

Origin of viruses: primordial replicators recruiting capsids from hosts

M Krupovic, Valerian Dolja, Eugene Koonin

▶ **To cite this version:**

M Krupovic, Valerian Dolja, Eugene Koonin. Origin of viruses: primordial replicators recruiting capsids from hosts. Nature Reviews Microbiology, 2019, 17 (7), pp.449-458. 10.1038/s41579-019-0205-6 . pasteur-02557191

HAL Id: pasteur-02557191

https://pasteur.hal.science/pasteur-02557191

Submitted on 30 Apr 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Origin of viruses: primordial replicators recruiting capsids from hosts

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

1 Unité Biologie Moléculaire du Gène chez les Extrêmophiles, Institut Pasteur, Paris, France
2 Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR, USA
3 National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD, USA

Correspondence to M.K. krupovic@pasteur.fr and E.V.K. koonin@ncbi.nlm.nih.gov

ABSTRACT

Viruses are ubiquitous parasites of cellular life and the most abundant biological entities on earth. It is widely accepted that viruses are polyphyletic but a consensus scenario for their ultimate origin(s) is still lacking. Traditionally, three different scenarios have been considered for the origin of viruses: descent from primordial, pre-cellular genetic elements, reductive evolution from cellular ancestors, and escape of genes from cellular hosts achieving partial replicative autonomy and becoming parasitic genetic elements. These three classical scenarios give different timelines for the origin of viruses and do not explain the provenance of the two key functional modules that are responsible, respectively, for virus genome replication and virion morphogenesis. Here we outline a 'chimeric' scenario under which different types of primordial, selfish replicons gave rise to viruses by recruiting host proteins for virion formation. We also propose that new groups of viruses have repeatedly emerged at all stages of the evolution of life, often, through the displacement of ancestral structural and genome replication genes.

[H1] Introduction

Viruses are the most abundant biological entities on our planet and have major roles in global ecology and evolution of the biosphere [1-4]. All cellular organisms, with the possible exception of some intracellular parasitic bacteria, harbour distinctive repertoires of viruses, that is, obligate intracellular genetic parasites that package their genomes into virus particles (virions)[5]. The ubiquity of viruses, combined with the theoretical argument that genetic parasites will inevitably emerge in replicator systems [6,7], implies that the entire course of the evolution of life is actually a story of virus-host coevolution [4,8-10]. Accordingly, evolution of life cannot be understood without elucidating the origin(s) of viruses, yet, these origins currently remain mysterious.

Historically, three distinct scenarios of virus origin have been considered [11-17] (Figure 1). According to the 'primordial virus world' or 'virus early' hypothesis, viruses are direct descendants of the first replicons that existed during the pre-cellular stage of the evolution of life. By contrast, in the 'reductive virus origin' or 'regression' scenario, viruses are the ultimate products of degeneration of ancestral cells that have lost their autonomy and transitioned to obligate intracellular parasitism. Finally, in the 'escaped genes' scenario, viruses evolved on multiple, independent occasions in different cellular organisms from host genes that acquired the capacity for (quasi)autonomous, selfish replication and infectivity. Accordingly, escaped bacterial, archaeal and eukaryotic genes are thought to have given rise to bacterial,

archaeal and eukaryotic viruses, respectively. The three scenarios seem to be mutually exclusive with respect to the origin of any particular group of viruses but different groups of viruses potentially could have evolved via different routes. Over the years, all three scenarios have been revised, elaborated, temporarily abandoned, and revisited.

Viruses use, effectively, all possible strategies of genome replication and expression, with different nucleic acid forms (single-stranded or double-stranded RNA or DNA) as the genome that is incorporated into virions. This diversity of genomic strategies in viruses contrasts with the uniformity observed in cellular organisms and seems to be most compatible with the possibility that the virus world descended directly from a pre-cellular stage of evolution [16,18]. By contrast, the discovery of protist-infecting giant viruses rival bacteria and archaea in terms of genome and particle size, and encode a variety of translation system components[19,20] inspired a revival of the regression hypothesis [21-24]. Finally, an updated version of the escape hypothesis posits that the first viruses 'escaped' not from modern but from primordial cells, predating the last universal cellular ancestor [25].

Formulating falsifiable hypotheses that would allow one to discriminate among the three scenarios of virus origin is no easy task. However, if clues are to be found, these likely will emerge from homologous relationships between virus and cellular host genes. Many virus genes, especially, in large viruses with double-stranded (ds) DNA genomes, have readily identifiable cellular homologs. However, most of these genes encode proteins that are involved in various forms of virus-host interactions, are conserved only among closely related viruses and seem to be relatively late acquisitions [20,26-29]. Thus, these genes reflect important aspects of virus evolution but are of little direct relevance for virus origins.

To explore the origins of viruses, it is necessary to investigate the provenance of the core virus genes that are responsible for the key virus-specific functions. A typical virus genome encompasses two core modules that consist, respectively, of genes encoding proteins required for genome replication and proteins involved in virion formation (hereafter we will call these replication module and morphogenetic module, respectively). In small viruses, these core modules include all or most of their genes whereas in large viruses, the core genes only represent a small fraction of their gene repertoire [16,30]. Understanding the origin of any virus group requires elucidation of the evolutionary roots of both core modules [31,32]. In this Opinion article, we review the evidence on the origins of the replication and morphogenetic modules of the most common and diverse groups of viruses, and we develop arguments for the emergence of the two core modules. We propose that the virus replication machinery arose from the primordial pool of genetic elements whereas the structural proteins seem to have been acquired from hosts at different stages of evolution, leading to a chimeric origin of viruses.

[H1] Origin of virus replication modules

Viral replication proteins often have no closely related homologs in extant cellular organisms with sequenced genomes [16,33]. Thus, it has been suggested that many of these proteins evolved in the pre-cellular world [4,16] or in primordial, now extinct, cellular lineages [11,25,34]. The list of the hallmark virus replication genes is short but spans multiple, diverse groups of viruses and related mobile genetic elements (MGEs) (Box 1). The prototypical virus replication hallmarks are the homologous RNA-dependent RNA polymerases (RdRp) and reverse transcriptases (RT) that, respectively, mediate the replication of all classes of RNA viruses and reverse-transcribing viruses as well as retroelements. Other hallmarks include a distinct superfamily of helicases, often fused to primases, protein-primed DNA polymerases, and the endonuclease involved in the initiation of rolling circle DNA replication (RCRE).

These hallmark proteins have only distant homologs in cellular organisms but are prevalent in capsidless MGEs (Box 1).

The RdRps and the RTs are particularly notable as potential relics of the primordial replicator pool. These enzymes catalyze RNA replication via RNA or DNA intermediates and these reaction contribute to specialized functions in cellular organisms (such as telomere biogenesis in eukaryotes [35]) but not the cellular genome replication [36,37]. The RdRps and RTs share the structural fold of the main catalytic domain with the DNA-dependent DNA polymerases that are involved in the replication of the genomes of archaea, eukaryotes and many dsDNA viruses as well as the RCRE and archaeo-eukaryotic primases (AEPs) [38,39], which are also commonly found in DNA viruses and plasmids from all three domains of cellular life [33,39,40]. This so-called 'palm' fold is related to the RNA-recognition motif (RRM) domain, one of the most common RNA-binding domains that is involved in various RNA biogenesis processes in all forms of cellular life [41].

We hypothesize that RRM was one of the earliest protein domains to evolve and was central to the origin and early evolution of both RNA and DNA replication (Figure 2). The origin of the RRM is likely to be rooted in the primordial RNA world [42] where it would serve as a cofactor to ribozymes including hypothetical ribozyme replicases. Subsequently, the RRM evolved a range of enzymatic functions, most notably, those of RdRp and RT. Although it is not clear which virus polymerase is older (ancestral), it has been suggested that virus RdRps could have evolved from a common ancestor shared with the RTs of group II introns [43,44], a highly diverse class of mobile retroelements that are widespread in bacterial genomes [45]. Indeed, RTs of group II introns are believed to be ancestral to the RTs of eukaryotic retroelements and reverse-transcribing viruses, and among virus RdRps, display closest sequence similarity to those of bacterial positive-sense RNA viruses (family *Leviviridae*)[43]. As a rule, RNA replication and reverse transcription are absent in cells revealing fundamental differences between cellular and viral life styles. The only known exceptions to this rule are telomerases and eukaryotic RdRps involved in the formation of telomeres and small RNAs (as a part of RNA interference and the antiviral response), respectively. Notably, cellular RdRps are not homologous to the viral enzymes and have the so-called double-psi barrel fold that is unrelated to the RRM and is instead homologous to the core catalytic domains of the DNA-dependent RNA polymerases that are involved in transcription in all domains of cellular life [46,47].

Many bacterial and archaeal viruses do not encode identifiable replicative enzymes of their own, apparently, having lost the primordial, RRM-based ones. Instead, these viruses rely on the host replication machinery, which is often, but not always, recruited to the virus genomes by various virus-encoded proteins [48]. In many cases, horizontally acquired cellular genes encode key components of the replisome, in particular, helicases and primases [33], that are responsible for the replisome assembly at the origin of replication. However, apart from a small minority of bacterial dsDNA viruses that possess replicative enzymes homologous to bacterial DNA polymerase III (family C)[33], there is no alternative to the RRM domain as the main replicative moiety of viruses and MGEs, indicating that the RRM is indeed the major legacy of the primordial pool of genetic elements.

The potential competing scenario that postulates escape(s) of the RRM domain(s) from modern-type cells at later stages of evolution seems far less likely. The primary argument against this scenario is that three key replicative enzymes found in broad classes of viruses and MGEs, namely, RdRp, RT and protein-primed DNA polymerase, have no counterparts in modern cells other than those that clearly were acquired from selfish elements such as, for example, the eukaryotic telomerase that originates from the RT of prokaryotic group II introns [49,50]. The RRM domains, especially, the enzymatically active versions,

possess readily recognizable sequence signatures, and therefore, origin of the virus replicative enzymes from undetected cellular ancestors is highly unlikely.

At present, we only have vague ideas on the pre-cellular stages of evolution. However, at least three aspects can be assumed logically. First, primordial genetic elements had small genomes; conceivably, soon after the advent of translation, most replicons resembled modern small viruses, encoding only one to a few proteins [16,18]. Second, the pool of primordial replicators necessarily included genetic parasites. Not only are such parasites ubiquitous in the modern biosphere, but a strong theoretical argument suggests that their emergence is an intrinsic feature of evolving replicators. In mathematical models of replicator evolution, stable equilibria necessarily include parasitic genetic elements [6,7]. To put the argument in a deliberately simplistic form, as soon as there is a resource that potentially can be exploited without investing in it (such as the translation system), parasites (cheaters) that hijack that resource instead of replenishing it will necessarily evolve. However, it is important to emphasize that viruses do not typically turn into ultimate parasites but rather retain some components of the replication machinery that control genome replication and maintain the partial autonomy of virus reproduction. Third, within the RNA world and during the transition to the DNA-RNA-protein world, these elements would use different forms of nucleic acids including single-stranded and double-stranded RNA as well as emerging ssDNA and dsDNA genomes. Effectively, all possible forms of genomes would be tried out at this stage, in part, because diversification of the genomic strategy is one of the means to limit the deleterious effects of genetic parasites [51]. Admittedly, the RNA world is a hypothesis but, to our knowledge, it is the only one that resolves, at least conceptually, the chicken and egg problem that seems to doom any other scenario of the origin of nucleic acid-protein reproducers and is, furthermore, compatible with the growing body of experimental results on versatile catalytic activities of ribozymes [52,53]. Further, it is useful to note that the scenario of early evolution outlined here is not contingent on any specific form of compartmentalization that would confine the primordial replicator pool. It is negligible for our line of argument whether this pool evolved in inorganic compartments or lipid membrane-bounded vesicles, which might be called protocells [54-58].

Thus, one of the few claims we can confidently make about the primordial genetic pool is that it was replete with diverse, competing, 'virus-like' genetic elements, some of which exploited resources produced by others. But, were there viruses, that is, genetic parasites that encode not only components of the replication machinery but also proteins required for virion formation? Indeed, the ability to form a virion is often considered a signature of the viral state that distinguishes viruses from other MGEs [59,60]. The role of the virion in a virus reproduction cycle includes both protection of the genome while outside a host cell and delivery of the genome into the cell. The entry function typically involves interaction between the virus particle and a receptor on the cell surface. The relevance of this function at the stage of the primordial genetic pool is uncertain because no valid data on the nature of the compartments that hosted that pool exists. However, the protective role of virus particles could be potentially important at this early stage of evolution. The virion could both stabilize the viral genome in adverse physicochemical conditions and protect it from targeted degradation by competing replicons. Given this apparently plausible hypothesis for the origin of viruses in the pre-cellular era, the problem becomes, primarily, empirical: can we trace the origin of virion components, and in particular, capsid proteins?

[H1] Diversity of virus capsids

About 60% of the recognized virus taxa have icosahedral capsids, which is unsurprising because the icosahedron has the largest volume to surface area ratio, closest to that of a sphere, the most

thermodynamically favorable three-dimensional shape, and generates the maximum enclosed volume for shells comprised of a given size subunit [61,62]. 50-90% of virus particles detected in the global marine virosphere have an average diameter of 50 nm and belong to the icosahedral tailless morphotype [63]. The vast majority of the remaining 10-50% are bacterial and archaeal viruses with icosahedral capsids and helical tails (order *Caudovirales*). Apart from icosahedra, numerous viruses, especially, those infecting plants and archaea, have elongated, rod-shaped or filamentous capsids [64-68] whereas others, such as numerous negative-strand RNA viruses of animals, have helical nucleocapsids [69-71]. Finally, there is a long 'tail' of odd-shaped virions such as those found in numerous viruses of hyperthermophilic archaea [72]. Thus, the diversity of widespread shapes among viral capsids is quite limited. But what about the major structural proteins that form these viral particles? Analysis of the evolutionary relationships between capsid proteins (CPs) and especially, their origins is a non-trivial task due to the rapid sequence divergence among these proteins. Sequence conservation among the CPs drops off rapidly even at short evolutionary distances because, in this case, selection maintains primarily the structural fold rather than the sequence [73]. Accordingly, structural comparisons of the CPs are far more informative than even the most sensitive sequence comparisons [74]. Such structural comparisons have uncovered far reaching structural unity between the CPs of diverse viruses that infect hosts from different domains and kingdoms of cellular life [73,75,76]. The number of structural folds, at least, widespread ones, represented in viral CPs is fairly small. For example, for viruses with dsDNA genomes from 20 families that account for a large part of the virosphere only 5 distinct structural folds of the CPs have been discovered [75]. Furthermore, small viruses with ssRNA and ssDNA genomes share homologous CPs [77]. Such findings demonstrate the antiquity of the CPs and evolutionary connections between the respective viruses, which can reflect both ancient common ancestry and horizontal gene exchange [73,75,78].

The other side of the coin, however, is that similar capsid geometries do not necessarily reflect homologous relationships between viruses [79]: for example, icosahedral capsids emerged at least 11 times during virus evolution from unrelated CPs with drastically different folds, from all-α to all-β [76]. Viral CPs forming helical (nucleo)capsids are equally diverse [76]. This observation is crucial for understanding the evolution of CPs because it refutes the hypothesis that the same fold, such as the jelly-roll fold, which is most common among icosahedral viruses [76,78], could evolve by convergence, driven by geometrical constraints on the capsid [79]. Furthermore, not all folds that are known to produce icosahedral shells are used by viruses for capsid formation. For example, bacterial microcompartments and certain enzymes, such as luminazine synthase, assemble into icosahedral cages but their shell proteins are structurally unrelated to known viral proteins [80-83]. Thus, the extant number of capsid protein folds in the virosphere likely represents the number of distinct occasions on which the capsids have evolved rather than the limits of the folding space.

[H1] Cellular ancestry of capsids

Most of the viral CPs do not have easily detectable homologs among cellular proteins [84], raising difficult questions about their origins. One possible scenario involves de novo origin of the CPs in the genomes of primordial selfish replicons through mechanisms such as overprinting and diversification [85]. Alternatively, CPs evolved from ancestral proteins that originally performed cellular functions but were subsequently recruited for capsid formation and underwent substantial acceleration of evolution.

Notwithstanding this apparent fast evolution of virus structural proteins, our recent comprehensive analysis of the evolutionary connections of the major virion proteins, including capsid, nucleocapsid and matrix proteins, using sensitive methods for sequence and structure comparison, indicates that many

CPs evolved from ancestral cellular proteins on multiple, independent occasions [76]. The natural history of the jelly-roll fold illustrates well both multiple cellular origins and the subsequent extent of evolution (Figure 3). The single jelly-roll (SJR) is the most common fold in viral CPs, forming capsids in viruses from about one third of virus families, mostly in eukaryotic ssRNA and ssDNA viruses. High-resolution CP structures are available for many SJR-CPs [73,78], and searching protein structure databases with these structures as queries revealed numerous and diverse homologs from bacteria, archaea and eukaryotes [76]. The cellular SJR domains are functionally diverse, including a variety of carbohydrate-binding proteins [86], nucleoplasmins or nucleophosmins [87] and various proteins of the tumor necrosis factor (TNF) superfamily [88]. Although obtaining a robust phylogenetic tree for the SJR domain might not be feasible due to the lack of sufficient sequence conservation, clustering by structural similarity suggests two or three independent points of origin for the SJR-CPs, possibly, from carbohydrate-binding proteins. Indeed, some of the cellular SJR proteins, such as the TNF homologs, can assemble into structures resembling virus particles [89]. Furthermore, ancestral viruses would benefit substantially from capture of a carbohydrate-binding protein that provides both genome protection and receptor-binding capacity. Accordingly, many extant viruses bind to specific glycan receptors directly through the SJR-CP [90-93]. The proposed direction of SJR recruitment, from cells to viruses, appears far more likely than the reverse direction, given the wide spread and functional diversity of the SJR protein in diverse cellular life forms and by contrast, their paucity among bacterial and archaeal viruses. Notably, whereas structural comparisons suggest the monophyly of SJR-CPs of eukaryotic ssRNA and ssDNA viruses, the SJR-CPs of bacterial microviruses and sphaerolipoviruses have been apparently acquired independently of each other and of eukaryotic viruses [76]. Within the eukaryotic positive-sense RNA viruses, SJR-CP evolution is largely congruent with the RdRp evolution, complemented by horizontal transfer of SJR-CP genes that is particularly frequent among the smallest RNA and DNA viruses [31,43,77,94]. Gene gain and loss analysis has suggested that the SJR-CP was ancestral in the eukaryotic RNA virome but subsequently was replaced, on many independent occasions, with various non-homologous capsid and nucleocapsid proteins, and even lost in some virus groups, leading to the emergence of capsidless 'viruses' [43].

The second most prevalent class of viral CPs includes the double jelly-roll (DJR) fold that consists of two consecutive SJRs [95]. The DJR-CP is found in about 10% of known virus taxa, including numerous dsDNA viruses that infect bacteria, archaea and eukaryotes. Most of the ubiquitous, icosahedral, tailless dsDNA phages detected in marine environments [63] likely have capsids assembled from the DJR-CP [96]. This possibility is reinforced by the recent discovery of the remarkable diversity of DJR-encoding bacterial and archaeal viruses in viromes and microbial (meta)genomes [97]. No DJR proteins of cellular origin have been detected suggesting that the DJR CP evolved from a SJR-CP by gene duplication in an ancestral virus genome [98]. Bacterial and archaeal viruses that belong to the family *Sphaerolipoviridae* might represent an intermediate step on the evolutionary path from SJR to the DJR capsids [99,100]. These viruses have two SJR-CPs that form homo- and heterodimers and the vertical orientation of the β-strands in the capsids matches DJR-CPs but differs from that of SJR-CPs of small RNA and DNA viruses [73]. Consistent with a SJR-CP gene duplication giving rise to the DJR-CP, the two CPs of sphaerolipoviruses show a closer structural similarity to each other than to any other CPs [76,100]. Among the SJR-CPs, those of sphaerolipoviruses are weakly similar to homologs from other viruses and instead cluster with nucleoplasmins and nucleophosmins, which implies independent origin from a cellular SJR ancestor [76].

We have also identified several other cases of apparent evolution of major virion proteins via recruitment of ancestral cellular proteins. For example, the CPs of alphaviruses evolved from a chymotrypsin-like protease closely related to the polyprotein-processing protease of flaviviruses [76,101], whereas the nucleocapsid protein of reverse-transcribing viruses of the recently introduced order *Ortervirales* [102] originated from a zinc-knuckle domain that is widespread in functionally diverse cellular

proteins [76]. In archaea, one of the nucleocapsid proteins of tristromaviruses has evolved from a truncated Cas4-like nuclease [103]. Similarly, matrix proteins, which form a protein layer between the nucleocapsid and the viral envelope, of arenaviruses, mononegaviruses and retroviruses evolved from the zinc-binding RING domain, cyclophilins and a DNA-binding domain of integrases, respectively [76]. Furthermore, the pan-eukaryotic fusogen HAP2 has been recently shown to be homologous to class II viral fusion proteins [104-106], which in many RNA viruses assemble into an external icosahedral shell [107,108]. Although the directionality of evolution is still debated in this case, given the ubiquity of HAP2 in eukaryotes in contrast to the limited distribution of class II fusogens in viruses, a cellular origin seems more likely, with viruses acquiring the fusogenic capacity at a later stage. Together with the evolutionary scenarios for the jelly-roll CPs outlined above, all these cases convey the same message, that of multiple origins of major virion proteins from cellular ancestors.

Virus capsid proteins apparently evolved from cellular ancestors at different stages of evolution, from early ones, shortly after the emergence of modern-type cells, to comparatively late stages, after the origin of eukaryotes or even the origin of animals. Although pinpointing the times at which distinct lineages of virus capsid proteins emerged is not an easy task, important clues can be derived from the host ranges of the extant viruses. Given the high prevalence of the SJR and DJR CPs among viruses infecting all three domains of cellular life, these proteins most likely emerged during or shortly after the advent of modern-type bacterial and archaeal cells. By contrast, the exclusive yet ubiquitous presence of reverse-transcribing viruses in eukaryotes implies that the nucleocapsid and capsid domains of the Gag polyprotein were acquired by ancestral reverse-transcribing virus at an early stage of eukaryogenesis [109]. Furthermore, the matrix protein domain, which is specific to members of the family *Retroviridae*, was likely appended to the Gag polyprotein following the divergence of retroviruses from other orterviruses, probably after the emergence of vertebrates. The apparent absence of negative-sense RNA viruses in bacteria, archaea and protists, which contrasts with their high prevalence in animals, strongly suggests that these viruses and their particular nucleocapsids co-emerged with metazoa and later spread to plants [110,111]. A similar scenario is likely for the recruitment of protease as the CP by alphaviruses [76,101] and of the RING domain by arenaviruses, whose host ranges are limited to animals.

The hypothesis that cellular proteins can be recruited to function as virus CPs draws additional support from recent breakthroughs in computational modeling of large protein assemblies, protein engineering and laboratory evolution experiments. Artificial, megadalton-scale two-component, 120-subunit icosahedral protein complexes have been designed computationally [112] and, following introduction of positively charged residues on the inner surface of the shells, these cages could encapsidate their own mRNA genomes [113]. Importantly, under selective pressure, such nucleoprotein assemblies quickly evolved for improved genome packaging, stability, and in vivo circulation time, all properties relevant for natural virions. A similar result was obtained with the naturally occurring, non-viral icosahedral shells of lumazine synthase from *Aquifex aeolicus* [114]. Appending a cationic peptide to the lumazine synthase enabled specific recognition of packaging signals on cognate mRNAs and subsequent evolutionary optimization led to virus-like capsids that were large enough to accomodate the cognate full-length RNA genome in vivo and protected the cargo RNA molecules from nucleases [114]. These studies demonstrate the relative ease of the emergence of virus-like entities from artificial, computationally designed as well as natural non-viral proteins.

Admittedly, the likely cellular ancestors are not detectable for all virus structural proteins so far, including several widespread varieties. In particular, the capsid proteins of the highly abundant tailed bacterial and archaeal viruses have cellular homologs that form intracellular nanocompartments in many bacteria and archaea, known as encapsulins [115]. In this case, there is no evidence of the direction of

evolution and a cellular origin remains a distinct possibility. For capsid proteins of rod-shaped and filamentous plant viruses as well as the capsid protein of most viruses of hyperthermophilic archaea, no cellular homologs have been identified. *De novo* origin of some of these 'orphan' virion proteins cannot be discarded as a distinct route of virus evolution. Nevertheless, overall, the exaptation of cellular proteins seems to be a common theme.

[H1] The fourth path for virus origin: an amalgam of the primordial and escaped genes scenarios

The analyses of the evolution of the replicative and structural modules of viruses discussed above suggest a 'chimeric' scenario for the origin of viruses, which is distinct from each of the three traditional scenarios (Figure 1) but combines features of the primordial and escaped genes scenarios (Figure 4). In this model, the pre-cellular and early cellular stages of evolution include various forms of parasitic replicators with RNA and DNA genomes. Emergence of such selfish elements seems to be an intrinsic feature of replicator systems [6,7]. At those early stages of evolution, the reproduction strategies of the parasitic elements resembled those of present day plasmids and transposons. Highly conserved transposases are widely represented in all cellular organisms[116], suggesting that integration into the host genome is an ancient strategy of genetic parasites. There is no evidence that the key replication proteins of small selfish replicators, such as RdRp, RT or RCRE, were ever encoded by bona fide cellular genes, and the same seems to apply to transposases. Thus, the replication modules of viruses seem to originate from the primordial genetic pool although the long course of their subsequent evolution involved many displacements by replicative genes from their cellular hosts.

This evidence in support of the primordial origins of the key components of the virus replication machinery contrasts the findings on the provenance of the components of the translation system (along with numerous other proteins of diverse functions) encoded by giant viruses [19,24]. Although the cell regression scenario of virus origin has been boosted by the discovery of these genes, detailed phylogenetic analyses suggest incremental, convergent capture of the translation-related genes from different eukaryotic hosts during the evolution of different giant viruses from smaller virus ancestors [20,117]. This scenario is further supported by the recent discovery of multiple ribosomal protein genes in several bacterial viruses with moderately-sized genomes [118]. Thus, the evolution of giant viruses, irrespective of the numerous interesting and puzzling aspects of their genome layout and biology, can be accommodated in the evolutionary scenario proposed here. Also, no evidence exists for the possible origin of viruses from intracellular parasitic bacteria. As intracellular parasitic or symbiotic bacteria have evolved numerous times and have independently given rise to extremely reduced forms, including organelles [119,120], the absence of bacteria-derived viruses suggests that the evolutionary path from a cell to a virus is impracticable.

The first CPs – and hence the first true viruses – most likely evolved as a result of recruitment of carbohydrate-binding or nucleic acid-binding proteins from cells that were advanced enough to encode multiple proteins with these functions. As most common viral capsids have simple, symmetrical, thermodynamically favorable shapes, such as icosahedra or helices, the structural transitions involved in the recruitment of cellular proteins as CPs could have been relatively minor. The multiple recruitments of unrelated proteins as CPs forming icosahedral capsids seem to support this view. Another argument for easy formation of capsid-like structures is the recruitment of two distinct proteins as structural units for capsid-like intracellular microcompartments [83]. Recent results from experimental evolution of virus-like particles through relatively small modification of cellular enzymes, primarily by introduction of positively charge residues interacting with nucleic acids, further reinforce this line of argument [114].

As emphasized above, genetic parasites are an integral feature of replicator systems and, accordingly, of all cellular life forms. These parasites are fundamentally diverse, spanning a wide range of relationships with the host, ranging from symbiotic, such as many plasmids and moderate, integrating viruses, to extremely aggressive, such as numerous lytic viruses [121,122]. Our current model posits that genetic parasites started out as relatively cooperative commensals or symbionts but, subsequently, on multiple occasions, recruited cellular structural proteins to evolve into elaborate selfish agents employing diverse but, generally, highly efficient reproduction strategies. The tight evolutionary link between viruses and capsidless MGEs is the core of our model of virus origin and is amply supported by comparative analysis of contemporary genetic parasites in all domains of cellular life [77,123,124].

REFERENCES

1 Danovaro, R. *et al.* Virus-mediated archaeal hecatomb in the deep seafloor. *Sci Adv* **2**, e1600492 (2016).
2 Chow, C. E. & Suttle, C. A. Biogeography of Viruses in the Sea. *Annu Rev Virol* **2**, 41-66 (2015).
3 Cobián Güemes, A. G. *et al.* Viruses as Winners in the Game of Life. *Annu Rev Virol* **3**, 197-214 (2016).
4 Koonin, E. V. & Dolja, V. V. A virocentric perspective on the evolution of life. *Curr Opin Virol* **3**, 546-557 (2013).
5 Raoult, D. & Forterre, P. Redefining viruses: lessons from Mimivirus. *Nat Rev Microbiol* **6**, 315-319 (2008).
6 Koonin, E. V., Wolf, Y. I. & Katsnelson, M. I. Inevitability of the emergence and persistence of genetic parasites caused by evolutionary instability of parasite-free states. *Biol Direct* **12**, 31 (2017).
7 Iranzo, J., Puigbo, P., Lobkovsky, A. E., Wolf, Y. I. & Koonin, E. V. Inevitability of Genetic Parasites. *Genome Biol Evol* **8**, 2856-2869 (2016).
8 Koonin, E. V. Viruses and mobile elements as drivers of evolutionary transitions. *Philos Trans R Soc Lond B Biol Sci* **371** (2016).
9 Forterre, P. & Prangishvili, D. The major role of viruses in cellular evolution: facts and hypotheses. *Curr Opin Virol* **3**, 558-565 (2013).
10 Frank, J. A. & Feschotte, C. Co-option of endogenous viral sequences for host cell function. *Curr Opin Virol* **25**, 81-89 (2017).
11 Forterre, P. The origin of viruses and their possible roles in major evolutionary transitions. *Virus Res* **117**, 5-16 (2006).
12 Luria, S. E. & Darnell, J. E. *General virology*. (Wiley, 1967).
13 Sapp, J. The prokaryote-eukaryote dichotomy: meanings and mythology. *Microbiol Mol Biol Rev* **69**, 292-305 (2005).
14 Flugel, R. M. The precellular scenario of genovirions. *Virus Genes* **40**, 151-154 (2010).
15 Forterre, P. & Prangishvili, D. The origin of viruses. *Res Microbiol* **160**, 466-472 (2009).
16 Koonin, E. V., Senkevich, T. G. & Dolja, V. V. The ancient Virus World and evolution of cells. *Biol Direct* **1**, 29 (2006).
17 Morse, S. S. in *The evolutionary biology of viruses* (ed Morse S. S.) 1–28 (Raven Press, 1994).
18 Holmes, E. C. What does virus evolution tell us about virus origins? *J Virol* **85**, 5247-5251 (2011).
19 Abrahao, J. *et al.* Tailed giant Tupanvirus possesses the most complete translational apparatus of the known virosphere. *Nat Commun* **9**, 749 (2018).

20 Schulz, F. *et al.* Giant viruses with an expanded complement of translation system components. *Science* **356**, 82-85 (2017).
21 Abergel, C., Legendre, M. & Claverie, J. M. The rapidly expanding universe of giant viruses: Mimivirus, Pandoravirus, Pithovirus and Mollivirus. *FEMS Microbiol Rev* **39**, 779-796 (2015).
22 Nasir, A. & Caetano-Anolles, G. A phylogenomic data-driven exploration of viral origins and evolution. *Sci Adv* **1**, e1500527 (2015).
23 Colson, P., La Scola, B., Levasseur, A., Caetano-Anolles, G. & Raoult, D. Mimivirus: leading the way in the discovery of giant viruses of amoebae. *Nat Rev Microbiol* **15**, 243-254 (2017).
24 Abrahao, J. S., Araujo, R., Colson, P. & La Scola, B. The analysis of translation-related gene set boosts debates around origin and evolution of mimiviruses. *PLoS Genet* **13**, e1006532 (2017).
25 Forterre, P. & Krupovic, M. in *Viruses: Essential Agents of Life* (ed G. Witzany) 43-60 (Springer Science+Business Media, 2012).
26 Fridman, S. *et al.* A myovirus encoding both photosystem I and II proteins enhances cyclic electron flow in infected Prochlorococcus cells. *Nat Microbiol* **2**, 1350-1357 (2017).
27 Roux, S. *et al.* Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537**, 689-693 (2016).
28 Pushkarev, A. *et al.* A distinct abundant group of microbial rhodopsins discovered using functional metagenomics. *Nature* **558**, 595-599 (2018).
29 Ahlgren, N. A., Fuchsman, C. A., Rocap, G. & Fuhrman, J. A. Discovery of several novel, widespread, and ecologically distinct marine Thaumarchaeota viruses that encode amoC nitrification genes. *Isme J* **13**, 618-631 (2019).
30 Iranzo, J., Krupovic, M. & Koonin, E. V. The Double-Stranded DNA Virosphere as a Modular Hierarchical Network of Gene Sharing. *MBio* **7**, e00978-00916 (2016).
31 Koonin, E. V., Dolja, V. V. & Krupovic, M. Origins and evolution of viruses of eukaryotes: The ultimate modularity. *Virology* **479-480**, 2-25 (2015).
32 Krupovic, M. & Bamford, D. H. Order to the viral universe. *J Virol* **84**, 12476-12479 (2010).
33 Kazlauskas, D., Krupovic, M. & Venclovas, C. The logic of DNA replication in double-stranded DNA viruses: insights from global analysis of viral genomes. *Nucleic Acids Res* **44**, 4551-4564 (2016).
34 Forterre, P. 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 U S A* **103**, 3669-3674 (2006).
35 Autexier, C. & Lue, N. F. The structure and function of telomerase reverse transcriptase. *Annu Rev Biochem* **75**, 493-517 (2006).
36 te Velthuis, A. J. Common and unique features of viral RNA-dependent polymerases. *Cell Mol Life Sci* **71**, 4403-4420 (2014).
37 Venkataraman, S., Prasad, B. & Selvarajan, R. RNA Dependent RNA Polymerases: Insights from Structure, Function and Evolution. *Viruses* **10**, 76 (2018).
38 Mönttinen, H. A., Ravantti, J. J. & Poranen, M. M. Common Structural Core of Three-Dozen Residues Reveals Intersuperfamily Relationships. *Mol Biol Evol* **33**, 1697-1710 (2016).
39 Iyer, L. M., Koonin, E. V., Leipe, D. D. & Aravind, L. Origin and evolution of the archaeo-eukaryotic primase superfamily and related palm-domain proteins: structural insights and new members. *Nucleic Acids Res* **33**, 3875-3896 (2005).
40 Kazlauskas, D. *et al.* Novel Families of Archaeo-Eukaryotic Primases Associated with Mobile Genetic Elements of Bacteria and Archaea. *J Mol Biol* **430**, 737-750 (2018).
41 Clery, A., Blatter, M. & Allain, F. H. RNA recognition motifs: boring? Not quite. *Curr Opin Struct Biol* **18**, 290-298 (2008).
42 Gilbert, W. Origin of life: The RNA world. *Nature* **319**, 618 (1986).
43 Wolf, Y. I. *et al.* Origins and Evolution of the Global RNA Virome. *MBio* **9**, e02329-02318 (2018).

44 Koonin, E. V., Wolf, Y. I., Nagasaki, K. & Dolja, V. V. The Big Bang of picorna-like virus evolution antedates the radiation of eukaryotic supergroups. *Nat Rev Microbiol* **6**, 925-939 (2008).
45 McNeil, B. A., Semper, C. & Zimmerly, S. Group II introns: versatile ribozymes and retroelements. *Wiley Interdiscip Rev RNA* **7**, 341-355 (2016).
46 Iyer, L. M., Koonin, E. V. & Aravind, L. Evolutionary connection between the catalytic subunits of DNA-dependent RNA polymerases and eukaryotic RNA-dependent RNA polymerases and the origin of RNA polymerases. *BMC Struct Biol* **3**, 1 (2003).
47 Salgado, P. S. *et al.* The structure of an RNAi polymerase links RNA silencing and transcription. *PLoS Biol* **4**, e434 (2006).
48 Weigel, C. & Seitz, H. Bacteriophage replication modules. *FEMS Microbiol Rev* **30**, 321-381 (2006).
49 Novikova, O. & Belfort, M. Mobile Group II Introns as Ancestral Eukaryotic Elements. *Trends Genet* **33**, 773-783 (2017).
50 Agrawal, R. K., Wang, H. W. & Belfort, M. Forks in the tracks: Group II introns, spliceosomes, telomeres and beyond. *RNA Biol* **13**, 1218-1222 (2016).
51 Takeuchi, N., Hogeweg, P. & Koonin, E. V. On the origin of DNA genomes: evolution of the division of labor between template and catalyst in model replicator systems. *PLoS Comput Biol* **7**, e1002024 (2011).
52 Ren, A., Micura, R. & Patel, D. J. Structure-based mechanistic insights into catalysis by small self-cleaving ribozymes. *Curr Opin Chem Biol* **41**, 71-83 (2017).
53 Lee, K. Y. & Lee, B. J. Structural and Biochemical Properties of Novel Self-Cleaving Ribozymes. *Molecules* **22**, E678 (2017).
54 Joyce, G. F. & Szostak, J. W. Protocells and RNA Self-Replication. *Cold Spring Harb Perspect Biol* **10**, a034801 (2018).
55 Lancet, D., Zidovetzki, R. & Markovitch, O. Systems protobiology: origin of life in lipid catalytic networks. *J R Soc Interface* **15**, 20180159 (2018).
56 Mulkidjanian, A. Y., Bychkov, A. Y., Dibrova, D. V., Galperin, M. Y. & Koonin, E. V. Origin of first cells at terrestrial, anoxic geothermal fields. *Proc Natl Acad Sci U S A* **109**, E821-830 (2012).
57 Martin, W., Baross, J., Kelley, D. & Russell, M. J. Hydrothermal vents and the origin of life. *Nat Rev Microbiol* **6**, 805-814 (2008).
58 Koonin, E. V. & Martin, W. On the origin of genomes and cells within inorganic compartments. *Trends Genet* **21**, 647-654 (2005).
59 Bamford, D. H. Do viruses form lineages across different domains of life? *Res Microbiol* **154**, 231-236 (2003).
60 Forterre, P., Krupovic, M. & Prangishvili, D. Cellular domains and viral lineages. *Trends Microbiol* **22**, 554-558 (2014).
61 Caspar, D. L. & Klug, A. Physical principles in the construction of regular viruses. *Cold Spring Harb Symp Quant Biol* **27**, 1-24 (1962).
62 Crick, F. H. & Watson, J. D. Structure of small viruses. *Nature* **177**, 473-475 (1956).
63 Brum, J. R., Schenck, R. O. & Sullivan, M. B. Global morphological analysis of marine viruses shows minimal regional variation and dominance of non-tailed viruses. *Isme J* **7**, 1738-1751 (2013).
64 Solovyev, A. G. & Makarov, V. V. Helical capsids of plant viruses: architecture with structural lability. *J Gen Virol* **97**, 1739-1754 (2016).
65 DiMaio, F. *et al.* Virology. A virus that infects a hyperthermophile encapsidates A-form DNA. *Science* **348**, 914-917 (2015).
66 Ptchelkine, D. *et al.* Unique architecture of thermophilic archaeal virus APBV1 and its genome packaging. *Nat Commun* **8**, 1436 (2017).

67 Zamora, M. *et al.* Potyvirus virion structure shows conserved protein fold and RNA binding site in ssRNA viruses. *Sci Adv* **3**, eaao2182 (2017).
68 DiMaio, F. *et al.* The molecular basis for flexibility in the flexible filamentous plant viruses. *Nat Struct Mol Biol* **22**, 642-644 (2015).
69 Sun, Y., Li, J., Gao, G. F., Tien, P. & Liu, W. Bunyavirales ribonucleoproteins: the viral replication and transcription machinery. *Crit Rev Microbiol* **44**, 522-540 (2018).
70 Sun, Y., Guo, Y. & Lou, Z. A versatile building block: the structures and functions of negative-sense single-stranded RNA virus nucleocapsid proteins. *Protein Cell* **3**, 893-902 (2012).
71 Jamin, M. & Yabukarski, F. Nonsegmented Negative-Sense RNA Viruses-Structural Data Bring New Insights Into Nucleocapsid Assembly. *Adv Virus Res* **97**, 143-185 (2017).
72 Prangishvili, D. *et al.* The enigmatic archaeal virosphere. *Nat Rev Microbiol* **15**, 724-739 (2017).
73 Abrescia, N. G., Bamford, D. H., Grimes, J. M. & Stuart, D. I. Structure unifies the viral universe. *Annu Rev Biochem* **81**, 795-822 (2012).
74 Greene, L. H. *et al.* The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Res* **35**, D291-297 (2007).
75 Krupovic, M. & Bamford, D. H. Double-stranded DNA viruses: 20 families and only five different architectural principles for virion assembly. *Curr Opin Virol* **1**, 118-124 (2011).
76 Krupovic, M. & Koonin, E. V. Multiple origins of viral capsid proteins from cellular ancestors. *Proc Natl Acad Sci U S A* **114**, E2401-E2410 (2017).
77 Krupovic, M. Networks of evolutionary interactions underlying the polyphyletic origin of ssDNA viruses. *Curr Opin Virol* **3**, 578-586 (2013).
78 Rossmann, M. G. & Johnson, J. E. Icosahedral RNA virus structure. *Annu Rev Biochem* **58**, 533-573 (1989).
79 Moreira, D. & López-García, P. Ten reasons to exclude viruses from the tree of life. *Nat Rev Microbiol* **7**, 306-311 (2009).
80 Sasaki, E. *et al.* Structure and assembly of scalable porous protein cages. *Nat Commun* **8**, 14663 (2017).
81 Ladenstein, R., Fischer, M. & Bacher, A. The lumazine synthase/riboflavin synthase complex: shapes and functions of a highly variable enzyme system. *Febs J* **280**, 2537-2563 (2013).
82 Kerfeld, C. A., Aussignargues, C., Zarzycki, J., Cai, F. & Sutter, M. Bacterial microcompartments. *Nat Rev Microbiol* **16**, 277-290 (2018).
83 Krupovic, M. & Koonin, E. V. Cellular origin of the viral capsid-like bacterial microcompartments. *Biol Direct* **12**, 25 (2017).
84 Cheng, S. & Brooks, C. L., 3rd. Viral capsid proteins are segregated in structural fold space. *PLoS Comput Biol* **9**, e1002905 (2013).
85 Sabath, N., Wagner, A. & Karlin, D. Evolution of viral proteins originated de novo by overprinting. *Mol Biol Evol* **29**, 3767-3780 (2012).
86 Boraston, A. B., Bolam, D. N., Gilbert, H. J. & Davies, G. J. Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochem J* **382**, 769-781 (2004).
87 Eitoku, M., Sato, L., Senda, T. & Horikoshi, M. Histone chaperones: 30 years from isolation to elucidation of the mechanisms of nucleosome assembly and disassembly. *Cell Mol Life Sci* **65**, 414-444 (2008).
88 Locksley, R. M., Killeen, N. & Lenardo, M. J. The TNF and TNF receptor superfamilies: integrating mammalian biology. *Cell* **104**, 487-501 (2001).
89 Liu, Y. *et al.* Crystal structure of sTALL-1 reveals a virus-like assembly of TNF family ligands. *Cell* **108**, 383-394 (2002).

90 Shen, S., Bryant, K. D., Brown, S. M., Randell, S. H. & Asokan, A. Terminal N-linked galactose is the primary receptor for adeno-associated virus 9. *J Biol Chem* **286**, 13532-13540 (2011).
91 Neu, U., Bauer, J. & Stehle, T. Viruses and sialic acids: rules of engagement. *Curr Opin Struct Biol* **21**, 610-618 (2011).
92 Maginnis, M. S. Virus-Receptor Interactions: The Key to Cellular Invasion. *J Mol Biol* **430**, 2590-2611 (2018).
93 Liu, Y. *et al.* Sialic acid-dependent cell entry of human enterovirus D68. *Nat Commun* **6**, 8865 (2015).
94 Shi, M. *et al.* Redefining the invertebrate RNA virosphere. *Nature* (2016).
95 Benson, S. D., Bamford, J. K., Bamford, D. H. & Burnett, R. M. Does common architecture reveal a viral lineage spanning all three domains of life? *Mol Cell* **16**, 673-685 (2004).
96 Kauffman, K. M. *et al.* A major lineage of non-tailed dsDNA viruses as unrecognized killers of marine bacteria. *Nature* **554**, 118-122 (2018).
97 Yutin, N., Backstrom, D., Ettema, T. J. G., Krupovic, M. & Koonin, E. V. Vast diversity of prokaryotic virus genomes encoding double jelly-roll major capsid proteins uncovered by genomic and metagenomic sequence analysis. *Virol J* **15**, 67 (2018).
98 Abrescia, N. G. *et al.* Insights into virus evolution and membrane biogenesis from the structure of the marine lipid-containing bacteriophage PM2. *Mol Cell* **31**, 749-761 (2008).
99 Rissanen, I. *et al.* Bacteriophage P23-77 capsid protein structures reveal the archetype of an ancient branch from a major virus lineage. *Structure* **21**, 718-726 (2013).
100 Santos-Perez, I. *et al.* Structural basis for assembly of vertical single beta-barrel viruses. *Nat Commun* **10**, 1184 (2019).
101 Kuhn, R. J. & Rossmann, M. G. Structure and assembly of icosahedral enveloped RNA viruses. *Adv Virus Res* **64**, 263-284 (2005).
102 Krupovic, M. *et al. Ortervirales*: New virus order unifying five families of reverse-transcribing viruses. *J Virol* **92**, e00515-00518 (2018).
103 Krupovic, M., Cvirkaite-Krupovic, V., Prangishvili, D. & Koonin, E. V. Evolution of an archaeal virus nucleocapsid protein from the CRISPR-associated Cas4 nuclease. *Biol Direct* **10**, 65 (2015).
104 Pinello, J. F. *et al.* Structure-Function Studies Link Class II Viral Fusogens with the Ancestral Gamete Fusion Protein HAP2. *Curr Biol* **27**, 651-660 (2017).
105 Fedry, J. *et al.* The Ancient Gamete Fusogen HAP2 Is a Eukaryotic Class II Fusion Protein. *Cell* **168**, 904-915 e910 (2017).
106 Valansi, C. *et al.* Arabidopsis HAP2/GCS1 is a gamete fusion protein homologous to somatic and viral fusogens. *J Cell Biol* **216**, 571-581 (2017).
107 Guardado-Calvo, P. & Rey, F. A. The Envelope Proteins of the Bunyavirales. *Adv Virus Res* **98**, 83-118 (2017).
108 Modis, Y. Relating structure to evolution in class II viral membrane fusion proteins. *Curr Opin Virol* **5**, 34-41 (2014).
109 Krupovic, M. & Koonin, E. V. Homologous Capsid Proteins Testify to the Common Ancestry of Retroviruses, Caulimoviruses, Pseudoviruses, and Metaviruses. *J Virol* **91**, e00210-00217 (2017).
110 Dolja, V. V. & Koonin, E. V. Metagenomics reshapes the concepts of RNA virus evolution by revealing extensive horizontal virus transfer. *Virus Res* **244**, 36-52 (2018).
111 Shi, M., Zhang, Y. Z. & Holmes, E. C. Meta-transcriptomics and the evolutionary biology of RNA viruses. *Virus Res* **243**, 83-90 (2018).
112 Bale, J. B. *et al.* Accurate design of megadalton-scale two-component icosahedral protein complexes. *Science* **353**, 389-394 (2016).
113 Butterfield, G. L. *et al.* Evolution of a designed protein assembly encapsulating its own RNA genome. *Nature* **552**, 415-420 (2017).

114 Terasaka, N., Azuma, Y. & Hilvert, D. Laboratory evolution of virus-like nucleocapsids from nonviral protein cages. *Proc Natl Acad Sci U S A* **115**, 5432-5437 (2018).
115 Nichols, R. J., Cassidy-Amstutz, C., Chaijarasphong, T. & Savage, D. F. Encapsulins: molecular biology of the shell. *Crit Rev Biochem Mol Biol* **52**, 583-594 (2017).
116 Craig, N. L. *et al. Mobile DNA III*. 3rd ed edn, (ASM Press, 2015).
117 Koonin, E. V., Krupovic, M. & Yutin, N. Evolution of double-stranded DNA viruses of eukaryotes: from bacteriophages to transposons to giant viruses. *Ann N Y Acad Sci* **1341**, 10-24 (2015).
118 Mizuno, C. M. *et al.* Numerous cultivated and uncultivated viruses encode ribosomal proteins. *Nat Commun* **10**, 752 (2019).
119 Casadevall, A. Evolution of intracellular pathogens. *Annu Rev Microbiol* **62**, 19-33 (2008).
120 López-García, P., Eme, L. & Moreira, D. Symbiosis in eukaryotic evolution. *J Theor Biol* **434**, 20-33 (2017).
121 Koonin, E. V. & Starokadomskyy, P. Are viruses alive? The replicator paradigm sheds decisive light on an old but misguided question. *Stud Hist Philos Biol Biomed Sci* **59**, 125-134 (2016).
122 Jalasvuori, M. Vehicles, replicators, and intercellular movement of genetic information: evolutionary dissection of a bacterial cell. *Int J Evol Biol* **2012**, 874153 (2012).
123 Koonin, E. V. & Dolja, V. V. Virus world as an evolutionary network of viruses and capsidless selfish elements. *Microbiol Mol Biol Rev* **78**, 278-303 (2014).
124 Iranzo, J., Koonin, E. V., Prangishvili, D. & Krupovic, M. Bipartite Network Analysis of the Archaeal Virosphere: Evolutionary Connections between Viruses and Capsidless Mobile Elements. *J Virol* **90**, 11043-11055 (2016).
125 Gladyshev, E. A. & Arkhipova, I. R. A widespread class of reverse transcriptase-related cellular genes. *Proc Natl Acad Sci U S A* **108**, 20311-20316 (2011).
126 Arkhipova, I. R. Using bioinformatic and phylogenetic approaches to classify transposable elements and understand their complex evolutionary histories. *Mob DNA* **8**, 19 (2017).
127 Krupovic, M., Beguin, P. & Koonin, E. V. Casposons: mobile genetic elements that gave rise to the CRISPR-Cas adaptation machinery. *Curr Opin Microbiol* **38**, 36-43 (2017).
128 Koonin, E. V. & Krupovic, M. Polintons, virophages and transpovirons: a tangled web linking viruses, transposons and immunity. *Curr Opin Virol* **25**, 7-15 (2017).
129 Chandler, M. *et al.* Breaking and joining single-stranded DNA: the HUH endonuclease superfamily. *Nat Rev Microbiol* **11**, 525-538 (2013).
130 Zhao, L., Rosario, K., Breitbart, M. & Duffy, S. Eukaryotic Circular Rep-Encoding Single-Stranded DNA (CRESS DNA) Viruses: Ubiquitous Viruses With Small Genomes and a Diverse Host Range. *Adv Virus Res* **103**, 71-133 (2019).
131 Erdmann, S., Tschitschko, B., Zhong, L., Raftery, M. J. & Cavicchioli, R. A plasmid from an Antarctic haloarchaeon uses specialized membrane vesicles to disseminate and infect plasmid-free cells. *Nat Microbiol* **2**, 1446-1455 (2017).
132 Filée, J. & Forterre, P. Viral proteins functioning in organelles: a cryptic origin? *Trends Microbiol* **13**, 510-513 (2005).

Acknowledgments

EVK is supported through the intramural program of the U.S. National Institutes of Health. MK was supported by l'Agence Nationale de la Recherche (ANR) (France) project ENVIRA (#ANR-17-CE15-0005-01).

Author contributions

All authors researched data for the article, contributed substantially to discussion of content, wrote the article and reviewed or edited the manuscript before submission.

Conflicts of interest

The authors declare no conflicts of interest.

Publisher's statement

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

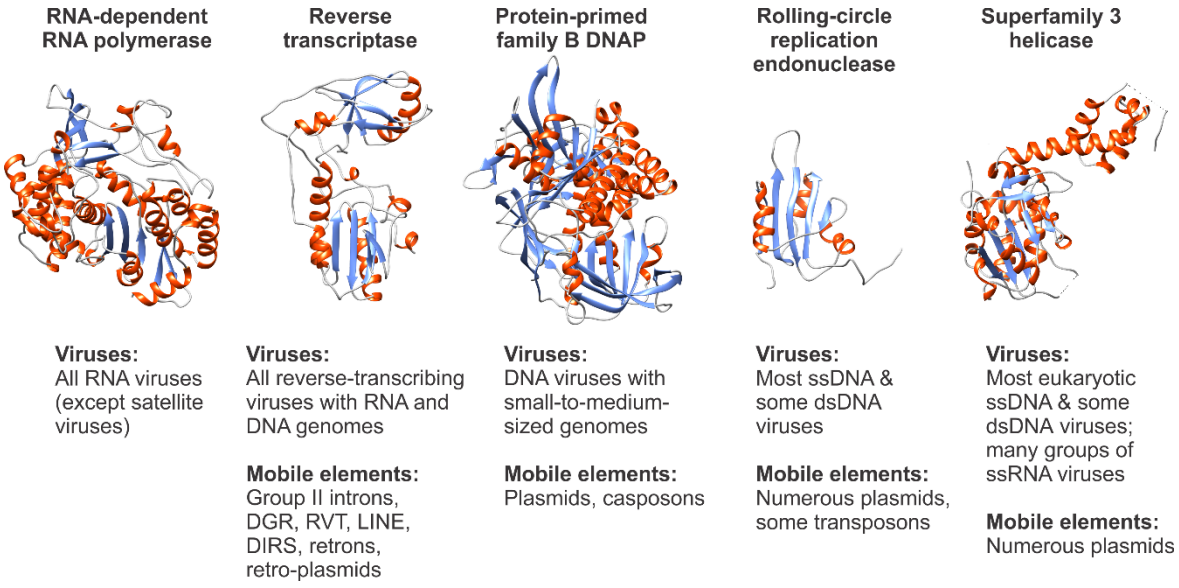
Reviewer information

Nature Reviews Microbiolog thanks Raul Andino, Purificación López-García, Didier Raoult and Yong-Zhen Zhang for their contribution to the peer review of this work.

Box 1. Hallmark viral replication genes.

Virus hallmark genes encode key proteins involved in genome replication and virion morphogenesis. These genes have no close homologs in cellular life forms but are shared by diverse viruses, and connect the evolutionary network in the virus world [16,30]. For example, the network of dsDNA viruses is held together by 14 hallmark genes, including those encoding different families of major capsid proteins, genome packaging ATPases, virion maturation proteases and various proteins involved in viral genome replication [30]. The reach of replication genes is particularly broad and connects not only viruses with different nucleic acid types but also links viruses and non-viral MGEs. Most notable are 5 broadly distributed virus hallmark genes involved in genome replication (Figure; structures are colored by secondary structure: α-helices, red; β-strands, blue; random coil, grey). RdRp is universally conserved across RNA viruses and, for this virus type, can be considered a universal marker equivalent to the 16S ribosomal RNA genes of cellular organisms (RdRp of poliovirus (PDB id: 1ra7) is shown in the figure). Indeed, RdRp allows studying the evolution of all RNA viruses in a single framework, despite the vast diversity of these viruses in other functional modules, most notably, those for morphogenesis [43]. RTs (catalytic fragment of Moloney murine leukemia virus RT (PDB id: 1mml) is shown in the figure) are encoded by all reverse-transcribing viruses with both RNA and DNA genomes (the latter sometimes unofficially referred to as the 'pararetroviruses') [102]. Unlike RdRps, RTs are not restricted to viruses but are found across a wide range of archaeal, bacterial and eukaryotic non-viral MGEs, including group II introns, diversity-generating retroelements (DGR), LINE, DIRS retrotransposons, retrons, retro-plasmids and others [123,125,126]. Furthermore, RTs have been recruited by cellular organisms on multiple independent occasions to perform specialized functions not linked with genome replication, for example, as telomerases[35]. Protein-primed family B DNA polymerases (PolBs; pPolB of bacteriophage phi29 (PDB id: 2py5) is shown in the figure) are homologous to the RNA-primed PolBs responsible for genome replication of archaea, eukaryotes and large dsDNA viruses. However, they have unique subdomains for strand displacement, a function performed by a helicase during cellular DNA replication, and are primed by a protein covalently attached to the termini of linear viral genomes rather than by an RNA primer. These polymerases are found in viruses with small-to-moderately-sized dsDNA or ssDNA genomes (10-50 kbp) [33]and besides viruses, also in linear cytoplasmic and mitochondrial plasmids of fungi and plants and in transposon-like elements called casposons [127,128]. Rolling-circle replication endonucleases (RCRE) of the HUH superfamily [129] are among the most widely spread replication proteins of viruses with small dsDNA and especially ssDNA genomes infecting hosts in all three domains of life [77,130] (RCRE of porcine circovirus

2 (PDBid: 2hw0) is shown in figure). Homologous RCRE are responsible for the replication of plasmids and transposition of certain transposons in bacteria, archaea, and eukaryotes [129]. Notably, in polyomaviruses and papillomaviruses (dsDNA genomes), the ancestral RCRE has lost the catalytic activity and now binds the replication origin [31]. Finally, hexameric superfamily 3 helicases (S3H; S3H of bovine papillomavirus type 1 (PDB id: 5a9k) is shown in figure) are involved in the replication of a wide range of viruses with ssRNA, ssDNA and dsDNA genomes, which range in size from 2 kbp in circoviruses to more than 1 Mbp in mimiviruses. The S3H domain is often fused to other functional domains involved in genome replication, including RCRE (in ssDNA viruses, polyomaviruses and papillomaviruses) and AEP or DnaG-like primases (in diverse dsDNA viruses) [33,39,77]. Similar multidomain proteins containing the S3H domain are also abundant in diverse bacterial and archaeal plasmids [40]. The distribution of the hallmark replication proteins in both viruses and non-viral MGEs illuminates the evolutionary connection between the two classes of genetic parasites and suggests that evolutionary transitions between the two are possible and likely had a key role in the emergence of the first viruses. A potential ongoing transition of a non-viral MGE into a virus is exemplified by a plasmid from an Antarctic haloarchaeon that uses specialized membrane vesicles to disseminate and infect plasmid-free cells [131].



Box Figure.

Figures and legends

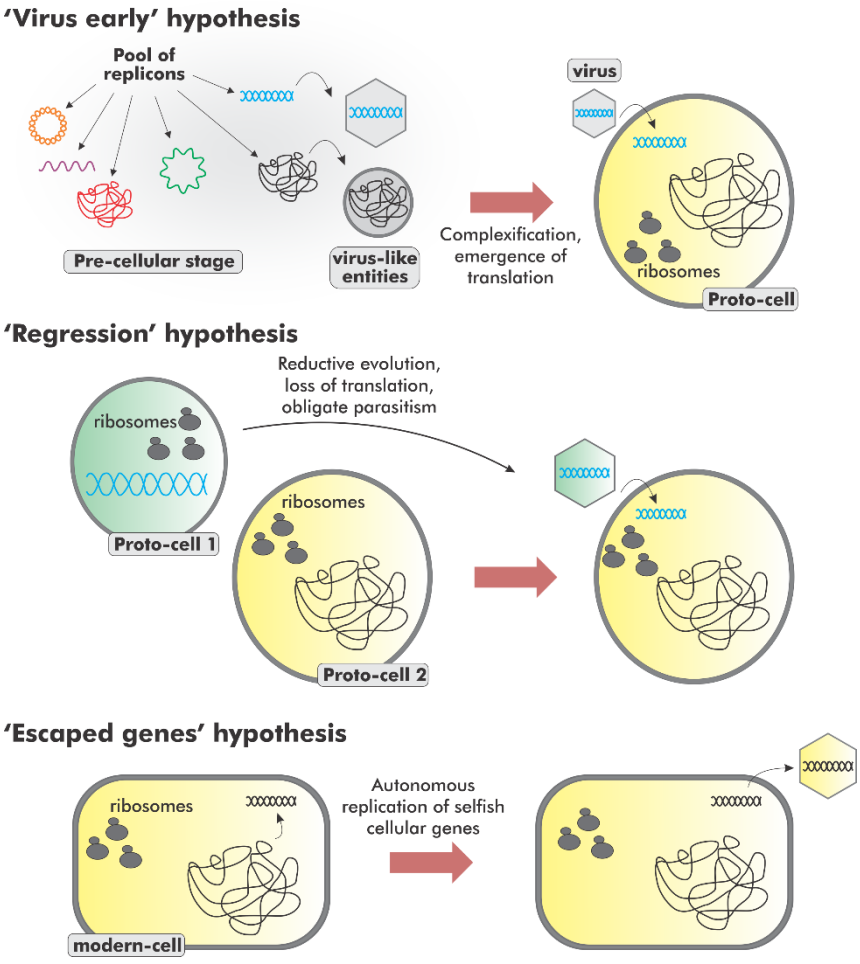


Figure 1. The three major scenarios for the origin of viruses. The 'virus early' hypothesis (top) assumes that viruses evolved from early replicative elements that preceded the first cellular life forms. The 'regression' hypothesis (middle) suggests that viruses emerged through the degeneration of cells that then assumed a parasitic lifestyle. Finally, the 'escaped genes' hypothesis (bottom) proposes that cellular genes acquired the ability for 'selfish' replication and spread.

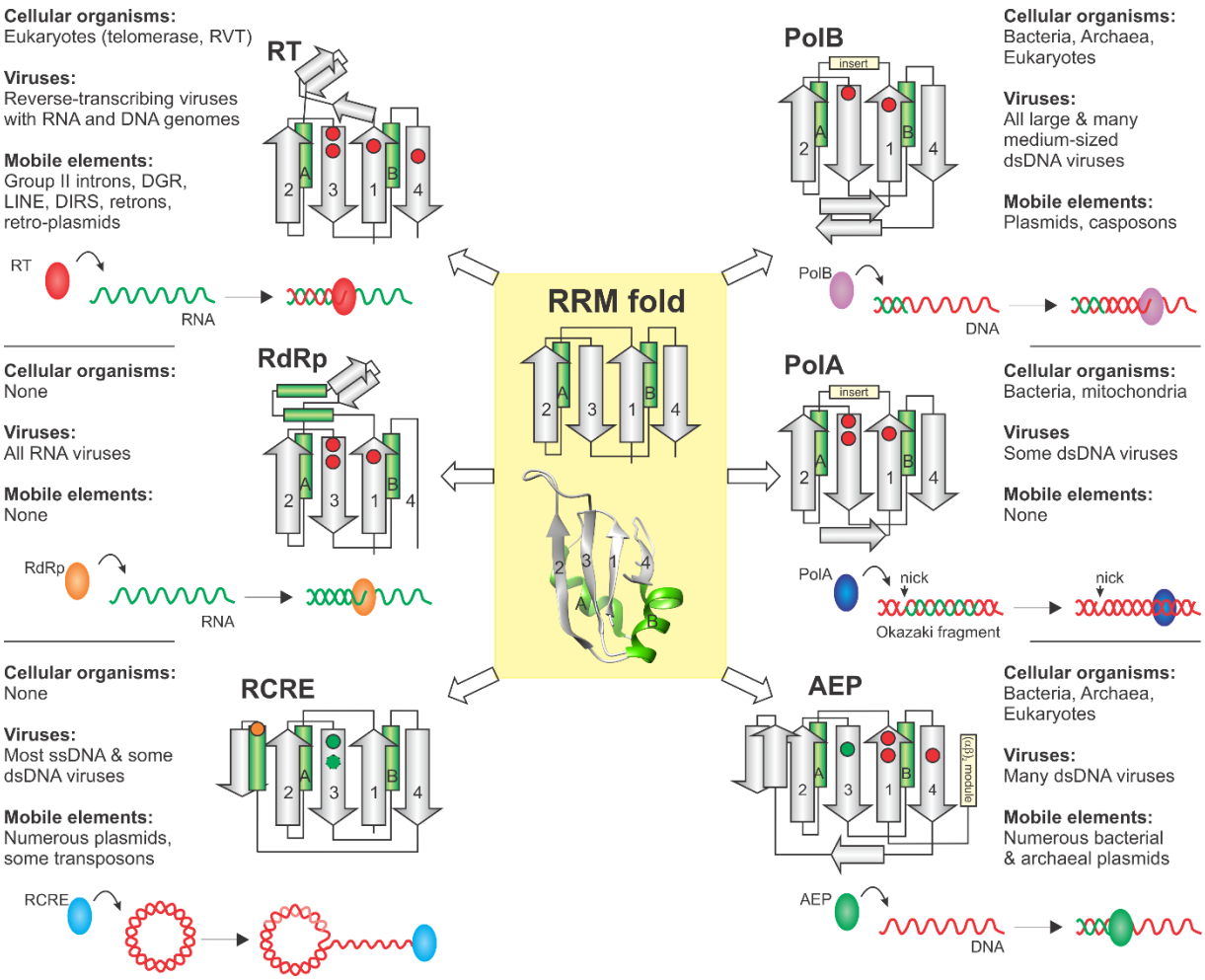


Figure 2. Evolution of viral and cellular replication modules from the ancestral RNA recognition motif. The RNA recognition motif (RRM) is one of the most common RNA-binding domains and occurs in all forms of cellular life. Structurally related domains are widespread in many viruses and mobile genetic elements. Replication enzymes containing this module include RT, reverse transcriptase; RdRp, RNA-dependent RNA polymerase; RCRE, rolling circle replication initiation endonuclease; PolA and PolB, DNA-dependent DNA polymerases of families A and B, respectively; and AEP, archaeo-eukaryotic primase. The schematic of the catalyzed reaction is shown for each enzyme, with red and green wavy lines representing DNA and RNA strands, respectively. Topologies of the proteins are shown with arrows for β-strands and green rectangles for α-helices, whereas larger insertions are depicted with yellow rectangles. Red, green and orange circles represent catalytic Asp, His and Tyr residues, respectively. In bacteria, PolA participates in the replication of the lagging DNA strand by connecting Okazaki fragments and in various repair processes, whereas in certain viruses, it functions as the replicative DNA polymerase that was also recruited from a bacteriophage for mitochondrial genome replication in eukaryotes [132].

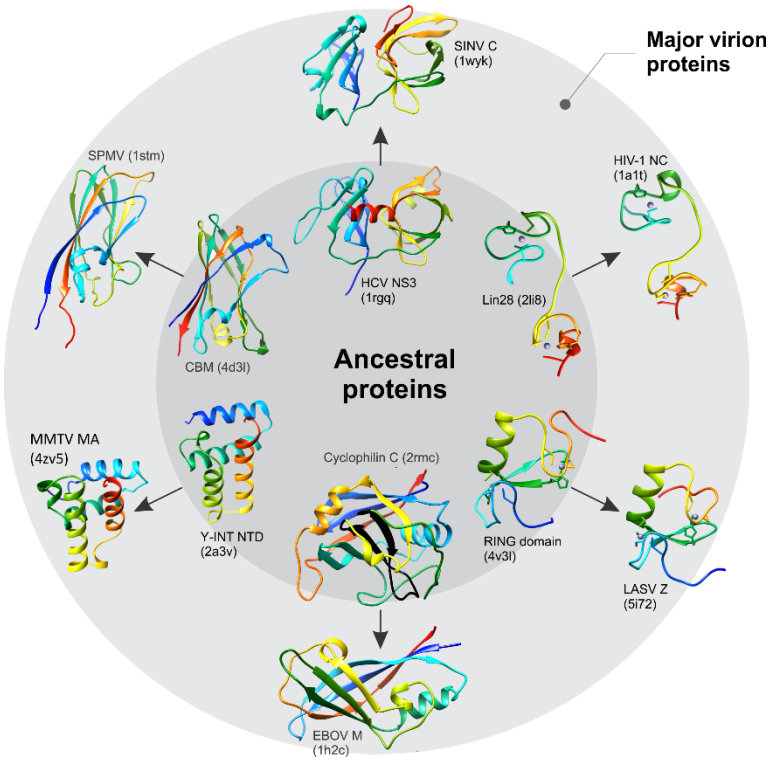


Figure 3. **Major structural virus proteins and their cellular homologs.** Many viral proteins that contribute to forming virions likely are derived from cellular carbohydrate-binding or nucleic acid-binding proteins, although some cellular proteins, as in the case of chymotrypsin-like proteases, are initially recruited to function as viral enzymes and only subsequently adapted for virion formation. SINV C, capsid protein of Sindbis virus (genus *Alphavirus*, family *Togaviridae*; PDB id: 1wyk); HIV-1 NC, nucleocapsid protein of human immunodeficiency virus 1 (family *Retroviridae*; PDB id: 1a1t); LASV Z, matrix protein Z of Lassa virus (family *Arenaviridae*; PDB id: 5i72); EBOV M, matrix protein of Ebola virus (family *Filoviridae*; PDB id: 1h2c); MMTV MA, matrix protein of mouse mammary tumor virus (family *Retroviridae*; PDB id: 4zv5); SPMV CP, single jelly-roll capsid protein of satellite panicum mosaic virus (genus *Papanivirus*; PDB id: 1stm); HCV NS3, non-structural protein 3 of hepatitis C virus (family *Flaviviridae*; PDB id: 1rgq); Lin28, human pluripotency factor Lin28 (PDB id: 2li8); RING, RING domain of ubiquitin ligase E3 (PDB id: 4v3l); Cyclophilin C (PDB id: 2rmc); Y-INT NTD, N-terminal domain of a tyrosine superfamily integrase (PDB id: 2a3v); CBM, carbohydrate binding motif (PDB id: 4d3l). All structures are colored using the rainbow scheme from blue (N terminus) to red (C terminus).

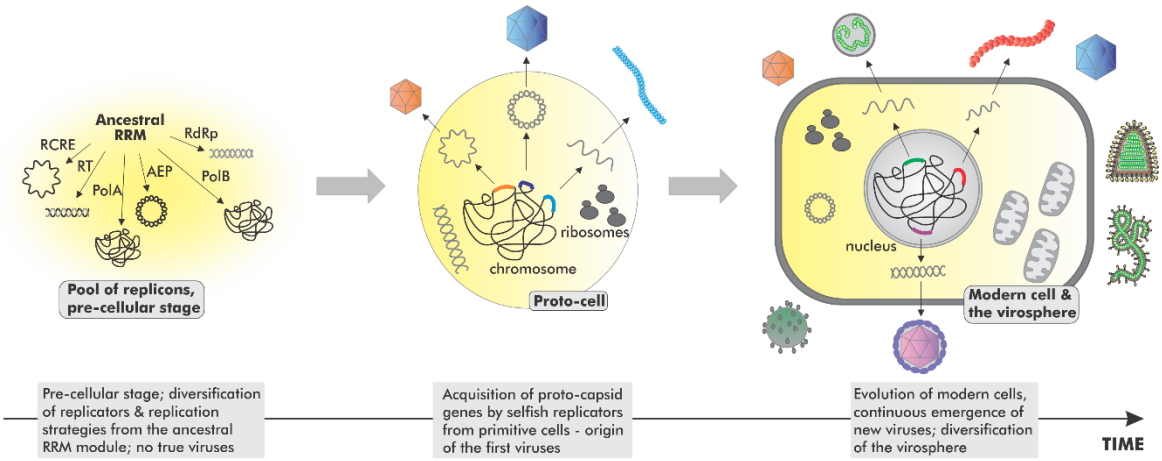


Figure 4. The chimeric scenario for the origin of viruses. In our model, the origin of viruses involves a two-stage process in which selfish replicators emerge before the first cellular life forms and then capture capsid protein genes from cellular organisms, which enables them to form virions. Continuing evolution and adoption of cellular genes contributes to further diversification of the virosphere.

Subject categories

Biological sciences / Microbiology / Virology / Viral evolution
[URI /631/326/596/2554]
Biological sciences / Microbiology / Virology / Virus structures
[URI /631/326/596/2148]
Biological sciences / Biochemistry / Proteins / Viral proteins
[URI /631/45/612/1256]

ToC blurb

The origin of viruses is an unsolved, controversial question. In this Opinion article, Krupovic, Dolja and Koonin propose a new scenario for the origin of viruses based on primordial, selfish replicators acquiring structural proteins from cells, enabling them to form virions.