Known phyla dominate the RNA virome in Tara Oceans data, not new phyla as claimed

Robert C. Edgar[1,*,‡]

[1]**Unaffiliated**

[‡]**Corresponding author: robert@drive5.com**

ABSTRACT

The Tara Oceans project claimed to identify five new RNA virus phyla which "are dominant in the oceans". However, their own analysis assigns 28,353 of their RdRp-containing contigs to known phyla vs. 886 (2.8%) to the five claimed new phyla combined. The five small "mega-clusters" classified as new virus phyla are consistent with endogenous insertions into host species, or with sub-groups of known virus phyla at ranks between class and sub-family, or with the noise I would expect to find in the tail of residual clusters generated from metatranscriptomics data. Their claim of 97% accuracy in recapitulating GenBank phylum and class annotations by clustering is found to be false.

Known phyla dominate the oceans

Tara Oceans Zayed et al. (2022) searched metatranscriptome assemblies of ocean samples for contigs containing viral RNA-dependent RNA polymerase (RdRp), the signature gene for *Orthornavirae*. I re-analyzed their Table S6, finding the results shown in my Table 1. The modal classification assigned to Tara contigs is "Not determined" (13,436 contigs), followed by *Pisuviricota* (13,014) and *Lenarviricota* (7,060), with 886 (2.8%) assigned to the claimed novel phyla. Thus, according to their own analysis, known phyla dominate their data.

Endogenous host insertions

Viral insertions into host genomes are well known from retroviruses. It is less widely known that host insertions also occur with RNA viruses having no DNA stage in their life cycle Holmes (2011). Fig. 1 shows two examples: an RdRp-like protein from *Proteus mirabilis*, a bacterium associated with Arctic cetacea Chan et al. (2001), and an acid transporter protein from *Symbiodinium microadriaticum*, a marine dinoflagellate. Both sequences have hits with low E-values to RdRp hidden Markov models (HMMs) from PFAM Bateman et al. (2004), and have high similarity to known A, B and C motifs in the RdRp palm domain. Three of the six essential catalytic residues are substituted in the transporter

protein. Thus, strong sequence similarity with viral RdRp is not sufficient to establish viral origin or polymerase function, and substitutions in the catalytic residues are suggestive of an endogenous host insertion with diverged function. Such substitutions are found in novel taxa claimed by Tara, including a D→E substitution in motif A of "Taraviricota" and a G→I substitution in motif C of "novel negarnaviricot class 67". In the Serratus project Edgar et al. (2022), my colleagues and I estimated that roughly 1% of our putative RdRps would be endogenous insertions, noting that these insertions cannot be distinguished from viral RdRp by sequence alone. This is in the ballpark of the 2.8% of putative RdRps assigned to novel phyla by Tara Oceans, and in my view the conservative conclusion is that the unclassified Tara sequences could be endogenous insertions rather than viral RdRps. In the rest of this paper, I will consider whether their claims of novel phyla are reasonable assuming that these sequences are actually viral RdRps, emphasizing that I do not believe this assumption is well-supported by Tara Oceans' data.

Mega-cluster analysis claiming new phyla

Translated RdRp sequences from Tara contigs were combined with RdRps from GenBank plus another recent ocean study Wolf et al. (2020). The combined RdRps were reduced to 6,238 non-redundant "centroids" at 50% identity by UCLUST Edgar (2010). The centroids were then clustered by MCL Enright et al. (2002), giving 19 "mega-clusters". Each mega-cluster was assigned a "mega-taxon" based on the majority taxon or taxa according to annotations of its GenBank sequences. Table 2 is a summary of mega-clusters reported in their Table S7. This shows that the first two clusters are large and correspond to phyla according to the authors' annotation, but thereafter cluster sizes decrease rapidly, splitting phyla into progressively smaller fragments. In the tail of small clusters where new phyla are claimed, the mega-taxa assigned by the authors are at ranks ranging from class (cluster 15, *Vidaverviricetes*) to sub-family (clusters 9 and 11, *Spinareovirinae* and *Sedoreovirinae*). The 515 of 3,538 GenBank centroids (15%) classified as "Not determined" by Table S6 are excluded from the mega-cluster analysis in Table S7, leaving open the possibility that they would cluster with the claimed new phyla. Thus, the authors' mega-cluster analysis is consistent with the more conservative hypothesis that the claimed new phyla belong to novel subgroups, or to known but unnamed subgroups, within known phyla at ranks between class and sub-family.

Phylum and class for RNA viruses

Prior to 2017, the highest rank for RNA viruses was order. ICTV introduced class and phylum ranks into their system on the basis of a phyologenetic tree from a controversial paper Wolf et al. (2018), which I will call "Wolf2018" here. The Wolf2018 tree was constructed from RdRp sequences for 4,617

RNA virus species using a combination of automated and manual methods which were not fully explained. This approach has been criticized on the grounds that there is not enough information in an RdRp alignment to reconstruct ancient phylogeny Holmes and Duchêne (2019). I agree, and in my view there are two additional fatal flaws in the Wolf2018 approach as a basis for defining higher taxonomic ranks: (1) their methods are not explained well enough to enable independent replication of their results, and (2) no consideration was given to scaling the methods to accommodate the deluge of new RNA species revealed by metagenomics. If by their own standards extensive manual intervention was required to align 4,617 RdRps, how would they classify the 131,957 novel species found by Serratus Edgar et al. (2022)? As a preliminary and tentative attempt to resolve ancient RNA virus phylogeny, their approach seems reasonable, but not as a basis for a radical expansion of ICTV taxonomy. Strangely, ICTV chose to define phyla which are not monophyletic in the Wolf2018 tree (*Duplornaviricota* is paraphyletic with *Negarnaviricota*), in apparent violation of ICTV rule 3.3.1 which calls for taxa to be "monophyletic groups" King et al. (2011). This puzzling choice is likely to mislead biologists, as illustrated by the Tara authors who incorrectly state that "*Duplornaviricota* and *Negarnaviricota* were strongly monophyletic [in the Wolf2018 tree]".

Phylum and class for new RNA viruses

It seems self-evident that taxonomy should be based on objectively observable traits, but ICTV did not provide criteria enabling a biologist to classify a virus or its RdRp sequence to a phylum or class, or to determine whether a virus or its RdRp sequence belongs to a new phylum or a new class. This appears to violate ICTV rule 3.5 which states "[t]axa will be established only when representative member viruses are sufficiently well characterized and described in the published literature so as to allow them to be identified unambiguously and the taxon to be distinguished from other similar taxa." King et al. (2011). As noted above, estimating phylum-level RNA virus phylogeny from RdRp sequences has recently been assessed to have a degree of difficulty somewhere between impossible Holmes and Duchêne (2019) and beyond the capabilities of state-of-the-art algorithms, requiring extensive manual intervention Wolf et al. (2018). It would therefore be a revolutionary and surprising advance in phylogenetics if a generic clustering algorithm solved this problem.

Distance-based clustering

A matrix of evolutionary distances can be used to estimate phylogeny. Here, "evolutionary distance" has a precise technical meaning: the length of the shortest path between two leaves in the phylogenetic tree (measured by the number of mutations, e.g. the number of substitutions per site) must be proportional to the distance between the sequences at those leaves. Then it is possible to work backwards from

a pair-wise distance matrix to reconstruct the tree if distance estimates are sufficiently accurate Durbin et al. (1998). RNA virus mutation rates, measured as μ = substitutions per nucleotide per cell infection, vary by at least two orders of magnitude, quite likely more given the difficulty of measuring the rate and the small number of viruses currently characterized Peck and Lauring (2018). For example, Human rhinovirus 14 has $\mu = 4.8 \times 10^{-4}$ Wang et al. (1998), roughly 300 times faster than Influenza B virus with $\mu = 1.7 \times 10^{-6}$ Nobusawa and Sato (2006). While a tree can still be inferred from an evolutionary distance matrix when rates are highly variable in different branches, a generic clustering algorithm cannot reconstruct monophyletic subgroups from such a matrix, as shown by the example in Fig. 2. Instead, a neighbor-joining algorithm is needed Saitou and Nei (1987); Durbin et al. (1998), i.e. one which is phylogeny-aware. It is therefore clear *a priori* that the Tara Oceans clustering approach cannot reconstruct monophyletic groups because the starting point is a matrix of pair-wise BLAST bit-scores, which are not evolutionary distances Koski and Golding (2001), and MCL is a generic clustering algorithm which is entirely ignorant of phylogeny.

Recapitulating GenBank taxonomy

The authors claim that their classification method based on MCL clusters "nearly completely recapitulated the previously established phylogeny-based ICTV-accepted taxonomy at the phylum and class ranks (97% agreement)". Here, the authors are referring to RdRp taxonomy annotations in GenBank, most of which are tentative assignments from sequence made by individual submitters, not official classifications by ICTV. Only a small subset of known RNA virus species have been classified by ICTV; these have phyla and class assignments based on a controversial and possibly unreproducible tree that surely has many incorrect edges, especially in deeper branches. It is therefore highly unlikely that the true tree would be consistent with GenBank taxonomy with 3% conflicts or less, and practically inconceivable that a generic clustering method based on BLAST bit-scores could do this. The claim is in fact false (explained later below).

Identifying new phyla

The Tara authors proceed to make a much stronger claim, namely that small residual mega-clusters which are not classified to mega-taxa are new phyla. A skeptical reader might wonder why an unclassified cluster must be a new phylum, especially considering that the rank of a mega-cluster can be as low as subfamily, but the Tara paper does not explain. To fill the gap, I will present my own view of how new phyla might be identified. If phyla are defined by phylogenetic trees such as Wolf2018, then discrimination of in-groups (within existing phyla) vs. out-groups (potential new phyla) relative to existing phyla requires a convincing prediction of the correct branching order at distances of several substitu-

tions per site where it is controversial whether the best phylogenetics methods work at all. Predicting branching order is more difficult than predicting large approximate clades (as attempted in Wolf2018) because large subgroups may be resolved without resolving the order in which they divided. ICTV has not defined objective traits for existing phyla, so in fact even if an out-group could be identified, there are no official criteria to decide whether it extends an existing phylum or should be classified as a new phylum. From this perspective, the Tara authors are making the extraordinary claim that MCL clusters have sufficient precision in resolving evolutionary events from billions of years ago that they can identify five small out-groups that are separated from existing phyla and from each other by sufficiently long branches that it is self-evident they should all be classified as new phyla.

High bootstrap values do not confirm RdRp trees

An apparent counter-argument to the difficulty of estimating deep RNA virus phylogeny is found in the high bootstrap values reported by Wolf2018 and Tara Oceans for their phylum-level trees. However, I have previously shown Edgar (2021) that the high bootstrap confidence scores reported for the Wolf2018 tree are almost certainly artifacts of systematic errors in their alignment. High bootstraps reported by Tara Oceans for their phylum-level tree can be similarly explained as due to systematic errors arising from its progressive construction from phylum- and claimed-phylum level sub-alignments and consensus sequences. This construction is strongly biased towards making intra-phylum alignments more similar than inter-phylum alignments, creating divisions which will then be reflected in the tree. The same bias towards phylum divisions is likely to be found in several alignment columns in which case they will be reproduced in a large majority of bootstrap replicates, giving spuriously high confidence values.

False claim of 97% agreement with GenBank taxonomy

The "97% agreement" claim by Tara refers to the circled value of adjusted Rand index (ARI) 0.97 at inflation (inf) 1.1 in their Fig. 1B, reconstructed in my Fig. 3 here. Tara's script to generate Fig. 1B is Figure1B_FigS3/Adjust_Rand_phyla_and_classes.R in repository https://bitbucket.org/MAVERICLab/global-rna-virus-evolution-2021/. Inspection of the source code indicated that the graph for ARI is obtained by comparing mega-taxa with MCL clusters. Mega-taxa are names assigned to the 19 MCL clusters generated at inf=1.1 as shown in the first and last columns of my Table 2. I verified this by writing my own python scripts to calculate ARI from their data (as shown in my Fig. 3). This exercise confirmed that ARI=0.97 is obtained from their Tables S6 and S7 at inf=1.1 by comparing mega-taxa with mega-clusters after excluding GenBank centroids classified as "Not determined". Thus, 97% agreement is not between MCL clusters and GenBank taxonomy at phylum and

class as stated; rather, it is between mega-taxa (derived from majority GenBank taxa in MCL clusters at inf=1.1) and mega-clusters (which are exactly MCL clusters at inf=1.1). Agreement of MCL clusters with GenBank phylum is much lower, e.g. 20% errors at inf=1.1. This probably under-estimates phylum disagreement because some "Not determined" GenBank centroids are readily assigned to mega-taxa from their GenBank taxonomies, but were excluded from clustering. For example, AAS86764.1 is annotated as *Pisuviricota* by GenBank but is "Not determined" according to Tara Table S6, possibly because AAS86764.1 did not cluster with *Pisuviricota* in a previous iteration of mega-cluster/mega-taxon optimization. I could not verify this because Tara's method for assigning mega-taxa is not described and no code is provided.

Noise cannot be filtered with 100% accuracy

The path from unidentified sequence to novel virus is fraught with pitfalls. Tara Oceans' data generation and analysis included sequencing, assembly, identification of viral RdRp and classification into known and novel taxa. All of these steps are error-prone. Sequencing libraries have reverse-transcriptase and PCR errors, and sequencing *per se* has base call errors. Assembly errors cause chimeric contigs and other artifacts. Low-coverage contigs may propagate incorrect nucleotide base calls through to the translated sequence. Viral RdRp may be misidentified due to false-positive HMM hits with low E-values; in particular, endogenous host insertions are indistinguishable from viral RdRp. Error filtering methods are not perfect, and any classification method, whether it is based on clustering or based on alignments and trees, will have a non-zero error rate. Combined noise from imperfect data and imperfect classification algorithms should be expected to generate somewhere around a few percent of unclassified sequences due entirely to data errors and/or classification false negatives after all quality control measures are applied. The tail of small MCL clusters where Tara claims new phyla look like the noisy residue I would expect to see after attempting to classify metatranscriptomics sequences.

Conclusion

The small clusters claimed by Tara Oceans as novel phyla are consistent with host sequences due to endogenous insertions. Even if they could be shown to be viral, clustering is not capable of classifying to phylum, and the leap from unclassified RdRp-like sequence to novel virus phylum is not justified. In fact, phylum identification for RNA viruses is currently impossible in principle because ICTV has not provided objective criteria for classification.

Code availability

Scripts to reproduce tables and my Fig. 3 are deposited at https://github.com/rcedgar/tara_oceans

References

Babaian, A. and Edgar, R. C. (2021). Ribovirus classification by a polymerase barcode sequence. *BioRxiv*.

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., et al. (2004). The pfam protein families database. *Nucleic acids research*, 32(suppl_1):D138–D141.

Chan, O., Mukerjee, J., Kinoshita, R., and Yuen, C. (2001). Microbial flora of blowhole samples in captive bottlenose dolphins (tursiops truncatus aduncas-type) in hong kong, 1993–2000. In *Proceedings of the 32nd Annual Conference of the International Association for Aquatic Animal Medicine. Tampa, Florida*, pages 24–25.

Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. (1998). *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than blast. *Bioinformatics*, 26(19):2460–2461.

Edgar, R. C. (2021). Muscle v5 enables improved estimates of phylogenetic tree confidence by ensemble bootstrapping. *bioRxiv*.

Edgar, R. C., Taylor, J., Lin, V., Altman, T., Barbera, P., Meleshko, D., Lohr, D., Novakovsky, G., Buchfink, B., Al-Shayeb, B., et al. (2022). Petabase-scale sequence alignment catalyses viral discovery. *Nature*, 602(7895):142–147.

Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic acids research*, 30(7):1575–1584.

Holmes, E. C. (2011). The evolution of endogenous viral elements. *Cell host & microbe*, 10(4):368–377.

Holmes, E. C. and Duchêne, S. (2019). Can sequence phylogenies safely infer the origin of the global virome? *MBio*, 10(2):e00289–19.

King, A. M., Lefkowitz, E., Adams, M. J., and Carstens, E. B. (2011). *Virus taxonomy: ninth report of the International Committee on Taxonomy of Viruses*, volume 9. Elsevier.

Koski, L. B. and Golding, G. B. (2001). The closest blast hit is often not the nearest neighbor. *Journal of molecular evolution*, 52(6):540–542.

Nobusawa, E. and Sato, K. (2006). Comparison of the mutation rates of human influenza a and b viruses. *Journal of virology*, 80(7):3675–3678.

Peck, K. M. and Lauring, A. S. (2018). Complexities of viral mutation rates. *Journal of virology*, 92(14):e01031–17.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4):406–425.

Wang, W., Lee, W.-M., Mosser, A. G., and Rueckert, R. R. (1998). Win 52035-dependent human rhinovirus 16: assembly deficiency caused by mutations near the canyon surface. *Journal of virology*, 72(2):1210–1218.

Wolf, Y. I., Kazlauskas, D., Iranzo, J., Lucía-Sanz, A., Kuhn, J. H., Krupovic, M., Dolja, V. V., and Koonin, E. V. (2018). Origins and evolution of the global rna virome. *MBio*, 9(6):e02329–18.

Wolf, Y. I., Silas, S., Wang, Y., Wu, S., Bocek, M., Kazlauskas, D., Krupovic, M., Fire, A., Dolja, V. V., and Koonin, E. V. (2020). Doubling of the known set of rna viruses by metagenomic analysis of an aquatic virome. *Nature microbiology*, 5(10):1262–1270.

Zayed, A. A., Wainaina, J. M., Dominguez-Huerta, G., Pelletier, E., Guo, J., Mohssen, M., Tian, F., Pratama, A. A., Bolduc, B., Zablocki, O., et al. (2022). Cryptic and abundant marine viruses at the evolutionary origins of earth's rna virome. *Science*, 376(6589):156–162.

Tables and figures

Nr. Tara Ocean RdRp-contigs	*Classification according to Table S6*
13436	Not determined
13014	Pisuviricota
7060	Lenarviricota;others
5543	Kitrinoviricota
1971	Wei-like
1133	Negarnaviricota;Polyploviricotina
655	Duplornaviricota;Chrymotiviricetes
644	**New_12at1.1 (*"Taraviricota"*)**
528	Negarnaviricota;Haploviricotina
237	Duplornaviricota;Sedoreovirinae
154	Lenarviricota;Allassoviricetes
119	**New_14at1.1 (*"Pomiviricota"*)**
118	Permuto-like
105	**New_19at1.1 (*"Arctiviricota"*)**
40	Kitrino-like
29	Duplornaviricota;Vidaverviricetes
24	New_17at1.1 (Lenar-like)
14	**New_16at1.1 (*"Paraxenoviricota"*)**
4	**New_18at1.1 (*"Wamoviricota"*)**

Table 1. Tara Ocean contigs assigned to each mega-taxon according to their Table S6, showing that classified RdRp-containing contigs are dominated by known phyla. Mega-taxa corresponding to claimed new phyla are highlighted. The numbers shown here were generated by the following Linux command:

```
grep "This study" TableS6.tsv | cut -f8 | sort | uniq -c | sort -nr
```

"Mega-cluster"	Centroids	Rank according to "Mega-taxon"	"Mega-taxon"
1	2101	Phylum	Pisuviricota
2	1655	Phylum	Kitrinoviricota
3	1040	Split phylum	Lenarviricota;others
4	371	Split phylum	Lenarviricota;Allassoviricetes
5	269	Sub-phylum	Negarnaviricota;Polyploviricotina
6	235	Class	Duplornaviricota;Chrymotiviricetes
7	225	Sub-phylum	Negarnaviricota;Haploviricotina
8	105	?	"Wei-like"
9	52	Sub-family	Duplornaviricota;Spinareovirinae
10	50	?	"Permuto-like"
11	44	Sub-family	Duplornaviricota;Sedoreovirinae
12	**36**	Claimed phylum	**"Taraviricota"**
13	24	?	"Kitrino-like"
14	**9**	Claimed phylum	**"Pomiviricota"**
15	8	Class	Duplornaviricota;Vidaverviricetes
16	**5**	Claimed phylum	**"Paraxenoviricota"**
17	4	?	"Lenar-like"
18	**3**	Claimed phylum	**"Paraxenoviricota"**
19	**2**	Claimed phylum	**"Arctiviricota"**

Table 2. Summary of mega-clusters and mega-taxa in Tara Table S7. “Rank according to Mega-taxon” is my interpretation of the ICTV rank of the mega-taxon name. “Centroids” is the number of centroid sequences in each mega-cluster, i.e. its size. Note that sizes decrease rapidly down the table, and known phyla become increasingly fragmented. The claimed new phyla (highlighted) are found in the size range associated with ranks between class (cluster 15) and sub-family (clusters 9 and 11).

```
>WP_185168479.1 RdRp-like protein, partial [Proteus mirabilis]
   A:115-126(16.8)      B:179-192(21.5)      C:222-229(15.4)
   LCCDYSRFDGFL    <52> SGFPLTVVVNSILN  <29> VYGDDNLI  [115]
   .|.|||+||+ |         || |+|||.||| |       ||||||||
   iagDySkFDssl         SGsplTsidNSivN       vyGDDnii   << PSSM logos
     ▲   ▲              ▲                      ▲▲▲
   RdRP_1 E=1.6e-28 (PF00680)


>OLQ09244.1 CMP-sialic acid transporter 5 [Symbiodinium microadriaticum]
   A:235-246(19.9)      B:317-330(18.2)      C:349-356(14.9)
   ISLDFKKWCQHF    <70> GGEGIDQKLWTLAT  <18> GQGDNQVI  [122]
   ||.|+.|||||+         |+||| ||+||| |       +|||||||
   islDasKWNnnq         glqGllqktSSllh       vqSDDqvi   << PSSM logos
     ▲   X              ▲                      XX
   Mononeg_RNA_pol E=2.5e-15 (PF00946)
```

Figure 1. Examples of endogenous host insertions. The figure shows PSSM alignments of the A, B and C motifs in the RdRp palm domain by palmscan Babaian and Edgar (2021). WP_185168479.1 is an RdRp-like protein from *Proteus mirabilis*. OLQ09244.1 is an acid transporter protein from *Symbiodinium microadriaticum*. Both sequences have hits with low E-values to PFAM RdRp models. In WP_185168479.1, the six essential catalytic residues D-[DN]-G-[GS]DD are present (gray triangles), while in OLQ09244.1 three of the six have substitutions (black Xs). These examples show that strong sequence similarity to viral RdRp is not sufficient to establish viral origin or polymerase function, and that substitutions in the catalytic residues are suggestive of an endogenous host insertion with diverged function. I suggest that the novel phyla claimed by Tara are likely to be host insertions similar to those shown here.

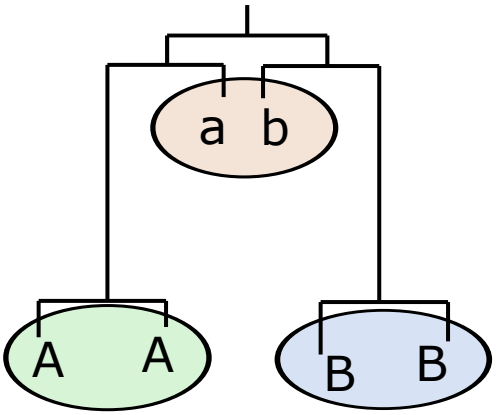

Figure 2. Generic clustering cannot reconstruct phylogeny. When evolutionary rates are highly variable, as is the case for RNA viruses, generic clustering algorithms combine nearest neighbors rather than evolutionary neighbors, as shown in this example with two clades *A* and *B*. Slow-evolving members *a* and *b* are more similar to each other than to fast-evolving members of their own clades, and are therefore placed together by a naive clustering algorithm.

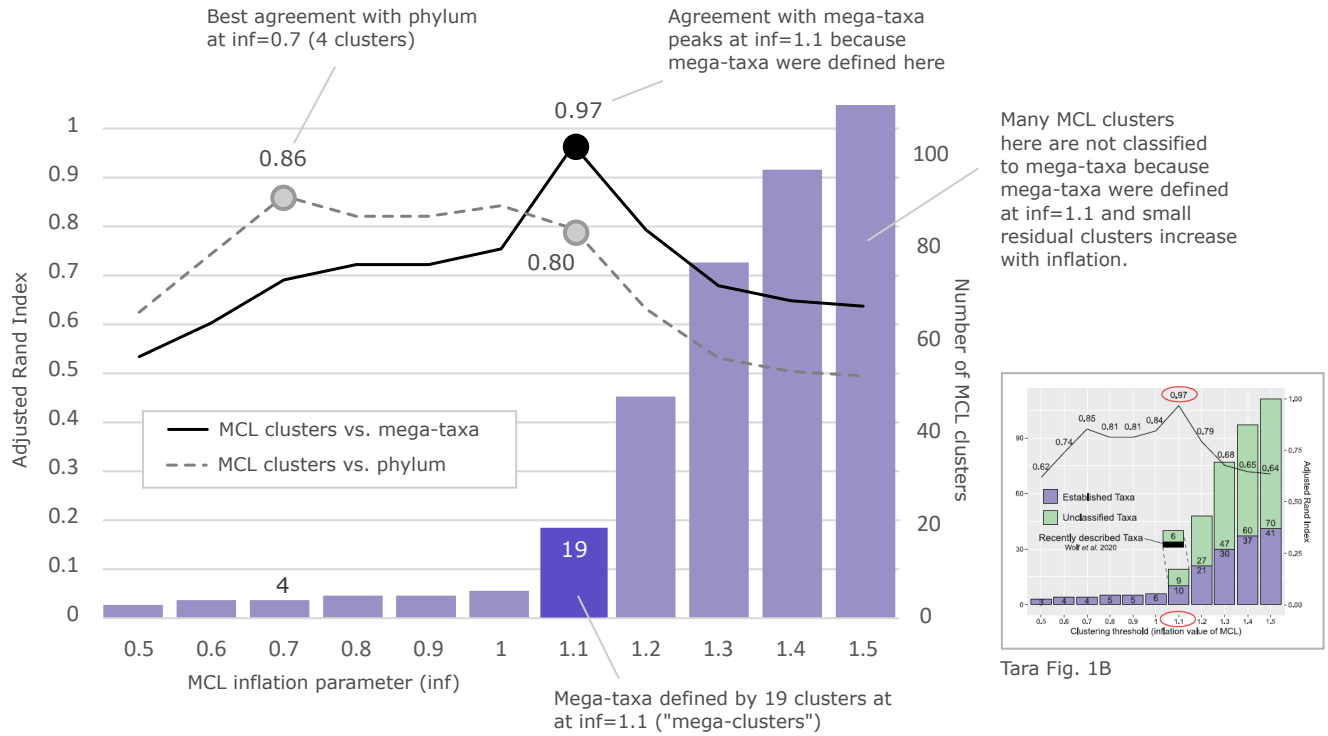

Figure 3. My reconstruction of Tara Fig. 1B. In my version, accessions and mega-taxon names were assigned to to MCL clusters using Tara Tables S6 and S7; phylum annotations were downloaded from GenBank. The histogram shows the number of MCL clusters obtained at inflation (inf) values from 0.5 to 1.5. The solid line shows the adjusted Rand index (ARI) of MCL clusters at different inf values compared to mega-taxa. Mega-clusters are the 19 MCL clusters generated at inf=1.1. The mega-taxon of a mega-cluster is the majority ICTV taxon (or taxa) found in the mega-cluster (see my Table 2), thus there are 19 mega-taxa corresponding 1:1 to MCL clusters at inf=1.1. The agreement between MCL clusters and mega-taxa trivially peaks at inf=1.1 because this is where mega-taxa were defined. The dotted line (added in my version) shows the ARI of MCL clusters compared to GenBank annotations of phyla. The best agreement between MCL clusters and GenBank phylum is at inf=0.7 with four clusters giving ANI=0.86, while inf=1.1 gives ANI=0.80 for the 19 mega-clusters vs. phylum, raising the question of why Tara chose to base their system on clusters at inf=1.1. In Tara's version, histogram bars are stacked to show the number of MCL clusters which are classified to mega-taxa (lower, purple) and the number of MCL clusters which are not classified to mega-taxa (upper, green). The number of unclassified clusters (though not the total number of sequences in them) increases with inflation due to an increasing number of small residual clusters in the tail of the size distribution. The clusters in the top stacked bar (unclassified MCL clusters) represent a much smaller number of centroid sequences compared to the lower stacked bar (classified MCL clusters). The number of unclassified centroid sequences is fixed by the mega-taxa assignments at inf=1.1 and therefore does not change with inflation. The apparent emergence of supposedly novel clusters with increasing inflation has no obvious relevance to the identification of novel phyla, which is solely based on the clusters at inf=1.1.