (*Varidnaviria*) show the closest structural similarity to different groups of cellular SJR proteins (Krupovic and Koonin, 2017). Together this suggests that SJR CPs have been recruited by viruses from cellular ancestral proteins on at least two, and likely more, independent occasions. To function as CPs, the recruited SJR proteins must acquire the ability to interact with the virus nucleic acid. Typically, this is achieved via positively charged terminal extensions of the CPs, which are missing in their cellular ancestors (Figure 2A; Requiao et al., 2020).

The CPs of the diverse dsDNA viruses in the kingdom *Bamfordvirae* of the realm *Varidnaviria* consist of two "jelly-roll" domains and thus are known as double jelly-roll (DJR) CPs (Figure 1; Krupovic et al., 2022). A search for possible cellular ancestors of the DJR-CP identified several families of such proteins, with DUF2961-family glycoside hydrolases showing the closest structural similarity to the DJR CPs of *Bamfordvirae* (Figure 1; Krupovic et al., 2022).

All of the numerous tailed icosahedral viruses of the realm *Duplodnaviria*, possess homologous CPs with the HK97 fold (after the phage for which the capsid structure was solved first) (Duda and Teschke, 2019). No cellular homologs of the HK97 CPs that could be their direct ancestors have yet been identified (apart from encapsulins that were likely derived from phage CPs), but a more complex relationship with two families of cellular proteins has been discovered (Holm, 2020). Apparently, the HK97 CP evolved from

bacterial proteins known as dodecins, via insertion of an uncharacterized domain (Figure 1). Dodecins are involved in flavin storage in bacteria and form dodecameric spherical structures that can be considered (mini)capsid-like (Grininger et al., 2006).

Recent protein structure comparisons of these 3 most abundant viral CPs highlight one of the major avenues of virus evolution, namely, exaptation. As originally formulated by Gould and Vrba, exaptation involves repurposing of biological entities, such as animal organs, for functions distinct form the original function (Gould and Vrba, 1982; Gould, 1997). At the molecular level, the concept of exaptation readily applies to evolutionary repurposing (refunctionalization) of proteins and RNA molecules. In the case of virus CPs, “pre-adapted” cellular proteins, i.e., endowed with the propensity for aggregation into ordered, symmetrical multimeric structures, and in some cases, carbohydrate affinity, were repurposed for unique virus functions. Such aggregation is a widespread feature of proteins as demonstrated by experiments on construction of artificial virus-like capsids from different enzymes (Bale et al., 2016; Tetter et al., 2021). The changes during exaptation of CPs are substantial and could even include distortions to the protein fold. This results in only moderate structural similarities between viral CPs and their apparent cellular ancestors and usually prevents recognition of ancestral relationships at the sequence level.

Thus, the ancestry of the structural VHPs appears to differ from the likely origins of the key replicative enzymes in protocellular systems. For each of the three widespread CPs that form icosahedral capsids discussed above, specific cellular ancestors could be identified. The respective families of cellular proteins that were refunctionalized by viruses for structural and morphogenetic roles are present and highly diversified in all three domains of life, suggesting that they had already emerged by the time of the LUCA (Krupovic et al., 2020a). Therefore, these types of VHPs, most likely, evolved from cellular ancestors early in the evolution of life, but after the advent of modern-type cells. For other structural components of virions, later origins from cellular ancestors can be inferred as discussed below.

The apparent temporal separation of the origins of the virus replication machinery and the structural proteins implies that, in the early evolution of life, bona fide viruses were preceded by capsid-less, virus-like, selfish genetic elements. Notably, such primordial elements seem to be recapitulated by extant “capsid-less viruses”, in particular, narnaviruses and mitoviruses, which represent simple replicators that only encode the RdRP (Koonin and Dolja, 2014; Koonin et al., 2021a).

Pervasive exaptation of host proteins by viruses

The recruitment of the three hallmark CPs set the stage for continued exaptation of cellular proteins for virus functions as exemplified by the ancient exaptation of the virus genome packaging ATPases (see below). The subsequent, more than 3 billion years long evolution of viruses was replete with additional, diverse cases of exaptation. In general, two modes of exaptation can be distinguished: i) the exapted protein performs a virus-specific function without direct precedent among cellular proteins, and ii) the function of

the exapted protein is adapted to the requirements of virus reproduction, but the biochemical activity persists – in this case, exaptation is manifest at the biological but not at the biochemical level.

Exaptation accompanied by radical functional change

The first route of exaptation that typically involves a major structural change to the protein, beyond straightforward recognition in some cases, is particularly characteristic of virion components, the functions of which are unique to viruses. In addition to the three structural VHPs discussed above, a notable case is presented by the giant pandoraviruses that lost the ancestral DJR CP of *Nucleocytovirecetes* (and with it, the icosahedral capsid itself). Instead, pandoraviruses recruited an inactivated cellular GH16-family glycoside hydrolase unrelated to the DJR CP as one of their two major virion proteins forming asymmetrical, pitcher-like shells and thus recapitulating the ancient route of exaptation (Krupovic et al., 2020b; Legendre et al., 2018). In molliviruses, the closest relatives of pandoraviruses that retain the DJR CP, the ortholog of the inactivated glycoside hydrolase is a minor component of the virion (Krupovic et al., 2020b; Legendre et al., 2015). This case seems to reflect a general trend whereby cellular genes are initially captured and fixed in a virus due to the immediate fitness benefits they provide for certain aspects of the infection cycle, followed by stepwise refunctionalization, first as minor and subsequently as major virion components.

Yet another, more complicated twist in the exaptation of proteins for the function of CP is the recruitment of a chymotrypsin-like protease to replace the ancestral SJR CP in alphaviruses, a genus within *Togaviridae*, a family of animal positive-sense RNA viruses in the phylum *Kitrinoviricota*. The sequence and structure of the alphavirus CP is most similar to that of the protease from another positive-sense RNA virus family, *Flaviviridae*, in which this protein is not a virion component, but rather, performs a function typical of virus proteases, namely, polyprotein processing (see below) (Wahaab et al., 2021). The alphavirus CP performs a single cleavage liberating the CP from the structural polyprotein, followed by inactivation of the protease (Aggarwal et al., 2014). Convergent with the case of the SJR CPs, recruitment of the protease as the alphavirus CP involved acquisition of a positively charged, non-structured N-terminal region (Lulla et al., 2013) (Figure 2A).

A similar evolutionary scenario applies to the origin of the proteins that form the tail tubes of bacterial and archaeal viruses of the class *Caudoviricetes* and are homologous to a virus serine protease with a unique fold (Fokine and Rossmann, 2016). In this case, the directionality of exaptation is unclear, and it cannot be ruled out that the protease evolved from the major tail tube protein. Homologous, enzymatically active proteases of this fold are also encoded by most of the viruses in *Herviviricetes* (second class of the realm *Duplodnaviria*) (Cheng et al., 2004). These proteases, known as assemblins in herpesviruses, catalyze the proteolytic maturation of the major capsid protein or cleavage of the scaffolding protein (Fokine and Rossmann, 2016).

Another remarkable case of exaptation occurred in the evolution of mimiviruses, where two closely related GMC (glucose-methanol-choline)-type oxidoreductases were

repurposed for structural roles, forming the external glycosylated fibrils decorating the icosahedral capsid and the helical nucleocapsid fibers that condense the 1.2 Mb virus genome (Villalta et al., 2022). Although oxidoreductase activity could not be demonstrated, all active site residues are conserved and FAD cofactor essential for activity is stably bound by the virus protein (Klose et al., 2015). The GMC-type oxidoreductases are not conserved in other members of the *Nucleocytoviricota* and likely were captured by the ancestor of mimiviruses from bacteria.

Many more cases of emergence of virion proteins via exaptation of host and virus proteins have been documented (Figure 2B). One notable example is the matrix protein of retroviruses, which is a derivative of the DNA-binding helix-turn-helix domain of integrases (Krupovic and Koonin, 2017). Another case in point are the matrix proteins of the viruses in the order *Mononegavirales* in the phylum *Negarnaviricota* that appear to have been derived from cyclophilins, molecular chaperones with peptidyl-prolyl-isomerase activity (Krupovic and Koonin, 2017). By contrast, the matrix protein of arenaviruses from the order *Bunyavirales* within the *Negarnaviricota* was exapted from a RING domain of E3 ubiquitin ligases (Krupovic and Koonin, 2017). Together with the alphavirus CP discussed above, exaptation of a cyclophilin illustrates a notable trend in the evolution of virus proteins, namely, recruitment of enzymes for structural roles accompanied by loss of the enzymatic activity.

Most of the time, discovery of the ancestry of virus proteins that evolve via exaptation of cellular proteins for functions that are unrelated to the original ones comes as a surprise and often requires application of the most sensitive methods for protein sequence and/or structure comparison. A relevant example is the poxvirus protein F12, which is involved in virus egress from infected cells (Carpentier et al., 2017), that was shown to be a derived, inactivated DNA polymerase (Yutin et al., 2014). Another poxvirus protein with yet unknown functions, F16, is an inactivated serine recombinase (Senkevich et al., 2011). Similar to the case of F12, the herpesvirus protein UL8 is an inactivated family B DNA polymerase, which was recruited to function as a non-enzymatic, yet essential component of the helicase-primase complex that is involved in multiple protein-protein interactions (Kazlauskas and Venclovas, 2014). A remarkable case of enzyme exaptation for a structural role in viruses are the major proteins of the nucleus-like shell formed by certain jumbo phages for protection of the virus DNA from cellular defense systems, such as restriction endonucleases and CRISPR-Cas (Guan and Bondy-Denomy, 2020). Structural comparisons showed that the major shell protein, chimallin, comprises two domains, the N-terminal domain with an α+β fold similar to that of an uncharacterized bacterial protein, and the C-terminal domain that is derived from a GCN5-related N-acetyltransferase most similar to *E. coli* AtaT and homologous to tRNA-acetylating toxins (Laughlin et al., 2022). Notably, as in many other cellular enzymes exapted by viruses, the acetyltransferase active site residues of chimallin are mutated, apparently abrogating the enzymatic activity.

It appears likely that the extent of radical exaptation in virus evolution is substantially underappreciated because of major changes of the involved protein sequences and even structures. Many more such cases can be expected to be uncovered through

comprehensive prediction of virus protein structures using the new, powerful methods, such as AlphaFold2 (Jumper et al., 2021; Mirdita et al., 2022) and RoseTTAFold (Baek et al., 2021). Thus, exaptation of preexisting cellular and virus proteins, including various enzymes, appears to be a major if not the main route of evolution of virus structural proteins.

Exaptation retaining biochemical activity

Numerous cases of exaptation of host proteins for roles in virus reproduction involve less dramatic changes than those observed in structural proteins, whereby the biochemical activity of the proteins is retained although the specific functional context changes. A notable case of such conservative exaptation are the ATPases that function as motors for energy-dependent packaging of virus genomes into capsids. These ATPases are (nearly) universal among the dsDNA viruses of the realms *Varidnaviria* and *Duplodnaviria* and apparently represent ancient acquisitions by viruses that occurred independently, shortly after these viruses emerged as the result of the CP exaptation. Virus genome packaging ATPases show a flavor of exaptation distinct from that of the virion proteins discussed above, and therefore, the ancestry of these virus proteins is more readily traceable. DNA packaging into virions as such is a unique virus function without direct counterparts in cellular life forms. Nevertheless, the packaging ATPases retain significant sequence similarity to their likely cellular ancestors with not only the structural fold but also the specific motifs involved in ATP binding and hydrolysis including the canonical Walker A and B sites being conserved (Gorbalenya and Koonin, 1989). The packaging ATPase of the viruses in the realm *Varidnaviria* belongs to the FtsK superfamily of ATPases (Iyer et al., 2004) that pump bacterial and plasmid DNA into daughter cells during cell division (Guo et al., 2016). Filamentous ssDNA bacteriophages of the kingdom *Loebvirae* (realm *Monodnaviria*) also encode FtsK-like ATPases which, however, instead of packaging the virus dsDNA into icosahedral capsids, pump the virus ssDNA through the cytoplasmic membrane of the host during virion extrusion (Roux et al., 2019). Arguably, the exaptation of this ATPase was facilitated by the mechanistic similarity between the two DNA pumping processes.

In the realm *Duplodnaviria*, the dsDNA-packaging ATPase known as the large terminase subunit, is distantly related to superfamily 2 helicases and contains an additional RNase H fold nuclease domain (Feiss and Rao, 2012). In this case, exaptation involved a more substantial modification of the exapted protein, including fusion of the ATPase and nuclease domains. Nevertheless, the ancestral relationship is readily recognizable through the conservation of the sequence motifs in the ATPase domain.

Most RNA viruses do not encode genome packaging enzymes, and nucleic acids co-assemble with major structural proteins to form virions. The dsRNA bacteriophages of the family *Cystoviridae* (ream *Riboviria*) are the only currently known exception to this trend and encode packaging enzymes related to the superfamily 4 helicases (El Omari et al., 2013).

Superfamily 3 helicase of parvoviruses presents a special case of recent exaptation. Unlike in other eukaryotic ssDNA viruses of the realm *Monodnaviria*, in which S3H is primarily involved in genome replication (Tarasova et al., 2021), in parvoviruses, it is additionally responsible for virus genome packaging into preformed empty capsids (King et al., 2001). Conversely, the poxvirus S3H that is fused to a primase and is essential for genome replication is additionally implicated in the ATP-dependent uncoating of the virus genome upon infection (Kilcher et al., 2014).

Similarly to numerous helicases and other NTPases, all virus packaging ATPases form pentameric or hexameric ring-shaped assemblies with a central channel through which the virus DNA or RNA genome is translocated (Hong et al., 2014). Thus, recruitment of these enzymes for the function in virion assembly capitalizes on the utility of the ternary structures of the ancestral cellular enzymes for this virus function.

Notably, virus helicases were exapted for more than just genome packaging and uncoating. For instance, in alphaviruses, an SF1 helicase nsP2, in addition to genome unwinding, also catalyzes the first step of the capping reaction, namely, dephosphorylation of the triphosphorylated RNA 5’ end, which is performed by a dedicated RNA 5′-triphosphatase in most other viruses and cellular organisms (Ahola et al., 2021).

Another common case of conservative exaptation are proteases that have been captured by diverse viruses and are involved in proteolytic processing of virus protein precursors. In many families of positive-sense RNA viruses in the phyla *Kitrinoviricota* and *Pisuviricota*, either the entire genomic RNA or its large portion encoding non-structural proteins are expressed as a polyprotein that contains one or more protease domains (Koonin et al., 2015). These viruses encode proteases of two expansive families with unrelated folds, papain-like and chymotrypsin-like (MEROPS clans CA and PA, respectively), and some viruses encode both types of proteases. The reverse-transcribing RNA viruses of the kingdom *Pararnaviria* also produce polyproteins that are cleaved by a virus-encoded aspartate protease (MEROPS clan AA) unrelated to other virus proteases (Dunn et al., 2002).

For some of the RNA virus proteases, the origin from specific families of cellular proteases is traceable. In particular, the chymotrypsin-like proteases in the order *Picornavirales* were apparently derived from the HtrA family of bacterial proteases (Koonin et al., 2008). The aspartate protease of the reverse-transcribing viruses seems to originate from the Ddi1 protease, a highly conserved component of the eukaryotic ubiquitin signaling network (Krylov and Koonin, 2001; Sirkis et al., 2006). Given that the ancestral cellular enzymes are not involved in polyprotein processing or capsid protein maturation, which is a quintessential virus function, the numerous cases of protease recruitment by viruses clearly represent exaptation. However, their biochemical activity is retained, which explains the readily detectable sequence conservation.

However difficult staging of major evolutionary events might be in general, in this case, the logic of evolution seems to dictate a specific scenario for the temporal order of the origin of virus proteases and polyproteins. Given that the single, 5'-terminal translation initiation site is an ancestral eukaryotic feature (Hinnebusch and Lorsch, 2012), it appears most likely that virus polyproteins evolved first and were initially processed by a host protease, such as a deubiquitinating enzyme, which was recruited by a virus *in trans*. The protease genes were subsequently captured by viruses, thus providing the capability of efficient polyprotein processing *in cis*.

Although viruses with DNA genomes typically do not encode polyproteins, most of the viruses of eukaryotes in the realm *Varidnaviria* encode a cysteine protease (MEROPS clan CE) that is required for the proteolytic maturation of capsid proteins (Moyer et al., 2016). This protease is specifically related to the eukaryotic deubiquitinating enzyme Ulp1 that probably was recruited by varidnaviruses at an early stage of their evolution in eukaryotes (Koonin and Yutin, 2019). As mentioned above, duplodnaviruses encode assemblin-like serine proteases (MEROPS clan SH) with a unique fold that could have originated in virus genomes (Zuhlsdorf and Hinrichs, 2017). However, in some tailed bacteriophages of the class *Caudoviricetes,* this ancestral protease was replaced on multiple occasions by bacterial ClpP-like proteases (MEROPS clan SK) (Liu and Mushegian, 2004).

Extramural versus intramural exaptation

Viruses have ample opportunities to sample the genes from their host genome, which is typically considerably larger than the virus genome. Such capture of genes 'from the outside' can be dubbed 'extramural exaptation'. However, not only host genes are exapted during virus evolution. Indeed, several well documented cases involve exaptation of virus genes, an evolutionary phenomenon we refer to as 'intramural exaptation', whereby an ancestral virus gene is refunctionalized, often after a duplication, followed by neofunctionalization of one of the copies or subfunctionalization of both (Lynch and Katju, 2004; Lynch et al., 2001), or by evolving an additional function of the same protein. The major tail protein and assemblin of duplodnaviruses as well as a major structural protein of pandoraviruses exapted from an inactivated GH16 glycoside hydrolase discussed above all fall into this category. Another notable example is the exaptation of E1B-55K, a multifunctional non-structural oncoprotein of mastadenoviruses (realm *Varidnaviria*), from a duplicated gene encoding the LH3-like minor capsid protein (Marabini et al., 2021).

In closteroviruses, which are among the viruses with the largest, most complex of the known ssRNA genomes (Dolja et al., 2006), the major CP forming the helical virus body was triplicated and two of the paralogs were exapted for virus cell-to-cell movement (Napuli et al., 2003). Furthermore, in several closteroviruses, the papain-like leader protease was duplicated, with one paralog involved in genome replication and the other one facilitating virus long-distance transport within plants (Liu et al., 2009). From a mechanistic point of view, intramural exaptation involving duplication of preexisting virus genes is likely to be more efficient compared to illegitimate recombination with the host genome, especially for RNA viruses. Experimental evolution studies have shown that at least some gene duplications are deleterious in RNA viruses (Willemsen et al., 2016),

suggesting that there is a substantial selection pressure for neofunctionalization or loss of one of the duplicates. In dsDNA viruses, gene duplications readily occur and can be followed by equally rapid homologous recombination-driven loss of extra gene copies via the evolutionary process dubbed “genomic accordion” (Elde et al., 2012; Filée, 2013).

The other route of intramural exaptation involves recruitment of essential virus proteins involved in genome replication and virion formation for additional roles, a phenomenon known as moonlighting or gene sharing (Copley, 2014; Piatigorsky and Wistow, 1989). The prime examples are virus helicases that, without compromising their roles in genome replication, are adopted for additional ATP-dependent processes, such as genome packaging or uncoating, or capping reactions (see above).

The special case of exaptation of host genes for virus counter-defense

All viruses, without exception, face multiple lines of host defense and launch their own counter-defense programs. The RNA and ssDNA viruses in the realms *Riboviria* and *Monodnaviria* that have small genomes often do not encode dedicated counter-defense proteins, relying instead on moonlighting by essential virus proteins. However, notwithstanding the genome compactness, other viruses of these realms encode dedicated counter-defense proteins, such as RNA interference (RNAi) suppressors in many plant, fungal and animal viruses. These suppressors encompass a variety of functional modes and unrelated protein structures attesting to their convergent evolution in multiple virus taxa (Jin et al., 2022). A substantial fraction of RNAi suppressors contain the dsRNA binding domain (dsRBD), a widespread RNA-binding domain in most cellular life forms (Tian and Mathews, 2003). There is little doubt that the dsRBD-containing RNAi suppressors originated via conservative exaptation, but the exact source is difficult to identify due to the small size of dsRBD and rapid evolution of virus genes.

In contrast, viruses with large dsDNA genomes in the realms *Varidnaviria* and *Duplodnaviria* encompass numerous dedicated counter-defense genes, the nature of which is dictated by host biology. Counter-defense systems have been studied in great detail in chordopoxviruses, in which about 100 of the approximately 200 encoded proteins target host defense (Bratke et al., 2013; Senkevich et al., 2021). Multiple chordopoxvirus proteins block different innate immunity pathways, in particular, the interferon-mediated virus resistance, the ubiquitin-proteasome system, and apoptosis, often acting as dominant negative inhibitors (Bratke et al., 2013). For example, vaccinia virus (VACV) protein E3 contains a dsRBD and functions as a dominant negative inhibitor of the PKR kinase through the formation of unproductive complexes with dsRNA thereby preventing PKR activation (Marq et al., 2009). An unrelated protein, K3, is a structural mimic of the translation initiation factor eIF2a and inhibits PKR via an unrelated mechanism, by directly binding the kinase (Elde et al., 2009). Notably, chordopoxviruses encode four families of defense proteins, each including several paralogs. Proteins in two of these families contain repetitive structures, namely, Kelch and TPR repeats, and the other two families contain Bcl-2 domains inhibiting different stages of apoptosis and PIE (Poxvirus Immune

Evasion) domains, respectively (Senkevich et al., 2021). The Bcl-2 proteins are a typical case of recruitment of host proteins as dominant negative inhibitors of the host defense pathways, in this case, apoptosis (Kvansakul et al., 2017). Phylogenetic analysis shows that these protein families evolved by serial gene duplication during poxvirus evolution, following the initial capture by an ancestral chordopoxvirus (Senkevich et al., 2021). Additionally, some chordopoxviruses encode proteins containing DEATH domain that also function as dominant negative apoptosis inhibitors (Bratke et al., 2013).

The origins of most of the chordopoxviruses counter-defense proteins can be generically traced to animal ancestors, but canot be pinned down to the level of specific genes (Senkevich et al., 2021). Taken together, these observations indicate that typically virus counter-defense proteins are substantially altered upon recruitment by viruses, including core fold rearrangements in some cases. However, there are some exceptions to this trend, when an apparently recently acquired protein retains high sequence similarity to an obvious ancestor in the host, e.g., serpins, inhibitors of caspases that prevent apoptosis (Bratke et al., 2013). The repertoire of poxvirus counter-defense proteins depends on the host biology. Indeed, fish poxviruses, the deepest branch among chordopoxviruses, lack all the characteristic counter-defense proteins present in the rest of the group and instead encode a roughly equivalent number of proteins without detectable similarity in the current databases (Gjessing et al., 2015). By implication, these proteins can be suspected to function in counter-defense, but no specific indications can be gleaned from the analysis of their sequences.

The suite of poxvirus counter-defense genes illustrates a key trend of virus-host coevolution, namely, exaptation of host defense systems or their components for virus functions, primarily counter-defense. Among bacterial and archaeal viruses, more cases of this strategy, dubbed “guns for hire” (Koonin et al., 2020b), have been identified including recruitment of CRISPR systems targeting host defense systems and of mini-CRISPR arrays that hijack host Cas proteins for inter-virus competition (Faure et al., 2019; Medvedeva et al., 2019).

Direct recruitment of host enzymes

Apart from the clear cases of exaptation discussed above that involve either a radical or a more conservative change in function, sequence and structure of the respective proteins, many host proteins were recruited without much functionally relevant modification. Such cases do not seem to qualify as exaptation because, although the recruited proteins function in the context of virus reproduction, their activities are (almost) fully preserved. Not unexpectedly, such direct recruitment of host proteins commonly involves proteins, often with enzyamatic activity, that function in virus genome replication and expression, such as DNA helicases, DNA polymerases (DNAP), primases, DNA ligases, ssDNA binding proteins and more (Kazlauskas et al., 2016). Enzymes of nucleotide biosynthesis, such as nucleoside and nucleotide kinases, thymidylate synthases and ribonucleoitde reductases, belong to the same category of proteins that

perform essentially the same roles in hosts and viruses, and accordingly, in most cases, retain significant sequence similarity (Liu et al., 2021).

In the case of each of these enzymatic activities, different viruses convergently captured distantly related or even non-homologous host genes (Kazlauskas et al., 2016). For example, different members of *Nucleocytoviricota* encode either an NAD-dependent or a distantly related ATP-dependent ligase (or no ligase at all), the former apparently being the ancestral form that was displaced by the ATP-dependent ligase several times during the evolution of these large viruses (Yutin and Koonin, 2009). Similarly, different tailed phages in the class *Caudoviricetes* encode either a B family DNAP or the distantly related family A DNAP, or the unrelated family C DNAP, or no DNAP at all (Kazlauskas et al., 2016; Yutin et al., 2021). The dsDNA viruses in the realms *Varidnaviria* and *Duplodnaviria* also encode either bacterial-type DnaG-like primase or an AEP, and DNA helicases of each of the 4 superfamilies (Kazlauskas et al., 2016).

In most cases, inferring direct ancestors of enzymes recruited by viruses is difficult, but there are notable exceptions. For instance, in archaea, phylogenetic analysis has shown that viruses of different groups have recruited replicative MCM helicases from their respective hosts on several independent occasions (Krupovic et al., 2010). These examples should suffice to show that convergent, independent acquisition of distinct, non-homologous or distantly related enzymes with the same activities is the dominant trend in the evolution of the replication machineries of large dsDNA viruses. Notably, however, none of these enzymes is universal across *Varidnaviria* or *Duplodnaviria*, emphasizing the interchangeability of the respective functionalities between virus-encoded and host proteins, especially, in viruses of prokaryotes that have facile access to the host replication apparatus. Members of *Nucleocytoviricota* that replicate inside virus factories within the cytosol of eukaryotic host cells present a different case; these viruses recruited and retained most if not all proteins required for their replication (Koonin and Yutin, 2019).

Although, as discussed above, a major theme in virus-host coevolution is exaptation of host genes for escaping host defenses, direct recruitment of host enzymes to complement cellular processes is common as well. As in the case of exaptation, the repertoire of recruited genes is determined by host biology. Two iconic cases are the recruitment of photosystem components by cyanophages (Fridman et al., 2017; Sullivan et al., 2006) and the capture of translation system components by members of *Nucleocytoviricota*, some of which encode the full complement of proteins required for translation except for ribosomal proteins (Abrahão et al., 2018; Koonin and Yutin, 2019). In each of these cases, the recruited enzymes do not seem to perform any virus-specific functions, but rather support the functions of the respective host systems, presumably, to avoid their shutdown during virus infection. While ribosome components are conspicuously missing in *Nucleocytoviricota*, some of the large dsDNA bacteriophages did capture genes encoding host ribosomal proteins (Mizuno et al., 2019). In this case, however, it remains unclear whether replacement of host ribosomal proteins by phage-specific ones helps the phage to outcompete cellular translation or complement it.

Dark matter of virus proteomes

Rapid evolution that often obscures ancestry is a salient feature of virus genes, which is further enhanced by exaptation. As a result, evolutionary constraints on protein sequence and structure are substantially relaxed. Thus, virus genomes also contain “dark matter”, genes whose provenance is obscure (Yin and Fischer, 2008). Even the small genomes of many viruses in the realms *Riboviria* and *Monodnaviria* encode some dark matter proteins, often small ones, without any detectable homologs, such as RNAi suppressors discussed above. In large dsDNA genomes of viruses within *Duplodnaviria* and especially *Varidnaviria*, dark matter accounts for a substantial fraction of the genes, up to 90% in the giant pandoraviruses (Legendre et al., 2018).

At least two substantially different routes of evolution are likely to account for the dark matter of virus genomes. The first is radical exaptation of host genes that so far remains undetected. The history of virus genome analysis shows that detection of ancestral relationships for virus proteins critically depends on the sensitivity of the methods used for sequence and structure analysis. Comprehensive application of the recently developed powerful methods for protein structure prediction to virus proteins will uncover numerous unsuspected cases of radical exaptation.

The second scenario for the origin of the dark matter of virus genomes is *de novo* emergence of protein-coding genes. The smoking guns for this route of evolution are proteins that emerge via a process termed “overprinting”. Overprinting refers to a process where nucleotide substitutions occur within preexisting virus genes and subsequent translation from such alternative reading frame induces expression of a novel protein with gain of function and evolutionary fixation (Pavesi, 2021). Overprinting is typical of viruses with small genomes. The epitome of overprinting are the proteins of leviviruses involved in bacterial cell wall lysis during virus egress from infected cells. These small hydrophobic proteins evolved on multiple, independent occasions within different levivirus genes (Chamakura and Young, 2020). Similar cases of overprinting are typical of many other RNA and ssDNA viruses (Pavesi, 2021). In viruses with large dsDNA genomes, overprinting has not been extensively studied. However, many if not most of these viruses encode numerous small proteins with no detectable homologs. Some of these proteins have been shown to counteract host defense systems. For instance, anti-CRISPR proteins that are encoded by many bacterial and archaeal viruses often form large cassettes in the virus genomes (Pawluk et al., 2018). *De novo* origin of at least some of such genes appears likely even if difficult to demonstrate unequivocally.

Generally, given the characteristic arms race-driven rapid evolution of genes involved in defense and counter-defense, it seems likely that much of the dark matter in virus genomes may be involved in counter-defense.

The method in this madness

With all the enormous diversity of viruses and their genes, we believe that the many routes of evolution described here can be accommodated by a simple and logical conceptual

framework. Evolution of viruses seems to have been seeded by a small core set of essential proteins involved in genome replication that appear to be traceable to the pre-cellular stage of evolution. Apart from these apparent primordial proteins, evolution of viruses can be described as a history of gene capture from the hosts as well as other mobile genetic elements for various roles in virus reproduction, often followed by horizontal spread of the exapted genes among viruses. The genes captured from the hosts follow two distinct evolutionary paths: direct recruitment, when proteins perform the same functions in virus reproduction that their ancestors perform in cells, and exaptation, when host genes are repurposed for virus-specific functions (Figure 3). Exaptation can be further subdivided into conservative, whereby the biochemical activity of the exapted protein is retained, even though its biological function is unique to viruses, and radical, when the virus function seems unrelated to cellular one.

Both direct recruitment and exaptation of host genes for virus functions display pervasive convergence whereby unrelated or distantly related genes are independently acquired by widely different viruses. This high prevalence of convergence appears to be determined by the relatively narrow range of functions involved in viral replication and expression. Functional diversity is substantially greater among genes involved in virus-host interactions, but even in these cases, convergent recruitment of inhibitors of key defense systems, often components of these systems themselves (guns for hire), is common (Koonin et al., 2020b). Complemented by *de novo* emergence of virus genes, the few evolutionary trends described here seem to account for the entire observed diversity of viruses, their genomes and genes (Figure 3).

Generally, the larger the virus genome the more auxiliary functions it encodes. Virus genomes grow as a result of acquisition of various auxiliary genes, which increase virus fitness or operational autonomy even though they are largely redundant with functions available from the host cells. The logic behind such apparent functional redundancy in viruses is likely to involve the ability to overcome host-imposed restriction of access to these functions within the framework of antiviral defense. Additionally, such redundant functions can boost the metabolic potential of the host to sustain active virus reproduction and ensure the localization of the respective proteins optimal for virus reproduction via dedicated subcelluar targeting signals.

RNA and DNA viruses have substantially different opportunities for genome expansion. The genome size of RNA viruses appears to be limited to ~40 kb (Saberi et al., 2018), conceivably, due to the intrinsically lower chemical stability of RNA molecules as well as biological and evolutionary factors, such as trade-offs between replication rate and fidelity, and activity of RNA-sensing host defense systems (Ferron et al., 2021). These size limitations dictate the functional repertoire of genes that can be accommodated in RNA virus genomes, so that only the essential functions are afforded. By contrast, the size and functional complexity of dsDNA virus genomes rivals those of cellular genomes. Indeed, viruses with the largest genomes encode nearly complete metabolic pathways, such as glycolysis and the TCA cycle (Moniruzzaman et al., 2020), protein glycosylation machineries (Notaro et al., 2021), key proteins involved in intracellular trafficking including actin (Da Cunha et al., 2022), kinesin (Subramaniam et al., 2020), and myosin (Kijima et

al., 2021), translation system (see above), and much more. In the case of dsDNA viruses, quantification of the fitness in relation to various virus traits suggests that evolution of larger genomes is dictated by benefits of increased infection efficiency, broader host range, potentially increased attachment success and decreased decay rate – traits particularly important in resource-limited, low host population density environments (Edwards et al., 2021).

Overall, the logic of virus evolution is defined by the key biological feature of viruses, namely their obligate intracellular parasitism. This lifestyle affords viruses with ample opportunities for direct appropriation as well as exaptation of a vast repertoire of host activities, but also dictates the necessity of overcoming the vast array of host defense systems. Conversely, viral genes were captured by hosts on many occasions, and some of the protein functionalities that have evolved in the virus genome context are exapted by the hosts, attesting to the extensive two-way transfer of genes and functions between viruses and cells (Filée and Forterre, 2005; Koonin and Krupovic, 2018). Detailed understanding of the logic of virus evolution should help probing virus-host coevolution and development of antivirus therapeutics.

Author contributions

EVK, VVD and MK wrote the manuscript.

Acknowledgements

E.V.K. is supported by the Intramural Research Program of the National Institutes of Health of the USA (National Library of Medicine). M.K. was supported by l'Agence Nationale de la Recherche grant ANR-21-CE11-0001-01. V.V.D. was partially supported by an NIH/NLM/NCBI Visiting Scientist Fellowship.

Declaration of Interests

The authors declare no competing interests.

References

Abrahão, J., Silva, L., Silva, L.S., Khalil, J.Y.B., Rodrigues, R., Arantes, T., Assis, F., Boratto, P., Andrade, M., Kroon, E.G.*, et al.* (2018). Tailed giant Tupanvirus possesses the most complete translational apparatus of the known virosphere. Nat Commun *9*, 749.

Aggarwal, M., Dhindwal, S., Kumar, P., Kuhn, R.J., and Tomar, S. (2014). trans-Protease activity and structural insights into the active form of the alphavirus capsid protease. J Virol *88*, 12242-12253.

Ahola, T., McInerney, G., and Merits, A. (2021). Alphavirus RNA replication in vertebrate cells. Adv Virus Res *111*, 111-156.

Aiewsakun, P., Katzourakis, A. (2016). Time-dependent rate phenomenon in viruses. J Virol 90, 7184-7195.

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G.R., Wang, J., Cong, Q., Kinch, L.N., Schaeffer, R.D.*, et al.* (2021). Accurate prediction of protein structures and interactions using a three-track neural network. Science *373*, 871-876.

Bale, J.B., Gonen, S., Liu, Y., Sheffler, W., Ellis, D., Thomas, C., Cascio, D., Yeates, T.O., Gonen, T., King, N.P.*, et al.* (2016). Accurate design of megadalton-scale two-component icosahedral protein complexes. Science *353*, 389-394.

Baltimore, D. (1971). Expression of animal virus genomes. Bacteriol Rev *35*, 235-241.

Bratke, K.A., McLysaght, A., and Rothenburg, S. (2013). A survey of host range genes in poxvirus genomes. Infect Genet Evol *14*, 406-425.

Carpentier, D.C.J., Van Loggerenberg, A., Dieckmann, N.M.G., and Smith, G.L. (2017). Vaccinia virus egress mediated by virus protein A36 is reliant on the F12 protein. J Gen Virol *98*, 1500-1514.

Chamakura, K.R., and Young, R. (2020). Single-gene lysis in the metagenomic era. Curr Opin Microbiol *56*, 109-117.

Cheng, H., Shen, N., Pei, J., and Grishin, N.V. (2004). Double-stranded DNA bacteriophage prohead protease is homologous to herpesvirus protease. Protein Sci *13*, 2260-2269.

Copley, S.D. (2014). An evolutionary perspective on protein moonlighting. Biochem Soc Trans *42*, 1684-1691.

Da Cunha, V., Gaia, M., Ogata, H., Jaillon, O., Delmont, T.O., and Forterre, P. (2022). Giant Viruses Encode Actin-Related Proteins. Mol Biol Evol *39*, msac022.

Dolja, V.V., Kreuze, J.F., and Valkonen, J.P. (2006). Comparative and functional genomics of closteroviruses. Virus Res *117*, 38-51.

Domingo, E., Garcia-Crespo, C., Lobo-Vega, R., and Perales, C. (2021). Mutation Rates, Mutation Frequencies, and Proofreading-Repair Activities in RNA Virus Genetics. Viruses *13*, 1882.

Duda, R.L., and Teschke, C.M. (2019). The amazing HK97 fold: versatile results of modest differences. Curr Opin Virol *36*, 9-16.

Dunn, B.M., Goodenow, M.M., Gustchina, A., and Wlodawer, A. (2002). Retroviral proteases. Genome Biol *3*, REVIEWS3006.

Edwards, K.F., Steward, G.F., and Schvarcz, C.R. (2021). Making sense of virus size and the tradeoffs shaping viral fitness. Ecol Lett *24*, 363-373.

El Omari, K., Meier, C., Kainov, D., Sutton, G., Grimes, J.M., Poranen, M.M., Bamford, D.H., Tuma, R., Stuart, D.I., and Mancini, E.J. (2013). Tracking in atomic detail the functional specializations in viral RecA helicases that occur during evolution. Nucleic Acids Res *41*, 9396-9410.

Elde, N.C., Child, S.J., Eickbush, M.T., Kitzman, J.O., Rogers, K.S., Shendure, J., Geballe, A.P., and Malik, H.S. (2012). Poxviruses deploy genomic accordions to adapt rapidly against host antiviral defenses. Cell *150*, 831-841.

Elde, N.C., Child, S.J., Geballe, A.P., and Malik, H.S. (2009). Protein kinase R reveals an evolutionary model for defeating viral mimicry. Nature *457*, 485-489.

Faure, G., Shmakov, S.A., Yan, W.X., Cheng, D.R., Scott, D.A., Peters, J.E., Makarova, K.S., and Koonin, E.V. (2019). CRISPR-Cas in mobile genetic elements: counter-defence and beyond. Nat Rev Microbiol *17*, 513-525.

Feiss, M., and Rao, V.B. (2012). The bacteriophage DNA packaging machine. Adv Exp Med Biol *726*, 489-509.

Ferron, F., Sama, B., Decroly, E., and Canard, B. (2021). The enzymes for genome size increase and maintenance of large (+)RNA viruses. Trends Biochem Sci *46*, 866-877.

Filée, J. (2013). Route of NCLDV evolution: the genomic accordion. Curr Opin Virol *3*, 595-599.

Filée, J., and Forterre, P. (2005). Viral proteins functioning in organelles: a cryptic origin? Trends Microbiol *13*, 510-513.

Fokine, A., and Rossmann, M.G. (2016). Common Evolutionary Origin of Procapsid Proteases, Phage Tail Tubes, and Tubes of Bacterial Type VI Secretion Systems. Structure *24*, 1928-1935.

Forterre, P., Prangishvili, D. (2009) The great billion-year war between ribosome- and capsid-encoding organisms (cells and viruses) as the major source of evolutionary novelties. Ann N Y Acad Sci *1178*, 65-77.

Fridman, S., Flores-Uribe, J., Larom, S., Alalouf, O., Liran, O., Yacoby, I., Salama, F., Bailleul, B., Rappaport, F., Ziv, T.*, et al.* (2017). A myovirus encoding both photosystem I and II proteins enhances cyclic electron flow in infected Prochlorococcus cells. Nat Microbiol *2*, 1350-1357.

Gjessing, M.C., Yutin, N., Tengs, T., Senkevich, T., Koonin, E., Ronning, H.P., Alarcon, M., Ylving, S., Lie, K.I., Saure, B.*, et al.* (2015). Salmon Gill Poxvirus, the Deepest Representative of the Chordopoxvirinae. J Virol *89*, 9348-9367.

Gorbalenya, A.E., and Koonin, E.V. (1989). Viral proteins containing the purine NTP-binding sequence pattern. Nucleic Acids Res *17*, 8413-8440.

Gould, S.J., and Vrba, E. (1982). Exaptation—a Missing Term in the Science of Form. Paleobiology *8*, 4-15.

Gould, S.J. (1997). The exaptive excellence of spandrels as a term and prototype. Proc Natl Acad Sci U S A 94, 10750-10755.

Grininger, M., Zeth, K., and Oesterhelt, D. (2006). Dodecins: a family of lumichrome binding proteins. J Mol Biol *357*, 842-857.

Guan, J., and Bondy-Denomy, J. (2020). Intracellular organization by jumbo bacteriophages. J Bacteriol *203*, e00362-00320.

Guo, P., Noji, H., Yengo, C.M., Zhao, Z., and Grainge, I. (2016). Biological Nanomotors with a Revolution, Linear, or Rotation Motion Mechanism. Microbiol Mol Biol Rev *80*, 161-186.

Hinnebusch, A.G., and Lorsch, J.R. (2012). The mechanism of eukaryotic translation initiation: new insights and challenges. Cold Spring Harb Perspect Biol *4*, a011544.

Holm, L. (2020). DALI and the persistence of protein shape. Protein Sci *29*, 128-140.

Hong, C., Oksanen, H.M., Liu, X., Jakana, J., Bamford, D.H., and Chiu, W. (2014). A structural model of the genome packaging process in a membrane-containing double stranded DNA virus. PLoS Biol *12*, e1002024.

Iyer, L.M., Makarova, K.S., Koonin, E.V., and Aravind, L. (2004). Comparative genomics of the FtsK-HerA superfamily of pumping ATPases: implications for the origins of chromosome segregation, cell division and viral capsid packaging. Nucleic Acids Res *32*, 5260-5279.

Jin, L., Chen, M., Xiang, M., and Guo, Z. (2022). RNAi-Based Antiviral Innate Immunity in Plants. Viruses *14*, 432.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Zidek, A., Potapenko, A.*, et al.* (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583-589.

Kazlauskas, D., Krupovic, M., and Venclovas, C. (2016). The logic of DNA replication in double-stranded DNA viruses: insights from global analysis of viral genomes. Nucleic Acids Res *44*, 4551-4564.

Kazlauskas, D., and Venclovas, C. (2014). Herpesviral helicase-primase subunit UL8 is inactivated B-family polymerase. Bioinformatics *30*, 2093-2097.

Kijima, S., Delmont, T.O., Miyazaki, U., Gaia, M., Endo, H., and Ogata, H. (2021). Discovery of Viral Myosin Genes With Complex Evolutionary History Within Plankton. Front Microbiol *12*, 683294.

Kilcher, S., Schmidt, F.I., Schneider, C., Kopf, M., Helenius, A., and Mercer, J. (2014). siRNA screen of early poxvirus genes identifies the AAA+ ATPase D5 as the virus genome-uncoating factor. Cell Host Microbe *15*, 103-112.

King, J.A., Dubielzig, R., Grimm, D., and Kleinschmidt, J.A. (2001). DNA helicase-mediated packaging of adeno-associated virus type 2 genomes into preformed capsids. EMBO J *20*, 3282-3291.

Klose, T., Herbst, D.A., Zhu, H., Max, J.P., Kenttamaa, H.I., and Rossmann, M.G. (2015). A Mimivirus Enzyme that Participates in Viral Entry. Structure *23*, 1058-1065.

Koonin, E.V., and Dolja, V.V. (2014). Virus world as an evolutionary network of viruses and capsidless selfish elements. Microbiol Mol Biol Rev *78*, 278-303.

Koonin, E.V., Dolja, V.V., and Krupovic, M. (2015). Origins and evolution of viruses of eukaryotes: The ultimate modularity. Virology *479-480*, 2-25.

Koonin, E.V., Dolja, V.V., Krupovic, M., and Kuhn, J.H. (2021a). Viruses Defined by the Position of the Virosphere within the Replicator Space. Microbiol Mol Biol Rev *85*, e0019320.

Koonin, E.V., Dolja, V.V., Krupovic, M., Varsani, A., Wolf, Y.I., Yutin, N., Zerbini, F.M., and Kuhn, J.H. (2020a). Global organization and proposed megataxonomy of the virus world. Microbiol Mol Biol Rev *84*, e00061-00019.

Koonin, E.V., and Krupovic, M. (2018). The depths of virus exaptation. Curr Opin Virol *31*, 1-8.

Koonin, E.V., Krupovic, M., and Agol, V.I. (2021b). The Baltimore Classification of Viruses 50 Years Later: How Does It Stand in the Light of Virus Evolution? Microbiol Mol Biol Rev *85*, e0005321.

Koonin, E.V., Makarova, K.S., Wolf, Y.I., and Krupovic, M. (2020b). Evolutionary entanglement of mobile genetic elements and host defence systems: guns for hire. Nat Rev Genet *21*, 119-131.

Koonin, E.V., Wolf, Y.I., Nagasaki, K., and Dolja, V.V. (2008). The Big Bang of picorna-like virus evolution antedates the radiation of eukaryotic supergroups. Nat Rev Microbiol *6*, 925-939.

Koonin, E.V., and Yutin, N. (2019). Evolution of the Large Nucleocytoplasmic DNA Viruses of Eukaryotes and Convergent Origins of Viral Gigantism. Adv Virus Res *103*, 167-202.

Krupovic, M., Dolja, V.V., and Koonin, E.V. (2019). Origin of viruses: primordial replicators recruiting capsids from hosts. Nat Rev Microbiol *17*, 449-458.

Krupovic, M., Dolja, V.V., and Koonin, E.V. (2020a). The LUCA and its complex virome. Nat Rev Microbiol *18*, 661-670.

Krupovic, M., Gribaldo, S., Bamford, D.H., and Forterre, P. (2010). The evolutionary history of archaeal MCM helicases: a case study of vertical evolution combined with hitchhiking of mobile genetic elements. Mol Biol Evol *27*, 2716-2732.

Krupovic, M., and Koonin, E.V. (2017). Multiple origins of viral capsid proteins from cellular ancestors. Proc Natl Acad Sci U S A *114*, E2401-E2410.

Krupovic, M., Makarova, K.S., and Koonin, E.V. (2022). Cellular homologs of the double jelly-roll major capsid proteins clarify the origins of an ancient virus kingdom. Proc Natl Acad Sci U S A *119*, e2120620119.

Krupovic, M., Yutin, N., and Koonin, E. (2020b). Evolution of a major virion protein of the giant pandoraviruses from an inactivated bacterial glycoside hydrolase. Virus Evol *6*, veaa059.

Krylov, D.M., and Koonin, E.V. (2001). A novel family of predicted retroviral-like aspartyl proteases with a possible key role in eukaryotic cell cycle control. Curr Biol *11*, R584-587.

Kvansakul, M., Caria, S., and Hinds, M.G. (2017). The Bcl-2 Family in Host-Virus Interactions. Viruses *9*, 290.

Laughlin, T.G., Deep, A., Pricharda, A.M., Seitz, C., Gub, Y., Enustuna, E., Suslov, S., Khanna, K., Birkholz, E.A., Armbruster, E.*, et al.* (2022). Architecture and self-assembly of the jumbo bacteriophage nuclear shell. https://wwwbiorxivorg/content/101101/20220214480162v1.

Legendre, M., Fabre, E., Poirot, O., Jeudy, S., Lartigue, A., Alempic, J.M., Beucher, L., Philippe, N., Bertaux, L., Christo-Foroux, E.*, et al.* (2018). Diversity and evolution of the emerging Pandoraviridae family. Nat Commun *9*, 2285.

Legendre, M., Lartigue, A., Bertaux, L., Jeudy, S., Bartoli, J., Lescot, M., Alempic, J.M., Ramus, C., Bruley, C., Labadie, K.*, et al.* (2015). In-depth study of Mollivirus sibericum, a new 30,000-y-old giant virus infecting Acanthamoeba. Proc Natl Acad Sci U S A *112*, E5327-5335.

Liu, J., and Mushegian, A. (2004). Displacements of prohead protease genes in the late operons of double-stranded-DNA bacteriophages. J Bacteriol *186*, 4369-4375.

Liu, Y., Demina, T.A., Roux, S., Aiewsakun, P., Kazlauskas, D., Simmonds, P., Prangishvili, D., Oksanen, H.M., and Krupovic, M. (2021). Diversity, taxonomy, and evolution of archaeal viruses of the class *Caudoviricetes*. PLoS Biol *19*, e3001442.

Liu, Y.P., Peremyslov, V.V., Medina, V., and Dolja, V.V. (2009). Tandem leader proteases of Grapevine leafroll-associated virus-2: host-specific functions in the infection cycle. Virology *383*, 291-299.

Lulla, V., Kim, D.Y., Frolova, E.I., and Frolov, I. (2013). The amino-terminal domain of alphavirus capsid protein is dispensable for viral particle assembly but regulates RNA encapsidation through cooperative functions of its subdomains. J Virol *87*, 12003-12019.

Lynch, M., and Katju, V. (2004). The altered evolutionary trajectories of gene duplicates. Trends Genet *20*, 544-549.

Lynch, M., O'Hely, M., Walsh, B., and Force, A. (2001). The probability of preservation of a newly arisen gene duplicate. Genetics *159*, 1789-1804.

Marabini, R., Condezo, G.N., Krupovic, M., Menendez-Conejero, R., Gomez-Blanco, J., and San Martin, C. (2021). Near-atomic structure of an atadenovirus reveals a conserved capsid-binding motif and intergenera variations in cementing proteins. Sci Adv *7*, eabe6008.

Marq, J.B., Hausmann, S., Luban, J., Kolakofsky, D., and Garcin, D. (2009). The double-stranded RNA binding domain of the vaccinia virus E3L protein inhibits both RNA- and DNA-induced activation of interferon beta. J Biol Chem *284*, 25471-25478.

Medvedeva, S., Liu, Y., Koonin, E.V., Severinov, K., Prangishvili, D., and Krupovic, M. (2019). Virus-borne mini-CRISPR arrays are involved in interviral conflicts. Nat Commun *10*, 5204.

Mihara, T., Koyano, H., Hingamp, P., Grimsley, N., Goto, S., and Ogata, H. (2018). Taxon Richness of "Megaviridae" Exceeds those of Bacteria and Archaea in the Ocean. Microbes Environ *33*, 162-171.

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2022). ColabFold: making protein folding accessible to all. Nat Methods *19*, 679-682.

Mizuno, C.M., Guyomar, C., Roux, S., Lavigne, R., Rodriguez-Valera, F., Sullivan, M.B., Gillet, R., Forterre, P., and Krupovic, M. (2019). Numerous cultivated and uncultivated viruses encode ribosomal proteins. Nat Commun *10*, 752.

Moniruzzaman, M., Martinez-Gutierrez, C.A., Weinheimer, A.R., and Aylward, F.O. (2020). Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. Nat Commun *11*, 1710.

Moyer, C.L., Besser, E.S., and Nemerow, G.R. (2016). A Single Maturation Cleavage Site in Adenovirus Impacts Cell Entry and Capsid Assembly. J Virol *90*, 521-532.

Napuli, A.J., Alzhanova, D.V., Doneanu, C.E., Barofsky, D.F., Koonin, E.V., and Dolja, V.V. (2003). The 64-kilodalton capsid protein homolog of Beet yellows virus is required for assembly of virion tails. J Virol *77*, 2377-2384.

Notaro, A., Coute, Y., Belmudes, L., Laugeri, M.E., Salis, A., Damonte, G., Molinaro, A., Tonetti, M.G., Abergel, C., and De Castro, C. (2021). Expanding the Occurrence of Polysaccharides to the Viral World: The Case of Mimivirus. Angew Chem Int Ed Engl *60*, 19897-19904.

Pavesi, A. (2021). Origin, Evolution and Stability of Overlapping Genes in Viruses: A Systematic Review. Genes (Basel) *12*, 809.

Pawluk, A., Davidson, A.R., and Maxwell, K.L. (2018). Anti-CRISPR: discovery, mechanism and function. Nat Rev Microbiol *16*, 12-17.

Piatigorsky, J., and Wistow, G.J. (1989). Enzyme/crystallins: gene sharing as an evolutionary strategy. Cell *57*, 197-199.

Requiao, R.D., Carneiro, R.L., Moreira, M.H., Ribeiro-Alves, M., Rossetto, S., Palhano, F.L., and Domitrovic, T. (2020). Viruses with different genome types adopt a similar strategy to pack nucleic acids based on positively charged protein domains. Sci Rep *10*, 5470.

Roux, S., Krupovic, M., Daly, R.A., Borges, A.L., Nayfach, S., Schulz, F., Sharrar, A., Matheus Carnevali, P.B., Cheng, J.F., Ivanova, N.N.*, et al.* (2019). Cryptic inoviruses revealed as pervasive in bacteria and archaea across Earth's biomes. Nat Microbiol *4*, 1895-1906.

Saberi, A., Gulyaeva, A.A., Brubacher, J.L., Newmark, P.A., and Gorbalenya, A.E. (2018). A planarian nidovirus expands the limits of RNA genome size. PLoS Pathog *14*, e1007314.

Senkevich, T.G., Koonin, E.V., and Moss, B. (2011). Vaccinia virus F16 protein, a predicted catalytically inactive member of the prokaryotic serine recombinase superfamily, is targeted to nucleoli. Virology *417*, 334-342.

Senkevich, T.G., Yutin, N., Wolf, Y.I., Koonin, E.V., and Moss, B. (2021). Ancient Gene Capture and Recent Gene Loss Shape the Evolution of Orthopoxvirus-Host Interaction Genes. mBio *12*, e0149521.

Sirkis, R., Gerst, J.E., and Fass, D. (2006). Ddi1, a eukaryotic protein with the retroviral protease fold. J Mol Biol *364*, 376-387.

Subramaniam, K., Behringer, D.C., Bojko, J., Yutin, N., Clark, A.S., Bateman, K.S., van Aerle, R., Bass, D., Kerr, R.C., Koonin, E.V.*, et al.* (2020). A New Family of DNA Viruses Causing Disease in Crustaceans from Diverse Aquatic Biomes. mBio *11*, e02938-02919.

Sullivan, M.B., Lindell, D., Lee, J.A., Thompson, L.R., Bielawski, J.P., and Chisholm, S.W. (2006). Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. PLoS Biol *4*, e234.

Tarasova, E., Dhindwal, S., Popp, M., Hussain, S., and Khayat, R. (2021). Mechanism of DNA Interaction and Translocation by the Replicase of a Circular Rep-Encoding Single-Stranded DNA Virus. mBio *12*, e0076321.

Tetter, S., Terasaka, N., Steinauer, A., Bingham, R.J., Clark, S., Scott, A.J.P., Patel, N., Leibundgut, M., Wroblewski, E., Ban, N.*, et al.* (2021). Evolution of a virus-like architecture and packaging mechanism in a repurposed bacterial protein. Science *372*, 1220-1224.

Tian, B., and Mathews, M.B. (2003). Phylogenetics and functions of the double-stranded RNA-binding motif: a genomic survey. Prog Nucleic Acid Res Mol Biol *74*, 123-158.

Villalta, A., Schmitt, A., Estrozi, L.F., Quemin, E.R.J., Alempic, J.-M., Lartigue, A., Pražák, V., Belmudes, L., Vasishtan, D., Colmant, A.M.S.*, et al.* (2022). The giant Mimivirus 1.2 Mb genome is elegantly organized into a 30 nm helical protein shield. https://wwwbiorxivorg/content/101101/20220217480895v1.

Wahaab, A., Mustafa, B.E., Hameed, M., Stevenson, N.J., Anwar, M.N., Liu, K., Wei, J., Qiu, Y., and Ma, Z. (2021). Potential Role of Flavivirus NS2B-NS3 Proteases in Viral Pathogenesis and Anti-flavivirus Drug Discovery Employing Animal Cells and Models: A Review. Viruses *14*, 44.

Willemsen, A., Zwart, M.P., Higueras, P., Sardanyes, J., and Elena, S.F. (2016). Predicting the Stability of Homologous Gene Duplications in a Plant RNA Virus. Genome Biol Evol *8*, 3065-3082.

Yin, Y., and Fischer, D. (2008). Identification and investigation of ORFans in the viral world. BMC Genomics *9*, 24.

Yutin, N., Benler, S., Shmakov, S.A., Wolf, Y.I., Tolstoy, I., Rayko, M., Antipov, D., Pevzner, P.A., and KooninE.V. (2021). Analysis of metagenome-assembled viral genomes from the human gut reveals diverse putative CrAss-like phages with unique genomic features Nature Communications *12*, 1044.

Yutin, N., Faure, G., Koonin, E.V., and Mushegian, A.R. (2014). Chordopoxvirus protein F12 implicated in enveloped virion morphogenesis is an inactivated DNA polymerase. Biol Direct *9*, 22.

Yutin, N., and Koonin, E.V. (2009). Evolution of DNA ligases of nucleo-cytoplasmic large DNA viruses of eukaryotes: a case of hidden complexity. Biol Direct *4*, 51.

Zhukovsky, E.A., Lee, J.O., Villegas, M., Chan, C., Chu, S., and Mroske, C. (2004). TNF ligands: is TALL-1 a trimer or a virus-like cluster? Nature *427*, 413-414; discussion 414.

Zuhlsdorf, M., and Hinrichs, W. (2017). Assemblins as maturational proteases in herpesviruses. J Gen Virol *98*, 1969-1984.

Figures

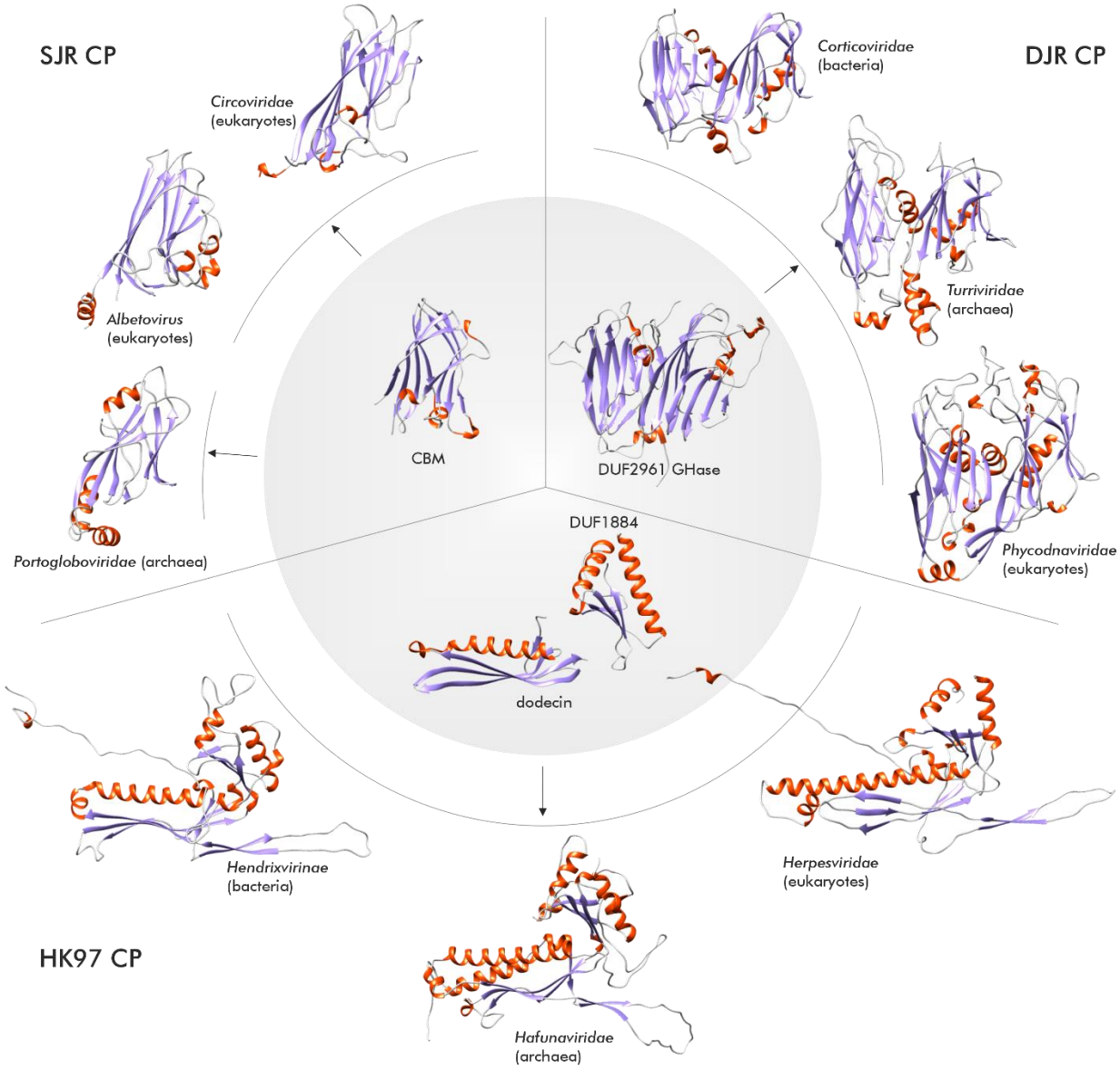

Figure 1. **Origin of the three major types of viral capsid proteins from cellular ancestors**. The three sections of the figure depict the three structural folds, namely, single jelly-roll (SJR), double jelly-roll (DJR) and HK97 fold, which each form icosahedral virus capsids. The likely cellular ancestors are depicted within a grey area, whereas the corresponding capsid proteins are shown on the periphery. For each fold, representatives from viruses infecting hosts from different domains of life are shown, when available. Only the floor domain of the herpesvirus capsid protein is shown. Archaeal HK97 capsid protein (*Hafunaviridae*) is represented by an AlphaFold2 structural model (Liu et al., 2021). The structures are colored according to the secondary structure: α-helices, red; β-strands, blue; random coils, grey. CBM, carbohydrate-binding module; DUF, domain of unknown function.

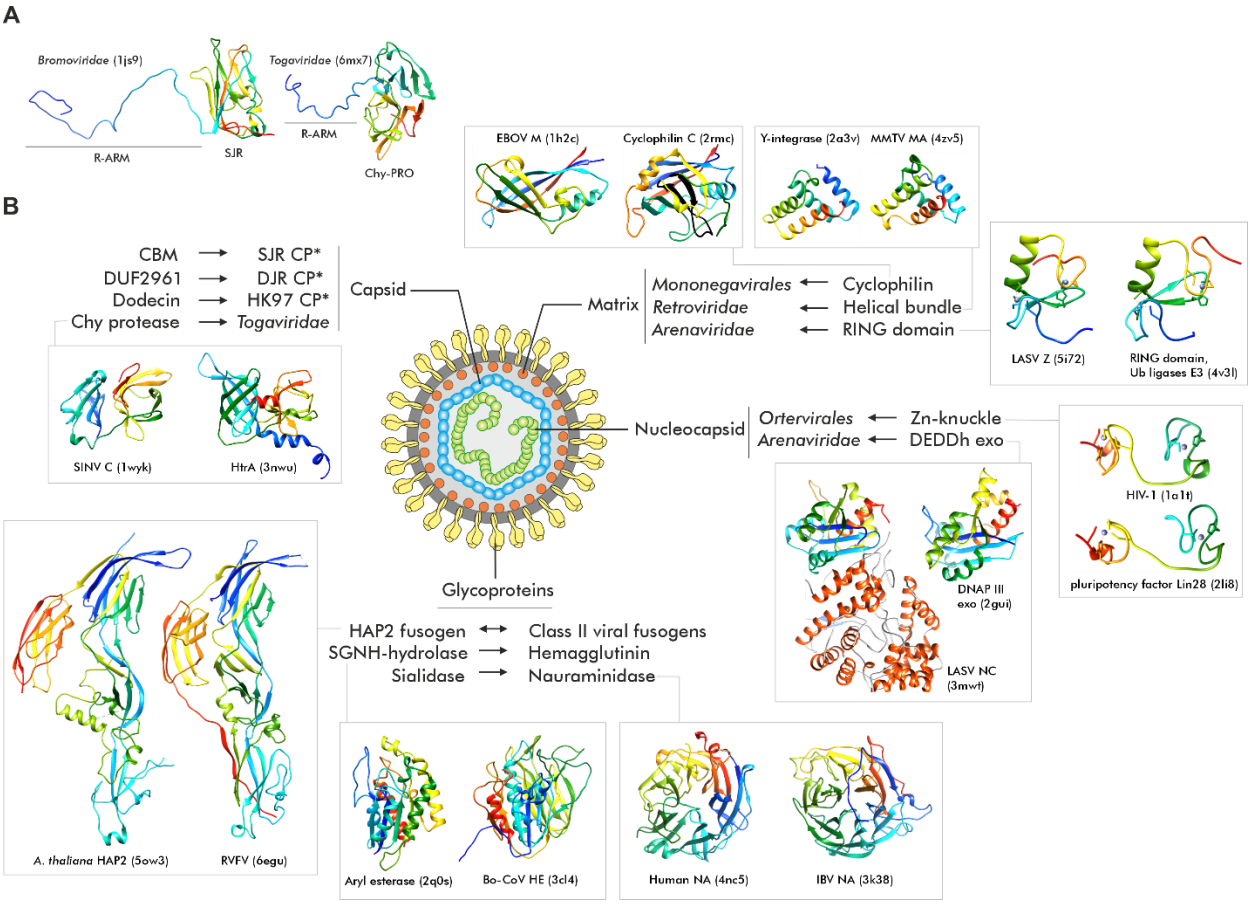

Figure 2. **Examples of structure conservation and variation in exapted virion proteins**. (A) Convergent acquisition of positively charged, unstructured regions (R-arms) by unrelated capsid proteins (CPs) of RNA viruses. Left: single jelly-roll (SJR) CP; right, chymotrypsin protease (Chy-PRO)-like CP. (B) Diversity of exapted virion proteins. The schematic of a virion shows the locations of the major virion proteins; not all of these proteins are necessarily present in virions of different viruses. In each box, the cellular ancestor and the corresponding exapted virion protein are shown side by side. Asterisks denote CPs shown in Figure 1. The direction of exaptation is indicated by arrows. In the case of class II fusogens, the directionality of evolution is unclear. The corresponding PDB accession numbers are indicated under each structure. All structures are colored using the rainbow scheme: from blue N-terminus to red C-terminus. For the Lassa virus (LASV) nucleocapsid (NC) protein, only the exonuclease domain is shown using rainbow the scheme for more convenient comparison, whereas the rest of the protein is shown in red. Abbreviations: EBOV, Ebola virus; MMTV, mouse mammary tumor virus; HIV-1, human immunodeficiency virus 1; IBV, infectious bronchitis virus; Bo-CoV, bovine coronavirus; RVFV, Rift Valley fever virus; SINV, Sindbis virus; NA, neuraminidase; HE, hemagglutinin esterase; M/MA/Z, matrix protein.

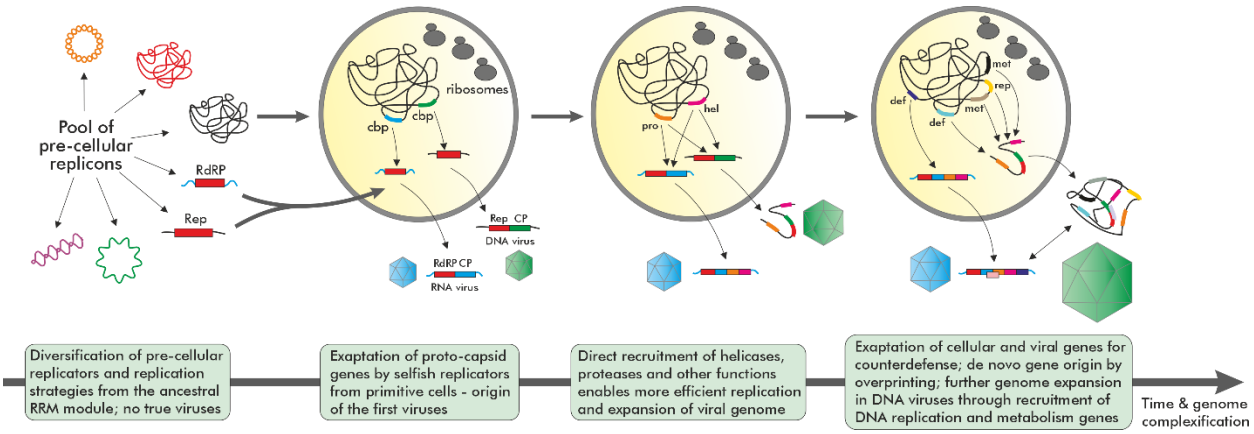

Figure 3. **Routes of virus gene evolution: primordial core, exaptation and direct recruitment of host proteins**. Gradual acquisition of genes from different sources coincides with the genome and capsid size expansion, especially in DNA viruses. Evolution of viruses starts with the recruitment of replication proteins from the pool of pre-cellular replicators, followed by exaptation of proteins from proto-cells for capsid formation, likely, at the pre-LUCA stage of evolution. Virus genome complexification involves recruitment of various cellular helicases and proteases. In parallel and subsequent to that, throughout the evolution of viruses, additional genes are acquired, including those responsible for counter-defense, and those further promoting viral genome replication and metabolism. Abbreviations: RdRP, RNA-directed RNA polymerase; cbp, carbohydrate binding protein; CP, capsid protein; pro, proteases; hel, helicases; def, defense proteins; met, metabolism proteins; rep, genome replication functions.