Did Viruses Evolve As a Distinct Supergroup from Common Ancestors of Cells?

Ajith Harish[1,*], Aare Abroi[2], Julian Gough[3], and Charles Kurland[4]

[1]Structural and Molecular Biology Group, Department of Cell and Molecular Biology, Biomedical Center, Uppsala University, Sweden

[2]Estonian Biocentre, Riia 23, Tartu 51010, Estonia

[3]Computational Genomics Group, Department of Computer Science, University of Bristol, The Merchant Venturers Building, UK

[4]Microbial Ecology, Department of Biology, Lund University, Sweden

*Corresponding author: E-mail: ajith.harish@gmail.com, ajith.harish@icm.uu.se.

Accepted: July 21, 2016

Abstract

The evolutionary origins of viruses according to marker gene phylogenies, as well as their relationships to the ancestors of host cells remains unclear. In a recent article Nasir and Caetano-Anollés reported that their genome-scale phylogenetic analyses based on genomic composition of protein structural-domains identify an ancient origin of the "viral supergroup" (Nasir et al. 2015. A phylogenomic data-driven exploration of viral origins and evolution. Sci Adv. 1(8):e1500527.). It suggests that viruses and host cells evolved independently from a universal common ancestor. Examination of their data and phylogenetic methods indicates that systematic errors likely affected the results. Reanalysis of the data with additional tests shows that small-genome attraction artifacts distort their phylogenomic analyses, particularly the location of the root of the phylogenetic tree of life that is central to their conclusions. These new results indicate that their suggestion of a distinct ancestry of the viral supergroup is not well supported by the evidence.

Key words: tree of life, origins of viruses, systematic error, rooting artifact, small genome attraction, homoplasy.

Introduction

The debate on the ancestry of viruses is still undecided: In particular, it is still unclear whether viruses evolved before their host cells or if they evolved more recently from the host cells. The virus-early hypothesis posits that viruses predate or coevolved with their cellular hosts (Wessner 2010). Two alternatives describe the virus-late scenario: (i) progressive evolution also known as the escape hypothesis and (ii) regressive evolution or reduction hypothesis. Both propose that viruses evolved from their host cells (Wessner 2010). According to the first of these two virus-late models, viruses evolved from their host cells through gradual acquisition of genetic structures. The other alternative suggests that viruses, like host-dependent endoparasitic bacteria, evolved from free-living ancestors by reductive evolution. The recent discovery of the so-called giant viruses with double-stranded DNA genomes that parallel endoparasitic bacteria with regards to genome size, number of genes, and particle size revived the reductive evolution hypothesis. However, there are so far no identifiable "universal" viral genes that are common to viruses such as the ubiquitous cellular genes. In other words, examples of common viral components that are analogous to the ribosomal RNA and ribosomal protein genes, which are common to cellular genomes, are not found. This is one compelling reason that phylogenetic tests of the "common viral ancestor" hypotheses seem so far inconclusive.

Recently, Nasir and Caetano-Anollés (2015) employed phylogenetic analysis of whole-genomes and gene contents of thousands of viruses and cellular organisms to test the alternative hypotheses. The authors conclude that viruses are an ancient lineage that diverged independently and in parallel with their cellular hosts from a universal common ancestor (UCA). They reiterate their earlier claim (Nasir et al. 2012) that viruses are a unique lineage, which predated or coevolved with the last UCA of cellular lineages (LUCA) through reductive evolution rather than through more recent multiple origins. Their claims are based on analyses of statistical- and phyletic-distribution patterns of protein domains, classified

as superfamilies (SFs) in Structural Classification of Proteins (SCOP) (Murzin et al. 1995).

A re-examination of Nasir and Caetano-Anollés' phylogenomic approach (Nasir et al. 2012; Nasir and Caetano-Anollés 2015) suggests that small genomes systematically distort their phylogenetic reconstructions of the tree of life (ToL), especially the rooting of trees. Here the ToL is described as the evolutionary history of contemporary genomes: as a tree of genomes or a tree of proteomes (ToP) (Snel et al. 1999; Yang et al. 2005; Fang et al. 2013; Kurland and Harish 2015). The bias due to highly reduced genomes of parasites and endosymbionts in genome-scale phylogenies has been known for over a decade (Snel et al. 1999; Yang et al. 2005). In fact, prior to the recent proposal (Nasir and Caetano-Anollés 2015) these authors recognized the anomalous effects of including small genomes in reconstructing the ToL in analyses that were limited to cellular organisms (Kim and Caetano-Anollés 2011) or which included giant viruses (Nasir et al. 2012): As they say "In order to improve ToP reconstructions, we manually studied the lifestyles of cellular organisms in the total dataset and excluded organisms exhibiting parasitic (P) and obligate parasitic (OP) lifestyles, as their inclusion is known to affect the topology of the phylogenetic tree" (Nasir et al. 2012). But, they may not have adequately addressed this problem, particularly when the samplings include viral genomes that are likely to further exacerbate bias due to small genomes (Nasir et al. 2012; Nasir and Caetano-Anollés 2015). For this reason we systematically tested the reliability of the phylogenetic trees, especially the rooting approach favored by Nasir and Caetano-Anollés (2015). This approach depends critically on a hypothesized ancestor to root the ToP (ToL), but that ancestor is not identified empirically. Rather, it is assumed a priori to be an empty set (Nasir and Caetano-Anollés 2015).

We show here in several independent phylogenetic reconstructions that a rooting based on a hypothetical "all-zero" ancestor—an ancestor that is assumed to be an empty set of protein domains—creates specific phylogenetic artifacts: In this particular approach (Nasir et al. 2012; Nasir and Caetano-Anollés 2015) implementing the all-zero ancestor artifactually leads the root to be located amongst the taxa with the smallest genomes (proteomes) very much like the classical false rooting due to long-branch attractions (LBA) in gene trees (Forterre and Philippe 1999; Gouy et al. 2015).

Results and Discussion

We emphasize at the outset of this study that virtually all the evolutionary interpretations based on phylogenetic reconstructions depend on a reliable identification of the root of a tree (Graham et al. 2002; Morrison 2006). In particular, the rooting of a tree will determine the branching order of species, and define the ancestor–descendant relationships between taxa as well as the derived features of characters (character states). In effect, the root polarizes the order of evolutionary changes along the tree with respect to time. By the same token, the rooting of a tree distinguishes ancestral states from derived states among the different observed states of a character. However, determining which of the observed states of a character is ancestral and which one(s) are derived presents a chicken-and-egg problem. If ancestral states are identified directly, for example from fossils, characters can be polarized a priori with regard to determining the tree topology. A priori polarization supports intrinsic rooting. However, direct identification of ancestral states, particularly for extant genetic data is generally not possible except in rare cases where fossilized ancient DNA sequences are available. Ancient DNA recovered from Neanderthal humans, wooly mammoths, and giant virus (from permafrost) are well-known examples dating back to about 30–40,000 years (Poinar et al. 2006; Green et al. 2010; Legendre et al. 2014).

For the overwhelming majority of genes and genomes that are analyzed routinely, there are no known fossil references and divergence time spans hundreds of millions of years. Repeated substitutions at the same sites weaken the strength of the phylogenetic signal (Gouy et al. 2015). In such cases, sophisticated probabilistic models of nucleotide and amino acid substitution based on empirically derived non-reversible substitution patterns are highly effective in rooting gene-trees directly (Huelsenbeck et al. 2002; Yap and Speed 2005). Despite the obvious advantages of nonreversible models, they are much underutilized due to the computational sophistication required to implement such models. For instance, non-reversible models of DNA substitution have been available since the mid-1990s but rarely employed (Yang and Roberts 1995; Yap and Speed 2005). Consequently conventional phylogenetic methods predominantly use time-reversible (undirected) models of character evolution and they only compute unrooted trees. For example, the roots of the iconic ribosomal RNA ToL as well as most concatenated gene-trees are not determined directly (Woese et al. 1990; Pace 1997; Ciccarelli et al. 2006; Williams et al. 2012).

Accordingly, conventional approaches to character polarization are indirect and the rooting of trees is normally a two-stage process. The most common rooting method is the outgroup comparison method, which is based on the premise that character-states common to the ingroup (study group) and a closely related sister-group (the outgroup) are likely to be ancestral to both. Therefore, in an unrooted tree the root is expected to be positioned on the branch that connects the outgroup to the ingroup. In this way, the tree (and characters) may be polarized a posteriori (Morrison 2006; Wheeler 2012). However, there are no known outgroups for the ToL. In the absence of natural outgroups, pseudo-outgroups are used to root the ToL (Wheeler 2012). The best-known case is the root grafted onto the unrooted ribosomal RNA ToL based on presumed ancient (pre-LUCA) gene duplications (Schwartz and Dayhoff 1978; Woese et al. 1990). Here the paralogous proteins act as reciprocal outgroups that root each other. Unlike

gene duplications used to root gene-trees, the challenge of identifying suitable outgroups becomes more acute for genome-trees.

To meet this challenge Nasir and Caetano-Anollés use a hypothetical ancestor (pseudo-outgroup): an artificial taxon constructed from presumed ancestral states. For each character (SF), the state "0" or "absence" of a SF is assumed to be "ancestral" a priori. This artificial "all-zero" taxon is used to independently locate the root of an unrooted ToL. Further, they use the Lundberg rooting method, in which hypothetical ancestors (or outgroups) are not included in the initial tree reconstruction. The Lundberg method involves estimating an unrooted tree for the ingroup taxa only, and then attaching a hypothesized ancestor or outgroup(s) (when available) to the tree a posteriori to determine the position of the root (Swofford and Begle 1993). Unrooted trees describe relatedness of taxa based on graded compositional similarities of characters (and states). Accordingly, we can expect the "all-zero" ancestor to cluster among genomes (proteomes) in which the smallest number of SFs is present. The latter are the proteomes described by the largest number of "0s" in the data matrix.

The instability of rooting with an all-zero ancestor becomes clear when the smallest proteome in a given taxon sampling varies in the rooting experiments (figs. 1 and 2). Rooting experiments were preformed both for SF occurrence (presence/absence) patterns and for SF abundance (copy number) patterns. However, we present results for the SF abundance patterns, as in Nasir and Caetano-Anollés (2015). Throughout, we refer to genomic protein repertoires as proteomes. Proteome size related as the number of distinct SFs in a proteome (SF occurrence) is depicted next to each taxon for easy comparison in figures 1 and 2. Phylogenetic analyses were carried out as described in (Nasir et al. 2012; Nasir and Caetano-Anollés 2015) (see Material and Methods). SFs that are shared between proteomes of viruses and cellular organisms were used as the characters (fig. 1*a*) as in (Nasir and Caetano-Anollés 2015). Initially no viruses are included in tree reconstructions and here the root was placed within the Archaea, which has the smallest proteome (*Pyrococcus horikoshii*; 503 SFs) among the supergroups (fig. 1*b*). When a still smaller bacterial proteome (*Rickettsia prowazekii*; 420 SFs) was included, the position of the root as well as the branching order changed. In this case, the bacteria were split into two groups and the root was placed within one of the bacterial groups (fig. 1*c*). Further, when a much smaller archaeal proteome was included (*Nanoarchaeum equitans*; 228 SFs), the root was relocated to a branch leading to the now smallest proteome (fig. 1*d*). Note that the newly included taxa, both bacteria and archaea are host-dependent symbionts with reduced genomes. Supplementary table S1, Supplementary Material online contains a list of all taxa in figure 1 their NCBI taxonomy IDs, species names, and proteome sizes (as SF occurrence).

Similarly including viruses in the analyses draws the root towards the smaller viral proteomes (fig. 2). As in the rooting experiments in figure 1, a group of DNA viruses (107–175 SFs) was introduced (supplementary table S1, Supplementary Material online). These DNA viruses have larger proteomes than do the RNA viruses, but they are much smaller than most known endosymbiotic bacteria (fig. 2*d*). Again, the root was repositioned within the DNA viruses group (fig. 2*a*). Following this experiment, two extremely reduced endosymbiotic bacteria (*Ca. Nasuia deltocephalinicola* and *Ca. Tremblaya princeps*; 107 SFs each) classified as betaproteobacteria were included. These further displaced the root closer to the smallest set of proteomes (fig. 2*b*). Finally, a set of genomes from four RNA viruses (4–17 SFs) as in supplementary table S1, Supplementary Material online, was introduced into the genome samplings and they rooted the tree within the RNA viruses (fig. 2*c*). These results challenge the conclusion drawn previously that the proteomes of RNA viruses are more ancient than proteomes of DNA viruses (Nasir and Caetano-Anollés 2015). In addition, the results contradict the purported antiquity of viral proteomes as such. Rather, the data suggest that there are severe artifacts generated by genome size-bias due to the inclusion of the viral proteomes in the analysis. These artifacts are expressed as grossly erroneous rootings caused by small-genome attraction (SGA) in the Lundberg rooting using the hypothetical all-zero ancestor.

Including the all-zero ancestor in the analysis either implicitly (defined by the ANCSTATES option in PAUP*) or explicitly (as a taxon in the data matrix) does not make a difference to the tree topology and rooting. We note that including the hypothetical ancestor during tree estimation amounts to a priori character polarization and prespecification of the root. In addition, the position of the root was the same in the different rooting experiments when the all-zero ancestor was explicitly specified as the outgroup to root trees using the outgroup method (see supplementary figs. S1 and S2, Supplementary Material online). These rooting experiments reveal a strong bias in rooting that favors small genomes irrespective of the different rooting methods—outgroup rooting, Lundberg rooting, and intrinsic rooting—used to root the ToL with the "all-zero" hypothetical ancestor. This SGA artifact is comparable to the better-known LBA artifact that is associated with compositional bias of nucleotides or of amino acids that distort gene-trees (Gouy et al. 2015).

The use of artificial outgroups is not uncommon in rooting experiments when rooting is ambiguous (Graham et al. 2002). Artificial taxa are either an all-zero outgroup or an outgroup constructed by randomizing characters and/or character-states of real taxa. Although rooting experiments with multiple real outgroups, or randomized artificial outgroups that simulate loss of phylogenetic signal can minimize the ambiguity in rooting, the all-zero outgroup has proved to be of little use (Graham et al. 2002; Wheeler 2012). Conclusions based

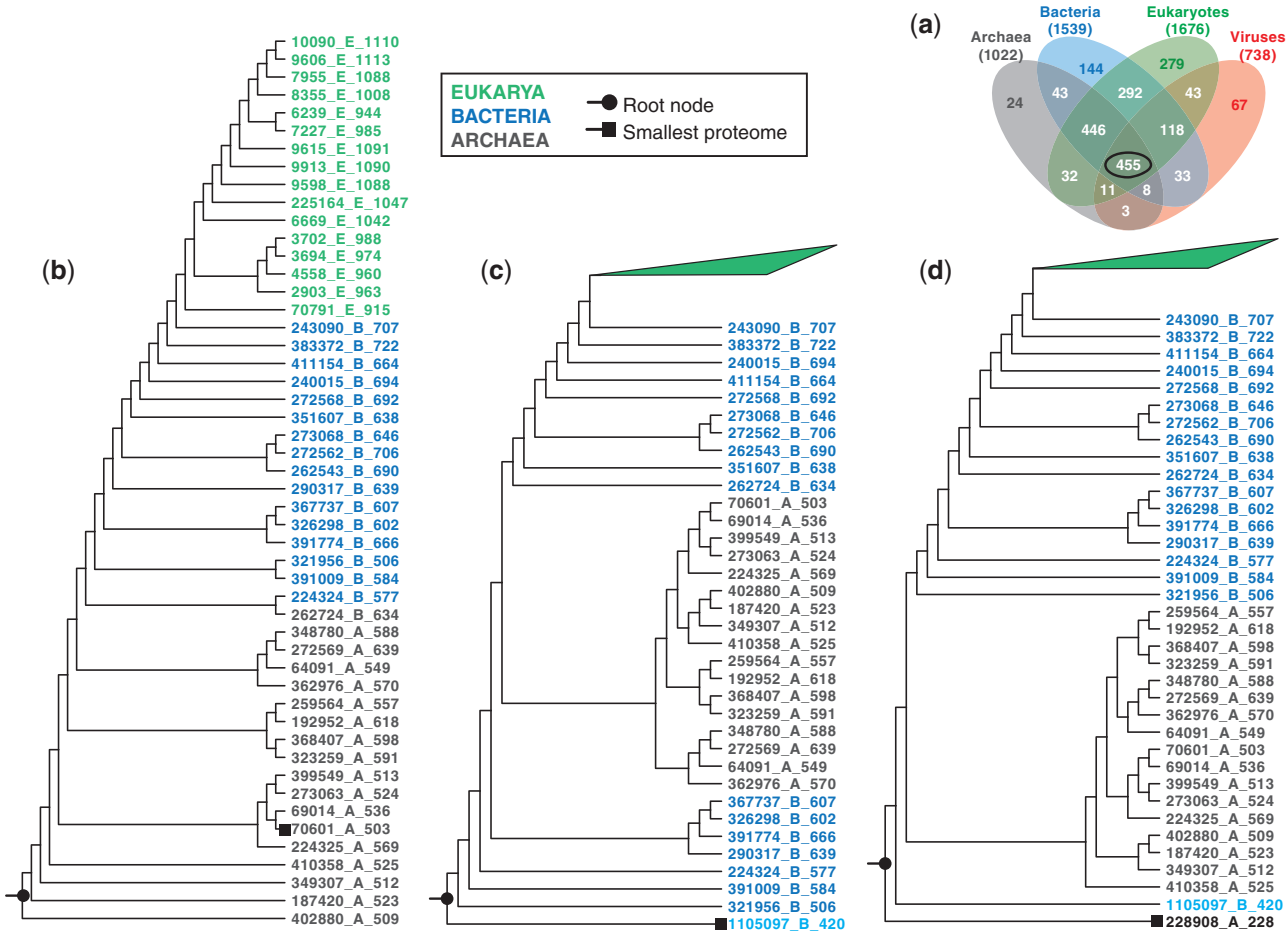

Fig. 1.—Implementing an "all-zero" pseudo-ancestor [2] severely distorts rooting of the ToL. Rooted trees using the Lundberg rooting method were reconstructed from a subset of 368 taxa (proteomes) sampled in Nasir and Caetano-Anollés (2015), which included 17 taxa each from Archaea, Bacteria, Eukarya, and 9 taxa from the virus groups (V). (*a*) Venn diagram shows the 455 SFs shared between cells and viruses (Archaea, Bacteria, Eukarya and Viruses; ABEV), which were used to reconstruct trees. (*b*) Single most parsimonious tree of ABE taxa rooted within Archaea. (*c, d*) New taxa, which represent the smallest proteome after inclusion, were progressively included in size order. The position of the root node changed accordingly to the branch corresponding to a group (or taxon) with the smallest proteome, which is Bacteria (*c*), Archaea (*d*); the Eukarya section is collapsed since tree topology is unaffected. Taxa are described by their NCBI taxonomy ID, taxonomic affiliation (A, B, E or V) and proteome size in terms of the number of distinct SFs present in the genome. To compare the position of the root node trees are drawn to show branching patterns only, branch lengths are not proportional to the quantity of evolutionary change.

on an all-zero outgroup are often refuted when empirically grounded analysis with real taxa are carried out (Wheeler 2012). Indeed, the present rooting experiments (figs. 1 and 2) clearly show that the position of the root depends on a group of small genomes or the smallest genome in the sample when a hypothetical all-zero ancestor or outgroup is used. In effect, the rooting approach favored by Nasir and Caetano-Anollés (2015) is not reliable.

Nevertheless, small proteome size is not an irreconcilable feature of genome-tree reconstructions (Snel et al. 1999; Yang et al. 2005; Harish et al. 2013). Small genome attraction artifacts may be observed when highly reduced proteomes of obligate endosymbionts are included in analyses with common samplings (Snel et al. 1999; Yang et al. 2005; Nasir et al. 2012; Harish et al. 2013). However, their untoward effects are only observed in the shallow end of the tree where the endosymbionts might be clustered with unrelated groups. In such cases the deep divergences or clustering of major branches are unperturbed (Yang et al. 2005; Harish et al. 2013). Such size-biases are readily corrected by normalizations that account for genome size (actually specific SF content in this case) (Snel et al. 1999; Yang et al. 2005; Harish et al. 2013).

Though size-bias correction for SF content phylogeny is known to be reliable, both for measures based on distance (Yang et al. 2005) and those based on frequencies of character distribution (Harish et al. 2013), Nasir and Caetano-Anollés do not apply such corrections. They only attend to the proteome size variations associated with SF abundances (Nasir et al. 2012; Nasir and Caetano-Anollés 2015). Instead

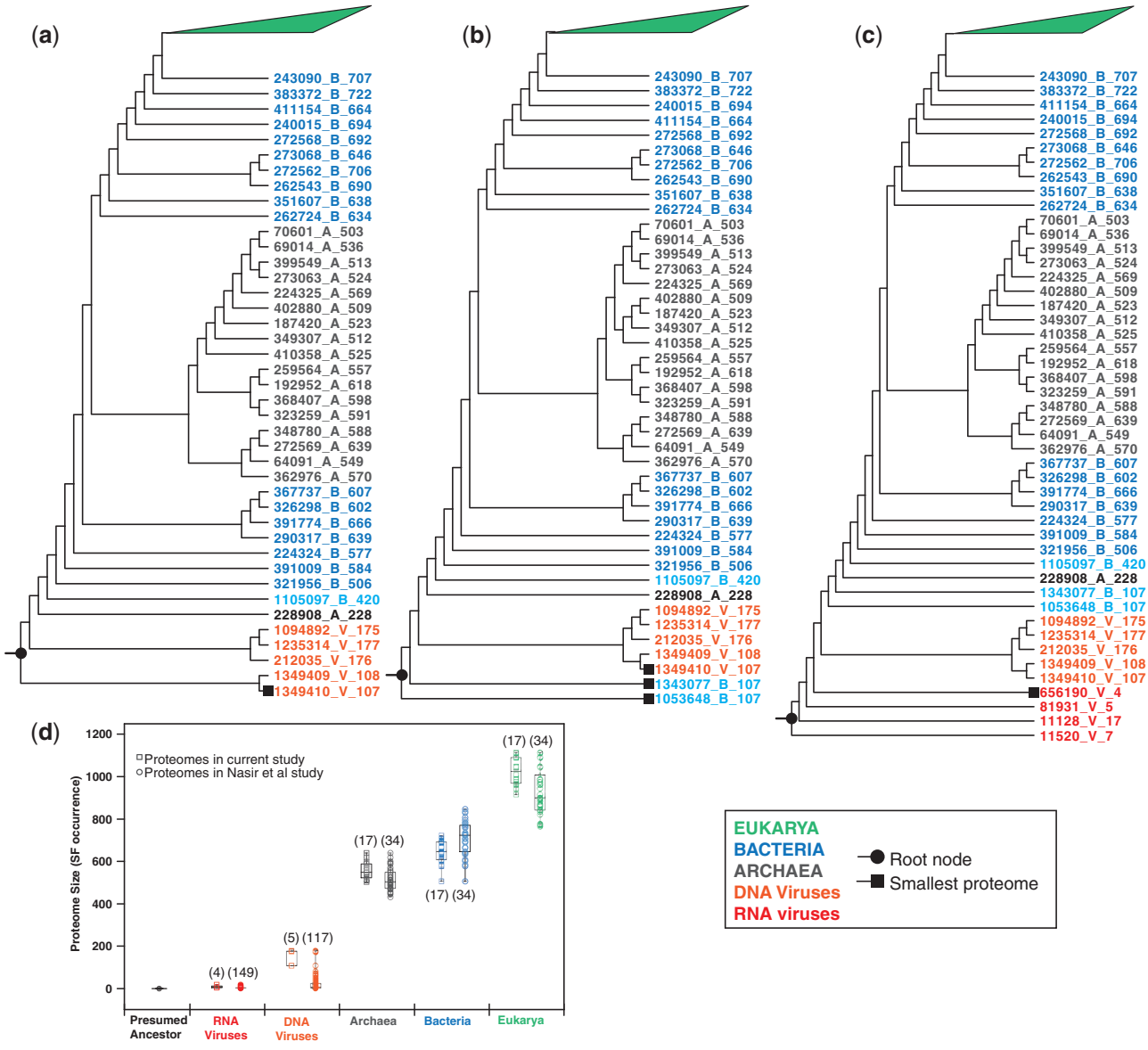

Fig. 2.—Rooting experiments (continued from fig. 1) show rooting bias of "all-zero" pseudo-ancestor towards small proteomes. (*a*–*c*) New taxa, which represent the smallest proteome after inclusion, were progressively included in size order, where the smallest proteome was from mega DNA viruses (*a*), Bacteria (*b*), and RNA viruses (*c*). Details in trees are same as in figure 1. (*d*) Comparison of proteome sizes of sampled taxa in terms of SF occurrence used to estimate trees in this study and in Nasir and Caetano-Anollés (2015). Numbers in parentheses above each group indicate the number of proteomes.

of accounting for novel taxon-specific SFs in their model of evolution, the authors choose to exclude potentially problematic small proteomes of parasitic bacteria and to include only the proteomes of "free-living" cellular organisms in their analyses (Nasir et al. 2012; Nasir and Caetano-Anollés 2015). But all viruses are parasites and obviously even more extreme examples of minimal proteomes.

This problem is further exacerbated by the uneven and largely incomplete annotation of SF domains in viral proteins (Abroi and Gough 2011; Abroi 2015). In fact, many viral "proteomes" that were sampled in Nasir and Caetano-Anollés (2015) are as small as a single SF. It is not clear why the inclusion of small viral proteomes was not recognized as even more problematic than the inclusion of small parasitic bacterial proteomes, in spite of the previous assertion of these authors that small proteomes should be excluded (Kim and Caetano-Anollés 2011). Nevertheless, including small viral proteomes is inconsistent with specifically excluding small cellular proteomes in the ToL, especially when hypotheses of reductive evolution are considered. Screening taxa based on "lifestyle" (free-living or parasitic) seems unwarranted since extreme reductive genome evolution, sometimes called genome

streamlining, is not limited to host-adapted parasitic bacteria but is common in free-living bacteria as well as eukaryotes (Andersson and Kurland 1998; Dujon et al. 2004; Giovannoni et al. 2014).

There is no theoretical reason to expect monotonic increase in complexity with time during evolution of lineages (Szathmáry and Smith 1995) nor is there empirical evidence to suggest that early evolution appeared to be a linear progression of simple to complex evolutionary forms (Rokas 2013). Yet, inferences about the root of the ToL often relate to a presumption of simple (primitive) to complex progression that aligns with the traditional principles of the *scala naturae* (Mayr 1982; Forterre and Philippe 1999; Gouy et al. 2015). Such common, but untested assumptions also motivate the rather simplistic models of proteome evolution in Nasir and Caetano-Anollés (2015), which is stated as "*ToP were rooted by the minimum character state* [i.e. "0" or absence of a SF], *assuming that modern proteomes evolved from a relatively simpler urancestral organism that harbored only few FSFs*"; see also Kim and Caetano-Anollés (2011) and Nasir et al. (2012).

In summary, the "all-zero" or "all-absent" hypothetical ancestor is neither empirically grounded nor biologically meaningful. The assumption that the absence of a SF ("0") is the ancestral state is a failure to distinguish between "ancestral absence" and "derived absence (loss)", and that failure creates a potential confusion of homology with analogy. Similar artifacts due to strong compositional biases in viral gene sequences are encountered when simple models of sequence evolution are used for phylogenetic inferences because such models fail to account for the substantial influence of nonphylogenetic signal (Moreira and López-García 2015). In fact, due to the very high rates of sequence evolution in viruses, sequence composition in informational genes used as phylogenetic markers to place viruses in the ToL seem to approach random sequence composition (Moreira and López-García 2015). Previous proposals for a fourth domain of life have identified viruses as a distinct group in a ToL determined in phylogenies of informational genes, but these were refuted by analyses using better models of sequence evolution that minimize phylogenetic noise (Williams et al. 2011; Moreira and López-García 2015). These careful studies show that over-simplified models frequently fail to distinguish homology from homoplasy, a lapse that may distort their resulting phylogenetic inferences (Philippe et al. 2011; Anisimova et al. 2013).

In addition, homoplasy in genome-scale phylogenies could be due to horizontal gene transfer (HGT). Rampant HGT between viruses and their hosts, often host-to-virus transfer, is well known (Yutin et al. 2014; Moreira and López-García 2015). Evidence for virus-to-virus transfers across distant viral groups such as RNA and DNA viruses, which were considered to be nonexistent, is growing (Stedman 2015). Despite this, Nasir and Caetano-Anollés (2015) conclude that a set of 68 SFs common to archaeal, bacterial, and eukaryote viruses as well as to their hosts corresponds to a conserved ancestral core of SFs, which are likely to be present in the common ancestor of cells and viruses "proto-virocells". They claim that the observed patterns of SF-sharing are unlikely to be due to host-to-virus HGT. They argue that multiple, independent instances of similar HGT in specific host–virus associations that are separated by large evolutionary distances are improbable. However, a closer look at the genomic distribution of the 68 "core-SFs" shows that they are widely distributed in their cellular hosts, but very sparsely distributed in the viruses sampled by Nasir and Caetano-Anollés (2015). These 68 SFs are present on average in 80% of cellular genomes but only in 3% of viral genomes according to the genomic SF frequency (*f* value) in Figures 3 and S3, Table S5 in Nasir and Caetano-Anollés (2015).

Furthermore, the distribution of the 68 core-SFs specifically in dsDNA viruses is higher on average (13% of genomes) compared with other viruses sampled by them (Nasir and Caetano-Anollés 2015). Indeed, 49 of 68 core-SFs are unique to dsDNA viruses and 32 of these are found in Mimivirus genes. The latter are known to be acquired by cell-to-virus HGT, either from the host amoeba or from bacteria that parasitize the host amoeba (Moreira and Brochier-Armanet 2008). It is important and relevant to note that all giant viruses isolated thus far are dsDNA viruses and that almost all of them are associated with cellular hosts belonging to a single genus: *Acanthamoeba*. Therefore, for these 68 SFs, the distribution pattern is consistent with their polyphyletic origins in viral genomes. However, the conclusion that these SFs were likely to be present in the LUCA of cellular lineages and subsequently lost in roughly 20% of the lineages is consistent with their widespread distribution in cellular genomes. It remains to be seen if giant viruses are associated with other cellular hosts or limited to very specific groups.

In addition to the ToP, the authors use a so-called tree of domains (ToD) to support their conclusion that proteomes of viruses are ancient and that proteomes of RNA viruses are particularly ancient. The ToD is projected as the evolutionary trajectory of individual SFs. Such projections are used as proxies to determine the relative antiquity or novelty of SFs (Nasir and Caetano-Anollés 2015). The ToD like the ToP is also rooted with a presumed ancestor (although using an opposite polarity compared with ToP). That rooting may be an artifact as for the ToP. Much more serious than potential artifacts in ToD is an egregiously bad assumption that is explicitly contradicted in the SCOP hierarchal classification: This is the notion that the ToD describes evolutionary relationships between SFs in the same way the ToL describes genealogy of species.

Evolutionary relationships between SFs with different folds in the SCOP classification have not been observed (Murzin et al. 1995; Gough et al. 2001). Physicochemical protein folding experiments and the corresponding statistical analyses of sequence evolution patterns, including simulations of protein

folding are all consistent with the observations that the sequence-structure space of SFs is discontinuous (Oliveberg and Wolynes 2005; Wolynes 2015). Empirical data indicate that the evolutionary transition from one SF to another through gradual changes implied in the ToD is unlikely, if at all feasible. This makes the ToD hypothesis, which assumes that all SFs are related to one another by common ancestry untenable. Thus the ToD contradicts the very basis upon which SFs are classified in the SCOP hierarchy (Murzin et al. 1995; Gough et al. 2001). The ToD is therefore uninterpretable as an evolutionary history of individual SFs. Accordingly, the ToD cannot reflect the "relative ages" of SFs nor can it support the inferred antiquity of viruses in the ToP.

Unlike phylogenetic trees that describe the evolution of individual proteomes, Venn diagrams, SF sharing patterns and summary statistics of SF frequencies among groups of proteomes only depict generalized trends. Multiple evolutionary scenarios can be invoked a priori to explain the general trends without any phylogenetic analyses (Abroi and Gough 2011; Abroi 2015). Although such patterns may be suggestive, for example the inference of a "conserved core of SFs in proto-virocells", these do not by themselves support reliable phylogenetic inferences. Thus, the inferences in (Nasir and Caetano-Anollés 2015) based on statistical distributions of SFs alone are speculative, at best.

In summary, the proposed rooting for the ToL (Nasir et al. 2012; Nasir and Caetano-Anollés 2015) is affected by clearly identifiable artifacts. Likewise, their supporting data and analyses seem to be biased by limited sampling and highly skewed SF distributions. Indeed, the data presented here undermine the inferred relative antiquity of viruses in the ToL. Although highly conserved, complex, genome-scale characters such as protein folds provide distinct advantages over fast evolving gene-sequence characters, the simplistic models of SF evolution implemented by Nasir et al. are demonstrably prone to phylogenetic artifacts. The SGA artifacts identified here are similar to the LBA artifacts identified previously in marker gene analyses using deficient models (Raoult et al. 2004; Claverie and Ogata 2009), which were later refuted by rigorous tests and analyses with appropriate models (Moreira and Brochier-Armanet 2008; Moreira and López-García 2015).

In effect, we suggest that Nasir et al.'s phylogenetic approach (Nasir et al. 2012; Nasir and Caetano-Anollés 2015) provides neither a test nor a confirmation of any one of the hypotheses for the origins of viruses (Wessner 2010). Despite its importance, reconciling the extensive genetic and morphological diversity of viruses as well as their evolutionary origins remains to be done (Wessner 2010; Forterre et al. 2014). Better methods and empirical models are required to test whether a multiplicity of scenarios or a single overarching hypothesis can account for the origins of viruses.

Material and Methods

Here, genomic protein repertoires are referred to as proteomes. We re-analyzed a subset of the 368 proteomes sampled in Nasir and Caetano-Anollés (2015) for phylogenetic rooting of diverse cells and their viruses. Here, we sampled all the 102 cellular proteomes containing 34 each from Archaea, Bacteria, and Eukaryotes, respectively as well as 16 viral proteomes from Nasir and Caetano-Anollés (2015). For the latter, we note that the DNA virus proteomes were substantially larger than those of RNA viruses in terms of the number of identified SFs. In addition we included for comparison some of the smallest known proteomes of Archaea and Bacteria not included in Nasir and Caetano-Anollés (2015). Roughly, half of the sampled proteomes were analyzed (figs. 1 and 2) for computational simplicity. Results did not vary when all the sampled taxa were included (see supplementary figs. S3 and S4, Supplementary Material online). Rooting experiments were preformed both with SF occurrence (presence/absence) patterns and SF abundance (copy number) patterns; however, we present results for the SF abundance patterns as in Nasir and Caetano-Anollés (2015). In addition to the Lundberg rooting procedures carried out as in Nasir et al. (2012) and Nasir and Caetano-Anollés (2015), rooting experiments were repeated by including the all-zero taxon in the tree reconstruction process implicitly (using the ANCSTATES option in PAUP*) and explicitly as a taxon in the data matrix. Further, when the all-zero taxon was explicitly included, rooting experiments were also repeated with the outgroup rooting method. Phylogenetic reconstructions were carried out using maximum parsimony criterion implemented in PAUP* ver. 4.0b10 (Swofford 2003) with heuristic tree searches using 1,000 replicates of random taxon addition and tree bisection reconnection branch swapping. Trees were rooted by Lundberg method, outgroup method or intrinsically rooted (by including the hypothetical all-zero ancestor in tree searches). NCBI taxonomy ID, species names, and proteome size (as SF occurrence) for the taxa analyzed here are listed in supplementary tables S1 and S2, Supplementary Material online.

Acknowledgments

We thank anonymous referees for their constructive suggestions. A.H. acknowledges support from The Swedish Research Council (to Måns Ehrenberg) and the Knut and Alice Wallenberg Foundation, RiboCORE (to Måns Ehrenberg and Dan Andersson), A.A. acknowledges support from Basic research financing to Estonian Biocentre and C.G.K. acknowledges support from the Nobel Committee for Chemistry of the Royal Swedish Academy of Sciences.

Supplementary Material

Supplementary tables S1, S2 and S5 and figures S1–S4 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Literature Cited

Abroi A. 2015. A protein domain-based view of the virosphere-host relationship. Biochimie. 119:231–243.

Abroi A, Gough J. 2011. Are viruses a source of new protein folds for organisms? - Virosphere structure space and evolution. BioEssays 33:626–635.

Andersson SGE, Kurland CG. 1998. Reductive evolution of resident genomes. Trends Microbiol. 6:263–268.

Anisimova M, et al. 2013. State-of the art methodologies dictate new standards for phylogenetic analysis. BMC Evol Biol. 13:161.

Ciccarelli FD, et al. 2006. Toward automatic reconstruction of a highly resolved tree of life. Science 311:1283–1287.

Claverie J-M, Ogata H. 2009. Ten good reasons not to exclude viruses from the evolutionary picture. Nat Rev Micro. 7:615.

Dujon B, et al. 2004. Genome evolution in yeasts. Nature 430:35–44.

Fang H, et al. 2013. A daily-updated tree of (sequenced) life as a reference for genome research. Sci Rep. 3: 2015.

Forterre P, Krupovic M, Prangishvili D. 2014. Cellular domains and viral lineages. Trends Microbiol. 22:554–558.

Forterre P, Philippe H. 1999. Where is the root of the universal tree of life? BioEssays 21:871–879.

Giovannoni SJ, Cameron Thrash J, Temperton B. 2014. Implications of streamlining theory for microbial ecology. ISME J. 8:1553–1565.

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. 313:903–919.

Gouy R, Baurain D, Philippe H. 2015. Rooting the tree of life: the phylogenetic jury is still out. Phil Trans R Soc B. 370:20140329.

Graham SW, Olmstead RG, Barrett SCH. 2002. Rooting phylogenetic trees with distant outgroups: a case study from the Commelinoid Monocots. Mol Biol Evol. 19:1769–1781.

Green RE, et al. 2010. A draft sequence of the neandertal genome. Science 328:710–722.

Harish A, Tunlid A, Kurland CG. 2013. Rooted phylogeny of the three superkingdoms. Biochimie. 95:1593–1604.

Huelsenbeck JP, Bollback JP, Levine AM. 2002. Inferring the root of a phylogenetic tree. Syst Biol. 51:32–43.

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

Kurland CG, Harish A. 2015. Structural biology and genome evolution: an introduction. Biochimie. 119:205–208. doi: http://dx.doi.org/10.1016/j.biochi.2015.10.023

Legendre M, et al. 2014. Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. Proc Natl Acad Sci. 111:4274–4279.

Mayr E. 1982. The growth of biological thought. Cambridge (MA): The Belknap Press of Harvard University Press.

Moreira D, Brochier-Armanet C. 2008. Giant viruses, giant chimeras: the multiple evolutionary histories of Mimivirus genes. BMC Evol Biol. 8:1–10.

Moreira D, López-García P. 2015. Evolution of viruses and cells: do we need a fourth domain of life to explain the origin of eukaryotes? Philos Trans R Soc Lond B Biol Sci. 370:

Morrison DA. 2006. Phylogenetic analyses of parasites in the new millennium. Adv Parasitol. 63:1–124.

Murzin AG, Brenner SE, Hubbard T, Chothia C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol. 247:536–540.

Nasir A, Caetano-Anollés G. 2015. A phylogenomic data-driven exploration of viral origins and evolution. Sci Adv. 1(8): e1500527.

Nasir A, Kim K, Caetano-Anolles G. 2012. Giant viruses coexisted with the cellular ancestors and represent a distinct supergroup along with superkingdoms Archaea, Bacteria and Eukarya. BMC Evol Biol. 12:156.

Oliveberg M, Wolynes PG. 2005. The experimental survey of protein-folding energy landscapes. Q Rev Biophys. 38:245–288.

Pace NR. 1997. A molecular view of microbial diversity and the biosphere. Science 276:734–740.

Philippe H, et al. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. 9(3): e100060.

Poinar HN, et al. 2006. Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. Science 311:392–394.

Raoult D, et al. 2004. The 1.2-Megabase genome sequence of mimivirus. Science 306:1344–1350.

Rokas A. 2013. My oldest sister is a sea walnut? Science 342:1327–1329.

Schwartz R, Dayhoff M. 1978. Origins of prokaryotes, eukaryotes, mitochondria, and chloroplasts. Science 199:395–403.

Snel B, Bork P, Huynen MA. 1999. Genome phylogeny based on gene content. Nat Genet. 21:108–110.

Stedman KM. 2015. Deep recombination: RNA and ssDNA virus genes in DNA virus and host genomes. Annu Rev Virol. 2:203–217.

Swofford DL. 2003. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sunderland (MA): Sinauer Associates.

Swofford DL, Begle PB. 1993. PAUP: phylogenetic analysis using parsimony: user's manual (Version 3.1). Champaign (IL): Illinois Natural History Survey.

Szathmáry E, Smith JM. 1995. The major evolutionary transitions. Nature 374:227–232.

Wessner D. 2010. The origins of viruses. Nat Educ. 3:37.

Wheeler WC. 2012. Systematics: a course of lectures. Hoboken (NJ): John Wiley & Sons, Ltd.

Williams TA, Embley TM, Heinz E. 2011. Informational gene phylogenies do not support a fourth domain of life for nucleocytoplasmic large DNA viruses. PLoS One 6:e21080.

Williams TA, Foster PG, Nye TMW, Cox CJ, Embley TM. 2012. A congruent phylogenomic signal places eukaryotes within the Archaea. Proc R Soc B Biol Sci. 279(1749): 4870–4879.

Woese CR, Kandler O, Wheelis ML. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci U S A. 87:4576–4579.

Wolynes PG. 2015. Evolution, energy landscapes and the paradoxes of protein folding. Biochimie. 119:218–230.

Yang S, Doolittle RF, Bourne PE. 2005. Phylogeny determined by protein domain content. Proc Natl Acad Sci U S A. 102:373–378.

Yang Z, Roberts D. 1995. On the use of nucleic acid sequences to infer early branchings in the tree of life. Mol Biol Evol. 12:451–458.

Yap VB, Speed T. 2005. Rooting a phylogenetic tree with nonreversible substitution models. BMC Evol Biol. 5:1–8.

Yutin N, Wolf YI, Koonin EV. 2014. Origin of giant viruses from smaller DNA viruses not from a fourth domain of cellular life. Virology 466–467:38–52.

Associate editor: Purificación López-García