Bacterial Origin and Reductive Evolution of the CPR Group

Rijja Hussain Bokhari[1,†], Nooreen Amirjan[1,†], Hyeonsoo Jeong[2], Kyung Mo Kim[3], Gustavo Caetano-Anollés[4], and Arshan Nasir [1,5,*]

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

[2]School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA

[3]Division of Polar Life Sciences, Korea Polar Research Institute, Incheon, Republic of Korea

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

[5]Theoretical Biology & Biophysics Group, Los Alamos National Laboratory, Los Alamos, New Mexico

*Corresponding author: E-mail: arshan.nasir@gmail.com.

Accepted: January 31, 2020

[†]These authors contributed equally to this work.

Abstract

The candidate phyla radiation (CPR) is a proposed subdivision within the bacterial domain comprising several candidate phyla. CPR organisms are united by small genome and physical sizes, lack several metabolic enzymes, and populate deep branches within the bacterial subtree of life. These features raise intriguing questions regarding their origin and mode of evolution. In this study, we performed a comparative and phylogenomic analysis to investigate CPR origin and evolution. Unlike previous gene/protein sequence-based reports of CPR evolution, we used protein domain superfamilies classified by protein structure databases to resolve the evolutionary relationships of CPR with non-CPR bacteria, Archaea, Eukarya, and viruses. Across all supergroups, CPR shared maximum superfamilies with non-CPR bacteria and were placed as deep branching bacteria in most phylogenomic trees. CPR contributed 1.22% of new superfamilies to bacteria including the ribosomal protein L19e and encoded four core superfamilies that are likely involved in cell-to-cell interaction and establishing episymbiotic lifestyles. Although CPR and non-CPR bacterial proteomes gained common superfamilies over the course of evolution, CPR and Archaea had more common losses. These losses mostly involved metabolic superfamilies. In fact, phylogenies built from only metabolic protein superfamilies separated CPR and non-CPR bacteria. These findings indicate that CPR are bacterial organisms that have probably evolved in an Archaea-like manner via the early loss of metabolic functions. We also discovered that phylogenies built from metabolic and informational superfamilies gave contrasting views of the groupings among Archaea, Bacteria, and Eukarya, which add to the current debate on the evolutionary relationships among superkingdoms.

Key words: candidate phyla radiation, tree of life, phylogenetics, reductive evolution, protein structure.

Introduction

The recent drive in single-cell genomics and genome-resolved metagenomics has led to the discovery and description of several novel microbial lineages that are challenging the once widely accepted evolutionary relationships of organisms. The known diversity and biological roles of both the archaeal and bacterial domains of life have significantly expanded following the discoveries of Asgard (Spang et al. 2015; Zaremba-Niedzwiedzka et al. 2017), DPANN (Rinke et al. 2013), and the candidate phyla radiation (CPR) (Brown et al. 2015) groups in Archaea and Bacteria. Asgard archaea encode several eukaryote-specific proteins involved in cytoskeleton remodeling and phagocytosis. Their intriguing genomic features have led to rejuvenated debates regarding the evolutionary relationship of Archaea with Eukarya (sister group or ancestors?) (Nasir et al. 2015, 2016; Spang and Ettema 2016; Da Cunha et al. 2017, 2018; Spang et al. 2018; Williams et al. 2020). Recently, a decade-long effort led to the description of the first cultured representative of Asgard archaea, which is a significant milestone in microbiology (Imachi et al. 2020).

In turn, DPANN and CPR are large and seemingly monophyletic expansions within the archaeal and bacterial domains of life, respectively, that are united by small genome and physical sizes and possibly symbiotic lifestyles (Castelle et al. 2018). Both groups harbor many hitherto-uncultivated representatives and lack numerous key biosynthetic capabilities

Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution 2020. This work is written by US Government employees and is in the public domain in the US.

(Castelle and Banfield 2018). Phylogenetic analysis based on 16S ribosomal RNA (rRNA) genes or concatenation of ribosomal proteins indicates that they occupy basal or deep branches in their respective trees of life (Hug et al. 2016) posing intriguing questions regarding their origin and mode of evolution (Castelle and Banfield 2018). A recent analysis on the comparative distribution of protein families in CPR and non-CPR bacteria confirmed the distinctive nature of CPR genomes within bacteria (Méheust et al. 2019). CPR have now been described in a range of environments indicating that they are quite widespread (He et al. 2015; Castelle et al. 2018; Orsi et al. 2018; Starr et al. 2018).

Although gene and protein sequences are useful tools to study evolution, they can sometimes fail to resolve very ancient evolutionary relationships (Caetano-Anollés and Nasir 2012). There are three major limitations of sequence-based phylogenetic analyses. First, the number of "universal" genes conserved across a wide range of organisms (and viruses) is likely very small. For example, the 120 single-copy marker genes conserved across >90,000 prokaryotic genomes in Parks et al. (2018) may represent only a small fraction of the average bacterial genome. This number is expected to decline further with the discovery of novel microbial lineages. Second, genes are susceptible to saturating mutations, which can obscure reconstruction of evolutionary events (Penny and Zhong 2015). Third, protein domain gains, losses, inversions, and rearrangements are common evolutionary processes in genomes (Nasir et al. 2014b; Moore and Bornberg-Bauer 2012). These evolutionary events can complicate the recovery of a reliable sequence alignment, especially when genes encoding multidomain proteins are aligned from fast-evolving and highly diverged taxa (Nasir et al. 2016; Caetano-Anollés et al. 2017). In general, statistically recognizable sequence similarity fades over time (Sober and Steel 2002; Penny and Zhong 2015) and even homologous genes and proteins may sometimes be classified as too divergent by the sequence-based approaches. This may lead to incomplete clusterings of proteins in homologous protein family sets.

Fortunately, a large number of protein structures have now been resolved by experimental biologists and deposited into public databases. The Protein Data Bank (Rose et al. 2015) already hosts >150,000 protein structures (September 2019). Protein structures evolve at least three to ten times slower than gene and protein sequences (Illergård et al. 2009) owing to their direct link to biological function. This places stronger evolutionary pressure on structures to remain conserved for relatively longer periods of time. Thus, protein structure comparison among highly diverged and fast-evolving organisms can reveal remote homologies and probably resolve evolutionary relationships that may be invisible to sequence-based analyses (Caetano-Anollés and Nasir 2012).

Here, we pooled the completely sequenced and draft proteomes from Archaea, Bacteria (non-CPR and CPR), Eukarya, and nucleocytoplasmic large DNA viruses (NCLDV) (Iyer et al. 2001, 2006) and analyzed their evolutionary relationships. We scanned each sampled proteome against a precomputed library of profile hidden Markov models (HMMs) (Gough et al. 2001; Gough and Chothia 2002) to detect significant hits to fold superfamily (FSF) domains. FSFs, as defined by the Structural Classification of Proteins Database (ver. 1.75) (Fox et al. 2014), include distantly related protein domains (sequence identity can be <15%) that exhibit strong structural and functional homologies indicative of common origin (Fox et al. 2014). FSFs are thus useful proxies for 3D protein structure and provide a coarse-grained view to evaluate proteome evolution. We compared FSF assignments across proteomes and specifically investigated the origin and mode of evolution of CPR relative to non-CPR bacteria, which we will refer to as "well-described bacteria" (WDB) for brevity and contrast. Our experiments confirmed the highly reduced nature of CPR proteomes shaped by loss of several metabolic domains, their placement as deep-branching bacteria in the phylogenomic trees, and Archaea-like evolution shaped by early reductive evolution (Wang et al. 2007; Wang, Kurland, et al. 2011). The majority of these results are confirmatory of previous findings (Méheust et al. 2019) and support the use of protein structure information in resolving deep evolutionary relationships.

Materials and Methods

Retrieval and Filtering of Genomic Data Sets

We retrieved protein sequences encoded by the 2,430 completely sequenced genomes of Archaea ($n = 149$) and WDB (2,281) from Jeong et al. (2016). We have previously analyzed these genomes exhaustively for their participation in horizontal gene transfer (HGT) (Jeong and Nasir 2017; Jeong et al. 2019). The archaeal data set did not include many newly described Asgard, DPANN, and TACK lineages. Therefore, an additional 126 Archaea belonging to these lineages were added to the archaeal data set. This increased the archaeal proteome count from 149 to 275. Separately, we retrieved precalculated FSF assignments for a total of 383 eukaryotic and 3,460 viral proteomes from Nasir and Caetano-Anollés (2015). The data set in Nasir and Caetano-Anollés (2015) included viruses sampled from all replicon types followed by exhaustive screening and filtering of low-quality genomes. The data set therefore represents the best curated collection of viral and eukaryotic genome assemblies with precomputed FSF assignments. As we needed all genomes to match with at least phylum- or family-level taxonomy, we downloaded the NCBI taxonomy file (last accessed September 7, 2018, https://www.ncbi.nlm.nih.gov/genome/browse#!/overview/) that contained taxonomy information for >30,000 sequenced genomes. A total of 317/383 eukaryotes and 1,360 double-stranded DNA (dsDNA) viruses from NCLDV (Iyer et al. 2001, 2006) and other major viral families matched exactly with the taxonomy database and were retained. Separately, we

downloaded nine proteomes for newly discovered giant viruses such as Tupanviruses (Abrahão et al. 2018) and some Pandoraviruses that were not in the data set of Nasir and Caetano-Anollés (2015) and added to the virus count. Finally, we downloaded 777 CPR genomes from NCBI (Bio Project: 273161) (Brown et al. 2015) and matched them to phyla names using the online ggKbase resource (https://ggkbase.berkeley.edu/CPR-complete-draft/organisms). The final data set included 275 Archaea, 2,281 WDB, 317 Eukaryotes, 777 CPR bacteria, and 1,369 dsDNA viruses (16 major families/orders), comprising a total of 5,019 proteomes (supplementary table S1, Supplementary Material online).

Assignment of Superfamilies to Protein Sequences

All archaeal, WDB and CPR, and newly added viral proteomes were scanned against the HMM libraries representing proteins of known structure hosted by the SUPERFAMILY database (ver. 1.75) (Wilson et al. 2009) using the "superfamily.pl" PERL wrapper script to detect significant FSF domains ($E < 0.0001$). FSF hits for eukaryotes and other viruses were taken directly from Nasir and Caetano-Anollés (2015) (assigned using the same thresholds) and were added to the data set, as explained above. A total of 1,943 nonredundant significant FSFs were detected in 5,019 sampled genomes (supplementary table S2, Supplementary Material online).

Phylogeny Reconstruction and Calculation of Protein Domain Gain and Loss Events

Maximum parsimony (MP) was used to search for the most optimal unrooted phylogenomic tree to describe the evolution of proteomes (taxa) using PAUP* ver. 4.0b10 (Swofford 2002). MP was chosen because it performs as well or better than model/probability-based methods for less restrictive models of discrete morphological character evolution (Goloboff et al. 2018). FSF *occurrence* (i.e., presence or absence) and *abundance* (total redundant count) were considered as possible discrete character states. For *abundance*, raw counts were rescaled, normalized, and log-transformed on a scale from 0 to 23 (0-9 and A-9 possible character states in PAUP*) to account for unequal genome sizes and variances, as previously described (Wang and Caetano-Anollés 2006). Trees were rooted using two criteria: 1) the Lundberg direct rooting method that attaches the root *a posteriori* to an optimal unrooted tree in the most parsimonious manner (Lundberg 1972), reviewed in Caetano-Anollés et al. (2018); and 2) the indirect outgroup rooting method by the addition of an outgroup proteome outside the ingroup taxa. The reliability of phylogenetic splits was evaluated by running 1,000 bootstrap (BS) replicates and summarized using the "sumtrees" program in Dendropy (ver. 4.4.0) (Sukumaran and Holder 2010). Trees were visualized and edited using Dendroscope (ver. 3.5.10) (Huson et al. 2007) and Adobe Illustrator (ver. CC 2019). Character changes along the branches of trees were recorded by setting the "chglist" option to "yes" in the /describetrees command of PAUP*.

Functional Annotations

Molecular functions were assigned to FSF domains based on the SUPERFAMILY functional annotation scheme (for SCOP ver. 1.73) (Chothia et al. 2003; Vogel et al. 2004, 2005). These annotations assign the most dominant molecular function to each FSF and were retrieved from http://supfam.org/SUPERFAMILY/function.html.

Measurement of FSF Evolutionary Age and Spread

FSF evolutionary age was calculated by a node distance (*nd*) value. The *nd* is the distance of each FSF domain (tip node) from the root node and is given on a scale from 0 (most ancient) to 1 (most recent). The *nd* values correlate well with the geological record and exhibit a clock-like behavior for the evolution of protein folds (Wang, Jiang, et al. 2011). The *nd* values were taken from a previous phylogenomic tree of FSF domains published in Nasir and Caetano-Anollés (2015).

FSF spread was given by the *f*-value, defined by the number of proteomes encoding an FSF domain divided by the total number of proteomes in that supergroup. The *f*-value therefore indicates how widespread a particular FSF is among supergroup proteomes. An *f*-value of 1 implies presence in all proteomes of that supergroup, whereas values closer to 0 imply rare presence or complete absence or loss (since not gaining can also be considered a form of loss) (Wang et al. 2007).

Evolutionary Principal Coordinate Analysis

A data matrix comprising of 500 proteomes and FSF presence/absence (i.e., 1/0) data for 460 FSF domains was prepared. For proteome selection, 100 proteomes were randomly and equally sampled each from Archaea, WDB, CPR, Eukarya, and viruses. For FSF selection, only FSFs that were detected in each of the five supergroups were kept (reduced number from 1,943 to 460). This data matrix was imported into XLSTAT (Fahmy and Aubry 2003) for evolutionary principal coordinate analysis (evoPCO) (see Nasir and Caetano-Anollés [2015] for a previous implementation). FSF presence/absence data were transformed by multiplying with $1 - nd$. This transformation ensured that the most ancient FSF domain, c.37.1 (the P-loop containing NTP hydrolase) that had *nd* value of 0, was not excluded from the analysis. A pairwise distance matrix (Euclidean distance) was calculated for all proteomes. This distance matrix was used to determine the three most significant loadings that described how component parts (FSFs) contribute to the history of systems (proteomes).

Results

CPR Proteomes Are Part of the Bacterial Proteomic Repertoire

We analyzed 5,019 proteomes from five "supergroups," including 275 Archaea, 2,281 WDB, 777 CPR, 317 Eukarya, and 1,369 dsDNA viruses. These proteomes collectively encoded a total of 1,943 nonredundant FSF domains (supplementary tables S1 and S2, Supplementary Material online). We refer to proteome sets as "supergroups" instead of "domains" or "superkingdoms" because their evolutionary rank remains a matter of debate, and also, to avoid confusion with the usage of FSF domains in this study. When convenient, we abbreviate proteome sets in Archaea as (*A*), WDB as (*B*), CPR as (*C*), Eukarya as (*E*), and viruses as (*V*).

A total of 31 possible combinations exist for FSF sharing between the five supergroups, 29 of which were nonzero (fig. 1*A*). No FSFs were either unique to CPR or shared uniquely by Archaea and CPR. The majority of FSF domains ($n = 460$) were shared by all five supergroups (i.e., the *ABCEV* Venn group) comprising 23.6% of the total FSFs (supplementary table S2, Supplementary Material online). The very large number of universally shared FSFs in viruses has previously been linked to their origin from ancient cells (Nasir and Caetano-Anollés 2015). The next largest Venn group was *ABCE*, that is, the cell-only Venn group that contained 374 additional FSFs (fig. 1*A*), again a very large subset unifying cells. Therefore, a total of 834/990 (84%) of FSF domains detected in CPR proteomes were either *ABCEV* or *ABCE* (i.e., universal/shared) and only 16% were shared between CPR and 1–3 other supergroups. In terms of the two-supergroup sharing combinations, CPR shared 15 FSFs uniquely with WDB, 4 with Eukarya, 1 with viruses, and none with Archaea (fig. 1). Thus, FSF sharing patterns did not suggest that CPR merited a unique supergroup status. They did not possess any supergroup-specific domains, which is an important criterion to recognize new branches on the tree of life, and their maximum two-supergroup sharing was with WDB (highlighted in supplementary table S2, Supplementary Material online).

Next, we merged FSFs from CPR and Bacteria (i.e., WDB + CPR) to collapse the five-way Venn diagram into a four-way diagram representing only four supergroups (i.e., Archaea, Bacteria, Eukarya, and viruses) and 15 possible combinations (fig. 1*B*). The number of bacterial FSFs increased by only 1.22% by the addition of only 19 CPR FSFs to the bacterial repertoire (from 1,550 to 1,569) (supplementary table S3, Supplementary Material online). The most surprising addition was the ribosomal protein L19e (FSF a.94.1) known to be absent in WDB but present in Archaea and Eukarya (Lecompte et al. 2002). Interestingly, 18 out of the 19 new additions were previously assigned to either archaeal or eukaryotic proteomes (or their combinations, i.e., *ACE*, *ACEV*, *CEV*, *ACV*, and *CE* Venn groups in fig. 1*A*) and four had informational functional roles (supplementary table S3, Supplementary Material online). Archaea and Eukarya share an extended pool of ribosomal protein families ($n = 33$) that is absent in Bacteria (Lecompte et al. 2002), recently reviewed by Gaia et al. (2018). The overlap between the informational machineries (or more specifically the specialization of ribosomes) in Archaea/Eukarya versus Bacteria has repeatedly featured in the evolutionary scenarios describing the diversification of life (Forterre 2013; Williams et al. 2013). Assignment of L19e to CPR now transfers one of the Archaea-/Eukarya-specific ribosomal proteins to Bacteria. L19e, however, was only detected in 1/777 CPR proteomes with a significant *E*-value of <0.0001. In six other CPR proteomes, L19e had a somewhat higher assignment *E*-value of 0.00785. These hits were therefore excluded from our analysis as we chose a very stringent cutoff of <0.0001 to assign any FSF to a proteome (see Materials and Methods). The assignment of L19e to CPR could therefore be genuine although alternative explanations such as gain via HGT or erroneous HMM assignment cannot be ruled out either.

The Patchy Makeup of CPR Proteomes

Venn group classifications rely on the detection of an FSF in even a single proteome in that supergroup (e.g., L19e in one CPR proteome). These cases of rare presences of protein domains in very few members of a supergroup can also be assimilated to nonvertical evolutionary events such as convergent evolution or HGT or past evolutionary bottlenecks. Therefore, we revised the five-way Venn diagram (fig. 1*A*) by keeping only the core FSFs (i.e., FSFs present in >70% proteomes of a supergroup). Significant changes to the sharing patterns were observed (fig. 1*C*). The core proteomes represented a very tiny fraction of the original proteomes in all supergroups, except in Eukarya. For example, core bacterial FSFs were 469 for WDB and 206 for CPR in comparison to 1,550 and 990 indicating reductions of ~70% and ~79%, respectively. Similarly, the numbers decreased from 1,090 in Archaea to 327 (70% reduction) and from 669 to only 1 in viruses (fig. 1*C*). In other words, the *ABCEV* Venn group decreased from 460 to 1 because only a single FSF domain (the P-loop containing NTP hydrolase) was detected in >70% of sampled viral proteomes. These decreases could indicate either strong and consistent reductive tendencies in the prokaryotic and viral proteomes or lineage-specific gains that are not widespread among supergroups. Their ultimate effects are therefore the patchy distributions of many FSFs within the prokaryotic and viral supergroups. In contrast, the decrease in Eukarya was considerably smaller (53%, fig. 1*C*).

The number of core supergroup-specific domains decreased significantly from 158 to 57 in WDB, whereas it increased in Eukarya from 264 to 347 and in CPR from 0 to 4 (fig. 1*C*). The four CPR-specific core domains included FSFs d.24.1 ["Pili subunits"], a.265.1 ["Fic-like"], d.52.10 ["EspE

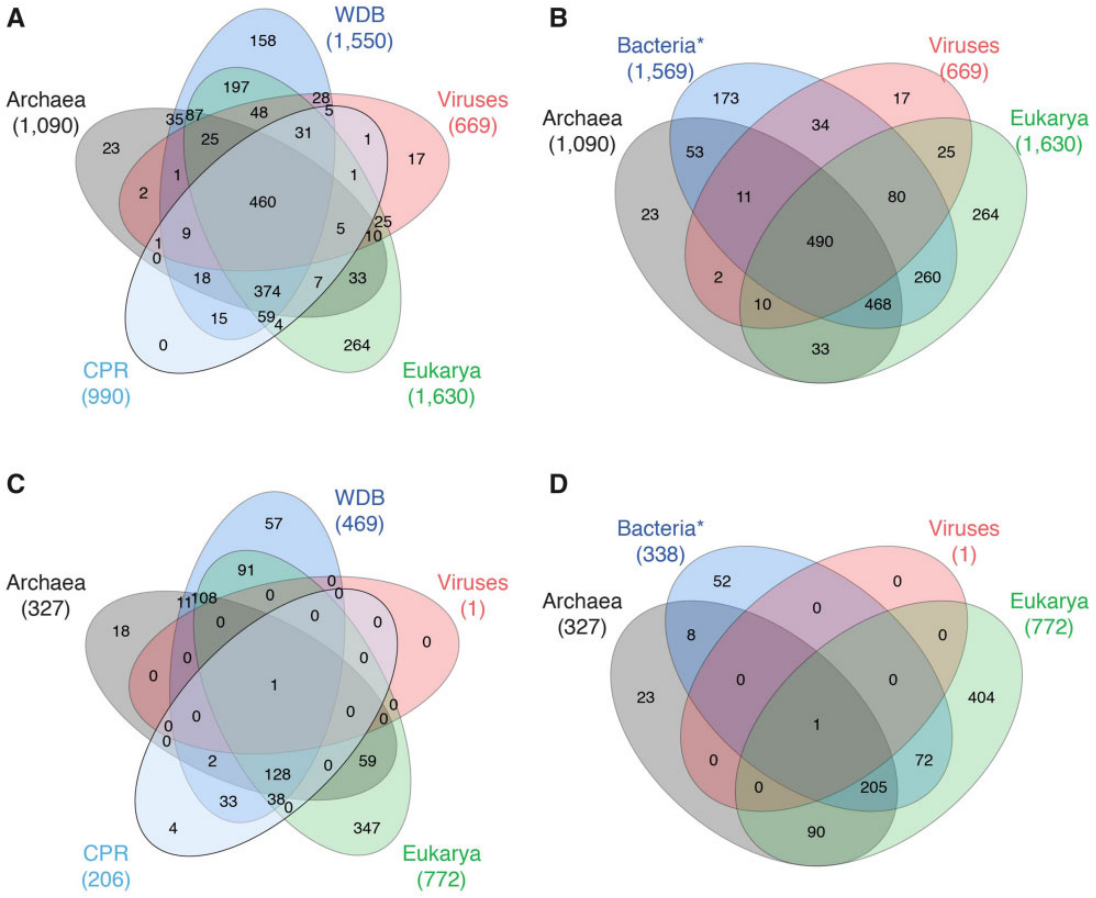

Fig. 1.—FSF sharing patterns in supergroups. (*A*) Distribution of all 1,943 FSFs in 5,019 proteomes from five supergroups. (*B*) Distribution of all FSFs in four supergroups. WDB and CPR pooled into Bacteria (indicated by asterisk). (*C*) Distribution of core FSFs in five supergroups. (*D*) Distribution of core FSFs in four supergroups. WDB and CPR pooled into Bacteria (indicated by asterisk). Numbers in parenthesis indicate total nonredundant FSFs detected in each supergroup.

N-terminal domain-like"], and a.29.9 ["LemA-like"]. Interestingly, the pili-subunits families were also found to be uniquely widespread among CPR in the analysis of Méheust et al. (2019). These proteins are likely involved in cell-to-cell adhesion and could therefore be essential for the predicted episymbiotic lifestyle of CPR (Méheust et al. 2019). A total of 267 out of 469 core WDB FSFs (57%) were absent in CPR, of which 153 (57%) performed metabolic functions such as amino acid, carbohydrate, lipid, and coenzyme metabolism and oxidation/reduction reactions (supplementary table S4, Supplementary Material online). This is confirmatory of previous findings of lack of key biosynthetic abilities in CPR (Brown et al. 2015; Castelle et al. 2018). A total of 150 out of these 153 metabolic FSFs were, however, assigned to the *ABCE* ($n = 75$), *ABCEV* (74), or *BCE* (1) Venn groups in figure 1*A* indicating that these FSFs were not completely absent in CPR but were present in very few CPR proteomes indicating a very patchy metabolic repertoire that indicates past or ongoing reductive evolution in CPR. The three remaining core bacterial FSFs that were completely absent from CPR included FSFs a.4.2 ["Enzyme I of the PEP: sugar phosphotransferase system HPr-binding (sub)domain"], d.50.2 ["Porphobilinogen deaminase (hydroxymethylbilane synthase), C-terminal domain"], and d.94.1 ["HPr-like"]. The former two were assigned to *ABE* and the latter to *ABEV* Venn groups (supplementary table S4, Supplementary Material online).

Finally, merging FSFs from CPR and WDB resulted in a core of 338 bacterial FSFs (fig. 1*D*). Notably, figure 1*B* suggested that FSF domains detected exclusively in Bacteria (CPR + WDB) and Eukarya (i.e., the *BE* Venn group) were significantly greater than the *AE* or *AB* Venn groups (260 vs. 33 and 53 FSFs) indicating a stronger evolutionary affiliation between Bacteria and Eukarya (Wang, Kurland, et al. 2011; Nasir and Caetano-Anollés 2013) rather than the more traditionally accepted Archaea and Eukarya sisterhood (Woese et al. 1990). Figure 1*D*, which is based on only the core FSF domain distribution in four supergroups revised these numbers. *BE* comprised only 72 core FSFs versus 90 in *AE* and 8 in *AB*. The 90 AE core domains included numerous metabolic and informational protein domains ($n = 29$ and 23, respectively) that were

previously detected in very few (i.e., <30%) members of WDB and CPR bacteria (supplementary table S5, Supplementary Material online). Hence, a focus on core domains indicates greater FSF sharing between Archaea and Eukarya whereas a focus on all domains indicates greater FSF sharing between Bacteria and Eukarya (fig. 1*D* vs. fig. 1*B*). Core domains, however, possibly ignore several lineage-specific gains or domains lost very early in evolution, scenarios that must be tested more formally.

CPR Bacteria Pursue an Economy-Driven Persistence Strategy

Proteome evolution is constrained by trade-offs between economy, flexibility, and robustness (Yafremava et al. 2013). To investigate these persistence strategies in our sampled proteomes, we used FSF *use* (i.e., the number of unique FSF domains detected in a proteome) as proxy for economy and FSF *reuse* (i.e., the total redundant count of FSF domains detected in a proteome) as proxy for flexibility. FSF *reuse* covers evolutionary processes such as gene duplication and HGT that can increase the count of same protein domain in proteomes. The ratio of flexibility to economy was taken as proxy for robustness (Nasir et al. 2014b). In all supergroups, we observed a roughly linear relationship between flexibility and economy (Adj. $R^2 > 0.6$ for all supergroups; fig. 2*A*). A unit increase in economy led to a 1.19-unit increase in flexibility in viruses, 1.7 in Archaea, 1.42 in CPR, 2.13 in WDB, and 3.93 in Eukarya (fig. 2*A*). Thus, CPR were intermediates between Archaea and WDB. Viruses represented the extreme of economy and eukaryotes the extreme of flexibility (see supplementary table S6, Supplementary Material online, for actual values). CPR pushed persistence toward economy and, in this regard, they were distinct from WDB that showed considerable flexibility. Indeed, CPR also had lowest robustness relative to all other cellular supergroups ($P < 0.05$, Wilcoxon two-tailed rank-sum test) (fig. 2*B*).

CPR Bacteria Evolve in an Archaea-like Manner

A push of persistence toward economy could either be the cause or effect of past or ongoing reductive evolution in CPR proteomes, as suspected previously (Castelle and Banfield 2018). Many known cellular species, specifically those that are endosymbionts of eukaryotic cells, and Archaea, are believed to have experienced reductive evolution in their history (McCutcheon and Moran 2012; Forterre 2013; Wolf and Koonin 2013). A similar model has also been proposed for giant virus evolution (Nasir, Kim, and Caetano-Anolles 2012; Nasir, Kim, and Caetano-Anollés 2012; Claverie and Abergel 2016). In general, FSFs that originated very early in evolution or at time 0 should have greater chances to multiply in genomes and spread to other lineages relative to recently originated domains. Therefore, we evaluated the spread of each FSF domain (*f*-value) in the evolutionary timeline (*nd*) for all supergroups. In these plots, deviations from the straight line can be indications of competing evolutionary scenarios such as genome reduction, HGT, extinctions, and evolutionary bottlenecks (fig. 3). These diagrams can therefore sometimes reveal more information than phylogenetic trees.

The *nd* versus *f* diagrams did not recover a linear pattern between the two variables, as expected from the strong effects of nonvertical evolutionary processes that are the fabric of organism and viral evolution (fig. 3). In Archaea, the distribution was extremely patchy with an early period marked by the high spread of ancient FSF domains followed by numerous periods of decline and rise in the *f*-values with the progression of time (fig. 3). In WDB, the relationship was somewhat linear but two clear halves could be identified. The first half included ancient FSF domains that are now widespread among WDB. The latter half included many rarely spread FSF domains in addition to several protein domains that have increased *f*-values recently (likely due to HGT homogenizing WDB proteomes) (Soucy et al. 2015) (fig. 3). Remarkably, the *nd* versus *f* distribution in CPR resembled that of Archaea, rather than WDB. As Archaea have likely experienced reductive evolution in the past (Forterre 2013), we extrapolate that CPR bacteria are evolving in an Archaea-like manner. Merging FSFs from CPR and WDB recovered a pattern similar to that of WDB, likely because CPR adds only~33% proteomes (777–2,281) to Bacteria and hence the *f*-values that use the number of total proteomes present in a supergroup as denominator are not significantly affected by their inclusion. Finally, Eukarya and viruses exhibited extreme tendencies in proteomic growth over time. In Eukarya, the pattern was clearly bimodal. Increases in *f*-values could be observed both very early and very late in evolution. Whereas, viruses exhibited no identifiable pattern. Very few FSFs were present in all viruses (only two that had *f*-value >0.6), and the majority were unique to specific lineages or even genomes.

Distinct Metabolism between CPR and WDB Bacteria

CPR exhibited Archaea- and virus-like persistence strategies and patterns of proteomic growth (figs. 2 and 3). To formally inspect these features, we divided each supergroup into major subgroups and extracted protein domains belonging to two major categories of molecular functions, "metabolism" and "information," as assigned by the SUPERFAMILY database (Chothia et al. 2003; Vogel et al. 2004; Vogel and Chothia 2006). We produced data matrices with sets of proteomes as rows and metabolic or informational protein domains as columns and generated heatmaps revealing the spread (*f*-values) of protein domains in proteome subgroups (fig. 4). In the metabolic heatmap (fig. 4*A*), we recovered two major clusters. The bottom cluster comprised of all CPR along with Planctobacteria, DPANN, NCLDV, and Tenericutes. All these non-CPR proteomes harbor small genome sizes and likely evolve via reductive evolution. The top cluster included

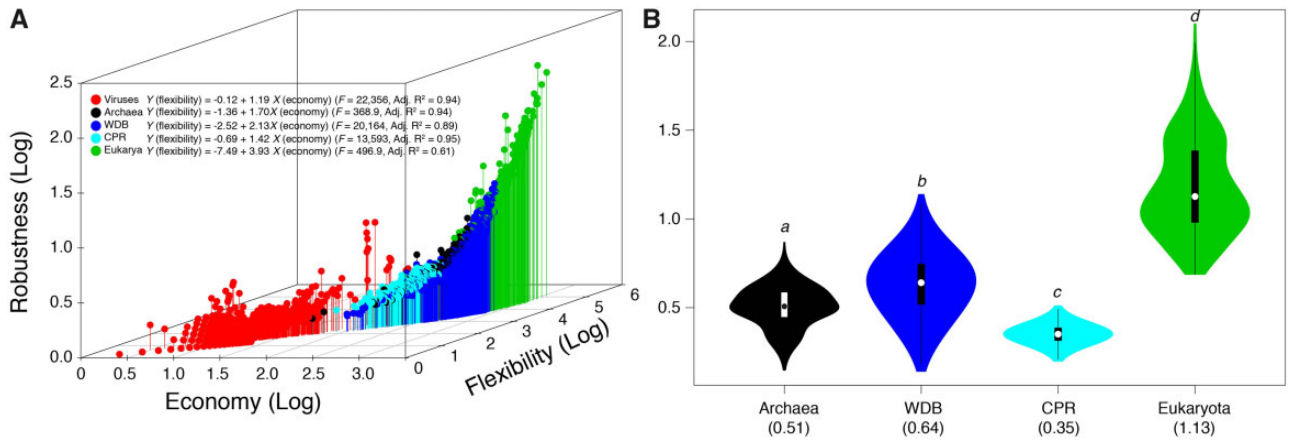

Fig. 2.—Persistence strategies in proteomes. (*A*) A 3D scatterplot displays relationships between economy, flexibility, and robustness for the five supergroups. Regression functions are listed. (*B*) Violin plots illustrate distribution of robustness parameter only for cellular supergroups. Numbers in parenthesis indicate median robustness (log-transformed). Different letters in italics indicate statistical significance ($P < 0.05$, Wilcoxon two-tailed rank-sum test).

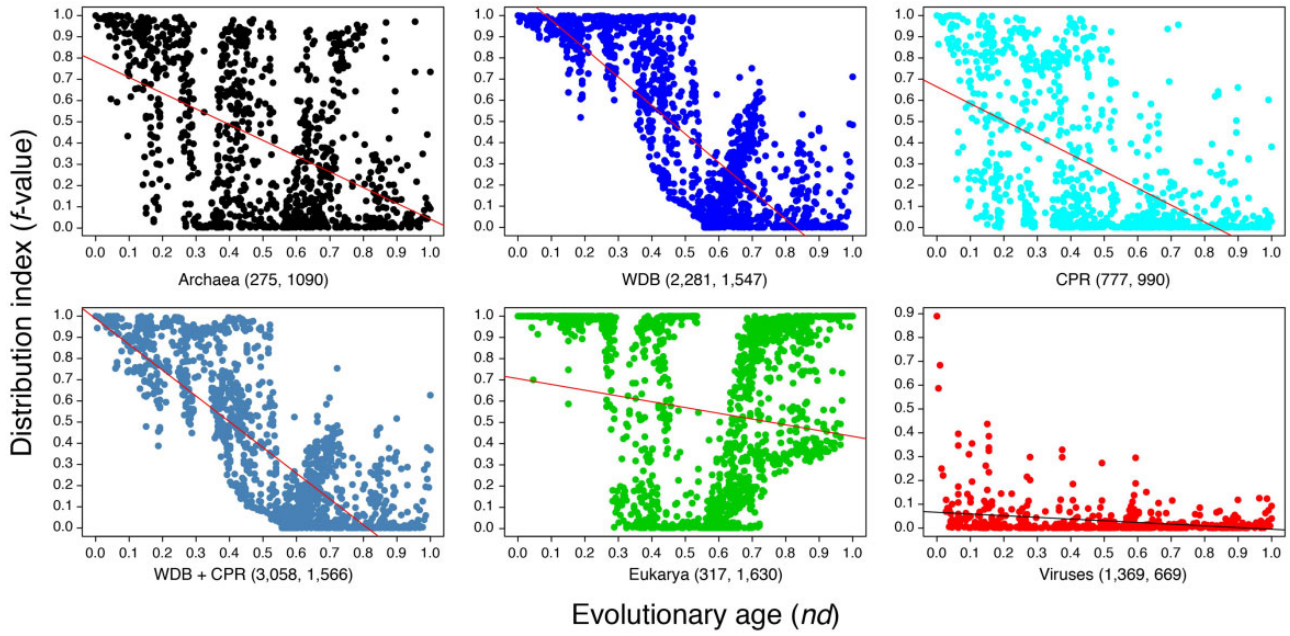

Fig. 3.—The spread (*f*-value) of FSFs in evolutionary time (*nd*) for all supergroups. The red line highlights the poor linear fit between the two variables and massive scatter around the line in all supergroups. Numbers in parentheses indicate the total number of genomes and FSF domains studied in each comparison (left, right).

proteomes from Archaea, WDB, and Eukarya. Bacteria and Eukarya are believed to share similar metabolism relative to Archaea. However, based on the *f*-values of metabolic protein domains present, prokaryotes clustered together. This is likely because of the higher *f*-values in the eukaryotic proteomes (fig. 3) relative to prokaryotes and also because of the relatively stronger reductive evolutionary tendencies in Archaea (Wang, Kurland, et al. 2011). Nevertheless, the metabolic heatmap highlighted that metabolic protein domains were rarely present or encoded by the bottom cluster comprising majorly of CPR. Hence, CPR bacteria can be distinguished from other cellular lineages, and especially WDB, on the basis of metabolism, which is confirmatory of previous findings (Brown et al. 2015; Méheust et al. 2019).

In turn, the informational heatmap also recovered two major clusters (fig. 4*B*). The bottom cluster included all WDB and CPR genomes (hence unified Bacteria). Here, Tenericutes clustered to the bottom of the CPR subcluster similar to the

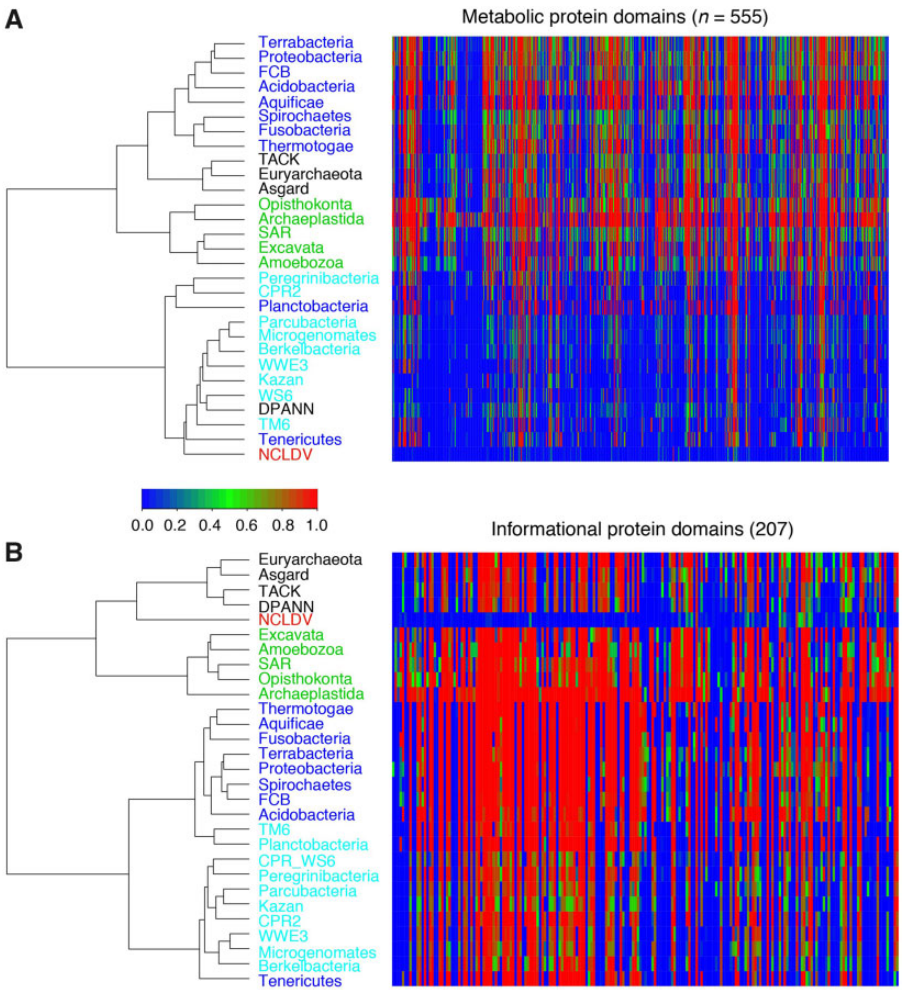

Fig. 4.—Clustering of proteome subgroups based on expression of metabolic and informational FSF domains. Heatmaps display the *f*-values of metabolic (*A*) and informational (*B*) protein domains. Dendrograms on the left show clustering patterns (Ward's hierarchical clustering method based on Euclidean distance matrix). Subgroups comprising <6 proteomes were not considered for this analysis. This excluded Kazan from CPR ($n = 5$ genomes) and Amoebozoa (5) and "other protists" (2) from Eukarya. For viruses, only NCLDV were included.

metabolic heatmap. However, DPANN did not. Thus, no mixing of Archaea and Bacteria was observed when informational domains were considered. The top cluster included Archaea and Eukarya as sister groups, whereas NCLDV clustered at the base of Archaea (albeit with very contrasting and lower *f*-values) (fig. 4*B*). The experiment therefore confirmed that the key separation between CPR bacteria and WDB was in metabolism. In CPR, very few metabolic protein domains are encoded and with very low *f*-values. These numbers match the distributions followed by DPANN, Tenericutes, and to an extreme extent by the NCLDV.

Metabolism and information are broad functional categories covering a wide range of molecular functions. According to the SUPERFAMILY functional classification scheme (Chothia et al. 2003; Vogel et al. 2004; Vogel and Chothia 2006), metabolism can be subdivided into 15 subcategories and information into 7. Figure 5 lists the mean *f*-values of all metabolic and informational domains corresponding to each subcategory in metabolism and information for all organisms and viruses. CPR consistently lacked FSFs involved in several major metabolism categories (see open or partially filled circles in fig. 5). For example, CPR had lower expression of FSFs involved in amino acids, nucleotide, and carbohydrate metabolism and transport, and oxidation/reduction relative to other sets of proteomes (excluding NCLDV) (fig. 5). They also completely lacked FSFs involved in electron transfer and transport, along with the absence of cell envelope metabolism and transport FSFs (fig. 5 and supplementary table S7, Supplementary Material online, for actual values). Indeed, none of the hitherto-reported CPR genomes encodes components necessary for cell envelope lipid synthesis (Castelle and Banfield 2018). For informational roles, CPR completely lacked FSFs involved in chromatin structure and dynamics and had lower spread of transcription-related FSFs. The

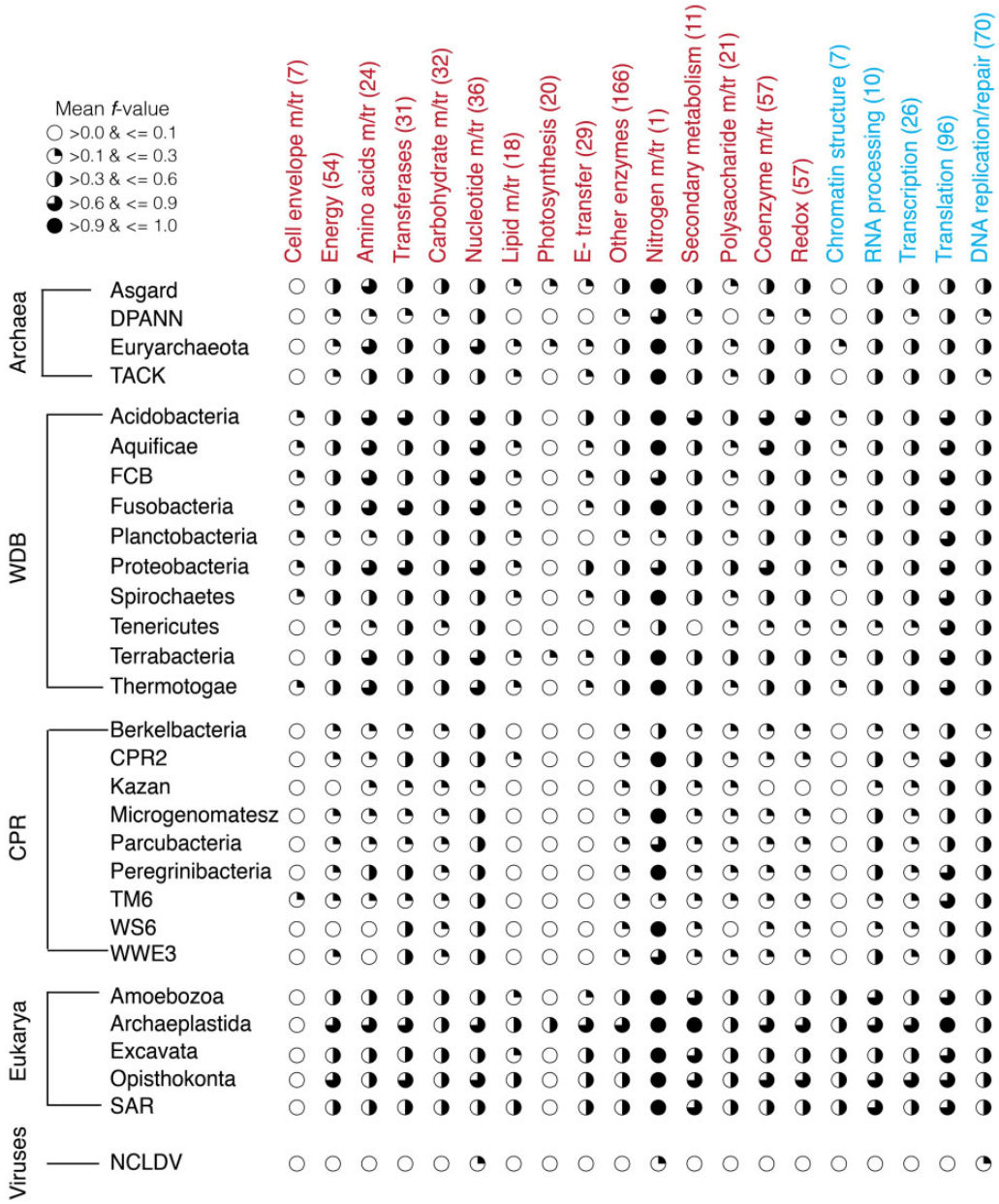

Fig. 5.—Presence/absence and expression of FSF domains corresponding to major subcategories of metabolism (red) and information (blue) detected in our sampled proteomes. Expression was calculated by averaging the *f*-values of all FSF domains in a particular subcategory. Numbers in parenthesis indicate total number of FSFs annotated to that subcategory (latest annotation on SCOP ver. 1.73 protein domains). See supplementary table S7, Supplementary Material online, for actual values. m/tr, metabolism and transport.

spread of FSFs involved in translation and DNA replication/repair was, however, roughly comparable to other prokaryotic microorganisms (fig. 5).

Protein Domain Gains Outnumber Losses in the Evolution of All Supergroups

To formally study the evolution of proteomes in each supergroup, we generated 100 random samples of proteomes for each supergroup (500 total samples). Each sample consisted of 100 randomly picked proteomes (taxa) from that supergroup except for viruses, for which we always kept all NCLDV ($n = 85$) and added 15 more viruses from four other viral subgroups (i.e., Caudovirales, Herpesvirales, Adenoviridae, and Baculoviridae). The subgroup proportions in random samples (approximately) matched their proportions in the full data set except for balancing and rounding purposes (see supplementary table S8, Supplementary Material online, for

subgroup proportions in the original and random data sets). Each sample had FSF domains (characters) that were detected in only that supergroup (based on fig. 1*A*). For phylogenomic analysis, we considered FSF *use* (*occurrence*) and FSF *reuse* (*abundance*) as two character state models and the Lundberg (Lundberg 1972) and outgroup rooting criteria (see Materials and Methods). For viruses, only NCLDV donated outgroup proteomes. For WDB, outgroup choices came from a filtered set of WDB genomes where multiple strains of the same species were excluded (reduced 2,281 WDB into 1,154). The filtered data set thus included all bacterial subgroups present in the full data set and ensured that diverse bacteria would be randomly picked as outgroups in each tree reconstruction. Each sample thus produced four different phylogenomic trees (two different rooting criteria and two different character state models) yielding a total of 2,000 phylogenomic trees reconstructed using the MP optimality criterion (see Materials and Methods). In each tree, we counted how many times every FSF domain character was gained or lost on the many branches of the supergroup phylogeny. For every FSF, we calculated the sum of score, summing up gains and losses of that FSF on all 400 trees for that supergroup. The sum of score could be <0 (indicating loss), equal to 0 (neutral), or >0 (indicating that the character was gained in the evolution of that supergroup).

In all supergroups, gains significantly outnumbered losses, especially in viruses (519 vs. 17, fig. 6*A*). These gains reflect increases in the domain *use* and *reuse* via multiple evolutionary processes such as HGT, gene duplication, and domain innovation and recruitment that have accumulated over billions of years. In turn, losses represent (possibly) irreversible one-time events in organism evolution and hence comprise a smaller proportion relative to gains. Losses are further underestimated because protein domains that are now completely absent from extant proteomes (due to evolutionary bottlenecks, extinction, and ancient losses in the last common ancestors) were not part of the supergroup FSF repertoires (fig. 1*A*) and hence not considered in tree reconstructions (supplementary table S9, Supplementary Material online, for gain and loss statistics in all supergroups).

In CPR, the proportions of domains in either the "lost" or "lost+neutral" categories were the highest ($>22\%$) followed by Archaea and WDB (fig. 6*B*). A total of 49 FSF domains were common losses in Archaea and CPR relative to 38 common losses in WDB and CPR (fig. 6*C*). In turn, 112 gains were common to CPR and WDB and 79 between CPR and Archaea (fig. 6*C*). This implied that both Archaea and CPR, in general, lost similar FSFs whereas CPR and WDB, in general, gained similar FSFs, providing further support to the initial idea that CPR bacteria have evolved in an Archaea-like manner . In contrast, neutral FSFs were all unique gains in Archaea, WDB, and CPR (see supplementary table S10, Supplementary Material online, for FSF IDs). Note also that a significant number of FSFs (i.e., 133 out of 669, 19.8%) were classified as neutral in viruses (fig. 6*B*). Remarkably, 69% of the neutral viral FSFs (92 out of 133) evolved relatively late in evolution (*nd* > 0.5) and 33% (45 out of 133) performed unknown or viral functions, in addition to several that are involved in toxins, defense, antiviral response, and protein–protein interactions (supplementary table S9, Supplementary Material online) suggesting that these could be relatively recent lineage-specific virus gains crucial for viral pathogenicity. Therefore, we calculated the gain-to-loss ratios for each FSF domain for three distinct periods in evolutionary time. The *early* period was marked by *nd* values >0 but ≤ 0.3, *middle* marked by *nd* >0.3 but ≤ 0.7, and *late* marked by *nd* >0.7 but ≤ 1. Indeed, the median values always increased in the order, *early*, *middle*, and *late* (except for Eukarya) and the differences were greatest for CPR (fig. 6*D*) suggesting protein gains accumulated over evolutionary time (Nasir et al. 2014b) and marking the *early* periods by relatively greater reductive tendencies in prokaryotic and viral proteomes (fig. 6*D*).

Our experiment with the random selections of outgroups from diverse organisms and viruses also helped us identify which proteomes were the most parsimonious additions as outgroups to the supergroup trees (highlighted rows in supplementary table S11, Supplementary Material online, fig. 7). In Archaea, Uhrbacteria (CPR) yielded the most parsimonious trees for either *occurrence*- or *abundance*-based tree reconstructions. For WDB, the most parsimonious outgroups were *Sulfolobus islandicus* (Crenarchaeota) for *abundance* and *Halorubrum lacusprofundi* (Euryarchaeota) for *occurrence*. For CPR, Thaumarchaeota and DPANN Nanoarchaeota yielded the most parsimonious trees for *abundance* and *occurrence*, respectively. Finally, for Eukarya these choices were CPR Yanofskybacteria and Proteobacteria, and for viruses, these were SAR apicomplexans and unknown CPR Parcubacteria (supplementary table S11, Supplementary Material online). These experiments can possibly help inform the rooting (correct outgroup?) of the tree of life. In the network depiction of these relationships (fig. 7), the hubs are Archaea and CPR, indicating that they either are the most ancient organisms or harbor highly tailored proteomes.

Evolutionary Relationships

Next, we reconstructed phylogenomic trees to describe the evolution of proteomes from all supergroups. Taxa and character selection choices can influence phylogenetic relationships (Zwickl and Hillis 2002; Heath et al. 2008). Therefore, we attempted to minimize these biases by randomly sampling an equal number of 100 proteomes from each supergroup and repeating the tree reconstruction exercise using four different choices of character subsets. As it has become standard practice to build phylogenies from informational genes (e.g., 16S rRNA, ribosomal proteins) (Woese and Fox 1977; Woese et al. 1990; Hug et al. 2016), we first concatenated 209 FSF domains involved in informational processes (chromatin

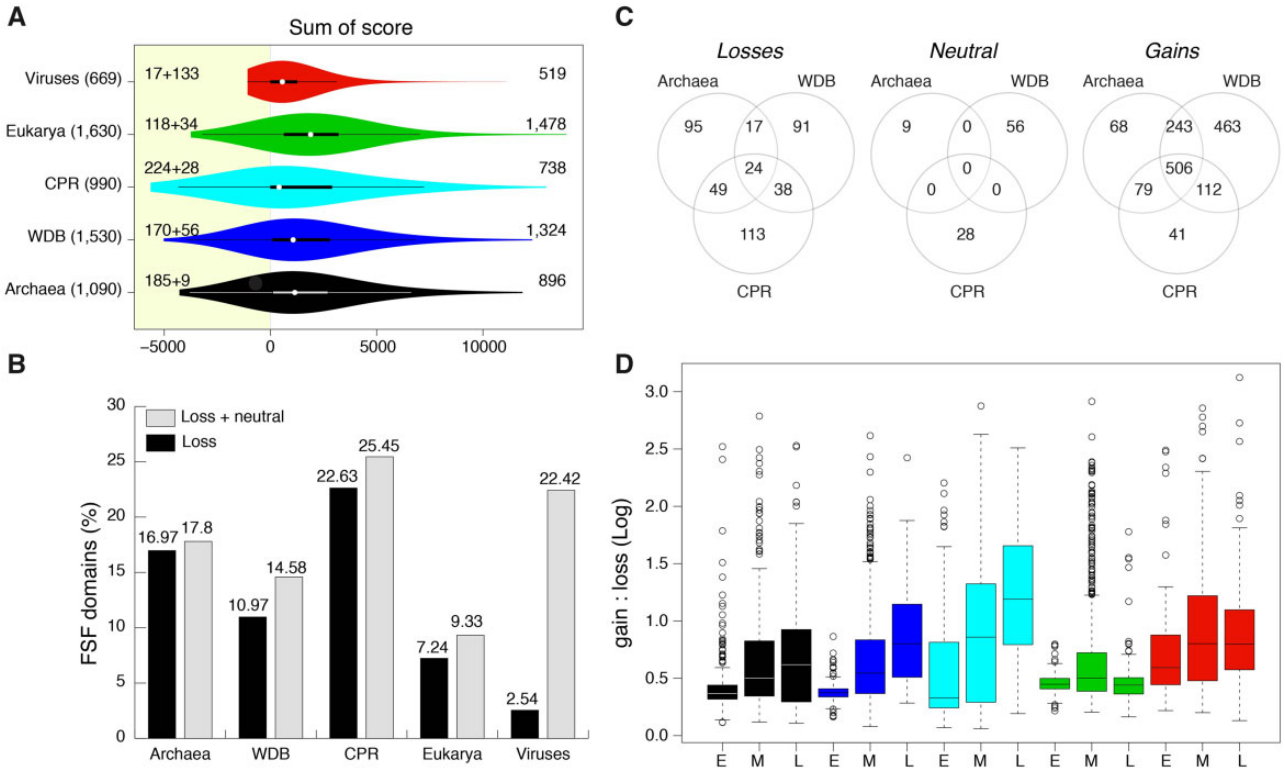

Fig. 6.—Patterns of protein domain gain and loss in supergroups. (*A*) Violin plots display sum of gains and losses for each FSF domain (character) in supergroup proteome (taxa) trees. Gains and losses were summed up from each of the 400 tree reconstructions generated for each supergroup. White circles indicate group medians. Numbers in parentheses indicate total FSF domains in each supergroup. Numbers inside the figure indicate numbers of FSFs classified as lost, neutral, or gained based on the sum of score, respectively. Shaded region highlights lost or neutral domains. (*B*) Bar plots indicate proportions of protein domains classified as either lost or lost + neutral in each supergroup. Proportions calculated by dividing the number of lost + neutral FSFs by the total number of FSFs in each supergroup. (*C*) Venn diagrams indicate how many FSFs classified as either lost, neutral, or gained were common among Archaea, WDB, and CPR (see supplementary table S10, Supplementary Material online, for FSF IDs). (*D*) Boxplots display distributions of log-transformed gain to loss ratios for protein domains in the *early*, *middle*, and *late* evolutionary periods for each supergroup. FSF domains for which *nd* values were not available ($n = 3$) and for which sum of losses was equal to 0 were excluded from the analysis. Horizontal lines in each box indicate group medians. E, *early*; M, *middle*, L, *late*.

structure, DNA replication, recombination and repair, RNA binding and processing, transcription, translation, and tRNA metabolism) to generate a phylogenomic data matrix that contained 500 proteomes (taxa) and 209 informational FSFs (characters). We considered both FSF *occurrence* and *abundance* as possible character state models and therefore built two separate phylogenomic trees. As there is no valid outgroup to root the trees of life, we restricted our phylogenies to the Lundberg rooting method (Lundberg 1972).

In the *occurrence* tree, viruses branched off early from the cellular supergroups in a paraphyletic manner with 96% BS (fig. 8*A*). Within the cellular supergroups, CPR clustered with WDB (although some CPR proteomes were oddly placed at the base of the archaeal–eukaryal subtree) and occupied basal positions within the bacterial subtree. Archaea and Eukarya clustered together, similar to most sequence-based phylogenies of informational genes (Da Cunha et al. 2017). The monophyly of Eukarya was supported by 85% BS, whereas Archaea were paraphyletic. Similar clustering patterns were also recovered in the *abundance* reconstruction, albeit with higher BS support for paraphyletic viruses (97%) and monophyletic Eukarya (99%) (fig. 8*A*). As the metabolic heatmap had separated CPR and WDB (fig. 4), we repeated the phylogenomic reconstruction using a concatenation of 555 metabolic FSF characters (fig. 8*B*). The metabolic phylogeny indeed separated CPR and WDP in both the *occurrence* and *abundance* reconstructions. The metabolic phylogeny overall indicated "five" distinct supergroups rather than four and suggested distinct origins of CPR. Moreover, the metabolic phylogeny supported a closer evolutionary relationship between Bacteria and Eukarya in contrast to the informational phylogeny that supported the traditional "Woesian" Archaea–Eukarya sisterhood (Woese et al. 1990). We have previously recovered Bacteria–Eukarya clustering to the exclusion of Archaea in a number of protein structure and function-based phylogenomic studies (Kim and Caetano-Anollés 2011, 2012; Kim, Nasir, Hwang, et al. 2014; Nasir et al. 2014a; see Caetano-Anollés et al. 2014; Staley and

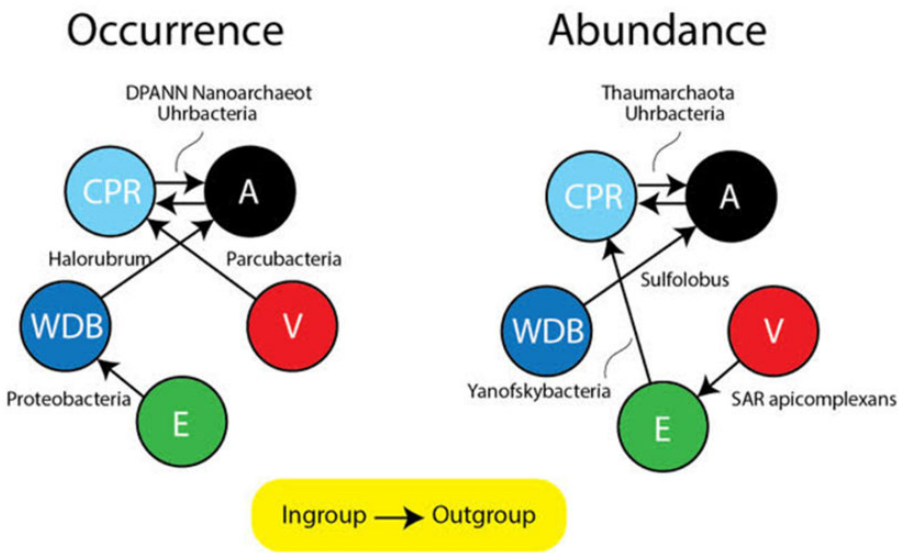

Fig. 7.—A network depiction of the most parsimonious ingroup–outgroup relationships.

Caetano-Anollés 2018 for recent reviews). Therefore, to further inspect the disagreement between the informational and metabolic protein phylogenies, we reconstructed two additional phylogenies using *all* (1,943) and *universal* or *ABCEV* (460) characters (fig. 8*C*).

In the *all* phylogeny based on either FSF *occurrence* or *abundance*, we recovered clustering patterns that largely resembled the informational tree. Viruses branched off early, again in a paraphyletic manner (>95% BS). Archaea and Eukarya formed sister groups, whereas CPR and WDB formed a paraphyletic distinct group. CPR again occupied the basal positions in the bacterial subtree (fig. 8*C*). In fact, these patterns were also conserved in the *universal* phylogeny built from FSF *occurrence* (fig. 8*D*). We observed monophyletic Archaea and monophyletic Eukarya (85% BS). Up till now, *occurrence* and *abundance* reconstructions largely agreed with each other. However, and surprisingly, *abundance* reconstruction for universal characters recovered clustering patterns that were similar to the metabolic phylogeny (fig. 8*D*).

In summary, we always observed distinct paraphyletic origins of viruses regardless of the choice of the character subset or character state model. The relationships among cellular supergroups largely depended on the choice of character subset used to build the tree. The phylogenies obtained from either informational or all protein domains supported Archaea/Eukarya sisterhood, similar to several previous sequence-based phylogenetic reconstructions. In turn, phylogenies built from metabolic protein domains or the *reuse* of universal characters indicated a "five-way" tree of life where CPR were separated from WDB and WDB and Eukarya formed sister groups. None of the trees, however, indicated an origin of Eukarya from within Archaea (Spang et al. 2015; Zaremba-Niedzwiedzka et al. 2017), although Archaea were paraphyletic in most reconstructions.

Finally, to inspect the evolutionary relationships within each supergroup, we extracted subtrees for each supergroup from the *all* phylogeny (fig. 8*C*) built from FSF *occurrence*. Within the archaeal subtree, DPANN occupied basal positions. Asgard clustered at the base of Euryarchaeota, and Thermococcales (Group I Euryarchaeota) were separated from the rest of Euryarchaeota (fig. 9*A*). In Eukarya, Opisthokonta formed a unified group (fig. 9*D*). Within Opisthokonta, monophyly of Fungi was supported by 91% BS and monophyly of Metazoa was supported by 72% BS. Land plants clustered with green algae (77% BS), whereas basal positions were occupied by amoebozoa, apicomplexans (SAR), and kinetoplasts (Excavata). The clustering patterns for WDB, CPR, and viruses are also shown in figure 9.

A Multidimensional Scaling Approach to Cluster Proteomes in Evolutionary Space

As the phylogenetic trees revealed contrasting results, we used a different method to analyze the evolutionary groupings of proteomes. The proteomes are made up of individual protein domains (parts), where each domain (part) contributes some functionality to the functioning of the overall system. In the case of proteomes, these parts were added at different timepoints in evolution. Therefore, if we know the evolutionary ages of all parts in a system, we can infer the evolutionary age of the system (Caetano-Anollés et al. 2018). We therefore used the *nd* values of *universal* FSF domains as proxies for their evolutionary ages and used multidimensional scaling approaches to analyze the clustering of proteomes in 3D space (fig. 10). This method has been previously named the

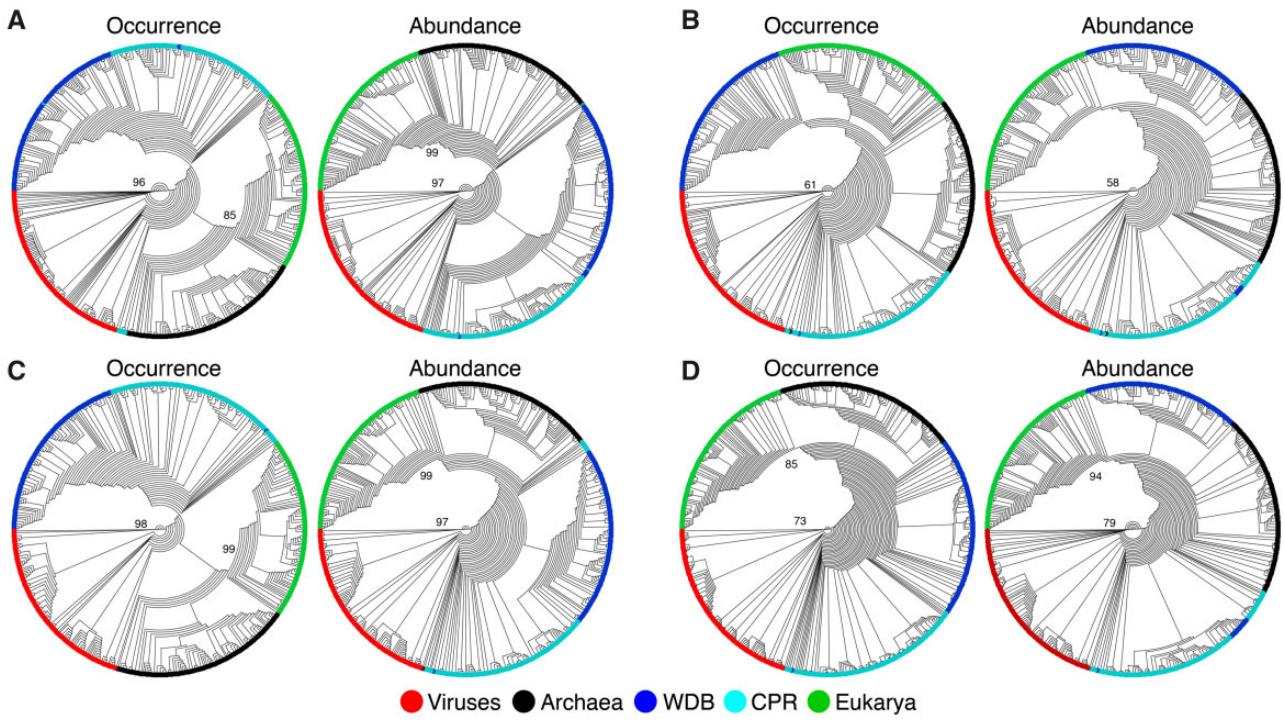

Fig. 8.—Evolutionary relationships among supergroups inferred from informational, metabolic, all, and universal domain character sets. (*A*) Single most parsimonious phylogenomic trees based on FSF *occurrence* ($n = 209$, parsimony informative = 206, tree length = 3,716, retention index = 0.87, $g_1 = -0.106$) and *abundance* ($n = 209$, parsimony informative = 206, tree length = 16,766, retention index = 0.83, $g_1 = -0.18$) of informational FSF domains. (*B*) Single most parsimonious phylogenomic trees based on FSF *occurrence* ($n = 555$, parsimony informative = 536, tree length = 13,022, retention index = 0.81, $g_1 = -0.08$) and *abundance* ($n = 555$, parsimony informative = 536, tree length = 59,240, retention index = 0.79, $g_1 = -0.06$) of metabolic FSF domains. (*C*) Single most parsimonious phylogenomic trees based on FSF *occurrence* ($n = 1{,}943$, parsimony informative = 1,803, tree length = 34,027, retention index = 0.82, $g_1 = -0.11$) and *abundance* ($n = 1{,}943$, parsimony informative = 1,803, tree length = 152,628, retention index = 0.80, $g_1 = -0.08$) of *all* FSF domains. (*D*) Single most parsimonious phylogenomic trees based on FSF *occurrence* ($n = 460$, parsimony informative = 460, tree length = 14,683, retention index = 0.80, $g_1 = -0.24$) and *abundance* ($n = 460$, parsimony informative = 460, tree length = 75,681, retention index = 0.80, $g_1 = -0.18$) of *universal* FSF domains. Taxa names not displayed, as they would not be legible. Numbers on branches indicate BS support values only for deepest splits, when available.

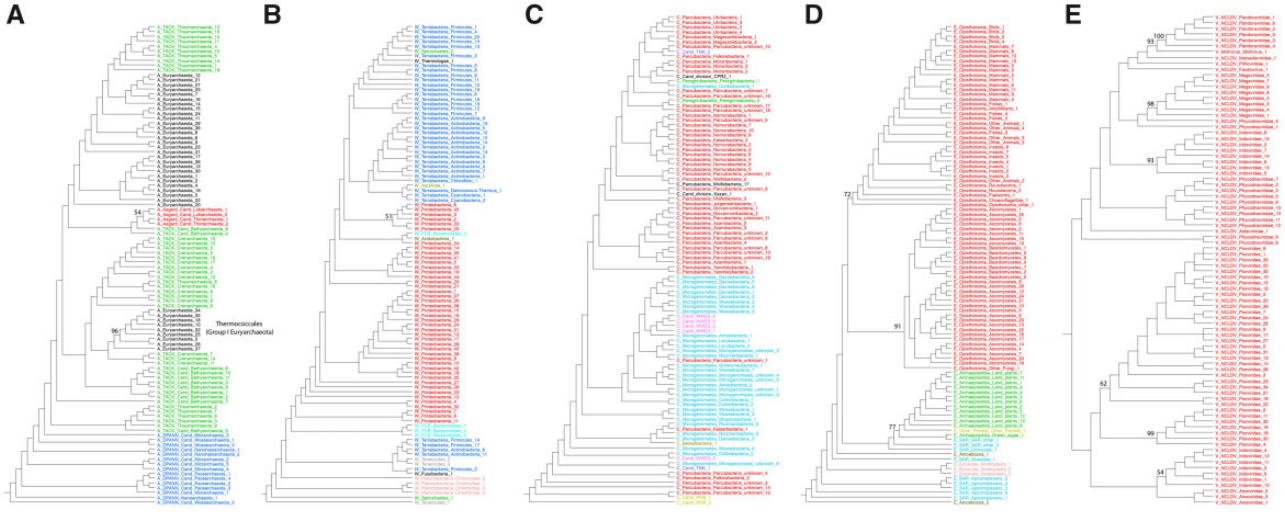

Fig. 9.—Phylogenomic trees of individual supergroups. Trees for Archaea (A), WDB (B), CPR (C), Eukarya (E), and viruses (V) were extracted from the *all* phylogeny (FSF *occurrence*) in figure 8C. BS support values are shown, when available. Taxa names are replaced by codes to ease understanding of clustering patterns.

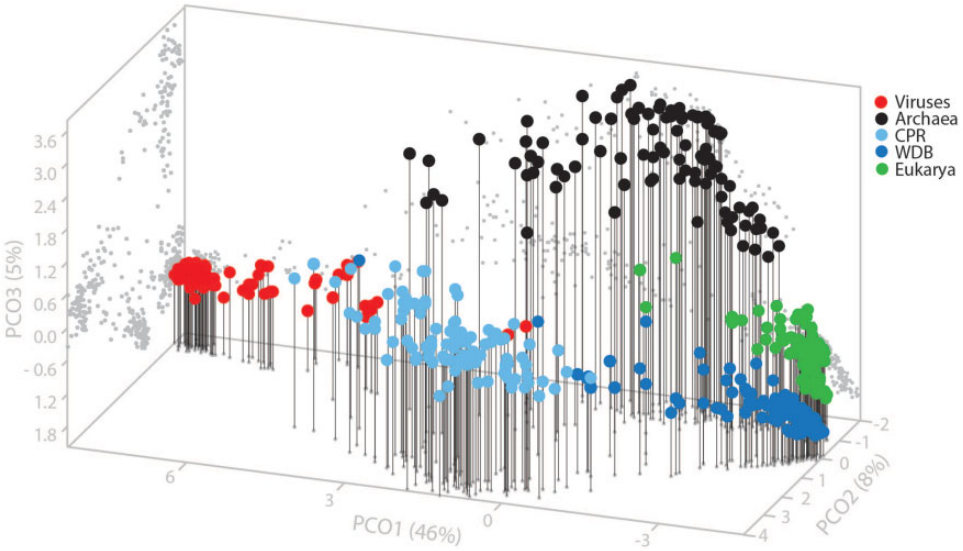

Fig. 10.—The temporal flow and principle of continuity from viruses to eukaryotes. Three most significant axes are displayed. Proteome data points were colored as previously described. Numbers in parenthesis indicate the percentage variability explained by each axis.

evoPCO (Nasir and Caetano-Anollés 2015). The PCO1, which explained 46% of total variability in the data set, indicated a temporal flow from viruses at the extreme left to WDB and Eukarya at the extreme right (fig. 10). Remarkably, CPR and Archaea were in the middle of the temporal flow suggesting that both have been subjected to the same kind of evolutionary constraints. This temporal flow supports the evolutionary principal of continuity or gradual shift from viruses to eukaryotes in the evolutionary timeline. The PCO3 accounted for only 5% of the variability, but it nicely dissected Archaea from the rest. PCO3 also revealed a decreasing straight line unifying the clouds of CPR and WDB, with WDB ending lower than Eukarya (fig. 10). This line can be taken as additional evidence in favor of the inclusion of CPR with WDB. Finally, PCO2, which accounted for 8% of the variability dissected three clouds: CPR, WDB, and viruses–Eukarya–Archaea. The PCO2–PCO3 projection at the left tells almost the entire story to complement PCO1, that is, four clouds of CPR, WDB, Archaea, and Eukarya–viruses.

Discussion

In this study, we compared the distribution of protein domain FSFs in >5,000 proteomes sampled from viruses, CPR, WDB, Archaea, and Eukarya. The data set thus included highly diverged, distantly related, and many fast-evolving proteomes sampled from all supergroups. These complex data sets are difficult to analyze using traditional approaches based on gene/protein sequence alignments. The obvious challenges are to identify sets of orthologous genes conserved across a wide range of proteomes and then to reliably align those genes/proteins to produce a workable alignment for phylogenetic studies. As shown by Holmes and Duchêne (2019), recovering such alignments can sometimes become an impossible task. We therefore mapped protein sequences to predefined SCOP superfamilies (Andreeva et al. 2007) rather than classifying or clustering protein sequences into families or superfamilies *de novo*. The unit of classification in SCOP is protein domains rather than whole proteins. In our view, this distinction is important as protein domains are independent evolutionary units within proteins that can be gained/lost at unequal rates (Nasir et al. 2014b). Importantly, the various families within a SCOP superfamily often have little or negligible sequence identities. Instead, they are united by the presence of conserved structural cores/backbones and molecular functions, which support their high conservation (Andreeva et al. 2007). The benefits of using molecular structure and a protein domain-based census in evolutionary studies have been discussed previously (Caetano-Anollés and Nasir 2012). For brevity, the method is free from the challenges posed by alignment-dependent phylogenetic methods that can be biased by low-quality alignments containing several gaps (Holmes and Duchêne 2019), imbalanced taxon sets (Nasir et al. 2016), and the failure to sample all members of the homologous protein sets, when data sets include diverged proteomes. On balance, protein structures may fail to properly resolve the relatively recent evolutionary relationships (e.g., HIV evolution) owing to their very high conservation, as also evident from the poorly resolved phylogenetic trees of individual supergroups (fig. 9). In such circumstances, protein/gene sequences will undoubtedly outperform protein structures.

Our objective was to place CPR proteomes in the tree of life along with proteomes from other prokaryotes, viruses, and eukaryotes. Previous phylogenetic studies positioned CPR to the base of the bacterial tree of life (Hug et al. 2016) and

separated CPR from WDB based on differences in the content and composition of metabolic protein families (Méheust et al. 2019). Our superfamily-focused analysis recovered largely similar results. Despite their "unusual" biology (Brown et al. 2015), we could confidently place CPR proteomes within the bacterial phylogeny in most tree reconstructions where they formed basal paraphyletic groups (indicating missing members) similar to previous sequence-based tree reconstructions (Hug et al. 2016). We noted that CPR added only 1.22% new FSFs to the bacterial repertoire and the majority of CPR-encoded FSFs included proteins that could be considered "universal" among cells (fig. 1*A*). Only four CPR-specific core FSFs were identified. These included pili subunits involved in cell-to-cell interactions, which could be important for establishing the episymbiotic lifestyles of CPR, as previously predicted (Méheust et al. 2019). There were no CPR-specific protein domains, which is an important criterion for the classification of unique group or domain status. The maximum two-group sharing between CPR and any other supergroup was between CPR and WDB. Thus, we conclude that CPR belong to the bacterial tree of life and are bona fide bacteria. However, and despite their bacterial affiliation, we argue that CPR have probably evolved in an Archaea-like manner. We observed that CPR and Archaea experienced common FSF losses over the course of evolution whereas CPR and WDB experienced common gains. The losses in CPR mostly involved metabolic protein domains. The clear distinction between WDB and CPR bacteria was indeed in their metabolic repertoires. In fact, a phylogeny reconstructed from only the metabolic protein domains separated WDB and CPR whereas a phylogeny reconstructed from informational FSFs grouped WDB and CPR (fig. 8).

The metabolic and informational phylogenies also differed in how they grouped Archaea, Bacteria, and Eukarya. For example, we recovered a canonical "Woesian" tree from *informational* and *all* protein domain character sets and not when *metabolic* FSF domains were considered (fig. 8). The phylogenies from *universal* characters further differed in the choice of character state model. They supported informational/all trees when *occurrence* was used and supported metabolic phylogeny when *abundance* was used. These are interesting and conflicting observations that bring to light the issue of character sampling and overall show that tree of life reconstruction is a challenging task. Informational machineries are very similar in Archaea and Eukarya (Lecompte et al. 2002), and hence their trees support the topologies of rRNA (also part of informational processes) and ribosomal protein concatenation trees (Hug et al. 2016). In turn, the *all* domain character set includes several domains that are unique to a particular supergroup (264 in Eukarya, 158 in WDB, 23 in Archaea, 17 in viruses, and 0 in CPR, altogether comprising roughly one-fourth of total FSFs). These characters cannot be compared across proteomes of other supergroups as they are absent from their proteomic repertoires. Following the same rationale, the, *universal* characters are detected in all supergroups and hence may provide a fairer comparison. These subsets support the Bacteria–Eukarya sisterhood and the early paraphyletic rise of Archaea (Staley and Caetano-Anollés 2018) when we use *abundance*-based character state model. *Abundance* is a more complex and complete model than *occurrence* that merely describes presence/absence. These issues are important to recognize for phylogeny reconstructions as there is no one-size-fits-all solution to produce trees of life. Although some of our trees indicated Archaea–Eukarya sisterhood, none of the trees supported the "two-domain" or "eocyte" tree that proposes the origin of eukaryotes from within Archaea (Lake et al. 1984; Spang et al. 2015; Zaremba-Niedzwiedzka et al. 2017). This view was boosted by the recent description of Asgard archaeal lineages that encode several eukaryote-specific proteins and were proposed to be the ancestors of eukaryotes (the two-domain tree) (Zaremba-Niedzwiedzka et al. 2017; see Nasir et al. 2016; Da Cunha et al. 2017, 2018; Spang et al. 2018; Williams et al. 2020 for ongoing debate and work).

CPR are characterized by small genomes and cell sizes and possibly a preference for (epi)-symbiotic lifestyles (Castelle and Banfield 2018; Castelle et al. 2018). We have previously argued that the inclusion of (obligate)-parasitic organisms with free-living members of the same supergroup in the tree of life can sometimes cause conflicts (Nasir et al. 2017). There is a tendency in the (obligate)-parasitic organisms to reduce their genome sizes by the loss of metabolic protein domains, as they increase dependency on their hosts (Nasir et al. 2011). Parasitic organisms can thus harbor a somewhat different profile of character state distribution relative to the free-living members of the same supergroup, which may confuse evolutionary groupings within a supergroup. This problem, however, does not extend to members across different supergroups because proteome profiles within a supergroup are always more similar than profiles between supergroups, regardless of the organism lifestyle. For example, each supergroup possesses several domains that are unique to that supergroup (fig. 1*A*). This is evident from figure 8 where DPANN archaea and CPR bacteria, despite having reduced and smaller genomes, are not clustered together and cause no distortions to the tree phylogeny. In other words, the groupings of proteomes in phylogenomic trees is not based on genome size, as incorrectly understood by Harish et al. (2016) and Harish and Kurland (2017), but is derived from the profile distribution of shared characters across proteomes (Nasir et al. 2017; Caetano-Anollés et al. 2018; Caetano-Anollés et al. 2019). This issue does not apply to viruses, as there are no free-living viruses to cause conflict within the viral supergroup (Nasir et al. 2017). We note that when testing empirical support for fully reversible models of character state evolution of the standard ordered characters that we here use, models matched the reconstructed frequencies of character change and were faithful to the distribution of serial

homologies in Wagner parsimony trees built from domain counts in proteomes (Caetano-Anollés et al. 2019). In sharp contrast, nonreversible models of the type used by Harish et al. (2016) and Harish and Kurland (2017) countered trends in the data they had to explain, violating the triangle inequality of distances, and attracted organisms with large proteomes to the base of the rooted trees. Our models were free of these artifacts, see also Kim, Nasir, and Caetano-Anollés (2014). We also note that biases induced by parasitic lifestyle and problems of character independence challenge tree of life reconstructions and can be offset by reconstructing phylogenies that describe the evolution of parts that are shared or are unique to biological systems, such as the evoPCO-based numerical analyses (Caetano-Anollés et al. 2018).

Among other contentious issues, our phylogenies placed viruses at the base of the tree of life (fig. 8), a result we have previously obtained (Nasir and Caetano-Anollés 2015; Colson et al. 2018). Viruses have been difficult to classify and compare with cellular organisms primarily due to the fast evolution of viral genes and their smaller genomes. The discovery of "giant" viruses (La Scola 2003; Philippe et al. 2013; Abergel et al. 2015; Abrahão et al. 2018) that surpass many cells in particle and genome size and encode numerous proteins related to protein translation (Colson et al. 2018) and a focus on using protein structures (more conserved) to classify viruses (Abrescia et al. 2012; Nasir and Caetano-Anollés 2017) has apparently overcame some of these limitations. Some authors believe that giant viruses have primarily evolved via gene capture from host cells and are "gene robbers" (Moreira and Brochier-Armanet 2008; Moreira and Lopez-Garcia 2009). In our view, these are oversimplified generalizations of how viruses evolve because the majority of viral genes lack any identifiable homolog in host cells (Legendre et al. 2018, 2019) and viruses encode several virus-specific protein folds that are absent from cells (fig. 1). These facts are difficult to reconcile with the "virus pickpocket" paradigm, as argued by Forterre (2016). In fact, virus-to-cell gene transfer may possibly exceed gene transfer in the opposite direction (Forterre 2011; Malik et al. 2017). Philosophically, the status of viruses as nonliving border-line entities is now being abandoned (Forterre 2016) and the distinction between the intracellular and extracellular stages of virus reproduction cycle is now being emphasized to recognize viruses as living organisms and their potential as gene creators, not robbers (Forterre 2011).

Finally, we took precomputed FSF assignments of eukaryotic genomes from a previous study (Nasir and Caetano-Anollés 2015). It can be argued that several eukaryotic genome assemblies, especially of mammals, are regularly updated. This is, however, not expected to drastically change the landscape of FSF assignment differences between eukaryotes and prokaryotes and viruses. Our experience with using FSF counts in proteomes over the past decade and a half has shown that genome assignments tend to be reliable and stable over time and even when using different classification schemes such as SCOP versus CATH (Bukhari and Caetano-Anollés 2013). Although the actual numbers of estimated FSFs in supergroups can differ from one study to another, the relative patterns of differences and similarities tend to remain conserved. Moreover, all CPR genomes analyzed in this study were sequenced at the "draft" level. In comparison, the majority of viral, WDB, and eukaryotic proteomes were "complete" and of "reference" quality. These facts can somewhat bias the numerical estimates of protein superfamily content and composition in proteomes. Though, a total of 510 out of ∼800 CPR genomes analyzed in this study were sequenced to a "near-complete" level. We are therefore hopeful that we have not missed a significant number of CPR domains. Availability of more and complete genome assemblies will undoubtedly improve our inferences. Finally, maximum-likelihood methods now allow gain and loss tracings of protein domains in phylogenies (Librado et al. 2012). These experiments can supplement parsimony-based analyses of protein domain gain and loss in phylogenies. This is a challenge we wish to undertake in a separate study.

In conclusion, the protein structure-based view of CPR proteomes supports their distinct status within the bacterial domain of life (Méheust et al. 2019). CPR proteomes are characterized by the loss of key metabolic features (Castelle et al. 2018) and form basal branches in the bacterial trees of life (Hug et al. 2016). CPR have likely evolved via dramatic reductive evolution, tendencies typically seen in the archaeal proteomes (Forterre 2013). These results therefore largely reconcile with the previous views regarding CPR origins and evolution (Méheust et al. 2019) and support the idea that protein structures can supplement, support, and even improve evolutionary inferences (Caetano-Anollés and Nasir 2012).

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We thank Patrick Forterre for critical reading of the manuscript. AN is recipient of the Los Alamos National Laboratory Oppenheimer Fellowship (20180751PRD3). Research was supported by a grant from the Collaborative Genome Program (20140428) funded by the Ministry of Oceans and Fisheries, Korea to K.M.K., and grants from the National Science Foundation (OISE-1132791) and the National Institute of Food and Agriculture of the United States Department of Agriculture (ILLU-802-909 and ILLU-483-625) to G.C.-A.

Literature Cited

Abergel C, Legendre M, Claverie J-M. 2015. The rapidly expanding universe of giant viruses: mimivirus, pandoravirus, pithovirus and mollivirus. FEMS Microbiol Rev. 39(6):779–796.

Abrahão J, et al. 2018. Tailed giant Tupanvirus possesses the most complete translational apparatus of the known virosphere. Nat Commun. 9(1):749.

Abrescia NGA, Bamford DH, Grimes JM, Stuart DI. 2012. Structure unifies the viral universe. Annu Rev Biochem. 81(1):795–822.

Andreeva A, et al. 2007. Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res. 36(Database):D419–D425.

Brown CT, et al. 2015. Unusual biology across a group comprising more than 15% of domain Bacteria. Nature 523(7559):208–211.

Bukhari SA, Caetano-Anollés G. 2013. Origin and evolution of protein fold designs inferred from phylogenomic analysis of CATH domain structures in proteomes. PLoS Comput Biol. 9(3):e1003009.

Caetano-Anollés D, Nasir A, Kim KM, Caetano-Anollés G. 2019. Testing empirical support for evolutionary models that root the tree of life. J Mol Evol. 87(2–3):131–142.

Caetano-Anollés G, et al. 2014. Archaea: the first domain of diversified life. Archaea 2014:1–26.

Caetano-Anollés G, et al. 2017. The compressed vocabulary of the proteins of archaea. In: Witzany G, editor. Biocommunication of archaea. Cham (Switzerland): Springer. p. 147–174.

Caetano-Anollés G, Nasir A. 2012. Benefits of using molecular structure and abundance in phylogenomic analysis. Front Genet. 3:172.

Caetano-Anollés G, Nasir A, Kim KM, Caetano-Anollés D. 2018. Rooting phylogenies and the tree of life while minimizing ad hoc and auxiliary assumptions. Evol Bioinform Online. 14:117693431880510.

Castelle CJ, Banfield JF. 2018. Major new microbial groups expand diversity and alter our understanding of the tree of life. Cell 172(6):1181–1197.

Castelle CJ, et al. 2018. Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. Nat Rev Microbiol. 16(10):629–645.

Chothia C, Gough J, Vogel C, Teichmann SA. 2003. Evolution of the protein repertoire. Science 300(5626):1701–1703.

Claverie J-M, Abergel C. 2016. Giant viruses: the difficult breaking of multiple epistemological barriers. Stud Hist Philos Biol Biomed Sci. 59:89–99.

Colson P, et al. 2018. Ancestrality and mosaicism of giant viruses supporting the definition of the fourth TRUC of microbes. Front Microbiol. 9:2668.

Da Cunha V, Gaia M, Gadelle D, Nasir A, Forterre P. 2017. Lokiarchaea are close relatives of Euryarchaeota, not bridging the gap between prokaryotes and eukaryotes. PLoS Genet. 13(6):e1006810.

Da Cunha V, Gaia M, Nasir A, Forterre P. 2018. Asgard archaea do not close the debate about the universal tree of life topology. PLoS Genet. 14(3):e1007215.

Fahmy T, Aubry P. 2003. XLSTAT-Pro (version 7.0). Society Addinsoft.

Forterre P. 2011. Manipulation of cellular syntheses and the nature of viruses: the virocell concept. C R Chim. 14(4):392–399.

Forterre P. 2013. The common ancestor of archaea and eukarya was not an archaeon. Archaea 2013:1–18.

Forterre P. 2016. To be or not to be alive: how recent discoveries challenge the traditional definitions of viruses and life. Stud Hist Philos Biol Biomed Sci. 59:100–108.

Fox NK, Brenner SE, Chandonia JM. 2014. SCOPe: Structural Classification of Proteins–extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res. 42(D1):D304–D309.

Gaia M, Da Cunha V, Forterre P. 2018. The tree of life. Cham: Springer. p. 55–99. doi: 10.1007/978-3-319-69078-0_3.

Goloboff PA, Torres A, Arias JS. 2018. Weighted parsimony outperforms other methods of phylogenetic inference under models appropriate for morphology. Cladistics 34(4):407–437.

Gough J, Chothia C. 2002. SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. Nucleic Acids Res. 30(1):268–272.

Gough J, Karplus K, Hughey R, Chothia C. 2001. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J Mol Biol. 313(4):903–919.

Harish A, Abroi A, Gough J, Kurland C. 2016. Did viruses evolve as a distinct supergroup from common ancestors of cells? Genome Biol Evol. 8(8):2474–2481.

Harish A, Kurland CG. 2017. Empirical genome evolution models root the tree of life. Biochimie 138:137–155.

He X, et al. 2015. Cultivation of a human-associated TM7 phylotype reveals a reduced genome and epibiotic parasitic lifestyle. Proc Natl Acad Sci U S A. 112(1):244–249.

Heath TA, Hedtke SM, Hillis DM. 2008. Taxon sampling and the accuracy of phylogenetic analyses. J Syst Evol. 46:239–257.

Holmes EC, Duchêne S. 2019. Can sequence phylogenies safely infer the origin of the global virome? mBio. 10(2):e00289–19.

Hug LA, et al. 2016. A new view of the tree of life. Nat Microbiol. 1:16048.

Huson DH, et al. 2007. Dendroscope: an interactive viewer for large phylogenetic trees. BMC Bioinformatics 8(1):460.

Illergård K, Ardell DH, Elofsson A. 2009. Structure is three to ten times more conserved than sequence–a study of structural response in protein cores. Proteins 77(3):499–508.

Imachi H, et al. 2020. Isolation of an archaeon at the prokaryote-eukaryote interface. Nature 577: 519–525.

Iyer LM, Aravind L, Koonin EV. 2001. Common origin of four diverse families of large eukaryotic DNA viruses. J Virol. 75(23):11720–11734.

Iyer LM, Balaji S, Koonin EV, Aravind L. 2006. Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. Virus Res. 117(1):156–184.

Jeong H, Arif B, Caetano-Anollés G, Kim KM, Nasir A. 2019. Horizontal gene transfer in human-associated microorganisms inferred by phylogenetic reconstruction and reconciliation. Sci Rep. 9(1):5953.

Jeong H, et al. 2016. HGTree: database of horizontally transferred genes determined by tree reconciliation. Nucleic Acids Res. 44(D1):D610–D619.

Jeong H, Nasir A. 2017. A preliminary list of horizontally transferred genes in prokaryotes determined by tree reconstruction and reconciliation. Front Genet. 8:112.

Kim K, Caetano-Anollés G. 2011. The proteomic complexity and rise of the primordial ancestor of diversified life. BMC Evol Biol. 11:140.

Kim KM, Caetano-Anollés G. 2012. The evolutionary history of protein fold families and proteomes confirms that the archaeal ancestor is more ancient than the ancestors of other superkingdoms. BMC Evol Biol. 12(1):13.

Kim KM, Nasir A, Caetano-Anollés G. 2014. The importance of using realistic evolutionary models for retrodicting proteomes. Biochimie 99:129–137.

Kim KM, Nasir A, Hwang K, Caetano-Anollés G. 2014. A tree of cellular life inferred from a genomic census of molecular functions. J Mol Evol. 79(5–6):240–262.

La Scola B. 2003. A giant virus in amoebae. Science 299(5615):2033–2033.

Lake JA, Henderson E, Oakes M, Clark MW. 1984. Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes. Proc Natl Acad Sci U S A. 81(12):3786–3790.

Lecompte O, Ripp R, Thierry JJ-C, Moras D, Poch O. 2002. Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. Nucleic Acids Res. 30(24):5382–5390.

Legendre M, et al. 2018. Diversity and evolution of the emerging *Pandoraviridae* family. Nat Commun. 9(1):2285.

Legendre M, et al. 2019. Pandoravirus celtis illustrates the microevolution processes at work in the giant *Pandoraviridae* genomes. Front Microbiol. 10:430.
Librado P, Vieira FG, Rozas J. 2012. BadiRate: estimating family turnover rates by likelihood-based methods. Bioinformatics 28(2):279–281.
Lundberg JG. 1972. Wagner networks and ancestors. Syst Zool. 21(4):398–413.
Malik SS, Azem-e-Zahra S, Kim KM, Caetano-Anollés G, Nasir A. 2017. Do viruses exchange genes across superkingdoms of life? Front Microbiol. 8:2110.
McCutcheon JP, Moran NA. 2012. Extreme genome reduction in symbiotic bacteria. Nat Rev Microbiol. 10(1):13–26.
Méheust R, Burstein D, Castelle CJ, Banfield JF. 2019. The distinction of CPR bacteria from other bacteria based on protein family content. Nat Commun. 10(1):4173.
Moore AD, Bornberg-Bauer E. 2012. The dynamics and evolutionary potential of domain loss and emergence. Mol Biol Evol. 29(2):787–796.
Moreira D, Brochier-Armanet C. 2008. Giant viruses, giant chimeras: the multiple evolutionary histories of *Mimivirus* genes. BMC Evol Biol. 8(1):12.
Moreira D, Lopez-Garcia P. 2009. Ten reasons to exclude viruses from the tree of life. Nat Rev Microbiol. 7:306–311.
Nasir A, Caetano-Anollés G. 2013. Comparative analysis of proteomes and functionomes provides insights into origins of cellular diversification. Archaea 2013:1–13.
Nasir A, Caetano-Anollés G. 2015. A phylogenomic data-driven exploration of viral origins and evolution. Sci Adv. 1(8):e1500527.
Nasir A, Caetano-Anollés G. 2017. Identification of capsid/coat related protein folds and their utility for virus classification. Front Microbiol. 8:380.
Nasir A, Kim KM, Caetano-Anolles G. 2012. Giant viruses coexisted with the cellular ancestors and represent a distinct supergroup along with superkingdoms Archaea, Bacteria and Eukarya. BMC Evol Biol. 12(1):156.
Nasir A, Kim KM, Caetano-Anolles G. 2015. Lokiarchaeota: eukaryote-like missing links from microbial dark matter? Trends Microbiol. 23(8):448–450.
Nasir A, Kim KM, Caetano-Anollés G. 2012. Viral evolution: primordial cellular origins and late adaptation to parasitism. Mob Genet Elements. 2(5):247–252.
Nasir A, Kim KM, Caetano-Anollés G. 2014a. A phylogenomic census of molecular functions identifies modern thermophilic archaea as the most ancient form of cellular life. Archaea 2014:1–15.
Nasir A, Kim KM, Caetano-Anollés G. 2014b. Global patterns of protein domain gain and loss in superkingdoms. PLoS Comput Biol. 10(1):e1003452.
Nasir A, Kim KM, Caetano-Anollés G. 2017. Phylogenetic tracings of proteome size support the gradual accretion of protein structural domains and the early origin of viruses from primordial cells. Front Microbiol. 8:1178.
Nasir A, Kim K, Da Cunha V, Caetano-Anollés G. 2016. Arguments reinforcing the three-domain view of diversified cellular life. Archaea 2016:1–11.
Nasir A, Naeem A, Khan MJ, Lopez-Nicora HD, Caetano-Anollés G. 2011. Annotation of protein domains reveals remarkable conservation in the functional make up of proteomes across superkingdoms. Genes (Basel) 2(4):869–911.
Orsi WD, Richards TA, Francis WR. 2018. Predicted microbial secretomes and their target substrates in marine sediment. Nat Microbiol. 3(1):32–37.
Parks DH, et al. 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol. 36(10):996–1004.
Penny D, Zhong B. 2015. Two fundamental questions about protein evolution. Biochimie 119:278–283.
Philippe N, et al. 2013. Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. Science 341(6143):281–286.
Rinke C, et al. 2013. Insights into the phylogeny and coding potential of microbial dark matter. Nature 499(7459):431–437.
Rose PW, et al. 2015. The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. Nucleic Acids Res. 43(D1):D345–D356.
Sober E, Steel M. 2002. Testing the hypothesis of common ancestry. J Theor Biol. 218(4):395–408.
Soucy SM, Huang J, Gogarten JP. 2015. Horizontal gene transfer: building the web of life. Nat Rev Genet. 16(8):472–482.
Spang A, et al. 2015. Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature 521(7551):173–179.
Spang A, et al. 2018. Asgard archaea are the closest prokaryotic relatives of eukaryotes. PLoS Genet. 14(3):e1007080.
Spang A, Ettema T. 2016. Microbial diversity: the tree of life comes of age. Nat Microbiol. 1:16056.
Staley JT, Caetano-Anollés G. 2018. Archaea-first and the co-evolutionary diversification of domains of life. BioEssays 40(8):1800036.
Starr EP, et al. 2018. Stable isotope informed genome-resolved metagenomics reveals that *Saccharibacteria* utilize microbially-processed plant-derived carbon. Microbiome 6(1):122.
Sukumaran J, Holder MT. 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics 26(12):1569–1571.
Swofford DL. 2002. Phylogenomic analysis using parsimony and other programs (PAUP*). Version 4.0b10. Sunderland (MA): Sinauer.
Vogel C, Berzuini C, Bashton M, Gough J, Teichmann SA. 2004. Supradomains: evolutionary units larger than single protein domains. J Mol Biol. 336(3):809–823.
Vogel C, Chothia C. 2006. Protein family expansions and biological complexity. PLoS Comp Biol. 2(5):e48.
Vogel C, Teichmann SA, Pereira-Leal J. 2005. The relationship between domain duplication and recombination. J Mol Biol. 346(1):355–365.
Wang M, Caetano-Anollés G. 2006. Global phylogeny determined by the combination of protein domains in proteomes. Mol Biol Evol. 23(12):2444–2454.
Wang M, Jiang Y-Y, et al. 2011. A universal molecular clock of protein folds and its power in tracing the early history of aerobic metabolism and planet oxygenation. Mol Biol Evol. 28(1):567–582.
Wang M, Kurland CG, Caetano-Anollés G. 2011. Reductive evolution of proteomes and protein structures. Proc Natl Acad Sci U S A. 108(29):11954–11958.
Wang M, Yafremava LS, Caetano-Anolles D, Mittenthal JE, Caetano-Anolles G. 2007. Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world. Genome Res. 17(11):1572–1585.
Williams TA, Cox CJ, Foster PG, Szöllősi GJ, Embley TM. 2020. Phylogenomics provides robust support for a two-domains tree of life. Nat Ecol Evol. 4(1):138–147.
Williams TA, Foster PG, Cox CJ, Embley TM. 2013. An archaeal origin of eukaryotes supports only two primary domains of life. Nature 504(7479):231–236.
Wilson D, et al. 2009. SUPERFAMILY–sophisticated comparative genomics, data mining, visualization and phylogeny. Nucleic Acids Res. 37(Suppl 1):D380–D386.
Woese CR, Fox GE. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A. 74(11):5088–5090.
Woese CR, Kandler O, Wheelis ML. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci U S A. 87(12):4576–4579.
Wolf YI, Koonin EV. 2013. Genome reduction as the dominant mode of evolution. BioEssays 35(9):829–837.

Yafremava LS, et al. 2013. A general framework of persistence strategies for biological systems helps explain domains of life. Front Genet. 4:16.
Zaremba-Niedzwiedzka K, et al. 2017. Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature 541(7637):353–358.
Zwickl DJ, Hillis DM. 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst Biol. 51(4):588–598.

Associate editor: Eric Bapteste