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. When reads are mapped to species-like clusters of contigs, the most abundant claimed phylum "Taraviricota" accounts for only 3.8% of the reads and the second-most abundant "Artiviricota" for 0.36% of reads. Viral origin of novel contigs is not definitively established. Claims of high accuracy in taxonomy classification are shown to be overstated.

Introduction

In 2017, fewer than 5,000 RNA virus species were known Wolf et al. (2018). Since then, high-throughput metagenomics has dramatically expanded the known RNA virome with studies such as Yangshan Deep-Water Harbour Wolf et al. (2020), Serratus Edgar et al. (2022) and Tara Oceans Zayed et al. (2022) collectively adding more than 140,000 new species. These new viruses are classified primarily on the basis of their RNA-dependent RNA polymerase (RdRp) gene sequences, with clusters at 90% amino acid identity (viral Operational Taxonomic Units, vOTUs) serving as proxies for species.

Claims made by the Tara Ocean study

The Tara Ocean authors claimed to identify five novel phyla. According to their Abstract, "'[s]pecies'-rank abundance determination revealed that viruses of the new phyla 'Taraviricota' ... and 'Arctiviricota' are widespread and dominant in the oceans". These "phyla" (quotes distinguish proposed taxa from ICTV nomenclature) were identified using a new, sequence-based classification system which "nearly completely recapitulated the previously established phylogeny-based ICTV-accepted taxonomy at the phylum and class ranks (97% agreement)". Here, I investigate these claims by re-analyzing their data.

RNA virus taxonomy

RNA virus taxonomy is defined by the International Committee on Taxonomy of Viruses (ICTV). Prior to 2017, the highest taxonomic rank was order. Subsequently, ICTV introduced class and phylum ranks into their system on the basis of an estimated phyologenetic tree from a controversial paper Wolf et al. (2018), which I call "Wolf2018" here. Five phyla were defined corresponding to major branches in this tree (Fig. 1), which was constructed from RdRp sequences for 4,617 RNA virus species. 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), and also that the tree is not reproducible and probably substantially incorrect Edgar (2021). Further, *Duplornaviricota* is paraphyletic with *Negarnaviricota* according to Wolf2018 (Fig. 1), violating 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 Ocean authors who incorrectly state that "*Duplornaviricota* and *Negarnaviricota* were strongly monophyletic [in the Wolf2018 tree]".

Identifying new phyla

The adoption of phyla based on an estimated RdRp tree raises the question of whether novel phyla can be identified in an idealized scenario where an updated tree is available and assumed to be correct, or at least good enough for this purpose. This problem is illustrated in Fig. 2. A subtree (call it x) is a candidate to be a new phylum if it is monophyletic and falls outside of existing phyla. However, it is also reasonable to extend the closest existing phylum to include x. Absent defining traits for phylum assignment other than estimated trees, both options are valid and the choice is ambiguous.

Tara "megataxonomy"

Tara's classification method is sketched in my Fig. 3. Translated RdRp sequences from Tara contigs were combined with RdRps from Yangshan Deep-Water Harbour and GenBank. The combined RdRps were reduced to 13,109 non-redundant "centroids" at 50% identity by UCLUST Edgar (2010). The centroids were then clustered by MCL Enright et al. (2002) from a matrix of pair-wise blastp bit scores, giving 19 "megaclusters". Each megacluster was assigned a "megataxon" based on the majority taxon or taxa according to annotations of its GenBank sequences.

Tara vOTU analysis

Tara's vOTU workflow is summarized in my Fig. 4. Raw data comprises sequencing reads for 1,703 RNA-seq runs (BioProject PRJEB402). These were assembled, and 44,799 RdRp-containing (RdRp+) contigs were identified by hidden Markov model (HMM) search. Translated RdRp sequences were

clustered at 90% amino acid (aa) identity, yielding 5,504 vOTUs. Reads were mapped to RdRp+ contigs using bowtie2 Langmead and Salzberg (2012). vOTU and megataxon abundances per sample were calculated from the bowtie2 mapping.

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. 5 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 viral RdRp HMMs from PFAM Bateman et al. (2004), and have high similarity to known A, B and C motifs te Velthuis (2014) in the RdRp palm domain. Three of the six essential catalytic residues are substituted in the transporter protein. Thus, a low E-value hit to a viral RdRp HMM is not sufficient to establish viral origin or polymerase function, and substitutions in catalytic residues are suggestive of an endogenous host insertion with diverged function.

RdRp sequences are highly diverged

RNA viruses evolve rapidly Peck and Lauring (2018), as reflected in the the Wolf2018 tree where the distance between major branches is ~ 6 amino acid substitutions per site. Protein sequence alignment becomes ambiguous in the so-called twilight zone at $\sim 35\%$ aa identity Rost (1999), corresponding to ~ 0.65 substitutions per site, while RdRps from different phyla are separated by $\sim 10\times$ greater sequence divergence. Accurate alignment and credible tree estimation for diverse RdRp sequences is thus somewhere between very challenging Wolf et al. (2018), and impossible Holmes and Duchêne (2019).

Clade identification by distance-based clustering

Tara's MCL clustering is based on a matrix of BLASTP bit-scores. To estimate phylogeny (as opposed to heuristic groupings by similarity), evolutionary distances are required. Here, "evolutionary distance" has a precise technical meaning: the length of the shortest path between two leaves in the tree must be additive, i.e. proportional to the pair-wise distance between the sequences at those leaves. Then it is possible to work backwards from a pair-wise distance matrix to reconstruct the tree, providing that the distance estimates are sufficiently accurate Durbin et al. (1998). An additive distance is obtained by the number of mutations along a branch, which can be approximated by substitutions alone which is a reasonable when the typical number of mutations per site is $\ll 1$. At larger distances, other muta-

tions (insertions, deletions, permutations, short duplications etc.) may be more prevalent, and multiple substitutions per site cannot be distinguished from single substitutions, so the estimated number of mutations (branch lengths) by any method necessarily becomes increasingly inaccurate. At $\gg 1$ mutations per site, as implied by the Wolf2018 tree, distance estimates are highly uncertain. 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 because nearest neighbors are not necessarily evolutionary neighbors, as shown by the example in Fig. 6. 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 clustering BLASTP bit-scores using MCL cannot credibly reconstruct monophyletic subtrees (i.e., evolutionary clades as opposed to nearest-neighbor groups) because BLASTP bit-scores are not evolutionary distances Koski and Golding (2001), and MCL is a generic clustering algorithm which is entirely ignorant of trees, evolution and phylogeny.

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.

New "phyla" do not dominate Tara contigs

Tara's megataxon assignments for RdRp+ contigs are provided in their supplementary tables S6 and S7. The number of contigs per megataxon is reproduced in my Table 1, showing that the claimed novel "phyla" account for only a small fraction of contigs, in apparent contradiction of the main claim of the paper that they "dominate the ocean". The claim states that dominance was determined from "species-rank abundance". Abundances were calculated by CoverM (https://github.com/wwood/CoverM), though this data is not provided to the best of my knowledge. CoverM is designed to calculate read depth and thereby estimate species abundances in DNA sequencing, where read depth of non-repetitive genomic sequence provides a reasonable estimate of relative abundance of cells. Applied to RNA-seq data, the interpretation is unclear because a single infected host may generate arbitrarily large numbers

of RdRp+ transcripts. Even so, it would be surprising if a small minority of contigs accounted for a large majority of reads across hundreds of independent samples. To investigate, I performed my own read mapping analysis.

Methods

Read mapping using diamond

I mapped all Tara RNA-seq reads to vOTUs with the diamond Buchfink et al. (2015) alignment module of Serratus Edgar et al. (2022) with default settings. The mapping reference comprised 5,426 RdRp aa vOTU representative sequences. Alignments were post-processed to discard hits below 90% aa identity. vOTUs were assigned to "megataxa" by Tara supplementary table S6.

Geolocation mapped using bowtie2

I used bowtie2 to map reads from a selected geolocation (lat. 72.5N, long. 44.1W) where Tara's Fig. 4 shows a very high abundance of "Taraviricota" (my Fig. 7). Samples at this location are SAMEA4397209, SAMEA4397168 and SAMEA4397112; the corresponding sequencing runs are ERR3586697, ERR3586749, ERR3586696, ERR3586718, ERR3586695 and ERR3586748. The mapping reference comprised all contigs assigned to vOTUs, i.e. those inferred to have complete or near-complete RdRp sequences. The bowtie2 command-line options were taken from the Tara supplementary material. The bowtie2 SAM files were filtered using a custom python script (tara_bowtie2_mega_counts.py) which discards alignments if any of the following conditions are true: $> 25\%$ of the read was soft-clipped, the alignment had $< 90\%$ nt identity, or the aligned sequence was $> 50\%$ poly-A or poly-T.

Reconstruction of Tara Fig. 1B

To confirm my understanding of megaclusters and "megataxa", I implemented a python script (adjusted_rand_index.py) to reproduce Tara Fig. 1B.

Results

New "phyla" do not dominate reads mapped to vOTUs

The "Reads" column of my Table 1 shows the total number of reads mapped by diamond to each "megataxon" at minimum 90% aa identity. The number of reads per "megataxon" correlates well with the number of contigs and number of vOTUs, as would normally be expected. The fractions of reads mapped to "Taraviricota" and "Arctiviricota" are 3.7% and 1.3% respectively.

“Taravicicota” has very low abundance at the selected geolocation

Results are shown in my Fig. 7. 173 / 838,302 (0.02%) RdRp+ reads were mapped to “Taraviricota” vOTUs by diamond and 177 reads by bowtie2. The good agreement between the mapping by diamond (translated alignment) and bowtie2 (nucleotide alignment) indicates that my bowtie2 SAM filtering script works well (no filtering is required for diamond because it reports an E-value while bowtie2 provides only MAPQ which assumes that reads are derived from the reference genome). Thus, I find “Taraviricota” to have very low abundance, in stark contrast to the high abundance shown in Tara Fig. 4. Before filtering, bowtie2 mapped 1,700,272 reads from this geolocation to “Taraviricota” vOTUs, primarily due to poly-As. This observation suggests that the discrepancy between my results and Tara’s could be explained if their filtering of low-complexity alignments was not sufficiently stringent.

My reconstruction of Tara Fig. 1B

Results are shown in my Fig. 8. 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 “megataxa”. Megaclusters are the 19 MCL clusters generated at inf=1.1. The “megataxon” of a megacluster is the majority ICTV taxon, or taxa, found in the megacluster (see my Table 1), thus there are 19 “megataxa” corresponding 1:1 to MCL clusters at inf=1.1. The agreement between MCL clusters and “megataxa” trivially peaks at inf=1.1 because this is where “megataxa” 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 ARI=0.80 for the 19 mega-clusters vs. phylum. In Tara’s version, histogram bars are stacked to show the number of MCL clusters which are classified to “megataxa” (lower, purple) and the number of MCL clusters which are not classified to “megataxa” (upper, green); the relevance of this analysis at inflation values other than 1.1 is unclear. 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 (similar to megaclusters 10 though 19 in my Table 1). 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 “megataxa” construction at inf=1.1 (my Fig. 3) and therefore does not change with inflation.

Claim of 97% accuracy for taxonomy assignment

Tara claimed that their classifier “nearly completely recapitulated the previously established phylogeny-based ICTV-accepted taxonomy at the phylum and class ranks (97% agreement)” on the basis of the

ARI=0.97 value reported in their Fig. 1B. However, the agreement underlying the claim is not the fraction of correct classifications of phylum and class with GenBank taxonomy annotations as implied by the statement; rather, it is the adjusted Rand index obtained by comparing "megataxa" and megaclusters. "Megataxa" do not correspond to phylum and class separately or to phylum and class together; rather they are a heterogeneous collection of phyla, sub-phyla, polyphyletic groups (e.g. "Lenarviricota, others"), classes, unassigned ranks (e.g. "Wei-like"), and sub-families. The method is summarized in my Fig. 3. After choosing an inflation value, a combination of one, two or more ICTV taxon names was assigned to each cluster by manually determining a good match. GenBank sequences which did not match were discarded (for example, AAS86764 and CAA76872 are classified as *Pisuviricota* by GenBank but excluded from the megaclusters), and the process was iterated. High agreement between megaclusters and "megataxa" is guaranteed by construction, because there are no prior constraints on which subset of taxa can be assigned to a cluster, and mismatches are discarded. This can be interpreted as extreme over-training of an ad-hoc classifier on all available training data. Measuring agreement between megaclusters and "megataxa" is therefore not predictive of accuracy when classifying new sequences. As explained in my Introduction, a generic clustering method is not appropriate for phylogenetic analysis, and it is therefore clear *a priori* that this approach will have low classification accuracy on sequences outside its training set. A more realistic measurement of classifier accuracy could have been obtained by hold-out validation.

Discussion

Discrepancy in abundance calculation

My Fig. 7 shows a huge discrepancy in the abundances of "Taraviricota" and "Arctiviricota" between Tara's analysis and my analysis at a selected geolocation. This discrepancy can be resolved by comparing bowtie2 alignments before and after filtering. The total size of the relevant FASTQ files is 175Gb uncompressed, so the end-to-end analysis from read mapping through abundance calculation can be reproduced in a few hours. My analysis scripts and the bowtie2 SAM mapping files for this geolocation are included in the data deposited at https://zenodo.org/record/7194888. If the Tara authors make their corresponding data available, it should be straightforward to resolve the discrepancy.

Viral origin of RdRp-like sequences

As noted in the Introduction, RdRp sequences can be inserted into host genomes and low E-value hits to viral RdRp HMMs therefore do not suffice to establish viral origin (Fig. 5). Substitutions in catalytic residues are indicative of host insertions, and 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 based on the number of viral RdRp HMM hits to non-viral GenBank proteins, 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 hypothesis should be that the highly diverged sequences of claimed "phyla" could be endogenous insertions rather than viral RdRps.

RNA virus phylum and class taxa are *de facto* undefined

It seems self-evident that taxonomy should be based on objectively observable traits, as required by 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." However, ICTV did not provide criteria enabling a biologist to classify a virus or its RdRp sequence to an existing phylum or class, or to determine whether a virus or its RdRp sequence belongs to a new phylum or a new class. In the current ICTV system, phylum and class taxa are defined entirely by reference to the Wolf2018 tree, with no provision for reproducing the tree independently or extending it to accommodate new sequences. In my view, this was a fundamental mistake. While the principles of taxonomy are endlessly controversial, it is surely necessary to provide objective criteria enabling definitive classification of an organism either inside one existing taxon or outside all existing taxa at any given rank. The Tara authors' proposal to add many new high-ranking taxa would compound this problem by adding new phyla and classes which are essentially undefined by common-sense standards.

Sister groups are not necessarily new phyla

High bootstraps reported by Tara Oceans for their phylum-level tree are likely to be due to systematic errors arising from its progressive construction from consensus sequences of megaclusters. This construction is strongly biased towards making intra-megacluster alignments more similar than inter-megacluster alignments, creating divisions which will then be reflected in the tree. The same bias 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 Edgar (2021). Disregarding these concerns for the sake of argument, Fig. 9 shows a subtree of Tara Ocean's global tree. This subtree shows (1) "Paraxenoviricota" and "Wamoviricota" are monophyletic, (2) they are distinct clades, and (3) they are outside *Kitrinoviricota*. However, these observations are not sufficient to show that they are distinct phyla. Alternatively, "Paraxenoviricota" and "Wamoviricota" could be defined as sub-clades of

one new phylum, or *Kitrinoviricota* could be extended to cover the entire subtree. See also my Fig. 2.

Conclusions

The Tara authors' own analysis reports that their claimed novel "phyla" comprise only a small fraction of contigs and vOTUs. My results show that they also comprise a small fraction of the RNA-seq reads, contradicting the claim that they "dominate the ocean". The novel sequences are clearly homologs of viral RdRp, but viral origin is not established as they are also consistent with endogenous host insertions. The "megataxon" classification system proposed by Tara is fundamentally flawed and does not provide a credible method for assigning taxonomy to new sequences. Even if the evidence provided by Tara is accepted, it is not sufficient to establish new phyla as they can equally well be accommodated by extending existing phyla.

Code and data availability

Code and data are deposited at https://github.com/rcedgar/tara_oceans and https://zenodo.org/record/7194888.

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.

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using diamond. *Nature methods*, 12(1):59–60.

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.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357–359.

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.

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein engineering*, 12(2):85–94.

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.

te Velthuis, A. J. (2014). Common and unique features of viral rna-dependent polymerases. *Cellular and molecular life sciences*, 71(22):4403–4420.

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.

Table and figures

Mega-taxon	*Mega-cluster*	*Rank*	*Centroids*	*Contigs*	*vOTUs*	*Reads*
(Not determined)	(None)	(None)	6871	13436	1375	6.7%
Pisuviricota	1	Phylum	2101	13014	1514	48.1%
Kitrinoviricota	2	Phylum	1655	5543	758	18.3%
Lenarviricota;others	3	Split phylum	1040	7060	1068	11.4%
Lenarviricota;Allassoviricetes	4	Class	371	154	24	2.4%
Negarnaviricota;Polyploviricotina	5	Sub-phylum	269	1133	129	3.1%
Duplornaviricota;Chrymotiviricetes	6	Class	235	655	110	0.4%
Negarnaviricota;Haploviricotina	7	Sub-phylum	225	528	75	0.3%
Wei-like	8	?	105	1971	385	3.5%
Duplornaviricota;Spinareovirinae	9	Sub-family	52	0	0	0%
Permuto-like	10	?	50	118	22	0.1%
Duplornaviricota;Sedoreovirinae	11	Sub-family	44	237	32	0.1%
New_12at1.1 (Taraviricota)	**12**	**"Phylum"**	**36**	**644**	78	**3.7%**
Kitrino-like	13	?	24	40	19	0.1%
New_14at1.1 (Pomiviricota)	**14**	**"Phylum"**	**9**	**119**	**27**	**0.4%**
Duplornaviricota;Vidaverviricetes	15	Class	8	29	7	<0.1%
New_16at1.1 (Paraxenoviricota)	**16**	**"Phylum"**	**5**	**14**	**10**	**0.2%**
New_17at1.1 (Lenar-like)	17	?	4	24	5	<0.1%
New_18at1.1 (Wamoviricota)	**18**	**"Phylum"**	**3**	**4**	**3**	**<0.1%**
New_19at1.1 (Arctiviricota)	**19**	**"Phylum"**	**2**	**105**	**5**	**1.3%**

Table 1. Summary of megaclusters and megataxa. “Megataxon” is the name assigned by the Tara authors to an MCL cluster at inflation=1.1. “Megacluster” is the MCL cluster number 1,2 ... 19 at inflation=1.1. “Rank” is the rank implied by the megataxon name. “Centroids” is the number of centroid sequences in each mega-cluster, i.e. its size. “Contigs” is the number of contigs assigned to each mega-cluster by Tara Table S6. “vOTUs” is the number of vOTUs assigned to each mega-cluster. “Reads” is the percentage of reads mapped by diamond to a vOTU centroid with minimum 90% aa identity according to my analysis. Claimed novel “phyla” are highlighted.

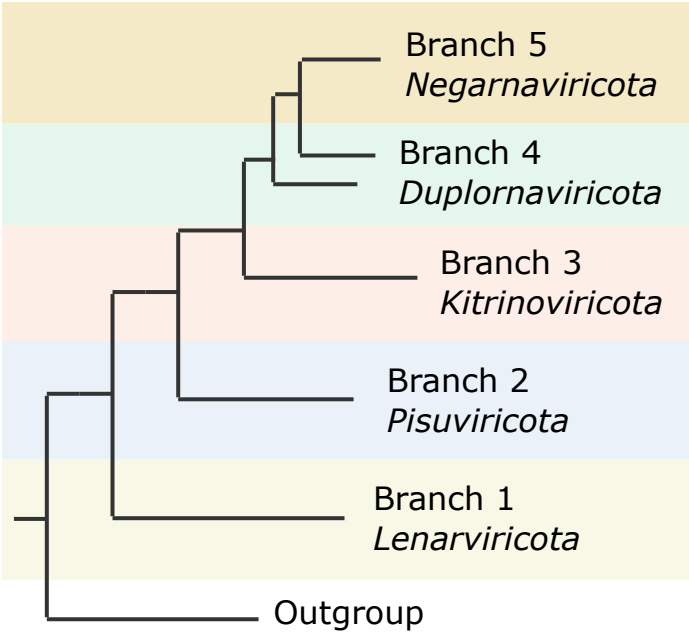

Figure 1. Topology of Wolf2018 tree. Wolf2018 identified five major branches in their tree, which were subsequently adopted as phyla by ICTV. Note that *Duplornaviricota* is not monophyletic. The height of the tree is ~ 3 subs./site, thus the distance between major branches is ~ 6 subs./site.

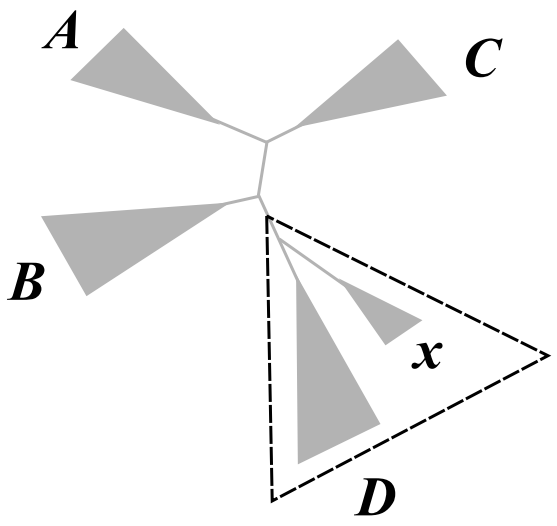

Figure 2. Identifying a new phylum. In this hypothetical example, the tree has four existing phyla *A*, *B*, *C* and *D*. A new group, *x*, is monophyletic and falls outside all existing phyla. Here, there are two choices: *x* is a new phylum, or the most closely related phylum (here, *D*) is extended to include *x*.

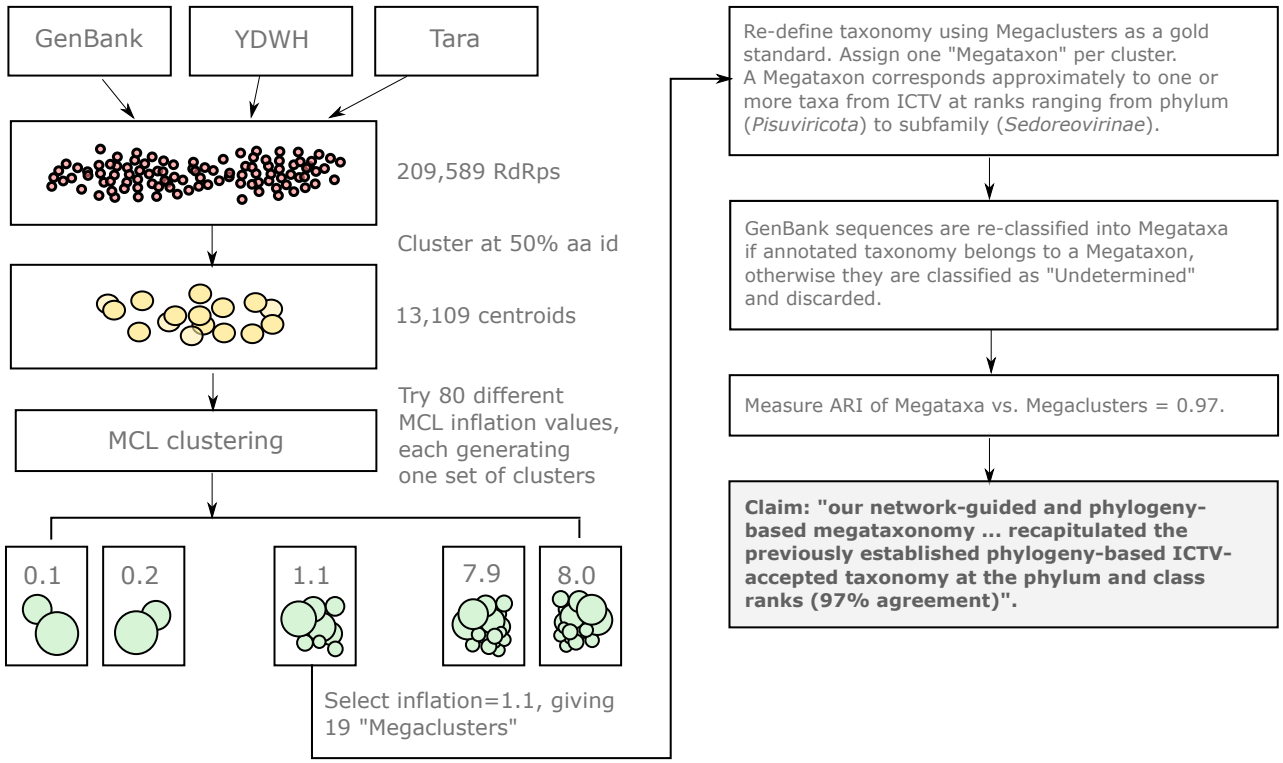

Figure 3. Flowchart of classification by Tara Oceans. RdRp amino acid sequences were collected from GenBank, Yangshan Deep-Water Harbour (YDWH) and Tara contigs. These were clustered twice: first at 50% aa identity, giving 13,109 centroids, then centroids were clustered by MCL giving 19 “megaclusters” at inflation=1.1. For each megacluster, the consensus taxonomy of its GenBank sequences (the “megataxon” of the cluster) was identified. This procedure was iterated. In each round, GenBank sequences with taxonomy/megataxon disagreements were discarded from the next round of MCL clustering so that the final megaclusters had high agreement between taxonomy and megataxonomy. Agreement was measured by adjusted Rand index (ARI).

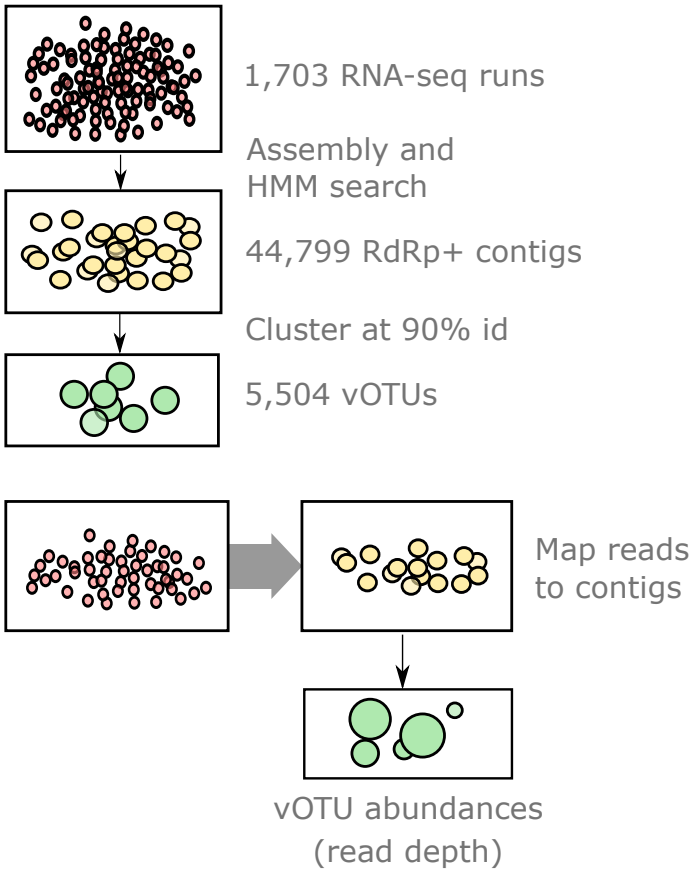

Figure 4. Flowchart of abundance calculation by Tara Oceans. RNA-seq reads were assembled separately for each run. HMMs were used to identify RdRp-containing contigs. Full-length and near-full-length translated RdRp sequences were clustered at 90% amino acid identity to construct species-like vOTUs. Reads were mapped to contigs using bowtie2 to obtain abundances. Contig sequences are available, but otherwise no code or supporting data is provided for this analysis to the best of my knowledge. Abundance is calculated by read depth, which estimates the relative frequency of RNA transcripts.

```
>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
   Mononeg_RNA_pol E=2.5e-15 (PF00946)
```

Figure 5. 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.

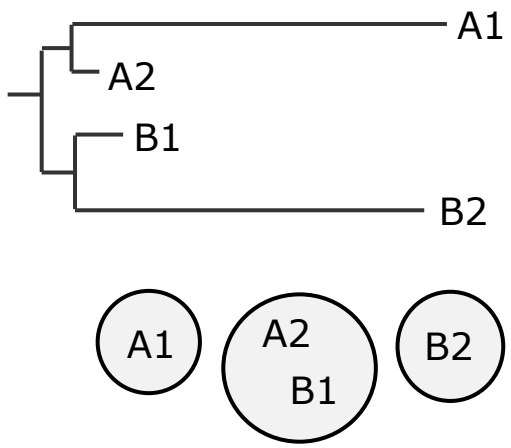

Figure 6. Generic clustering cannot reconstruct phylogeny. When evolutionary rates are highly variable, as is the case for RNA viruses, generic clustering algorithms combine nearest neighbors rather than evolutionary neighbors, as shown in this example.

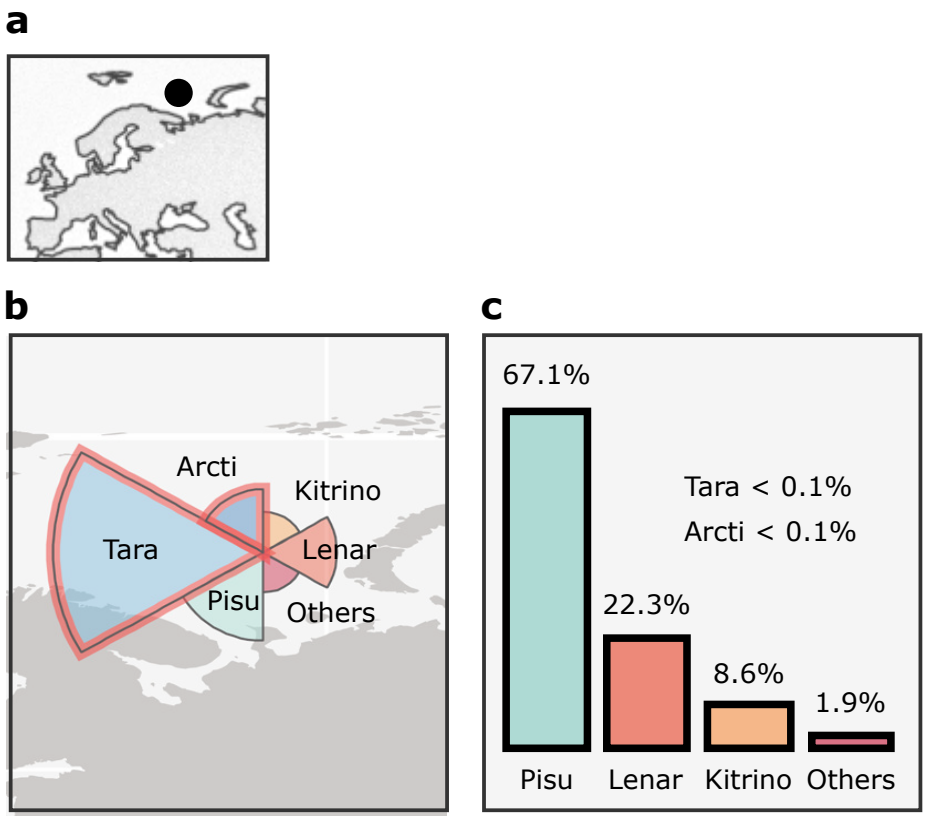

Figure 7. Phylum abundances at a selected geolocation. Analysis of samples from latitude 72.5 degrees North, longitude 44.1 East. **a** Location shown by black dot. **b** Depiction of this location in Tara Fig. 4. **c** Histogram showing the fraction of reads mapped to each phylum according to my analysis (Tara="Taraviricota", Kitrino=*Kitrinoviricota* and so on). Note the extreme discrepancy in relative abundances of "Taraviricota" and "Arctiviricota".

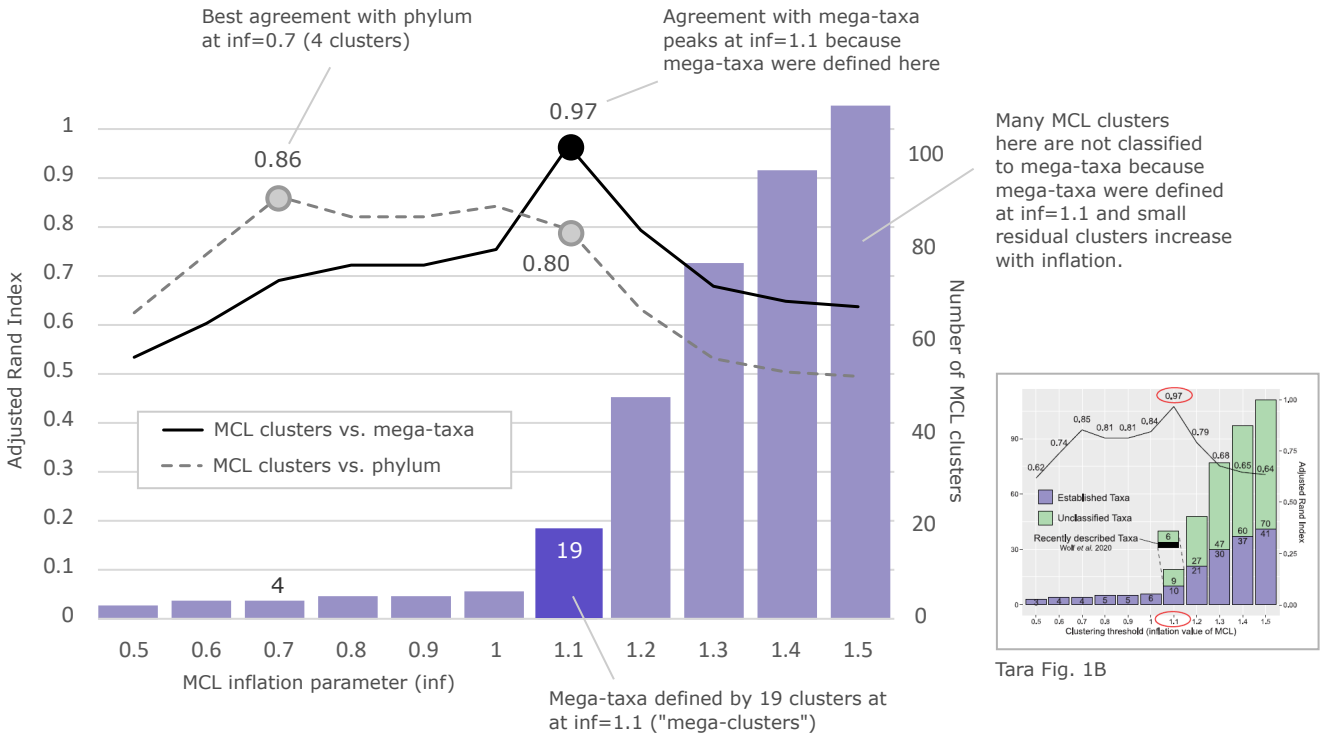

Figure 8. My reconstruction of Tara Fig. 1B. In my version, accessions and mega-taxon names were assigned to to MCL clusters using Tara Tables S6 and S7; phylum annotations were downloaded from GenBank.

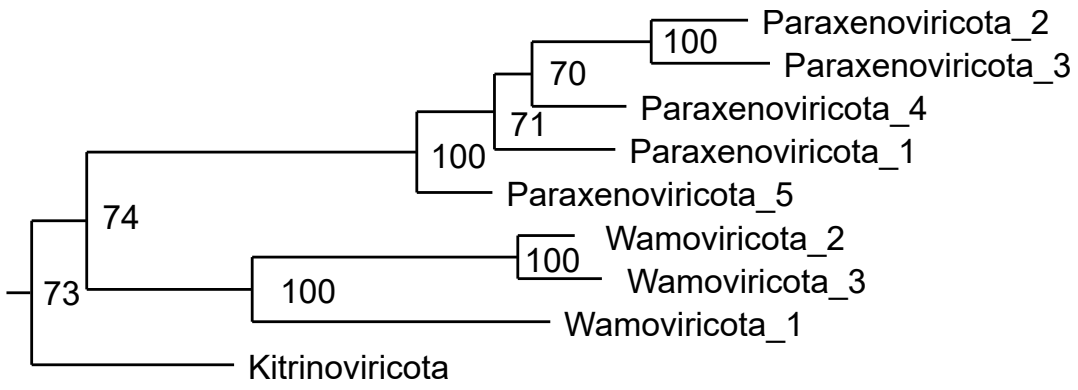

Figure 9. Phylum boundaries are ambiguous. This figure shows a subtree of Tara Oceans' global tree (their Fig. 3A). Assuming this tree is credible and the bootstrap values are informative, then it shows (1) "Paraxenoviricota" and "Wamoviricota" are monophyletic, (2) they are distinct clades, and (3) they are outside *Kitrinoviricota*. However, it does not show that they should be distinct phyla. It is also consistent with defining "Paraxenoviricota" and "Wamoviricota" as sub-clades of a new phylum, or extending *Kitrinoviricota* to include both new groups (see also my Fig. 2).