bioRxiv preprint doi: https://doi.org/10.1101/050179; this version posted April 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

No 'small genome attraction' artifact: A response to Harish et al. 'Did viruses evolve as a distinct supergroup from common ancestors of cells?'

Arshan Nasir[1], Kyung Mo Kim[2] and Gustavo Caetano-Anollés[3*]

[1]Department of Biosciences, COMSATS Institute of Information Technology, Islamabad 45550, Pakistan
[2]Biological Resource Center, Korea Research Institute of Bioscience and Biotechnology, Jeongeup, Republic of Korea
[3]Evolutionary Bioinformatics Laboratory, Department of Crop Sciences, University of Illinois at Urbana-Champaign, Urbana, IL 61820, USA

*Corresponding author: gca@illinois.edu

In a recent eLetter and associated preprint, Harish, Abroi, Gough and Kurland criticized our structural phylogenomic methods, which support the early cellular origin of viruses. Their claims include the argument that the rooting of our trees is artifactual and distorted by small genome (proteome) size. Here we uncover their aprioristic reasoning, which mingles with misunderstandings and misinterpretations of cladistic methodology. To demonstrate, we labeled the phylogenetic positions of the smallest proteomes in our phylogenetic trees and confirm that the smallest genomes were neither attracted towards the root nor caused any distortions in the four-supergroup tree of life. Their results therefore stem from confusing outgroups with ancestors and handpicking problematic taxa to distort tree reconstruction. In doing so, they ignored the details of our rooting method, taxa sampling rationale, the plethora of evidence given in our study supporting the ancient origin of the viral supergroup and also recent literature on viral evolution. Indeed, our tree of life uncovered many viral monophyletic groups consistent with ICTV classifications and showed remarkable evolutionary tracings of virion morphotypes onto a revealing tree topology.

Harish et al. would like to see the origin of Eukarya at the base of the Tree of Life (ToL) [1]. So, in their commentary [2,3], they begin by questioning our phylogenomic analysis, which is supported by large-scale structural and functional data and well-established comparative genomics, phylogenomics, and multidimensional scaling approaches [4]. Their writings [2,3] fail to acknowledge recent literature, including our recent invalidation of both their rationale [5] and their own phylogenetic methodologies [6], which we showed counter modern evolutionary thinking. While our critique remains unanswered, here we recap how their argumentation and misguided experimentation is used to *"mud the waters"* (*sensu* [2]) of cladistics understanding. A quick fact-checking exercise is presented in Table 1 and described in the following sections.

1. Their claim that our rooting approach uses outgroup taxa is incorrect. We do not *"use a hypothetical pseudo-outgroup, ... an artificial all-zero taxon ... to root the ToL"* [2] nor *"an ancestor that is assumed to be an empty set of protein domains"* as outgroup

bioRxiv preprint doi: https://doi.org/10.1101/050179; this version posted April 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

to *"create specific phylogenetic artifacts"* [3]. No outgroup taxon (presumably extant or artificial) was ever used/defined in our study [4], demonstrating they are confusing outgroups with ancestors (Table 1). We are therefore puzzled by the indirectly rooted ToLs they build in their attempt to mimic and undermine our methods [3], as well as the putative sensitivity to addition of new taxa, especially because these taxa represent organisms that engage in extreme forms of cellular endosymbiosis (please read below for the rationale for why then viruses were included). Even their tree searches were conducted differently, failing to minimize Farris' *f*-values and being dependent on the location of the root. Contrary to their claims, our rooting method is grounded in early and well-established cladistic formalizations [7,8] and is *direct* because it polarizes character transformations with information solely present in the taxa being studied (the ingroup), distinguishing ancestral from derived character states and rooting the trees both empirically and *a posteriori*. In fact, and crucially, our rooted trees comply with Weston's generality criterion [9,10], which states that as long as ancestral characters are preponderantly retained in descendants, ancestral character states will always be more general than their derivatives given their nested hierarchical distribution in rooted phylogenies. Biologically, domain structures spread by recruitment in evolution when genes duplicate and diversify, genomes rearrange, and genetic information is exchanged. This is a process of accumulation and retention of iterative homologies, such as serial homologues and paralogous genes [10], which is global, universal and largely unaffected by proteome size. This same process is widely used to generate rooted phylogenies from paralogous gene sequences. The Lundberg method [8], which does not attach outgroup taxa to the ingroup as they claim, simply enables rooting by the generality criterion [11].

2. Their confusion of a priori and a posteriori character polarization questions their understanding of cladistic methodology. Their claim that we use the pseudo-outgroup to polarize character state changes *a priori* is inconsistent with our methodology of first reconstructing an undirected network and then polarizing character transformations with a direct method that complies with Weston's rule [4]. They miss the fact that rooting is not a neutral procedure. While the length of the most parsimonious trees is unaffected by the position of the root, making *a priori* polarization unnecessary [7], rooting impacts the homology statements of the undirected networks [7,8]. *"The length of a tree is unaffected by the position of the root but is certainly not unaffected by the inclusion of a root"* [12].

3. Their claim that small genome size affects rooting and induces attraction artifacts is conceptually and empirically false. Their recitation that organisms and viruses with small genomes (irrespective of their taxonomic affiliation) would be attracted to basal branches of our trees is incorrect. During searches of tree space and prior to rooting, we optimize character change in unrooted networks. This allows unrestricted gains and losses of domain occurrence or abundance throughout their branches. Thus, *rooting plays no role in defining unrooted tree topology* and cannot be distorted by genome size, which is a property of taxa (not individual characters changing in trees). This was already described in our supplementary text [4] and made explicit in a recent phylogenetic reconstruction study [13]. Polarization is only applied empirically *a posteriori*: *(a)* considering character spread in nested branches while accounting unproblematically for homoplasy (Weston's rule), *(b)* searching for most parsimonious solutions with Lundberg

bioRxiv preprint doi: https://doi.org/10.1101/050179; this version posted April 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

while treating homologies as taxic hypotheses, and *(c)* allowing gradual and punctuated build-up of evolutionary emergence of protein structures, including gain and loss, that complies with the principle of spatiotemporal continuity (PC), Leibnitz's *lex continui*. These three mutually supportive technical and biological axiomatic criteria were confirmed experimentally by Venn group distribution in ToDs and by visualizing clouds of proteomes in temporal space (*see* Figs. 5 and 8 of [4]). Felsenstein's suggestion of inverse polarization [14] of our ordered (Wagner) characters, which can be polarized in only two directions, produced suboptimal trees (e.g., Figs. 3 and 4 of [6]). Empirically, plotting the node distance (*nd*) for each terminal node (i.e. taxa) from the root node of a ToL – on a scale from 0 (most basal) to 1 (most recent) –against the genome sizes of taxa showed substantial scatter (*rho* = 0.80), poor lineal fits (several peaks and troughs with 100 iterations of the LOWESS fitting method and a smoothing of q = 0.05), shallow monotonic increases (flat lines in Archaea and Bacteria), and no distortions/mixing of taxa among the four supergroups, viruses, Archaea, Bacteria and Eukarya. Figure 1A shows the plot for viral taxa located at the base of the rooted ToL. Genomes of similar sizes were scattered throughout the *nd* axis suggesting that genome size was not a significant determinant of taxa position in our trees. Similarly, genome size scatter for individual *nd* increased towards the base of the trees, including scatter for supergroups (*rho* values of 0.55, 0.63, 0.75 and 0.85 for viruses, Archaea, Bacteria and Eukarya respectively), debunking the alleged basal attraction artifact. The order of appearance of supergroups matched their proteomic complexity, from simple to complex, which also matched scaling patterns of use and reuse of structural domains in proteomes [6]. This emerging property of trees supports evolution's PC.

Labeling the phylogenetic positions of the smallest proteomes in our trees confirmed that the smallest genomes were neither attracted towards the root nor caused distortions in the four-supergroup ToL (see also Fig. 7A in [4]). Here, Figure 1B showcases the phylogenetic positions of the "smallest" proteomes in each of the four supergroups (Archaea, Bacteria, Eukarya, and viruses) in our ToL. For clarification, the smallest proteomes included *Ignicoccus hospitalis* (Archaea, 213 fold superfamilies), *Lactobacillus delbrueckii* (Bacteria, 261), *Ashbya gossypii* (Eukarya, 326), and the bat cyclovirus (virus, 1). We also labeled the 9 viruses (4 RNA and 5 dsDNA) that produced phylogenetic distortions in the trees of Harish et al. [3]. Following the Harish et al. [3] logic, one should expect for the smallest proteomes to "fight" for the basal position within each of the four-supergroup subtrees or appear together closest to the root, irrespective of their taxonomic affiliation [2,3]. Contrary to that, the smallest proteomes appeared at well-derived positions (see arrow pointers) and did not cause any phylogenetic distortions or mixing of taxa from different supergroups! Even the smallest virus (the 1.7kb bat cyclovirus encoding a single fold superfamily) did not appear at the cluster of basal RNA viruses but was clustered with its closest evolutionary relative, the Dragonfly cyclovirus at more derived positions (red arrow in Figure 1B and at *nd* = 0.11 in Figure 1A). Thus, claims that *"the rooting in viral lineages is an inevitable consequence of pre-specifying '0' or 'absence' as the ancestral state",* that *"the position of the root depends on the smallest genome in the sample"* [2], and that we *"recognized anomalous effects of including small genomes in reconstructing the ToL"* [3] are therefore conceptually and empirically false. The contradictory results obtained by Harish

bioRxiv preprint doi: https://doi.org/10.1101/050179; this version posted April 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

et al. [2,3] therefore stem from (i) their misunderstanding of our rooting approach and cladistics, and (ii) intentionally/unintentionally missing the taxa sampling rationale that was well explained in our study [4]. To clarify, we took great care in sampling an equal number of cellular organisms from each supergroup (34 each) and at least five viruses from each family/order (87 viral families and 266 viruses were represented in our trees [4]). Indeed, our ToL (Figure 7 in [4]) revealed many monophyletic groups of viruses that were consistent with ICTV classifications and was far more complicated than showed/claimed by Harish et al. [3]. In turn, and in their haste to claim phylogenetic distortions, Harish et al. [3] likely handpicked problematic taxa (only 9 viruses in comparison to our 266) to claim phylogenetic distortions and to misguide the interpretation of our phylogenomic methodology (please also see Section 5). This was a crucial mistake since we had explicitly stated that taxa must be sampled broadly for the success of genome composition based phylogenies (see supplementary material in [4]). Even, the addition of the extremely reduced *Rickettsia prowazeki* and *Nanoarchaeum equitans* genomes (originally excluded because of their lifestyles) to our dataset, did not cause any topological distortions to our trees (data not shown), in contradiction to claims of Harish et al. [3].

4. Confusion of characters and taxa bootstraps their preconceptions. In rushing their unsupported claim that the rooting of trees of structural domains (ToDs) is also unreliable and affected by small genome size, they wrongly considered ToDs as being *"uninterpretable in terms of the definition of the* (domain) *superfamilies which it comprises",* because homology *"within"* superfamilies *"can be ascertained based on similarity of sequence, structure and function"*. But superfamilies are the taxa and proteomes the characters, and definitions of taxa (superfamily hidden Markov models) do not need to follow either definitions of characters (superfamily growth in proteomes) or statements of homology tested in ToDs. What is *'uninterpretable'* however is the putative effect of genome size on ToDs, since each proteome embodies a character, which by definition (Kluge's Auxiliary Principle) is independent of others. So fiction bootstraps their preconceptions, including the idea that domains, the evolutionary units of proteins, do not evolve. Are 1,200 structural folds fortuitous findings or the makings of intelligent cause? Where does significant evolutionary signal of the ToDs, including a match to the geological record [15], come from? Even an exploration of the mapping of functions in genotype space shows the centrality of structure in defining evolutionary constraints [16].

5. Their analysis failed to avoid sampling pitfalls and problematic taxa. They confused exclusion of taxa engaged in obligate parasitism and symbiosis in cellular organisms with avoidance of genome size attraction artifacts, when in fact our intention was to exclude organisms with ill-defined hologenomes of holobiont collectives (the host and its associated organismal communities), which are known to complicate definitions of taxa [17,18]. No such attempt was extended to the viral supergroup since one hallmark of viruses is harboring a life cycle with strict dependence on a host. We explored their dataset and found that sampling of taxa was limited and imbalanced (Table 1) and included questionable taxa that were likely cherry-picked from extreme proteomic outliers (e.g. Fig. 1C in [4]), and sometimes selected outside our initial sampling (e.g. *Cand*. Nausia deltocephalinicola). For example, *Cand.* Tremblaya princeps included in

bioRxiv preprint doi: https://doi.org/10.1101/050179; this version posted April 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

their trees (Figure 2B in [3]) is part of a three-pronged endosymbiotic organismal system (bug in a bug in a bug) [19]. Its genome encodes only 55 universal domain superfamilies. It is not considered an independent organism since it depends on its host (*Planococcus citri*) and its endosymbiont (*Cand.* Moranella endobia) to synthesize essential metabolites [20]. To quote López-Madrigal et al. *"The genome sequence reveals that 'Ca. Tremblaya princeps' cannot be considered an independent organism but that the consortium with its gammaproteobacterial symbiotic associate represents a new composite living being"* [20]. Similarly, *Cand*. N. deltocephalinicola is an obligate endosymbiont of leafhoppers, which harbors the smallest known bacterial genome and encodes only 53 universal superfamilies [21]. These extreme proteomic outliers do not bias tree reconstructions because of their size nor induce *"grossly erroneous rootings"* [3]. Instead, their hologenomes arise from relatively modern genomic exchanges and recruitments likely resulting from complex trade-off relationships that complicate the dissection of their evolutionary origin and their definition as single taxon in phylogenetic data matrix. Phylogenetically, they represent 'problematic' taxa that should be excluded from analysis pending further understanding of their genetic make up. The intentional inclusion of problematic taxa is expected to generate biased reconstructions; see [22] for a dinosaur phylogeny example and the detection of problematic taxa with double decay analysis. Instead, our decision to exclude organisms that do not engage in free-living relationships avoids these kinds of pitfalls [23]. In contrast to the effect of lifestyles, comparing trees built from information-related and nonrelated domain families uncovered that tree reconstruction is refractory to biases resulting from horizontal gene transfer [23], a much more serious putative effect on tree reconstruction.

They state that our Venn diagrams and summary statistics of domain distributions in supergroups are unconvincing in light of other comparative genomic analyses [24,25]. However, Abroi and Gough [24] argued that viruses may be a source of new protein fold architectures, a conclusion strongly supported by our Venn analysis, and Abroi [25] showcases the distribution and sharing of domain superfamilies between viral replicon types and cells, which is remarkably consistent with our analysis [4]. There are no irreconcilable differences between these studies. Instead, our phylogenomic analysis dissects the alternative evolutionary scenarios that can be posed with the comparative method [4,25]. For example, we also mapped virus-host information to the Venn distribution of protein fold superfamilies and identified 68 fold superfamilies that were (remarkably) shared by archaeo-, bacterio-, and eukaryoviruses (Fig. 3 in [4]). These superfamilies included many ancient folds involved in cellular metabolism and hinted towards an origin of viruses predating the origin of modern cells. Harish et al. [3] therefore I ignore the plethora of evidence given in our article (Figs. 1-8 in [4]) supporting the ancient origin of viruses and instead rely on customized phylogenetic trees to claim invalidity of our results.

Conclusions. Harish et al. [2,3] believe we root our ToLs *a priori* with an indirect method and an outgroup taxon they confuse with an ancestor, when in fact we root our ToLs *a posteriori* using a direct method compliant both with Weston's rule and experimentation. They claim our rooting method attracts organisms and viruses with small genomes to the base of rooted trees when in fact tree topology is established during

bioRxiv preprint doi: https://doi.org/10.1101/050179; this version posted April 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

unrooted tree optimization and prior to character polarization. It is ironic that attempts to controvert our direct methods of character polarization come from authors that are themselves proponents of the use of polarized characters, but with arbitrary transformation costs carefully engineered *a priori* to attract large eukaryotic genomes to the base of their trees [1]. These characters violate the 'triangle inequality' [26], a fundamental property conferring metricity to phylogenetic distances. Its violation invalidates phylogenetic reconstruction [27]. Quoting the cladistics opinion of W. C. Wheeler [26] on the validity of characters with arbitrary transformation costs: *"Although the matrix character optimization algorithm does not require metricity, biologically odd results may occur otherwise. As an example, an additional state could be added to an existing set, with very low transformation cost to all other elements* ($\sigma_{k,0} < ½ \min \sigma_{i,j}$). *The median state at all internal 2 vertices (V \ L) would then be this new state for all trees, no matter what the leaf conditions were"* (page 184). He further illustrates the unforeseen consequences of optimization with non-metric distances using the well-known NP-hard 'traveling salesman problem', a salesman that wishes to visit a number of cities while minimizing travel time. This is a notorious task known to require considerable travelling optimization. The use of non-metric distances however makes a city have zero distance to all other cities, creating a 'wormhole' in space-time that allows to reach all cities at zero cost. Such bizarre property has dire consequences for the recovery of a correct tree and invalidates the Harish et al. [1] approach. Add to this the fact that their loss-favoring step matrices require that they be solely optimized on a rooted tree, making rooted tree reconstruction prone to 'large genome attraction' artifacts. In addition to this self-inconsistency, transformation costs also violate genomic scaling and processes responsible for scale-free behavior of proteins, challenging evolution's PC and artificially forcing biological innovations to the base of universal trees [6], effectively resulting in an ancestral clumping of irreducible complexity.

References

1. A. Harish, A. Tunlid, C. G. Kurland, Rooted phylogeny of the three superkingdoms. *Biochimie* **95**, 1593-1604 (2013).
2. A. Harish, A. Abroi, J. Gough, C.G. Kurland, eLetter concerning Nasir and Caetano-Anollés' conclusions about the origins of viruses. Sci Adv, eLetter http://advances.sciencemag.org/content/1/8/e1500527.e-letters#
3. A. Harish, A. Abroi, J. Gough, C.G. Kurland, Did viruses evolve as a distinct supergroup from common ancestors of cells? bioRxiv preprint server (posted Apr. 18, 2016), doi: http://dx.doi.org/10.1101/049171.
4. A. Nasir, G. Caetano-Anolles, A phylogenomic data-driven exploration of viral origins and evolution. *Sci. Adv.* **1**, e1500527 (2015).
5. A. Nasir, K.M. Kim, G. Caetano-Anolles, Authors' response to Harish et al.: No 'small genome attraction' artifact. Sci Adv, eLetter http://advances.sciencemag.org/content/1/8/e1500527.e-letters#
6. K. M. Kim, A. Nasir, G. Caetano-Anollés, The importance of using realistic evolutionary models for retrodicting proteomes. *Biochimie* **99**, 129-137 (2014).
7. J. S. Farris, Methods for computing Wagner trees. *Syst. Zool.* **19**, 83-92 (1970).
8. J. Lundberg, Wagner networks and ancestors. *Syst. Zool.* **18**, 1-32 (1972).

bioRxiv preprint doi: https://doi.org/10.1101/050179; this version posted April 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

9. P. H. Weston, Indirect and direct methods in systematics, in *Ontogeny and Systematics*, C, J. Humphries, Ed. (Columbia University Press: New York, 1988), pp. 27–56.
10. P. H. Weston, Methods for rooting cladistic trees, in *Models in Phylogeny Reconstruction*, Systematics Association Special Volume No. 52, D. J. Siebert, R. W. Scotland, D. M. Williams, Eds. (Clarendon Press, Oxford, 1994), pp. 125–155.
11. H. N. Bryant, Hypothetical ancestors and rooting in cladistic analysis. *Cladistics* **13**, 337-348 (1997).
12. A. V. Brower, M. C. de Pinna, Homology and errors. *Cladistics* **28**, 529-538 (2012).
13. A. Nasir, K. M. Kim, G. Caetano-Anollés, Global patterns of protein domain gain and loss in superkingdoms. *PLoS Comput. Biol.* **10**, e1003452 (2014).
14. J. Felsenstein, The statistical approach to inferring phylogeny and what it tells us about parsimony and character compatibility, in Cladistics: Perspectives on the Reconstruction of Evolutionary History, T. Duncan, T. F. Stuessy, Eds. (Columbia University Press, New York, 1984), pp. 169-191.
15. M. Wang, Y.-Y., Jiang, K.M. Kim, G. Qu, H.-F. Ji, H.-Y. Zhang, G. Caetano-Anollés, A molecular clock of protein folds and its power in tracing the early history of aerobic metabolism and planet oxygenation. *Mol. Biol. Evol.* **28**, 567-582 (2011).
16. E. Ferrada, A. Wagner, Evolutionary innovations and the organization of protein functions in genotype space. *PLoS ONE* **5**, e14172 (2010).
17. E. Rosenberg, I. Zilber-Rosenberg, The hologenome concept: human, animal and plant microbiota (Springer, 2013).
18. P. J. Keeling, Endosymbiosis: bacteria sharing the load. *Curr Biology* **21**, R623-R624 (2011).
19. J. P. McCutcheon, C. D. von Dohlen, An interdependent metabolic patchwork in the nested symbiosis of mealybugs. *Current Biology*. **21**, 1366-1372 (2011)
20. S. López-Madrigal, A. Lattore, M. Porcar, A. Moya, R. Gil, Complete genome sequence of "Candidatus Tremblaya princeps" strain PCVAL, an intriguing translational machine below the living-cell status. *J Bacteriol* **193** 19: 5587-5588 (2011)
21. G. M. Bennett, N. A. Moran, Small, smaller, smallest: the origins and evolution of ancient dual symbioses in a Phloem-feeding insect. *Genome Biol. Evol*. **5**, 1675-1688 (2013).
22. M. Wilkinson, J.L. Thorley, P. Upchurch. A chain is no longer than its weakest link: double decay analysis of phylogenetic hypotheses. *Syst. Biol*. **49**, 754-776 (2000).
23. K. M Kim, G. Caetano-Anollés. The evolutionary history of protein fold families and proteomes confirms that the archaeal ancestor is more ancient than the ancestors of other superkingdoms. *BMC Evol. Biol.* **12**, 13 (2012).
24. A. Abroi, J. Gough, Are viruses a source of new protein folds for organisms? - Virosphere structure space and evolution. *Bioessays* **33**, 626-635 (2011).
25. A. Abroi, A protein domain-based view of the virosphere–host relationship. *Biochimie* **119**, 231-243 (2015).
26. W. C. Wheeler, *Systematics: a course of lectures* (John Wiley & Sons, Hoboken, New Jersey, 2012).
27. W. C. Wheeler, The triangle inequality and character analysis. *Mol. Biol. Evol*. **10**, 707–712 (1993).

bioRxiv preprint doi: https://doi.org/10.1101/050179; this version posted April 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 1. Fact-checking the narrative of Harish et al. [2,3]. A somehow similar table fact-checking can be found in our eLetter response [5].

Fiction	Fact	Discussed in:
We root trees with the *"outgroup comparison method"* [2]	We do not rely on outgroup taxa and indirect methods. Instead, we root trees with Weston's generality criterion (a direct method) implemented using Lundberg [8].	Section 1
We use an *"an artificial all-zero taxon… an 'all-absent' hypothetical ancestor"* [2,3]	We do not combine outgroups and ancestors, an approach known to be invalid [11].	Section 1
"Including the hypothetical ancestor during tree estimation amounts to a priori character polarization" [3]	We polarize character transformations *a posteriori*, empirically and most parsimoniously, and complying with Weston's generality criterion.	Section 2
"Unrooted trees describe relatedness of taxa based on graded compositional similarities of characters" [3]	The search of tree space using maximum parsimony as an optimality criterion is defined by homology relationships not graded compositional similarities.	
"Accordingly, we can expect the 'all-zero' pseudo-outgroup to cluster with … proteomes described by the largest number of '0s' in the data matrix" [2,3]	During phylogenetic searches, we first optimize character change in unrooted networks using the Wagner algorithm [7]. The topology of rooted trees cannot be predicted from patterns in character state vectors of ingroup or outgroup taxa and cannot be affected by genome size.	Section 3
"The rooting in viral lineages is an inevitable consequence of pre-specifying '0' or 'absence' as the ancestral state" [2]	Character polarization plays no role in defining unrooted tree topology and cannot be distorted by genome size, a property of taxa, not characters. This claim is also debunked empirically by labeling the positions of smallest proteomes in our trees (please see the Figure).	Section 3
"Including viruses in the analyses draws the root towards the smaller viral proteomes" [3]	A simple node distance (*nd*) versus genome size plot dispels their putative 'small genome attraction artifact' (Figure 1A).	Section 3
"Detailed reexamination [of our] *phylogenomic approach suggests that small genomes systematically distort phylogenetic reconstructions"* [3]	They reconstruct trees (their Figs. 1 and 2) using an incorrect outgroup-driven mimic of our method to fictionalize both our conclusions and methodologies. Trees describes phylogenetic relationships between limited and imbalanced taxa selected from extreme outliers (instead of using objective taxon selection criteria already explained in [4]).	Sections 3 and 5
The tree of structural domains (ToD) *"appears uninterpretable in terms of the definition of superfamilies which it comprises"* [2]	Harish et al. confuse characters and taxa and extend their unsubstantiated claims to ToDs. To build ToDs, domain homologies are defined with robust hidden Markov models and delimit taxa. Their proteomic abundances are used as characters. ToDs contain significant phylogenetic signal that follows a molecular clock linking the molecular and geological records [15].	Section 4
"Half of the sampled proteomes were analyzed (Figs. 1 and 2) for computational simplicity" [3]	They included only 16 eukaryal, 17 archaeal, 17 bacterial, and 5-9 viral proteomes, which only represent 16% of our taxa [4]! Sampling was limited and imbalanced and taxa selection followed no rationale.	Section 5
"Rooting experiments (Figs. 1	When running their experimental mimic, they cherry-	Section 5

bioRxiv preprint doi: https://doi.org/10.1101/050179; this version posted April 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

and 2) clearly show that the position of the root depends on the smallest genome in the sample when an all-zero outgroup is used" [3]	picked extreme outliers, such as *Candidatus* Tremblaya princeps, and *Cand.* Nasuia deltocephalinicola, which were excluded from our phylogenies because taxon definition was compromised by extreme reduction and ill-defined hologenome make-up. They also included only 9 viruses to distort tree reconstruction while we included 266 viruses from tens of families following an already explained rationale for properly surveying taxonomic groups for the success of genome-composition phylogenies [4].	
The exclusion of highly reduced 'parasitic' proteomes appears to be inconsistent with the inclusion of viruses [3]	Our exclusion and inclusion of taxa followed clear rationale. Exclusion of organisms engaging in obligate parasitism and symbiosis and representing genomic outliers ensured integrity of definition of taxa. Inclusion of representatives of all viral groups portrayed the entire viral supergroup, which is unified by its parasitic lifestyle.	Section 5
"Inferences based on statistical distributions of superfamilies alone are unconvincing, especially in light of other recent analyses" [2]	The other recent analyses that are mentioned [24,25] failed to couple the comparative genomic studies to phylogenomic reconstructions, which in our case [4] helped weed out evolutionary scenarios (e.g. Fig. 5 of [25]) that were non-significant or incompatible with reconstructed history.	Section 6
"We emphasize that for an unrooted ToL as well as for rooting by other methods (referring to [1])*, small proteome size is not an irreconcilable feature of genome-trees"* [3]	Harish et al. are proponents of polarized characters with arbitrary transformation costs [1], which violate the 'triangle inequality' of phylogenetic distances and are engineered to attract large genomes to the base of their trees [6]. Their use of unrealistic evolutionary assumptions does have irreconcilable consequences for the correct reconstruction of trees [6].	Conclusion

bioRxiv preprint doi: https://doi.org/10.1101/050179; this version posted April 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

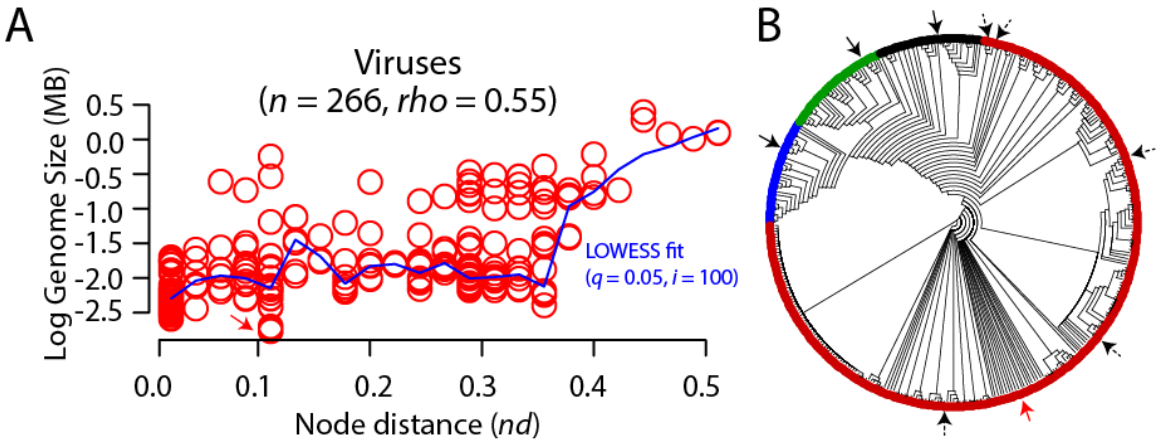

Figure 1. Trees of proteomes are robust and insensitive to the effects of genome size. **A.** Scatter plot describing the relationship between genome size and node distance (*nd*) for viral taxa used in our study [4]. For clarification, our data set included 266 viral proteomes including up to 5 viral representatives from each known viral family/order and belonging to each of the seven replicon types seen in viruses, as in [4]. The blue line describes the nature of the relationship, as determined by the Locally Weighted Regression Scatter Plot Smoothing (LOWESS) method, which obtains a smoothed curve by fitting successive regression functions. The red arrow indicates the smallest virus of the genomic set, the bat cyclovirus (1.7kb genome and encoding a single fold superfamily domain). High scatter values of the plot (*rho* = 0.55) indicate no 'small genome attraction' artifact that would pull small genomes toward the base of the tree, i.e. towards *nd* = 0. **(B)** The single most parsimonious tree (length = 45,935, retention index = 0.83, g_1= –0.31) describing the evolution of 102 cellular organisms and 266 viruses (described in [4]). The smallest proteomes in each cellular supergroup are represented by black arrows (see text for description). The smallest viral proteome (bat cyclovirus) is labeled with a red arrow. Viruses sampled by Harish et al. [3] are indicated with dashed arrows. The names of taxa are not shown because they would not be visible. Instead, the positions of terminals were colored according to supergroup, green (Eukarya), blue (Bacteria), black (Archaea) and red (viruses).