Alpine lakes harbor diverse, endemic and structurally unique RNA viruses

Lilin Wu[1, 2 #], Yongqin Liu[3, 4, 5 #], Wenqing Shi[1, 2], Tianyi Chang[1, 2], Pengfei Liu[3], Keshao Liu[4, 5], Yong He[6], Zhaorong Li[6], Mang Shi[7], Nianzhi Jiao[1, 2], Xiyang Dong[8] *, Qiang Zheng[1, 2] *

[1] College of Ocean and Earth Sciences and State Key Laboratory of Marine Environmental Science, Xiamen University, Xiamen 361005, China

[2] Fujian Key Laboratory of Marine Carbon Sequestration, Xiamen University, Xiamen 361005, China

[3] Center for the Pan-third Pole Environment, Lanzhou University, Lanzhou, 730000, China

[4] State Key Laboratory of Tibetan Plateau Earth System, Resources and Environment (TPESRE), Institute of Tibetan Plateau Research, Chinese Academy of Sciences, Beijing 100101, China

[5] University of Chinese Academy of Sciences, Beijing 100049, China

[6] Alibaba Cloud Intelligence, Alibaba Group, Hangzhou 310013, China

[7] Centre for Infection and Immunity Study (CIIS), School of Medicine (Shenzhen), Shenzhen Campus of Sun Yat-sen University, Sun Yat-sen University, Shenzhen 518107, China

[8] Key Laboratory of Marine Genetic Resources, Third Institute of Oceanography, Ministry of Natural Resources, Xiamen 361005, China

[#] These authors contributed equally to this work.

* Correspondence can be addressed to Qiang Zheng (zhengqiang@xmu.edu.cn); Xiyang Dong (dongxiyang@tio.org.cn).

Abstract

Alpine lakes, characterized by isolation, low temperatures, oligotrophic conditions, and intense ultraviolet radiation, remain a poorly explored ecosystem for RNA viruses. In this study, we utilized BLASTp, HMM searches, and the deep learning algorithm LucaProt to analyze metatranscriptomes of 20 water samples from Lake Nam Co, a high-altitude alkaline saline lake to uncover RNA viruses. We identified 742 RNA virus species, including 383 novel genus-level groups and 14 unique phylogenetic clades exclusively found in Lake Nam Co, significantly expanding the known diversity within the five phyla of *Orthornavirae*. The newly identified lineage, *f.0167*_Nam_Co, occupied an unknown evolutionary position between *Birnaviridae* and *Kitrinoviricota*, distinguishing from members of *Birnaviridae* and *Kitrinoviricota* by disappearance, rearrangement, and relocation of conserved motif C in RNA-dependent RNA polymerase. RNA viruses in Lake Nam Co exhibited distinct physicochemical properties and RNA secondary structures compared to those found in other ecosystems (*e.g*. estuarine and terrestrial environments). We discovered 13 previously unknown RNA virus families infecting prokaryotes, expanding beyond the currently recognized groups. The presence of genes encoding cytochrome c oxidase and ATP synthase in viral genomes suggested a strategy employed by RNA viruses to exploit host energy production mechanisms to ensure sufficient energy for their replication. Overall, this study sheds light on the diversity and evolutionary dynamics of RNA viruses in extreme plateau aquatic environments.

Introduction

Since the isolation of the first RNA virus from a crab in 1968, the study of RNA viruses has advanced significantly[1]. Early efforts focused on culturing virus particles, while recent researches leverage high-throughput sequencing data to uncover new viruses more efficiently[2]. Over the past two decades, comprehensive omics surveys have led to exponential growth in identifying RNA viruses, particularly through the identification of RNA-dependent RNA polymerases (RdRPs) encoding sequences in specific organisms[3-5] and environmental samples[6-9]. The methodologies for discovering RNA viruses from metatranscriptomes have evolved from traditional sequence similarity searches (*e.g.*, BLAST) and profile alignment-based approaches (*e.g.*, Hidden Markov Model, HMM) to more advanced techniques using deep learning frameworks[10-14].

In 2018, Wolf and colleagues identified five distinct orthornaviran branches[15], which were later recognized as five major phyla within *Orthornavirae*[16]. Recent discoveries of numerous novel groups have led to various proposed classifications of RNA viruses. For instance, an analysis of over 5000 metatranscriptomes has identified vast clades of RNA bacteriophages[17]. Additionally, a deep learning algorithm named LucaProt has revealed a hidden realm of viral diversity[18]. Although these novel groups have not been officially proposed, their significance in RNA virus retrieval and identification is undeniable. Nevertheless, it has been estimated that currently discovered orthornaviran RNA viruses only represent a small fraction of the total diversity, with many yet to be discovered[6, 19].

The abundance and diversity of RNA viruses vary greatly across different habitats, with their distribution closely related to environmental differences and specific hosts[14, 20]. Recent studies have explored RNA viruses in various ecological niches, such as hot springs, significantly expanding our understanding of RNA viral diversity[20, 21]. Freshwater ecosystems exhibit richer alpha diversity and higher RNA virus

abundance compared to other habitats[14], with cold conditions being more beneficial for viral survival[22]. Given this information, lakes in cold regions have the potential to harbor undiscovered RNA virus diversity. High-altitude lakes, in particular, represent a relatively extreme and isolated ecological niche, offering the potential to uncover novel clades of RNA viruses with distinct host specificities.

The Tibetan Plateau, often referred to as the "Third Pole" due to its extensive coverage of glaciers and high-altitude lakes beyond the polar regions[23], boasts over 1,400 lakes larger than 1.0 km^2 [24]. This region, characterized by unique topography, serves as a crucial water tower supplying numerous streams and rivers[25, 26]. Studies indicated that bacteria inhabiting the Tibetan Plateau have evolved the capacity to generate a diverse array of secondary metabolites with distinctive structural and functional characteristics, adapting to the region's harsh environmental conditions[26]. These prokaryotes are known to be infected by diverse DNA viruses, while no RNA virome investigations have been carried out in plateau environments[27]. Lake Nam Co, situated at an elevation of 4,718 meters, spans a surface area of around 2,000 square kilometers, making it the second-largest lake on the Tibet Plateau[28]. A study conducted on Lake Nam Co revealed complex patterns of microbial diversity and abundance across different water layers with distinct temperature and salinity stratifications[29], indicating the possibility of discovering new RNA viruses that infect a diverse and unique range of hosts.

In this study, we utilized traditional similarity-based methods—BLAST and HMM—along with the deep learning approach LucaProt, in conjunction with structure-based protein clustering analysis, to examine RdRP sequences from Lake Nam Co. By analyzing metatranscriptomes in this distinct and isolated high-altitude plateau lake, we aimed to investigate the diversity and key characteristics of RNA viruses in this habitat compared to other environments. Through a comprehensive examination encompassing the ICTV classification and recent taxonomic advancements, we discovered a significant array of novel and endemic RNA viruses,

substantially broadening our understanding of RNA viruses associated with microbial communities predominantly comprising bacteria in Lake Nam Co. Additionally, we identified multiple protein and genome characteristics that likely contribute to the adaptation of RNA viruses to extreme high-altitude environments.

Results and Discussion

Diverse microbial life in Lake Nam Co's cold, alkaline, and oligotrophic waters

Environmental analysis of the 20 water samples from four distinct depths (surface, upper-middle, lower-middle, and bottom) in Lake Nam Co reveals its high-altitude, cold, alkaline, moderately saline, and oligotrophic conditions (**Figure 1 and Tables S1, S2**)[23, 29]. The lake's significant exposure to intense sunlight is marked by a peak photosynthetically active radiation of 1,379 W m^{-2}. Temperatures at the surface drop significantly with depth, from 12.9°C to 3.6°C even during summer, while the pH levels remained stable between 9.3 and 9.4. Salinity showed minor fluctuations, ranging from 1.05 to 1.08‰. Dissolved oxygen levels increased with depth, from 11.1 to 13.6 mg L^{-1}. Dissolved organic carbon concentrations were narrowly constrained between 3.42 to 3.45 mg L^{-1}. Nutrient concentrations, including nitrate, phosphates, and ammonium, were below 1 μM, and generally increased from the surface to the bottom, with the lower-middle layer showing the highest concentrations (**Table S2**).

Based on sequencing of 16S rRNA and 18S rRNA transcripts in Lake Nam Co, the active microbial landscape reveals a significant bacterial and eukaryotic diversity (**Figures S1–S4**). At the class level, *Cyanobacteriia* dominantly comprises 28.0% to 68.7% of the bacterial community. Other notable contributors to the bacterial diversity include *Gammaproteobacteria*, *Bacteroidia*, *Alphaproteobacteria*, *Planctomycetes*, *Verrucomicrobiales*, *Actinobacteria*, *Bdellovibrionota*, *Desulfuromonadia*, and *Acidimicrobiia*. For the eukaryotic community, diversity spans several groups with *Spirotrichea* (ranging from 2.8% to 59.5%) and *Dinophyceae* (17.5% to 55.5%)

among the most prominent. Other key eukaryotic classes include *Bacillariophyta*, *Cryptophyceae*, *Arthropoda*, *Heterotrichea*, *Choanoflagellatea*, *MAST-2*, *Trebouxiophyceae*, *CONThreeP*, *Chrysophyceae*, *Rotifera*, *Zygnemophyceae*, *Litostomatea*, *Chytridiomycota*, and *Filosa-Sarcomonadea*.

Discovery of 14 novel RNA viral clades

Using BLAST and HMM, along with the deep learning algorithm LucaProt, followed by a series of validation processes (**Figure 1**), we identified 742 non-redundant complete RdRP sequences from 20 metatranscriptomes of Lake Nam Co. Among these, 724 RdRPs, which were subsequently used for phylogenetic analysis, were identified through both sequence-alignment-based and protein structure-based clustering analyses (**Figure 2A**), with the remaining 18 detected solely via structure-based clustering (**Figure 2B**). For the former 724 RdRPs, 334 RdRP sequences were assigned to families currently defined by ICTV, while another 390 belonged to families identified in the RVMT database[17] but not officially proposed (**Figures S5-S20 and Table S3**). Of the 18 RdRPs identified exclusively by structure-based clustering, 15 were identified to current phyla within *Orthornavirae* without classification into more refined taxonomic levels, while the remaining three could not be assigned to any existing *Orthornavirae* taxonomy (**Figures S21-S31**). From the total RdRP dataset, 602 newly identified RdRPs formed 383 genus-level novel groups at a 75% amino acid identity (AAI) [30], and 118 RdRPs formed 84 family-level novel groups at a 40% AAI[30]. Based on sequence and structural similarities (**Table S3 and Figure 1A**), except for the three previously mentioned unclassified RdRPs, the RdRP dataset comprised viruses from the five established phyla: 248 from *Lenarviricota*, 216 from *Pisuviricota*, 128 from *Kitrinoviricota*, 123 from *Duplornaviricota*, and 21 from *Negarnaviricota*. Additionally, the dataset includes three *Birnaviridae* viruses that could not be assigned to any of the 5 major phyla due to the permutation of three essential motifs forming the RdRP catalytic core[18, 30] (**Figure 1**).

156 Expansions were observed in various phyla through the phylogenetic analysis, and 14
157 novel and Lake Nam Co-specific phylogenetic clades (with ≥ 3 representative
158 sequences) at the sub-family or family level were identified. These clades shared less
159 than 50% AAI with RdRP reference sequences from other environments[9, 14, 17, 31]
160 (**Table S4 and Figures S5-S20**). In *Lenarviricota*, although the majority of newly
161 identified RdRP sequences were found (**Figure 2A I**), only two novel clades
162 (*Lenarviricota*_Clades_I-II) were discovered, corresponding to the families
163 *Narnaviridae* and *f.0388* from the RVMT database[17] (**Table S4 and Figures S5, S8**).
164 For *Pisuviricota*, six newly identified clades were observed: one in *Picornavirales*
165 (*Pisuviricota*_Clade_I) within *Pisoniviricetes* and five in *Durnavirales*
166 (*Pisuviricota*_Clades_II-VI) within *Duplopiviricetes* (**Figure 2A II**) (**Table S4 and**
167 **Figures S9, S12, S13**). In *Kitrinoviricota*, the primary expansion was observed in the
168 order *o.0045* of class *c.0034* (**Figure 2A III**), where 45 sequences formed an
169 independent clade (*Kitrinoviricota*_Clade_I)[17] (**Figure S14**). These sequences
170 diverged from the closest reference sequences in family *f.0176* with similarities
171 ranging from 29.4% to 36.3% (**Table S4**). In *Duplornaviricota*, 76 RdRP sequences,
172 mostly within the *Ghabrivirales* (**Figure 2A IV**), formed five novel clades
173 (*Duplornaviricota*_Clades_I-V) (**Table S4 and Figures S18-S20**). Among them, three
174 clades (*Duplornaviricota*_Clades_III, IV and V) corresponded to families *f.0295*,
175 *f.0286* and *f.0287* in the RVMT database, exhibiting low AAI (<40%) with other
176 sequences and were distantly clustered (**Figures S18 and S20**). In *Negarnaviricota*,
177 the fewest newly identified RdRP sequences were observed, predominantly belonging
178 to *Bunyavirales* within *Ellioviricetes* of *Polyploviricotina* (**Figure 2A V**), with no
179 novel clades discovered.

180 **Novel RdRPs with unique structural characteristics in Lake Nam Co**

181 The polymerase domain of RdRPs consists of three subdomains: palm, fingers, and
182 thumb[32]. To determine the taxonomy of sequence-alignment-based unclassified
183 RdRPs, which could not be recognized through sequence alignment, we predicted the

184 structures of the polymerase domain of all Lake Nam Co RdRPs and their closely
185 related reference sequences, and conducted structural clustering (**Figures 3 and 4**).

186 Through structural clustering and root mean squared deviation values (RMSD)
187 comparison, we assigned the phyla of 15 sequence-alignment-based unidentified
188 RdRPs, seven of which exhibited unique structural features in their core regions
189 (**Figures 3A and S21-S28**). For instance, Nam-Co_*Birnaviridae*_001 had a special
190 β-sheet structure between two typical α-helices in the non-conserved region between
191 motifs A and B (**Figures 3A I, S21**). Nam-Co_*Lenarviricota*_001 featured multiple
192 α-helices and irregular coils with distinctive amino acids in motifs A and B (**Figures**
193 **3A II, S22**). Nam-Co_*Pisuviricota*_004 and Nam-Co_*Pisuviricota*_007 exhibited
194 extra helix and distinctive β-sheet structures, respectively (**Figures 3A III, IV, S24B,**
195 **S25B**). Nam-Co_*Kitrinoviricota*_001 showed a shorter first α-helix between motifs A
196 and B, with an extra α-helix at the β-turn between motifs B and C (**Figure S26**).
197 These unique features highlight the diversity and endemism of RdRPs in Lake Nam
198 Co.

199 Three sequence-alignment-based unidentified RdRPs remain unclassified based on
200 structural clustering. Their RMSD values closely align with reference structures from
201 different phyla, indicating their unique structural features. For
202 Taxa_unknown_RdRP_001, the RMSD values of its structure compared to relatively
203 close reference structures belonging to *Pisuviricota* and *Kitrinoviricota* both exceed 2
204 (**Figures 3B I and S29**). It shared β-turn features with *Kitrinoviricota* reference and
205 similar α-helical structures with the *Pisuviricota* reference, though its helical structure
206 was longer and divided into three segments. Taxa_unknown_RdRP_002 referenced
207 *Lenarviricota* and *Kitrinoviricota*, with RMSD values of 2.694 and 3.259,
208 respectively. It shared more similarities with *Kitrinoviricota* reference in the
209 composition of motifs A and B, along with the β-turns at the edges of motifs A and C
210 (**Figures 3B II and S30**). Taxa_unknown_RdRP_003 was associated with
211 *Lenarviricota* and *Kitrinoviricota* references, with both RMSD values greater than 4,

212 suggesting a high degree of specificity. It presented a highly unique area comprising
213 two α-helices and two β-folds between motifs A and B (**Figures 3B III and S31**).
214 These three RdRPs, identified exclusively by structure-based clustering with remote
215 sequence and structural similarity, suggest the potential emergence of new phyla.

216 Besides the confirmation of phyla of sequence-alignment-based unidentified RdRPs,
217 structural clustering unveiled a unique RNA virus group in Lake Nam Co. Six RdRP
218 sequences, initially identified as *Kitrinoviricota* based on sequence alignments from
219 Lake Nam Co (referred to as *f.0167*_Nam_Co), did not cluster with other
220 *Kitrinoviricota* sequences. Instead, they formed a distinct cluster situated between
221 *Birnaviridae* and the five orthornaviran phyla according to protein structural
222 clustering analysis (**Figure 4A**). The RdRPs of *Birnaviridae* or related viruses are
223 permutated, meaning that their motifs might shift places (**Figure 4B**)[33]. The low
224 RMSD values (RMSD = 2.16) between *f.0167*_Nam_Co and *Birnaviridae* references,
225 along with the similar rearranged motif sequences (motifs C-A-B) suggest a similarity
226 between *f.0167*_Nam Co and *Birnaviridae* (**Figures 4B and 4C I**). Reference
227 sequences of *f.0167*_Nam_Co, previously classified in the family *f.0167* within the
228 class *c.0032* of *Kitrinoviricota*, were thought to contain the SDD sequence as motif C
229 with the general arrangement[17]. However, structural predictions revealed that these
230 sequences did not possess motif C in the palm region of the polymerase domain
231 (**Figure 4C II**). The distinct arrangement of motifs in *f.0167* between Lake Nam Co
232 and other aquatic environments[17, 34] reveals the evolutionary process of motif C
233 positions, reflecting reconstitution and loss in the conserved core region.

234 **Genome and protein features of RNA viruses adapting to alpine environments**

235 A total of 1,159 RNA viral contigs, representing viral genomes, were identified based
236 on the presence of genes encoding RdRPs. Among these, 166 contigs contained more
237 than three open reading frames (ORFs), with genome lengths ranging from 1,371 to
238 16,715 bases. Besides genes encoding RdRPs and capsid proteins, the majority of

these RNA virus genomes contained predominantly unannotated genes (**Figure S32**). Other annotated viral proteins encoded by RNA viruses in Lake Nam Co included movement proteins (MP), glycoproteins, helper-component proteinases (HCP), replicases, methyltransferases (MT), packaging NTPase P4, and peptidases S39 and A21 (**Figures 5A, S33A**).

Six of viral genomes contained auxiliary metabolic genes (AMGs), impacting various metabolic processes in hosts, including energy generation (*atpA*, *cyoA*), translation (*rplA*), phospholipid metabolism (*ugpQ*), and heme biosynthesis (*hemX*, *hemY*). (**Figure 5A and Table S8**). Among them, a virus belonging to the family *f.0102* within *Pisuviricota* carries a gene encoding cytochrome c oxidase subunit II in *Achlya* (**Figure 5A, Table S8**). An RNA virus identified as Nam-Co_family_51 within *Kitrinoviricota* carries a gene encoding ATP synthase CF1 alpha subunit of an unknown host (**Figure 5A, Table S8**). These two AMGs encode proteins that have been confirmed to function in the host's mitochondria[35] or chloroplasts[36], can affect the energy metabolism processes, which reflects a viral strategy to exploit host energy production. Genes related to other metabolism were also identified, such as AMGs encoding uroporphyrinogen-III C-methyltransferase and heme biosynthesis-associated TPR protein in *Hydrogenobacter*, are possibly associated with multiple processes like electron transfer and oxygen transport[37]. Besides AMGs, two genes associated with anti-prokaryotic immune systems, like APIS022 and AcrVIA7, suggest an ongoing arms race between viruses and their hosts[38] (**Figure 5A and Table S8**).

To illustrate the RNA virus protein characteristics in Lake Nam Co, we compared the physicochemical properties of capsid proteins and RdRPs from RNA viruses in Lake Nam Co, various terrestrial sites (16 provinces and regions in China)[8], and estuarine environments (Yangshan Port)[7], including hydrophobicity, polarity, mutability, transmembrane tendency, and refractivity (**Figure S34**). Capsid proteins from the two aquatic environments showed no obvious difference (wilcox test, $p > 0.05$) but were significantly different from terrestrial capsid proteins (wilcox test, $p < 0.05$).

267 Comparing to capsid proteins, Lake Nam Co RdRPs significantly differed from both
268 estuarine and terrestrial environments across all five indices (wilcox test, $p < 0.001$)
269 (**Figure S34**). For instance, Lake Nam Co RdRPs showed broader, more uniform
270 refractivity values (**Figure 5B)**, potentially linked to intense UV radiation[39, 40]. These
271 results demonstrate the consistency of capsid proteins' physicochemical properties in
272 the aquatic environment and the uniqueness of Lake Nam Co RdRPs.

273 Besides characteristics of given proteins, the genome features of RNA viruses in Lake
274 Nam Co were also shaped by its environment and hosts. We compared the RNA
275 secondary structural features of RNA viruses from Lake Nam Co with RVMT
276 references[17] within the same classes, analyzing hairpins, stems, interior loops,
277 multiloops, unpaired ratios, and minimum free energy (MFE)[41] (**Figures 5C, 5D and**
278 **S35**). We focused on taxonomic groups at the class level with over ten relatively
279 complete RNA genomes, averaging 90 sequences per class, including *Alsuviricetes*,
280 *Amabiliviricetes*, *c.0013*, *c.0031*, *c.0034*, *c.0069*, *Chrymotiviricetes*, *Duplopiviricetes*,
281 *Ellioviricetes*, *Howeltoviricetes*, *Leviviricetes*, *Miaviricetes*, and *Tolucaviricetes*
282 (**Figures S36-41**). Overall, the variation trends of secondary structural features along
283 classes of Lake Nam Co viral RNAs aligned with those of RVMT viral RNAs[17],
284 indicating stable differences across taxa (**Figures 5C and S36-41A**). However, within
285 viral classes, except *Ellioviricetes*, the secondary structural features of Lake Nam Co
286 viral RNAs exhibit distinct density distributions, indicating specificity compared to
287 RVMT references. (**Figures 5D and S36-41B**). For example, compared to RVMT
288 references[17], *Amabiliviricetes* viruses of Lake Nam Co have fewer hairpins (wilcox
289 test, $p < 0.05$) (**Figure S36B**) but more stems (wilcox test, $p < 0.01$) (**Figure S37B**)
290 and interior loops (wilcox test, $p < 0.05$) (**Figure S38B**). In comparison to RVMT
291 references[17], *Chrymotiviricetes* viruses of Lake Nam Co have fewer stems (wilcox test,
292 $p < 0.0001$) (**Figure S37B**), interior loops (wilcox test, $p < 0.01$) (**Figure S38B**),
293 multiloops (wilcox test, $p < 0.05$) (**Figures S39B**) but higher unpaired ratio (wilcox
294 test, $p < 0.001$) (**Figure S40B**) and MFE (wilcox test, $p < 0.0001$) (**Figure S41B**).

295 **Host range of RNA viruses predominantly targeting prokaryotic organisms**

296 We employed various methods to determine the relationship between RNA viruses
297 and their hosts in Lake Nam Co. Eighty-three RNA viruses were identified with direct
298 host evidence. Of these viruses, nine were predicted to infect eukaryotes, 73 to target
299 bacteria, and one to infect archaea (**Figure 6A**). Previous studies have identified only
300 three groups known to infect bacteria: *Leviviricetes*, *Cystoviridae*, and
301 *Picobirnaviridae*[4, 17, 42, 43], however, infection evidences in our study revealed 13
302 additional families in Lake Nam Co (**Table S7**). These 13 families included
303 *Totiviridae*, *Mitoviridae*, *Narnaviridae*, and *Benyviridae* currently defined by ICTV,
304 as well as *f.0028*, *f.0076*, *f.0102*, *f.0111*, *f.0176*, *f.0283*, *f.0286*, *f.0295*, and *f.0296*
305 identified in the RVMT database but not yet officially proposed[17]. Among these
306 families, *Benyviridae* and *f.0176* are within *Kitrinoviricota*, a phylum previously
307 unrecognized for bacterial infection.

308 To predict host range of RNA viruses that lack direct evidence, we predicted the
309 tendency of their infection by calculating and visualizing amino acid usage frequency
310 (AAUF) (**Figure S42A**) and di-/tetra-nucleotide usage frequency (D/TUF)[44, 45]
311 (**Figures 6B, 6C and S42B**). Visualization of AAUF showed that viruses
312 predominantly clustered independently, with a few overlapping with bacteria (**Figure**
313 **S42A**). In contrast, RNA viral D/TUF closely matched their potential hosts,
314 particularly bacteria, indicating a preference for infecting prokaryotes (**Figures 6B,**
315 **S42B**). According to the TUF aggregation of RNA viruses with microbes at the
316 phylum level, they primarily infected bacteria from *Bacteroidota*, *Actinomycetota*,
317 *Chloroflexota*, and *Cyanobacteriota* phyla (**Figure 6C**).

318 We analyzed the relative abundance of prokaryotic RNA viruses across samples to
319 investigate potential host and environmental influences (**Figures S43 and S44**).
320 Prokaryotic viruses were most abundant in upper-middle water layers and decreased
321 in deeper layers **(Figure S43)**. Distribution of individual prokaryotic viruses varied

322 across samples **(Figure S44)**, with the most significant differences observed between
323 sampling sites, followed by water layers. These differences align with the distribution
324 of prokaryotic host groups **(Figures S45 and S1).** The distribution of prokaryotic
325 RNA viruses in Lake Nam Co was also influenced by unique environmental
326 conditions. For instance, prokaryotic viruses were less active in surface waters
327 (**Figure S43**), likely due to high levels of UV radiation on the Tibetan Plateau[46],
328 which compromised virus integrity, especially in simpler prokaryotic hosts[47].
329 Analogously, their abundance was higher at lakeshores than at lake centers (**Figure**
330 **S43**), where suspended particles reduce UV effects[48].

331 Summary

332 Through comprehensive analyses, we elucidated the diversity and adaptive strategies
333 of RNA viruses in Lake Nam Co's unique ecosystem. Our results uncovered numerous
334 novel and endemic groups/clades of RNA viruses, which exhibited significant
335 specificity in their genomes, RNA secondary structures, and protein structures.
336 Contrary to conventional assumptions, our study revealed a significant bias of these
337 viruses towards bacterial hosts, with potential implications for archaeal infection as
338 well. Additionally, the RNA viruses in Lake Nam Co directly impacted host metabolic
339 processes through AMGs. Overall, this study sheds light on the diversity, adaptation,
340 and host interactions of RNA viruses in Lake Nam Co, providing insights into their
341 ecological roles and evolutionary dynamics in extreme alpine environments. However,
342 due to experimental limitations, we did not capture particles smaller than 0.2 μm,
343 which means our results primarily represent RNA viruses actively infecting hosts
344 within cells, missing free viruses dispersed in water. Additionally, the lack of direct
345 evidence from CRISPR-Cas limits our ability to confidently predict virus-host
346 linkages. Although constrained by the limited number of metatranscriptomic samples
347 from Lake Nam Co, our research highlights numerous novel RNA virus groups and
348 characterizes their genomic and proteomic features. Future work should include

349 collecting more metatranscriptomic samples to discover additional novel groups and
350 achieve a more comprehensive understanding of RNA virus-host interactions across
351 diverse environments.

352 # Methods

353 ## Sample collection

354 Water samples were collected from two sampling stations in Lake Nam Co, Tibet,
355 China (**Figure 1 and Table S1**) **-** the LS station near the lakeshore (90.98°E, 30.79°N)
356 and the LC station near the lake center (90.48°E, 30.46°N). At the LS station, 20 liters
357 of surface water (at a water depth of 0.5 m) was collected. Simultaneously, at the LC
358 station, samples were taken from four different layers - surface (0.5 m), upper-middle
359 (20 m), lower-middle (50 m), and bottom (80 m), with 20 liters collected from each
360 layer. All water samples were pre-filtrated using nylon sieves with a pore size of 60
361 μm and then filtered through polycarbonate membranes (Millipore) of different pore
362 sizes within 5 hours after collection. The membranes of 10 μm, 3 μm, 1.2 μm, and 0.2
363 μm were collected, rapidly frozen with liquid nitrogen, and stored at -80 °C until
364 RNA extraction. Additionally, 40 ml of filtrate was transferred to a 50 ml clean tube
365 and stored at -20 °C for inorganic nutrient analysis. For dissolved organic carbon
366 (DOC) analysis, water samples were filtered through 0.7-μm GF/F glass fiber filters
367 (precombusted at 450 °C for 4 hours, Whatman). The 20 ml filtered water samples
368 were collected directly into two 40 ml glass vials (acid-washed, rinsed with Milli-Q
369 water, and pre-combusted, CNW) and immediately stored at -20 °C for DOC
370 concentration measurements.

371 ## Measurement of environmental parameters

372 Solar radiation intensity was measured using Handheld TES-1333 solar power meter
373 digital radiation detector. Water temperature, pH, dissolved oxygen (DO), and salinity
374 were measured using in situ Conductivity-Temperature-Depth (CTD) sensors

375 (Sea-Bird Electronics, Bellevue, WA, USA). Inorganic nutrient concentrations,
376 including phosphate, nitrate, nitrite, and silicate, were measured using the PowerMon
377 Kolorimeter AA3 (Bran + Luebbe, Charlotte, NC, USA) following the
378 spectrophotometric methods[49]. The DOC concentrations were determined using a
379 Shimadzu TOC-VCPH analyzer through high-temperature (680 °C) catalytic
380 oxidation. Ammonium concentrations were measured spectrophotometrically using
381 the indophenol blue method.

382 **16S/18S rRNA transcripts amplicon sequencing and data processing**

383 Total RNA was extracted using the DNA/RNA Miniprep Kit (Zymo Research
384 Corporation) according to the manufacturer's instructions and subsequently stored at
385 -80 °C. One part of this RNA was reverse-transcribed into cDNA using random
386 hexamer primers and SuperScript III Reverse Transcriptase (Invitrogen), as per the
387 provided guidelines. To identify active microbial communities, the V4-V5 regions of
388 prokaryotic 16S rRNA were amplified using forward (515F
389 5'-GTGCCAGCMGCCGCGGTAA-3') and reverse (907R
390 5'-CCGTCAATTCMTTTRAGTTT-3') primers. Similarly, the V4 region of
391 microeukaryotic 18S rRNA was amplified with forward (TAReuk454FWD1
392 5'-CCAGCASCYGCGGTAATTCC-3') and reverse (TAReukREV3
393 5'-ACTTTCGTTCTTGATYRA-3') primers. Sequencing of the 16S/18S rRNA
394 transcripts was conducted on an Illumina NovaSeq 6000 system (pair-end sequencing,
395 2 × 250 bp) by Magigene (Guangdong, China). For the analysis, Qiime2 v2021.8.0
396 was employed to process the 16S/18S rRNA sequencing data from the 20 samples[50].
397 This process included merging, quality filtering, denoising, and the exclusion of
398 sequences originating from chloroplasts and mitochondria. Additionally, the software
399 supported the clustering and annotation of operational taxonomic units (OTUs),
400 enabling the profiling of the taxonomic composition.

401 **Metatranscriptomic sequencing and data processing**

402 After the RNA extraction, a part of total RNA was subjected to rRNA depletion using
403 the Ribo-off® rRNA Depletion Kit V2. This rRNA-depleted RNA was then
404 reverse-transcribed into cDNA using random hexamer primers and SuperScript III
405 Reverse Transcriptase (Invitrogen). Metatranscriptomic sequencing was performed on
406 the Illumina NovaSeq 6000 platform (pair-end sequencing, 2 × 150 bp) by Magigene
407 (Guangdong, China). For the metatranscriptomic sequencing data, Fastp v0.23.2 was
408 utilized for trimming and quality control, retaining only high-quality clean reads[51].
409 These clean reads were then mapped to the SortMeRNA database v4.3.4 [52] for
410 separating rRNA reads from non-rRNA reads using Bowtie2 v2.4.3 [53]. Non-rRNA
411 reads were assembled using MEGAHIT v1.2.9 [54] to generate assembled contigs,
412 employing both single-sample assembly and multiple-sample co-assembly (LC + LS),
413 resulting in a total of 22 assemblies. The getORF package from EMBOSS v6.6.0.0 [55]
414 and Prodigal v2.6.3 [56] were used for open reading frame (ORF) prediction from the
415 assembled contigs (using the 11th codon table), preserving ORFs longer than 200
416 amino acids.

417 **Identification of RdRP sequences**

418 To enhance the detection of RNA viruses within our datasets, we applied a
419 comprehensive approach combining traditional sequence alignment, hidden Markov
420 model (HMM) searches, and a deep learning algorithm (**Figure 1**). Specifically, we
421 utilized: (1) BLASTp analysis (Diamond v2.0.15.15363)[57] to compare our sequences
422 against a custom database comprising RdRP sequences gathered from four recent
423 studies on RNA virus diversity[9, 14, 17, 31]; (2) HMM searches (HMMER v3.3.2)[58]
424 targeting 77 different RdRP family profiles to identify potential matches[59]; (3) deep
425 learning analysis through LucaProt (https://github.com/alibaba/LucaProt)[14], focusing
426 on ORFs longer than 300 amino acids. Through this multifaceted approach, we
427 initially identified 4,021 potential RdRP sequences (**Figure 1**).

428 To refine our results for species-level analysis of RdRP, we applied CD-HIT version
429 4.8.1 to eliminate duplicates from the sequences, using a 90% amino acid identity
430 (AAI) similarity threshold[30, 60, 61], reducing the count to 3,089 non-redundant
431 candidate RdRP sequences. Further analysis for identifying three core RdRP motifs
432 (Motifs A, B, and C)[30] was conducted using PalmScan2[30]. Sequences exhibiting all
433 three motifs were considered viable RdRP candidates, resulting in a total of 809
434 sequences. These candidates underwent additional check against the non-redundant
435 (nr) database using Diamond BLASTp v2.0.15.15363 to ensure their viral origin. We
436 excluded sequences whose top BLAST hits (e-value < 0.001) did not match known
437 RNA virus proteins, thereby eliminating false positives and potential endogenous viral
438 elements[5]. This careful filtering process yielded 742 sequences that were identified as
439 RdRP sequences and selected for further analysis (**Figure 1**).

440 **Phylogenetic tree construction and taxonomic assignment of RdRPs**

441 The taxonomic classification of newly discovered RdRPs was determined through
442 sequence-alignment-based and protein structure-based clustering analyses. For the
443 sequence alignment approach, the determination of RdRP taxonomy was informed by
444 a synthesis of the top BLASTp hits across a range of databases, incorporating the
445 valuable contributions from the RVMT database[17], as well as research conducted by
446 Hou et al. [14], Charon et al. [31], and Zayed et al. [9, 62]. RdRPs with taxonomy references
447 here were considered to be sequence-alignment-based identified RdRPs, while others
448 were identified as sequence-alignment-based unidentified RdRPs.

449 For the phylogenetic analysis of sequence-alignment-based identified RdRPs, due to
450 the rapid evolution of RNA viruses and substantial differences in RdRp sequences
451 among different classes, multiple gaps occurred during alignment, hindering the
452 construction of a robust phylogenetic tree. Therefore, separate alignments were
453 performed for the five phyla of RNA viruses: *Duplornaviricota*, *Kitrinoviricota*,
454 *Pisuviricota*, *Lenarviricota*, and *Negarnaviricota*. Subsequently, maximum likelihood

455 phylogenetic trees for RdRp proteins were constructed. For classes with over 100
456 identified RdRp sequences, sub-alignments were performed based on evolutionary
457 distances and protein clustering numbers, followed by sub-trees construction. Based
458 on different classifications, the identified complete RdRPs was aligned with reference
459 sequences using MAFFT v 7.508[63]. Initial trimming of aligned regions was performed
460 using trimAl v 1.4.rev15[64]. Sequences were manually trimmed to the adjacent area of
461 core RdRP region (including the three conserved motifs A, B, and C) and sequences
462 causing alignment confusion were manually removed using Jalview v 2.11.3.2[65].
463 FastTree v 2.1.10 was used to construct maximum likelihood phylogenetic trees for
464 each phylum and class [66], the trees were then visualized using the Chiplot online
465 website (https://www.chiplot.online/)[67]. New taxonomy groups of Lake Nam Co were
466 then identified according to the amino acid similarity.

467 To conduct the structure verification of RdRP candidates and ascertain the
468 approximate clades of sequence-alignment-based unidentified RdRPs, Alphafold2 v
469 dtk22.04[68] was used to predict the structures of all RdRP candidates in Lake Nam Co
470 and their reference sequences[9, 14, 17, 31, 59]s. We conducted the paired structure
471 alignment based on template modeling score (TM-score) method[69] and the
472 structure-based clustering of all predicted RdRP structures[70]. RdRP candidates with a
473 TM-score over 5 compared to all references were excluded (Establishment of RdRP
474 dataset part). To eliminate the potential impact of non-core region sequences bias (too
475 much amino acid on the left or right) on structural similarity, all
476 sequence-alignment-based identified RdRP sequences were trimmed to the relative
477 conserved domain according to the sequence alignment results, while the
478 sequence-alignment-based unidentified RdRPs remain untrimmed before re-clustering
479 implemented the same way as described above. For a sequence-alignment-based
480 unidentified RdRP, the phylum was determined by the RdRP structure with the lowest
481 TM-score. If the structure with the lowest TM-score was a sequence-alignment-based
482 identified RdRP from Lake Nam Co, an additional reference structure from RdRP

483 identification databases[9, 14, 17, 31, 59] with the lowest TM-score was also included. The
484 phylum of the sequence-alignment-based unidentified RdRP was then determined
485 based on comparisons and visualization results of the region near the three motifs
486 with the reference structures[71, 72]. After integrating the phylum classification of
487 sequence-alignment-based identified RdRPs and sequence-alignment-based
488 unidentified RdRPs, the conservative motifs A, B, and C for different phyla of RNA
489 viruses were visualized using WebLogo v 2.1.10[73].

490 All reference sequences and newly identified RdRPs were subjected to palmprint
491 recognition by PalmScan2[30]. Palmprints were clustered using MMseq2 v 14.7e284 [74]
492 at 75% and 40% identity thresholds to identify clusters at the genus-level and
493 family-level[30] composed solely of newly identified RdRPs from Lake Nam Co.

494 **RNA virus genome annotation and distribution**

495 All contigs containing the identified RdRP of Lake Nam Co were considered as RNA
496 viral contigs. To ensure that the contigs represent indeed viral genomes and not
497 assembly errors, we used metagenomic sequencing reads from the same environment
498 to map to these contigs. Contigs that that had more than 10 reads mapped from DNA
499 sequencing data were excluded[14]. After redundancy reduction of nucleotide sequences
500 with a threshold of 0.95 using CD-HIT v4.8.1[60], these genomes were annotated using
501 tools such as Cenote-Taker v3.0.0[75], eggNOG-mapper v2.1.12[76], and dbCAN3
502 v4.0.0[77], along with annotation using DRAM v1.2.0[78] after preprocessing with
503 Virsorter2 v2.2.4[79]. The dbAPIS database was also utilized for the annotation of
504 Anti-prokaryotic Immune System (APIS) proteins in these RNA viruses[38].

505 The non-rRNA reads of all 20 samples were mapped to the identified genomes using
506 Bowtie2 v 2.4.3[53] before the abundance and diversity of RNA viruses in Lake Nam
507 Co were determined by MetaPop v 1.0.0 [80].

508 **Comparison of physicochemical property indices of RdRPs and capsid proteins**
509 **in RNA viruses from different environments**

510 RdRPs and capsid proteins of Yangshan estuary environment[74] and terrestrial
511 environment across China[75], together with the RdRPs and capsid proteins annotated in
512 Lake Nam Co were collected. The physicochemical properties indices, including
513 hydrophobicity, polarity, relative mutability, transmembrane tendency, and refractivity
514 were calculated refer to ProtScale website (https://web.expasy.org/protscale/)[76]. A
515 linear weighted model (window size of 9) was applied to calculate scores at different
516 positions in the protein sequences for each of the physicochemical properties. The
517 variation of individual protein indicators across positions were depicted by line plots,
518 while the comparison of all physicochemical properties indices between different
519 locations were depicted by violin plots drawing by R v 4.2.0.

520 **Comparison of RNA secondary structures between RNA viruses in Lake Nam Co**
521 **and global data**

522 Classes containing over 10 RNA virus sequences in Lake Nam Co were selected, and
523 100 viral contigs from these classes were randomly extracted from the RVMT
524 database[13] for comparison. RNA secondary structure features (**Figure S35**) of these
525 classes, including hairpins, stems, interior loops, multiloops, unpaired ratios, and
526 minimum free energy (MFE), were analyzed and visualized refer to the script
527 published urreeby Uri Neri's laboratory
528 (https://github.com/UriNeri/RNAfold_virus/tree/main).

529 **RNA Virus-Host relationship**

530 Leveraging iPhop v1.3.3[817] for viral host prediction, in addition with host information
531 of closely related viruses identified through sequence alignments, and annotated viral
532 sequences with host genes, we identified the potential hosts for Lake Nam Co RNA
533 viruses. Kaiju v 1.10.0 was employed to predict the taxonomic classification of all

contigs from Lake Nam Co. Amino acid usage frequency (AAUF) and the di-/tetra-nucleotide usage frequency (D/TUF) of the contigs exceeding 10,000 nucleotides along with the identified RNA viral contigs identified were then predicted and visualized utilizing scripts published in previous research[82].

References

1. Vago A Virus Disease in Crustacea. *Nature* **209**, 1290 (1966).
2. Bolduc, B. et al. Identification of novel positive-strand RNA viruses by metagenomic analysis of archaea-dominated Yellowstone hot springs. *J Virol* **86**, 5562-5573 (2012).
3. Shi, M. et al. High-Resolution Metatranscriptomics Reveals the Ecological Dynamics of Mosquito-Associated RNA Viruses in Western Australia. *J Virol* **91** (2017).
4. Callanan, J. et al. RNA Phage Biology in a Metagenomic Era. *Viruses* **10** (2018).
5. Shi, M. et al. Redefining the invertebrate RNA virosphere. *Nature* **540**, 539-543 (2016).
6. Dominguez-Huerta, G. et al. The RNA virosphere: How big and diverse is it? *Environ Microbiol* **25**, 209-215 (2023).
7. Wolf, Y.I. et al. Doubling of the known set of RNA viruses by metagenomic analysis of an aquatic virome. *Nat Microbiol* **5**, 1262-1270 (2020).
8. Chen, Y.M. et al. RNA viromes from terrestrial sites across China expand environmental viral diversity. *Nat Microbiol* **7**, 1312-1323 (2022).
9. Zayed, A.A. et al. Cryptic and abundant marine viruses at the evolutionary origins of Earth's RNA virome. *Science* **376**, 156–162 (2022).
10. Liu, F., Miao, Y., Liu, Y. & Hou, T. RNN-VirSeeker: A Deep Learning Method for Identification of Short Viral Sequences From Metagenomes. *IEEE/ACM Trans Comput Biol Bioinform* **19**, 1840-1849 (2022).
11. Miao, Y., Liu, F., Hou, T. & Liu, Y. Virtifier: a deep learning-based identifier for viral sequences from metagenomes. *Bioinformatics* **38**, 1216-1222 (2022).
12. Shang, J. & Sun, Y. CHEER: HierarCHical taxonomic classification for viral mEtagEnomic data via deep leaRning. *Methods* **189**, 95-103 (2021).
13. Sukhorukov, G. et al. VirHunter: A Deep Learning-Based Method for Detection of Novel RNA Viruses in Plant Sequencing Data. *Front Bioinform* **2**, 867111 (2022).
14. Hou, X. et al. Using artificial intelligence to document the hidden RNA virosphere. *bioRxiv*, 2023.2004.2018.537342 (2024).
15. Wolf, Y.I. et al. Origins and Evolution of the Global RNA Virome. *mBio* **9** (2018).
16. Koonin, E.V. et al. Global Organization and Proposed Megataxonomy of theVirus World. *Microbiology and Molecular Biology Reviews* **84**, e00061-00019 (2020).
17. Neri, U. et al. Expansion of the global RNA virome reveals diverse clades of bacteriophages. *Cell* **185**, 4023-4037 e4018 (2022).
18. Koonin, E.V., Kuhn, J.H., Dolja, V.V. & Krupovic, M. Megataxonomy and global ecology of the virosphere. *ISME J* **18** (2024).
19. Sun, J. et al. Diverse, abundant and stable coastal RNA viruses identified by spatiotemporal

metatranscriptomics. (2022).
20. Urayama, S.I. et al. Double-stranded RNA sequencing reveals distinct riboviruses associated with thermoacidophilic bacteria from hot springs in Japan. *Nat Microbiol* **9**, 514-523 (2024).
21. Le Lay, C., Stott, M.B., Shi, M., Sadiq, S. & Holmes, E.C. A metatranscriptomic analysis of geothermal hot springs reveals diverse RNA viruses including the phylum Lenarviricota. *Virology* **587**, 109873 (2023).
22. John, D.E. & Rose, J.B. Review of Factors Affecting Microbial Survival in Groundwater. *Environmental Science & Technology* **39**, 7345-7356 (2005).
23. Yao, T. et al. Third Pole Environment (TPE). *Environmental Development* **3**, 52-64 (2012).
24. Guoqing, Z. (A Big Earth Data Platform for Three Poles, 2019).
25. Xu, X., Lu, C., Shi, X. & Gao, S. World water tower: An atmospheric perspective. *Geophysical Research Letters* **35** (2008).
26. Cheng, M. et al. A genome and gene catalog of the aquatic microbiomes of the Tibetan Plateau. *Nat Commun* **15**, 1438 (2024).
27. Liu, Y. et al. Diversity and function of mountain and polar supraglacial DNA viruses. *Sci Bull (Beijing)* **68**, 2418-2433 (2023).
28. Huang, A. et al. Evaluating and Improving the Performance of Three 1-D Lake Models in a Large Deep Lake of the Central Tibetan Plateau. *J Geophys Res Atmos* **124**, 3143-3167 (2019).
29. Liu, K. et al. Vertical variation of bacterial community in Nam Co, a large stratified lake in central Tibetan Plateau. *Antonie Van Leeuwenhoek* **109**, 1323-1335 (2016).
30. Edgar, R.C. et al. Petabase-scale sequence alignment catalyses viral discovery. *Nature* **602**, 142-147 (2022).
31. Charon, J., Buchmann, J.P., Sadiq, S. & Holmes, E.C. RdRp-scan: A bioinformatic resource to identify and annotate divergent RNA viruses in metagenomic sequence data. *Virus Evol* **8**, veac082 (2022).
32. Yin, W. et al. Structural basis for inhibition of the RNA-dependent RNA polymerase from SARS-CoV-2 by remdesivir. *Science* **368**, 1499-1504 (2020).
33. Černý, J., Černá Bolfíková, B., Valdés, J.J., Grubhoffer, L. & Růžek, D. Evolution of tertiary structure of viral RNA dependent polymerases. *PLoS One* **9**, e96070 (2014).
34. Chen, I.A. et al. IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Res* **47**, D666-d677 (2019).
35. Cooper, C.E., Nicholls, P. & Freedman, J.A. Cytochrome c oxidase: structure, function, and membrane topology of the polypeptide subunits. *Biochem Cell Biol* **69**, 586-607 (1991).
36. Richter, M.L., Hein, R. & Huchzermeyer, B. Important subunit interactions in the chloroplast ATP synthase. *Biochim Biophys Acta* **1458**, 326-342 (2000).
37. Zhong, L. et al. Heme axial methionine fluxionality in Hydrogenobacter thermophilus cytochrome c552. *Proc Natl Acad Sci U S A* **101**, 8637-8642 (2004).
38. Yan, Y., Zheng, J., Zhang, X. & Yin, Y. dbAPIS: a database of anti-prokaryotic immune system genes. *Nucleic Acids Res* **52**, D419-D425 (2024).
39. Wommack, K.E., Hill, R.T., Muller, T.A. & Colwell, R.R. Effects of sunlight on bacteriophage viability and structure. *Appl Environ Microbiol* **62**, 1336-1341 (1996).
40. Mojica, K.D.A. & Brussaard, C.P.D. Factors affecting virus dynamics and microbial host–

616 virus interactions in marine environments. *FEMS Microbiology Ecology* **89**, 495-515 (2014).
617 41. Thiel, B.C., Beckmann, I.K., Kerpedjiev, P. & Hofacker, I.L. 3D based on 2D: Calculating
618 helix angles and stacking patterns using forgi 2.0, an RNA Python library centered on
619 secondary structure elements. *F1000Res* **8** (2019).
620 42. Boros, A. et al. Multiple divergent picobirnaviruses with functional prokaryotic
621 Shine-Dalgarno ribosome binding sites present in cloacal sample of a diarrheic chicken.
622 *Virology* **525**, 62-72 (2018).
623 43. Krishnamurthy, S.R. & Wang, D. Extensive conservation of prokaryotic ribosomal binding
624 sites in known and novel picobirnaviruses. *Virology* **516**, 108-114 (2018).
625 44. Le Lay, C. et al. Viral community composition of hypersaline lakes. *Virus Evol* **9**, vead057
626 (2023).
627 45. Di Giallonardo, F., Schlub, T.E., Shi, M. & Holmes, E.C. Dinucleotide Composition in Animal
628 RNA Viruses Is Shaped More by Virus Family than by Host Species. *J Virol* **91** (2017).
629 46. Mojica, K.D. & Brussaard, C.P. Factors affecting virus dynamics and microbial host-virus
630 interactions in marine environments. *FEMS Microbiol Ecol* **89**, 495-515 (2014).
631 47. Jacquet, S. & Bratbak, G. Effects of ultraviolet radiation on marine virus-phytoplankton
632 interactions. *FEMS Microbiol Ecol* **44**, 279-289 (2003).
633 48. Zhang, Y. et al. Spatial-temporal variations of euphotic depth and its ecological significance of
634 Lake Taihu. *Sci. China* **49**, 431-442 (2006).
635 49. Knap, A., Michaels, A., Close, A., Ducklow, H. & Dickson, A. Protocols for the joint global
636 ocean flux study (JGOFS) core measurements. *JGOFS, Reprint of the IOC Manuals and*
637 *Guides No. 29, UNESCO 1994* **19** (1996).
638 50. Bolyen, E. et al. Reproducible, interactive, scalable and extensible microbiome data science
639 using QIIME 2. *Nat Biotechnol* **37**, 852-857 (2019).
640 51. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor.
641 *Bioinformatics* **34**, i884-i890 (2018).
642 52. Kopylova, E., Noe, L. & Touzet, H. SortMeRNA: fast and accurate filtering of ribosomal
643 RNAs in metatranscriptomic data. *Bioinformatics* **28**, 3211-3217 (2012).
644 53. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**,
645 357-359 (2012).
646 54. Li, D., Liu, C.M., Luo, R., Sadakane, K. & Lam, T.W. MEGAHIT: an ultra-fast single-node
647 solution for large and complex metagenomics assembly via succinct de Bruijn graph.
648 *Bioinformatics* **31**, 1674-1676 (2015).
649 55. P, R., I, L. & A, B. EMBOSS: The European Molecular Biology Open Software Suite. *Trends*
650 *in Genetics* **16**, 276-277 (2000).
651 56. D, H. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification.
652 **11**, 119 (2010).
653 57. Buchfink, B., Xie, C. & Huson, D.H. Fast and sensitive protein alignment using DIAMOND.
654 *Nat Methods* **12**, 59-60 (2015).
655 58. Mistry, J., Finn, R.D., Eddy, S.R., Bateman, A. & Punta, M. Challenges in homology search:
656 HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res* **41**, e121
657 (2013).
658 59. Olendraite, I., Brown, K. & Firth, A.E. Identification of RNA Virus-Derived RdRp Sequences

in Publicly Available Transcriptomic Data Sets. *Mol Biol Evol* **40** (2023).

60. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150-3152 (2012).
61. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659 (2006).
62. Zayed, A.A., Wainaina, J.M. & Dominguez-Huert, G. Cryptic and abundant marine viruses at the evolutionary origins of Earth's RNA virome. *Science* **376**, 156–162 (2022).
63. Katoh, K. & Standley, D.M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772-780 (2013).
64. Capella-Gutierrez, S., Silla-Martinez, J.M. & Gabaldon, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-1973 (2009).
65. Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M. & Barton, G.J. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189-1191 (2009).
66. Price, M.N., Dehal, P.S. & Arkin, A.P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).
67. Xie, J. et al. Tree Visualization By One Table (tvBOT): a web application for visualizing, modifying and annotating phylogenetic trees. *Nucleic Acids Res* **51**, W587-W592 (2023).
68. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).
69. Zhang, C., Shine, M., Pyle, A.M. & Zhang, Y. US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nat Methods* **19**, 1109-1115 (2022).
70. Huang, J. et al. Discovery of deaminase functions by structure-based protein clustering. *Cell* **186**, 3182-3195 e3114 (2023).
71. DeLano, W.L. Pymol: An open-source molecular graphics tool. *Protein Crystallogr* **40**, 82-92 (2002).
72. Schrodinger, LLC (2015).
73. Crooks, G.E., Hon, G., Chandonia, J.M. & Brenner, S.E. WebLogo: a sequence logo generator. *Genome Res* **14**, 1188-1190 (2004).
74. Steinegger, M. & Soding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* **35**, 1026-1028 (2017).
75. Tisza, M.J., Belford, A.K., Dominguez-Huerta, G., Bolduc, B. & Buck, C.B. Cenote-Taker 2 democratizes virus discovery and sequence annotation. *Virus Evol* **7**, veaa100 (2021).
76. Cantalapiedra, C.P., Hernandez-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol Biol Evol* **38**, 5825-5829 (2021).
77. Zheng, J. et al. dbCAN3: automated carbohydrate-active enzyme and substrate annotation. *Nucleic Acids Res* **51**, W115-W121 (2023).
78. Shaffer, M. et al. DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Res* **48**, 8883-8900 (2020).
79. Guo, J. et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* **9**, 37 (2021).

80. Gregory, A.C. et al. MetaPop: a pipeline for macro- and microdiversity analyses and visualization of microbial and viral metagenome-derived populations. *Microbiome* **10**, 49 (2022).
81. Roux, S. et al. iPHoP: An integrated machine learning framework to maximize host prediction for metagenome-derived viruses of archaea and bacteria. *PLoS Biol* **21**, e3002083 (2023).
82. Ni, Y., Xu, T., Yan, S., Chen, L. & Wang, Y. Hiding in plain sight: The discovery of complete genomes of 11 hypothetical spindle-shaped viruses that putatively infect mesophilic ammonia-oxidizing archaea. *Environ Microbiol Rep* **16**, e13230 (2024).

Data availability

The transcriptome and microbial community data have been uploaded to the National Center for Biotechnology Information under accession number PRJNA1089960.

Author contributions

Z.Q., D.X., and L.Y. initiated and designed the study. W.L., L.K., X.J., and L.P. carried out field sampling. W.L. performed total RNA extraction. W.L., Z.Q., D.X., and S.W. performed all analyses of metagenomes, metatranscriptomes, and RNA viral communities with the help of L.Z. and S.M.. W.L., Z.Q., D.X., and others led the writing of the manuscript, with result interpretations from J.N. and C.T.. C.T. and S.M. helped revise the paper. All authors contributed to editing the text and approved the final version.

Acknowledgements

This work was supported by National Natural Science Foundation of China (42188102, 42222604, 92351303 and 92251306), National Key Research and Development Project of the Ministry of Science and Technology of China (Grant No. 2021QZKK0102).

Competing interests

The authors declare no competing interests.

731 ## Figure legends

732 **Figure 1. Sampling locations and RNA virus detection process.** The diagram outlines the
733 workflow from data collection and processing to the construction of the RdRP dataset for Lake
734 Nam Co. The methods used for retrieving RdRP sequences include BLASTp, HMM searches, and
735 the deep learning algorithm LucaProt. In total, 742 non-redundant complete RdRP sequences were
736 identified from 20 metatranscriptomes. Of these, 724 RdRPs were identified through both protein
737 structure-based clustering and sequence-alignment-based analyses, while the remaining 18 were
738 detected exclusively via structure-based clustering. The RdRP dataset consists of RNA viruses
739 from the five orthornaviran phyla and the *Birnaviridae* family, which is not assigned to the five
740 major phyla. The pie chart displays the distribution of RNA virus species across the five
741 orthornaviran phyla and *Birnaviridae*, based on combined sequence-alignment and protein
742 structure clustering analyses.

743

744 **Figure 2. Classification and structural information of RdRP dataset in Lake Nam Co.**
745 (A) Phylogenetic trees of 724 sequence-alignment-based identified RdRPs. From I to V, they
746 represent phyla *Lenarviricota*, *Pisuviricota*, *Kitrinoviricota*, *Duplornaviricota*, and
747 *Negarnaviricota*, respectively. RdRPs found in Lake Nam Co are highlighted in red, while the
748 reference sequences are shown in black. (B) Schematic representation of 18
749 sequence-alignment-based unidentified RdRPs that detected solely via protein structure-based
750 clustering. The "thumb", "palm", and "fingers" regions of RdRP are colored in pink, yellow, and
751 blue, respectively. Within the "palm" region, catalytic center motifs A, B, and C are colored in
752 purple, green, and orange, respectively.

753

754 **Figure 3. Schematic representation of catalytic core in the sequence-alignment-based**
755 **unidentified RdRP structures.**
756 (A) Unique structures of the "palm" region of sequence-alignment-based unidentified RdRPs that
757 could be identified at the phyla level based on protein structural clustering. (B) Structures of the
758 "palm" region of RdRPs that remain unidentified at the phyla level through either

759 sequence-alignment-based or protein structural-based clustering analyses. Motifs from
760 sequence-alignment-based unidentified RdRPs are colored in red, motifs from the RVMT
761 reference structure are colored in dark blue, and motifs representing sequence-alignment-based
762 identified RdRPs in Lake Nam Co are colored in light blue. The yellow dashed circle indicates
763 unique structural regions of sequence-alignment-based unidentified RdRPs from Lake Nam Co.

764

765 **Figure 4. RdRP structure-based clustering, motif arrangements and structural features.**
766 **(**A) Structure-based clustering results of the RdRP dataset in Lake Nam Co and references. **(**B)
767 Composition and arrangement of RdRP motifs at the phylum level in Lake Nam Co, with the
768 motifs of *f.0167*_Nam_Co from *Kitrinoviricota* exhibiting a permuted arrangement shown at the
769 bottom. **(**C) Schematic representation of the RdRP structure of *f.0167* from *Kitrinoviricota*, with
770 overlaid structures from Lake Nam Co (I) and the RVMT database (II), annotated with the number of
771 overlaid sequences. The "thumb," "palm," and "fingers" regions of RdRP are colored in pink, yellow,
772 and blue, respectively. Within the "palm" region, catalytic center motifs A, B, and C are colored in
773 purple, green, and orange, respectively. Motifs in the enlarged details are displayed in stick mode.

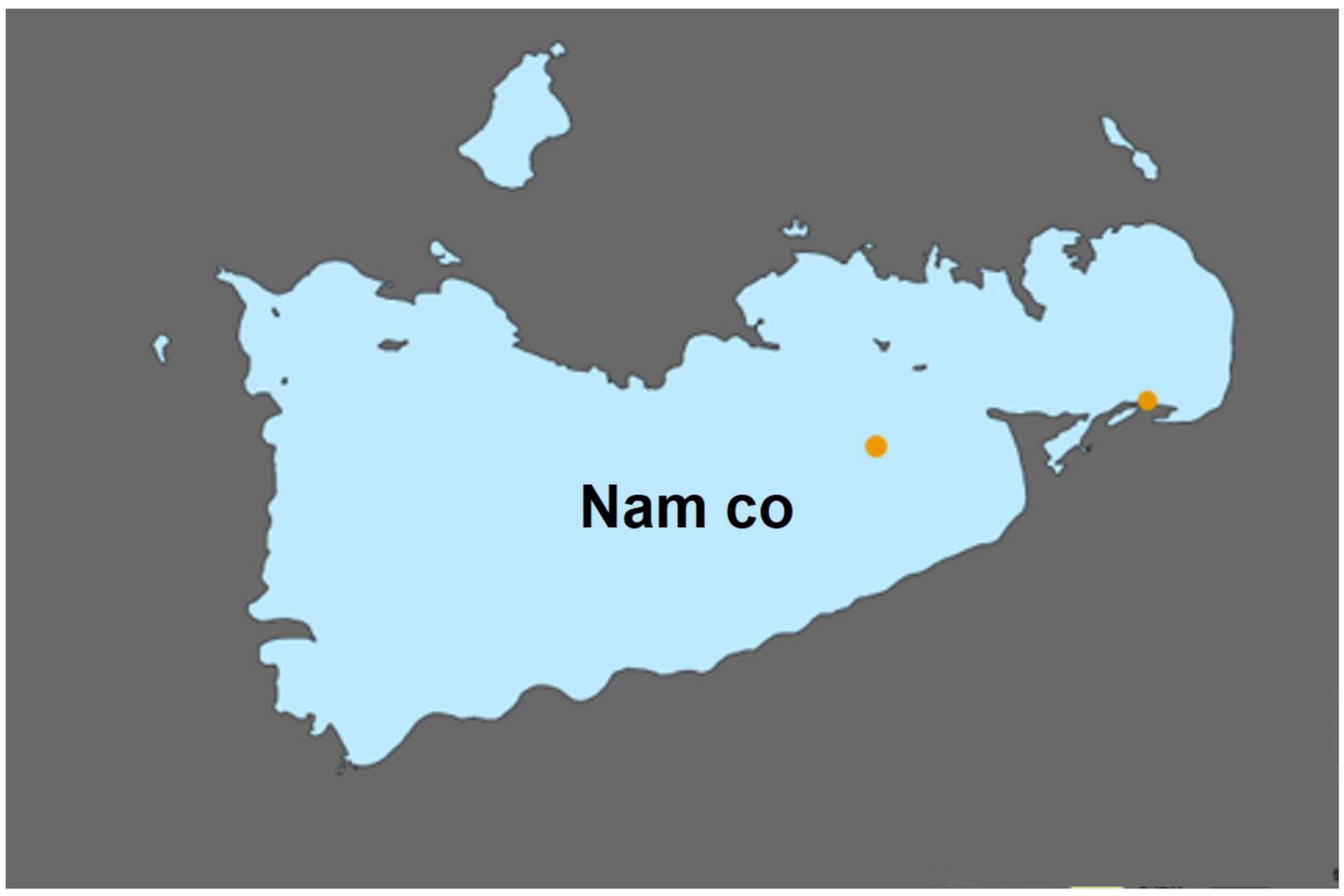
774

775 **Figure 5. Features of genomes and proteins of RNA viruses in Lake Nam Co.**
776 (A) The RNA virus genomes containing viral and host genes in Lake Nam Co. Genes encoding
777 RdRPs are colored in red, genes encoding capsid or coat proteins are colored in blue, other viral
778 genes were colored in green, host genes were colored in yellow. (B) Comparison of refractivity
779 indices of RNA virus proteins from Lake Nam Co with those from terrestrial and estuarine
780 environments. RdRPs and capsid proteins of RNA viruses from different locations are shown in
781 gradient blue and green, respectively. The capsid proteins of Lake Nam Co RNA viruses and those
782 from estuarine environments show no significant difference in refractivity, but they do differ
783 significantly from terrestrial environments. The RdRPs of Lake Nam Co RNA viruses show
784 significant refractivity differences from both estuarine and terrestrial environments. (C) The
785 distribution of normalized minimum free energy (MFE) of RNA secondary structures with respect

786 to viral classes. MFE distributions of RNA viruses from different classes in Lake Nam Co (colored
787 in gradient purple) and from the RVMT database (colored in gradient blue-green) show consistent
788 trends across classes. (D) Comparison of normalized MFE of RNA secondary structures between Lake
789 Nam Co and the RVMT database at the level of viral classes. MFE of RNA viruses from the same class
790 in Lake Nam Co (colored in purple) and the RVMT database (colored in green) show significant
791 differences between nine classes.

792

793 **Figure 6. Host prediction for RNA viruses in Lake Nam Co.**
794 (A) Virus-host classification of RNA viruses with specific hosts. Most of the determined hosts are
795 prokaryotes (72 hosts), and nine and one eukaryotic and archaeal hosts, respectively, were
796 identified. (B) UMAP visualization of tetra-nucleotide usage frequency (TUF) of RNA viruses and
797 microbes in Lake Nam Co at the super-kingdom level. Possible hosts with similar tetra-nucleotide
798 usage frequency to RNA viruses cluster together in UMAP visualization. Most of RNA viruses
799 cluster with prokaryotes (purple dots), while a limited portion of RNA viruses cluster with
800 eukaryotes (blue dots). (C) UMAP visualization of TUF of RNA viruses and microbes in Lake
801 Nam Co at the phylum level. Possible hosts with similar tetra-nucleotide usage frequency to RNA
802 viruses cluster together in UMAP visualization. RNA viruses infecting bacteria primarily comprise
803 *Bacteroidota*, *Cyanobacteriota*, *Chloroflexota*, and *Actinomycetota*.



20 Water Samples
≈ 400 L

20 Meta-transcriptomes

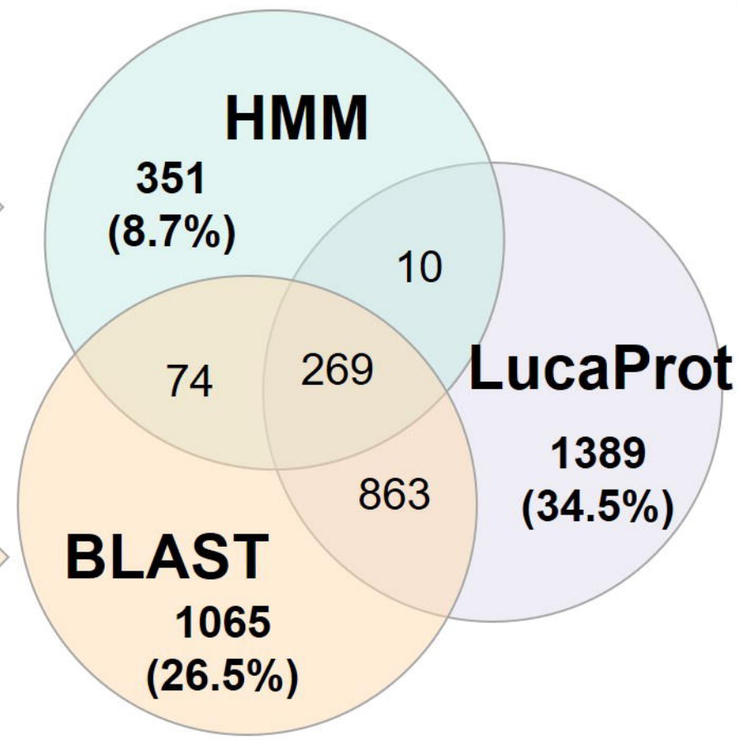
~8 million assembled contigs

~5 million ORFs

RdRP searching

Olendraite et al., 2023

Neri et al., 2022
Hou et al., 2023
Charon et al., 2022
Zayed et al., 2022



4021 potential RdRP seqs

Redundancy removal

RdRP candidates
3089 seqs

Motif check

Complete RdRP candidates
775 seqs

Structure check

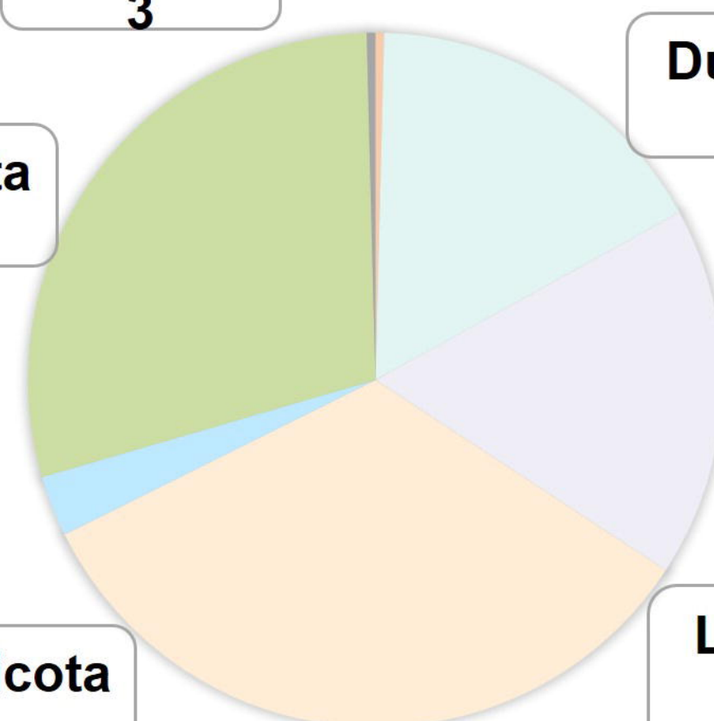
Structure verified RdRPs
742
RdRP dataset

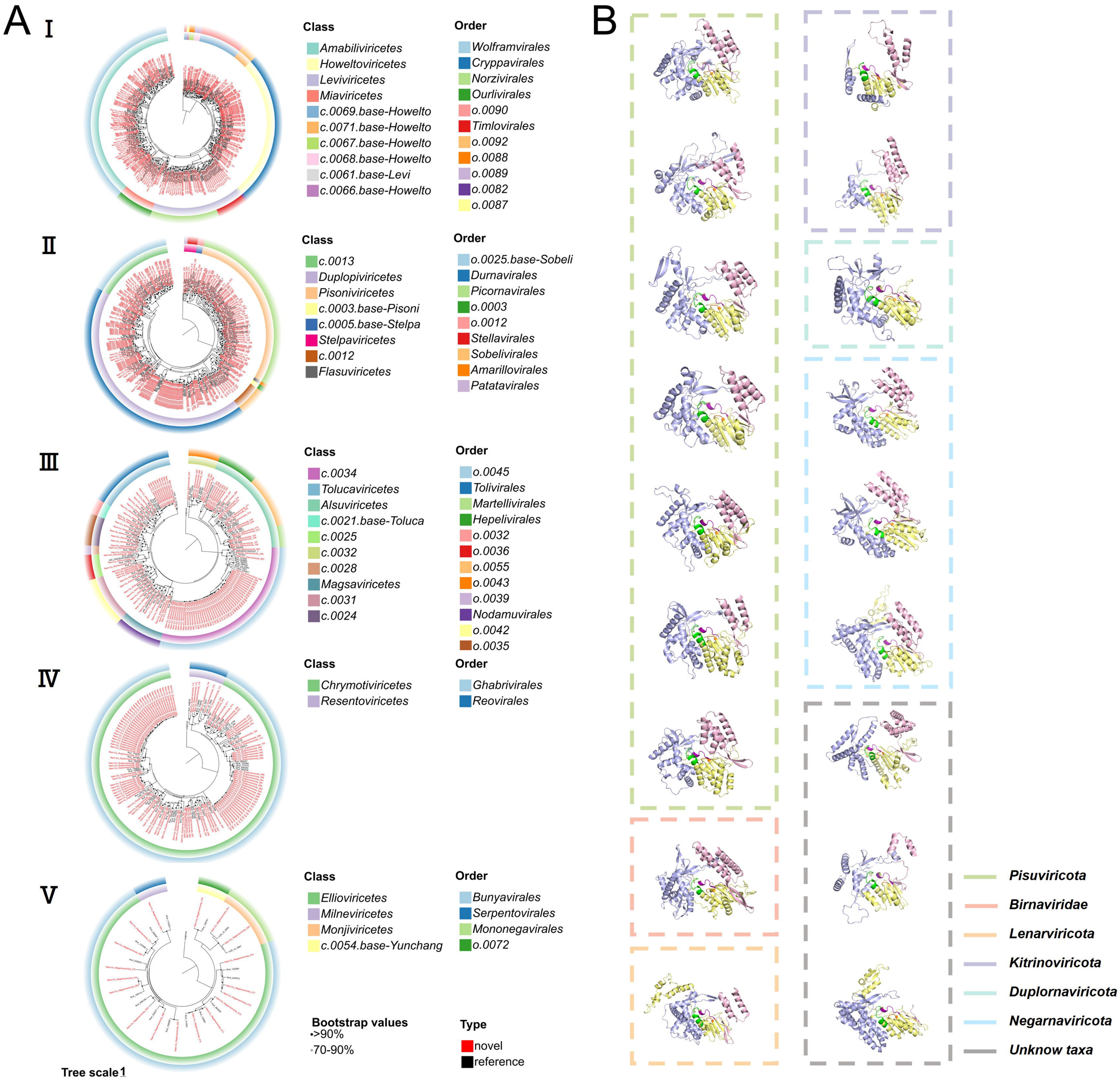
Taxonomy assignment

Sequence-alignment-based identified RdRPs
724

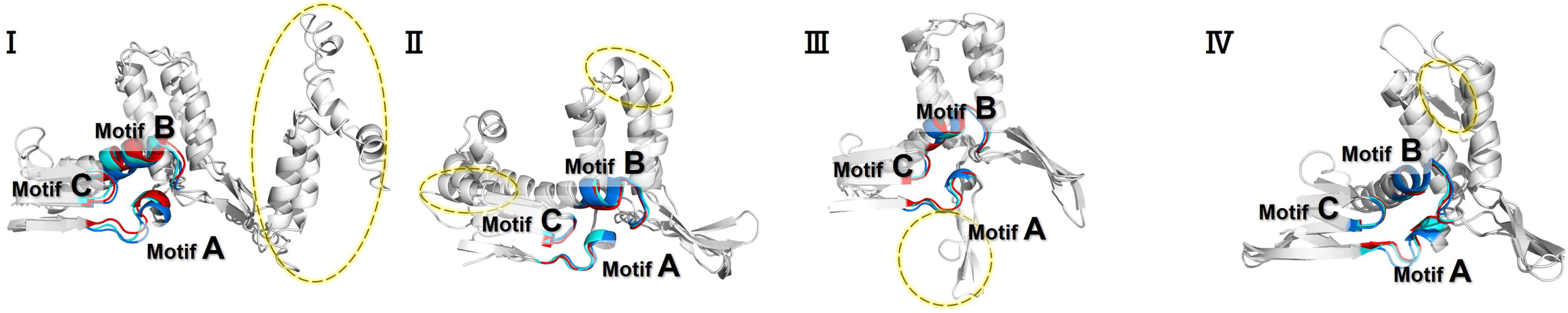
Sequence-alignment-based unidentified RdRPs
18

RDRP DATASET

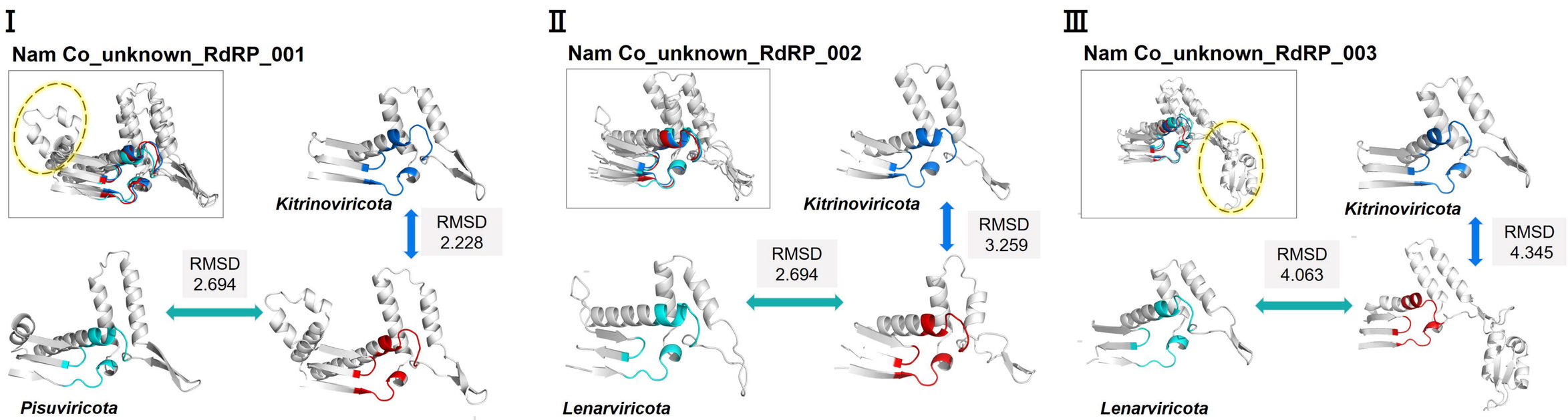


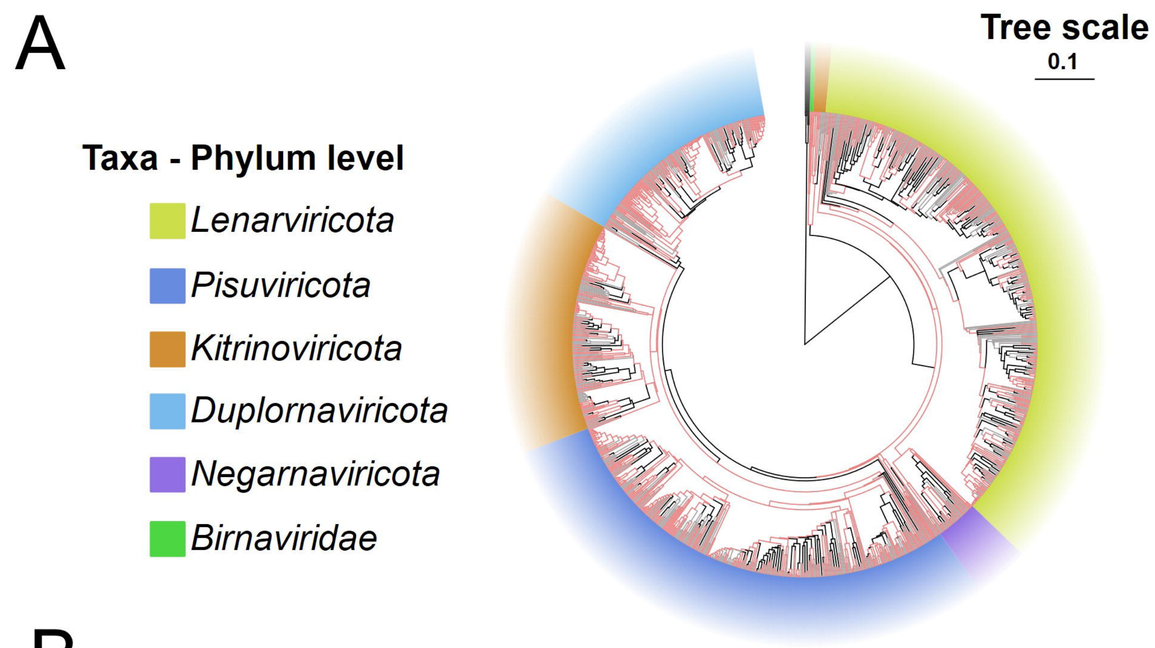

A
I
Class
Amabiliviricetes
Howeltoviricetes
Leviviricetes
Miaviricetes
c.0069.base-Howelto
c.0071.base-Howelto
c.0067.base-Howelto
c.0068.base-Howelto
c.0061.base-Levi
c.0066.base-Howelto
Order
Wolframvirales
Cryppavirales
Norzivirales
Ourlivirales
o.0090
Timlovirales
o.0092
o.0088
o.0089
o.0082
o.0087
II
Class
c.0013
Duplopiviricetes
Pisoniviricetes
c.0003.base-Pisoni
c.0005.base-Stelpa
Stelpaviricetes
c.0012
Flasuviricetes
Order
o.0025.base-Sobeli
Durnavirales
Picornavirales
o.0003
o.0012
Stellavirales
Sobelivirales
Amarillovirales
Patatavirales
III
Class
c.0034
Tolucaviricetes
Alsuviricetes
c.0021.base-Toluca
c.0025
c.0032
c.0028
Magsaviricetes
c.0031
c.0024
Order
o.0045
Tolivirales
Martellivirales
Hepelivirales
o.0032
o.0036
o.0055
o.0043
o.0039
Nodamuvirales
o.0042
o.0035
IV
Class
Chrymotiviricetes
Resentoviricetes
Order
Ghabrivirales
Reovirales
V
Class
Ellioviricetes
Milneviricetes
Monjiviricetes
c.0054.base-Yunchang
Order
Bunyavirales
Serpentovirales
Mononegavirales
o.0072
Bootstrap values
>90%
70-90%
Type
novel
reference
Tree scale 1
B
Pisuviricota
Birnaviridae
Lenarviricota
Kitrinoviricota
Duplornaviricota
Negarnaviricota
Unknow taxa

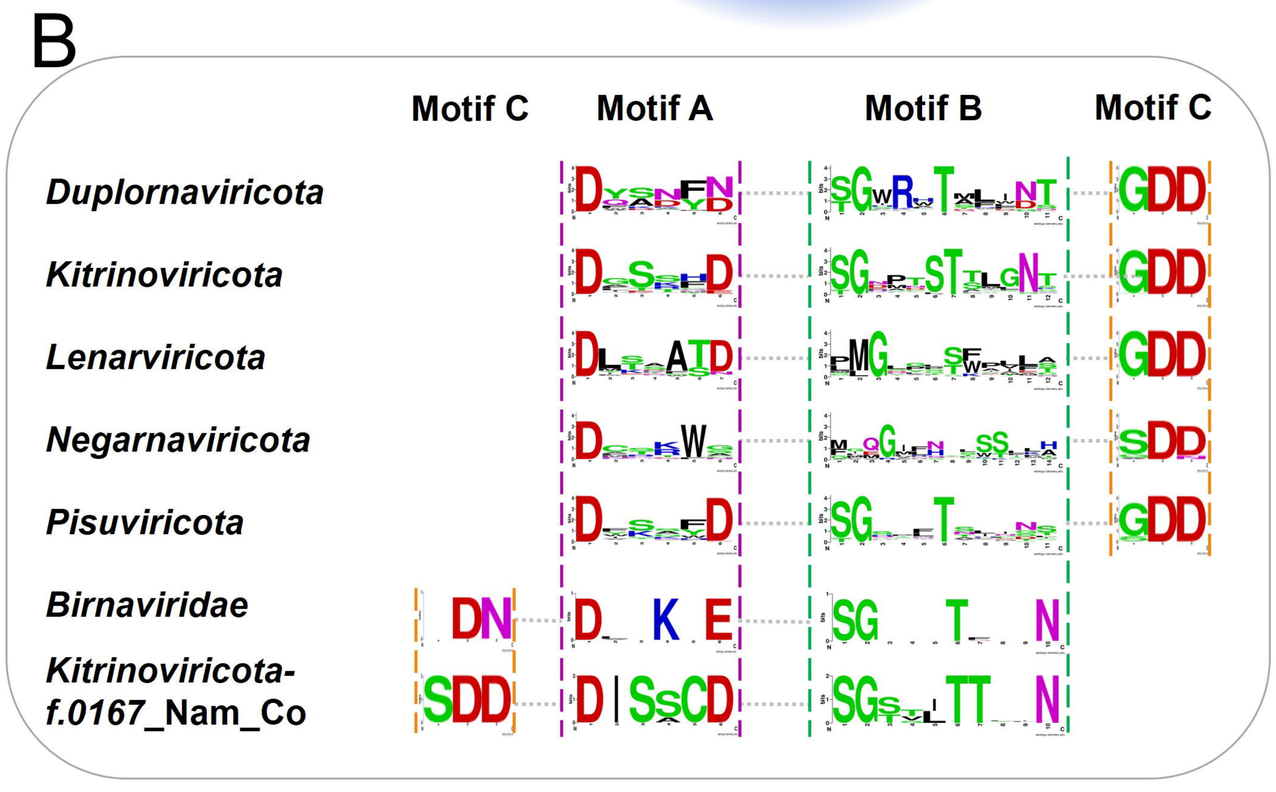
A

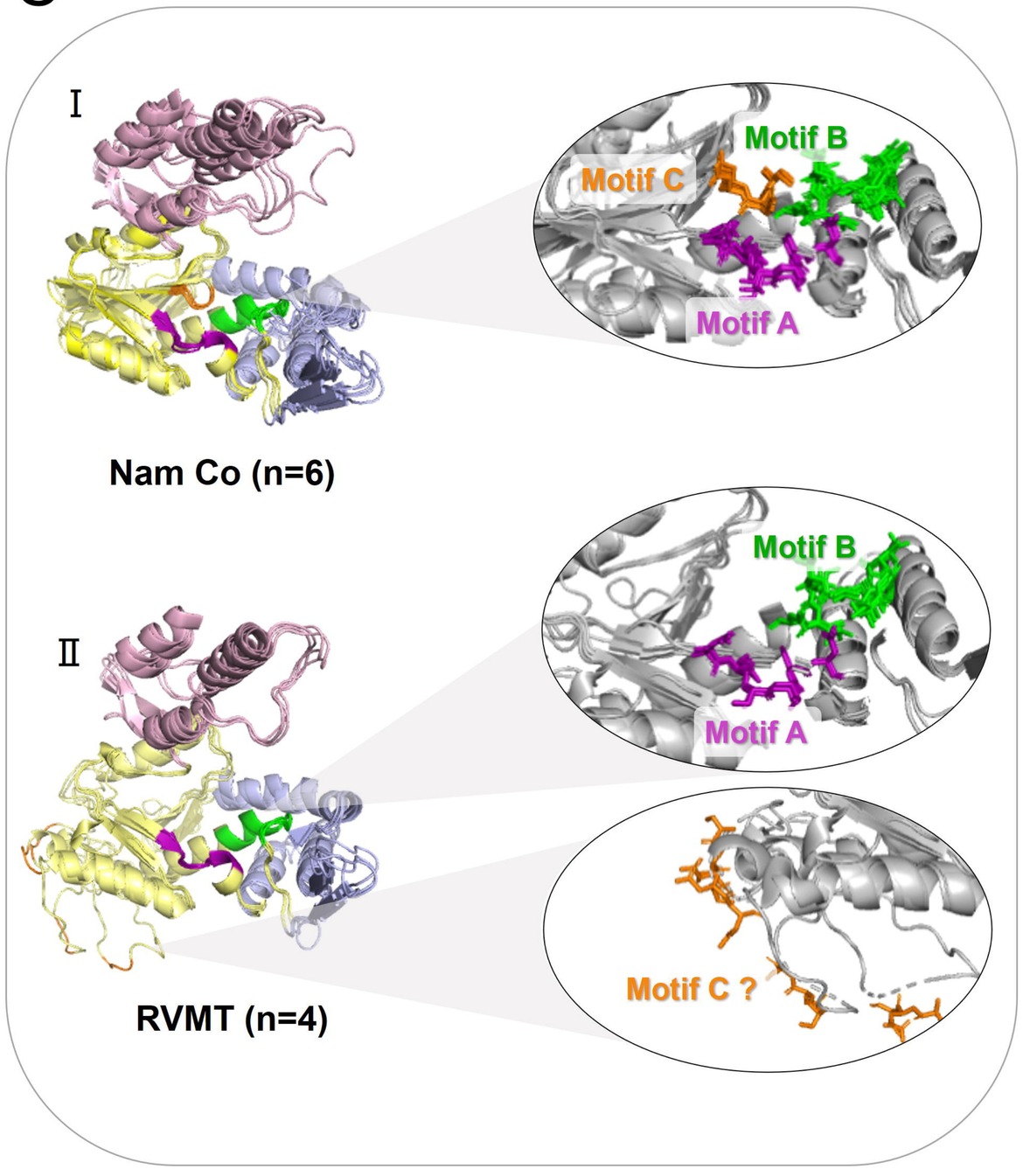

I
Motif B
Motif C
Motif A
Nam Co_Lenarviricota_RdRP_001
II
Motif B
Motif C
Motif A
Nam Co_Pisuviricota_RdRP_004
III
Motif B
Motif C
Motif A
Nam Co_Pisuviricota_RdRP_007
IV
Motif B
Motif C
Motif A
Nam Co_Birnaviridae_RdRP_001

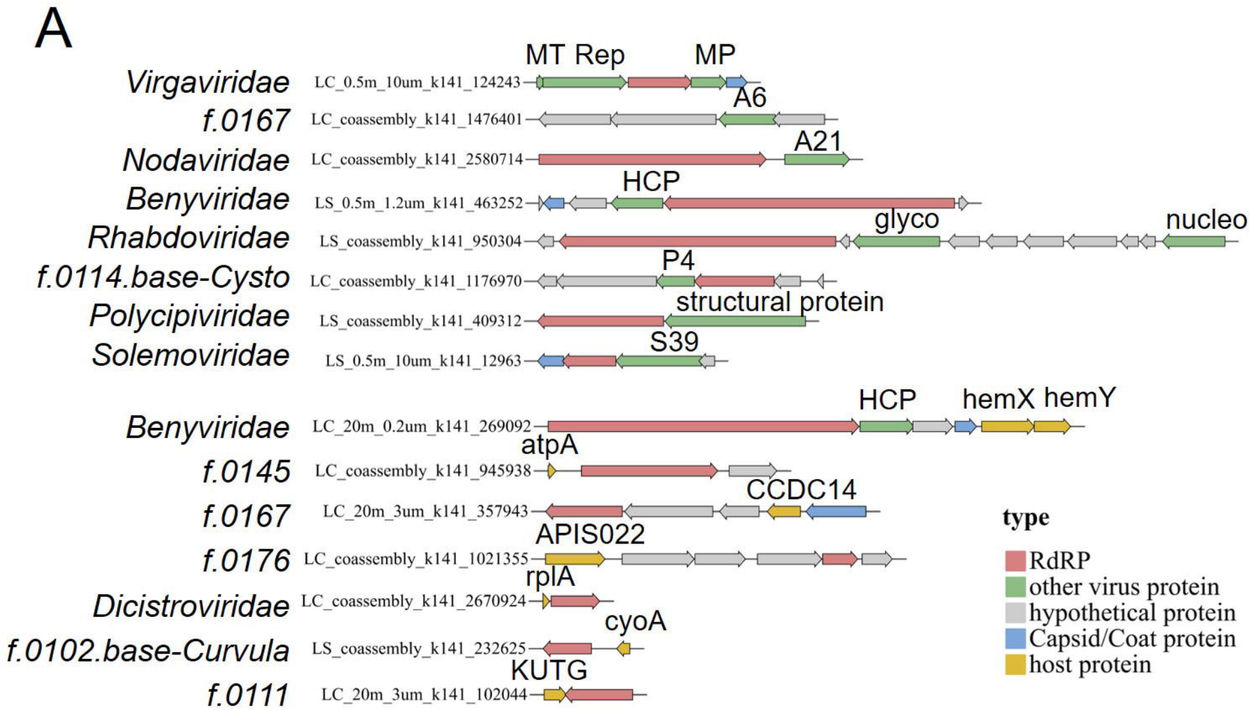
B

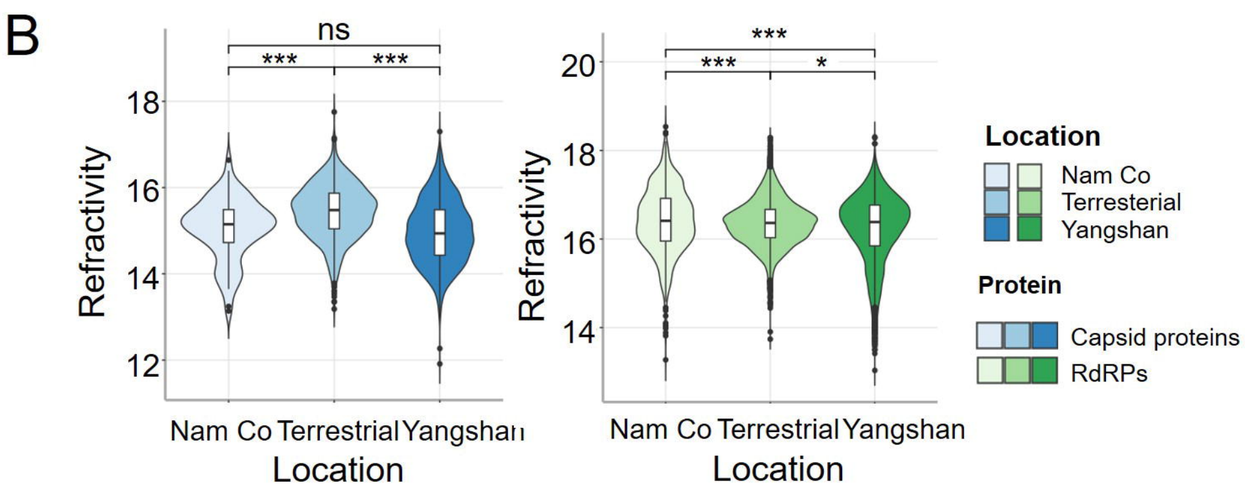

I
Nam Co_unknown_RdRP_001
Kitrinoviricota
RMSD 2.228
RMSD 2.694
Pisuviricota
II
Nam Co_unknown_RdRP_002
Kitrinoviricota
RMSD 3.259
RMSD 2.694
Lenarviricota
III
Nam Co_unknown_RdRP_003
Kitrinoviricota
RMSD 4.345
RMSD 4.063
Lenarviricota

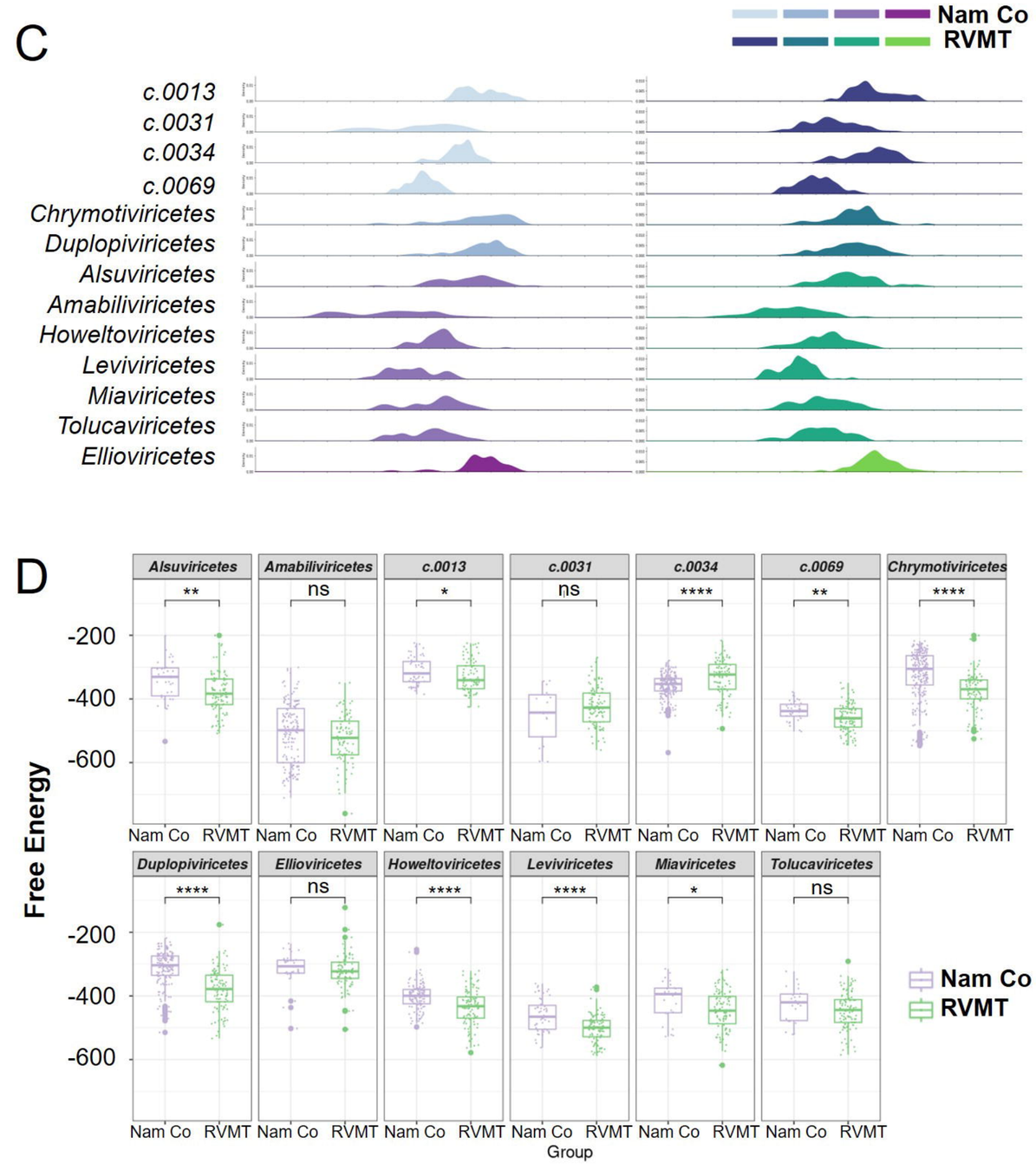

A
Tree scale
0.1
Taxa - Phylum level
Lenarviricota
Pisuviricota
Kitrinoviricota
Duplornaviricota
Negarnaviricota
Birnaviridae


B
Motif C
Motif A
Motif B
Motif C
Duplornaviricota
Kitrinoviricota
Lenarviricota
Negarnaviricota
Pisuviricota
Birnaviridae
Kitrinoviricota-f.0167_Nam_Co


C
I
Motif B
Motif C
Motif A
Nam Co (n=6)
II
Motif B
Motif A
Motif C ?
RVMT (n=4)







* $P < 0.05$, ** $P < 0.01$, *** $p < 0.001$, **** $P < 0.0001$

A

Search method | Superkingdom of Host | Phylum of Host | Family of Host | Phylum of Virus | Class of Virus

Host genes (3)
iPhop (18)
Related viruses (62)

Bacteria (72)
Eukaryota (10)
Archaea (1)

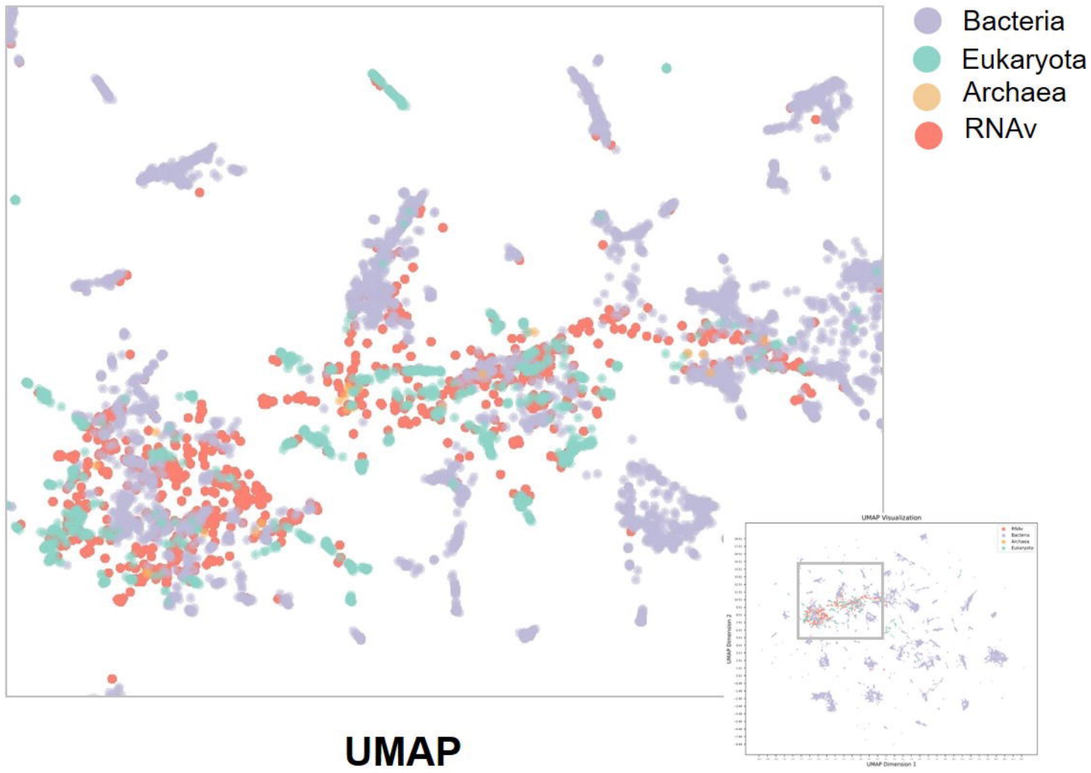
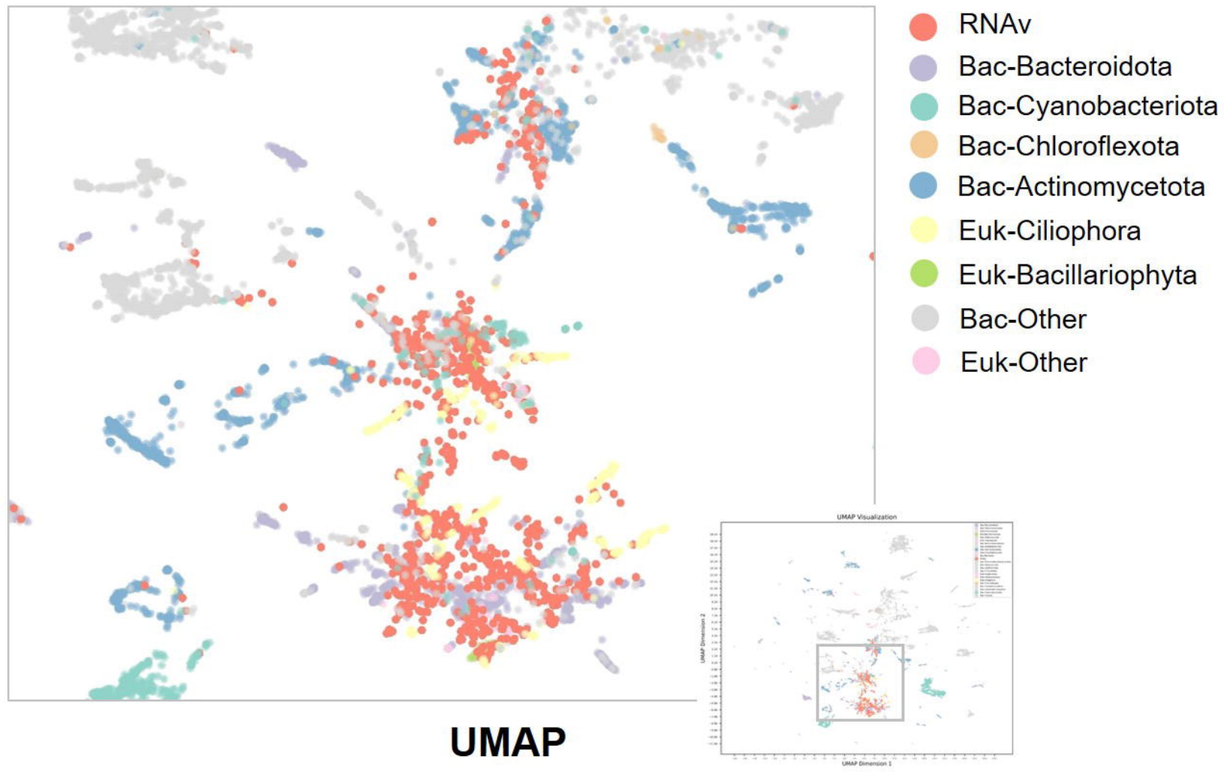
Actinomycetota (2)
Bacillota (3)
Bacteroidota (9)
Pseudomonadota (5)
Oomycota (1)
Arthropoda (2)
Chordata (1)
Streptophyta (6)
Methanobacteriota (1)

Ilumatobacteraceae (1)
Mycobacteriaceae (1)
Desulfitobacteriaceae (1)
UBA660 (1)
CAG-826 (1)
Marinifilaceae (1)
Bacteroidaceae (2)
Flavobacteriaceae (4)
Paludibacteraceae (2)
Midichloriaceae (1)
Comamonadaceae (1)
GCA-002716945 (1)
Gallionellaceae (1)
Pseudomonadaceae (1)
Saprolegniaceae (1)
Aphididae (1)
Gerroidea (1)
Hominidae (1)
Cucurbitaceae (1)
Fabaceae (1)
Poaceae (1)
Solanaceae (3)
Methanothermobacteraceae (1)

Duplornaviricota (8)
Kitrinoviricota (7)
Lenarviricota (54)
Negarnaviricota (2)
Pisuviricota (12)

Chrymotiviricetes (8)
Alsuviricetes (6)
c.0034 (1)
Leviviricetes (51)
Howeltoviricetes (2)
Amabiliviricetes (1)
Monjiviricetes (1)
unsure (1)
Duplopiviricetes (6)
c.0013 (1)
Pisoniviricetes (4)
Stelpaviricetes (1)

B



C