VIROLOGY

# A phylogenomic data-driven exploration of viral origins and evolution


10.1126/sciadv.1500527

**Arshan Nasir* and Gustavo Caetano-Anollés†**

**The origin of viruses remains mysterious because of their diverse and patchy molecular and functional makeup. Although numerous hypotheses have attempted to explain viral origins, none is backed by substantive data. We take full advantage of the wealth of available protein structural and functional data to explore the evolution of the proteomic makeup of thousands of cells and viruses. Despite the extremely reduced nature of viral proteomes, we established an ancient origin of the "viral supergroup" and the existence of widespread episodes of horizontal transfer of genetic information. Viruses harboring different replicon types and infecting distantly related hosts shared many metabolic and informational protein structural domains of ancient origin that were also widespread in cellular proteomes. Phylogenomic analysis uncovered a universal tree of life and revealed that modern viruses reduced from multiple ancient cells that harbored segmented RNA genomes and coexisted with the ancestors of modern cells. The model for the origin and evolution of viruses and cells is backed by strong genomic and structural evidence and can be reconciled with existing models of viral evolution if one considers viruses to have originated from ancient cells and not from modern counterparts.**

## INTRODUCTION

The origin and evolution of viruses remain difficult to explain. This stems from numerous philosophical and technical issues, including an experimental focus on single genes and consequent failure to take into account the complete makeup of viral proteomes. Perhaps the most challenging problems plaguing the deep evolutionary studies of viruses are the fast evolution and high mutation rates of most viral genes [especially RNA viruses (*1*)]. This makes it difficult to unify viral families especially using sequence-based phylogenetic analysis. For example, the latest (2014) report of the International Committee on the Taxonomy of Viruses (ICTV) recognizes 7 orders, 104 families, 23 subfamilies, 505 genera, and 3186 viral species (*2*). Under this classification, viral families belonging to the same order have likely diverged from a common ancestral virus. However, only 26 viral families have been assigned to an order, and the evolutionary relationships of most of them remain unclear. The number of viral families without an order is expected to continuously increase, especially with the discovery of novel viruses from atypical environments, and because genes of many viral families do not exhibit significant sequence similarities (*3*). In fact, homologous proteins often diverge beyond recognition at sequence level after a relatively long evolutionary time has passed (*4*). In such cases, traditional sequence-based homology searches [for example, Basic Local Alignment Search Tool (BLAST)] and alignment software perform very poorly. However, the three-dimensional (3D) packing of amino acid side chains in cores of protein structural domains retains its arrangement over long evolutionary periods (*5*). Because homologous proteins often maintain 3D folds and biochemical properties, they can still be recognized at the structure and function levels (*6*–*11*). However, given the massive genetic and phenotypic viral diversity, this task has remained a big challenge.

One scheme for classifying protein domains based on their structural, functional, and evolutionary relationships is the Structural Classification of Proteins (SCOP) database (*9*). SCOP is considered the "gold standard" in the classification of protein domains with known 3D structures and provides useful evolutionary information on domains grouped into fold families (FFs) and fold superfamilies (FSFs). FFs include domains that are typically more than 30% identical in their sequence composition. In turn, FSFs group FFs with common 3D structural cores and biochemical properties, albeit with low sequence identity (could be <15%). Hence, FSFs are more conserved in evolution and suitable for deep evolutionary comparisons (*5*, *12*). This is demonstrated by the fact that nearly half a million protein sequences in UniProtKB/Swiss-Prot (*13*) map only to ~2300 FSFs (SCOP 1.75). Empirically, it has been shown that protein structure is at least 3 to 10 times more conserved than sequence (*14*). Moreover, the conserved 3D core of FSF domains rarely (that is, 0.4 to 4%) evolves by convergent evolution (*15*). A focus on FSF domains also puts bounds on the molecular diversity of viruses and cellular organisms. This and other advantages (*16*) make FSF domains reliable phylogenetic characters for evolutionary studies, especially when the focus is to reconstruct the deep evolutionary history of life [as shown previously (*5*, *17*–*19*)].

Here, we analyzed a total of 5080 completely sequenced proteomes from cells and viruses and assigned FSF domains to their proteins using structure-based hidden Markov models (HMMs) defined by the SUPERFAMILY database (version 1.75) (*20*). The viral data set included 3460 proteomes from 1649 double-stranded DNA (dsDNA), 534 single-stranded DNA (ssDNA), 166 double-stranded RNA (dsRNA), 991 single-stranded RNA (ssRNA) (881 plus sense and 110 minus sense), and 120 retrotranscribing (56 ssRNA-RT and 64 dsDNA-RT) viruses (table S1), whereas the cellular data set included 1620 proteomes from 122, 1115, and 383 organisms from the superkingdoms Archaea, Bacteria, and Eukarya, respectively (table S2). Applying both comparative genomic and phylogenomic strategies, we asked a number of crucial questions: Can we quantify viral diversity? How many unique protein folds exist in the virosphere? What is the predominant direction of gene transfer (cell-to-virus or virus-to-cell)? Are viruses infecting different organisms evolutionarily related? Can we identify protein folds that define viral lineages based on common virion architectures? Are viruses monophyletic or polyphyletic? Where do viruses lie on the tree of life? What were the earliest replicons?

Analysis revealed that, despite exhibiting high levels of diversity, viral proteomes retain traces of ancient evolutionary history that can

Evolutionary Bioinformatics Laboratory, Department of Crop Sciences and Illinois Informatics Institute, University of Illinois, Urbana, IL 61801, USA.
*Present address: Department of Biosciences, COMSATS Institute of Information Technology, Islamabad 45550, Pakistan.
†Corresponding author. E-mail: gca@illinois.edu

be recovered using advanced bioinformatics approaches. The most parsimonious hypothesis inferred from proteomic data suggests that viruses originated from multiple ancient cells that harbored segmented RNA genomes and coexisted with the ancestors of modern cells. We refer to the viral ancestors as "proto-virocells" to emphasize the cellular nature of ancient viruses and to distinguish them from modern virocells that produce elaborate virions [a virocell is any ribocell that, upon viral infection, produces viral progeny instead of dividing by binary fission; sensu (*21*, *22*)]. This implies the existence of ancient cellular lineages common to both cells and viruses before the appearance of the "last universal cellular ancestor" that gave rise to modern cells. According to our data, the prolonged pressure of genome and particle size reduction eventually reduced virocells into modern viruses (identified by the complete loss of cellular makeup), whereas other coexisting cellular lineages diversified into modern cells. The cellular nature of viruses is restored when modern viruses (re)take control of the cellular machinery of modern cells or when they integrate into cellular genomes. The model for the origin and evolution of the "viral supergroup" (a collection of seven viral subgroups defined by replicon type and replication strategy), as described in the Baltimore classification (*23*), captures the many aspects of viral diversity (for example, host preferences, viral morphologies, and proteomic makeup) and, as we show, is backed by strong support from molecular data.

## RESULTS

### Viral supergroup behaves similarly to cellular superkingdoms in terms of FSF sharing patterns

A total of 1995 significant FSF domains ($E < 0.0001$) were detected in ~11 million proteins of 5080 proteomes sampled from cells and viruses. A four-set Venn diagram showed that roughly two-thirds of the total FSFs (1279 of 1995) were detected only in cellular proteomes (that is, A, B, E, AB, AE, BE, and ABE Venn groups), whereas the remaining FSFs (716) either were shared between cells and viruses, represented by "XV" Venn groups (that is, AV, BV, EV, ABV, AEV, BEV, and ABEV), or were unique to viruses (V) (Fig. 1A). Viruses shared FSFs with each and every Venn group (that is, there were no zeros), indicating that Venn diagrams can be extended to include four groups, instead of three, without any oddities. The most populated Venn groups of universal FSFs found in both cells and viruses (ABEV) or shared by Archaea, Bacteria, and Eukarya (ABE) had 442 and 457 FSFs, respectively. The large size of the ABEV group, which is one-fifth of the total FSFs (442), suggests the coexistence of ancient viruses and cells, very much like the large size of ABE strengthening the hypothesis of a common origin of modern cells. In turn, FSFs unique to superkingdoms and viruses (that is, A, B, E, and V groups) indicate possible later gains specific for each supergroup. These gains were more common in Eukarya (283 FSFs) and Bacteria (154 FSFs)

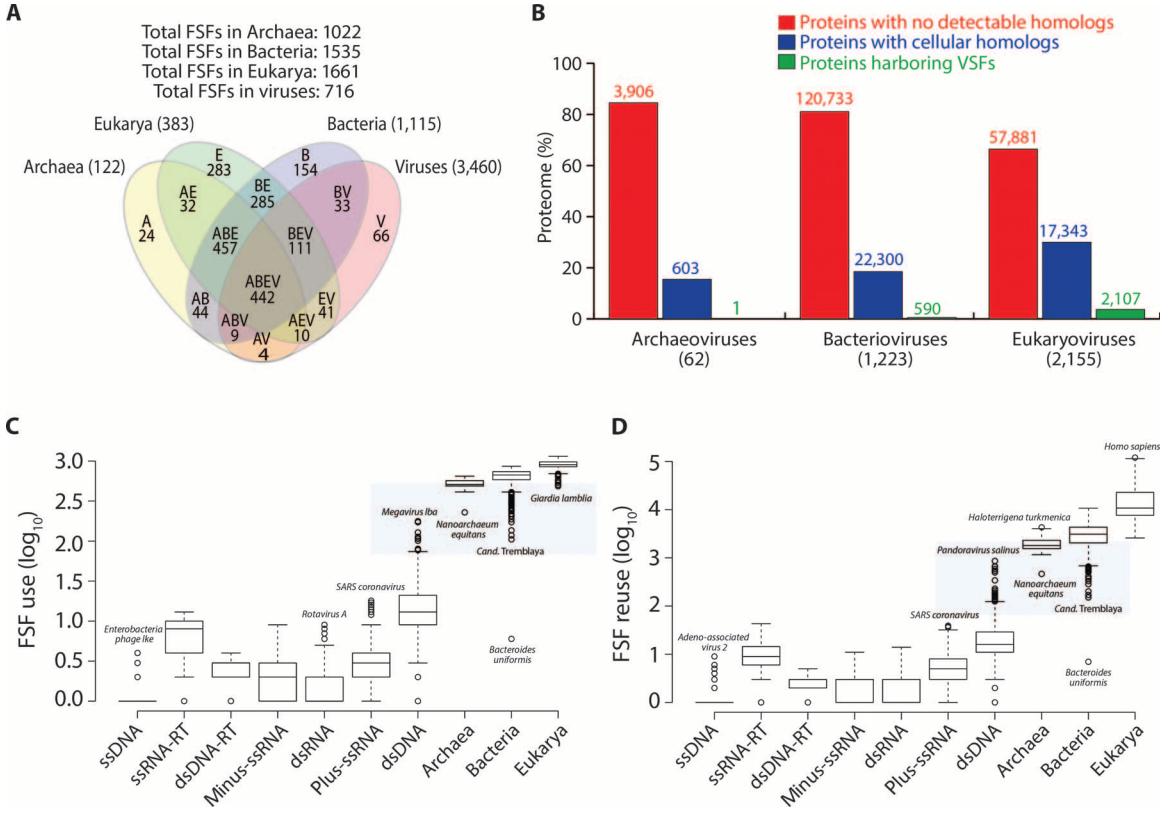


**Fig. 1. FSF sharing patterns and makeup of cellular and viral proteomes.** (**A**) Numbers in parentheses indicate the total number of proteomes that were sampled from Archaea, Bacteria, Eukarya, and viruses. (**B**) Barplots comparing the proteomic composition of viruses infecting the three superkingdoms. Numbers in parentheses indicate the total number of viral proteomes in each group. Numbers above bars indicate the total number of proteins in each of the three classes of proteins. VSFs are listed in Table 1. (**C** and **D**) FSF use and reuse for proteomes in each viral subgroup and in the three superkingdoms. Values given in logarithmic scale. Important outliers are labeled. Shaded regions highlight the overlap between parasitic cells and giant viruses.

**Table 1. VSFs and their distribution in the viral supergroup.** FSFs in boldface could be potential VSFs based on the criterion described in the text. FSFs were referenced by either SCOP ID or css. For example, the P-loop containing NTP hydrolase FSF is c.37.1, where "c" is the α/β class of secondary structure present in the protein domain, "37" is the fold, and "1" is the FSF.

| SCOP ID | SCOP css | Venn group | FSF description | Distribution |
|---|---|---|---|---|
| 69070 | a.150.1 | V | Anti-sigma factor AsiA | dsDNA |
| 55064 | d.58.27 | V | Translational regulator protein regA | dsDNA |
| 48493 | a.120.1 | V | Gene 59 helicase assembly protein | dsDNA |
| 89433 | b.127.1 | V | Baseplate structural protein gp8 | dsDNA |
| 69652 | d.199.1 | V | DNA binding C-terminal domain of the transcription factor MotA | dsDNA |
| 56558 | d.182.1 | V | Baseplate structural protein gp11 | dsDNA |
| 49894 | b.28.1 | V | Baculovirus p35 protein | dsDNA |
| 160957 | e.69.1 | V | Poly(A) polymerase catalytic subunit–like | dsDNA |
| 51289 | b.85.5 | V | Tlp20, baculovirus telokin-like protein | dsDNA |
| 88648 | b.121.6 | V | Group I dsDNA viruses | dsDNA |
| 161240 | g.92.1 | V | T-antigen–specific domain–like | dsDNA |
| 118208 | e.58.1 | V | Viral ssDNA binding protein | dsDNA |
| 54957 | d.58.8 | V | Viral DNA binding domain | dsDNA |
| 51332 | b.91.1 | V | E2 regulatory, transactivation domain | dsDNA |
| 56548 | d.180.1 | V | Conserved core of transcriptional regulatory protein vp16 | dsDNA |
| 90246 | h.1.24 | V | Head morphogenesis protein gp7 | dsDNA |
| 47724 | a.54.1 | V | Domain of early E2A DNA binding protein, ADDBP | dsDNA |
| 57917 | g.51.1 | V | Zn binding domains of ADDBP | dsDNA |
| 49889 | b.27.1 | V | Soluble secreted chemokine inhibitor, VCCI | dsDNA |
| 89428 | b.126.1 | V | Adsorption protein p2 | dsDNA |
| 82046 | b.116.1 | V | Viral chemokine binding protein m3 | dsDNA |
| 158974 | b.170.1 | V | WSSV envelope protein-like | dsDNA |
| 47852 | a.62.1 | V | Hepatitis B viral capsid (hbcag) | dsDNA-RT |
| 111379 | f.47.1 | V | VP4 membrane interaction domain | dsRNA |
| 48345 | a.115.1 | V | A virus capsid protein alpha-helical domain | dsRNA |
| 69908 | e.35.1 | V | Membrane penetration protein mu1 | dsRNA |
| 75347 | d.13.2 | V | Rotavirus NSP2 fragment, C-terminal domain | dsRNA |
| 69903 | e.34.1 | V | NSP3 homodimer | dsRNA |
| 75574 | d.216.1 | V | Rotavirus NSP2 fragment, N-terminal domain | dsRNA |
| 58030 | h.1.13 | V | Rotavirus nonstructural proteins | dsRNA |
| 49818 | b.19.1 | V | Viral protein domain | dsRNA, minus-ssRNA, plus-ssRNA |
| 88650 | b.121.7 | V | Satellite viruses | ssDNA |
| 48045 | a.84.1 | V | Scaffolding protein gpD of bacteriophage procapsid | ssDNA |
| 50176 | b.37.1 | V | N-terminal domains of the minor coat protein g3p | ssDNA |
| 75404 | d.213.1 | V | VSV matrix protein | Minus-ssRNA |
| 118173 | d.293.1 | V | Phosphoprotein M1, C-terminal domain | Minus-ssRNA |

continued on next page

| SCOP ID | SCOP css | Venn group | FSF description | Distribution |
|---|---|---|---|---|
| 69922 | f.12.1 | V | Head and neck region of the ectodomain of NDV fusion glycoprotein | Minus-ssRNA |
| 101089 | a.8.5 | V | Phosphoprotein XD domain | Minus-ssRNA |
| 58034 | h.1.14 | V | Multimerization domain of the phosphoprotein from Sendai virus | Minus-ssRNA |
| 50012 | b.31.1 | V | EV matrix protein | Minus-ssRNA |
| 48145 | a.95.1 | V | Influenza virus matrix protein M1 | Minus-ssRNA |
| 143021 | d.299.1 | V | Ns1 effector domain–like | Minus-ssRNA |
| 161003 | e.75.1 | V | Flu NP-like | Minus-ssRNA |
| 160453 | d.361.1 | V | PB2 C-terminal domain–like | Minus-ssRNA |
| 101156 | a.30.3 | V | Nonstructural protein ns2, Nep, M1 binding domain | Minus-ssRNA |
| 160892 | d.378.1 | V | Phosphoprotein oligomerization domain–like | Minus-ssRNA |
| 56983 | f.10.1 | V | Viral glycoprotein, central and dimerization domains | Plus-ssRNA |
| 101257 | a.190.1 | V | *Flavivirus* capsid protein C | Plus-ssRNA |
| 103145 | d.255.1 | V | Tombusvirus P19 core protein, VP19 | Plus-ssRNA |
| 89043 | a.178.1 | V | Soluble domain of poliovirus core protein 3a | Plus-ssRNA |
| 110304 | b.148.1 | V | Coronavirus RNA binding domain | Plus-ssRNA |
| 101816 | b.140.1 | V | Replicase NSP9 | Plus-ssRNA |
| 140367 | a.8.9 | V | Coronavirus NSP7–like | Plus-ssRNA |
| 143076 | d.302.1 | V | Coronavirus NSP8–like | Plus-ssRNA |
| 144246 | g.86.1 | V | Coronavirus NSP10–like | Plus-ssRNA |
| 103068 | d.254.1 | V | Nucleocapsid protein dimerization domain | Plus-ssRNA |
| 117066 | b.1.24 | V | Accessory protein X4 (ORF8, ORF7a) | Plus-ssRNA |
| 143587 | d.318.1 | V | SARS receptor binding domain–like | Plus-ssRNA |
| 159936 | d.15.14 | V | NSP3A-like | Plus-ssRNA |
| 160099 | d.346.1 | V | SARS Nsp1–like | Plus-ssRNA |
| 140506 | a.30.8 | V | FHV B2 protein–like | Plus-ssRNA |
| 144251 | g.87.1 | V | Viral leader polypeptide zinc finger | Plus-ssRNA |
| 141666 | b.164.1 | V | SARS ORF9b–like | Plus-ssRNA |
| 55671 | d.102.1 | V | Regulatory factor Nef | ssRNA-RT |
| 56502 | d.172.1 | V | gp120 core | ssRNA-RT |
| 57647 | g.34.1 | V | HIV-1 VPU cytoplasmic domain | ssRNA-RT |
| **49749** | **b.121.2** | **EV** | **Group II dsDNA viruses VP** | **dsDNA** |
| **103417** | **e.48.1** | **EV** | **Major capsid protein VP5** | **dsDNA** |
| **140713** | **a.251.1** | **EV** | **Phage replication organizer domain** | **dsDNA** |
| **161008** | **e.76.1** | **EV** | **Viral glycoprotein ectodomain–like** | **dsDNA, minus-ssRNA** |
| **110132** | **b.147.1** | **EV** | **BTV NS2-like ssRNA binding domain** | **dsRNA** |
| **82856** | **e.42.1** | **EV** | **L-A virus major coat protein** | **dsRNA** |
| **140809** | **a.260.1** | **EV** | **Rhabdovirus nucleoprotein–like** | **Minus-ssRNA** |
| **101399** | **a.206.1** | **EV** | **P40 nucleoprotein** | **Minus-ssRNA** |
| **55405** | **d.85.1** | **EV** | **RNA bacteriophage capsid protein** | **Minus-ssRNA** |

continued on next page

| SCOP ID | SCOP css | Venn group | FSF description | Distribution |
|---|---|---|---|---|
| 68918 | a.140.4 | BV | Recombination endonuclease VII, C-terminal and dimerization domains | dsDNA |
| 50017 | b.32.1 | BV | gp9 | dsDNA |
| 58046 | h.1.17 | BV | Fibritin | dsDNA |
| 56826 | e.27.1 | BV | Upper collar protein gp10 (connector protein) | dsDNA |
| 161234 | g.91.1 | BV | E7 C-terminal domain–like | dsDNA |
| 140919 | a.263.1 | BV | DNA terminal protein | dsDNA |
| 89064 | a.179.1 | BV | Replisome organizer (g39p helicase loader/inhibitor protein) | dsDNA |
| 160570 | d.368.1 | BV | YonK-like | dsDNA |
| 51327 | b.90.1 | BV | Head binding domain of phage P22 tailspike protein | dsDNA |
| 141658 | b.163.1 | BV | Bacteriophage trimeric proteins domain | dsDNA |
| 64210 | d.186.1 | BV | Head-to-tail joining protein W, gpW | dsDNA |
| 51274 | b.85.2 | BV | Head decoration protein D (gpD, major capsid protein D) | dsDNA |
| 159865 | d.186.2 | BV | XkdW-like | dsDNA |
| 101059 | a.159.3 | BV | B-form DNA mimic Ocr | dsDNA |
| 58091 | h.4.2 | BV | *Clostridium* neurotoxins, "coiled-coil" domain | dsDNA |
| 47681 | a.49.1 | BV | C-terminal domain of B transposition protein | dsDNA |
| 58059 | h.2.1 | BV | Tetramerization domain of the Mnt repressor | dsDNA |
| 54328 | d.15.5 | BV | Staphylokinase/streptokinase | dsDNA |
| 64465 | d.196.1 | BV | Outer capsid protein sigma 3 | dsRNA |
| 57987 | h.1.4 | BV | Inovirus (filamentous phage) major coat protein | ssDNA |
| 160940 | e.66.1 | BEV | Api92-like | dsDNA |
| 160459 | d.362.1 | BEV | BLRF2-like | dsDNA |
| 109859 | a.214.1 | BEV | NblA-like | dsDNA |
| 54334 | d.15.6 | BEV | Superantigen toxins, C-terminal domain | dsDNA |
| 51225 | b.83.1 | BEV | Fiber shaft of virus attachment proteins | dsDNA, dsRNA |
| 49835 | b.21.1 | BEV | Virus attachment protein globular domain | dsDNA, dsRNA |
| 50203 | b.40.2 | BEV | Bacterial enterotoxins | dsDNA, ssDNA |
| 111474 | h.3.3 | BEV | Coronavirus S2 glycoprotein | dsDNA, plus-ssRNA |
| 56831 | e.28.1 | BEV | Reovirus inner layer core protein p3 | dsRNA |
| 109801 | a.30.5 | AV | Hypothetical protein D-63 | dsDNA |
| 161229 | g.90.1 | ABV | E6 C-terminal domain–like | dsDNA |
| 74748 | a.154.1 | ABV | Variable surface antigen VlsE | dsDNA |
| 143602 | d.321.1 | ABEV | STIV B116-like | dsDNA |
| 58064 | h.3.1 | ABEV | Influenza hemagglutinin (stalk) | dsDNA, minus-ssRNA |

than in Archaea (24 FSFs) and viruses (66 FSFs) (Fig. 1A). The 66 virus-specific FSFs (VSFs) include domains involved in viral pathogenicity such as binding to host DNA and receptors, manipulating host immune systems, and encapsulating viral genomes with capsid proteins (Tables 1 and 2). VSFs uniquely identify the viral supergroup on a scale comparable to that of Archaea, Bacteria, and Eukarya, each of which also encodes its own set of unique FSFs (Fig. 1A). In fact, VSFs were 2.75-fold greater in number than the number of specific FSFs in Archaea, which is a bona fide superkingdom.

## VSFs are underestimated in our census

Viral genomes often integrate into cellular genomes and contribute proteins to their makeup. These proteins become part of XV Venn groups. To detect such transfers, we looked at the molecular functions of each

**Table 2. Significantly enriched "biological process" GO terms in (66 +43) VSFs (FDR < 0.01).**

| GO ID | GO term | *Z* score | *P* | FDR |
|---|---|---|---|---|
| GO:0044415 | Evasion or tolerance of host defenses | 14.56 | $4.01 \times 10^{6}$ | $3.00 \times 10^{5}$ |
| GO:0050690 | Regulation of defense response to virus by virus | 14.56 | $4.01 \times 10^{6}$ | $3.00 \times 10^{5}$ |
| GO:0044068 | Modulation by symbiont of host cellular process | 13.8 | $5.72 \times 10^{6}$ | $3.00 \times 10^{5}$ |
| GO:0052572 | Response to host immune response | 13.14 | $7.86 \times 10^{6}$ | $3.02 \times 10^{5}$ |
| GO:0002832 | Negative regulation of response to biotic stimulus | 12.57 | $1.05 \times 10^{5}$ | $3.02 \times 10^{5}$ |
| GO:0052255 | Modulation by organism of defense response of other organism involved in symbiotic interaction | 12.57 | $1.05 \times 10^{5}$ | $3.02 \times 10^{5}$ |
| GO:0051805 | Evasion or tolerance of immune response of other organism involved in symbiotic interaction | 12.57 | $1.05 \times 10^{5}$ | $3.02 \times 10^{5}$ |
| GO:0019048 | Modulation by virus of host morphology or physiology | 12.06 | $1.36 \times 10^{5}$ | $3.53 \times 10^{5}$ |

FSF in every XV group and identified FSFs that were rare in the proteomes of the corresponding superkingdom(s). As a threshold, we selected only those FSFs in the XV groups that were detected in ≤2% of the total number of X proteomes. Using this stringent criterion, we identified 43 additional FSFs that could be potential candidates for VSFs (highlighted in Table 1; see table S3 for percentages). Remarkably, the list includes several proteins critical to viruses, such as components of viral capsid/coat architectures, envelope membranes, and proteins involved in viral entry and cellular attachment. For example, the "Group II dsDNA viruses VP" FSF [SCOP concise classification string (css) b.121.1], which is the "double jelly-roll" capsid fold signature of many dsDNA viruses (*3*), and the "Major capsid protein VP5" FSF (e.48.1) of herpesviruses were categorized in the EV group, indicating that these FSFs were shared by eukaryotes and viruses. However, b.121.1 and e.48.1 were detected only in 5 of 383 (1.3%) and in 1 of 383 (0.3%) eukaryotic proteomes that were sampled, indicating a rare presence in eukaryotes. Because both FSFs are components of viral capsids and perform a "hallmark" viral function, their rare presence in eukaryotes is likely a result of horizontal gene transfer (HGT) from virus to host or a mistake in HMM assignment rather than shared innovation or vertical inheritance. Similarly, the "gp9" FSF (b.32.1) in the BV group helps in T4 bacteriophage attachment to its host, *Escherichia coli* (*24*). It was only detected in 1 of 1115 (0.08%) bacterial proteomes that were sampled, again suggesting either virus-to-host HGT or erroneous assignment. Remarkably, 20 of 33 (60.06%) BV FSFs were part of our selection, suggesting that a large number of BV FSFs originated in viruses. Because bacterioviruses are known to mediate gene exchange between bacterial species (*25*), our finding is biologically significant and less likely attributable to mistakes in HMM assignments. These observations suggest that VSFs are spreading to other Venn groups and that their number is expected to grow once a pool of more diverse viruses is sequenced and HGT-associated relationships are determined. Some of the "cell-only" Venn groups (that is, A, B, E, AB, AE, BE, and ABE) may also be contaminated with viral FSFs because a large number of viral FSFs remain unknown as a result of sampling biases and technical limitations in virus discovery in different species.

## VSFs originate independently in viral subgroups

Although VSFs were detected in all seven viral subgroups, they were mostly specific for them (Table 1). The exception was the "Viral protein domain" FSF (b.19.1) shared by dsRNA (rotaviruses), plus-ssRNA (coronaviruses), and minus-ssRNA viruses (influenza viruses) (Table 1). FSF b.19.1 is the β-sandwich domain in the capsid proteins of bluetongue virus and rotaviruses, where it facilitates virus attachment to the host cell (*26*, *27*). It is also present in the hemagglutinin glycoproteins of influenza viruses, helping recognize the cell surface receptor (*28*, *29*). Thus, it could be a unifying feature of most RNA viruses (read below). Extending the number of VSFs from 66 to 109 by considering the 43 potential VSFs as true VSFs did not change the overall picture (Table 1). Only six additional VSFs were shared by more than one viral subgroup, including mainly viral attachment proteins and envelope glycoproteins [with the possible exception of "Bacterial enterotoxins" FSF (b.40.2)]. Some of these could be candidates of virus-to-virus HGT during coinfection of a common host or could be vertically inherited from a common ancestor. Most VSFs were restricted to a single viral subgroup, suggesting that each genome type has evolved different VSFs to successfully carry out its reproductive cycle and that VSFs have evolved rather recently in viral lineages during infection cycles in host cells (confirmed below).

## Viral proteomes are enriched with proteins of unknown origin

It is sometimes argued that viral genomes only grow by acquiring genes from their hosts (*30*, *31*). To test whether this argument was supported by proteomic data, we classified viruses (according to their host type) into archaeoviruses, bacterioviruses, and eukaryoviruses, and studied their proteomic composition (Fig. 1B). In all cases, viral proteomes contained three classes of proteins: (i) those for which no structural relative was detected in the HMM library, (ii) those for which homologs existed in cellular proteomes, and (iii) proteins encoding VSFs. Class I proteins with no structural hits represented most viral proteins. Roughly 80% of prokaryotic proteins and 75% of eukaryoviral proteins belonged to this category (Fig. 1B), indicating that we know very little about the structures and functions of most viral proteins. Some of these could be very ancient and thus are no longer detectable by BLAST or structure-based HMMs. In turn, class II viral proteins are composed of FSFs that are also encoded by the proteins of their host cells. These could be either true orthologs or proteins acquired through HGT between cells and viruses. Finally, class III proteins encoding VSFs confirm that genes can originate in viruses and sometimes be transferred to cells, thus becoming class II proteins [a process that is now widely acknowledged by many authors (*5*, *32*–*35*)]. The global nature of viral proteomes must be considered when speculating

about viral origins because single-gene analyses do not provide a complete picture of viral evolution.

### Genome reduction: A better way to think about the viral mode of evolution

Analyses of FSF use (that is, total number of unique FSFs in a proteome) and reuse (total number of FSFs) (Fig. 1, C and D) revealed that giant viruses, such as *Megavirus lba* and *Pandoravirus salinus*, overlapped many parasitic and symbiotic microbial species (mostly Mycoplasma and Proteobacteria) (see table S4 for use and reuse values in all proteomes). To confirm and as a control, we plotted FSF use and reuse for viruses and only "free-living" organisms that eliminated the overlap between large dsDNA viruses and microbial parasites (fig. S1). Giant viruses were not too far away from archaeal species that also have experienced genome reduction in the past (*18*, *36*). The analysis suggests that proteomes of viruses, especially giant dsDNA viruses, are similar in size to many well-known cellular parasites and also share with them a similar lifestyle (that is, benefitting from host resources). This shows that one unifying property for cells and viruses could be common parasitic lifestyle. Because the small proteomes of cellular parasites are likely a result of reductive evolution (*37*–*39*), it would seem logical to extend this argument to the evolution of the viral supergroup [as previously argued (*40*–*42*)], albeit cautiously for RNA viruses with small proteome complements (read below).

### FSFs shared with viruses are more widespread in cellular proteomes

To infer the predominant direction of gene transfer (that is, virus to cell or cell to virus), we divided FSFs in each superkingdom into two sets: (i) those shared only with cells and (ii) those also shared with viruses. FSFs specific for each superkingdom (that is, A, B, and E Venn groups in Fig. 1A) were excluded because they represent gains unique to each superkingdom and de facto could not be subject to horizontal transfers unless they were later completely lost from the donor superkingdom. A total of 1022 FSFs were encoded by archaeal proteomes. After the exclusion of 24 Archaea-specific FSFs, 533 (52%) were shared only with Bacteria and Eukarya and 465 (45%) were also shared with viruses. Similarly, of 1535 total bacterial FSFs, 154 were Bacteria-specific, 786 (51%) were shared only with Archaea and Eukarya, and 595 (39%) were also shared with viruses. Finally, eukaryal proteomes encoded a total of 1661 FSFs, including 283 that were Eukarya-specific, 774 (47%) that were shared only with the superkingdoms Archaea and Bacteria, and 604 (36%) that were also shared with viruses. Next, we calculated a fractional ($f$) value to determine the spread of FSFs in the proteomes of each superkingdom (Fig. 2). The $f$ value gives the spread of each FSF in modern proteomes and ranges from 0 (complete absence in sampled proteomes) to 1 (present in all proteomes).

In all superkingdoms, FSFs shared with viruses were significantly more widespread in proteomes than those shared only with cells. The

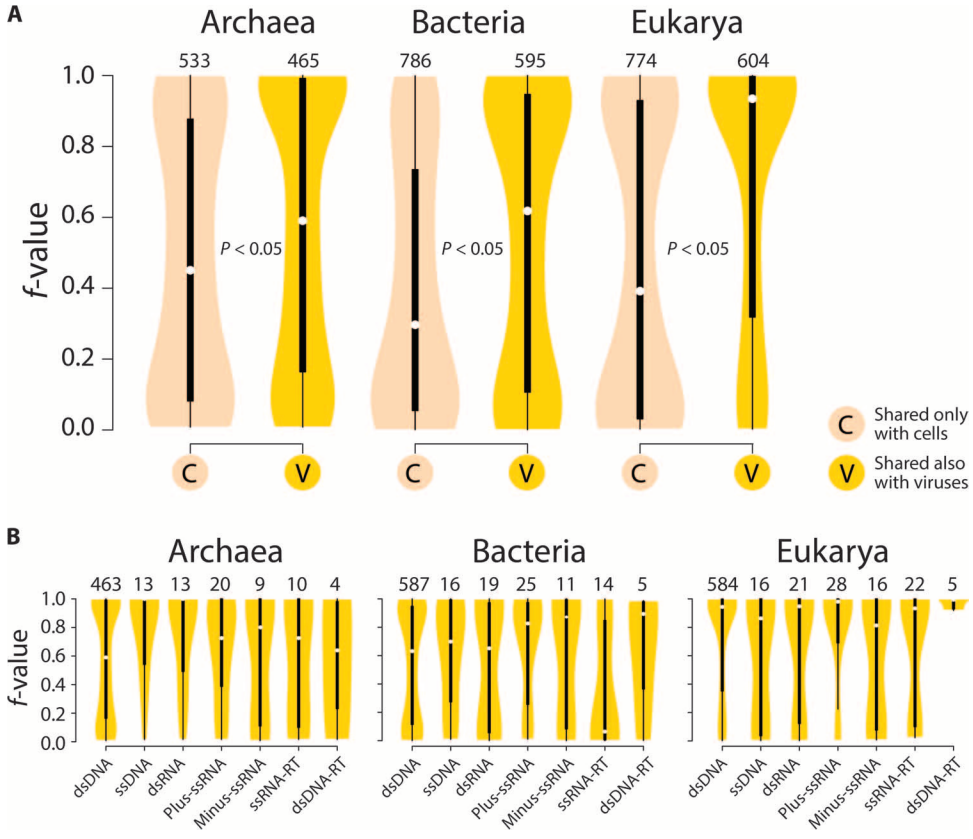


**Fig. 2. Spread of viral FSFs in cellular proteomes.** (**A**) Violin plots comparing the spread ($f$ value) of FSFs shared and not shared with viruses in archaeal, bacterial, and eukaryal proteomes. (**B**) Violin plots comparing the spread ($f$ value) of FSFs shared with each viral subgroup in archaeal, bacterial, and eukaryal proteomes. Numbers on top indicate the total number of FSFs involved in each comparison. White circles in each boxplot represent group medians. Density trace is plotted symmetrically around the boxplots.

median *f* value in Archaea for FSFs shared only with cells was 0.45, in comparison to 0.59 for FSFs shared with viruses (that is, a 31% increase in spread). Similarly, medians increased from 0.30 to 0.62 (up by 106%) in Bacteria and increased most significantly from 0.39 to 0.93 (up by 138%) in Eukarya (Fig. 2A). Regardless of the numerical differences between superkingdoms, FSFs shared with viruses were significantly more widespread in individual members of each superkingdom. One explanation is that viruses mediated the spread of these FSFs by serving as vehicles of gene transfer. It also suggests that viruses are very ancient and most likely infected the last common ancestor of each superkingdom because viral FSFs were present in a diverse array of cellular organisms ranging from small microbes to large eukaryotes. A breakdown by viral replicon type provided additional insights (Fig. 2B). In Archaea, nearly all of the viral FSFs were well represented in member species. Surprisingly, FSFs shared with RNA viruses were also enriched in archaeal proteomes. Because RNA viruses seemingly cannot carry out a productive infectious life cycle in Archaea (read below), it is unlikely that they picked these FSFs from archaeal hosts through HGT. In turn, it is more likely that RNA viruses infecting different superkingdoms share FSFs that were retained during their evolution from ancient cells. Similar patterns were also seen in bacterial proteomes (Fig. 2B). Remarkably, FSFs shared with each viral replicon type were almost universal (*f* approaching 1) among members of the Eukarya superkingdom. As we will now show, this is consistent with Eukarya hosting a large number of viruses from each replicon type.

## Viruses infecting the three superkingdoms share a conserved core of ancient FSFs

We calculated the "virus count" for each replicon type in major host groups to determine the virus-host relationships of viruses in our data set (Fig. 3A). The exercise revealed that most RNA viral subgroups were exclusive to eukaryotes (for example, minus-ssRNA and retrotranscribing viruses) (Fig. 3A). In turn, a large number of DNA viruses (mostly *Caudovirales*) infected prokaryotic hosts. The bias in the distribution of replicon types in superkingdoms (that is, DNA viruses in prokaryotes and RNA viruses in eukaryotes) leads to an interesting possibility about the early origin of RNA viruses and their loss in prokaryotes [see Discussion (*43*)]. Virus-host relationships have been described in detail previously (*43*–*45*). Here, the more relevant question was asked: Do viruses infecting distantly related hosts share common protein folds? To answer, we generated a Venn diagram describing viral FSF repertoires. FSFs that were shared by archaeoviruses (*a*), bacterioviruses (*b*), and eukaryoviruses (*e*) were pooled into the *abe* Venn group; those shared by viruses infecting two different superkingdoms were pooled into the *ab*, *ae*, or *be* group; and those unique to viruses infecting a single superkingdom were pooled into the *a*, *b*, and *e* groups (Venn group nomenclature avoids ambiguity with that of Fig. 1A) (Fig. 3B). We stress that FSFs in the *abe* group do not mean that these were present in a virus capable of infecting Archaea, Bacteria, and Eukarya. To date, no virus is known to infect organisms in more than one superkingdom. Instead, it simply refers to the count of FSFs that were shared between archaeoviruses, bacterioviruses, and eukaryoviruses.

We discovered that viruses infecting species in each of the three superkingdoms shared a total of 68 FSFs (Fig. 3B, *abe* group). A closer inspection revealed that these FSFs performed crucial metabolic functions (table S5) and were widespread in cellular proteomes ($f > 0.75$) (Fig. 3C). These FSFs originated very early in evolution (fig. S2, *abe* group) and were detected in a large number of viruses from each replicon type (Fig. 3C). In fact, 19 *abe* FSFs (28%) were shared by two or more viral subgroups. It is often argued that, because viruses infect all species, they must have originated before modern cells. Here, we show that viruses infecting the three superkingdoms have a very large and conserved structural core that is particularly enriched in crucial metabolic functions believed to be very ancient. This is a strong indication of both the ancient origin of viruses and their coexistence with ancient cells. An alternative explanation could be the transfer of these FSFs

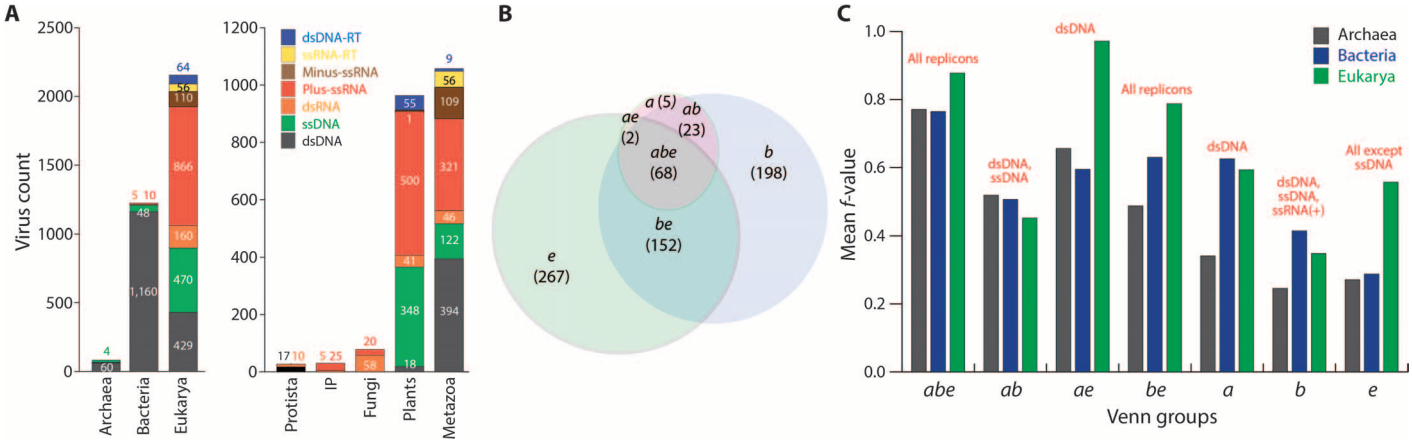


**Fig. 3. Virus-host preferences and FSF distribution in viruses infecting different hosts.** (**A**) The abundance of each viral replicon type that is capable of infecting Archaea, Bacteria, and Eukarya and major divisions in Eukarya. Virus-host information was retrieved from the National Center for Biotechnology Information Viral Genomes Project (*119*). Hosts were classified 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 3440 of the 3660 viruses that were sampled in this study. Two additional ssDNA archaeoviruses were added from the literature (*129*, *130*). Numbers on bars indicate the total virus count in each host group. (**B**) Venn diagram shows the distribution of 715 (of 716) FSFs that were detected in archaeoviruses, bacterioviruses, and eukaryoviruses. Host information on the Circovirus-like genome RW_B virus encoding the "Satellite viruses" FSF (b.121.7) was not available. (**C**) Mean *f* values for FSFs corresponding to each of the seven Venn groups defined in (B) in archaeal, bacterial, and eukaryal proteomes. Values were averaged for all FSFs in each of the seven Venn groups. Text above bars indicates how many different viral subgroups encoded those FSFs.

from modern cells to viruses through HGT. However, viruses do not infect hosts separated by large evolutionary distances [Fig. 3A; see also (*44*)]. Still these FSFs were detected in seemingly unrelated viruses. Moreover, roughly similar patterns were also observed for the *ab*, *ae*, and *be* FSFs (Fig. 3C and table S5). This greatly reduces confidence in cell-to-virus HGT because the probability of a large number of similar HGT events occurring in very different environments (that is, different hosts and viruses) is very unlikely.

However, a minor role for HGT cannot be ruled out. In fact, FSFs in *a*, *b*, or *e* Venn groups could be more influenced by HGT because they represent viruses infecting only a single superkingdom. For example, five FSFs that were detected only in archaeoviruses (Fig. 3B, *a* group) ["Ada DNA repair protein, N-terminal domain (N-Ada 10)" (g.48.1), "An anticodon binding domain of class I aminoacyl-tRNA synthetases" (a.97.1), "Carbamoyl phosphate synthetase, small subunit N-terminal domain" (c.8.3), "ArfGap/RecO-like zinc finger" (g.45.1), and "Hypothetical protein D-63" (a.30.5) FSFs (table S5)] appear more "cellular" than "viral" in nature. Here, the possibility that archaeoviruses picked these FSFs from archaeal hosts during infection cannot be ruled out with confidence. These FSFs were, however, more widespread in bacterial and eukaryal proteomes than in archaeal proteomes but were absent in their respective viruses (Fig. 3C). This could be a result of the loss of viral lineages from Bacteria and Eukarya or from reductive evolution in Archaea itself (*18*, *36*), which would again negate HGT. In turn, *b* and *e* FSFs were more represented in bacterial and eukaryal proteomes, respectively (as expected), and did not have very high *f* values (Fig. 3C). Specifically, 198 FSFs unique to bacterioviruses could be a result of HGT in either direction in Bacteria and viruses, especially because bacterioviruses are known to mediate gene exchange between bacterial species (for example, the 60% BV FSFs that could be potential VSFs; Table 1) and most of these FSFs originated very late in evolution (fig. S2, *b* group). Similar patterns were also observed for *e* FSFs (fig. S2, *e* group). Finally, only two FSFs ["DNA polymerase β, N-terminal domain–like" (a.60.6) and "Alkaline phosphatase–like" (c.76.1)] were shared by archaeoviruses and eukaryoviruses (*ae*). This is in line with previous understanding that eukaryoviruses are very distinct from archaeoviruses (*46*) and challenges the concept that eukaryoviruses originated from the merging of prokaryotic viruses [for example, (*45*); see Discussion]. In summary, the evolution of viruses follows a bidirectional route influenced by both the vertical inheritance of a structural core present in many distantly related viruses (that is, those infecting more than one superkingdom) and the HGT of FSFs from modern cells. The common core includes proteins mainly of cellular origin that likely originated in ancient cells.

### Testing capsid/coat structure–based viral lineages

Viruses infecting different organisms often use conserved 3D protein folds to produce capsids and show striking similarities in their virion architecture. These observations have led to the proposal of a structure-based viral taxonomy (*47*). Now, four major viral lineages have been defined mainly for icosahedral viruses (the most commonly seen capsid symmetry): "picornavirus-like," "PRD1/adenovirus-like," "HK97-like," and "BTV-like" (*47*). These lineages capture many viral families and attempt to simplify the overall diversity of the virosphere. Member viruses of the PRD1/adenovirus-like (characterized by the double jelly-roll fold) and HK97-like lineages infect species in the three superkingdoms, suggesting their ancient origin before the divergence of modern cells (*47*). To test this classification and to determine how the signature FSFs of each lineage distributed in our data set, we identified 22 capsid/coat–related FSFs using a keyword search of "capsid" and "coat" in SCOP 1.75 and in the literature (Table 3). Member FSFs of each major lineage, along with their abundance in cellular proteomes, are listed in Table 3. The HMM-based computational approach quickly reproduced the four major capsid-based viral lineages along with proposals for additions to some lineages (for example, negative-sense RNA viruses in picornavirus-like lineage) (table S6). Only very few members were missing (table S6), which could be a result of using a stringent criterion in assigning FSFs to viral proteins ($E < 0.0001$) that likely missed some hits but also protected from false-positive assignments. In short, FSFs identified in our study could be used as bait for quick assignment of viruses to major viral lineages defined by a common virion architecture and capsid/coat similarities (*47*).

### Viral hallmark architectures in cells

To confirm whether capsid/coat–related FSFs were indeed exclusive to viruses, we checked for the presence of 22 capsid/coat–related viral FSFs in the 1620 cellular proteomes that were sampled. Of the total 22 FSFs, 19 were either completely or nearly completely absent in cells (Table 3). Only the "Major capsid protein gp5" FSF (d.183.1) of *Caudovirales* (HK97-like lineage) was present in ~24% of cellular proteomes. The HK97-like fold has been detected in the shell-forming protein (encapsulin) of some archaeal protein nanocompartments that store metabolic enzymes (*48*). These nanocompartments are polyhedral protein shells that are morphologically similar to icosahedral viruses. Because archaeal and bacterial encapsulins are homologous, it is likely that prokaryotic microcompartments are closely related to ancient viral capsids (*49*). Those of bacterial carboxysomes are also morphologically similar to viral capsids (*50*) but are built from protein folds not yet detected in viruses (*51*). We identified two FSFs that are part of bacterial carboxysomes: (i) "Ccmk-like" (d.58.56) and (ii) "EutN/CcmL-like" (b.40.15) FSFs. Both had an *f* value of 0 in sampled viral proteomes, confirming a lack of overlap between carboxysomes and viral capsids. However, this could also be explained by the loss of an ancient capsid protein fold in modern viruses or an outcome of sampling biases (*49*). Alternatively, it is possible that viruses harboring similar folds exist in nature but remain to be discovered. An interesting analogy could also be made for eukaryotes where histone monomers assemble around DNA to produce chromatin structure. Remarkably, this process is mediated by histone chaperones that harbor the "jelly-roll" fold (*52*) that is abundant in icosahedral viruses. Thus, on the basis of current knowledge, although most viral capsid/coat FSFs have no SCOP structural relatives and lack cellular homologs (Table 3), rare capsid structural homologies in cellular proteomes suggest either instances of virus-to-host HGT or relics of the ancient coexistence of cells and viruses.

### FSF distributions in the viral supergroup are very patchy but highlight a major contribution from RNA viruses

Next, we explored how the 716 viral FSFs distributed between viral replicon types (Fig. 4). Most viral FSFs were only detected in dsDNA viruses (Fig. 4A). In comparison, proteomes of the ssDNA, ssRNA, dsRNA, and retrotranscribing groups were genetically poor. Roughly, 91% (649 of 716) of the total viral FSFs were unique to a single viral subgroup, and only ~9% (67) of the total viral FSFs were shared by more than one subgroup (Fig. 4A). The number of shared FSFs in each viral subgroup exceeded the number of unique FSFs, except for dsDNA and minus-ssRNA viruses. A seven-set Venn diagram made clear that each

**Table 3. FSFs involved in capsid/coat assembly processes in viruses.** FSFs that are completely absent in cellular proteomes are presented in boldface. Several other FSFs also have negligible *f* values in cells.

| SCOP ID | SCOP css | FSF description | Viral lineage | *f*-value in cells |
|---|---|---|---|---|
| 82856 | e.42.1 | L-A virus major coat protein | BTV-like | 0.00025 |
| 56831 | e.28.1 | Reovirus inner layer core protein p3 | BTV-like | 0.00019 |
| **48345** | **a.115.1** | **A virus capsid protein alpha-helical domain** | **BTV-like** | **0** |
| 56563 | d.183.1 | Major capsid protein gp5 | HK97-like | 0.2352 |
| 103417 | e.48.1 | Major capsid protein VP5 | HK97-like | 0.00006 |
| 88633 | b.121.4 | Positive stranded ssRNA viruses | Picornavirus-like | 0.00364 |
| 88645 | b.121.5 | ssDNA viruses | Picornavirus-like | 0.00099 |
| **88650** | **b.121.7** | **Satellite viruses** | **Picornavirus-like** | **0** |
| **88648** | **b.121.6** | **Group I dsDNA viruses** | **Picornavirus-like** | **0** |
| 49749 | b.121.2 | Group II dsDNA viruses VP | PRD1/adenovirus-like | 0.00031 |
| 47353 | a.28.3 | Retrovirus capsid dimerization domain–like | Other/unclassified | 0.00407 |
| 47943 | a.73.1 | Retrovirus capsid protein, N-terminal core domain | Other/unclassified | 0.00123 |
| 47195 | a.24.5 | TMV-like viral coat proteins | Other/unclassified | 0.00099 |
| 57987 | h.1.4 | Inovirus (filamentous phage) major coat protein | Other/unclassified | 0.00068 |
| 51274 | b.85.2 | Head decoration protein D (gpD, major capsid protein D) | Other/unclassified | 0.00049 |
| 64465 | d.196.1 | Outer capsid protein sigma 3 | Other/unclassified | 0.00006 |
| 55405 | d.85.1 | RNA bacteriophage capsid protein | Other/unclassified | 0.00006 |
| **48045** | **a.84.1** | **Scaffolding protein gpD of bacteriophage procapsid** | **Other/unclassified** | **0** |
| **47852** | **a.62.1** | **Hepatitis B viral capsid (hbcag)** | **Other/unclassified** | **0** |
| **101257** | **a.190.1** | ***Flavivirus* capsid protein C** | **Other/unclassified** | **0** |
| **50176** | **b.37.1** | **N-terminal domains of the minor coat protein g3p** | **Other/unclassified** | **0** |
| **103068** | **d.254.1** | **Nucleocapsid protein dimerization domain** | **Other/unclassified** | **0** |

viral subgroup shared FSFs with every other subgroup (the sole exception being ssDNA and dsDNA-RT viruses) but did so sparsely (Fig. 4A, Venn diagram). The diagram shows that there was no single FSF common to all viral subgroups (Fig. 4A). However, it also revealed that the minus-ssRNA and dsDNA groups circumscribed the most widely shared FSFs (traces highlighted in the Venn diagram) (Table 4).

The "DNA/RNA polymerases" FSF (e.8.1), which includes T7 RNA polymerase, RNA-dependent RNA polymerase of plus-sense and dsRNA viruses, reverse transcriptase, DNA polymerase I, and the catalytic domain of Y-family DNA polymerase, was detected in six of the seven subgroups (the exception being ssDNA viruses, which replicate by using the host's polymerase). In turn, "S-adenosyl-L-methionine dependent methyltransferases" (c.66.1) FSF was detected in five of the seven viral subgroups, except retrotranscribing viruses. Three additional FSFs, the "P-loop containing NTP hydrolase" (c.37.1), the "Ribonuclease H–like" (c.55.3) and the "Positive stranded ssRNA viruses" (b.121.4), were detected in four of the seven viral subgroups (Table 4). The c.37.1 FSF is one of the most abundant and widespread FSFs in modern cells. The c.55.3 superfamily includes many proteins involved in informational processes (including replication and translation) that are universal among cellular proteomes. This FSF was relatively widespread in viral subgroups but was absent in the proteomes of plus-ssRNA, dsRNA, and dsDNA-RT viruses. It was especially abundant in ssRNA-RT (79% of proteomes) and dsDNA (58%) viruses. The c.55.3 FSF also includes the catalytic domain of retroviral integrase, which is an important target to silence retroviral gene expression (*53*) and is medically important. In turn, b.121.4 is the jelly-roll fold, which is one of the most common topologies observed in viral capsid proteins (*3*, *54*). Finally, 10 FSFs were present in three of the six viral subgroups, whereas 52 were shared by two subgroups (Fig. 4A, Venn diagram, and Table 4).

The seven-set Venn diagram is analogous to a maze or logic puzzle that can be solved using Ariadne's thread logic (Fig. 4B). Metaphorically, threads keep track of evolutionary paths while traversing a maze sculpted by reductive loss. We define our Ariadne's threads as Venn subgroups of FSFs shared by two to six of the seven viral replicon types (there were no FSFs shared by all seven viral groups). These threads revealed that only 19 of the 120 possible Venn subgroups of shared FSFs existed (total Venn – internal groups: $2^7 - 1 = 127$), where 14 were shared by two to three viral groups. They make explicit how sparsely shared FSFs are in viral groups and uncover deep evolutionary patterns likely left by reductive evolution. Only 8 of 21 and only 6 of 35 possible subgroups shared by two and three viral groups, respectively, were pres-

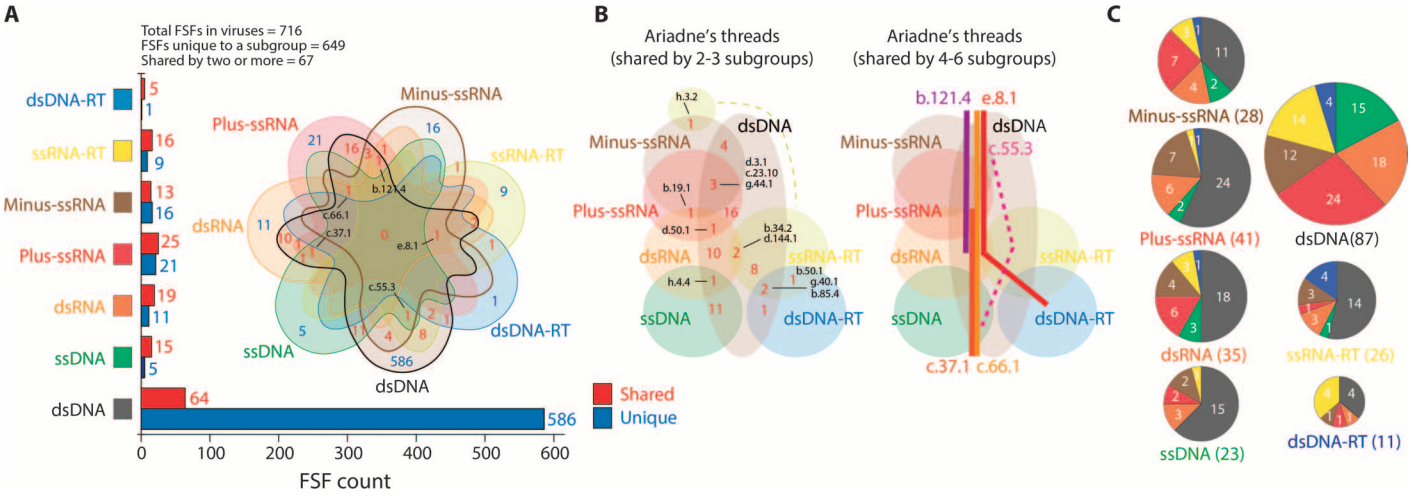


**Fig. 4. FSF distribution in the viral supergroup.** (**A**) Total number of FSFs that were either shared or uniquely present in each viral subgroup. A seven-set Venn diagram makes explicit the 127 ($2^7$ – 1) combinations that are possible with seven groups. (**B**) Ariadne's threads give the most parsimonious solution to encase all highly shared FSFs between different viral subgroups. Threads were inferred directly from the seven-set Venn diagram. FSFs identified by SCOP css. (**C**) Number of FSFs shared in each viral subgroup with every other subgroup. Pie charts are proportional to the size of the FSF repertoire in each viral subgroup.

ent. As expected, dsDNA viruses, which hold the largest proteomes and comparatively are minimally affected by reductive evolution, were part of 11 of these 14 Venn subgroups. Remarkably, 9 of 14 (64%) subgroups with their 39 FSFs (63%) involved minus-ssRNA, plus-ssRNA, and dsRNA replicons. Similarly, of 64 possible groups sharing four to seven replicon types, only five groups were present (lines in Ariadne's thread diagram; Fig. 4B), all including RNA viruses. As mentioned previously, these five groups represent polymerases, metabolic enzymes, ribonuclease, and capsid-associated FSFs. Because RNA viruses define most threads and their proteomes are the most reduced, they are most informative in explaining FSF distribution in the viral supergroup. This leads to the speculation that perhaps RNA viruses predated DNA viruses in evolution, which we confirm with phylogenetic methods below. Finally, a large number of FSFs were shared between DNA and RNA viruses (Fig. 4C), suggesting that the virosphere may not be as disjoint as previously thought. In fact, recombination between RNA and DNA viruses can sometimes generate "hybrid" viruses with DNA genomes but capsids typical of RNA viruses [for example, (*55*)].

In summary, the patchy distribution of FSFs in the viral supergroup revealed a significant overlap between viruses of different replicon types. Although most FSFs were unique to a particular subgroup, a large number of FSFs were shared between viruses belonging to different replicon types (Fig. 4).

## Phylogenomic analysis of FSF domains identifies two phases in the evolution of viruses

The reconstruction of phylogenomic trees of domains (ToD), which describe the evolution of the 1995 FSF domains (taxa) that were surveyed in the 5080 sampled proteomes (characters) (see Materials and Methods for the tree reconstruction protocol), showed that most viral FSFs originated very early in evolution (see the legend bar on top of ToD in Fig. 5A). Because of its highly unbalanced nature, ToD enabled the calculation of a "proxy" for the relative age of each FSF domain, which was defined as the node distance (nd) value. This value was derived simply by counting the number of nodes from a terminal taxon to the root node of the tree and by expressing the phylogenetic distance on a relative scale from 0 (most ancient) to 1 (most recent) [methodology discussed elsewhere (*18*)]. We have previously shown that nd is a reliable proxy for the evolutionary age of FSFs and describes a clock-like behavior of FSF evolution that is remarkably consistent with geological records (*56*). To uncover likely evolutionary scenarios, we plotted FSFs in each of the 15 Venn groups in Fig. 1A against their FSF ages (that is, nd values) (boxplots in Fig. 5A).

The ABEV Venn group, which includes 442 FSFs encoded by both cells and viruses, was the most ancient group and covered the entire nd axis. The P-loop containing NTP hydrolase (c.37.1) FSF was the first FSF to appear at nd = 0. The median nd was ~0.4, suggesting that at least 50% of ABEV FSFs originated very early in evolution and were shared by cells and viruses. This finding is remarkable and implies that some of the most ancient FSFs found in cells were also shared by very different groups of viruses, again suggesting the ancient coexistence of cells and viruses. In turn, the relatively longer tail on the right of the graph likely includes many FSFs of recent origin (nd > 0.63) that could have been gained in viruses from cells through HGT.

The ABEV group was followed by the appearance of the ABE group. The first FSF in ABE was the "ACT-like" FSF (d.58.18), which includes regulatory protein domains mainly involved in amino acid metabolism and transport. We propose that d.58.18 was most likely "lost" (or never gained) in ancient viruses because simultaneous gain in three superkingdoms is less likely than loss in just one superkingdom. By extension, the appearance of the BEV group with the inception of the "Lysozyme-like" FSF (d.2.1) at nd = 0.15 signals the loss of the first FSF in a cellular superkingdom (Archaea). Simply, the absence of an ancient FSF in one group (out of three or four groups) is more likely a result of reductive evolution than separate gains [as previously described (*18*)]. The previously reconstructed proteome of the last common ancestor of Archaea, Bacteria, and Eukarya was reported to encode a minimum of 70 FSFs (*57*). The most recent of those FSFs, "Terpenoid synthases" FSF (a.128.1), appeared at nd = 0.19 and was absent in all viruses, except one (African swine fever virus). These events demonstrate the early reductive

**Table 4. FSFs shared by different viral subgroups.**

| SCOP ID | SCOP css | FSF description | Distribution |
|---|---|---|---|
| 56672 | e.8.1 | DNA/RNA polymerases | dsDNA, dsRNA, dsDNA-RT, ssRNA-RT, minus-ssRNA, plus-ssRNA |
| 52540 | c.37.1 | P-loop containing nucleoside triphosphate hydrolases | dsDNA, dsRNA, ssDNA, plus-ssRNA |
| 53335 | c.66.1 | *S*-Adenosyl-L-methionine–dependent methyltransferases | dsDNA, dsRNA, ssDNA, minus-ssRNA, plus-ssRNA |
| 53098 | c.55.3 | Ribonuclease H–like | dsDNA, ssRNA-RT, ssDNA, minus-ssRNA |
| 88633 | b.121.4 | Positive stranded ssRNA viruses | dsDNA, dsRNA, minus-ssRNA, plus-ssRNA |
| 57850 | g.44.1 | RING/U-box | dsDNA, minus-ssRNA, plus-ssRNA |
| 51283 | b.85.4 | dUTPase-like | dsDNA, dsDNA-RT, ssRNA-RT |
| 56112 | d.144.1 | Protein kinase–like (PK-like) | dsDNA, dsRNA, ssRNA-RT |
| 54768 | d.50.1 | dsRNA binding domain–like | dsDNA, dsRNA, plus-ssRNA |
| 54001 | d.3.1 | Cysteine proteinases | dsDNA, minus-ssRNA, plus-ssRNA |
| 52266 | c.23.10 | SGNH hydrolase | dsDNA, minus-ssRNA, plus-ssRNA |
| 58100 | h.4.4 | Bacterial hemolysins | dsDNA, dsRNA, ssDNA |
| 49818 | b.19.1 | Viral protein domain | dsRNA, minus-ssRNA, plus-ssRNA |
| 57756 | g.40.1 | Retrovirus zinc finger–like domains | dsDNA, dsDNA-RT, ssRNA-RT |
| 50044 | b.34.2 | SH3 domain | dsDNA, dsRNA, ssRNA-RT |
| 57924 | g.52.1 | Inhibitor of apoptosis (IAP) repeat | dsDNA, plus-ssRNA |
| 50249 | b.40.4 | Nucleic acid binding proteins | dsDNA, ssDNA |
| 53041 | c.53.1 | Resolvase-like | dsDNA, ssDNA |
| 55550 | d.93.1 | SH2 domain | dsDNA, ssRNA-RT |
| 55464 | d.89.1 | Origin of replication binding domain, RBD-like | dsDNA, ssDNA |
| 56399 | d.166.1 | ADP ribosylation | dsDNA, ssDNA |
| 100920 | b.130.1 | Heat shock protein 70 kD (HSP70), peptide binding domain | dsDNA, plus-ssRNA |
| 47413 | a.35.1 | Lambda repressor–like DNA binding domains | dsDNA, ssDNA |
| 69065 | a.149.1 | RNase III domain–like | dsDNA, plus-ssRNA |
| 46785 | a.4.5 | Winged helix DNA binding domain | dsDNA, ssDNA |
| 53448 | c.68.1 | Nucleotide-diphospho-sugar transferases | dsDNA, dsRNA |
| 57997 | h.1.5 | Tropomyosin | dsDNA, dsRNA |
| 54236 | d.15.1 | Ubiquitin-like | dsDNA, ssRNA-RT |
| 47954 | a.74.1 | Cyclin-like | dsDNA, ssRNA-RT |
| 90229 | g.66.1 | CCCH zinc finger | dsDNA, minus-ssRNA |
| 103657 | a.238.1 | BAR/IMD domain–like | dsDNA, ssRNA-RT |
| 53067 | c.55.1 | Actin-like ATPase domain | dsDNA, plus-ssRNA |
| 47794 | a.60.4 | Rad51 N-terminal domain–like | dsDNA, ssDNA |
| 143990 | d.336.1 | YbiA-like | dsDNA, plus-ssRNA |
| 55811 | d.113.1 | Nudix | dsDNA, dsRNA |
| 51197 | b.82.2 | Clavaminate synthase–like | dsDNA, plus-ssRNA |
| 53756 | c.87.1 | UDP-glycosyltransferase/glycogen phosphorylase | dsDNA, dsRNA |

continued on next page

| SCOP ID | SCOP css | FSF description | Distribution |
|---|---|---|---|
| 81665 | f.33.1 | Calcium ATPase, transmembrane domain M | dsDNA, plus-ssRNA |
| 52949 | c.50.1 | Macro domain–like | dsDNA, plus-ssRNA |
| 53955 | d.2.1 | Lysozyme-like | dsDNA, dsRNA |
| 49899 | b.29.1 | Concanavalin A–like lectins/glucanases | dsDNA, dsRNA |
| 48371 | a.118.1 | ARM repeat | dsDNA, plus-ssRNA |
| 51126 | b.80.1 | Pectin lyase–like | dsDNA, plus-ssRNA |
| 47598 | a.43.1 | Ribbon-helix-helix | dsDNA, ssDNA |
| 50494 | b.47.1 | Trypsin-like serine proteases | dsDNA, plus-ssRNA |
| 55144 | d.61.1 | LigT-like | dsDNA, plus-ssRNA |
| 81296 | b.1.18 | E set domains | dsDNA, plus-ssRNA |
| 161008 | e.76.1 | Viral glycoprotein ectodomain–like | dsDNA, minus-ssRNA |
| 90257 | h.1.26 | Myosin rod fragments | dsDNA, dsRNA |
| 57501 | g.17.1 | Cystine-knot cytokines | dsDNA, ssRNA-RT |
| 54117 | d.9.1 | Interleukin 8–like chemokines | dsDNA, dsRNA |
| 58069 | h.3.2 | Virus ectodomain | ssRNA-RT, minus-ssRNA |
| 50630 | b.50.1 | Acid proteases | dsDNA-RT, ssRNA-RT |
| 47459 | a.38.1 | HLH, helix-loop-helix DNA binding domain | dsDNA, ssRNA-RT |
| 50939 | b.68.1 | Sialidases | dsDNA, minus-ssRNA |
| 55166 | d.65.1 | Hedgehog/DD peptidase | dsDNA, ssDNA |
| 51225 | b.83.1 | Fiber shaft of virus attachment proteins | dsDNA, dsRNA |
| 49835 | b.21.1 | Virus attachment protein globular domain | dsDNA, dsRNA |
| 111474 | h.3.3 | Coronavirus S2 glycoprotein | dsDNA, plus-ssRNA |
| 55658 | d.100.1 | L9 N-domain–like | dsDNA, dsDNA-RT |
| 55895 | d.124.1 | Ribonuclease Rh–like | dsDNA, plus-ssRNA |
| 52972 | c.51.4 | ITPase-like | dsDNA, plus-ssRNA |
| 57959 | h.1.3 | Leucine zipper domain | dsDNA, ssRNA-RT |
| 50203 | b.40.2 | Bacterial enterotoxins | dsDNA, ssDNA |
| 48208 | a.102.1 | Six-hairpin glycosidases | dsDNA, ssDNA |
| 50022 | b.33.1 | ISP domain | dsDNA, ssRNA-RT |
| 58064 | h.3.1 | Influenza hemagglutinin (stalk) | dsDNA, minus-ssRNA |

tendencies in early cellular lineages, especially in ancient cells leading to viruses and Archaea.

In comparison, FSFs unique to superkingdoms and the viral supergroup appeared much later (see the A, B, E, and V groups in Fig. 5A). These gains signaled the diversification of that superkingdom or supergroup. The late appearance of VSFs (V group in Fig. 5A) is interesting because it includes FSFs involved in viral pathogenicity (Tables 1 and 2). The phylogenomic analysis shows that VSFs originated at the same time or after the diversification of modern cells. Thus, they represent the time point when proto-virocells, under prolonged genome reduction pressure, completely lost their cellular nature and became fully dependent on emerging archaeal, bacterial, and eukaryal cells for reproduction. This idea is strengthened by the evolutionary appearances of the AV, BV, and EV groups soon after the FSFs of the superkingdom-specific A, B, and E groups, respectively (see patterned regions in Fig. 5A). We speculate that FSFs in the AV, BV, and EV groups either perform functions required by viruses to successfully infect their hosts (for example, BV FSFs that perform viral functions) or are simply HGT gains from their hosts. We have already discussed the composition of the BV group, which includes ~60% FSFs of viral origin (Table 1). Similarly, the AV and EV groups also include viral FSFs, albeit in lower numbers (Table 1). A Gene Ontology (GO) enrichment test on EV FSFs, however, showed that these were enriched in biological processes crucial for cellular development and regulation, such as GO:0048483 (autonomic

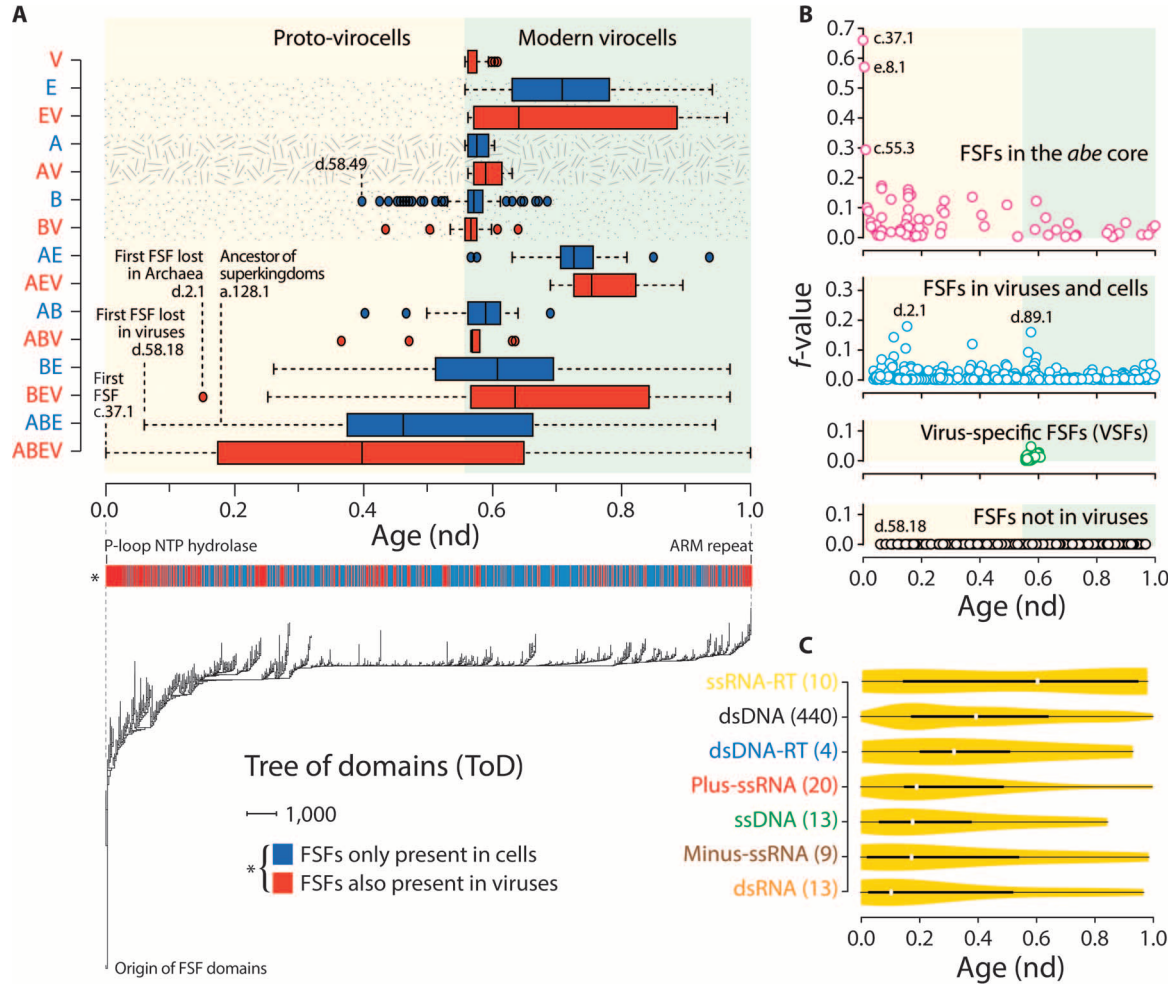


**Fig. 5. Phylogenomic analysis of FSF domains.** (**A**) ToD describe the evolution of 1995 FSF domains (taxa) in 5080 proteomes (characters) (tree length = 1,882,554; retention index = 0.74; $g_1 = -0.18$). The bar on top of ToD is a simple representation of how FSFs appeared in its branches, which correlates with their age (nd). FSFs were labeled blue for cell-only and red for those either shared with or unique to viruses. The boxplots identify the most ancient and derived Venn groups. Two major phases in the evolution of viruses are indicated in different background colors. Patterned area highlights the appearances of AV, BV, and EV soon after A, B, and E, respectively. FSFs are identified by SCOP css. (**B**) Viral FSFs plotted against their spread in viral proteomes (*f* value) and evolutionary time (nd). FSFs identified by SCOP css. (**C**) Distribution of ABEV FSFs in each viral subgroup along evolutionary time (nd). Numbers in parentheses indicate the total number of ABEV FSFs in each viral subgroup. White circles indicate group medians. Density trace is plotted symmetrically around the boxplots.

nervous system development), GO:0002062 (chondrocyte differentiation), and GO:0050921 (positive regulation of chemotaxis) (table S7). It is possible that this repertoire was provided to eukaryotes from viruses or was simply gained in eukaryoviruses from their eukaryotic hosts through HGT. In turn, no biological process was enriched in either AV or BV.

Next, we divided viral FSFs into four subgroups: (i) those shared between prokaryotic viruses and eukaryoviruses (that is, the *abe* core of Fig. 3B; table S5); (ii) other viral FSFs shared with cells (cyan circles); (iii) VSFs (green circles); and (iv) FSFs not detected in viral proteomes (black circles) (Fig. 5B). Generally, FSFs of the *abe* core were present in a greater number of viral proteomes (higher *f* values) and in more replicon types (fig. S3). Some of the most popular FSFs again included the P-loop containing NTP hydrolase (c.37.1), DNA/RNA polymerases (e.8.1), and Ribonuclease H–like (c.55.3) FSFs. In turn, FSFs shared with cells were relatively less widespread. However, the Lysozyme-like FSF (d.2.1) was detected in a large number of viruses (18%), mostly bacterioviruses. Lysozymes can penetrate bacterial peptidoglycan layers and facilitate viral entry. We speculate that this capability was transferred to eukaryotic cells from viruses to block bacterial infections in eukaryotes. Another relatively widespread FSF was the "Origin of replication binding domain, RBD-like" (d.89.1) FSF, which was detected in ~16% of the sampled viruses. Both the *abe* core and FSFs shared with cells spanned the entire nd axis. Thus, viral proteomes encode both very ancient and very derived FSFs. The former group was most likely inherited vertically from the common ancestor of cells and viruses, whereas the latter could be a result of recent HGT gains from cells or shared innovation. The enrichment of very ancient FSFs in the *abe* core present in viruses infecting the three superkingdoms provides strong support to their ancient origin. The origin of VSFs, on the other hand, marks the onset of modern virocell life cycles.

Results therefore highlight two important phases in viral evolution: (i) an early cell-like existence of viruses (the precursors of modern virocells) and (ii) a late transition to the viral mode, as we know it today.

### Proteomes of RNA viruses are more ancient than proteomes of DNA viruses

A series of experiments determined the relative age of each viral subgroup.

(i) Evidence from ToD. We zoomed into the most ancient core ABEV Venn group and separated FSFs belonging to each of the seven viral replicon types (Fig. 5C). In all viruses, regardless of the replicon type, median nd values were very low (see white circles), indicating that they shared ancient FSFs with cells. Likewise, each viral subgroup had a longer tail toward the right, suggesting that HGT may have played evolutionary roles only very recently. The most ancient ABEV repertoires were derived from dsRNA, minus-ssRNA, and ssDNA viruses, suggesting that they predated dsDNA viruses in evolution.

(ii) Ariadne's threads traced in evolutionary time. We traced FSF domain ages onto the threads of FSFs shared between viral subgroups (Fig. 6A). The oldest domains were spread in a transect that unified minus-ssRNA, plus-ssRNA, and dsRNA proteomes. This pattern was clearly evident in violin plots that describe FSF age in the threads along the early timeline of domain evolution (nd < 0.3) (Fig. 6B). Once again, the proteomes of minus-ssRNA viruses were particularly enriched in ancient domains, suggesting that perhaps ssRNA was involved in virocell origins (read below).

(iii) Evidence from trees of proteomes (ToP). To describe the evolutionary relationships between the proteomes of cells and the proteomes of viruses (taxa), we reconstructed ToP from the abundance and occurrence of 442 ABEV FSFs (phylogenetic characters). The ABEV group was selected because it includes many FSFs of ancient origin (median nd ~0.4; Fig. 5A), the entire *abe* core (Fig. 3B and table S5), and ancient FSFs in Ariadne's threads (Figs. 4B and 6A). Because biases in taxon sampling could influence tree reconstruction, we randomly sampled a set of 368 proteomes (taxa) from cells and viruses, including up to 5 viral species from each viral order or family and 34 proteomes corresponding to only free-living organisms in Archaea, Bacteria, and Eukarya. The rooted phylogeny dissected proteomes into four supergroups (Fig. 7A). Viruses formed a distinct paraphyletic group at the base of ToP that was distinguishable from cells by 76% bootstrap (BS) support. In turn, archaeal organisms were clustered paraphyletically in the more basal branches (black circles), whereas Bacteria and Eukarya formed monophyletic groups (blue and green circles) supported by 66% and 100% BS, respectively (Fig. 7A). This topology supported an ancient origin of both viruses and Archaea and a sister relationship between Bacteria and Eukarya, which goes against some gene sequence–based phylogenies (*58*–*60*) but is congruent with a number of structure- and function-based studies [discussed elsewhere (*18*, *61*–*65*)].

The most basal taxa corresponded to RNA and retrotranscribing viruses. These included well-known dsRNA viral families that have segmented genomes such as *Birnaviridae*, *Partitiviridae*, and *Picobirnaviridae* (2 segments); *Chrysoviridae* and *Quadriviridae* (4 segments); and *Reoviridae* (10 to 12 segments). *Nodaviridae* that have bipartite genomes (that is, two segments) and "capsidless" *Narnaviridae* (both plus-ssRNA viruses) also occupied the most basal positions in ToP along with dsRNA and dsDNA-RT viruses. Other very ancient viral groups included retrotranscribing viruses (*Caulimoviridae*, *Hepadnaviridae*, and *Retroviridae*), ssDNA viruses (*Anelloviridae* and *Inoviridae*), dsDNA viruses (*Plasmaviridae* and *Polydnaviridae*), ambisense arenaviruses, and minus-sense influenza viruses (Fig. 7A). It has been hypothesized that retrotranscribing viruses likely mediated the transition from an ancient RNA world to the modern DNA world (*66*). Remarkably, in our tree, retrotranscribing viruses originated before bona fide DNA viruses, validating the hypothesis (Fig. 7A). Another interesting position was that of polydnaviruses, which are "symbionts" of endoparasitic wasps (*67*). These viruses also encode segmented dsDNA genomes. These observations suggest the presence of segmented viral genomes (mostly RNA) in ancient cells and the late appearance of "capsid-encoding" and DNA viruses.

A tree of viruses (ToV) reconstructed from *abe* core FSFs (fig. S4) further confirmed an early origin in RNA viruses. Although the distribution patterns of replicon types were not entirely clear-cut, there was clear enrichment of RNA viral proteomes at the base of ToP, specifically minus-ssRNA and dsRNA viruses. This tree was poorly resolved partly as a result of the limited number of phylogenetic characters that were used to distinguish proteomes and largely as a result of the patchy distribution of *abe* FSFs in viral proteomes (a consequence of reductive evolution in viruses). Finally, grouping viruses by host type (that is, archaeoviruses, bacterioviruses, and eukaryoviruses) did not yield three

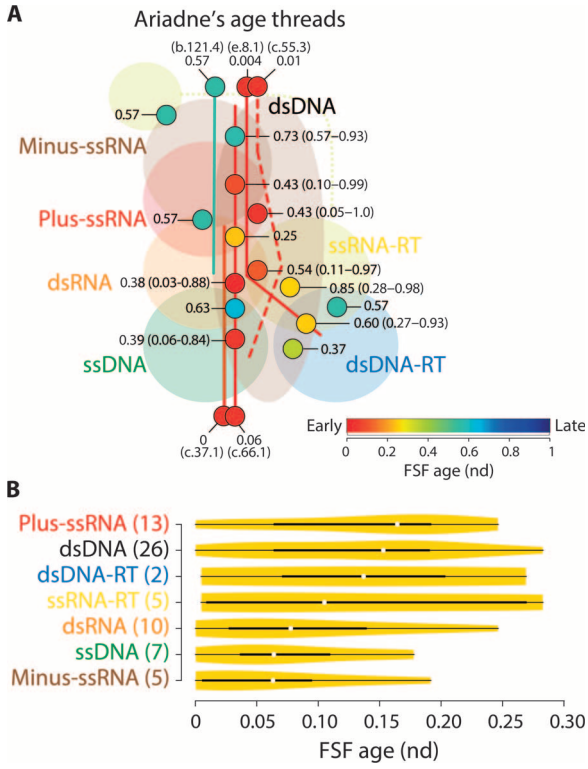


**Fig. 6. Ancient history of RNA viral proteomes.** (**A**) The length of Ariadne's threads (colored lines) identifies FSFs that were shared by more than three viral subgroups. Filled circles indicate FSFs shared between two or three viral subgroups. Numbers next to each circle give the mean nd of FSFs shared by each combination. Numbers in parentheses give the range between the most ancient and the most recent FSFs that were shared by each combination. (**B**) Distribution of the most ancient (nd < 0.3) ABEV FSFs in evolutionary timeline (nd) for each viral subgroup. Numbers in parentheses indicate the total FSFs in each viral subgroup. White circles indicate group medians. A density trace is plotted symmetrically around the boxplots.

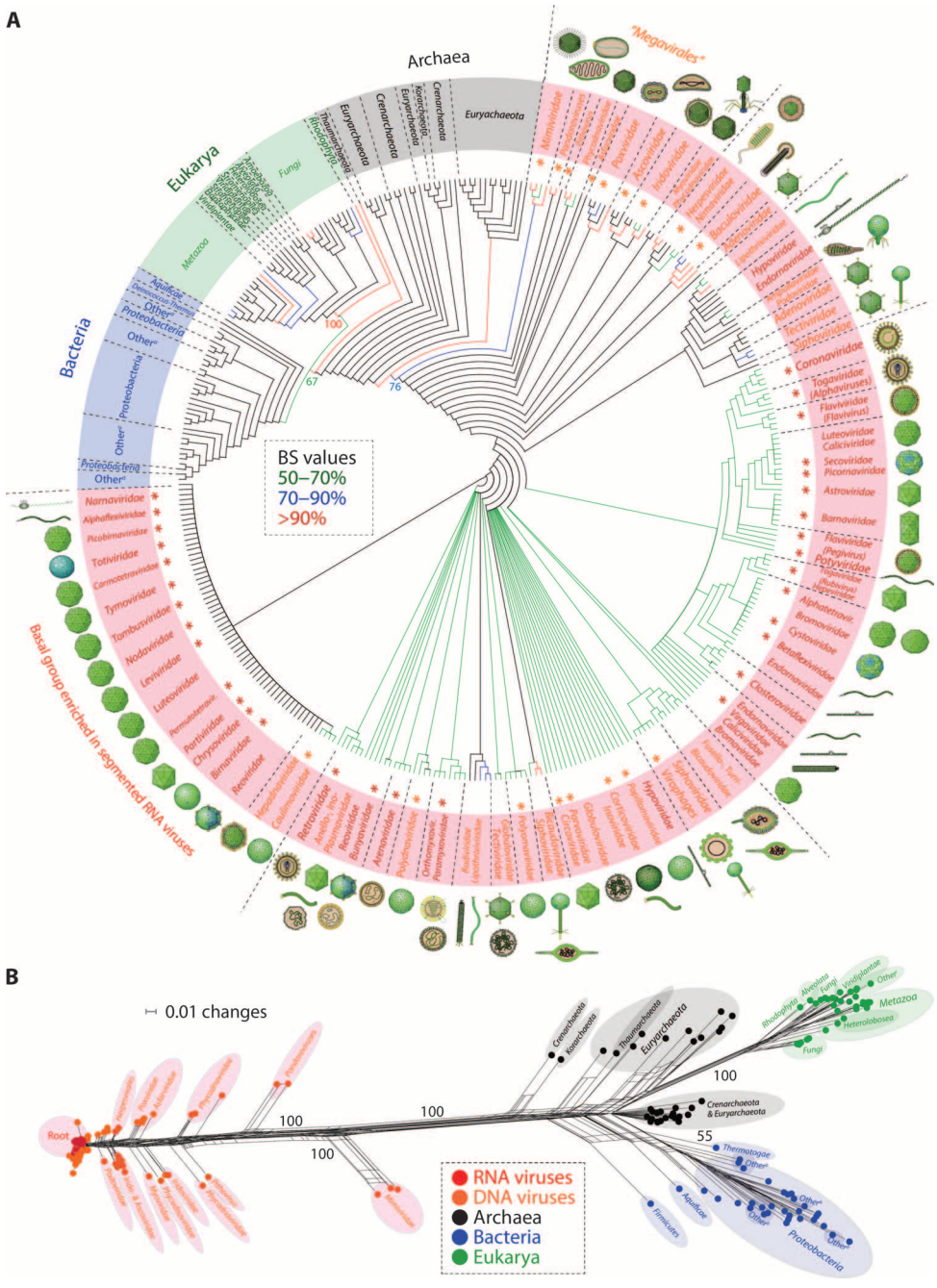


**Fig. 7. Evolutionary relationships between cells and viruses.** (**A**) ToP describing the evolution of 368 proteomes (taxa) that were randomly sampled from cells and viruses and were distinguished by the abundance of 442 ABEV FSFs (characters) (tree length = 45,935; retention index = 0.83; $g_1$ = −0.31). All characters were parsimony informative. Differently colored branches represent BS support values. Major groups are identified. Viral genera names are given inside parentheses. The viral order "Megavirales" is awaiting approval by the ICTV and hence written inside quotes. Viral families that form largely unified or monophyletic groups are labeled with an asterisk. Virion morphotypes were mapped to ToP and illustrated with images from the ViralZone Web resource (*131*). No picture was available for *Turriviridae*. [a]Actinobacteria, Bacteroidetes/Chlorobi, Chloroflexi, Cyanobacteria, Fibrobacter, Firmicutes, Planctomycetes, and Thermotogae. (**B**) A distance-based phylogenomic network reconstructed from the occurrence of 442 ABEV FSFs in randomly sampled 368 proteomes (uncorrected *P* distance; equal angle; least-squares fit = 99.46). Numbers on branches indicate BS support values. Taxa were colored for easy visualization. Important groups are labeled. [b]Actinobacteria, Bacteroidetes/Chlorobi, Chloroflexi, Cyanobacteria, Deinococcus-Thermus, Fibrobacter, Firmicutes, and Planctomycetes. [c]Amoebozoa Chromalveolata.

independent groups, suggesting that viruses, regardless of host type, could be structurally (and evolutionarily) more related to each other (fig. S5). It also suggests that viruses can jump hosts [for example, severe acute respiratory syndrome (SARS) and Ebola viruses, loss of RNA viruses in prokaryotes (*44*)], and thus, evolutionary relationships based on virus-host preferences should be considered with caution [sensu (*43*)].

We evaluated ToP phylogeny by comparing it against ICTV and structure-based classifications. ToP recovered some well-known relationships. For example, the genera *Flavivirus* (*Flaviviridae*) and *Alphavirus* (*Togaviridae*) were grouped together, suggesting their close evolutionary association (66% BS). In fact, alphaviruses were initially classified by the ICTV under *Flaviviridae* but were later assigned their own genera in *Togaviridae*. Both viral families show striking similarities in virion architecture (enveloped and spherical) and genome replication strategies (monopartite linear plus-ssRNA). Similarly, *Polyomaviridae*, *Closteroviridae*, *Coronaviridae*, and many others also formed individual monophyletic groups (indicated by asterisk in Fig. 7A). Another largely unified group was that of filamentous dsDNA archaeoviruses (*Rudiviridae* and *Lipothrixiviridae*) that have been classified under the order "Ligamenvirales" (*68*). Similarly, viral families in the "nucleocytoplasmic large DNA viruses" group (*Poxviridae*, *Phycodnaviridae*, *Ascoviridae*, *Asfarviridae*, *Iridoviridae*, and *Mimiviridae*) formed a paraphyletic group at the very derived positions. This group also included the recently discovered pandoraviruses (*69*) and pithoviruses (*70*) and the oddly placed bacteriophage (*Myoviridae*). The close grouping of all giant viruses supports the proposal of the viral order "Megavirales" (*71*) and a previous reconstruction (*19*). However, herpesviruses and *Caudovirales* that share the common HK97 capsid protein fold did not form a single group (*47*), but they were in close proximity (Fig. 7A). In turn, *Adenoviridae* and *Tectiviridae* that belong to the PRD1/adenovirus-like lineage were closely clustered. Similarly, *Totiviridae* and some *Reoviridae* of the BTV-like lineage occupied basal positions. Some members of the picornavirus-like lineage (for example, *Luteoviridae*, *Caliciviridae*, and *Picornaviridae*) and retrotranscribing viruses also clustered together, but clear-cut structure-based viral lineages did not materialize in ToP. Other discrepancies also existed with regard to viral families defined by the ICTV that did not form unified groups. However, ICTV classifications are subject to revisions and do not always yield evolutionarily informative classifications. In light of these, ToP reconstructed from the abundance of conserved FSF domains present a "third" and global view of the evolutionary relationships of viruses, which adds deep lineage relationships to ICTV and structure-based classifications.

ToP also provide interesting information regarding the evolution of virion morphotypes. Most basal branches were populated by spherical or filamentous virions (two of the simplest designs from the point of view of tensegrity). They gradually become more decorated, with additional features such as spikes and glycoproteins (retroviruses) in spherical virions, and rod-like designs (inoviruses) likely evolving from filamentous versions. Perhaps the rods and spheres combine to form the head-tail morphotype so abundant in prokaryotic viruses. Thus, mapping of virion morphotypes onto ToP likely hints toward the origin of viruses from a limited number of structural designs (*43*). However, we caution that morphological similarities may also stem from convergent evolution. At this point, we lack evidence to confirm homologies between different virion morphotypes. Nevertheless, the early appearance of spherical and filamentous virions harboring segmented RNA genomes is remarkable and worthy of further attention.

(iv) Evidence from distance-based networks. Typically, viral proteomes encode far less proteins and in lower abundance relative to the proteomes of cellular organisms (except for some giant viruses). To account for such differences and to test whether the phylogeny in Fig. 7A was not influenced either by HGT or by technical artifacts associated with our choice of phylogenetic model, we used FSF occurrence in distance-based networks reconstruction (Fig. 7B). The resulting topology still favored a "tree-like" structure (Fig. 7B), suggesting that the phylogeny of Fig. 7A was not influenced by processes that could artificially increase genomic abundance. Moreover, none of the viral proteomes clustered with their hosts (for example, plant RNA viruses did not group with plants), indicating that the predicted cellular nature of viruses was not attributable to HGT from their hosts but was likely a result of ancient coexistence. The phylogenomic network retained most evolutionary relationships defined earlier by ToP but also recovered a closer grouping of herpesviruses with *Podoviridae* (*Caudovirales*) that was not so clear in ToP derived from genomic abundance, supporting the proposal that the two viral groups are closely related (*47*, *72*).

(v) ToP derived directly from the age of protein domains. We also developed a multidimensional scaling approach to study the evolution of cells and viruses: the evolutionary principal coordinate (evoPCO) analysis (Fig. 8A). The evoPCO method combines the power of cladistic and phenetic approaches by calculating principal coordinates directly from temporal evolutionary distances between the proteomes of species (see Materials and Methods). The distance between proteomes reflects phylogenetic dissimilarity in the age of FSF domain repertoires (that is, nd values) and can be displayed in 3D temporal space, assuming that the age of an FSF is the age of the first instance of that FSF appearing in evolution. Because proteomes are biological systems that are made up of component parts (that is, FSFs in this case) but describe cellular organisms and viruses, each component (regardless of its abundance) contributes an age to the overall age of the cellular or viral system. This factor, when taken into account, results in a powerful projection of a multidimensional space of proteomes onto a 3D temporal space that allows visualization of evolutionary relationships.

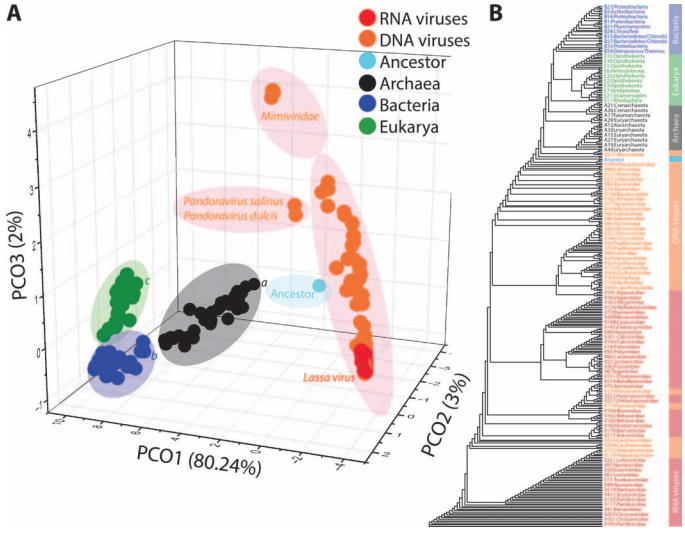


**Fig. 8. Evolutionary history of proteomes inferred from numerical analysis.** (**A**) Plot of the first three axes of evoPCO portrays evolutionary distances between cellular and viral proteomes. The percentage of variability explained by each coordinate is given in parentheses on each axis. The proteome of the last common ancestor of modern cells (*57*) was added as an additional sample to infer the direction of evolutionary splits. [a]*Ignicoccus hospitalis*, [b]*Lactobacillus delbrueckii*, [c]*Caenorhabditis elegans*. (**B**) A distance-based NJ tree reconstructed from the occurrence of 442 ABEV FSFs in randomly sampled 368 proteomes. Each taxon was given a unique tree ID (tables S1 and S2). Taxa were colored for quick visualization.

The evoPCO method revealed four clouds of proteomes in temporal space that correspond to viruses and to the three cellular superkingdoms (Fig. 8A). The first three coordinates explained ~85% of the total variability. Using the previously reconstructed proteome of the last common ancestor of modern cells as reference point (*57*), we inferred viruses as the most ancient supergroup, followed by Archaea, Bacteria, and Eukarya, in that order (Fig. 8A). This topology supports earlier results from comparative genomic and phylogenomic analyses, adding a fifth line of evidence in support of the early origin of viruses. Remarkably, *Lassa virus*, which belongs to *Arenaviridae* and harbors segmented RNA genomes, appeared at the most basal position of the evoPCO plot, supporting the early origin of segmented RNA viruses in ToP (Fig. 7A). Some giant viruses appeared more derived, supporting their ancient coexistence with cells (*19*, *73*). The topology and ordering of proteomes in evoPCO analysis were further supported by a distance-based neighbor-joining (NJ) tree (Fig. 8B) reconstructed directly from the temporal distance matrix, which retained the cohesive and ancient nature of the viral supergroup. The NJ tree made explicit the early origins of RNA viral families and was largely congruent with ToP recovered earlier (Fig. 7A), validating the power of the evoPCO strategy.

## DISCUSSION

### Viruses merit inclusion in the tree of life

The search for a "fourth" domain of life is not new [for example, (*74*, *75*)]. It has been the subject of intense debate in evolutionary biology [refer to (*76*–*83*) and references therein]. Here, we put forth the bold conjecture of a universal tree of life (uToL) that describes the evolution of cellular and viral proteomes (Figs. 7 and 8). Formally placing viruses in uToL is a daring task because many scientists even question whether viruses are living entities mainly because of (i) the lack of true viral metabolism and (ii) their inability to reproduce on their own (*76*, *84*). However, counterarguments have recently gained popularity, especially inspired by the study of "virus factories," which are intracellular structures formed by many giant viruses inside infected cells (*85*). Virus factories are "cell-like organisms" [sensu (*86*)] that are compartmentalized by a membrane, have ribosomes, obtain energy from mitochondria, and contain full information to successfully produce numerous virions (*85*). They are strikingly similar to many intracellular parasitic bacteria that also depend on host metabolism to reproduce. For these reasons, it has been argued that the true "self" of a virus is the intracellular virus factory of infected cells, which is metabolically active and should be contrasted with the extracellular and metabolically inert virion state. Specifically, virocells produce viral gametes (virions) that are functionally analogous to cellular gametes of sexually reproducing species, which fuse during fertilization. These viral gametes can then fertilize (infect) other cells [sensu (*86*)]. Thus, viruses should be considered "living" organisms that simply survive by means of an atypical reproduction method that requires infecting a cell [similar to obligate parasitism (*87*)].

The argument that viruses do not replicate or evolve independently of cells and hence should not be deemed worthy of living status (*84*) has been toned down because each species replicates and evolves in nature and requires coexistence with other life forms (*87*). In short, there is a need to broaden our definitions of "life" and to abandon viewing virions as viruses [sensu (*86*)]. In light of these arguments, we contend that it is legitimate to study viral origins and evolution on a scale comparable to that of Archaea, Bacteria, and Eukarya and to ask fundamental questions related to the evolutionary history of cells and viruses. Here, we propose that the encoded genetic makeup and its ancient history define the functionalities of virions, capsids, and replicons that are necessary to complete the reproduction cycle of a virus. We therefore study the proteomic composition of viral replicons to infer viral evolutionary trajectories (similar to how phylogenetic analysis of cellular genes tells us about their history), presenting a hypothesis of virus origin and evolution that is more compatible with virus biology and large-scale molecular data.

### Ancient cellular origin of viruses by reductive evolution

Our comparative and phylogenomic data refute the "virus-first" hypothesis or the precellular scenario for the origin of viruses [see (*80*) for a new version]. This proposal suggests an early origin of self-replicating viral replicons predating the origin of cells. However, the hypothesis is unsatisfactory because viruses are tightly associated with proteins (capsids) and must replicate in an intracellular environment to produce viral progeny. Fossil evidence also shows that cells appeared early in evolution (*88*, *89*). The large size of the ABEV group, which includes many FSFs of ancient origin and membrane proteins, is also incompatible with the virus-first scenario and suggests an ancient cellular origin of viruses (Figs. 1A and 5). Thus, our data can be better reconciled with either the "escape" hypothesis or the "reductive" hypothesis [see (*31*, *41*, *90*) for details], both of which associate viral origins with cells. Under the escape hypothesis, replicons in proto-virocells became autonomous and acquired virions to infect other ancient cells. In turn, the reduction hypothesis suggests loss of the primordial ribosomal machinery in proto-virocells and subsequent reduction into viruses. Although both hypotheses explain the origin of viruses in ancient cells, reduction seems to be more parsimonious with our data given the strong lifestyle resemblance of viruses to cellular parasites (that also evolve similarly) (*37*), the early loss of FSFs suggested by evolutionary timelines (Fig. 5A), and the discovery of giant viruses that overlap cellular parasites in genomic and physical features (Fig. 1, C and D) (*69*, *70*, *91*, *92*). Although reduction of modern cells into virions (enclosing few proteins) may seem far-fetched, this would be relatively more straightforward in ancient cells where ribosomal machinery and other cellular features were yet to fully materialize.

From the point of view of natural history, our bold conjecture simply invokes the existence of proto-virocells—additional cellular descendants of the last universal ancestor of both modern cells and viruses. The proto-virocells reduced into modern viruses, whereas their siblings diversified into Archaea, Bacteria, and Eukarya. It is important to distinguish proto-virocells from modern-day virocells. Although infection of modern-day virocells may result in virion synthesis and cell lysis (*22*), the proto-virocell genomes coexisted in the intracellular environment and reproduced either without lysis (similar to endogenous viruses in cellular genomes or plasmids trapped in cells) or by producing primitive forms of virions. In the absence of the jelly-roll fold and other capsid-associated viral folds (which appeared quite late in our timelines), primitive capsid-like structures could have been built from folds seen in prokaryotic protein nanocompartments or by formation and secretion of membrane vesicles. The prokaryotic protein compartments (such as encapsulins and carboxysomes) are polyhedral protein shells that are morphologically similar to icosahedral viruses [for example, (*48*)]. Modern-day viral capsids store nucleic acids, whereas prokaryotic protein compartments store enzymes. Perhaps the switch from storing proteins to storing nucleic acids facilitated

viral origins in an ancient cell (*48*). In turn, modern cells frequently secrete membrane vesicles to communicate with other cells. These vesicles are also morphologically similar to spherical viruses and can package viral genes and contribute toward viral infection (*93*). This vesicle secretion phenomenon is very ancient and could have played roles in the origins of ancient viruses [see (*94*, *95*) for other vesicle-related scenarios of viral origins]. Both scenarios explain how virions were synthesized in proto-virocells to export viral genetic information. Under this scenario, plasmids and other selfish genetic elements also originated from proto-virocells but did not acquire capsids and remained tightly integrated with the emerging ribocellular makeup.

Our phylogenomic reconstructions also show that proto-virocells initially harbored segmented RNA genomes and that proteomes of all seven kinds of viral replicons were enriched in ancient FSFs (Fig. 5C). Given the massive diversity in replicon type seen in modern viruses, it is likely that all kinds of replication strategies were used in proto-virocells. A logical outcome of this experimentation would be the discovery of many key replication-associated proteins and perhaps DNA itself in the virus world [an idea previously put forward by Forterre (*66*, *96*)].

In summary, the virus-mediated infection of (ancient or modern) cells is an old process that has evolved gradually over billions of years but materialized fully once lineages of organisms diversified. After the proto-virocells reduced to completely lose their cellular nature and acquired types of folds to build viral capsids (jelly-roll and other forms), modern viruses were born. The cellular nature of viruses is restored when modern viruses infect and replicate inside modern cells or when they become integral parts of their genomes. Whether the infection is modern or ancient, it has two interesting consequences: (i) viruses gain complete access to and control of the cellular machinery to create genes and protein folds, thus explaining the large size of class I proteins and VSFs (Fig. 1, A and B), and (ii) they can also pick old or newly created genes from cells, spreading them to other cellular lineages through virus-to-cell gene transfer (Fig. 2), if it provides selective advantages to the evolving host. Thus, our model, which explains the evolution of virocells, is biphasic and identifies an early cell-like phase followed by the emergence of modern-day viral lineages.

### Primacy of virus-to-cell gene transfer

It has been argued that viruses frequently pickpocket genes from cells and that this phenomenon explains their primary mode of evolution (*76*). However, our data and previous genomic analyses (*97*–*100*) strongly refute this idea and have revealed the abundance of unique genes (that is, class I proteins in Fig. 1B) in viral proteomes lacking cellular homologs. A large number of these proteins are likely very ancient and thus are no longer detectable either by BLAST or HMM-based searches, whereas the remaining proteins probably originated rather recently in viral lineages (for example, VSFs in Fig. 5A). In fact, genes are continuously created by viral lineages in infected cells during viral replication cycles, when viruses have full access to the cellular machinery to produce genes (*34*). This phenomenon has been greatly underestimated in the past but is now being acknowledged [for example, (*5*, *90*, *97*, *100*–*102*)]. Discovery of viruses from atypical habitats and hosts is expected to improve the Protein Data Bank (PDB) representation of viral structures and will no doubt increase our knowledge about class I proteins. In turn, alternative explanations, such as the rapid evolution of class I proteins in viruses after uptake from cells or acquisition from yet-to-be-discovered cellular species, are less satisfactory and account for only a minor fraction of class I proteins. For example, the former scenario is inconsistent with the presence of class II proteins that, surprisingly, remained robust to fast evolution in the same viral proteomes (*42*). Moreover, synonymous-to-nonsynonymous substitution rates for "unique" genes in giant DNA viruses did not vary significantly from the substitution rates of vertebrate proteins (*97*). In turn, the latter scenario posits a decrease in the number of VSFs with the sampling of more cellular genomes, which has not been observed [for example, comparison of (*19*) and the present study; also discussed in (*49*)]. Together, the major fraction of viral proteomes includes proteins with no detectable cellular homologs. This subset is likely indicative of the gene-creation abilities of viruses during the virocell life cycles.

In turn, there are three possible routes to explain the origin of class II proteins: (i) they originated in cells and transferred to viruses; (ii) they originated in viruses and transferred to cells; and (iii) they spread vertically from the common ancestor of cells and viruses or by means of shared innovation. Each of the three possibilities has known examples. However, the gene flow from virus to cell numerically exceeds the gene flow in the opposite direction [for example, transfer of the RNA polymerase gene of dsRNA viruses to eukaryotes (*103*), provirus genes integrated in mitochondria (*104*), endogenization of viruses (*105*), and syncytin protein involved in mammalian placenta development (*106*); see also (*22*)]. Similarly, Cortez *et al.* (*100*) showed that a significant fraction of archaeal and bacterial genes (~15 to 20%) are of foreign origin, most likely inherited from viruses or plasmids. This is not surprising given the abundance of integrated viral-like elements in eukaryotic genomes (for example, ~50% in humans). Our discovery of 66 bona fide VSFs and 43 additional VSFs that were hidden in cellular proteomes (Table 1) is additional support for this argument. Together, we argue that the gene-creation and gene-transfer abilities of viruses have been significantly underestimated by some authors [for example, (*76*)]. We falsify the idea that viral genomes only evolve by acquiring genes from host species.

### Early origin of segmented RNA viruses

The very basal viral groups in the uToL exemplified by ToP, ToV, and evoPCO reconstructions (Figs. 7A and 8 and fig. S4) included minus-ssRNA viruses and families of dsRNA viruses that harbor segmented genomes. This is remarkable. The influenza virus genome typically contains six to eight RNA segments and evolves by random genetic drift or by the reassortment of genome segments with other coinfecting influenza viruses. Thus, it is likely that proto-virocells had segmented RNA genomes that often "mated" by combining with other RNA segments. This is compatible with the proposal of Woese (*107*) that the earliest cells stored genes in the form of segmented RNAs. These findings support the general idea that RNA came before DNA. The ubiquity of the use of RNA primers in the synthesis of DNA and its deoxyribonucleotide precursors from ribonucleotide precursors of RNA (*108*) is additional support for this argument. The principle of continuity dictates that a possible shift from RNA to DNA was gradual and was likely mediated by retrotranscribing viruses (for example, Fig. 7A).

### Polyphyletic origin of diversified viruses

The uToL supports a polyphyletic origin for viruses. At least two kinds of virions (spherical and filamentous), both apparently unrelated, were the likely precursors of many complex virion morphotypes (Fig. 7A). The viral mode of life therefore originated more than once in evolution but always before the divergence of modern cells from a cellular stem line of descent. In turn, the support for a monophyletic origin is weak

but could be explained by reductive evolution, which is expected to confound the evolutionary patterns, especially if a long time has passed. The seven-set Venn diagram revealed a highly patchy distribution of FSFs in the viral supergroup. Although no FSF was shared by all seven viral subgroups, some ancient FSFs were shared by four to six viral subgroups including mostly RNA viruses. The data also identified a large core of *abe* FSFs that were shared by viruses infecting the three superkingdoms. Together, the most parsimonious explanation for structural data confirms an ancient cellular history of viruses and their origin from one or more virocell ancestors.

### Rise of the diversified cellular world

The early polyphyletic origin of RNA viruses is relevant to the discussion of the origin of eukaryotes and (especially) eukaryoviruses. It was recently proposed that eukaryotes originated either by "fusion" of two prokaryotic cells [for example, archaeon and bacterium (*109*, *110*)] or from a subgroup of Archaea [archaeal ancestor scenario (*111*–*113*)]. These scenarios logically constrain the origin of eukaryoviruses from the merging of prokaryotic viruses, as claimed by Koonin *et al.* (*45*). Here, we question these scenarios and argue that they are less parsimonious with our data and other observations [see also (*46*)].

First, both scenarios postulate a transition of one domain of life into another, which is incompatible with the membrane composition of domains of life and with their biochemical differences. For example, archaeal membrane lipids are different from bacterial and eukaryal membranes, which are perhaps better suited to the extreme ecological niche of archaeal species. Thus, fusion or archaeal-ancestor scenarios posit the transition of archaeal membrane into bacterial/eukaryal membrane, an event that has never been observed in archaeal lineages despite several documented episodes of the HGT of genes from Bacteria to Archaea (*114*). Second, both scenarios also posit the accelerated appearance of a remarkable number of eukaryote-specific folds (283 E FSFs; Fig. 1A), including several aspects of eukaryotic cellular biology that differ starkly with prokaryotes (for example, replacement of prokaryotic-like division by mitosis and decoupling of transcription and translation). Third, and perhaps most importantly, the origin of eukaryoviruses from prokaryotic viruses is less likely because many families of eukaryoviruses have no counterpart in either Archaea or Bacteria [for example, Fig. 3A; see also (*43*, *44*)].

If the host of the fusion event was an archaeon or if eukaryotes branched off from some archaeal phyla, one should expect eukaryoviruses to resemble archaeoviruses at the molecular and/or phenotypic (that is, virion architecture) levels and to recognize archaeal membranes for infectivity. This is clearly not the case. Only two FSFs were shared by archaeoviruses and eukaryoviruses involved in DNA replication/repair (a.60.6) and metabolism (c.76.1), both apparently unrelated to viral pathogenicity. Similarly, we recently compared the virion morphotype distribution in viruses and discovered that two morphotypes (rod-shaped and bacilliform) were unique to archaeoviruses and eukaryoviruses (*44*). However, close examination of the 3D folds of coat proteins of "rod-shaped" viruses does not suggest a common origin (*68*, *115*), and the same is probably also true for the "baciliform" morphotype (*44*). Moreover, member viruses of the two morphotypes did not cluster together in our ToP (Fig. 7A), suggesting that the observed phenotypic resemblance is more likely a result of convergence than divergence. The distribution and abundance of RNA viruses in eukaryotes are in disagreement with the aforementioned scenarios because of the paucity of RNA viruses in prokaryotes. Although RNA viruses are abundant in the superkingdom Eukarya (for example, dsRNA in Fungi, RNA in Plants, and retrotranscribing viruses in Plants and Metazoa), no RNA viruses are now known in Archaea (and are rare in Bacteria). Although Bolduc *et al.* (*116*) isolated putative RNA viruses from a metagenomic sample rich in archaeal organisms, their host tropism could not be established with confidence and they considered contamination as an alternative but unlikely explanation. These observations greatly reduce confidence in the emergence of eukaryotic RNA viruses from viruses of a prokaryotic ancestor and point toward an alternative scenario for the origin of modern cells that is linked to differential selection of the virosphere [see also (*43*, *44*, *46*)]. Because several archaeal members are characterized by thermophilic habitats and RNA is unstable at high temperatures, the apparent bias in the distribution of viruses in cellular superkingdoms (that is, DNA viruses in prokaryotes and RNA in eukaryotes; Fig. 3A) is better explained by early loss of RNA viruses in Archaea when they migrated to harsh temperatures. Perhaps it was a driving force behind this transition (*46*). Archaeal viruses, plasmids, insertion sequences, and antiviral defense (clustered regularly interspaced short palindromic repeats) closely resemble the mobilome and defense system in Bacteria. In turn, plasmids are rare in eukaryotes, and more sophisticated defense systems (mediated by small interfering RNA) are used against invading viruses (*46*). Even the archaeal member viruses of the PRD1/adenovirus-like and HK97-like lineages are more similar to bacterioviruses than to eukaryoviruses (*117*). This suggests that both Archaea and Bacteria have experienced similar selection pressure to get rid of RNA viruses early in evolution (*46*). Although Archaea migrated to warm temperatures to escape RNA viruses, the development of a thick peptidoglycan-containing cell wall in Bacteria likely blocked the entry of many viral families (*34*, *118*). In turn, Eukarya likely benefited from the interaction with RNA and retroviruses (triggering genomic rearrangements) and evolved toward complexity, as recently discussed by Forterre (*46*). In light of these arguments and our FSF data, the early origin of RNA viruses is a significant event in the history of life that triggered major evolutionary trends in coexisting cellular lineages and (perhaps) also led to the discovery of DNA genomes in ancient cells (through retrotranscribing viruses) (*66*).

### Limitations and some technical considerations

Our conclusions rely on the accuracy of HMMs to detect FSF domains in protein sequences and on the current definitions of SCOP, the influential gold standard in protein classification. The structural census could be the subject of biases in the genomes that have been sequenced so far and in our ability to appropriately survey viral and cellular biodiversity. In particular, there is a strong ascertainment bias toward the discovery and study of plant, vertebrate, and human viruses because of economic and medical reasons. Thus, the current picture of viruses and their hosts is largely incomplete.

We also stress that we focused on protein domain structure and not on protein sequence. We therefore avoided the time-erasing effect of mutations and the confounding convergent effects of historical patchworks present in multidomain protein sequences, which represent a substantial fraction of every proteome that has been sequenced (*16*). A global analysis involving both viral and cellular proteomes is perhaps only possible by focusing on domain structure and molecular function—characters that are relatively more conserved in evolution than gene sequence (*16*). Thus, our analysis provides an evolutionarily deep "structural" view that, as expected, is not always in line with the shallow "sequence" view of viral evolution. This fact should be taken into consideration when interpreting our results.

Although phylogenomic reconstructions depend on the choice of phylogenetic model and search strategy for optimal trees, our experience with these methodologies has shown that, in general, phylogenetic reconstructions are reliable. It can be argued that some ancestral FSFs were lost from our census because of historical bottlenecks that have characterized cellular evolution. To minimize such effects, we built trees of life from only ABEV FSFs that were present in all sampled groups with relatively higher abundance. Moreover, we have previously linked the evolutionary age of each FSF to its geological age (*56*). Loss of some ancestral FSFs as a result of extinction events would likely distort the molecular clock of protein folds. But no such distortions have been observed, suggesting that character sampling for phylogenetic studies is reliable. Moreover, occurrence- and abundance-based analyses provided largely congruent results, suggesting that both parameters of the structural census carry similar signatures of the evolutionary process (see text S1 for a discussion of the choice of the phylogenetic methods used in this study). We therefore assume that retrodiction statements are not biased by preconceptions of modernity in the extant features studied.

Finally, we stress that our conclusions are the "most likely" and "most parsimonious" scenarios inferred from both comparative genomic (for example, Venn diagrams and *f* values) and phylogenomic approaches (ToD, ToP, and evoPCO). Studying viral and cellular evolution is a difficult problem complicated by many logical and technical considerations. In light of these, we hope that our study will initiate further discussions of this topic and that a consensus regarding viral evolution will be reached in the near future to benefit both viral biology and taxonomy.

## MATERIALS AND METHODS

### Data retrieval

Viral protein sequences were retrieved from the National Center for Biotechnology Information Viral Genomes Project (June 2014) (*119*). A total of 190,610 viral proteins corresponded to proteomes of 3966 viruses. For simplicity, unclassified and unassigned phages and viruses, and deltaviruses that require helper coviruses to replicate in host tissues (for example, Hepatitis delta virus) were excluded from the analysis. Viral proteomes were scanned against SUPERFAMILY HMMs (*20*) to detect significant SCOP FSF domains ($E < 0.0001$). Proteomes with no hits were further excluded from the analysis. This yielded a final viral data set of 3460 viral proteomes. In turn, FSF assignments for 10,930,447 proteins in 1620 cellular organisms were directly retrieved from the local installation of the SUPERFAMILY MySQL database (release July 2014; version 1.75). A total repertoire of 1995 significant FSF domains were detected in the entire set of 5080 proteomes.

### Maximum parsimony tree reconstruction

Phylogenomic analysis was carried out as previously described (*12*, *120*). Specifically, we calculated the abundance (that is, the total count) of each FSF in every proteome. Raw abundance values were log-transformed and rescaled to ensure compatibility with PAUP* (version 4.0b10) (*121*). For example, the raw abundance value of FSF *a* in proteome *b* was log-transformed ($g_{ab}$) and divided by the maximum abundance value in that proteome ($g_{ab\_max}$). This was done for each FSF in every proteome. The transformed matrix was then rescaled from 0 to 23 to yield 24 possible character states for use in PAUP*

$$g_{ab\_normal} = \text{round}[\ln(g_{ab} + 1)/\ln(g_{ab\_max} + 1) \times 23].$$

Normalization and rescaling ensure compatibility with PAUP* and protect against the effects of unequal proteome sizes and variances. Maximum parsimony (MP) was used to reconstruct ToD and ToP. ToD described the evolution of FSF domains (taxa) using proteomes as characters. In turn, ToP resembled conventional phylogenies that described the evolution of proteomes (taxa) using FSF domain characters. Trees were rooted by the method of Lundberg (*122*), which does not require specification of any outgroup taxon. Instead, first, an unrooted network is calculated, which is rooted a posteriori by the branch yielding a minimum increase in tree length. For this purpose, ancestral character states were specified using the ANCSTATES command in PAUP*. ToD were polarized by the maximum character state, assuming that the more abundant and widespread FSFs should be more ancient relative to those with lower abundance and limited spread. In contrast, ToP were rooted by the minimum character state, assuming that modern proteomes evolved from a relatively simpler urancestral organism that harbored only few FSFs (*57*) (see text S1 for a discussion of phylogenetic assumptions and models). MP approximates maximum likelihood when phylogenetic characters evolve at different rates (*123*) and is appropriate for global proteome studies. BS analysis with 1000 replicates was performed to assess the reliability of deep evolutionary relationships. Trees were visualized with Dendroscope (version 3.2.8) (*124*).

### uToL reconstructions from the numerical analysis of FSF domain age

EvoPCO analysis was performed using Microsoft Excel XLSTAT plug-in (*125*). For this reconstruction, proteomes were treated as samples and FSFs as variables. Because proteomes are composed of FSFs of different ages (that is, nd values), we transformed the FSF occurrence matrix into an FSF occurrence* (1 − nd) matrix, making the matrix a multidimensional space of the evolutionary age of domains. The "reverse age" 1 − nd transformation ensured that we did not lose information about FSFs of very ancient origin (for example, c.37.1 that has an nd of 0 and could be confused with FSFs that were absent in a proteome). Similarly, the transformation ensured that FSF absences (domains that have not yet materialized) did not contribute age to the multidimensional temporal space. Next, we calculated Euclidean distances that described pairwise dissimilarity among proteomes. The pairwise phylogenetic distance matrix was used to calculate the first three principal coordinates that described maximum variability in data. Effectively, the evoPCO method provided the three most significant loadings that described how component parts (FSFs) contribute to the history of systems (proteomes). The evoPCO method should be considered "rooted" in time because the multidimensional space was centered on an nd parameter that correlates with geological time (*56*). For reference, we added the previously reconstructed proteome of the last common ancestor of modern cells (*57*) as an additional sample.

### Network and NJ reconstructions

Phylogenomic networks were generated using the Neighbor-Net algorithm (*126*) implemented in the SplitsTree package (version 4.13.1) (*127*). An NJ tree was calculated from the pairwise phylogenetic distance matrix using the "Phangorn" and "ape" packages in R version 2.15.2.

### Functional analysis

GO enrichment analysis was performed using the domain-centric GO resource (*128*). A list of FSFs was provided as input, and only the most significant and highly specific biological process GO terms that were enriched in the given set of FSFs were retrieved.

## SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/1/8/e1500527/DC1

Text S1. Phylogenetic assumptions and models.
Fig. S1. FSF use and reuse for proteomes in each viral subgroup and for free-living cellular organisms.
Fig. S2. Distribution of FSFs in each of the seven Venn groups defined in Fig. 3B along the evolutionary timeline (nd).
Fig. S3. Spread of *abe* core FSFs in viral subgroups.
Fig. S4. Evolutionary relationships within the viral subgroup.
Fig. S5. Evolutionary relationships between cells and viruses.
Table S1. List of viruses sampled in this study.
Table S2. List of cellular organisms sampled in this study.
Table S3. VSFs and their spread in cellular (X) proteomes.
Table S4. FSF use and reuse values for all proteomes.
Table S5. List of FSFs corresponding to each of the seven Venn groups defined in Fig. 3B.
Table S6. FSFs mapped to structure-based viral lineages.
Table S7. Significantly enriched "biological process" GO terms in EV FSFs (FDR < 0.01).
References (*132*–*137*)

## REFERENCES AND NOTES


1. E. Domingo, J. J. Holland, RNA virus mutations and fitness for survival. *Annu. Rev. Microbiol.* **51**, 151–178 (1997).
2. A. M. Q. King, M. J. Adams, E. B. Carstens, E. J. Lefkowitz, *Virus Taxonomy: Classification and Nomenclature of Viruses: Ninth Report of the International Committee on Taxonomy of Viruses* (Elsevier, San Diego, CA, 2012).
3. M. Krupovic, D. H. Bamford, Double-stranded DNA viruses: 20 families and only five different architectural principles for virion assembly. *Curr. Opin. Virol.* **1**, 118–124 (2011).
4. S. Balaji, N. Srinivasan, Comparison of sequence-based and structure-based phylogenetic trees of homologous proteins: Inferences on protein evolution. *J. Biosci.* **32**, 83–96 (2007).
5. A. Abroi, J. Gough, Are viruses a source of new protein folds for organisms? Virosphere structure space and evolution. *Bioessays* **33**, 626–635 (2011).
6. S. Balaji, N. Srinivasan, Use of a database of structural alignments and phylogenetic trees in investigating the relationship between sequence and structural variability among homologous proteins. *Protein Eng.* **14**, 219–226 (2001).
7. T. J. P. Hubbard, T. L. Blundell, Comparison of solvent-inaccessible cores of homologous proteins: Definitions useful for protein modelling. *Protein Eng.* **1**, 159–171 (1987).
8. C. Chothia, A. M. Lesk, The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823–826 (1986).
9. A. G. Murzin, S. E. Brenner, T. Hubbard, C. Chothia, SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540 (1995).
10. A. E. Todd, C. A. Orengo, J. M. Thornton, Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **307**, 1113–1143 (2001).
11. D. Lundin, A. M. Poole, B.-M. Sjöberg, M. Högbom, Use of structural phylogenetic networks for classification of the ferritin-like superfamily. *J. Biol. Chem.* **287**, 20565–20575 (2012).
12. D. Caetano-Anollés, K. M. Kim, J. E. Mittenthal, G. Caetano-Anollés, Proteome evolution and the metabolic origins of translation and cellular life. *J. Mol. Evol.* **72**, 14–33 (2011).
13. C. O'Donovan, M. J. Martin, A. Gattiker, E. Gasteiger, A. Bairoch, R. Apweiler, High-quality protein knowledge resource: SWISS-PROT and TrEMBL. *Brief. Bioinform.* **3**, 275–284 (2002).
14. K. Illergård, D. H. Ardell, A. Elofsson, Structure is three to ten times more conserved than sequence—A study of structural response in protein cores. *Proteins* **77**, 499–508 (2009).
15. J. Gough, Convergent evolution of domain architectures (is rare). *Bioinformatics* **21**, 1464–1471 (2005).
16. G. Caetano-Anollés, A. Nasir, Benefits of using molecular structure and abundance in phylogenomic analysis. *Front. Genet.* **3**, 172 (2012).
17. A. Harish, A. Tunlid, C. G. Kurland, Rooted phylogeny of the three superkingdoms. *Biochimie* **95**, 1593–1604 (2013).
18. M. Wang, L. S. Yafremava, D. Caetano-Anollés, J. E. Mittenthal, G. Caetano-Anollés, Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world. *Genome Res.* **17**, 1572–1585 (2007).
19. A. Nasir, K. M. Kim, G. Caetano-Anolles, Giant viruses coexisted with the cellular ancestors and represent a distinct supergroup along with superkingdoms Archaea, Bacteria and Eukarya. *BMC Evol. Biol.* **12**, 156 (2012).
20. J. Gough, K. Karplus, R. Hughey, C. Chothia, Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.* **313**, 903–919 (2001).
21. P. Forterre, The virocell concept, in: eLS (Wiley, Chichester, UK, 2012).
22. P. Forterre, Manipulation of cellular syntheses and the nature of viruses: The virocell concept. *C. R. Chim.* **14**, 392–399 (2011).
23. D. Baltimore, Expression of animal virus genomes. *Bacteriol. Rev.* **35**, 235–241 (1971).
24. V. A. Kostyuchenko, G. A. Navruzbekov, L. P. Kurochkina, S. V. Strelkov, V. V. Mesyanzhinov, M. G. Rossmann, The structure of bacteriophage T4 gene product 9: The trigger for tail contraction. *Structure* **7**, 1213–1222 (1999).
25. J. Liu, G. Glazko, A. Mushegian, Protein repertoire of double-stranded DNA bacteriophages. *Virus Res.* **117**, 68–80 (2006).
26. J. Grimes, A. K. Basak, P. Roy, D. Stuart, The crystal structure of bluetongue virus VP7. *Nature* **373**, 167–170 (1995).
27. M. Mathieu, I. Petitpas, J. Navaza, J. Lepault, E. Kohli, P. Pothier, B. V. Venkataram Prasad, J. Cohen, F. A. Rey, Atomic structure of the major capsid protein of rotavirus: Implications for the architecture of the virion. *EMBO J.* **20**, 1485–1497 (2001).
28. P. B. Rosenthal, X. Zhang, F. Formanowski, W. Fitz, C.-H. Wong, H. Meier-Ewert, J. J. Skehel, D. C. Wiley, Structure of the haemagglutinin-esterase-fusion glycoprotein of influenza C virus. *Nature* **396**, 92–96 (1998).
29. Y. Ha, D. J. Stevens, J. J. Skehel, D. C. Wiley, H5 avian and H9 swine influenza virus haemagglutinin structures: Possible origin of influenza subtypes. *EMBO J.* **21**, 865–875 (2002).
30. N. Yutin, Y. I. Wolf, E. V. Koonin, Origin of giant viruses from smaller DNA viruses not from a fourth domain of cellular life. *Virology* **466–467**, 38–52 (2014).
31. R. W. Hendrix, J. G. Lawrence, G. F. Hatfull, S. Casjens, The origins and ongoing evolution of viruses. *Trends Microbiol.* **8**, 504–508 (2000).
32. J. Filée, P. Forterre, Viral proteins functioning in organelles: A cryptic origin? *Trends Microbiol.* **13**, 510–513 (2005).
33. H. Brüssow, C. Canchaya, W.-D. Hardt, Phages and the evolution of bacterial pathogens: From genomic rearrangements to lysogenic conversion. *Microbiol. Mol. Biol. Rev.* **68**, 560–602 (2004).
34. P. Forterre, D. Prangishvili, The major role of viruses in cellular evolution: Facts and hypotheses. *Curr. Opin. Virol.* **3**, 558–565 (2013).
35. E. V. Koonin, V. V. Dolja, A virocentric perspective on the evolution of life. *Curr. Opin. Virol.* **3**, 546–557 (2013).
36. M. Wang, C. G. Kurland, G. Caetano-Anollés, Reductive evolution of proteomes and protein structures. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 11954–11958 (2011).
37. J. P. McCutcheon, N. A. Moran, Extreme genome reduction in symbiotic bacteria. *Nat. Rev. Microbiol.* **10**, 13–26 (2011).
38. A. Nasir, A. Naeem, M. J. Khan, H. D. L. Nicora, G. Caetano-Anollés, Annotation of protein domains reveals remarkable conservation in the functional make up of proteomes across superkingdoms. *Genes* **2**, 869–911 (2011).
39. N. A. Moran, Microbial minimalism: Genome reduction in bacterial pathogens. *Cell* **108**, 583–586 (2002).
40. C. I. Bandea, A new theory on the origin and the nature of viruses. *J. Theor. Biol.* **105**, 591–602 (1983).
41. C. I. Bandea, The origin and evolution of viruses as molecular organisms. *Nat. Preceedings* 10101/npre.2009.3886.1 (2009).
42. J.-M. Claverie, C. Abergel, Open questions about giant viruses. *Adv. Virus Res.* **85**, 25–56 (2013).
43. A. Nasir, F.-J. Sun, K. M. Kim, G. Caetano-Anollés, Untangling the origin of viruses and their impact on cellular evolution. *Ann. N. Y. Acad. Sci.* **1341**, 61–74 (2015).
44. A. Nasir, P. Forterre, K. M. Kim, G. Caetano-Anollés, The distribution and impact of viral lineages in domains of life. *Front. Microbiol.* **5**, 194 (2014).
45. E. V. Koonin, V. V. Dolja, M. Krupovic, Origins and evolution of viruses of eukaryotes: The ultimate modularity. *Virology* **479–480**, 2–25 (2015).
46. P. Forterre, The common ancestor of Archaea and Eukarya was not an archaeon. *Archaea* **2013**, 372396 (2013).
47. N. G. A. Abrescia, D. H. Bamford, J. M. Grimes, D. I. Stuart, Structure unifies the viral universe. *Annu. Rev. Biochem.* **81**, 795–822 (2012).
48. M. Sutter, D. Boehringer, S. Gutmann, S. Günther, D. Prangishvili, M. J. Loessner, K. O. Stetter, E. Weber-Ban, N. Ban, Structural basis of enzyme encapsulation into a bacterial nanocompartment. *Nat. Struct. Mol. Biol.* **15**, 939–947 (2008).
49. P. Forterre, D. Prangishvili, The origin of viruses. *Res. Microbiol.* **160**, 466–472 (2009).

50. T. O. Yeates, Y. Tsai, S. Tanaka, M. R. Sawaya, C. A. Kerfeld, Self-assembly in the carboxysome: A viral capsid-like protein shell in bacterial cells. *Biochem. Soc. Trans.* **35**, 508–511 (2007).
51. T. O. Yeates, M. C. Thompson, T. A. Bobik, The protein shells of bacterial microcompartment organelles. *Curr. Opin. Struct. Biol.* **21**, 223–231 (2011).
52. S. Dutta, I. V. Akey, C. Dingwall, K. L. Hartman, T. Laue, R. T. Nolte, J. F. Head, C. W. Akey, The crystal structure of nucleoplasmin-core: Implications for histone binding and nucleosome assembly. *Mol. Cell* **8**, 841–853 (2001).
53. M. Mizuno, K. Yasukawa, K. Inouye, Insight into the mechanism of the stabilization of Moloney murine leukaemia virus reverse transcriptase by eliminating RNase H activity. *Biosci. Biotechnol. Biochem.* **74**, 440–442 (2010).
54. M. S. Chapman, L. Liljas, Structural folds of viral proteins. *Adv. Protein Chem.* **64**, 125–196 (2003).
55. G. S. Diemer, K. M. Stedman, A novel virus genome discovered in an extreme environment suggests recombination between unrelated groups of RNA and DNA viruses. *Biol. Direct* **7**, 13 (2012).
56. M. Wang, Y.-Y. Jiang, K. Mo Kim, G. Qu, H.-F. Ji, J. E. Mittenthal, H.-Y. Zhang, G. Caetano-Anollés, 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 (2011).
57. K. M. Kim, G. Caetano-Anollés, The proteomic complexity and rise of the primordial ancestor of diversified life. *BMC Evol. Biol.* **11**, 140 (2011).
58. J. P. Gogarten, H. Kibak, P. Dittrich, L. Taiz, E. J. Bowman, B. J. Bowman, M. F. Manolson, R. J. Poole, T. Date, T. Oshima, Evolution of the vacuolar H+-ATPase: Implications for the origin of eukaryotes. *Proc. Natl. Acad. Sci. U.S.A.* **86**, 6661–6665 (1989).
59. J. P. Gogarten, L. Olendzenski, Orthologs, paralogs and genome comparisons. *Curr. Opin. Genet. Dev.* **9**, 630–636 (1999).
60. C. R. Woese, O. Kandler, M. L. Wheelis, Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. U.S.A.* **87**, 4576–4579 (1990).
61. J. T.-F. Wong, J. Chen, W.-K. Mat, S.-K. Ng, H. Xue, Polyphasic evidence delineating the root of life and roots of biological domains. *Gene* **403**, 39–52 (2007).
62. H. Xue, K.-L. Tong, C. Marck, H. Grosjean, J. T.-F. Wong, Transfer RNA paralogs: Evidence for genetic code–amino acid biosynthesis coevolution and an archaeal root of life. *Gene* **310**, 59–66 (2003).
63. H. Xue, S.-K. Ng, K.-L. Tong, J. T.-F. Wong, Congruence of evidence for a *Methanopyrus*-proximal root of life based on transfer RNA and aminoacyl-tRNA synthetase genes. *Gene* **360**, 120–130 (2005).
64. K. M. Kim, G. Caetano-Anollés, 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 (2012).
65. K. M. Kim, A. Nasir, K. Hwang, G. Caetano-Anollés, A tree of cellular life inferred from a genomic census of molecular functions. *J. Mol. Evol.* **79**, 240–262 (2014).
66. P. Forterre, The two ages of the RNA world, and the transition to the DNA world: A story of viruses and cells. *Biochimie* **87**, 793–803 (2005).
67. G. R. Burke, M. R. Strand, Polydnaviruses of parasitic wasps: Domestication of viruses to act as gene delivery vectors. *Insects* **3**, 91–119 (2012).
68. D. Prangishvili, M. Krupovic, A new proposed taxon for double-stranded DNA viruses, the order "Ligamenvirales". *Arch. Virol.* **157**, 791–795 (2012).
69. N. Philippe, M. Legendre, G. Doutre, Y. Couté, O. Poirot, M. Lescot, D. Arslan, V. Seltzer, L. Bertaux, C. Bruley, J. Garin, J. M. Claverie, C. Abergel, Pandoraviruses: Amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science* **341**, 281–286 (2013).
70. M. Legendre, J. Bartoli, L. Shmakova, S. Jeudy, K. Labadie, A. Adrait, M. Lescot, O. Poirot, L. Bertaux, C. Bruley, Y. Couté, E. Rivkina, C. Abergel, J.-M. Claverie, Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 4274–4279 (2014).
71. P. Colson, X. De Lamballerie, N. Yutin, S. Asgari, Y. Bigot, D. K. Bideshi, X.-W. Cheng, B. A. Federici, J. L. Van Etten, E. V. Koonin, B. La Scola, D. Raoult, "Megavirales", a proposed new order for eukaryotic nucleocytoplasmic large DNA viruses. *Arch. Virol.* **158**, 2517–2521 (2013).
72. M. L. Baker, W. Jiang, F. J. Rixon, W. Chiu, Common ancestry of herpesviruses and tailed DNA bacteriophages. *J. Virol.* **79**, 14967–14970 (2005).
73. A. Nasir, K. M. Kim, G. Caetano-Anollés, Viral evolution: Primordial cellular origins and late adaptation to parasitism. *Mob. Genet. Elements* **2**, 247–252 (2012).
74. D. Wu, M. Wu, A. Halpern, D. B. Rusch, S. Yooseph, M. Frazier, J. C. Venter, J. A. Eisen, Stalking the fourth domain in metagenomic data: Searching for, discovering, and interpreting novel, deep branches in marker gene phylogenetic trees. *PLOS One* **6**, e18011 (2011).
75. M. Boyer, M.-A. Madoui, G. Gimenez, B. La Scola, D. Raoult, Phylogenetic and phyletic studies of informational genes in genomes highlight existence of a 4th domain of life including giant viruses. *PLOS One* **5**, e15530 (2010).
76. D. Moreira, P. López-García, Ten reasons to exclude viruses from the tree of life. *Nat. Rev. Microbiol.* **7**, 306–311 (2009).
77. D. Moreira, C. Brochier-Armanet, Giant viruses, giant chimeras: The multiple evolutionary histories of Mimivirus genes. *BMC Evol. Biol.* **8**, 12 (2008).
78. D. Moreira, P. López-García, Comment on "The 1.2-megabase genome sequence of Mimivirus". *Science* **308**, 1114 (2005).
79. E. V. Koonin, On the origin of cells and viruses: Primordial virus world scenario. *Ann. N. Y. Acad. Sci.* **1178**, 47–64 (2009).
80. E. V. Koonin, T. G. Senkevich, V. V. Dolja, The ancient Virus World and evolution of cells. *Biol. Direct* **1**, 29 (2006).
81. D. Raoult, P. Forterre, Redefining viruses: Lessons from Mimivirus. *Nat. Rev. Microbiol.* **6**, 315–319 (2008).
82. J.-M. Claverie, H. Ogata, Ten good reasons not to exclude giruses from the evolutionary picture. *Nat. Rev. Microbiol.* **7**, 615 (2009).
83. J.-M. Claverie, H. Ogata, S. Audic, C. Abergel, K. Suhre, P.-E. Fournier, Mimivirus and the emerging concept of "giant" virus. *Virus Res.* **117**, 133–144 (2006).
84. P. López-García, The place of viruses in biology in light of the metabolism- versus-replication-first debate. *Hist. Philos. Life Sci.* **34**, 391–406 (2012).
85. R. R. Novoa, G. Calderita, R. Arranz, J. Fontana, H. Granzow, C. Risco, Virus factories: Associations of cell organelles for viral replication and morphogenesis. *Biol. Cell* **97**, 147–172 (2005).
86. J.-M. Claverie, Viruses take center stage in cellular evolution. *Genome Biol.* **7**, 110 (2006).
87. N. R. Hegde, M. S. Maddur, S. V. Kaveri, J. Bayry, Reasons to include viruses in the tree of life. *Nat. Rev. Microbiol.* **7**, 615 (2009).
88. D. Wacey, M. R. Kilburn, M. Saunders, J. Cliff, M. D. Brasier, Microfossils of sulphur-metabolizing cells in 3.4-billion-year-old rocks of Western Australia. *Nat. Geosci.* **4**, 698–702 (2011).
89. E. J. Javaux, C. P. Marshall, A. Bekker, Organic-walled microfossils in 3.2-billion-year-old shallow-marine siliciclastic deposits. *Nature* **463**, 934–938 (2010).
90. P. Forterre, The origin of viruses and their possible roles in major evolutionary transitions. *Virus Res.* **117**, 5–16 (2006).
91. B. La Scola, S. Audic, C. Robert, L. Jungang, X. de Lamballerie, M. Drancourt, R. Birtles, J. M. Claverie, D. Raoult, A giant virus in amoebae. *Science* **299**, 2033 (2003).
92. D. Arslan, M. Legendre, V. Seltzer, C. Abergel, J.-M. Claverie, Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 17486–17491 (2011).
93. D. G. Meckes Jr., N. Raab-Traub, Microvesicles and viral infection. *J. Virol.* **85**, 12844–12854 (2011).
94. M. Jalasvuori, J. K. Bamford, Structural co-evolution of viruses and cells in the primordial world. *Orig. Life Evol. Biosph.* **38**, 165–181 (2008).
95. P. Forterre, M. Krupovic, in *Viruses: Essential Agents of Life*, G. Witzany, Ed. (Springer, Dordrecht, Netherlands, 2012), pp. 43–60.
96. P. Forterre, The origin of DNA genomes and DNA replication proteins. *Curr. Opin. Microbiol.* **5**, 525–532 (2002).
97. H. Ogata, J.-M. Claverie, Unique genes in giant viruses: Regular substitution pattern and anomalously short size. *Genome Res.* **17**, 1353–1361 (2007).
98. D. Prangishvili, R. A. Garrett, E. V. Koonin, Evolutionary genomics of archaeal viruses: Unique viral genomes in the third domain of life. *Virus Res.* **117**, 52–67 (2006).
99. Y. Yin, D. Fischer, Identification and investigation of ORFans in the viral world. *BMC Genomics* **9**, 24 (2008).
100. D. Cortez, P. Forterre, S. Gribaldo, A hidden reservoir of integrative elements is the major source of recently acquired foreign genes and ORFans in archaeal and bacterial genomes. *Genome Biol.* **10**, R65 (2009).
101. P. Forterre, 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).
102. P. Forterre, D. Prangishvili, The great billion-year war between ribosome- and capsid-encoding organisms (cells and viruses) as the major source of evolutionary novelties. *Ann. N. Y. Acad. Sci.* **1178**, 65–77 (2009).
103. H. Liu, Y. Fu, D. Jiang, G. Li, J. Xie, J. Cheng, Y. Peng, S. A. Ghabrial, X. Yi, Widespread horizontal gene transfer from double-stranded RNA viruses to eukaryotic nuclear genomes. *J. Virol.* **84**, 11876–11887 (2010).
104. T. E. Shutt, M. W. Gray, Bacteriophage origins of mitochondrial replication and transcription proteins. *Trends Genet.* **22**, 90–95 (2006).
105. A. Katzourakis, R. J. Gifford, Endogenous viral elements in animal genomes. *PLOS Genet.* **6**, e1001191 (2010).
106. S. Mi, X. Lee, X. Li, G. M. Veldman, H. Finnerty, L. Racie, E. LaVallie, X. Y. Tang, P. Edouard, S. Howes, J. C. Keith Jr., J. M. McCoy, Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature* **403**, 785–789 (2000).
107. C. R. Woese, *Evolution from Molecules to Men* (Cambridge Univ. Press, Cambridge, UK, 1983), pp. 209–233.
108. D. Lundin, S. Gribaldo, E. Torrents, B. M. Sjoberg, A. M. Poole, Ribonucleotide reduction—Horizontal transfer of a required function spans all three domains. *BMC Evol. Biol.* **10**, 383 (2010).
109. P. López-García, D. Moreira, Metabolic symbiosis at the origin of eukaryotes. *Trends Biochem. Sci.* **24**, 88–93 (1999).
110. W. Martin, M. Müller, The hydrogen hypothesis for the first eukaryote. *Nature* **392**, 37–41 (1998).

111. C. J. Cox, P. G. Foster, R. P. Hirt, S. R. Harris, T. M. Embley, The archaebacterial origin of eukaryotes. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 20356–20361 (2008).
112. T. A. Williams, P. G. Foster, T. M. W. Nye, C. J. Cox, T. M. Embley, A congruent phylogenomic signal places eukaryotes within the Archaea. *Proc. R. Soc. B* **279**, 4870–4879 (2012).
113. L. Guy, T. J. G. Ettema, The archaeal 'TACK' superphylum and the origin of eukaryotes. *Trends Microbiol.* **19**, 580–587 (2011).
114. S. Nelson-Sathi, T. Dagan, G. Landan, A. Janssen, M. Steel, J. O. McInerney, U. Deppenmeier, W. F. Martina, Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 20537–20542 (2012).
115. A. Goulet, S. Blangy, P. Redder, D. Prangishvili, C. Felisberto-Rodrigues, P. Forterre, V. Campanacci, C. Cambillau, Acidianus filamentous virus 1 coat proteins display a helical fold spanning the filamentous archaeal viruses lineage. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 21155–21160 (2009).
116. B. Bolduc, D. P. Shaughnessy, Y. I. Wolf, E. V. Koonin, F. F. Roberto, M. Young, Identification of novel positive-strand RNA viruses by metagenomic analysis of archaea-dominated Yellowstone hot springs. *J. Virol.* **86**, 5562–5573 (2012).
117. M. Krupovič, P. Forterre, D. H. Bamford, Comparative analysis of the mosaic genomes of tailed archaeal viruses and proviruses suggests common themes for virion architecture and assembly with tailed viruses of bacteria. *J. Mol. Biol.* **397**, 144–160 (2010).
118. D. Prangishvili, The wonderful world of archaeal viruses. *Annu. Rev. Microbiol.* **67**, 565–585 (2013).
119. Y. Bao, S. Federhen, D. Leipe, V. Pham, S. Resenchuk, M. Rozanov, R. Tatusov, T. Tatusova, National Center for Biotechnology Information Viral Genomes Project. *J. Virol.* **78**, 7291–7298 (2004).
120. M. Wang, G. Caetano-Anollés, Global phylogeny determined by the combination of protein domains in proteomes. *Mol. Biol. Evol.* **23**, 2444–2454 (2006).
121. D. L. Swofford, *Phylogenomic Analysis Using Parsimony and Other Programs (PAUP*)* Ver 4.0b10 (Sinauer, Sunderland, MA, 2002).
122. J. G. Lundberg, Wagner networks and ancestors. *Syst. Biol.* **21**, 398–413 (1972).
123. B. Kolaczkowski, J. W. Thornton, Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* **431**, 980–984 (2004).
124. D. H. Huson, D. C. Richter, C. Rausch, T. Dezulian, M. Franz, R. Rupp, Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics* **8**, 460 (2007).
125. T. Fahmy, P. Aubry, *XLSTAT-Pro (Version 7.0)* (Society Addinsoft, Brooklyn, NY, 2008).
126. D. Bryant, V. Moulton, Neighbor-Net: An agglomerative method for the construction of phylogenetic networks. *Mol. Biol. Evol.* **21**, 255–265 (2004).
127. D. H. Huson, SplitsTree: Analyzing and visualizing evolutionary data. *Bioinformatics* **14**, 68–73 (1998).
128. H. Fang, J. Gough, dcGO: Database of domain-centric ontologies on functions, phenotypes, diseases and more. *Nucleic Acids Res.* **41**, D536–D544 (2013).
129. M. K. Pietilä, E. Roine, L. Paulin, N. Kalkkinen, D. H. Bamford, An ssDNA virus infecting archaea: A new lineage of viruses with a membrane envelope. *Mol. Microbiol.* **72**, 307–319 (2009).
130. T. Mochizuki, M. Krupovic, G. Pehau-Arnaudet, Y. Sako, P. Forterre, D. Prangishvili, Archaeal virus with exceptional virion architecture and the largest single-stranded DNA genome. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 13386–13391 (2012).
131. C. Hulo, E. de Castro, P. Masson, L. Bougueleret, A. Bairoch, I. Xenarios, P. Le Mercier, ViralZone: A knowledge resource to understand virus diversity. *Nucleic Acids Res.* **39** (Suppl. 1), D576–D582 (2011).
132. A. Nasir, K. M. Kim, G. Caetano-Anollés, Global patterns of protein domain gain and loss in superkingdoms. *PLOS Comput. Biol.* **10**, e1003452 (2014).
133. P. H. Weston, in *Ontogeny and Systematics*, C. J. Humphries, Ed. (Columbia Univ. Press, New York, 1988), pp. 27–56.
134. P. H. Weston, in *Models in Phylogeny Reconstruction*, R. W. Scotland, D. J. Siebert, D. M. Williams, Eds. (Clarendon Press, Oxford, UK, 1994), pp. 125–155.
135. G. Caetano-Anollés, M. Wang, D. Caetano-Anollés, J. E. Mittenthal, The origin, evolution and structure of the protein world. *Biochem. J.* **417**, 621–637 (2009).
136. J. P. Huelsenbeck, R. Nielsen, Effect of nonindependent substitution on phylogenetic accuracy. *Syst. Biol.* **48**, 317–328 (1999).
137. G. Caetano-Anollés, A. Nasir, K. Zhou, D. Caetano-Anollés, J. E. Mittenthal, F.-J. Sun, K. M. Kim, Archaea: The first domain of diversified life. *Archaea* **2014**, 590214 (2014).


**Acknowledgments:** We thank K. M. Kim, J. Mittenthal, M. Hudson, J. Ma, P. Forterre, and members of the Evolutionary Bioinformatics Laboratory for their support and valuable input. **Funding:** Research was supported by the National Science Foundation (grant OISE-1132791 to G.C.-A.) and the United States Department of Agriculture (grants ILLU-802-909 and ILLU-483-625 to G.C.-A.). A.N. was the recipient of Chateaubriand fellowship from the French Government and Dissertation Completion fellowship from the Graduate College of the University of Illinois. The research reported in this study is part of his doctoral dissertation. **Author contributions:** These authors contributed equally to this work. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data and information necessary to evaluate the conclusions of this paper are presented herein.

Submitted 27 April 2015
Accepted 30 June 2015
Published 25 September 2015
10.1126/sciadv.1500527

**Citation:** A. Nasir, G. Caetano-Anollés, A phylogenomic data-driven exploration of viral origins and evolution. *Sci. Adv.* **1**, e1500527 (2015).

ScienceAdvances
AAAS

# A phylogenomic data-driven exploration of viral origins and evolution

Arshan Nasir and Gustavo Caetano-Anollés (September 25, 2015)
*Sci Adv* 2015, 1:.
doi: 10.1126/sciadv.1500527