Archives of Virology (2020) 165:2177–2191
https://doi.org/10.1007/s00705-020-04724-1

REVIEW

The origin and evolution of viruses inferred from fold family structure

Fizza Mughal[1,2] · Arshan Nasir[3,4] · Gustavo Caetano-Anollés[1,2]

Received: 2 April 2020 / Accepted: 30 May 2020 / Published online: 3 August 2020

Abstract
The canonical frameworks of viral evolution describe viruses as cellular predecessors, reduced forms of cells, or entities that escaped cellular control. The discovery of giant viruses has changed these standard paradigms. Their genetic, proteomic and structural complexities resemble those of cells, prompting a redefinition and reclassification of viruses. In a previous genome-wide analysis of the evolution of structural domains in proteomes, with domains defined at the fold superfamily level, we found the origins of viruses intertwined with those of ancient cells. Here, we extend these data-driven analyses to the study of fold families confirming the co-evolution of viruses and ancient cells and the genetic ability of viruses to foster molecular innovation. The results support our suggestion that viruses arose by genomic reduction from ancient cells and validate a co-evolutionary 'symbiogenic' model of viral origins.

Virology in the era of giant viruses

The ongoing COVID-19 pandemic [1] is a respiratory illness caused by the rapid global transmission of SARS-CoV-2 [2], the seventh coronavirus known to infect humans. This ravaging disease illustrates the planetary consequences of recurrent episodes of zoonotic transmission from animals to human populations. Pandemics shape public perception. Viruses are seen as noxious agents of infection and death. Pandemics also poise philosophers and virologists to wonder about the origins of viruses and their ability to infect all cellular lineages on Earth [3–7]. Disagreements on whether viruses are living or nonliving persist despite more than a hundred years of virological research and recent data-driven breakthroughs in the field of evolutionary genomics [8–11]. In 2003, the discovery of 'giant' viruses [12] revived the debate and challenged epistemological foundations [7, 13]. The size of their genomes rivals that of several parasitic organisms from all three cellular domains (superkingdoms) of life [14]. Their virions, vehicles of transmission that embed the genetic material of the virus with a capsid (protein shell) and a lipid envelope, are cell-like and large enough to be visualized under a light microscope [14]. They can produce thousands of proteins, including proteins involved in hijacking host metabolism [15] and translation [13]. These and other features (some cell-like) were previously never thought to be associated with viruses [4]. Today, dozens of giant viruses have been discovered inhabiting a wide range of environments on Earth [16, 17]. They were hiding in plain sight [18].

The new findings were met with equal criticism and enthusiasm. Some scientists, including us, proposed that giant viruses are 'living' parts of complex cellular cycles and together with other viruses make up an additional supergroup or a 'fourth' domain of life [8, 9, 11, 19–22], whereas others dismissed their living status altogether and attributed their size to massive genetic transfer from cellular genomes [23–25]. Despite controversy, giant viruses have redefined virology in several ways. Building on previous ideas [26], new intelligent definitions of viruses have been proposed (e.g., the virocell and virion-factory concepts discussed

Handling Editor: Marc H. V. Van Regenmortel.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00705-020-04724-1) contains supplementary material, which is available to authorized users.

✉ Gustavo Caetano-Anollés
gca@illinois.edu

[1] Evolutionary Bioinformatics Laboratory, Department of Crop Sciences, University of Illinois at Urbana-Champaign, Urbana, IL, USA

[2] Illinois Informatics Institute, University of Illinois at Urbana-Champaign, Urbana, IL, USA

[3] Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM, USA

[4] Department of Biosciences, COMSATS University Islamabad, Islamabad, Pakistan

below) [4, 27] that highlight the distinction between viruses and virions [7, 28]. In parallel, we and others have integrated the genomes of giant viruses into phylogenetic and phylogenomic studies to investigate their origin, evolution, and co-evolution with their hosts [3, 9, 11, 14, 29–33], some fueled by the appreciation of the importance of protein structure [34, 35].

Here, we summarize our phylogenomic effort, which spans over a decade of investigations [3, 8–10, 29, 36]. Unlike traditional approaches, our strategy focuses on protein structural domains grouped by conserved three-dimensional structural backbones [35]. Protein structures are more resistant to evolutionary change than gene and protein sequences [34]. Their higher evolutionary conservation allows comparison of distantly related and fast-evolving genomes with a higher level of accuracy [35]. In turn, sequence alignments of highly divergent datasets often fail to preserve a sufficient and reliable evolutionary signal for downstream analysis. For example, a recent study showed that it was impossible to reliably align large blocks of RNA-dependent RNA polymerase (RdRp) protein sequences from several families of RNA viruses [37]. Here, we show how protein structures can outperform sequences in such cases by highlighting our past discoveries and updating an important study [9] by adding and analyzing newer genomes and structures.

Hybrid models of virus origins and evolution

One main question concerning viruses is the timing and mechanism of their origin. This problem has continued to capture the imagination of scientists and the public alike but has proven extremely difficult to answer. The major roadblock is the unusually high genetic and morphological diversity seen in hundreds (possibly thousands) of extant viral lineages [38, 39]. No single feature appears conserved across the virosphere, the world of virus diversity. This likely suggests that viruses originated multiple times in evolution, and possibly via more than one mechanism. The lack of unity also suggests that viruses are probably very ancient. Under these assumptions, three classical viewpoints have emerged in the scientific literature in various forms: (i) the *virus-first* hypothesis, which posits that viruses originated from pre-cellular genetic elements, (ii) the *reduction* hypothesis, in which viruses originated from cells via reductive evolution, leading to extraordinary genomic and physical streamlining, and (iii) the *escape* hypothesis, in which cellular genes escaped from cells and transformed into (enveloped) viruses (Fig. 1A). None of these models satisfactorily captures or explains the massive diversity of the virosphere. They all have shortcomings [3, 4]. Consequently, hybrid models that combine elements from the different viewpoints have recently gained popularity [3, 40]. We will discuss two such models (Fig. 1B).

The model of Krupovic et al. [40] combines elements from the *virus-first* and *escape* models, making it chimeric. The authors divide virus genes into two genetic modules, one involved in replication and the other in virion structure and assembly. The replication module originated in an 'ancient world' inhabited by primordial genetic elements that existed prior to cells [40]. This ancestrality explains the massive diversity seen in the replication strategies and genomes of modern-day viral lineages. In turn, the structural genes responsible for virion formation and assembly were gradually captured from cells, once they appeared in evolution [40]. Indeed, the single jelly-roll fold, the most common capsid protein fold in RNA viruses, has structural homologies to several cellular proteins, mainly carbohydrate-binding proteins [30]. While capsids hold a central place in any model of virion origins, capsid-like analogs are rare or absent in cells [9], and some capsid protein folds are unique to viruses [39]. In addition, the rise of genetics prior to cells implies an ancient nucleic acid world (typically an "RNA world"), which brings all its difficulties and is incompatible with significant evidence [41].

In contrast, the 'co-evolutionary' symbiogenic model of Nasir and Caetano-Anollés [3] proposes that viruses reduced from ancient cells with segmented RNA genomes as the world of communal cells was originating on the planet and was giving rise to the ancestors of Archaea, Bacteria, and Eukarya [9, 10]. Phylogenetically, the model thus invokes a fourth sister group arising with the three ancestors of cellular domains. Unlike its other siblings, this primordial group devolved into viruses by selective loss while remaining linked to cellular hosts via cycles of cellular internalization. The key difference with previous models therefore lies in the timing of genome origins and gradual unfolding of replication novelties in cells, not prior to them. In other words, instead of virus-like genetic elements evolving in a pre-cellular world, their genetics co-evolved alongside ancient cells. The symbiogenic model also proposes that RNA viruses originated first in evolution. Later, the retro-transcribing and DNA viruses either (co)evolved directly from the RNA viruses or evolved independently from the viral stem line of descent when RNA-based replication systems were gradually replaced by DNA-based replication counterparts. The model is compatible with one that links the cellular origin of giant viruses to the RNA-to-DNA evolutionary transition [27], prior to the radiation of cellular lineages [7]. The symbiogenic model also considers capsid proteins to have originated in the stem line at the onset of organismal lineage diversification or having been coopted from co-existing emerging cellular lineages.

The symbiogenic model draws inferences from the strong reductive tendencies that have become a hallmark feature of

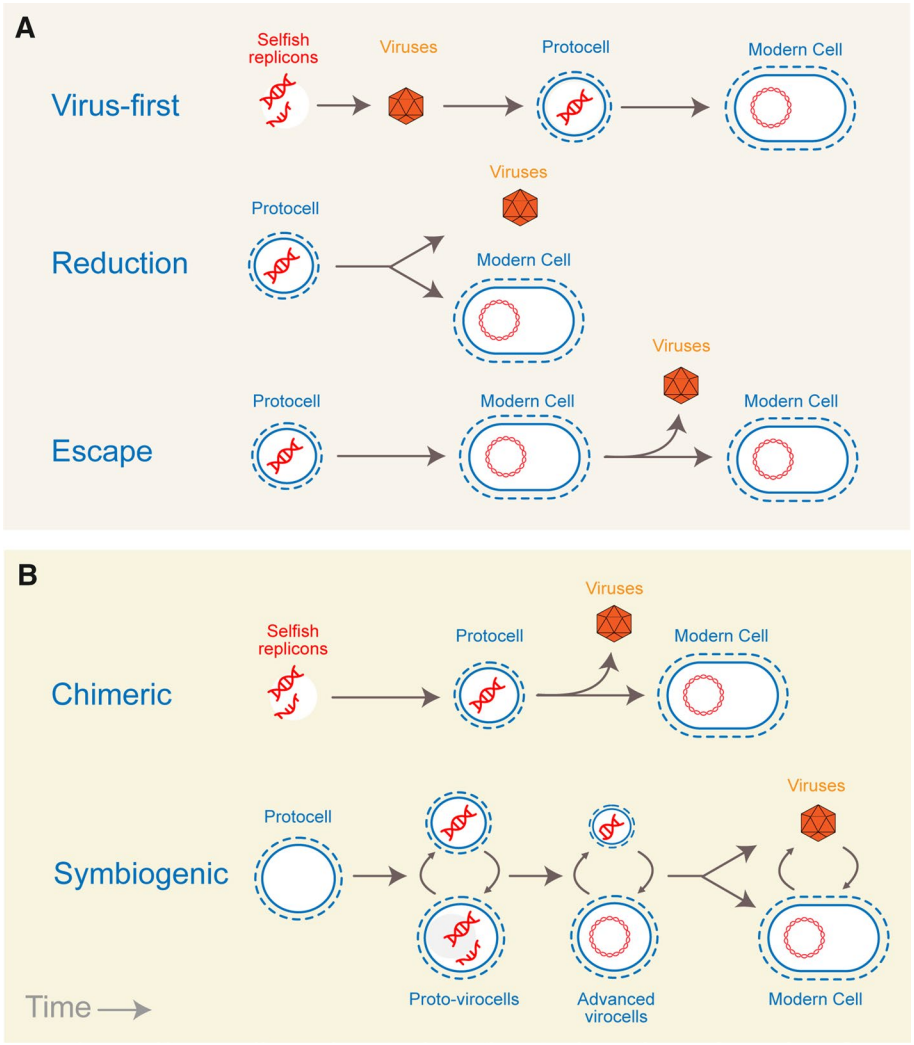


Fig. 1 Models of viral evolution. (A) Classic frameworks of viral evolution associate virus origins to the origin of the virion or its components. The *virus-first* hypothesis postulates ancient origins of viruses that preceded cellular life. The *reduction* hypothesis suggests that the origin of cells preceded that of viruses, with DNA viruses evolving via reduction in their genomes. The *escape* hypothesis deems viruses as selfish genetic elements that escaped control of cellular machinery and 'pickpocketed' genes via HGT. (B) Hybrid models of viral evolution support statements with comparative genomics and phylogenomic information. The hybrid 'chimeric' model put forth by Krupovic and coworkers [40] suggests an ancient origin of virus genomes and late recruitment of cellular proteins for capsid proteins to 'escape' from cells. The 'symbiogenic' model proposed by Nasir and Caetano-Anollés [3] hypothesizes that ancient cells coexisted with ancient virocells and that modern-day viruses evolved by genomic reduction

obligate parasitic and endosymbiotic organisms. It would be difficult to imagine that viruses, which are the ultimate parasites, would prefer any another route [7]. Moreover, the model offers no constraint for ancient viruses to propagate via virion synthesis or other means. The ancient cells simply produced vesicles that transported genetic cargo within the emergent cellular community (facilitating 'vesiduction' [42]). The vesicle-mediated transport is routinely observed in multicellular organisms and, interestingly, shows remarkable resemblances to virus exit pathways [43]. Thus, ancient virions were likely free-floating vesicles that mimicked modern-day virus transmission via virions. More-sophisticated virion structures evolved much later at a time when diversified cellular lineages appeared. The rise of virions is therefore explained as an enhancement of viral spread, very much as social media has enhanced the spread of news in human society. All aspects of this symbiogenic model of viral evolution are strongly supported by the data-driven comparative genomics and phylogenomic analyses we will now describe.

Comparative genomics supports cellular history in virus evolution

The cell-like existence of viruses in the distant past was strongly supported by our comparative genomic surveys, which compared the spread and distribution of fold superfamilies (FSFs) of protein structural domains in thousands of proteomes [9]. FSFs, as defined by the Structural Classification of Proteins (SCOP) gold standard [44], are groupings of one or several fold families (FFs). FFs include domains that have sequence and structural evidence of common origin. FFs are thus orthologous evolutionary units. In turn, FSFs group FFs with conserved structural cores and molecular functions. Thus, domains grouped into an FSF may have little or no sequence identity but strong structural and functional similarities, which indicate evolutionary relatedness. FSFs are thus more conserved than FFs but provide less resolution in exploring relatively recent evolutionary relationships.

We discovered that roughly one-fourth of total FSFs were shared between cells and viruses (the ABEV group) [9]. Several of these FSFs included ancient domains involved in metabolic functions and were components of cell membranes [9]. These findings endorsed the idea of a shared cellular history of viruses. Since the publication of this most recent study, an even greater sampling of proteomes and a focus on FFs rather than FSFs confirmed the overall comparative genomic patterns. Fig. 2 illustrates the distribution of 3,892 FF domains in 8,127 reference-quality proteomes. Two FFs embedded in the RdRp and spike complex of SARS-CoV-2, the betacoronavirus responsible for the COVID-19 pandemic, exemplify the structural entities of the proteomic census. In these studies, 139 archaeal, 1,734 bacterial, 210 eukaryal, and 6,044 viral proteomes were selected from the RefSeq database [45]. Again, the ABEV group included a significant fraction of the FF domain repertoire (979 out of 3,892, 25%). The next-largest group was the ABE group (899 out of 3,892, 23%), which supports a stem line of descent embodying universal 'cellular' ancestors. Note also how the Venn diagram is complete (there are no zeroes) and holds several supergroup-specific FFs in viruses, Archaea, Bacteria, and Eukarya. Remarkably, the virus supergroup had Venn distributions comparable to those of the three cellular domains.

Viruses complete the evolutionary picture

In absence of retrodiction (the use of phylogenetic methods to travel back in time), the Venn group distributions of either FSFs or FFs can help infer how the cellular domains diversified from the last universal common ancestor (LUCA) of life. This is another highly charged and controversial problem that divides evolutionary biologists. All kinds of phylogenetic trees have been published supporting distinct topologies of the 'tree of life'. The canonical view of a three-domain (3D) cellular world endorsed by the school of Carl R. Woese and supported by rRNA sequence analysis roots the tree of life in the branch leading to Bacteria and produces a sister-group relationship between Archaea and Eukarya [46]. The 3D view is backed by several ribosomal proteins that are conserved in Archaea and Eukarya but absent in Bacteria [47], although this may be due in part to incomplete sampling of the bacterial superkingdom, especially the surveys of CPR proteomes (e.g. [48]). The 3D tree was challenged by the discovery and subsequent phylogenetic analysis of the proteins encoded by the Åsgard superphylum of Archaea [49]. The Åsgard encode several eukaryote-specific proteins, including proteins involved in cytoskeleton formation and cell rearrangement, which brought to fame the old 'eocyte' tree hypothesis championed by James Lake and colleagues in the 1980s [50]. The eocyte tree is effectively

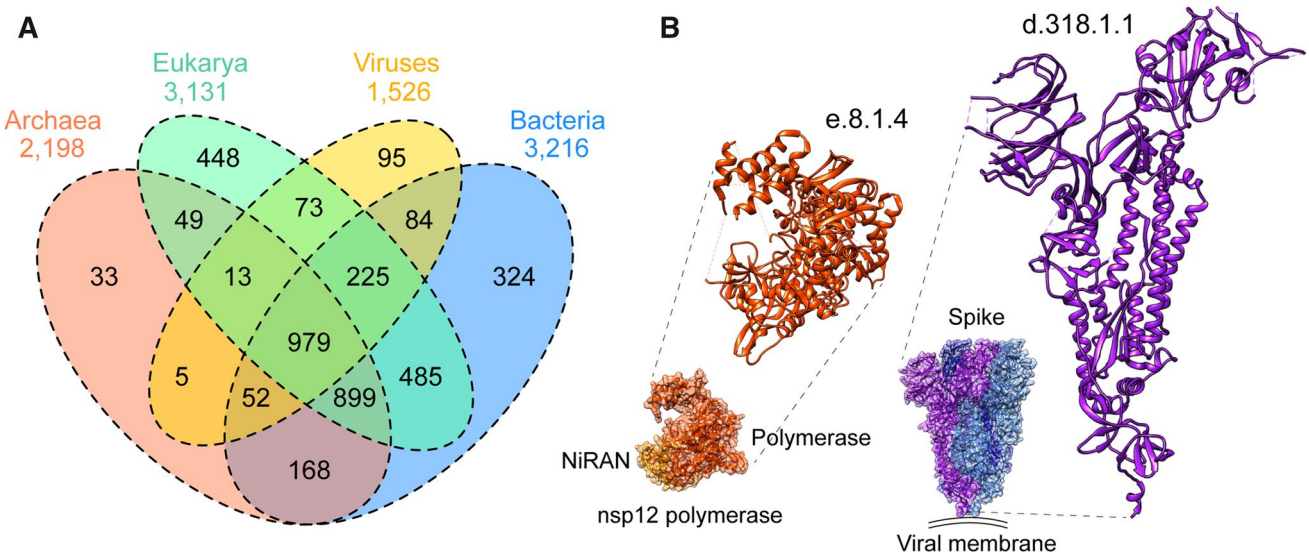


Fig. 2 Fold families (FFs) in proteomes. (A) Venn diagram describing the sharing of 3,892 FFs among 8,127 proteomes from cells and viruses. (B) Two FFs of the structural survey illustrate domain structures present in the human SARS-CoV-2 betacoronavirus. On the left, the enzyme that replicates the genome of the virus inside the lung tissue of the human host (PDB entry 6NUS) has two domains, a nidovirus-unique N-terminal extension with nucleotidyltransferase activity (NiRAN) and a polymerase domain holding the RNA-dependent RNA polymerase FF (e.8.1.4) [63], with a fold of very ancient origin. On the right, a structural model illustrates the three subunits of the spike complex of SARS-CoV-2 (PDB entry 6VSB), which makes the outer surface 'corona' that is responsible for recognizing crucial molecules of the lung and causing the COVID-19 disease [64]. Subunits are colored with different shades of blue and purple. They are made of a virus-specific protein FF (d.318.1.1), the evolutionarily recent SARS receptor-binding domain-like fold

a two-domain (2D) world scenario, which supports two primary domains, Archaea and Bacteria, with the merger of the two leading to the origin of Eukarya. The 2D tree is widely popular but has been challenged by a number of conceptual and methodological difficulties related to phylogenetic reconstruction [51–53]. The 2D tree has multiple forms that vary with regard to the nature of the prokaryotic host cell and the timing of a putative primordial 'phagocytosis' event [54]. Similarly, the 3D tree has two significant variations. In one model, the tree is rooted in the branch leading to Eukarya [55, 56]. In the other, the 'Archaea-first' model, the tree is rooted in branches leading to Archaea [57–59]. Rooted trees built from structural domains in proteomes defined with different classifications and classification levels [59–61] and from Gene Ontology (GO) terms in genomes [62], and inter-proteome and ribosomal protein similarities consistent with vertical transmission [63], support the Archaea-first model. In contrast, the 'Eukarya-first' 3D view was challenged on methodological and conceptual grounds related to phylogenetic rooting and the model of biological change [64, 65]. Despite the constant evolutionary presence of viruses across domains of life, all of these models and most studies exclude viruses. This yields an incomplete picture that fosters controversy.

Without any formal phylogenetic analysis, the Venn group distributions already help us rule out some of the above-mentioned scenarios. For example, a 2D tree implies a very strong genetic affiliation between Archaea and Eukarya or a lesser affiliation between Archaea and Bacteria. Fig. 2 shows that this is clearly not the case. Archaea and Eukarya share only 49 unique FFs (the AE group), while Archaea and Bacteria share 168 FFs (the AB group). In contrast, Bacteria and Eukarya share 485 FFs (the BE group). In absence of horizontal genetic exchange, their stronger genetic affiliation falsifies the expectation of a 2D scenario. A focus on viral proteomes provides an alternative perspective (Fig. 3A). Venn diagrams group viral FFs according to the cellular domain of the virus hosts. For example, *abe* indicates FFs conserved in archaeoviruses, bacterioviruses, and eukaryoviruses. Quite remarkably, the *abe* group shares a biologically and numerically significant core of 112 FFs. This is additional evidence that the virus mode of life likely evolved prior to the diversification of modern cells. Again, there is very little or no overlap between the prokaryotic viruses (Venn group *ab,* 56 FFs) or between the archaeoviruses and eukaryoviruses (*ae*, 6), when compared to FFs shared by bacterioviruses and eukaryoviruses (*be*, 352). If eukaryotes either descended from or are sister to Archaea, we should expect to see higher genetic similarity in either the two domains or their viruses. However, FF evidence is not consistent with such a view. The strongest similarities are observed between Bacteria and Eukarya and their viruses, which supports the Archaea-first 3D tree. We would, however, like to clarify that Venn group numbers can be affected by microbial proteomes that have not been sequenced [52, 66]. So far, however, the patterns of distribution we report here have remained robust over a decade of investigations.

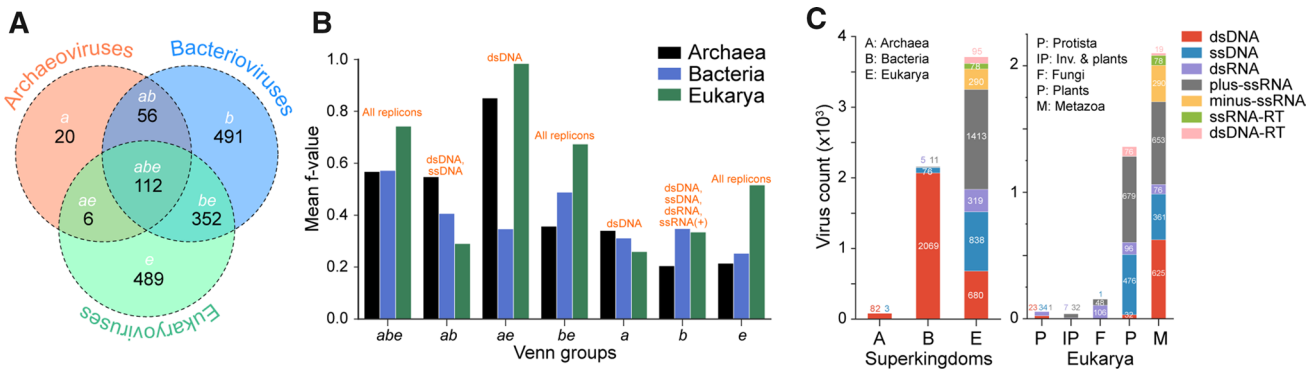
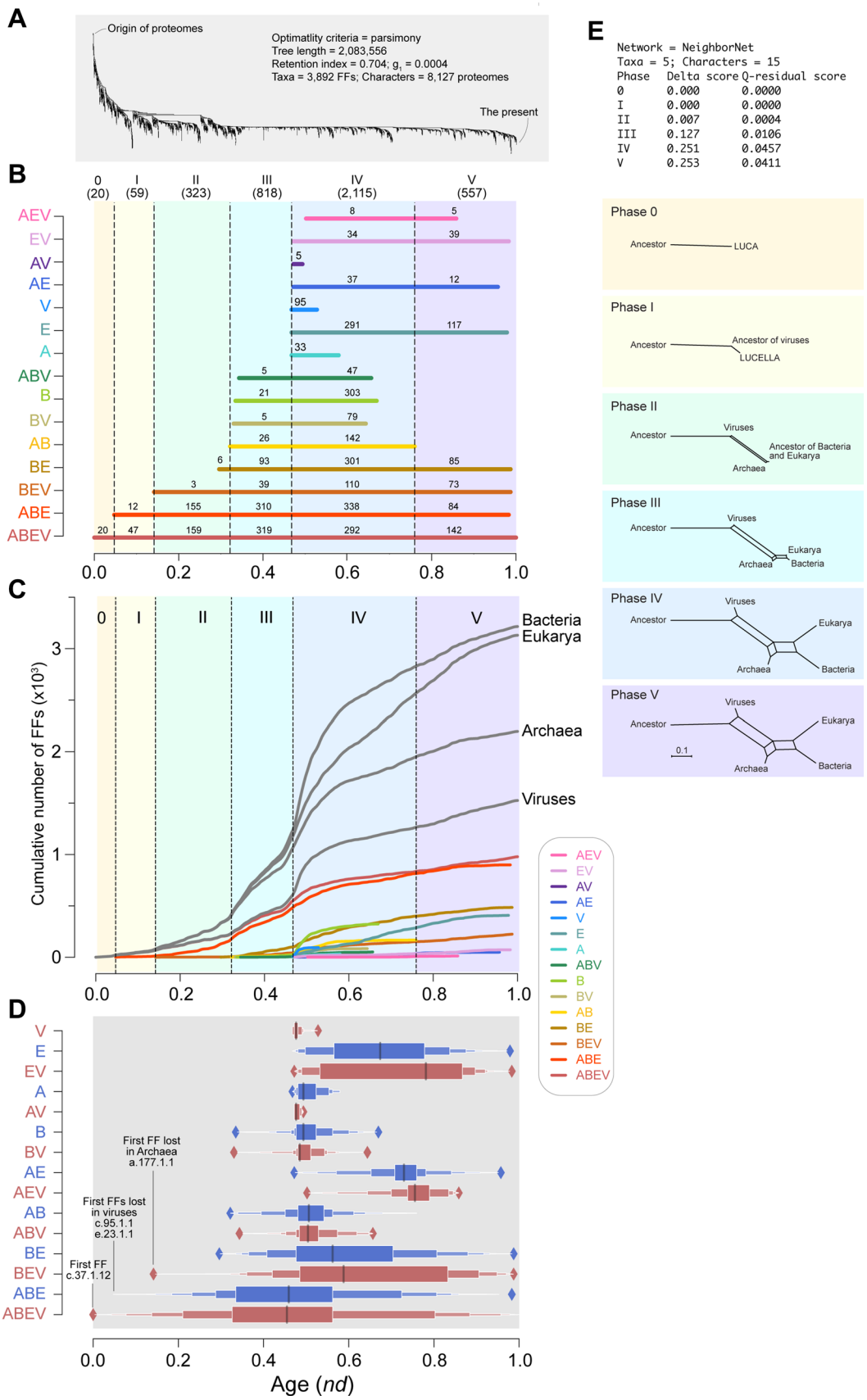


Fig. 3 FFs in the proteomes of viruses infecting different hosts. (A) Venn diagram describing the distribution of 1,526 FFs in archaeoviruses, bacterioviruses, and eukaryoviruses. FFs in the *abe* group do not indicate that these were present in a virus able to infect members of all three superkingdoms. They merely refer to the count of FFs shared among archaeoviruses, bacterioviruses, and eukaryoviruses. (B) Mean *f* values for FFs representing the seven Venn groups defined in panel A in archaeal, bacterial, and eukaryal proteomes. Text above bars indicates how many different viral subgroups possess those FFs. (C) The abundance of each viral replicon type by host superkingdom and major taxonomic groups in Eukarya. Hosts were grouped into Archaea, Bacteria, Protista (animal-like protists), Fungi, plants (all plants, blue-green algae, and diatoms), invertebrates and plants (IP), and Metazoa (vertebrates, invertebrates, and humans). Host information was available for 6,029 out of a total of 6,044 viruses in our dataset, and replicon information was available for 5,959 viruses. Numbers on bars represent the total virus count in each host group


A
Origin of proteomes
Optimatlity criteria = parsimony
Tree length = 2,083,556
Retention index = 0.704; g_1 = 0.0004
Taxa = 3,892 FFs; Characters = 8,127 proteomes
The present
B
0 (20)
I (59)
II (323)
III (818)
IV (2,115)
V (557)
AEV
EV
AV
AE
V
E
A
ABV
B
BV
AB
BE
BEV
ABE
ABEV
0.0
0.2
0.4
0.6
0.8
1.0
C
Cumulative number of FFs (x10³)
Bacteria
Eukarya
Archaea
Viruses
D
First FF lost in Archaea a.177.1.1
First FFs lost in viruses c.95.1.1 e.23.1.1
First FF c.37.1.12
Age (nd)
E
Network = NeighborNet
Taxa = 5; Characters = 15
Phase Delta score Q-residual score
0 0.000 0.0000
I 0.000 0.0000
II 0.007 0.0004
III 0.127 0.0106
IV 0.251 0.0457
V 0.253 0.0411
Phase 0
Ancestor
LUCA
Phase I
Ancestor of viruses
LUCELLA
Phase II
Viruses
Ancestor of Bacteria and Eukarya
Archaea
Phase III
Eukarya
Bacteria
Phase IV
Phase V
0.1

◂**Fig. 4** The evolutionary history of structural domains of the proteomes of cells and viruses. (A) Phylogenomic tree of structural domains reconstructing the evolutionary history of 3,892 FF domains (taxa) in 8,127 proteomes (characters). (B) Six phases in the evolutionary timeline of FFs, with time progressing from the origin of domains (*nd* = 0) to the present (*nd* = 1), showing age (*nd*) ranges for FFs in the Venn groups of Figure 2. Numbers in parentheses indicate the total number of FFs appearing in each phase. Numbers above the horizontal bars represent FFs appearing in a phase for the particular Venn group. (C) Cumulative frequency distribution of FFs along the evolutionary timeline. (D) Boxenplots show the distributions of FFs for each Venn group. Boxenplots are an alternative form of boxplots that plot quantiles beyond the quartile values to better represent large nonparametric datasets with fat-tailed distributions. Blue denotes cellular FFs, while red represents FFs that are either shared with or are exclusive to viruses. (E) Model of diversification of protein domains in proteomes derived from phylogenomic data. The network diagrams implemented with the NeighborNet algorithm in the SplitsTree package confirm the early evolutionary rise of viruses, followed by Archaea and then Bacteria and Eukarya. The last universal common ancestor (LUCA) and the last universal cellular ancestor (LUCELLA) are indicated at their first appearance in the networks. The delta score evaluates the amount of vertical phylogenetic signal present in the data (a score of 0 implies a fully bifurcating tree, and a score of 1 implies a full network)

Protein fold families recapitulate evolutionary history

The Venn diagrams are merely descriptive and can only help approximate the most parsimonious evolutionary scenario. While the molecular census can be biased by unequal samplings and/or uneven performance of bioinformatics programs on different genomic datasets [19, 51], comparative genomic approaches are necessarily limited by the effect of horizontal genetic exchange, which complicates any evolutionary inference derived from extant data that considers only vertical descent. These limitations prompt the use of data-driven phylogenomic methods that are capable of reconstructing the past from a census of structural domains in proteomes [9]. Using the extended genomic dataset described above, we now dissect different evolutionary phases in the history of cells and viruses by a method that traces the origin of the FFs or FSFs of each Venn group [60, 67]. In these experiments, phylogenomic trees are reconstructed that describe the evolution of structural domains. These trees differ from the more traditional trees that describe the evolution of the proteomes of cellular organisms and/or viruses. The leaves (taxa) of the trees are domains instead of proteomes. To calculate the relative age of each domain in a Venn group, a 'node distance' (*nd*) from the base of the tree to each leaf allows us to build evolutionary timelines of appearance of structural domains in proteomes, with *nd* = 0 representing the origin of protein domains and *nd* = 1 representing the present [9, 59, 68]. Ages can be calibrated with a molecular clock of folds that coverts the relative timeline into geological timescales by correlating domain ancestry with domain structures linked to markers of the geological record [69]. Six major evolutionary phases are evident in the timelines of FFs (Fig. 4A and B).

i. *The communal world (nd = 0–0.043).* This initial phase is the birth period of the 20 most ancient FF domains. These FFs are encoded by both viruses and cellular organisms (i.e., they are part of the ABEV group, Fig. 2) and participate in generic functions, involving small-molecule binding. The "ABC transporter ATPase domain-like" FF (c.37.1.12) was the first FF to appear during this time period, which makes it the most ancient FF domain. Overall, this period resembles a communal world inhabited by primordial cells containing a limited number of membrane-associated proteins.
ii. *The rise of viruses (nd = 0.043–0.137).* This phase begins with the loss of two FFs, the "Acetyl-CoA synthetase-like" (e.23.1.1) and "Thiolase-related" (c.95.1.1) protein domains in viruses. These domains were either lost in viruses or, less likely, were gained by the stem line of descent that eventually produced Archaea, Bacteria, and Eukarya. The phase therefore marks the onset of reductive evolution in the cellular ancestors of viruses, which led to a first split of the primordial cellular stem line. The majority of FFs emerging in this phase participate in core metabolic processes.
iii. *The birth of Archaea (nd = 0.137–0.318).* This phase begins with the first loss of a FF in Archaea, the "Sigma2 domain of RNA polymerase sigma factors" (a.177.1.1). This checkpoint therefore marks the split of archaeal ancestors and the beginning of reductive evolution in Archaea. A total of 323 FFs originated in this phase, mostly involved in metabolic functions.
iv. *Diversification of Bacteria (nd = 0.318–0.464).* Bacteria-specific FFs started to appear during this phase, some of which confer pathogenicity to bacteria. This middle period therefore marks the first diversification of an ancient lineage into modern organisms. Thus, while the archaeal lineage splits off earlier, it was the bacterial lineage that diversified first. This important and consistent finding reconciles both the canonical Woesian 3D view and the Archaea-first 3D scenario. A β-propeller FF structure of the ABEV group, the 'YVTN repeat' (b.69.2.3), also appeared in this period (*nd* = 0.347). This FF and a universal late-appearing immunoglobin-like β-sandwich structure (b.1.3.1; *nd* = 0.863) form the surface layer (S-layer) proteins typical of bacterial and archaeal 'protective coats' that wrap up and sieve the cell envelopes with their lattices [70]. Since their FFs are present in 35-48% of cellular organisms but only as single FF domains in

a few dsDNA viral lineages (a total of 20, including four giant pithoviruses) (Table S7), these wrapping structures may represent cellular ancestors of viral capsids. They are probably almost absent in viruses because they have been superseded through reductive evolution by the late appearance of viral capsids.

v. *Diversification of Archaea, Eukarya, and Viruses (nd = 0.464–0.77).* The lineages leading to viruses, Archaea, and Eukarya also diversified with the innovation of group-specific FFs. Remarkably, the gains by the viral supergroup included FFs that confer capsids and virulence to viral proteomes and were part of the virus-specific V group, which appeared early during this phase. Interestingly, the AV, BV, and EV groups soon followed the appearances of A, B, and E groups. These observations therefore suggest that viral pathogenesis started soon after the diversification of the host. A GO [71] enrichment analysis of FFs in AV, BV, and EV confirms that these protein domains enable viruses to infect hosts and ensure their survival (Table S8). The findings were expected; pathogenicity implies the ability to recognize and distinguish between different hosts, which is only possible if the hosts represent diversified lineages.

vi. *Growth in Eukarya (nd = 0.77–1.00).* A large number of FFs appearing in the final phase were specific to eukaryotic lineages, contributing to the late development of eukaryal proteomes as well as acquisition of advanced functionalities such as those involved in regulation, signaling and intracellular processes [60].

The FF timeline is consistent with our previous results [3]. It indicates that the viral supergroup had cellular origins and diversified from a stem line of descent very early in evolution. This primordial viral stem line then continued to evolve via reductive evolution. We have termed these cell-like entities 'proto-virocells' because they combined the modern autonomous replication of modern virus genomes inside a host (virocell) [28] with the ability to spread genetic (and cellular) wealth to other cells through ancient vesicle-like structures. Remarkably, the reductive tendencies were later also observed in the archaeal lineage and are common in modern day endosymbionts and obligate parasites [72]. To illustrate, Fig. 5 reveals an overlap of FF use and reuse values between bacterial parasites and giant viruses. The reduction model of virus evolution is also supported by the distribution of FFs in different virus replicon types of the Baltimore virus classification (Fig. 6). Very few, if any, FFs are conserved between the different virus subgroups. The dsDNA viruses harbor the maximum number of FFs. Out of the 1,526 FFs, ~92% (1,409) were unique to each viral subgroup, while only 8% (117) were shared by more than one subgroup. There were no FFs common to all subgroups. However, the "Reverse transcriptase" (e.8.1.2) FF was the most shared domain, present in five of the seven viral subgroups (Table S4). The data therefore show a very patchy distribution of FFs within the virosphere, which falsifies a monophyletic origin but can be reconciled with reductive evolution of viruses.

Interestingly, the patterns of emergence and loss of FF domains are also remarkably similar to the host preferences of viral lineages. Archaea and Bacteria encode smaller FF repertoires and also exhibit lower diversity in the number

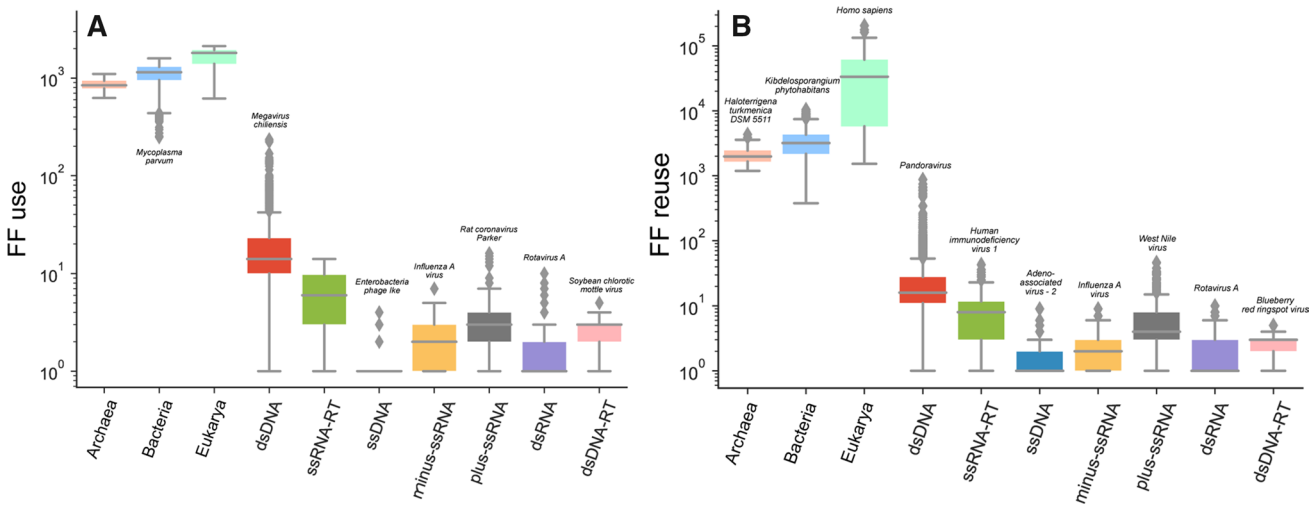


Fig. 5 Use (A) and reuse (B) of FFs in proteomes in each viral subgroup and in the three superkingdoms (N = 8,042 out of total 8,127 proteomes; 85 viruses with no replicon assignment have been excluded in this analysis). *Y*-axis values are provided in the logarithmic scale. Important outliers are labeled

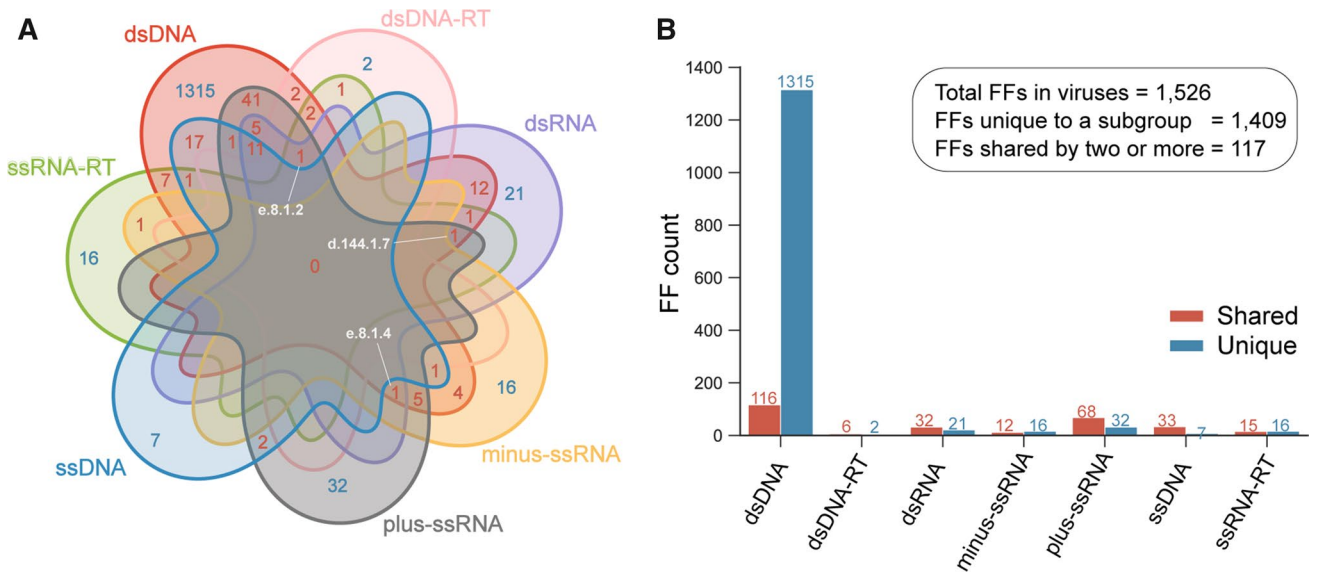


Fig. 6 Sharing of FFs among viruses. (A) A seven-set Venn diagram shows sharing patterns among the viral subgroups. (B) Total number of FFs that were either shared or exclusively occur in each viral subgroup

of virus interactions [73]. Eukaryotes encode the richest FF repertoires and also the maximum diversity in the number and types of virus interactions. For example, eukaryotes, especially animals, are infected by a wide range of RNA and retro-transcribing viruses that are completely absent in Archaea, or are rarely detected or absent in Bacteria, respectively [73, 74]. These biases also manifest in other domains. If the earliest cells contained segmented RNA genomes, as we propose [3, 9], the absence of RNA viruses in Archaea can be explained by the loss of ancient RNA viruses. Perhaps, RNA viruses triggered the early 'archaeal split' [73, 75]. These insights align with interpretations from the Venn group data (Fig. 2).

Virus-specific genes and virus-to-cell gene transfer

The number of virus particles in the universe probably exceeds the number of progeny of cellular organisms [76]. Despite their small genome sizes, collectively, the virus genetic pool likely exceeds that of cells by several-fold. The genes of this genetic pool could make their way to cellular genomes via virus-to-cell gene transfer through virus infection, outnumbering transfers of cellular genes into viral genomes. Such lateral transfers could occur at rates matching Avogadro's number, ~ 10^{24} viral infections per second in the oceans [77]. These concepts have been nicely illustrated before. Virus genes without any detectable cellular homolog or ORFans, virus-specific genes, constitute >90% of the genes in the proposed family "*Pandoraviridae*" [78], whose members possess the largest virus genomes ever discovered [14]. A very recent study identified an amoeba virus (Yaravirus) with a nearly complete 'ORFan genome' [79]. These virus-specific genes are starkly different even among members of the same virus family [14, 32]. These observations suggest these genes partake in highly dynamic processes of change and likely originate continuously in the virus genomes. Our global analyses also revealed similar patterns. On average, ~80% of proteins of prokaryotic viruses and ~60% of eukaryoviral proteins lacked domain assignments (Fig. 7A). These results indicate that ORFans are a feature of viruses infecting all three domains of cellular life. Thus, virus genomes are better characterized by the presence of ORFan genes rather than the relatively smaller subgroup of proteins shared between viruses and their host proteomes (Fig. 3A). Members of this second class of proteins, which accounts for 18-38%, do match structural domains detected in host proteomes. Their origin could be due to shared history, virus-to-cell gene transfer, or cell-to-virus gene transfer. A third class of proteins are the virus-specific proteins (VSFs) that match domains defined in the structure databases, but those domains have not been detected hitherto in any cellular proteome. The proteomic composition of the bulky proteomes belonging to the proposed order "*Megavirales*" also conformed to the same patterns (Fig. 7B). Together, the ORFans and virus-specific proteins overwhelmingly characterize viruses as gene creators rather than 'robbers' or hosts of genetic escapees from cells [78].

The mechanistic details of virus gene creation abilities are nicely illustrated by the virocell concept [28]. In brief, a virus, upon infection, transforms the 'ribocells' of

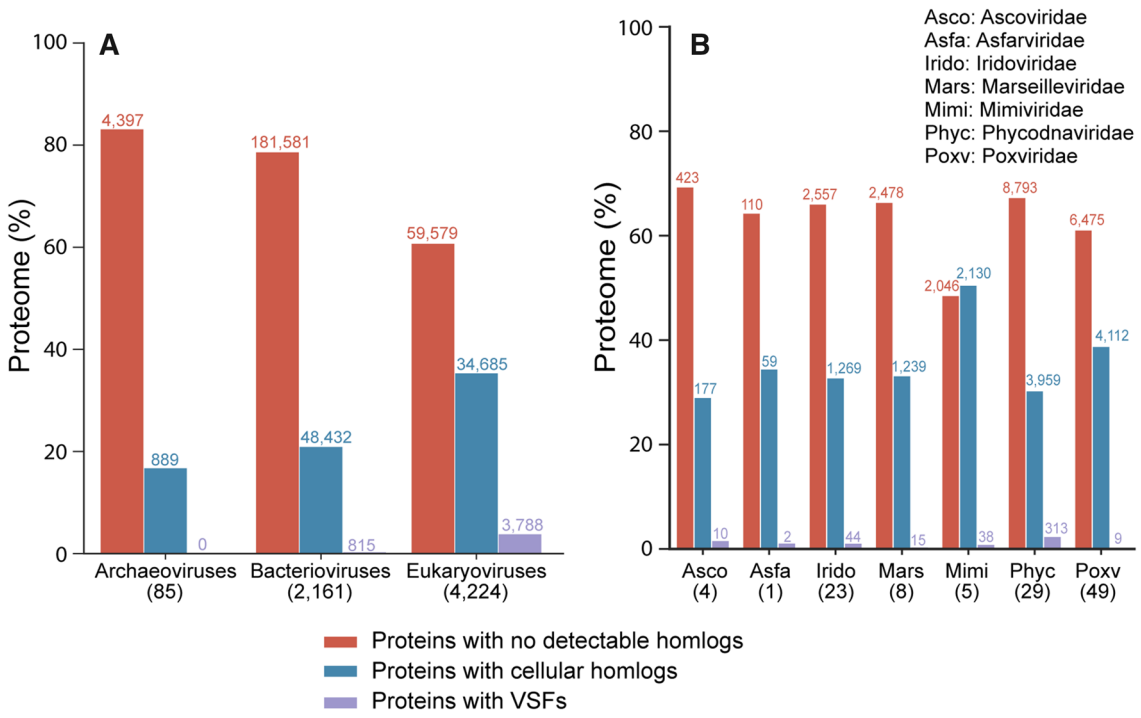


Fig. 7 Proteomic composition of viruses of hosts in the three superkingdoms (A) and in the proposed order "*Megavirales*" (B). Numbers in parentheses represent the total number of proteomes in each category. Numbers above bars show the protein count in each of the three groups of proteins

a ribosome-encoding cellular organism into a virocell that produces virions rather than dividing by binary fission. The virocell is metabolically active and has lost its identity. During this stage, the virus can produce a virus-factory-like system that controls the cellular machinery and can create and replicate genes using the same mechanisms used previously by ribocells [27]. Interestingly, the modern virocell may be the closest visualization of the ancient cellular ancestors of viruses, which we have termed 'proto-virocells'.

The sheer number of VSFs would create ample opportunities for several virus genes to be endogenized and domesticated by their cellular hosts. There are several such examples, such as the abundance of endogenized virus-like elements in mammalian genomes and other proteins that perform useful functions. Endogenous viral elements, specifically mammalian endogenous retroviruses, appear to boost antiviral immunity in their hosts [80]. Syncytins and Gag proteins derived from ancient retrotransposons are involved in placentation in mammals [80]. Gag proteins have also been found to participate in the evolution of the mammalian brain [80]. To formally test this idea, we identified FFs that were shared by XV Venn groups but with sparse occurrence in X proteomes, X being a proteome of Archaea, Bacteria, or Eukarya. We found an additional 64 putative VSFs using this approach. With the exception of five VSFs and three putative VSFs (Table S1), most VSFs were associated with a single viral subgroup. This matches the analysis of FSF domains and the proposal that each viral replicon type has evolved different VSFs for its survival, function and reproduction [9]. The list of new putative VSFs included several proteins responsible for viral core and structural components as well as cellular and nucleic acid binding (Table S5). All four EV FFs from this list (Table S1) represent viral structural proteins with low spread in eukaryotes (Table S5), indicating that these domains may have been acquired by eukaryotes via horizontal gene transfer or (less likely) could be a result of inaccurate hidden Markov model (HMM) assignment. Similarly, several of the BV domains identified as putative VSFs constitute viral proteins and encompass 65.63% of the putative VSFs (42 out of 64), indicating that most BV domains are derived from viruses. This observation is consistent with previous findings [9, 10], given that bacterioviruses are known to facilitate genetic transfers among organisms and thus drive evolution in bacteria.

The virus-to-cell direction of gene transfer is further supported by the comparison of two different FF distributions: (i) FFs that are only shared among cells and (ii) FFs that are shared between cells and viruses. We find that the latter group was more widely distributed in proteomes when spread was measured with an *f* index (Fig. 8), a finding that

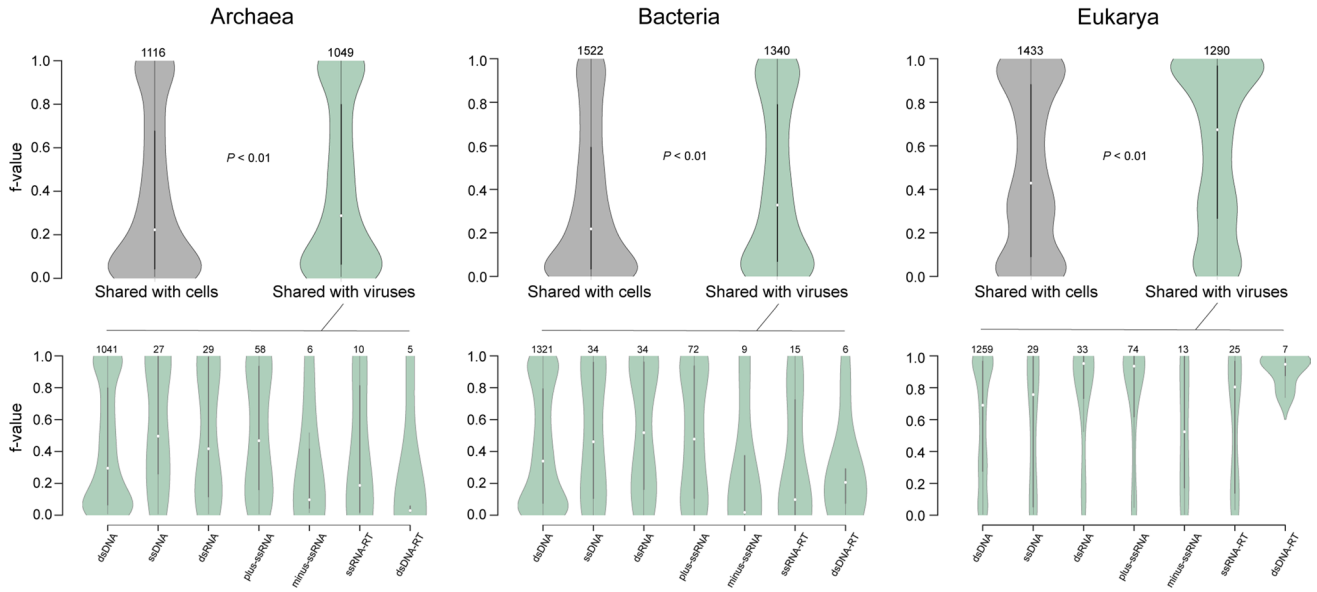


Fig. 8 Spread of viral FFs in cellular proteomes to study the direction of gene transfer (either virus-to-cell or cell-to-virus). Fractional (*f*) values measure the spread of each FF with values from 0 (absence in all proteomes analyzed) to 1 (presence in all proteomes). FFs belonging to A, B, and E Venn groups are excluded in this analysis, as they reflect superkingdom-specific gains. (A) Comparison of spread (*f* value) of FFs shared or not shared with viruses in archaeal, bacterial, and eukaryal proteomes (Wilcoxon rank sum test, two-tailed, $P < 0.01$). (B) Comparison of spread (*f* value) of FFs shared with each viral subgroup in archaeal, bacterial, and eukaryal proteomes. Numbers at the top represent the total number of FFs included in each comparison

confirms a similar analysis with domains defined at the FSF level [9]. The index ranges from 0 (absence) to 1 (presence in all genomes). The median *f* value for FFs in Archaea shared only with cells was 0.223, compared to the *f* value of 0.288 (29.1% increase) for FFs shared with viruses. Likewise, the sharing of FFs with viruses increased *f* values by 50.5% (from 0.218 to 0.328) in Bacteria and up to 57.1% (from 0.429 to 0.674) in Eukarya. Irrespective of the magnitude of the median increase among superkingdoms, FFs shared with viruses were always significantly more widespread than those only shared with cells. The presence of these viral FFs in an assortment of cellular proteomes, from microorganisms to large eukaryotes, suggests that viruses are very ancient and already transferred genes to the last common ancestor of each superkingdom [9]. Inspection of archaeal FFs shared with viral replicons showed an enrichment of all viral subgroups, including the presence of RNA viruses. RNA viruses are unable to successfully infect Archaea. It is therefore likely that RNA viruses infecting members of different superkingdoms exchanged FFs that were conserved in evolution from ancient cells [9]. Similarly, FFs shared with eukaryotes by each viral subgroup were greatly widespread, as revealed by high median *f* values (Fig. 8B), which is consistent with eukaryotic proteomes hosting a large number of viruses from each subgroup (Fig. 3C).

The negligible presence of capsid/coat proteins in cells uncovered by our analysis supports the role of virus-to-cell horizontal gene transfer events that mediate molecular innovation. Capsid/coat-related proteins are hallmark virus proteins that point to genetic innovation in viruses [29]. The cellular distribution of 37 FFs related to capsid/coat FSFs (identified using data from Nasir and Caetano-Anollés [9]) attests to this innovation. Only the "Major capsid protein gp5" FF (d.183.1.1) was present in roughly 29% of cellular proteomes (Table S3). Most of the capsid/coat FFs (27 out of 37) were completely absent or had a negligible presence in cellular proteomes. Their functions and design were developed in viruses and are still in the process of being recruited by cellular makeup [39].

Early origins of RNA viruses

RNA viruses are thought to have been the first to evolve among viruses, with origins rooted in ancient cells [7, 27, 81]. Our phylogenomic reconstructions using FSFs [9] and FFs, which we report here, support this hypothesis. Two different approaches give further credence to the ancient origin of viruses. First, surveying the age distribution of FFs in the ancient ABEV group allowed us to establish the relative age of each viral subgroup (Fig. S2A). With the exception of dsDNA-RT viruses, the *nd* medians were fairly low and distributions had long right, which that are expected when the effects of horizontal gene transfer are recent. The most ancient ABEV FFs belonged to plus-ssRNA and ssDNA, suggesting that they emerged prior to the very common

dsDNA viruses. Second, evolutionary principal coordinate (evoPCO) analysis, an multidimensional scaling strategy that determines the relative evolutionary ancestry of proteomes [9], revealed the ancient origins of RNA viruses. Domains become part of each proteome at different evolutionary timepoints. Their relative ages can measure the evolutionary makeup and give a general indication of the age of a proteome. The evoPCO methodology uses a data matrix of ages to generate a temporal view of proteome evolution. This view can become more powerful than the standard 'tree of life' visualization because it minimizes problems of independence of phylogenetic characters known to plague sequence analysis [19, 64]. Fig. 9 shows an evoPCO plot of FF data reflecting the 8,127 reference-quality proteomes of cellular organisms and viruses that we recently analyzed. The first three coordinates, which account for ~72% of total variance, revealed four distinct temporal clouds of proteomes dissecting viruses and each of the three superkingdoms (Fig. 9). The evolutionary direction of the clouds of data points uncovered a temporal flow from viruses to cellular organisms that unfolded along the PCO1 axis. The cloud of RNA viral proteomes at the back of the evoPCO plot was temporally the most basal. Two RNA (plus-ssRNA) viruses, Lasius neglectus virus 1 and Lygus lineolaris virus 1, belonging to the order *Picornavirales*, occupied the most basal position in the plot. The group of giant viruses was clearly dissected from the main viral cloud and was temporally closer to cellular proteomes, confirming results that suggest their late appearance in viral evolution [13, 31] and supporting the coexistence of ancient viruses with cells [8, 27]. All results reinforce the notion of an ancient origin of RNA viruses.

Conclusions

The 21st century is the century of data science. The continuous flood of data and information that flows from multiple data streams can vastly improve life on our planet, from predictions that impact weather and climate change to understanding the structure and behavior of artificial and natural biological systems that generate, process, store, use and communicate information. Evolutionary biologists are also blessed by the depth and breadth of expanding biological databases, from genome, metagenome, and microbiome repositories to gold standards of protein classification. These datasets provide unique opportunities to address age-old and puzzling mysteries. Here, we review our data-mining efforts to uncover the origins and evolution of viruses. Comparative genomic and phylogenomic approaches that take advantage of the three-dimensional atomic structure of proteins support the origin of viruses from cells predating the ancestors of cellular domains. These ancient cells gradually devolved into modern-day viruses via reductive evolution, matching well-known processes that operate in Archaea and endosymbiotic organisms. Contrary to popular belief, the genomes of viruses harbor an abundance of well-characterized virus-specific genes and encode numerous protein structures that carry significantly deep evolutionary information about a pervasive virus-to-cell genetic transfer of cellular innovations. Thus, viruses should be considered drivers of cellular evolution rather than minimalistic genetic parasites. They have played and continue to play major roles in the evolution of the living world. Following the breakthrough discovery of giant viruses, we should expect that the discovery of other

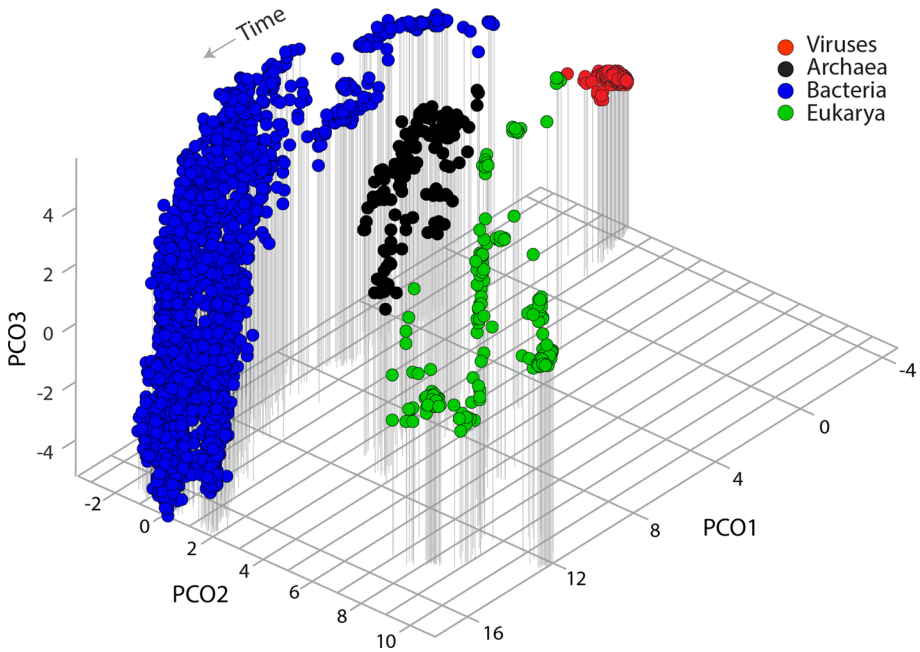


Fig. 9 Evolutionary principal coordinate (evoPCO) analysis describing the evolution of the proteomes of cellular organisms and viruses. The 3-dimensional scatter plot describes the temporal relationships of 139 archaeal, 1,740 bacterial, 210 eukaryal and 6,044 viral proteomes, making up a total of 8,127 reference-quality proteomes and involving the ages of 3,892 FFs. The multidimensional scaling analysis highlights in its first three most significant axes a temporal flow from viruses to cellular organisms

viruses inhabiting diverse hosts and habitats will help revise the way we think about viruses and the roles they play on our planet.

Acknowledgements The phylogenomic analyses presented in this review were conducted by FM as part of her doctoral thesis program at the Illinois Informatics Institute of the University of Illinois, Urbana-Champaign. Farzana Gul, a graduate student from the Department of Biosciences, COMSATS University Islamabad, Pakistan, helped with the initial draft of the paper. This work was supported by grants from the National Science Foundation (OISE-1132791) and the National Institute of Food and Agriculture of the United States Department of Agriculture (ILLU-802-909 and ILLU-483-625) to GCA. AN is supported by the U.S. Department of Energy LDRD program at Los Alamos National Laboratory (20180751PRD3). This work made use of the Illinois Campus Cluster and several Blue Waters supercomputing allocations awarded to GCA, computing resources that are operated by the Illinois Campus Cluster Program (ICCP) and the National Center for Supercomputing Applications (NCSA), respectively.

References

1. Shereen MA, Khan S, Kazmi A et al (2020) COVID-19 infection: origin, transmission, and characteristics of human coronaviruses. J Adv Res 24:91–98
2. Andersen KG, Rambaut A, Lipkin WI et al (2020) The proximal origin of SARS-CoV-2. Nat. Med. 26:450–452
3. Nasir A, Kim KM, Caetano-Anollés G (2012) Viral evolution: primordial cellular origins and late adaptation to parasitism. Mob Genet Elements 2:247–252. https://doi.org/10.4161/mge.22797
4. Forterre P (2016) To be or not to be alive: how recent discoveries challenge the traditional definitions of viruses and life. Stud Hist Philos Sci Part C 59:100–108. https://doi.org/10.1016/j.shpsc.2016.02.013
5. Dupré J, Guttinger S (2016) Viruses as living processes. Stud Hist Philos Sci Part C Stud Hist Philos Biol Biomed Sci 59:109–116. https://doi.org/10.1016/j.shpsc.2016.02.010
6. Koonin EV, Starokadomskyy P (2016) Are viruses alive? The replicator paradigm sheds decisive light on an old but misguided question. Stud Hist Philos Sci Part C Stud Hist Philos Biol Biomed Sci 59:125–134. https://doi.org/10.1016/j.shpsc.2016.02.016
7. Claverie JM, Abergel C (2016) Giant viruses: the difficult breaking of multiple epistemological barriers. Stud Hist Philos Sci Part C Stud Hist Philos Biol Biomed Sci 59:89–99. https://doi.org/10.1016/j.shpsc.2016.02.015
8. Nasir A, Kim K, Caetano-Anolles G (2012) Giant viruses coexisted with the cellular ancestors and represent a distinct supergroup along with superkingdoms Archaea, Bacteria and Eukarya. BMC Evol Biol 12:156. https://doi.org/10.1186/1471-2148-12-156
9. Nasir A, Caetano-Anollés G (2015) A phylogenomic data-driven exploration of viral origins and evolution. Sci Adv 1:e1500527. https://doi.org/10.1126/sciadv.1500527
10. Nasir A, Sun F-J, Kim KM, Caetano-Anollés G (2015) Untangling the origin of viruses and their impact on cellular evolution. Ann N Y Acad Sci 1341:61–74. https://doi.org/10.1111/nyas.12735
11. Colson P, Levasseur A, La Scola B et al (2018) Ancestrality and mosaicism of giant viruses supporting the definition of the fourth TRUC of microbes. Front Microbiol 9:2668. https://doi.org/10.3389/fmicb.2018.02668
12. La Scola B, Audic S, Robert C et al (2003) A giant virus in amoebae. Science 299:2033. https://doi.org/10.1126/science.1081867
13. Colson P, La Scola B, Levasseur A et al (2017) Mimivirus: leading the way in the discovery of giant viruses of amoebae. Nat Rev Microbiol 15:243–254
14. Philippe N, Legendre M, Doutre G et al (2013) Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. Science 341:281–286. https://doi.org/10.1126/science.1239181
15. Maynard ND, Gutschow MV, Birch EW, Covert MW (2010) The virus as metabolic engineer. Biotechnol J 5:686–694
16. Schulz F, Alteio L, Goudeau D et al (2018) Hidden diversity of soil giant viruses. Nat Commun 9:4881. https://doi.org/10.1038/s41467-018-07335-2
17. dos Andrade ACSP, Arantes TS, Rodrigues RAL et al (2018) Ubiquitous giants: a plethora of giant viruses found in Brazil and Antarctica. Virol J 15:22. https://doi.org/10.1186/s12985-018-0930-x
18. Brandes N, Linial M (2019) Giant viruses-big surprises. Viruses 11:404
19. Nasir A, Kim KM, Caetano-Anollés G (2017) Phylogenetic tracings of proteome size support the gradual accretion of protein structural domains and the early origin of viruses from primordial cells. Front Microbiol 8:1178. https://doi.org/10.3389/fmicb.2017.01178
20. Colson P, Gimenez G, Boyer M et al (2011) The giant Cafeteria roenbergensis virus that infects a widespread marine phagocytic protist Is a new member of the fourth domain of life. PLoS One 6:e18935. https://doi.org/10.1371/journal.pone.0018935
21. Legendre M, Arslan D, Abergel C, Claverie J-M (2012) Genomics of Megavirus and the elusive fourth domain of Life. Commun Integr Biol 5:102–106. https://doi.org/10.4161/cib.18624
22. Boyer M, Madoui M-A, Gimenez G et al (2010) Phylogenetic and phyletic studies of informational genes in genomes highlight existence of a 4th domain of life including giant viruses. PLoS One 5:e15530. https://doi.org/10.1371/journal.pone.0015530
23. Bäckström D, Yutin N, Jørgensen SL et al (2019) Virus genomes from deep sea sediments expand the ocean megavirome and support independent origins of viral gigantism. MBio 10:e02497-18. https://doi.org/10.1128/mBio.02497-18
24. Moreira D, Brochier-Armanet C (2008) Giant viruses, giant chimeras: The multiple evolutionary histories of Mimivirus genes. BMC Evol Biol 8:12. https://doi.org/10.1186/1471-2148-8-12
25. Moreira D, López-García P (2009) Ten reasons to exclude viruses from the tree of life. Nat Rev Microbiol 7:306–311. https://doi.org/10.1038/nrmicro2108
26. Bândea CI (1983) A new theory on the origin and the nature of viruses. J Theor Biol 105:591–602. https://doi.org/10.1016/0022-5193(83)90221-7
27. Claverie JM (2006) Viruses take center stage in cellular evolution. Genome Biol 7:110. https://doi.org/10.1186/gb-2006-7-6-110
28. Forterre P, Krupovic M (2012) The origin of virions and virocells: The escape hypothesis revisited. Viruses: essential agents of life. Springer, The Netherlands, pp 43–60
29. Malik SS, Azem-e-Zahra S, Kim KM et al (2017) Do viruses exchange genes across superkingdoms of life? Front Microbiol 8:2110. https://doi.org/10.3389/fmicb.2017.02110
30. Krupovic M, Koonin EV (2017) Multiple origins of viral capsid proteins from cellular ancestors. Proc Natl Acad Sci USA 114:E2401–E2410. https://doi.org/10.1073/pnas.1621061114
31. Koonin EV, Dolja VV (2014) Virus world as an evolutionary network of viruses and capsidless selfish elements. Microbiol Mol Biol Rev 78:278–303. https://doi.org/10.1128/mmbr.00049-13
32. Legendre M, Alempic J-M, Philippe N et al (2019) Pandoravirus celtis illustrates the microevolution processes at work in the giant

pandoraviridae genomes. Front Microbiol 10:430. https://doi.org/10.3389/fmicb.2019.00430
33. Guglielmini J, Woo AC, Krupovic M et al (2019) Diversification of giant and large eukaryotic dsDNA viruses predated the origin of modern eukaryotes. Proc Natl Acad Sci USA 116:19585–19592. https://doi.org/10.1073/pnas.1912006116
34. Abrescia NGA, Bamford DH, Grimes JM, Stuart DI (2012) Structure unifies the viral universe. Annu Rev Biochem 81:795–822. https://doi.org/10.1146/annurev-biochem-060910-095130
35. Caetano-Anollés G, Nasir A (2012) Benefits of using molecular structure and abundance in phylogenomic analysis. Front Genet 3:172. https://doi.org/10.3389/fgene.2012.00172
36. Nasir A, Kim KM, Caetano-Anollés G (2017) Long-term evolution of viruses: a Janus-faced balance. BioEssays 39:1700026. https://doi.org/10.1002/bies.201700026
37. Holmes EC, Duchêne S (2019) Can sequence phylogenies safely infer the origin of the global virome? MBio 10:e00289-19
38. Prangishvili D, Bamford DH, Forterre P et al (2017) The enigmatic archaeal virosphere. Nat Rev Microbiol 15:724–739
39. Nasir A, Caetano-Anollés G (2017) Identification of capsid/coat related protein folds and their utility for virus classification. Front Microbiol 8:380. https://doi.org/10.3389/fmicb.2017.00380
40. Krupovic M, Dolja VV, Koonin EV (2019) Origin of viruses: primordial replicators recruiting capsids from hosts. Nat Rev Microbiol 17:449–458. https://doi.org/10.1038/s41579-019-0205-6
41. Caetano-Anollés G, Seufferheld MJ (2013) The coevolutionary roots of biochemistry and cellular organization challenge the RNA world paradigm. J Mol Microbiol Biotechnol 23:152–177. https://doi.org/10.1159/000346551
42. Soler N, Forterre P (2020) Vesiduction: the fourth way of HGT. Environ Microbiol 1462–2920:15056. https://doi.org/10.1111/1462-2920.15056
43. Meckes DG, Raab-Traub N (2011) Microvesicles and viral infection. J Virol 85:12844–12854. https://doi.org/10.1128/jvi.05853-11
44. Lo Conte L, Ailey B, Hubbard TJ et al (2000) SCOP: a structural classification of proteins database. Nucleic Acids Res 28:257–259
45. O'Leary NA, Wright MW, Brister JR et al (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44:D733–D745. https://doi.org/10.1093/nar/gkv1189
46. Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci U S A 87:4576–4579. https://doi.org/10.1073/pnas.87.12.4576
47. Spang A, Saw JH, Jørgensen SL et al (2015) Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature 521:173–179. https://doi.org/10.1038/nature14447
48. Bokhari RH, Amirjan N, Jeong H et al (2020) Bacterial origin and reductive evolution of the CPR group. Genome Biol Evol 12:103–121
49. Zaremba-Niedzwiedzka K, Caceres EF, Saw JH et al (2017) Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature 541:353–358. https://doi.org/10.1038/nature21031
50. Lake JA, Henderson E, Oakes M, Clark MW (1984) Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes. Proc Natl Acad Sci USA 81:3786–3790. https://doi.org/10.1073/pnas.81.12.3786
51. Nasir A, Kim KM, Da Cunha V, Caetano-Anollés G (2016) Arguments reinforcing the three-domain view of diversified cellular life. Archaea 2016:1851865
52. Nasir A, Kim KM, Caetano-Anollés G (2015) Lokiarchaeota: Eukaryote-like missing links from microbial dark matter? Trends Microbiol 23:448–450
53. Da Cunha V, Gaia M, Nasir A, Forterre P (2018) Asgard archaea do not close the debate about the universal tree of life topology. PLoS Genet 14:e1007215. https://doi.org/10.1371/journal.pgen.1007215
54. Kurland CG, Collins LJ, Penny D (2006) Genomics and the irreducible nature of eukaryote cells. Science 312:1011–1014
55. Jun SR, Sims GE, Wu GA, Kim SH (2010) Whole-proteome phylogeny of prokaryotes by feature frequency profiles: An alignment-free method with optimal feature resolution. Proc Natl Acad Sci USA 107:133–138. https://doi.org/10.1073/pnas.0913033107
56. Harish A, Kurland CG (2017) Akaryotes and Eukaryotes are independent descendants of a universal common ancestor. Biochimie 138:168–183. https://doi.org/10.1016/j.biochi.2017.04.013
57. Wong JTF, Chen J, Mat WK et al (2007) Polyphasic evidence delineating the root of life and roots of biological domains. Gene 403:39–52. https://doi.org/10.1016/j.gene.2007.07.032
58. Di Giulio M (2007) The tree of life might be rooted in the branch leading to Nanoarchaeota. Gene 401:108–113. https://doi.org/10.1016/j.gene.2007.07.004
59. Wang M, Yafremava LS, Caetano-Anollés D et al (2007) Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world. Genome Res 17:1572–1585. https://doi.org/10.1101/gr.6454307
60. Kim KM, Caetano-Anollés G (2012) The evolutionary history of protein fold families and proteomes confirms that the archaeal ancestor is more ancient than the ancestors of other superkingdoms. BMC Evol Biol 12:13. https://doi.org/10.1186/1471-2148-12-13
61. Bukhari SA, Caetano-Anollés G (2013) Origin and evolution of protein fold designs inferred from phylogenomic analysis of CATH domain structures in proteomes. PLoS Comput Biol 9:e1003009. https://doi.org/10.1371/journal.pcbi.1003009
62. Kim KM, Nasir A, Hwang K, Caetano-Anollés G (2014) A tree of cellular life inferred from a genomic census of molecular functions. J Mol Evol 79:240–262. https://doi.org/10.1007/s00239-014-9637-9
63. Long X, Xue H, Wong JT-F (2019) Descent of Bacteria and Eukarya from an archaeal root of life. bioRxiv 2019:745372. https://doi.org/10.1101/745372
64. Caetano-Anollés G, Nasir A, Kim KM, Caetano-Anollés D (2018) Rooting phylogenies and the tree of life while minimizing ad hoc and auxiliary assumptions. Evol Bioinforma 14:1176934318805101
65. Caetano-Anollés D, Nasir A, Kim KM, Caetano-Anollés G (2019) Testing empirical support for evolutionary models that root the tree of life. J Mol Evol 87:131–142. https://doi.org/10.1007/s00239-019-09891-7
66. Bernard G, Pathmanathan JS, Lannes R et al (2018) Microbial dark matter investigations: how microbial studies transform biological knowledge and empirically sketch a logic of scientific discovery. Genome Biol Evol 10:707–715
67. Staley JT, Caetano-Anollés G (2018) Archaea-first and the co-evolutionary diversification of domains of life. BioEssays 40:1800036
68. Kim KM, Qin T, Jiang Y-Y et al (2012) Protein domain structure uncovers the origin of aerobic metabolism and the rise of planetary oxygen. Structure 20:67–76. https://doi.org/10.1016/j.str.2011.11.003
69. Wang M, Jiang Y-Y, Kim KM et al (2011) A universal molecular clock of protein folds and its power in tracing the early history of aerobic metabolism and planet oxygenation. Mol Biol Evol 28:567–582. https://doi.org/10.1093/molbev/msq232
70. Rodrigues-Oliveira T, Belmok A, Vasconcellos D et al (2017) Archaeal S-layers: overview and current state of the art. Front Microbiol 8:2597. https://doi.org/10.3389/fmicb.2017.02597

71. Zeng C, Zhan W, Deng L (2018) SDADB: A functional annotation database of protein structural domains. Database 2018:bay064. https://doi.org/10.1093/database/bay064
72. Nasir A, Naeem A, Khan MJ et al (2011) Annotation of protein domains reveals remarkable conservation in the functional make up of proteomes across superkingdoms. Genes (Basel) 2:869–911. https://doi.org/10.3390/genes2040869
73. Nasir A, Forterre P, Kim KM, Caetano-Anollés G (2014) The distribution and impact of viral lineages in domains of life. Front Microbiol 5:194. https://doi.org/10.3389/fmicb.2014.00194
74. Koonin EV, Dolja VV, Krupovic M (2015) Origins and evolution of viruses of eukaryotes: the ultimate modularity. Virology 479–480:2–25
75. Forterre P (2013) The common ancestor of archaea and eukarya was not an archaeon. Archaea 2013:372396
76. Suttle CA (2013) Viruses: unlocking the greatest biodiversity on Earth. Genome 56:542–544. https://doi.org/10.1139/gen-2013-0152
77. Suttle CA (2007) Marine viruses—major players in the global ecosystem. Nat Rev Microbiol 5:801–812. https://doi.org/10.1038/nrmicro1750
78. Legendre M, Fabre E, Poirot O et al (2018) Diversity and evolution of the emerging Pandoraviridae family. Nat Commun 9:2285. https://doi.org/10.1038/s41467-018-04698-4
79. Boratto PVM, Oliveira GP, Machado TB et al (2020) Yaravirus: A novel 80-nm virus infecting Acanthamoeba castellanii. Proc Natl Acad Sci USA 117(28):16579–16586. https://doi.org/10.1073/pnas.2001637117
80. Frank JA, Feschotte C (2017) Co-option of endogenous viral sequences for host cell function. Curr Opin Virol 25:81–89
81. Forterre P (2005) The two ages of the RNA world, and the transition to the DNA world: a story of viruses and cells. Biochimie 2005:793–803