nature microbiology

Resource

https://doi.org/10.1038/s41564-024-01876-7

Unravelling viral ecology and evolution over 20 years in a freshwater lake

Received: 16 February 2024

Accepted: 1 November 2024

Published online: 3 January 2025

Check for updates

Zhichao Zhou[1,2,3], Patricia Q. Tran[1,4], Cody Martin[1,5], Robin R. Rohwer[6], Brett J. Baker[6,7], Katherine D. McMahon[1,8] & Karthik Anantharaman[1,9,10]

As freshwater lakes undergo rapid anthropogenic change, long-term studies reveal key microbial dynamics, evolutionary shifts and biogeochemical interactions, yet the vital role of viruses remains overlooked. Here, leveraging a 20 year time series from Lake Mendota, WI, USA, we characterized 1.3 million viral genomes across time, seasonality and environmental factors. Double-stranded DNA phages from the class *Caudoviricetes* dominated the community. We identified 574 auxiliary metabolic gene families representing over 140,000 auxiliary metabolic genes, including important genes such as *psbA* (photosynthesis), *pmoC* (methane oxidation) and *katG* (hydrogen peroxide decomposition), which were consistently present and active across decades and seasons. Positive associations and niche differentiation between virus–host pairs, including keystone Cyanobacteria, methanotrophs and *Nanopelagicales*, emerged during seasonal changes. Inorganic carbon and ammonium influenced viral abundances, underscoring viral roles in both 'top-down' and 'bottom-up' interactions. Evolutionary processes favoured fitness genes, reduced genomic heterogeneity and dominant sub-populations. This study transforms understanding of viral ecology and evolution in Earth's microbiomes.

Viruses that infect bacteria and archaea (phages) are the most abundant biological entities in ecosystems. Phages can reshape microbial metabolism, drive nutrient cycling and influence global biogeochemical cycles[1,2]. Uncultivated viral genomes obtained from metagenomes have substantially enriched the collection of viruses in public databases and improved our understanding of viruses in nature[3]. In the largest public viral database to date (the Integrated Microbial Genome/Virus database v4, IMG/VR v4), freshwater lake viruses accounted for approximately 15% of all viral genomes, ranking them first among all environmental subtypes. Despite the substantial virus sequence deposition compared with other environments such as oceans and soil, viruses in freshwater lakes remain understudied. Globally, freshwater lakes are undergoing rapid change due to landscape and climate alterations. Microbial communities are foundational players in freshwater ecology[4], and characterizing the diversity, function, ecology and evolution of their viruses will improve our understanding of a major 'top-down' control of microbial communities.

Time-series studies have been adopted in the field of microbial ecology and have revealed microbial dynamics, population variation and ecological impacts of microorganisms on natural ecosystems.

[1]Department of Bacteriology, University of Wisconsin–Madison, Madison, WI, USA. [2]Institute for Advanced Study, Shenzhen University, Shenzhen, China. [3]Synthetic Biology Research Center, Shenzhen University, Shenzhen, China. [4]Freshwater and Marine Sciences Program, University of Wisconsin–Madison, Madison, WI, USA. [5]Microbiology Doctoral Training Program, University of Wisconsin–Madison, Madison, WI, USA. [6]Department of Integrative Biology, The University of Texas at Austin, Austin, TX, USA. [7]Department of Marine Science, Marine Science Institute, The University of Texas at Austin, Port Aransas, TX, USA. [8]Department of Civil and Environmental Engineering, University of Wisconsin–Madison, Madison, WI, USA. [9]Department of Integrative Biology, University of Wisconsin–Madison, Madison, WI, USA. [10]Department of Data Science and AI, Wadhwani School of Data Science and AI, Indian Institute of Technology Madras, Chennai, India. e-mail: karthik@bact.wisc.edu

https://doi.org/10.1038/s41564-024-01876-7

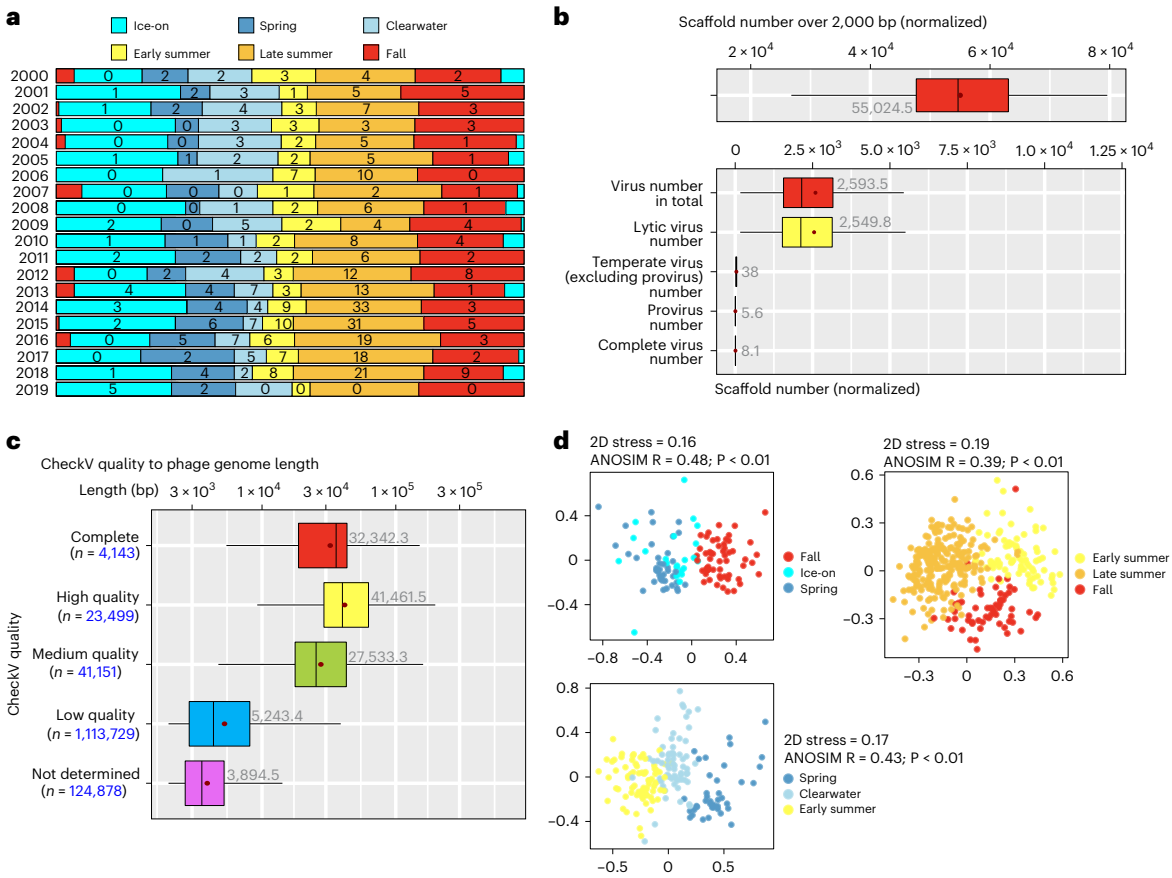


Fig. 1 | Statistics of viral scaffolds and genomes. a, The number of metagenomes obtained in each season across 20 years. **b**, Statistics of total metagenomic scaffolds and viral scaffolds across 471 metagenomes. Scaffold numbers were first normalized by 100 million reads/metagenome to overcome uneven sequencing depth across samples. **c**, Length and completeness of viruses after binning. CheckV quality to completeness range: complete (100%); high quality (90.0–100.0%); medium quality (50.0–89.99%); low quality (0.01–49.99%); not determined. The box plots in **b** and **c** show the median (central line), the mean (dark red dot with labelled values), the interquartile range (spanning from the 25th to the 75th percentiles) and the whiskers. For simplicity, outliers are not shown. **d**, Non-metric multidimensional scaling plots presenting viral genome distribution among seasons. The viral genome abundances were calculated at the family level. ANOSIM, analysis of similarity.

The establishment of long-term microbial and biogeochemical observation projects, such as Hawaii Ocean Time-series[5] and Bermuda Atlantic Time-series[6], have notably improved our understanding of long timescale influences of climate change and environmental alteration on microbial dynamics and matter and energy flows. In addition to allowing observations and analyses of temporal variation, time-series studies can also contribute to an understanding of evolutionary progression. For example, a 6 year time series of lake pelagic bacterial community composition in Lake Mendota (Wisconsin, USA) highlighted regular interannual dynamics and the link between the microbial community and seasonal drivers, which reflected the climate variation[7]. By harnessing the advantage of tracing population microdiversity changes, a 9 year time-series study of lake microorganisms in Trout Bog Lake (Wisconsin, USA) revealed the evolutionary processes of bacterial speciation, which were subjected to two distinctive evolutionary models coexisting in the same environment[8]. However, few time-series studies have been conducted to date involving a comprehensive characterization of viral communities, except for a recent ecogenomic characterization of virophages[9]. By harnessing time-series metagenomics of Lake Mendota and Trout Bog Lake in Wisconsin, USA, the characterization of 25 uncultivated virophages revealed virus–host relationships between virophages and giant viruses, and unravelled ecological and evolutionary patterns over multiple years[9].

There are known potential viral associations with geochemistry, primarily through the activity of auxiliary metabolic genes (AMGs)[10–13]. Functional and metabolic reprogramming of host metabolism by AMGs can maintain, drive or short-circuit important metabolic steps, providing viruses with fitness advantages[11,13–15]. The involvement of AMGs in freshwater ecosystems has been rarely reported, except for recent studies of methane oxidation and photosynthesis in freshwater lakes[16], compared with well-studied instances of photosynthesis[17–19], sulfur oxidation[12,20,21], ammonia oxidation[22] and ammonification[23] in the oceans. In addition, research linking how viral populations and their ecological functions are influenced by environmental factors remains elusive.

In this Article, we leveraged time-series metagenomes collected over 20 years (2000–2019; the 'TYMEFLIES' (Twenty Years of Metagenomes Exploring Freshwater Lake Interannual Eco/evo Shifts) metagenome project) to study freshwater viral diversity, ecology and their association with metabolism and their hosts.

Results

Freshwater lakes harbour enormous uncharacterized viral diversity

In this study, we analysed a total of 471 metagenome samples. For each year, we divided the samples into six seasons (Fig. 1a and Supplementary Table 1). These seasons—ice-on, spring, clearwater, early summer, late summer and fall—were defined by environmental data and most accurately represent microbial phenology[24,25]. Our analysis identified a total of 1,820,639 viral scaffolds, with an average of approximately 2,600

https://doi.org/10.1038/s41564-024-01876-7

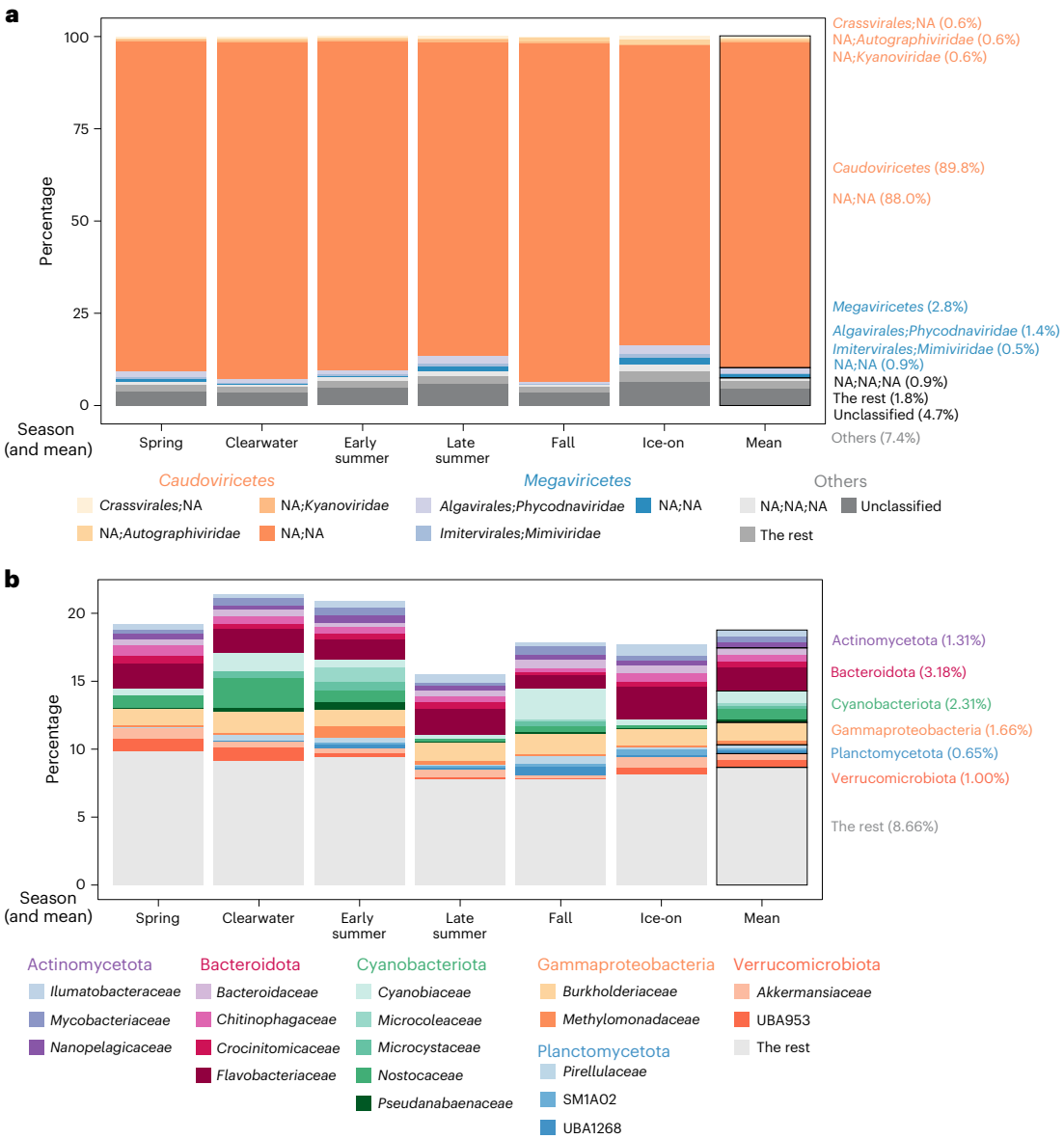


Fig. 2 | Seasonal abundance distribution of viruses and viral hosts. a, Seasonal abundance distribution of viruses at the family level. Families of relative abundance <0.5% across all seasons were combined as 'The rest'. The 'NA;NA' taxon within *Caudoviricetes* and *Megaviricetes* indicates families unclassified within the respective classes. Similarly, the 'NA;NA;NA' taxon under 'Others' represents other unclassified families. NA, not determined. **b**, Seasonal abundance distribution of viral hosts at the family level. Families of relative abundance <0.5% across all seasons were combined as 'The rest'. Unclassified host families were not depicted in the bar plot. In addition, each bar plot includes a 'Mean' bar, indicating the average percentage of relative abundances for all six seasons, with corresponding mean percentages clearly labelled in each plot.

viral scaffolds per metagenome (Fig. 1b and Supplementary Table 2). Applying a stringent binning approach, we obtained 1,307,400 vMAGs (viral metagenome-assembled genomes or viral bins) (Fig. 1c, Extended Data Fig. 1 and Supplementary Table 3). In this study, a substantial number of viral genomes were generated, constituting approximately one-quarter of the entries within the IMG/VR v4 database, which encompasses around 5.6 million high-confidence viruses[3].

The use of viral genome binning significantly improved the length and completeness of our viral sequence collections, as assessed by CheckV[26], leading to a notable enhancement in overall viral genome quality (Fig. 1c, Extended Data Fig. 1 and Supplementary Table 4). Furthermore, we clustered all vMAGs into 749,694 species based on 97% sequence identity. The number of species identified in our samples is comparable to approximately one-quarter of all virus species catalogued in the IMG/VR v4 database (high-confidence viruses) (749,694 versus 2,917,521), underscoring the substantial viral diversity in freshwater lakes. It is worth noting that the rarefaction curve of species did not plateau, suggesting that there is still a considerable amount of unknown viral diversity in freshwater lakes (Extended Data Fig. 1d). To delve deeper into the patterns observed, we examined the viral genome distribution at the family level, revealing distinct separation among viral communities across different seasons (Fig. 1d). Collectively, this not only highlights the richness of viral species but also emphasizes the seasonal and dynamic nature of viral communities.

Viral taxonomic classification revealed that dsDNA viruses in the class *Caudoviricetes* were the predominant group, followed by nucleocytoplasmic large DNA viruses in the class *Megaviricetes* (Fig. 2a). Most viral species identified within *Caudoviricetes* could not be taxonomically classified further, highlighting the need to investigate the extensive diversity of this class in freshwater environments. Less than

https://doi.org/10.1038/s41564-024-01876-7

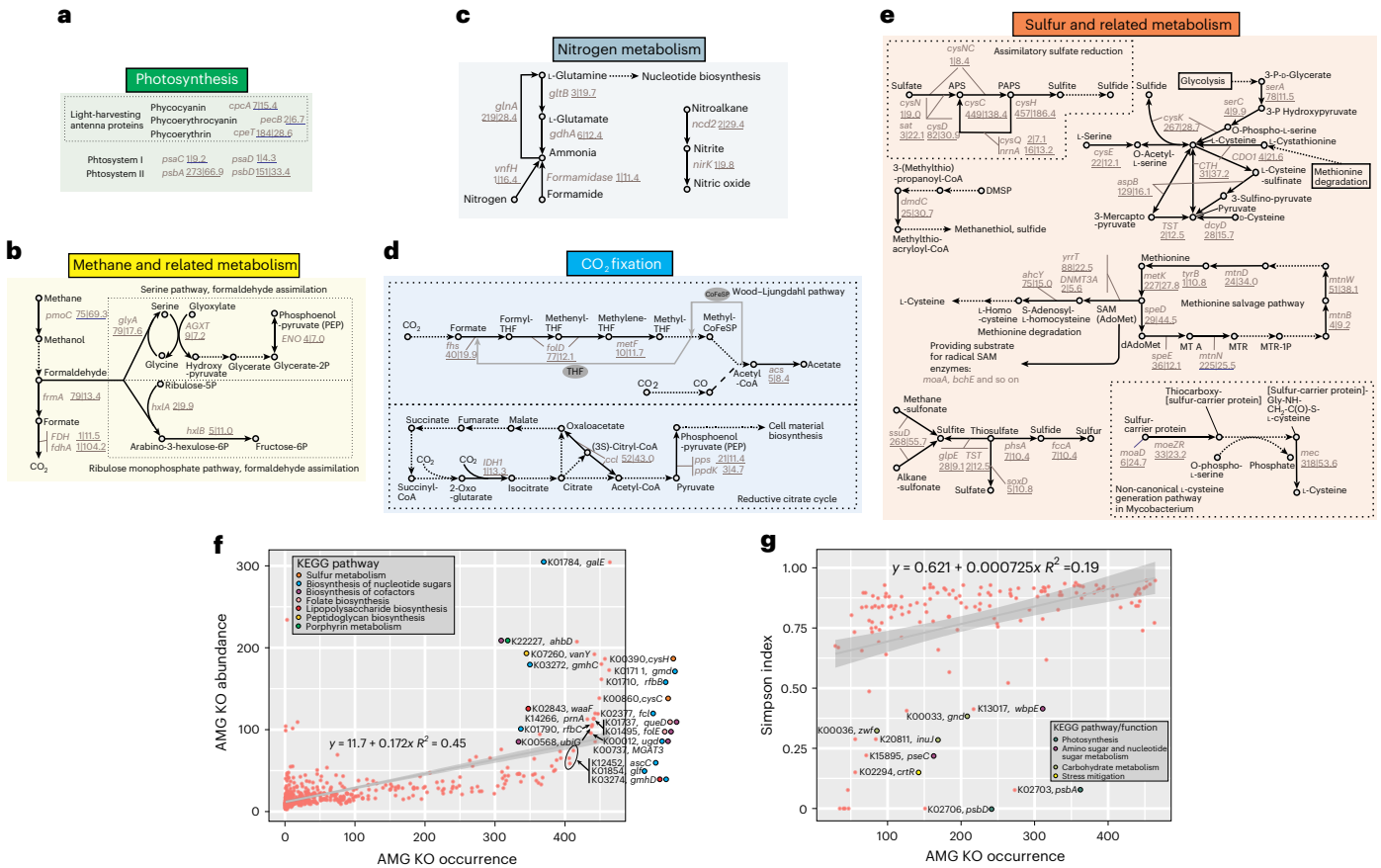


Fig. 3 | AMGs associated with carbon, nitrogen and sulfur metabolism in freshwater viruses and modes of distribution patterns of AMGs. **a**, AMGs involved in photosynthesis. **b**, AMGs involved in methane and related metabolism. **c**, AMGs involved in nitrogen metabolism. **d**, AMGs involved in CO_2 fixation. **e**, AMGs involved in sulfur and related metabolism. THF, tetrahydrofolate; APS, adenosine 5′-phosphosulfate; PAPS, 3′-phosphoadenosine 5′-phosphosulfate; DMSP, dimethylsulfoniopropionate; SAM, S-adenosylmethionine; MTA, methylthioadenosine; MTR, methylthioribose; dAdoMet, 5′-deoxyadenosylmethionine. Gene symbols are depicted in blue together with the occurrence and abundance values (occurrence, the number of metagenomes in which an AMG cluster can be found; abundance, the mean normalized abundance of AMG cluster carrying viruses in the metagenomes that this AMG cluster can be found). Dotted arrows indicate steps that are not encoded by AMGs. Detailed information on each AMG cluster can be found in Supplementary Table 6. The full set of AMG metabolisms and functions are depicted in Extended Data Fig. 2. **f**, The scatter plot of AMG cluster 'abundance to occurrence'. High occurrence refers to distribution in >400 metagenomes. AMG cluster KO dots are labelled with KO ID, gene name and the corresponding KEGG pathway. AMG cluster abundance was normalized by 100 million reads/metagenome. **g**, The correlation of Simpson index of host taxonomy (family level) to AMG cluster occurrence. AMG clusters with low Simpson indices (narrow host range) compared with the fitting curve are labelled with the KEGG pathway/function information (only AMG clusters distributed >50 metagenomes were considered). In panels **f** and **g**, the scatter plots show linear regression curves with corresponding equations and R^2 values, while the shaded error bands represent the 95% confidence intervals for the regression lines.

20% of the viral community received host prediction assignments (Fig. 2b). Across all six seasons, Bacteroidota were the most abundant host phylum, followed by Cyanobacteriota and Gammaproteobacteria, consistent with the dominant phyla identified in a parallel prokaryotic study[25]. The host prediction results revealed a diverse range of virus hosts, underscoring the extensive uncharted territory in this field.

AMG distribution patterns range from broad to narrow host specificity

We identified 143,751 AMGs from viral genomes, which clustered into 574 protein families (Kyoto Encyclopedia of Genes and Genomes (KEGG) ortholog groups, KOs, hereafter referred to as AMG clusters) (Supplementary Table 5). These protein families showed a diverse functional repertoire, comprising 12 distinct categories in total covering important biogeochemical transformations of C, N and S (Fig. 3, Extended Data Fig. 2, Supplementary Table 6 and Supplementary Results). This repertoire was larger than those reported in previous studies (34 clusters discovered in the Global Ocean Survey communities and 322 clusters discovered in pelagic and benthic communities of the Baltic Sea)[27,28], highlighting the high diversity of our recovered vMAGs. The roles played by freshwater AMGs encompass crucial processes such as photosynthesis, methane oxidation, CO_2 fixation for energy and carbon metabolisms, nitrogen metabolism for nucleotide biosynthesis, and sulfur metabolism for organosulfur degradation and sulfide production. Meanwhile, the viral genome and AMG expression results revealed that active viruses also contained actively expressed AMGs, indicating viral involvement in host metabolic functions and biogeochemical processes over time (Supplementary Results). We propose that, similar to other ecosystems, these AMGs likely provide substantial fitness benefits to viruses.

Two modes of viral auxiliary metabolism (namely, broad and narrow) were discovered based on their distribution across broad or narrow host ranges, respectively. Broad host range AMGs were high-abundance AMG protein families that showed a high-occurrence distribution (identified in >85% of all samples), were unaffected by seasonal change and were resilient to intra-population dynamics of specific AMG-carrying viral species (Fig. 3f and Extended Data Fig. 3). These AMG protein families were primarily involved in sulfur metabolism

and cofactor and folate biosynthesis. Viral auxiliary metabolism for organic and inorganic sulfur transformations often results in the production of sulfide as an end product[29], which benefits host survival, growth, amino acid synthesis, protein function and virion assembly (Supplementary Results).

Conversely, narrow host range AMGs revealed AMG protein families with limited host ranges that performed specific functions (Fig. 3g and Extended Data Fig. 4). Examples of these AMG families include *psbA/D*, which encode photosystem II reaction centre domains D1/D2[10,11,14,30]. These AMG families can maintain photosynthetic activity of infected cyanobacteria[13]. Previous research suggests cyanophages can either decouple carbon fixation from photosynthesis[13,31] or suppress the Calvin cycle and alter host metabolism towards the pentose phosphate pathway for NADPH generation and deoxynucleotide biosynthesis[32]. In our analyses, we also identified specific AMG families associated with the pentose phosphate pathway, such as *gnd* and *zwf*. Other examples include *crtG*, which is responsible for antioxidant production and mitigating stress from reactive oxygen species[33] in *Burkholderiaceae*, thereby enhancing the overall fitness of virocells, and *inuJ*, which encodes for enzymatic degradation of sucrose into glucose in *Chitinophagaceae*, potentially serving as a means of energy preservation to support viral propagation.

Distribution of AMGs in different members of a viral species

To understand how AMGs are distributed across viral populations, we analysed variation in AMG clusters within different members of a viral species. We found that ~30% of AMG clusters and species combinations had high presence ratios (defined as ratio of presence among all the members; top 75–100% quartile), indicating widespread distribution across viral species (Extended Data Fig. 3a). This pattern held steady regardless of species size.

Focusing on the largest species (4th quartile) with the highest AMG presence (75–100% quartile), we observed that high-occurrence AMG clusters had presence ratios above 95% and were more abundant, indicating that these clusters were consistently carried by multiple viruses across samples. Over the 20 year time series, these high-occurrence AMGs appeared consistently across different seasons (Extended Data Fig. 3b).

Seasonal patterns of virus–host dynamics and inter-viral competition

Combining two decades of relative abundance data revealed seasonal patterns of virus and host abundance. We examined three keystone microbial taxa: Cyanobacteria, Methanotrophs and *Nanopelagicales* (ultrasmall acI within Actinobacteriota). In our analysis, we considered the abundances of both AMG-containing and non-AMG-containing viruses (Fig. 4). Among the 13 examined AMG clusters, when the viral genome completeness was high (75–100%), the majority of species members contained the corresponding AMG cluster (>85%) (Fig. 4a). This underscores the efficacy of delineating AMG-containing species representatives to represent all AMG-containing viruses.

In three groups of Cyanobacteria, namely, *Planktothrix*, *Microcystis* and *Cyanobiaceae*, host and virus abundances were positively correlated (Fig. 4b). In addition, the peak abundance time points for these viruses (specifically, the non-AMG-containing fractions) and hosts showed consistent patterns over two decades. In the case of *Planktothrix*, the virus onset date lagged the host onset date by ~20 days, while the onset time points for viruses and hosts in *Microcystis* and *Cyanobiaceae* remained nearly identical. Previous research indicates host physiology and habitat controls influence viral progeny and suggests that fast-growing hosts could provide more resources for viral production[34,35]. Therefore, one plausible explanation is that the *Planktothrix* virus population shows an extended lag phase, allowing for substantial replication when hosts achieve notable abundance levels at the start of the annual growth cycle. This observation suggests a potential competition for the overlapping ecological niche, particularly concerning light and nutrient resources[36], in which both viruses and hosts actively participate. This dynamic pattern provides valuable insights for studying the mechanisms underpinning niche competition and temporal succession.

Similar to patterns observed in Cyanobacteria, four methanotroph genera also had positive host-to-virus abundance correlations and seasonal abundance patterns (Fig. 4c). It is worth noting that UBA10906 emerged as the most abundant genus, outcompeting the other three genera. Its peak abundance occurs later in the summer, extending into the fall and lasting longer compared with the other three genera. Specifically, we noted that the ratio of *Methylocystis* viruses containing *pmoC* (particulate methane monooxygenase subunit C) to their hosts was of a similar magnitude to the ratio of *Methylocystis* viruses lacking *pmoC* to hosts. By contrast, the comparisons involving the other three genera consistently showed that *pmoC*-containing viruses were less abundant by one order of magnitude than viruses lacking *pmoC*. Viral-encoded *pmoC* has the potential to augment aerobic methane oxidation[16]. An earlier study in soils showed that *Methylocystis* viruses were the most abundant viruses receiving the CH_4-derived carbon in soil microcosm incubations fuelled by ^{13}C-CH_4[37].

Similarly, two *Nanopelagicales* (acI group) genera showed a positive correlation between their host-to-virus abundance (Fig. 4d). The abundance of *Planktophila* during summer seasons is probably suppressed due to high concentrations of H_2O_2 produced by abiotic photochemical actions and biotic cyanobacterial and algal metabolisms which peak during this period[38]. This suppression leads to a noticeable decline in *Planktophila* populations. While catalases encoded by *katG* are vital for reducing H_2O_2 levels and stabilizing *Planktophila* growth[39], the low abundance of *katG*-containing viruses infecting *Planktophila* suggests they are insufficient in bolstering catalase activity to counteract the stress from elevated H_2O_2. Furthermore, haem is an essential cofactor in the generation of catalases. The *ahbD*-containing *Planktophila* virus abundance also represented a declining trend in late summer and fall (data not shown), indicating a constraint in virus-assisted haem synthesis for catalase production. Consequently, despite the presence of these viruses, the high H_2O_2 concentrations during summer likely contribute substantially to the observed decline in *Planktophila* abundance.

Among these three host groups with important biogeochemical roles, the ratios of AMG-containing viruses to hosts were consistently one to two orders of magnitude lower than those of non-AMG-containing viruses to hosts. Considering that most previous studies did not enumerate the abundance of AMG-containing viruses in nature, there might have been an excessive emphasis on the importance of viral AMGs associated with the metabolism of specific substrates (such as *pmoC* for methanotrophs in methane utilization) or enhancing rate-limiting enzymes (such as *psbA* for Cyanobacteria to optimize photosynthesis). Furthermore, we propose that the previous belief that AMG-containing viruses are more prevalent and important in the community primarily stems from isolated viruses and metagenomes from the open ocean and other marine environments. For instance, almost all Myoviruses and over half of Podoviruses infecting Cyanobacteria are believed to have *psbA* in their genomes, and 89% of recruited Cyanopodovirus scaffolds from Global Ocean Sampling Expedition (GOS) datasets contain *psbA* genes[40]. However, based on the findings of this study from a freshwater environment, observations of relative abundances suggest that non-AMG-containing viruses overwhelmingly prevailed and closely mirrored the seasonal fluctuations of their hosts. This implies that non-AMG-containing viruses make up the majority of viral communities. Overall, we propose that the magnitude of influence of AMGs on viral fitness, metabolism, ecosystem function, biogeochemistry and their adaptations to hijacking hosts likely needs re-evaluation in future research[41]. Furthermore, comprehensive studies are essential to explore the interactions between non-AMG-containing viruses and hosts.

https://doi.org/10.1038/s41564-024-01876-7

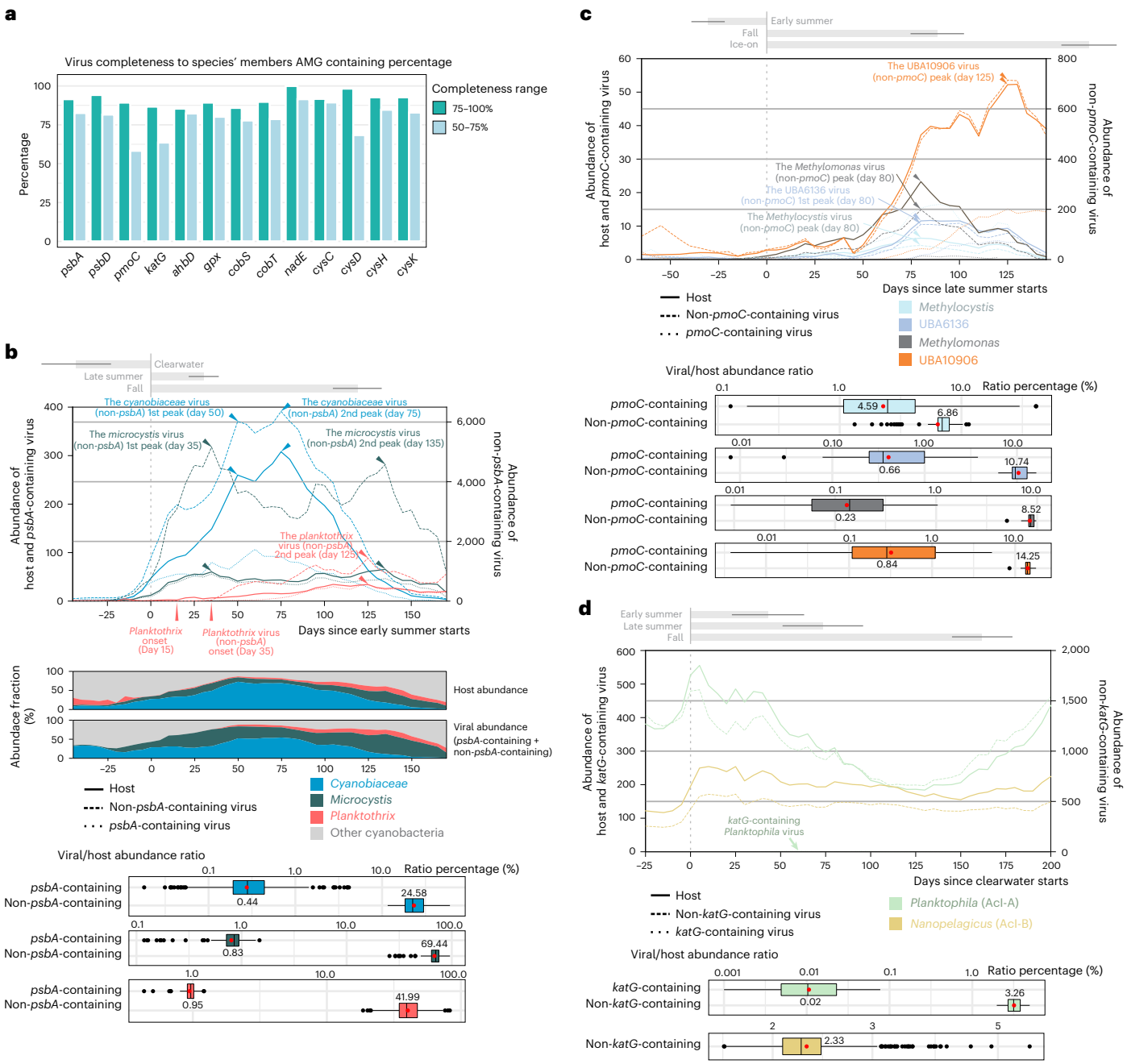


Fig. 4 | Viral–host correlation and seasonal change patterns for Cyanobacteria, methanotrophs and Nanopelagicales. a, Bar plot representing comparison of viral genome completeness to percentage of species' members that contain AMGs. Species with representative genomes containing the corresponding AMGs were considered. **b**, Seasonal change of viral and host abundances for Cyanobacteria groups. The first line chart depicts abundances of viruses and hosts of *Cyanobiaceae*, *Microcystis* and *Planktothrix*. Individual lines represent mean abundances derived from interpolated values calculated at intervals of 5 days. The second chart depicts abundance fractions of *Cyanobiaceae*, *Microcystis*, *Planktothrix* and other Cyanobacteria during seasonal change. The third box plot depicts viral/host abundance ratios of *Cyanobiaceae*, *Microcystis* and *Planktothrix* across all time points (with biological replicate numbers $n = 241, 330, 106$). **c**, Seasonal change of viral and host abundances for methanotroph genera. The first line chart depicts abundances of viruses and hosts of four methanotroph genera. The second box plot depicts viral/host abundance ratios of four methanotroph genera (with biological replicate numbers $n = 81, 50, 32, 75$). **d**, Seasonal change of viral and host abundances for *Nanopelagicales* genera. The first line chart depicts abundances of viruses and hosts of two *Nanopelagicales* genera. The second box plot depicts viral/host abundance ratios of two *Nanopelagicales* genera (with biological replicate numbers $n = 113, 615$). The bar plots on the top of each line chart of **b**, **c** and **d** indicate the means and standard deviations of the start day of corresponding seasons (with biological replicate number $n = 20$ from 20 years). The box plots within **b**, **c** and **d** show the median (central line), the mean (dark red dot with labelled values), the interquartile range (spanning from the 25th to the 75th percentiles), the whiskers and the outliers.

Temporally variable viruses have a high contribution to the AMG pool

To better understand the ecological and evolutionary roles of biogeochemically important viral populations, we analysed the coverage of viral species representatives. In line with seasonal patterns of AMG cluster abundance (Supplementary Dataset 1), viral species containing *psbA* peaked in late summer (Supplementary Table 7), leading us to select late summer as a representative time point for each year.

https://doi.org/10.1038/s41564-024-01876-7

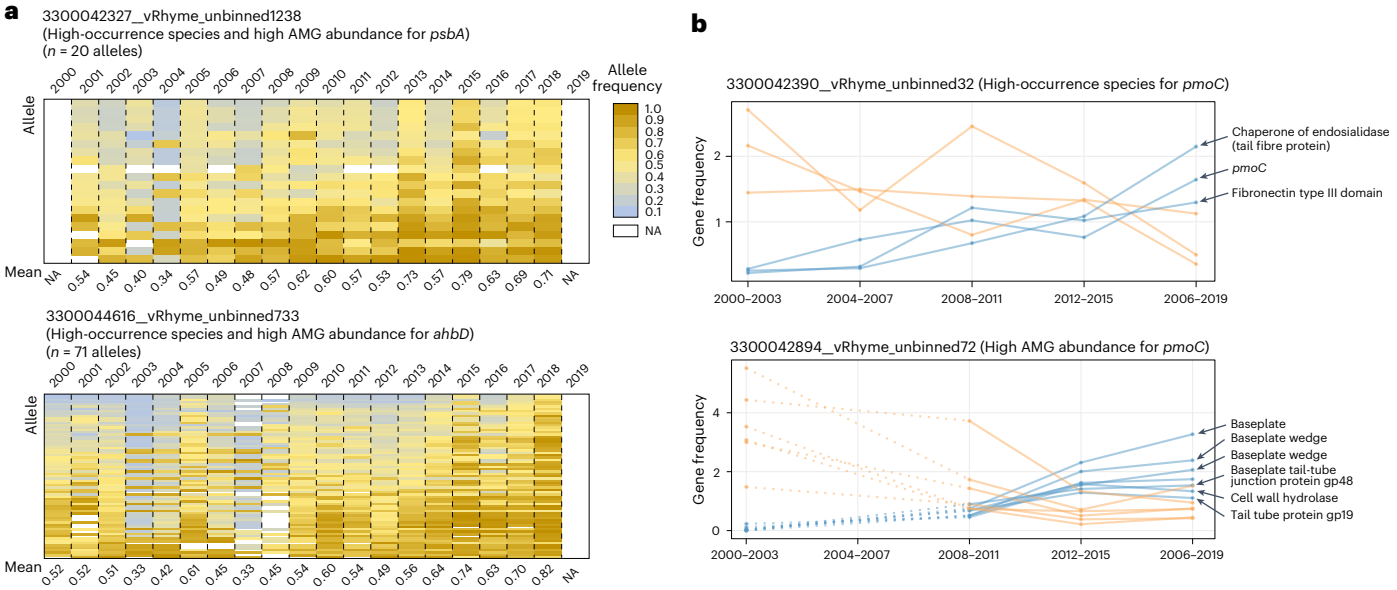


Fig. 5 | Microdiversity changes of persistent viral populations. a, Pattern of SNP allele frequency variation. SNP alleles are organized in ascending order based on the mean allele frequencies over 20 years, with each row denoting an individual SNP. The mean allele frequencies across the genome are labelled below each column accordingly. The allele frequency is depicted by the proportion of reads aligning to the reference allele—the predominant allele in the corresponding viral genome for 2018. **b**, Gene frequency change pattern. The *y* axis represents gene frequency, determined by dividing gene coverage by the average coverage of all other genes within the genome. Genes showing a mean frequency change of ≥1.0 between 2000–2003 and 2016–2019 are considered to have significantly increased or decreased in frequency. Intervals devoid of meaningful gene frequency values are represented with dash lines in the graph.

Time-series data revealed that four *psbA*-containing viral species persisted throughout 15 or more of the 20 years (Extended Data Fig. 5a). However, these persistent species were not always the most abundant in terms of AMG content. Only two of the four had high AMG abundance (≥10%). Similarly, *pmoC*-, *katG*- and *ahbD*-containing viral species with high occurrence differed from those with high AMG abundance (except for one *katG*-containing viral species and two *ahbD*-containing viral species) (Extended Data Fig. 5b–d). This indicates that persistent species containing AMGs are not typically the most abundant annually; instead, the pool of AMGs is driven by species that fluctuate year to year. As AMGs regulate key host functions, these findings suggest that temporally variable viral species may play an important role in critical biogeochemical processes, such as photosynthesis and methane oxidation.

Viral persistence is linked to evolutionary progressions and soft genome-wide sweeps

To examine intra-population diversity among persistent viral species, we studied viral species with AMGs encoding four important functions (*psbA* for photosynthesis, *pmoC* for methane oxidation, *katG* for reducing H_2O_2 stress and *ahbD* for haem synthesis) (Supplementary Table 7 and Supplementary Results). We studied the normalized abundance of viral scaffolds on a per-nucleotide basis (see details in Methods). In our analysis of the 471 metagenomes, four of the six *psbA*-containing viral species showed a positive correlation between species abundance and nucleotide diversity ($P < 0.05$) (Supplementary Table 8). Moreover, two *ahbD*-containing viral species characterized by high occurrence showed a positive correlation between species abundance and single-nucleotide polymorphism (SNP) density ($P < 0.05$).

We subsequently expanded our analysis to encompass all viral species containing AMGs. Of these, 221 out of 865 species with valid nucleotide diversity results and 262 out of 776 species with valid SNP density results showed significant positive correlations with viral abundance, respectively. By contrast, only 23 out of 865 species with valid nucleotide diversity results and 12 out of 776 species with valid SNP density results showed significant negative correlations with viral abundance, respectively (Supplementary Table 8). These findings suggest that viral intra-population diversity is mainly governed by the neutral theory[42], wherein an augmented population size generally leads to increased nucleotide diversity and SNP density. Positively selected genes within these persistent viral populations (Supplementary Table 9) encoded enzymes associated with purine biosynthesis, viral RNA synthesis, DNA repair, controlling of cellular and viral DNA and messenger RNA turnover, transcriptional regulation and bacterial cell wall penetration, as well as auxiliary metabolisms related to photosynthesis (*psbA*), haem synthesis (*ahbD*) and folate biosynthesis (*moaA*). Such findings suggest a mechanism of viral fitness selection associated with viral infection and host regulation, virion replication and host metabolism redirection or augmentation[2,12,43].

In certain viral species, whole-genome genetic heterogeneity gradually decreased (Fig. 5a), as evidenced by the fact that persistent *psbA*- and *ahbD*-containing viral species showed an increasing SNP allele frequency over time, as shown by linear regression (high regression slope) and Spearman's rank correlation tests (significant *P* value) (Supplementary Table 10). We examined the genes containing these increasing-frequency SNP alleles. Six of the nine viral genes were positively selected within these two vMAGs; three were annotated with important functions (*psbA* and *ahbD* for auxiliary metabolisms, restriction endonuclease type II-like genes for host genome degradation, nucleotide recycling for viral replication[44] and exclusion of superinfections[45]), contributing to viral fitness. This indicates that, while these high-occurrence viral species persist over time, their sub-populations have changed. Specifically, selection favoured some sub-populations with advantageous alleles. Nevertheless, the mean allele frequencies across genomes remain relatively low (~0.7), implying that either the genome-wide sweep remains ongoing or viral populations undergo a 'soft sweep', wherein selection favoured a few sub-populations from large, diverse populations[46–48]. Bacteriophages typically have higher genomic diversity and recombination rates than bacteria[49,50]. Due to the high microdiversity that existed before the start of this study, it will take a longer time for sub-populations with selection advantages to take over the population. Concurrently, an elevated recombination

https://doi.org/10.1038/s41564-024-01876-7

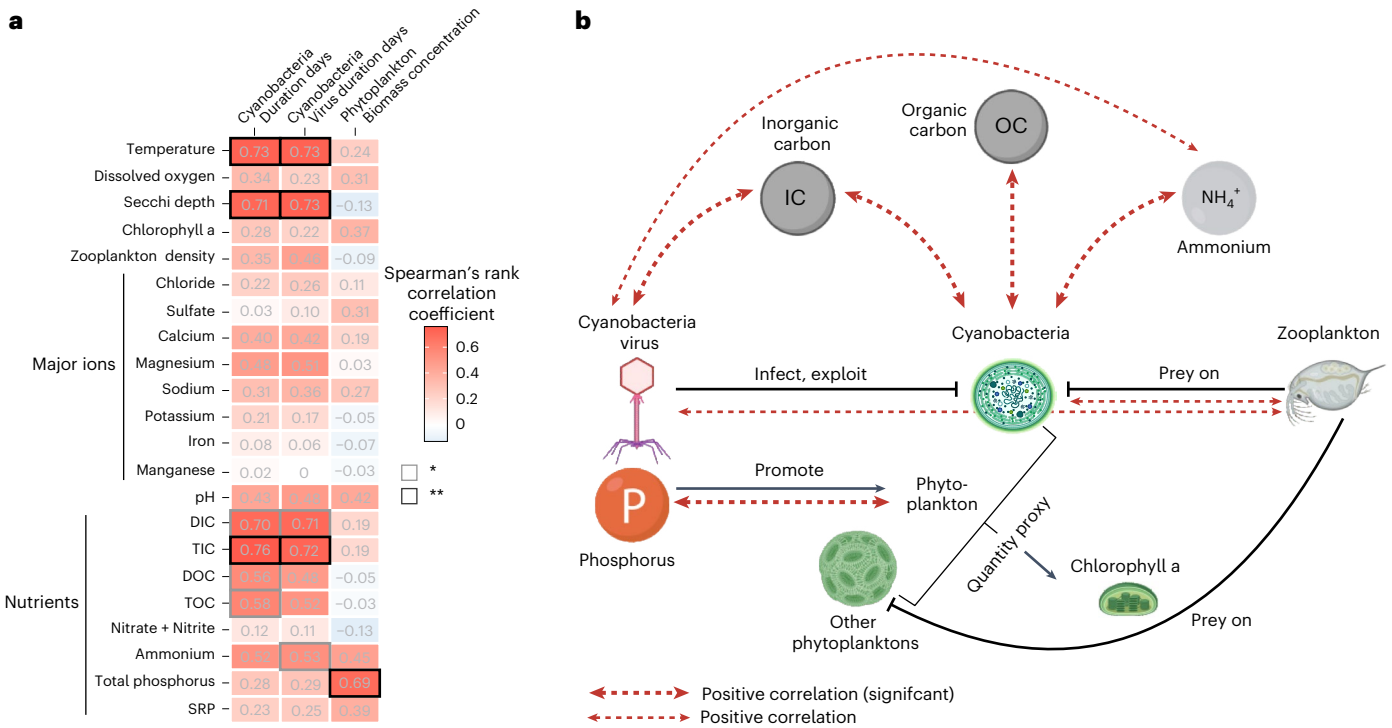


Fig. 6 | Correlations between viruses, hosts and environmental parameters across the time series. a, Heat map representing the Spearman's rank correlation between environmental parameters and cyanobacteria, cyanobacteria virus, phytoplankton biomass concentration and zooplankton density. Spearman's rank correlation coefficient was calculated to assess the relationship between two lists, with a two-tailed test applied to determine significance. Spearman's rank correlation coefficients are labelled in individual cells, and significance levels are indicated by grid borders, with black borders denoting $^{**}P < 0.01$ and grey borders denoting $^{*}P < 0.05$. IC, inorganic carbon; OC, organic carbon; DIC, dissolved inorganic carbon; TIC, total inorganic carbon; DOC, dissolved organic carbon; TOC, total organic carbon. **b**, Schematic diagram showing the interactive connections among the environment, viruses, bacteria, phytoplankton and zooplankton.

rate seems to counteract selection, promoting recombination among sub-populations with distinct micro-niches, thereby preserving genome-wide diversity.

In addition, for some viral populations, certain genes have either increased or decreased gene frequencies over time (Fig. 5b and Supplementary Table 11). The increasing-frequency gene repertoire encodes structural proteins, such as chaperones of endosialidase (tail fibre proteins for initial absorption of virus into the host[51]), baseplate, baseplate wedge and tail tube proteins; viral core function proteins, such as fibronectin type III containing protein (probably for virus–cell surface interaction[52]) and cell wall hydrolase (for host cell wall degradation and facilitating bacteriolysis and virion release[53]); and an AMG protein (PmoC). This indicates the importance of virus structural proteins, viral infection proteins and auxiliary metabolic proteins in strengthening viral fitness. This scenario suggests a similar genome-wide selection pattern in that certain sub-populations that harboured important functional genes in the lake before this study (before 2000) gradually became dominant in the populations from 2000 to 2019. Collectively, despite viral populations having a high level of diversity and rate of recombination, selections for genes with fitness advantages and genome-wide selections still play an important role in viral population dynamics.

Environmental constraints shape viral communities via top-down and bottom-up controls

Host dynamics are controlled by both top-down (for example, grazing by protists, viral lysis) and 'bottom-up' (for example, water temperature, nutrient concentrations) drivers[4]. We expect these dynamics to also manifest in measured viral abundances and potential viral roles. We focused on Cyanobacteria and their viruses to explore whether available limnological measurements could explain their dynamics. The number of duration days in which the Cyanobacteria and Cyanobacteria virus abundances were >20% of their peak abundances were related to the environmental parameters using Spearman's rank correlation test. The number of duration days should reflect an integrated influence of the environmental conditions during the summer season. As expected, water temperature and Secchi depth (a measure of water clarity) were positively correlated with both Cyanobacteria and their viruses (Fig. 6a and Supplementary Table 12).

These relationships reflect the intricate balance of both top-down and bottom-up factors that shape the environment, viruses, bacteria and predators within the ecosystem. Phosphorus, acting as a bottom-up factor, stimulates phytoplankton growth[54]. This increase in phytoplankton biomass forms the basis for further ecological interactions. Inorganic carbon serves as the primary carbon source for cyanobacteria through photoautotrophy, promoting both cyanobacteria and viral proliferation in a bottom-up manner. In addition, several cyanobacteria can assimilate organic carbon concurrently during photosynthesis, and the mixotrophic metabolism accelerates the growth of Cyanobacteria[55,56]. This aligns with the observed positive correlation between organic carbon and Cyanobacteria (Fig. 6b). Ammonium also correlated positively with cyanobacterial virus abundance. As the primary nitrogen source for assimilation by Cyanobacteria[57], it promotes Cyanobacteria growth, which in turn supports the proliferation of Cyanobacteria viruses. Conversely, reflecting a top-down control, as Cyanobacteria abundance increases, zooplankton density increases, showing a typical predator–prey dynamic. The increased cyanobacterial abundance provides more hosts for cyanobacteria viruses, which leads to the elevation of virus abundance as evident from the observed correlations in virus-to-host abundance. These top-down interactions are critical in regulating the populations within the ecosystem.

Discussion

Our study highlights the enormous volume of unknown viral diversity found in a single temperate freshwater lake and additionally suggests that other freshwater systems, such as tropical lakes, are likely purveyors of viruses playing important roles in nutrient and biogeochemical transformations that require further investigation.

Few studies have focused on the evolution and environmental analyses of viral population dynamics[9], especially over long timescales such as for a two-decade time-series study of the natural environment conducted here. In this study, persistently distributed viral populations of high occurrence underwent both positive gene selection and genome-wide selection. Three evolutionary processes were inferred: selection favoured genes associated with fitness, genomic heterogeneity decreased over time and sub-populations carrying certain genes became dominant. Similar to a lake green sulfur bacterial population in which SNP variations were slowly purged and some genes were either swept through or lost within the population over time[8], our study indicates the universality of evolutionary processes in both viruses and microorganisms. These processes can be jointly explained by the concept that some sub-populations with advantageous traits acquired through mutations or horizontal gene transfer outcompete others and become predominant in the observed populations[8,46,47], which appear to be 'stable' when only viewed from a macrodiversity perspective.

In the evolutionary arms race between viruses and their hosts, 'kill-the-winner' and other forms of dynamics frequently occur, causing fluctuations in the abundance of various viral strains[58]. Despite these fluctuations, certain viral species persist over extended periods and show high occurrence over time, indicating their evolutionary success in adapting to changing environmental conditions. These high-occurrence viral species may represent a 'royal family' viral species in the model used to explain the kill-the-winner dynamics[59], where certain sub-populations with enhanced viral fitness have descendants that become dominant in subsequent kill-the-winner cycles. It is probable that these high-occurrence viral species maintain a stable presence at the coarse diversity level while undergoing continuous genomic and physiological changes at the microdiversity level. For example, the selection of viral genes associated with resistance and counter-resistance results in enhanced bacterial cell wall penetration, initial absorption into the host cell and virus–cell surface interactions. Therefore, the sustained interactions and co-evolution of viruses and hosts over time suggest better adaptation of highly abundant viral species to local environmental conditions.

Concurrently, we identified environmental factors, such as inorganic carbon and ammonium, that might indirectly influence viral abundance through virus–host interactions. Our observations suggest a complex interplay of bottom-up controls, such as nutrient availability and primary production, and top-down controls such as predator–prey dynamics. Overall, our findings underscore the necessity for further research on viruses in microbiomes and ecosystems and for a holistic approach that places viral studies in the broader context of biodiversity, virus–host interactions and the physico-chemical constraints existing in natural environments.

Methods and materials

Samples

In this study, 471 water filter samples were collected from a pelagic integrated 12 m depth zone in Lake Mendota, Madison, WI, USA (GPS: 43.0995, −89.4045). Lake Mendota is a eutrophic freshwater lake located in Madison, WI (size, 39.4 km^2; average depth, 12.8 m; pH, 8.5) and an important component within the North Temperate Lakes Long Term Ecological Research project started in 1981. The samples were collected over several time points across different seasons each year, and the total sample period spanned 20 years (2000–2019). For each sample date, an approximately 250 ml integrated water sample was collected by filtering through a 0.2 µm pore size polyethersulfone Supor filter (Pall Corporation)[58]. Filters were stored at −80 °C for long-term storage. For omics sequencing, DNA extraction was conducted by using the FastDNA Spin Kit (MP Biomedicals) with minor modifications.

Environmental parameters

Environmental parameters were acquired from the sampling station at Lake Mendota (GPS: 43.0988, −89.4054) through the North Temperate Lakes Long Term Ecological Research program (https://lter.limnology.wisc.edu/) and are available through the Environmental Data Initiative (https://edirepository.org/), including water temperature[60–64], dissolved oxygen[60,61,64], Secchi depth[63,65,66], major ions[67], limnological nutrients[68], chlorophyll-a[69], phytoplankton biomass[70] and zooplankton density[71].

$$\mathrm{Days}_{>20\%\ \mathrm{peak\ abundance}} \sim ?\left(\frac{\mathrm{Days}_{\mathrm{Early\ summer}}}{\mathrm{Days}_{\mathrm{Early\ summer+Late\ summer}}} \times \mathrm{Env\ para}_{\mathrm{Early\ summer}} + \frac{\mathrm{Days}_{\mathrm{Late\ summer}}}{\mathrm{Days}_{\mathrm{Early\ summer+Late\ summer}}} \times \mathrm{Env\ para}_{\mathrm{Late\ summer}}\right) \times r \quad (1)$$

$$r = \frac{\mathrm{Days}_{\mathrm{Early\ summer+Late\ summer}}}{\mathrm{Days}_{\mathrm{Early\ summer+Late\ summer}}} \quad (2)$$

The two equations presented above show the method to assess potential correlations between the duration of days (Days) in which the abundance is maintained at >20% of the peak abundance (for both cyanobacteria and cyanobacteria viruses) and the average environmental parameters (Env para) throughout summer seasons. These environmental parameters were obtained from both Early summer and Late summer, and their significance was weighted based on the dates of each summer season. In equation (1), the symbol '~?' denotes evaluating the presence of potential correlations with significant support for the contents on both sides of the equation. In equation (2), 'r' represents the ratio used to standardize the summer dates for each year in relation to the mean summer dates. Duration days were determined by computing the abundance profile using an interpolation function for each year, with intervals of 5 days. The interpolation function was applied to abundance data within the time range of −45 to 160 days (since the start of early summer for each year). The years that could not meet the entire time range were excluded from the correlation analysis. The Spearman's rank correlation test with Fisher z-transformation was used, and the P value was provided. The mean value of each parameter for every season was calculated by averaging all the measurements within that specific season (some missing values were denoted as 'NA').

Metagenome sequencing and processing

Extracted DNA from 471 samples was submitted to the Department of Energy Joint Genome Institute (DOE JGI) (Walnut Creek, CA) for metagenomic sequencing. Illumina regular fragments with ~300 bp length were made for metagenome library construction; afterward, the high-throughput sequencing was conducted using the Illumina NovaSeq S4 (Illumina), yielding paired-end reads of 150 bp each and approximately 1.5×10^8 reads (accounting for both ends) per sample. The metagenome assembly was performed by metaSPAdes v3.14.1[72] and annotated by the IMG annotation pipeline (IMGAP) v5.0.20[73] implemented in the Integrated Microbial Genomes & Microbiomes system (https://img.jgi.doe.gov/m/).

Virus identification and genome binning

VIBRANT v1.2.1[74] was used to identify and annotate virus (bacteriophage) scaffolds from metagenomic assemblies with default settings. Only viral scaffolds (including both provirus and non-provirus) longer than 2,000 bp were used for downstream analysis. vRhyme v1.0.0[75] was used to reconstruct vMAGs from the identified viral scaffolds with default settings. The following four criteria were used to refine the best

vMAG (or bin) collection suggested by the default result of vRhyme: (1) Proviruses identified by VIBRANT were excluded from binning. (2) Two or more temperate (non-provirus) viral scaffolds cannot be in the same bin. (3) Viral scaffolds identified by CheckV v0.8.1[26] (database checkv-db-v0.6) as 'Complete' were excluded from binning. (4) The maximum number of bin redundancy should be ≤1. Any predicted bins that did not meet the above four criteria were split into individual viral scaffolds. CheckV was also used to estimate the vMAG quality with default settings. As CheckV can only process single-contig viral genomes, for each vMAG, we first linked vMAG scaffolds with 1,500 'N's to make temporary 'single-contig' viral genomes.

Virus clustering

We first clustered all viral genomes into families and genera using the gene sharing and amino acid identity (AAI) method[76]. Specifically, an all-vs-all DIAMOND BLASTP (v0.9.14.115) was performed for all virus genome protein sequences with the settings of '--evalue 1e-5 --max-target-seqs 10000 --query-cover 50 --subject-cover 50'. Then, the gene-sharing numbers between each pair of genomes, as well as the average amino acid identity of shared proteins, were parsed from DIAMOND BLASTP results. Edges (based on the minimum values of gene sharing and AAI) and nodes (viral genomes) were parsed accordingly for both families and genera, then filtered and subjected to MCL-based (v14.137) network clustering[77]. The edge filtering criteria and settings of the Markov clustering (MCL) inflation factor were adopted from a previous publication[76] (https://github.com/snayfach/MGV/tree/master/aai_cluster).

For non-singleton genera, we further clustered them into species. dRep v3.2.2[78] was used for dereplicating all viruses within each genus with the settings of '-l 2000 --ignoreGenomeQuality -pa 0.8 -sa 0.95 -nc 0.85 -comW 0 -conW 0 -strW 0 -N50W 0 -sizeW 1 -centW 0'. The resulting representatives (the best representative viral genomes picked according to genome length) together with singleton genera and individual viral genomes that were not assigned to any genera were the final collection of species.

Taxonomic classification

We combined three approaches to conduct taxonomic classification. For the first and second approaches, we adopted the procedure as described in the instructions as suggested previously[3]. For searching against National Center for Biotechnology Information (NCBI) Reference Sequence (RefSeq) viral proteins, DIAMOND BLASTP v0.9.14.115 was used to BLAST against NCBI RefSeq viral proteins (2023-01-13 release)[79] using all the TYMEFLIES viral proteins with settings of 'blastp -evalue 1e-5 --query-cover 50 --subject-cover 50 -k 10000'. For any viral genome with ≥30% of proteins having significant hits to NCBI RefSeq viral proteins, a ≥50% majority taxonomy was assigned based on the taxonomy (using reformatted International Committee on Taxonomy of Viruses (ICTV) taxonomy with eight ranks) of the best hits of individual proteins. For the viral orthologous groups (VOG) marker hidden Markov model (HMM) searching approach, hmmsearch (HMMER v3.1b2[80]) was used to search against VOG database v97 (2021-04-19 release, http://vogdb.org) using all the TYMEFLIES viral proteins. Only 587 VOG marker HMM profiles were used as the reference for taxonomic classification[3]. The criteria for positive hits were score ≥40 and $E < 1 \times 10^{-5}$. The taxonomy of a viral genome was obtained based on individual markers detected using a simple plurality rule if multiple hits were present. For the third approach, geNomad v1.5.1 was used to annotate viruses and get taxonomy from the annotation result using default settings[81]. For each vMAG, we first linked vMAG scaffolds with 1,000 Ns to make temporary single-contig viral genomes to meet the input requirement of geNomad.

If a viral genome was not assigned taxonomy by any of the above three approaches while it was placed in a genus with the other member(s) assigned using the NCBI RefSeq viral protein searching approach, the lowest common ancestor (LCA) of this genus was used as the taxonomic classification (in this case, the deepest LCA rank is limited to the genus level). All the aforementioned taxonomic classification approaches were labelled in accordance with the taxonomy obtained. If overlaps occurred, the top-order approach was given the highest priority.

Host prediction

We used three approaches for predicting the hosts of viruses. For the first approach, iPHoP v1.2.0 (re. [82]) was used to predict the host from all viruses (N-linked sequences to make temporary single-contig viral genomes) using the default settings. The TYMEFLIES species representative metagenome-assembled genomes (2,855 MAGs total, dereplicated by dRep with 96% sequence identity cut-off) were added to the default iPHoP database 'Sept_21_pub'[25]. The host prediction to genome results (based on host-based tools, including 'blast', 'CRISPR', and 'iPHoP-RF' results) were finally assigned with the following rules: (1) If blast or CRISPR results were obtained for one virus genome, the result with the highest confidence score was assigned as the final result. (2) If only iPHoP-RF results were obtained for one virus genome, the result with the highest confidence score was assigned as the final result. For the second approach, we predicted the viral host based on the AMG identity match between AMGs and microbial counterpart genes. The viral AMGs were identified, filtered and summarized (for details, refer to the following section). The counterpart genes in prokaryotic hosts (with the same KEGG Orthology) were parsed out from all TYMEFLIES MAGs. For any viruses that had AMGs connected to their counterparts in TYMEFLIES MAGs (potential viral contigs were excluded as mentioned above) based on the cut-off of sequence identity ≥60% (with DIAMOND BLASTP options of '--query-cover 70 --subject-cover 70'), the AMG viral–host connections were established. Similarly, such multiple viral–host connections based on AMGs for a virus were aggregated, and the lowest rank with ≥80% consensus was determined as the host taxonomy. For the third approach, a TYMEFLIES MAG that contained the scaffold where the provirus was located was determined as the host.

If a viral genome was not assigned a host by any of the three approaches while it was placed in a species with the other member(s) assigned, the LCA of the host taxonomy for this species was used. Note that only the provirus, AMG host prediction and blast or CRISPR-based iPHoP results were used to get host predictions from other species members. All the results were labelled with corresponding host taxonomy prediction approaches. The overlapped host taxonomies were resolved based on the following order of priority: (1) provirus within a host genome; (2) blast-based iPHoP result; (3) CRISPR-based iPHoP result; (4) AMG match to host genome; (5) iPHoP-RF result; (6) derived from species host taxonomy.

AMG summary

The AMGs identified by VIBRANT in the above section were first filtered according to the following criteria: (1) Edge-located AMGs (AMGs located at either end of a scaffold) were filtered. (2) AMGs that had any KEGG v-score or Pfam v-score (assigned by VIBRANT) ≥1 were filtered. (3) AMGs with flanking genes (four genes on either the upstream or downstream sites) having a KEGG v-score <0.25 were filtered. (4) AMGs with annotation by COG category as 'T' or 'B' were filtered. The filtered AMGs were then summarized by adding the information, including date and season, KO hit and name, Pfam hit and name and KEGG metabolism, pathway and module. The AMG cluster occurrence (the number of metagenomes in which an AMG cluster can be found) and abundance (the mean normalized abundance of AMG cluster containing viruses in the metagenomes that this AMG cluster can be found) were obtained by summarizing AMG cluster containing viruses and were used to make scatter plots to find potential relationships using R v4.1.3 (R library 'ggpmisc').

The summary of AMG cluster abundance for each season (a season from 20 years combined) and each year-season (a season from each year; for example, '2000-Spring') was conducted using the following steps: (1) Calculate AMG cluster abundance (normalized by 100 million reads per metagenome) in each metagenome. (2) Calculate AMG cluster abundance for each season by adding up AMG cluster abundances from all metagenomes of this season (the AMG cluster abundance was normalized by the number of metagenomes in this season). (3) Calculate AMG cluster abundances for each year-season by adding up AMG cluster abundances from all metagenomes of this year-season (the AMG cluster abundance was normalized by the number of metagenomes in this year-season). (4) Generate AMG cluster abundance trend plots for each season (20 years combined) and each year-season (resulting in 20 facets for all 20 years) using R (R library 'ggpubr').

AMG cluster variation

In non-singleton species, we investigated the AMG cluster variation by calculating the presence ratio of the AMG clusters across all the viral members within the species. To investigate how the viral genome completeness influences AMG cluster variation within a viral species, we selected several important AMG clusters for analysis. Viral genomes were categorized into five completeness levels: '75–100% complete', '50–75% complete', '25–50% complete', '0–25% complete' and NA. For the species with its species representative genome containing a specific AMG cluster, we calculated the AMG cluster containing ratio (as a percentage) by dividing the number of AMG cluster containing viral genomes by the total number of viral genomes within each completeness category. The results that represent the relation of virus completeness to the AMG cluster containing percentage for species' members across all selected AMG clusters were plotted using bar plots in R (R library 'ggplot2').

The influence of species size (number of viral genomes in the species) on the AMG cluster variation was analysed by dividing the combinations of AMG cluster and species into four quartiles according to the species size. The combination of AMG cluster and species was used for AMG cluster variation analysis because some species can contain multiple AMG clusters. The mean AMG cluster presence ratio for each AMG cluster from the AMG cluster and species combinations of the first quartile (75–100%) of AMG cluster presence ratio category (the highest presence ratio) with the species size in the fourth quartile (the largest species size) was calculated. It was then plotted against the AMG cluster count fraction (the percentage of occurrences of a single AMG cluster among all AMG clusters within a species) to illustrate the relationship. The presence tables of AMG clusters for each season (20 years combined) were obtained for individual AMG clusters. They were compared with the available metagenomes for each season. The percentage of AMG cluster containing metagenome number over the total metagenome number in each season was calculated for each high-occurrence AMG cluster (distributed >400 metagenomes).

AMG cluster carrying viruses and host diversity

To get the alpha diversity of viruses and their hosts for each AMG cluster, we used the family-level taxonomy. Viruses with uninformative assignments (for example, 'Unclassified', NA;NA, and 'o__;f__') were excluded. To evenly reflect alpha diversity, 100 viruses with informative family assignment and 25 viruses with informative host family assignment were randomly selected for each AMG cluster. Any AMG clusters that could not meet the required number of viruses were excluded. Alpha diversities (represented by Simpson indices) of viruses and viral hosts were obtained by R (R library 'vegan'), and they were plotted against AMG cluster occurrence to find potential relationships (R library ggpmisc).

AMG coverage ratio, viral genome abundance and MAG abundance calculation

The mapping reference for viral abundance calculation was the collection of viral species representative genomes. These genomes were also the longest among species members. The AMG counterpart gene-containing microbial scaffolds from all metagenomes were also included to avoid potential mis-mapping of microbial reads to viral AMGs. The mapping process was conducted by Bowtie 2 v2.4.5 using all metagenomic reads with default settings. The resulting bam files were subjected to viral abundance and microdiversity analysis by MetaPop v0.0.60[83] using the settings of '--id_min 93 --snp_scale both' (gene files from the above VIBRANT analysis were used in place of self-annotation by MetaPop; modifications were made to the gene files to adapt them to MetaPop requirements).

A custom script 'cov_by_region.py' was used to parse the site-specific depth file (within '04.Depth_per_Pos' directory of MetaPop result from the above section) to get the AMG coverage and viral scaffold coverage (excluding all the AMG regions). We obtained the normalized viral genome coverage by first calculating the average of all its scaffold coverage values (excluding all the AMG regions) and then normalizing it by setting each metagenome read number as 100 million. Note that all scaffold coverages from a viral genome needed to pass the cut-off of ≥0.01; otherwise, we assigned this viral genome as 'absent'.

After summarizing viral genome coverages across all the metagenomes, we set a custom viral genome presence cut-off as follows: coverage ≥0.33 and breadth ≥50%. Then, based on these 'present' viral genomes, we obtained the corresponding viral genome coverage (referred to as 'abundance') and AMG coverage values across all the metagenomes, as well as the AMG coverage ratios (AMG coverage divided by its corresponding viral genome coverage). Using the same criteria for scaffold coverage cut-off and viral genome coverage and breadth cut-offs, we calculated the viral genome abundance for the other non-AMG-containing viruses.

TYMEFLIES species representative MAGs were used as the mapping reference for conducting metagenomic read mapping using Bowtie 2 with default settings. CoverM v0.7.0 was used to calculate contig abundance using the settings of '--min-read-percent-identity 93 -m metabat'. The MAG was assigned as present in each metagenome with a breadth cut-off of 10%. The MAG abundance was calculated by computing the average contig abundance using the ratios of contig length to genome length. The MAG taxa abundance (at the family level) for each season was summarized using similar methods described above (refer to the second paragraph of the section 'AMG summary').

Virus composition pattern analysis

The abundance of viruses at the family level across all metagenomes was summarized. Subsequently, we generated non-metric multidimensional scaling plots using R (R library vegan and ggplot2). These plots represent the ordination of metagenomes based on pairwise distances of virus composition among all the metagenomes. According to the metagenome-to-season corresponding relationship, the analysis of similarity (ANOSIM) test was conducted to inspect whether there was a statistical difference among metagenomes from different seasons using R (R library vegan) with options of 'distance = 'bray', permutations = 9999'.

Virus and host association analysis

For the calculation of Cyanobacteria virus and host abundances, we mainly focused on the three Cyanobacteria groups: *Cyanobiaceae*, *Microcystis* and *Planktothrix*. Within each group, we computed the abundances of MAGs, *psbA*-containing viral genomes and non-*psbA*-containing viral genomes, specifically from the 0 day of early summer for each year. Subsequently, we used an interpolation function to generate abundance profiles for each year, with 5 day intervals. To plot the mean curves for virus and host abundances, we initially obtained the mean values of abundance percentages (normalized by the highest abundance within 1 year) for each time point. We then multiplied these mean abundance percentages by the highest abundance within that respective year. This approach was implemented to

mitigate the impact of substantial abundance fluctuations from year to year. The mean value for each time point was calculated based on valid abundances from at least 3 years. Subsequently, for each of the Cyanobacteria groups, we plotted the mean abundance curves for all viruses and hosts in a single line chart frame using Python 3 (Python library 'matplotlib').

For the calculation of methanotroph virus and host abundances, we mainly focused on the four methanotroph genera: *Methylocystis*, UBA6136, *Methylomonas* and UBA10906 (Supplementary Fig. 1). Similarly, within each genus, we computed the abundances of MAGs, *pmoC*-containing viral genomes and non-*pmoC*-containing viral genomes, specifically from the 0 day of late summer for each year. The remaining methods were consistent with those described in the preceding paragraph to plot the mean abundance curves for all viruses and hosts. For the calculation of *Nanopelagicales* virus and host abundances, we mainly focused on two genera: *Planktophila* and *Nanopelagicus*. Similarly, within each genus, we computed the abundances of MAGs, *katG*-containing viral genomes and non-*katG*-containing viral genomes, specifically from the 0 day of clearwater for each year. The remaining methods were consistent with those described in the preceding paragraph to plot the mean abundance curves for all viruses and hosts.

To calculate the viral-to-host abundance ratios for each group mentioned in the previous paragraphs, we considered pairs of virus and host abundances that met specific criteria. Specifically, at each time point, both the virus and host abundance percentages were required to be non-null (not 'nan') and greater than 10% to be considered valid pairs. Subsequently, box plots were created to show the viral-to-host abundance ratio distribution for each group using R (R library 'ggplots2' and 'scales') with the ratio range displayed with log-transformed values.

Microdiversity analysis

The microdiversity parameters were parsed based on the results of MetaPop (the 'Microdiversity' folder). Only viral scaffolds that passed the requirement of microdiversity calculation (breadth ≥70% and depth ≥10) were taken into consideration. Similar to the methods described in the above sections, the following microdiversity parameters for each metagenome, each year-season, each season and/or each year were calculated and summarized accordingly: nucleotide diversity (pi) for viral genome and viral genes, SNP density for viral genomes, rates of non-synonymous (pN) and synonymous (pS) polymorphism (pN/pS) for viral genes and fixation index (F_{ST}) for viral scaffolds. The correlations between viral genome abundance with nucleotide diversity and SNP density were calculated by Python 3 using Spearman's rank correlation test with the *P* value provided.

To investigate the populational genetic alteration between summer and winter (Late summer versus Ice-on), and between the beginning and ending years (2000–2003 versus 2016–2019), we conducted similar MetaPop analyses by aggregating the relevant metagenomic reads. Likewise, we calculated and summarized the four microdiversity parameters accordingly.

Yearly SNP allele frequency was calculated by parsing SNPs across the full length of the viral genome. The 'reference' alleles were chosen to be the predominant alleles of viral genomes in 2018. The choice was made because 2018 has the most metagenomes ($n = 44$) in the latter years, and a yearly changing trend from the beginning to the ending years can be depicted simply. SNP allele frequency was the percentage of reads matching the reference allele at each SNP locus.

Yearly gene frequency was calculated to reflect the gene relative abundance change in the viral species population along the time series. Gene frequency was estimated as the coverage of each gene divided by the mean coverage of all other genes in the genome. To set the detection limit for genome coverage, the mean coverage of all genes in the genome was required to be ≥5. Genes of length <450 bp were excluded from the analysis. In addition, to avoid the coverage variation influenced by the 'all-to-all' read mapping method (the default setting of Bowtie 2), the positions within the first and last 150 bp of a scaffold were excluded from coverage calculations. To get a statistically meaningful gene frequency, there was another requirement that the gene number in a genome with a valid coverage (not NA) should be over 50% of the total gene number. A gene frequency of 1.0 indicates that, statistically compared with the other genes in the genome, each virus in the population encodes one copy of the gene. Gene frequencies were considered significantly increased or decreased within a population if the change in gene frequency was ≥1.0. The yearly changing trend of gene frequency was fitted to the linear regression by Python 3.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The metagenomic datasets (including assemblies and raw reads) are all available under JGI Proposal ID 504350 at the platform of the Integrated Microbial Genomes & Microbiomes system (https://img.jgi.doe.gov/m/). The retrieved viral genomes were deposited in NCBI Bioproject PRJNA1130067. At the same time, the TYMEFLIES viral genomes and related properties, including annotations for viral proteins, taxonomic classification, host prediction and virus clustering results, are available via Figshare at https://figshare.com/articles/dataset/TYMEFLIES_vMAGs_and_related_properties/24915750 (ref. 84). The raw environmental parameter spreadsheets are available in the Environmental Data Initiative (https://edirepository.org/) database.

Code availability

Codes used in this project are available via GitHub at https://github.com/AnantharamanLab/TYMEFLIES_Viral.

References

1. Tran, P. Q. & Anantharaman, K. Biogeochemistry goes viral: towards a multifaceted approach to study viruses and biogeochemical cycling. *mSystems* **6**, e01138–21 (2021).
2. Rosenwasser, S., Ziv, C., van Creveld, S. G. & Vardi, A. Virocell metabolism: metabolic innovations during host–virus interactions in the ocean. *Trends Microbiol.* **24**, 821–832 (2016).
3. Camargo, A. P. et al. IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. *Nucleic Acids Res.* **51**, D733–D743 (2023).
4. McMahon, K. D. & Newton, R. J. Pelagic bacteria, archaea, and viruses. in *Wetzel's Limnology* 4th edn (eds Jones, I. D. & Smol, J. P.) Ch. 23, 705–757 (Academic, 2024).
5. Karl, D. M. & Church, M. J. Microbial oceanography and the Hawaii Ocean Time-series programme. *Nat. Rev. Microbiol.* **12**, 699–713 (2014).
6. Steinberg, D. K. et al. Overview of the US JGOFS Bermuda Atlantic Time-series Study (BATS): a decade-scale look at ocean biology and biogeochemistry. *Deep Sea Res. II* **48**, 1405–1447 (2001).
7. Shade, A. et al. Interannual dynamics and phenology of bacterial communities in a eutrophic lake. *Limnol. Oceanogr.* **52**, 487–494 (2007).
8. Bendall, M. L. et al. Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. *ISME J.* **10**, 1589–1601 (2016).
9. Roux, S. et al. Ecogenomics of virophages and their giant virus hosts assessed through time series metagenomics. *Nat. Commun.* **8**, 858 (2017).
10. Bragg, J. G. & Chisholm, S. W. Modeling the fitness consequences of a cyanophage-encoded photosynthesis gene. *PLoS ONE* **3**, e3550 (2008).

11. Mann, N. H., Cook, A., Millard, A., Bailey, S. & Clokie, M. Bacterial photosynthesis genes in a virus. *Nature* **424**, 741 (2003).
12. Kieft, K. et al. Ecology of inorganic sulfur auxiliary metabolism in widespread bacteriophages. *Nat. Commun.* **12**, 1–16 (2021).
13. Puxty, R. J., Evans, D. J., Millard, A. D. & Scanlan, D. J. Energy limitation of cyanophage development: implications for marine carbon cycling. *ISME J.* **12**, 1273–1286 (2018).
14. Lindell, D., Jaffe, J. D., Johnson, Z. I., Church, G. M. & Chisholm, S. W. Photosynthesis genes in marine viruses yield proteins during host infection. *Nature* **438**, 86–89 (2005).
15. Hellweger, F. L. Carrying photosynthesis genes increases ecological fitness of cyanophage in silico. *Environ. Microbiol.* **11**, 1386–1394 (2009).
16. Chen, L.-X. et al. Large freshwater phages with the potential to augment aerobic methane oxidation. *Nat. Microbiol.* **5**, 1504–1515 (2020).
17. Lindell, D. et al. Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. *Proc. Natl Acad. Sci. USA* **101**, 11013–11018 (2004).
18. Ruiz-Perez, C. A., Tsementzi, D., Hatt, J. K., Sullivan, M. B. & Konstantinidis, K. T. Prevalence of viral photosynthesis genes along a freshwater to saltwater transect in Southeast USA. *Environ. Microbiol. Rep.* **11**, 672–689 (2019).
19. Sullivan, M. B. et al. Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. *PLoS Biol.* **4**, e234 (2006).
20. Roux, S. et al. Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell- and meta-genomics. *eLife* **3**, e03125 (2014).
21. Anantharaman, K. et al. Sulfur oxidation genes in diverse deep-sea viruses. *Science* **344**, 757–760 (2014).
22. Ahlgren, N. A., Fuchsman, C. A., Rocap, G. & Fuhrman, J. A. Discovery of several novel, widespread, and ecologically distinct marine Thaumarchaeota viruses that encode amoC nitrification genes. *ISME J.* **13**, 618–631 (2019).
23. Cassman, N. et al. Oxygen minimum zones harbour novel viral communities with low diversity. *Environ. Microbiol.* **14**, 3043–3065 (2012).
24. Rohwer, R. R., Hale, R. J., Vander Zanden, M. J., Miller, T. R. & McMahon, K. D. Species invasions shift microbial phenology in a two-decade freshwater time series. *Proc. Natl Acad. Sci. USA* **120**, e2211796120 (2023).
25. Rohwer, R. R. et al. Two decades of bacterial ecology and evolution in a freshwater lake. *Nat. Microbiol.* https://doi.org/10.1038/s41564-024-01888-3 (2025).
26. Nayfach, S. et al. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **39**, 578–585 (2021).
27. Sharon, I. et al. Comparative metagenomics of microbial traits within oceanic viral communities. *ISME J.* **5**, 1178–1190 (2011).
28. Heyerhoff, B., Engelen, B. & Bunse, C. Auxiliary metabolic gene functions in pelagic and benthic viruses of the Baltic Sea. *Front. Microbiol.* **13**, 863620 (2022).
29. Kieft, K. et al. Virus-associated organosulfur metabolism in human and environmental systems. *Cell Rep.* **36**, 109471 (2021).
30. Nguyen, A. A. et al. CpeT is the phycoerythrobilin lyase for Cys-165 on β-phycoerythrin from *Fremyella diplosiphon* and the chaperone-like protein CpeZ greatly improves its activity. *Biochim. Biophys. Acta* **1861**, 148284 (2020).
31. Puxty, R. J., Millard, A. D., Evans, D. J. & Scanlan, D. J. Viruses inhibit CO_2 fixation in the most abundant phototrophs on earth. *Curr. Biol.* **26**, 1585–1589 (2016).
32. Thompson, L. R. et al. Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. *Proc. Natl Acad. Sci. USA* **108**, E757–E764 (2011).
33. Sachindra, N. M. et al. Radical scavenging and singlet oxygen quenching activity of marine carotenoid fucoxanthin and its metabolites. *J. Agric. Food Chem.* **55**, 8516–8522 (2007).
34. Parada, V., Herndl, G. J. & Weinbauer, M. G. Viral burst size of heterotrophic prokaryotes in aquatic systems. *J. Mar. Biol. Assoc. UK* **86**, 613–621 (2006).
35. Middelboe, M. Bacterial growth rate and marine virus–host dynamics. *Microb. Ecol.* **40**, 114–124 (2000).
36. Burson, A., Stomp, M., Greenwell, E., Grosse, J. & Huisman, J. Competition for nutrients and light: testing advances in resource competition with a natural phytoplankton community. *Ecology* **99**, 1108–1118 (2018).
37. Lee, S. et al. Methane-derived carbon flows into host–virus networks at different trophic levels in soil. *Proc. Natl Acad. Sci. USA* **118**, e2105124118 (2021).
38. Yoon, H., Kim, H.-C. & Kim, S. Long-term seasonal and temporal changes of hydrogen peroxide from cyanobacterial blooms in fresh waters. *J. Environ. Manage.* **298**, 113515 (2021).
39. Kim, S., Kang, I., Seo, J.-H. & Cho, J.-C. Culturing the ubiquitous freshwater actinobacterial acI lineage by supplying a biochemical 'helper' catalase. *ISME J.* **13**, 2252–2263 (2019).
40. Zheng, Q., Jiao, N., Zhang, R., Chen, F. & Suttle, C. A. Prevalence of *psbA*-containing cyanobacterial podoviruses in the ocean. *Sci. Rep.* **3**, 3207 (2013).
41. Warwick-Dugdale, J., Buchholz, H. H., Allen, M. J. & Temperton, B. Host-hijacking and planktonic piracy: how phages command the microbial high seas. *Virol. J.* **16**, 15 (2019).
42. Nei, M., Suzuki, Y. & Nozawa, M. The neutral theory of molecular evolution in the genomic era. *Annu. Rev. Genomics Hum. Genet.* **11**, 265–289 (2010).
43. Howard-Varona, C. et al. Phage-specific metabolic reprogramming of virocells. *ISME J.* https://doi.org/10.1038/s41396-019-0580-z (2020).
44. Agarkova, I. V., Dunigan, D. D. & Van Etten, J. L. Virion-associated restriction endonucleases of chloroviruses. *J. Virol.* **80**, 8114–8123 (2006).
45. Jeudy, S. et al. The DNA methylation landscape of giant viruses. *Nat. Commun.* **11**, 2657 (2020).
46. Cohan, F. M. Bacterial species and speciation. *Syst. Biol.* **50**, 513–524 (2001).
47. Cohan, F. M. & Perry, E. B. A systematics for discovering the fundamental units of bacterial diversity. *Curr. Biol.* **17**, R373–R386 (2007).
48. Shapiro, B. J. & Polz, M. F. Microbial speciation. *Cold Spring Harb. Perspect. Biol.* **7**, a018143 (2015).
49. Hatfull, G. F. Bacteriophage genomics. *Curr. Opin. Microbiol.* **11**, 447–453 (2008).
50. Mavrich, T. N. & Hatfull, G. F. Bacteriophage evolution differs by host, lifestyle and genome. *Nat. Microbiol.* **2**, 17112 (2017).
51. Schulz, E. C. et al. Structural basis for the recognition and cleavage of polysialic acid by the bacteriophage K1F tailspike protein EndoNF. *J. Mol. Biol.* **397**, 341–351 (2010).
52. Bork, P. & Doolittle, R. F. Proposed acquisition of an animal protein domain by bacteria. *Proc. Natl Acad. Sci. USA* **89**, 8990–8994 (1992).
53. Parisien, A., Allain, B., Zhang, J., Mandeville, R. & Lan, C. Q. Novel alternatives to antibiotics: bacteriophages, bacterial cell wall hydrolases, and antimicrobial peptides. *J. Appl. Microbiol.* **104**, 1–13 (2008).
54. Ahern, K. S., Ahern, C. R. & Udy, J. W. In situ field experiment shows *Lyngbya majuscula* (cyanobacterium) growth stimulated by added iron, phosphorus and nitrogen. *Harmful Algae* **7**, 389–404 (2008).
55. Kang, R. et al. Interactions between organic and inorganic carbon sources during mixotrophic cultivation of *Synechococcus* sp. *Biotechnol. Lett.* **26**, 1429–1432 (2004).

56. Muñoz-Marín, M. D. C., López-Lozano, A., Moreno-Cabezuelo, J. Á., Díez, J. & García-Fernández, J. M. Mixotrophy in cyanobacteria. *Curr. Opin. Microbiol.* **78**, 102432 (2024).
57. Flores, E. & Herrero, A. Nitrogen assimilation and nitrogen control in cyanobacteria. *Biochem. Soc. Trans.* **33**, 164–167 (2005).
58. Rohwer, R. R. & McMahon, K. D. Lake iTag measurements over nineteen years, introducing the limony dataset. Preprint at *bioRxiv* https://doi.org/10.1101/2022.08.04.502869 (2022).
59. Breitbart, M., Bonnain, C., Malki, K. & Sawaya, N. A. Phage puppet masters of the marine microbial realm. *Nat. Microbiol.* **3**, 754–766 (2018).
60. Rohwer, R. R. & McMahon, K. D. Lake Mendota Microbial Observatory temperature, dissolved oxygen, pH, and conductivity data, 2006-present. *Environmental Data Initiative* https://doi.org/10.6073/PASTA/7E533C197ED8EBD27777A89A2C8D7DFE (2022).
61. Magnuson, J. J., Carpenter, S. R. & Stanley, E. H. North temperate lakes LTER: physical limnology of primary study lakes 1981 - current. *Environmental Data Initiative* https://doi.org/10.6073/PASTA/316203040EA1B8ECE89673985AB431B7 (2021).
62. Magnuson, J., Carpenter, S. & Stanley, E. North temperate lakes LTER: high frequency water temperature data - Lake Mendota Buoy 2006 - current. *Environmental Data Initiative* https://doi.org/10.6073/PASTA/8CEFF296AD68FA8DA6787076E0A5D992 (2020).
63. Robertson, D. Lake Mendota water temperature secchi depth snow depth ice thickness and meterological conditions 1894 - 2007. *Environmental Data Initiative* https://doi.org/10.6073/PASTA/F20F9A644BD12E4B80CB288F1812C935 (2016).
64. Magnuson, J. J., Carpenter, S. R. & Stanley, E. H. Lake Mendota multiparameter sonde profiles: 2017 - current. *Environmental Data Initiative* https://doi.org/10.6073/PASTA/5F15BF453851987FC030B2F07A110B21 (2021).
65. Rohwer, R. R. & McMahon, K. D. Lake Mendota microbial observatory secchi disk measurements 2012-present. *Environmental Data Initiative* https://doi.org/10.6073/PASTA/3B650E19D28CBC7B9ED631F0A7878033 (2022).
66. Magnuson, J. J., Carpenter, S. R. & Stanley, E. H. North temperate lakes LTER: secchi disk depth; other auxiliary base crew sample data 1981 - current. *Environmental Data Initiative* https://doi.org/10.6073/PASTA/26FA98B39F9758FDA2109021F5B88076 (2021).
67. Magnuson, J. J., Carpenter, S. R. & Stanley, E. H. North temperate lakes LTER: chemical limnology of primary study lakes: major ions 1981 - current. *Environmental Data Initiative* https://doi.org/10.6073/pasta/bb563f16c7338fdb3ddf82057ef43cc6 (2023).
68. Magnuson, J. J., Carpenter, S. R. & Stanley, E. H. North temperate lakes LTER: chemical limnology of primary study lakes: nutrients, pH and carbon 1981 - current. *Environmental Data Initiative* https://doi.org/10.6073/PASTA/325232E6E4CD1CE04025FA5674F7B782 (2023).
69. Magnuson, J., Carpenter, S. & Stanley, E. North temperate lakes LTER: chlorophyll - Madison Lakes area 1995 - current. *Environmental Data Initiative* https://doi.org/10.6073/PASTA/F9C2E1059BCF92F138E140950A3632F2 (2022).
70. Magnuson, J. J., Carpenter, S. R. & Stanley, E. H. North temperate lakes LTER: phytoplankton - Madison Lakes area 1995 - current. *Environmental Data Initiative* https://doi.org/10.6073/PASTA/43D3D401AF88CC05C6595962BDB1AB5C (2022).
71. Magnuson, J., Carpenter, S. & Stanley, E. North temperate lakes LTER: zooplankton - Madison Lakes area 1997 - current. *Environmental Data Initiative* https://doi.org/10.6073/PASTA/D5ABE9009D7F6AA87D1FCF49C8C7F8C8 (2022).
72. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
73. Chen, I.-M. 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).
74. Kieft, K., Zhou, Z. & Anantharaman, K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* **8**, 90 (2020).
75. Kieft, K., Adams, A., Salamzade, R., Kalan, L. & Anantharaman, K. vRhyme enables binning of viral genomes from metagenomes. *Nucleic Acids Res.* **50**, e83 (2022).
76. Nayfach, S. et al. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat. Microbiol.* **6**, 960–970 (2021).
77. van Dongen, S. & Abreu-Goodger, C. Using MCL to extract clusters from networks. *Methods Mol. Biol.* **804**, 281–295 (2012).
78. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).
79. O'Leary, N. A. et al. Reference Sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).
80. Finn, R. D. et al. HMMER web server: 2015 update. *Nucleic Acids Res.* **43**, W30–W38 (2015).
81. Camargo, A. P. et al. Identification of mobile genetic elements with geNomad. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-023-01953-y (2023).
82. 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).
83. 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).
84. Zhou, Z. et al. TYMEFLIES vMAGs and related properties. *Figshare* https://figshare.com/articles/dataset/TYMEFLIES_vMAGs_and_related_properties/24915750 (2023).

Acknowledgements

We thank the local support for fieldwork conducted in Lake Mendota, WI, as a site of a long-term lake ecological study for North Temperate Lakes Long Term Ecological Research, including the following people as sampling leads: A. Kent, T. Yannarell, A. Shade, S. Jones, R. Newton, G. Wolfe, E. K. Read, L. Beversdorf and J. Mutschler; and initial Microbial Observatory lead, E. W. Triplett. This research was supported by the National Science Foundation grant number DBI2047598 (K.A.), US Department of Agriculture National Institute of Food and Agriculture under Hatch project 1025641 (K.A.), Simons Foundation Investigator in Aquatic Microbial Ecology Award LI-SIAME-00002001 (B.J.B.) and research funds from Synthetic Biology Research Center of Shenzhen University (Z.Z.). C.M. was supported by a National Science Foundation Graduate Research Fellowship. R.R.R. was supported by a National Science Foundation Postdoctoral Research Fellowship in Biology (DBI-2011002). Sequencing and initial sequence datasets processing were carried out at the US DOE JGI (CSP 504350). The work (proposal: CSP 504350) conducted by the US DOE JGI (https://ror.org/04xm1d337), a Department of Energy Office of Science User Facility, is supported by the Office of Science of the US Department of Energy operated under contract number DE-AC02-05CH11231.

Author contributions

Z.Z., K.D.M. and K.A. conceived the project. R.R.R. performed DNA extraction and sequencing. Z.Z., C.M. and P.Q.T. conducted bioinformatic analyses, statistical analyses, visualization of results and content organization. Z.Z. and K.A. wrote the manuscript draft. All authors (Z.Z., K.D.M., K.A., R.R.R., B.J.B., C.M., P.Q.T.) reviewed the results and edited and approved the manuscript.

Competing interests
The authors declare no competing interests.

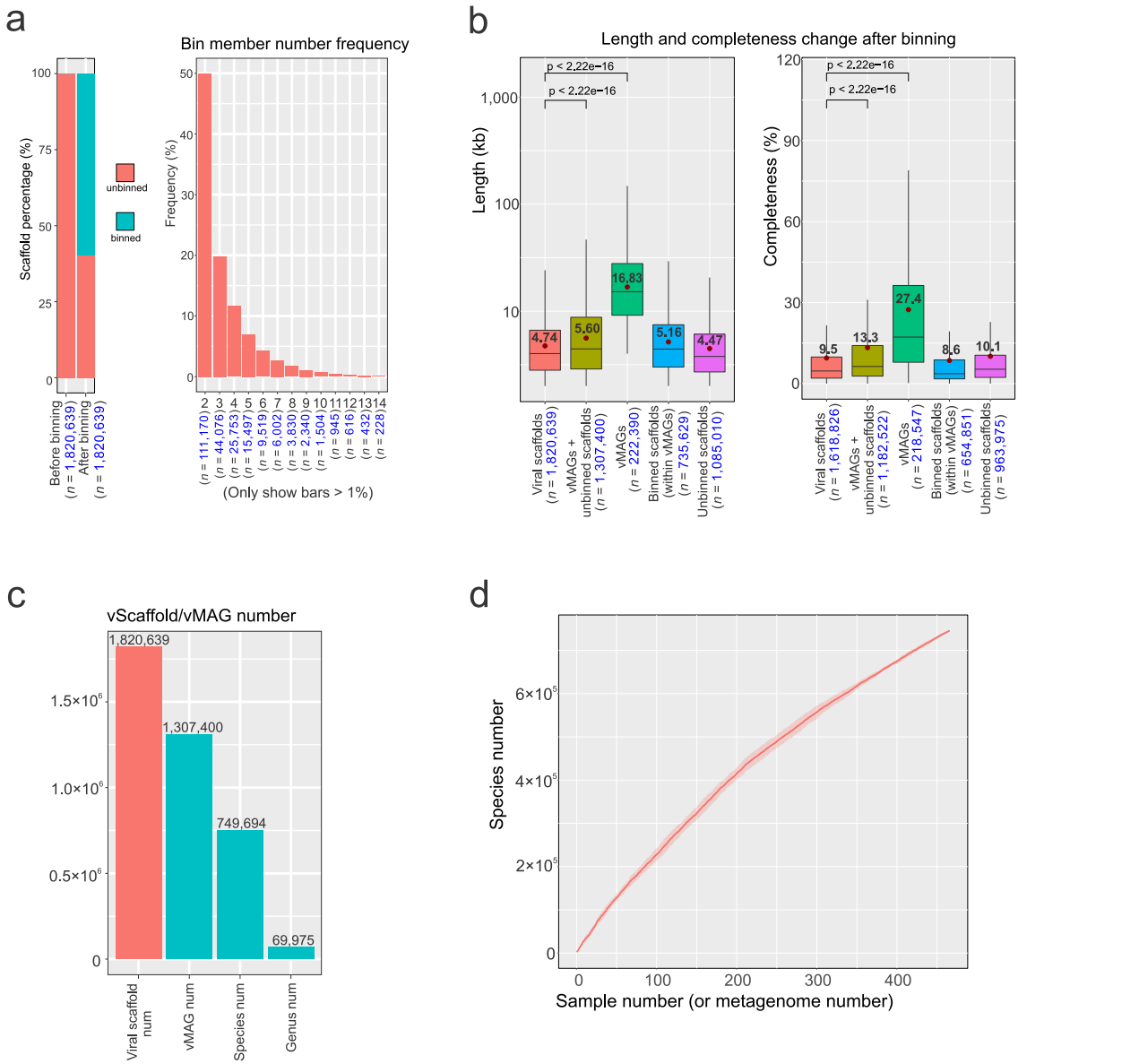
Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41564-024-01876-7.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41564-024-01876-7.

Correspondence and requests for materials should be addressed to Karthik Anantharaman.

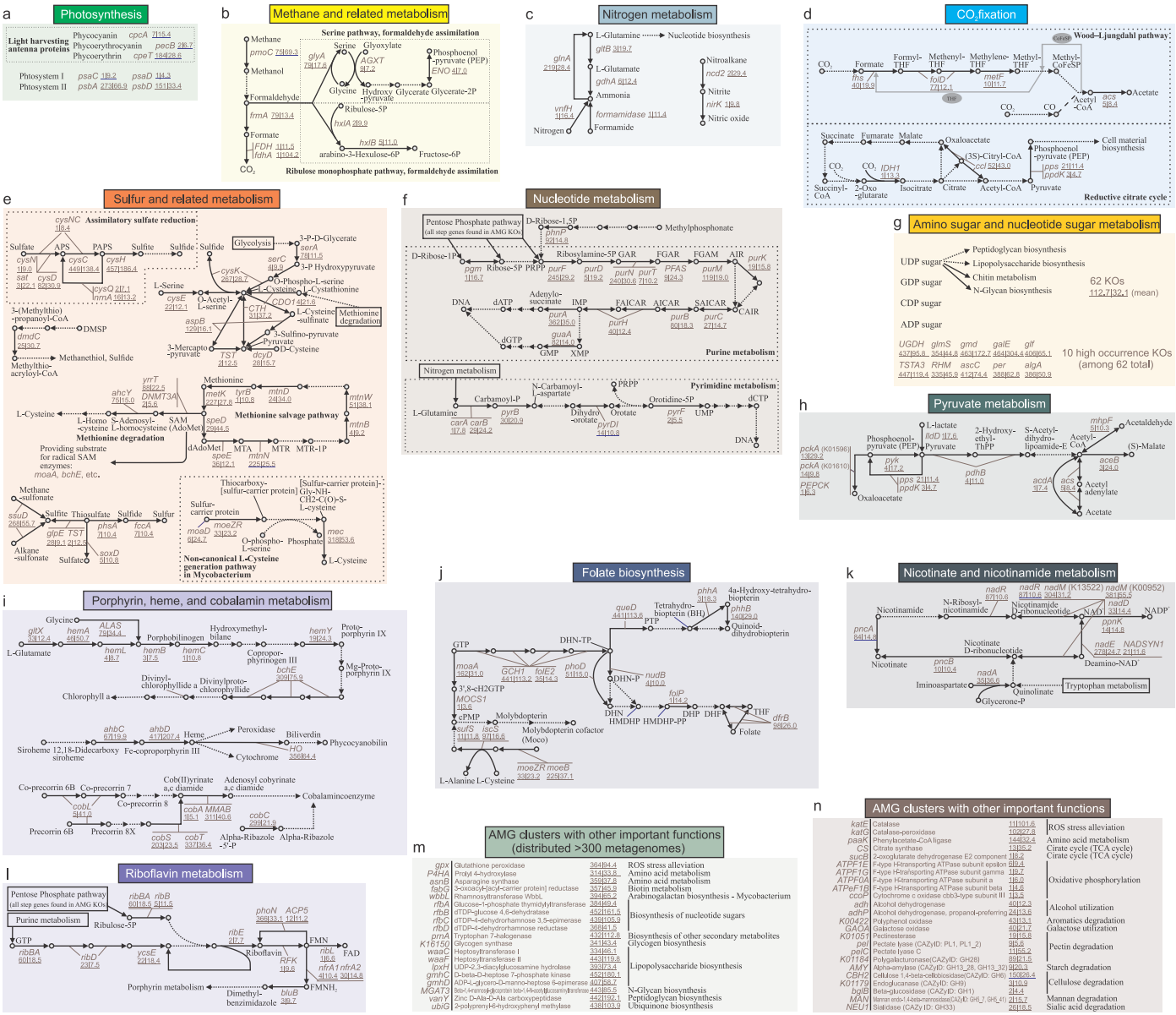
Peer review information *Nature Microbiology* thanks Timothy Ghaly, Andrew Millard and David Pearce for their contribution to the peer review of this work. Peer reviewer reports are available.

https://doi.org/10.1038/s41564-024-01876-7



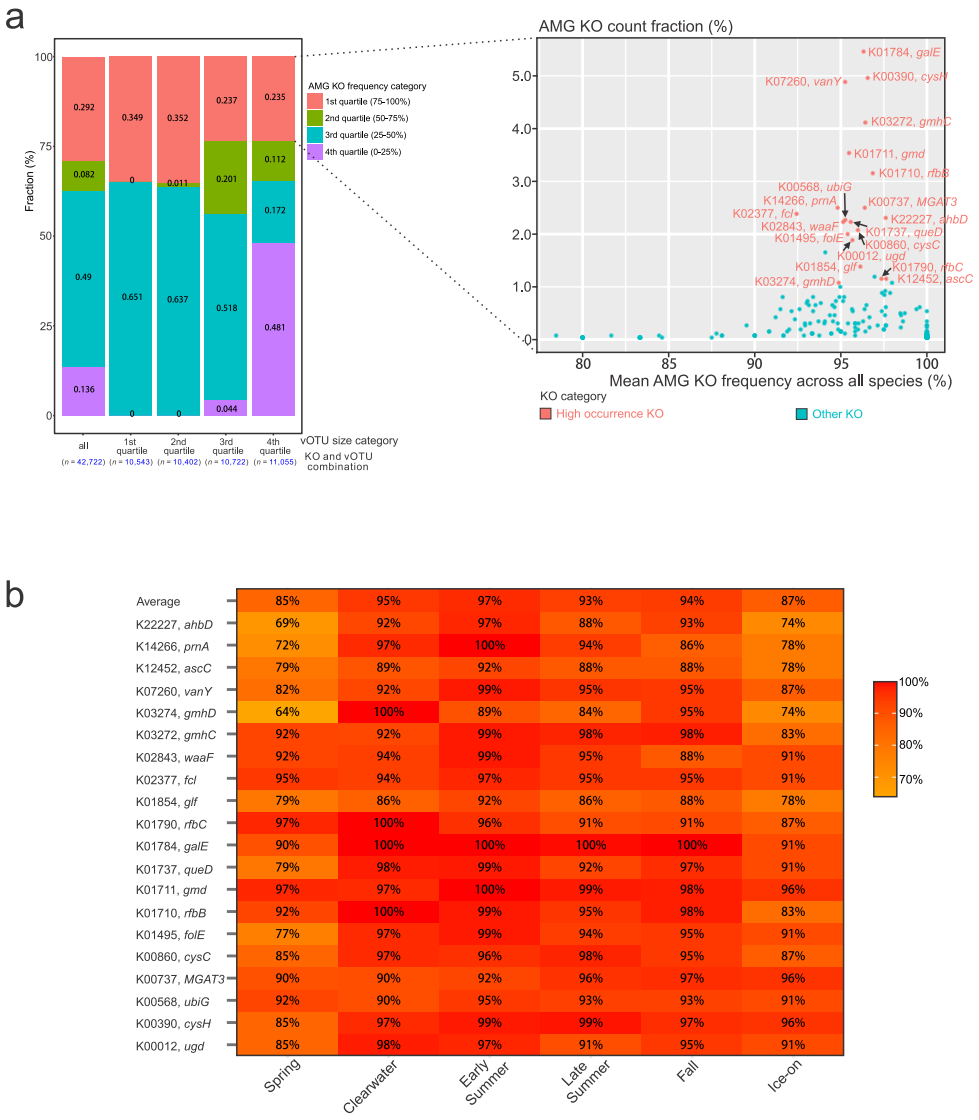
Extended Data Fig. 1 | Summary of viral scaffolds and genomes. a Binned/unbinned scaffold percentage after binning by vRhyme and bin member number frequency for all bins (vMAGs). Only bin member numbers with frequencies > 1% are shown in the bar plot. Numbers of scaffolds and numbers of bins are labeled accordingly. **b** Length and completeness change after binning, and CheckV quality to viral genome length distribution. Viral scaffold or/and vMAG (viral genome) numbers are labeled accordingly. "Viral scaffolds": total viral scaffolds before binning; "vMAGs+unbinned scaffolds": vMAGs and unbinned scaffolds after binning; "vMAGs": vMAGs after binning; "Binned scaffolds (within vMAGs)": binned scaffolds (the scaffolds that are in the vMAGs) after binning; "Unbinned scaffolds": unbinned scaffolds after binning. Statistical significance was assessed using two-sided t-tests for the indicated comparisons, with p-values indicating significance between comparisons. **c** The number of viral scaffolds, vMAGs, species, and genera. **d** The rarefaction curve of species-level vOTU numbers. Ten replicates with a random starting sample were made to generate error bars.

https://doi.org/10.1038/s41564-024-01876-7



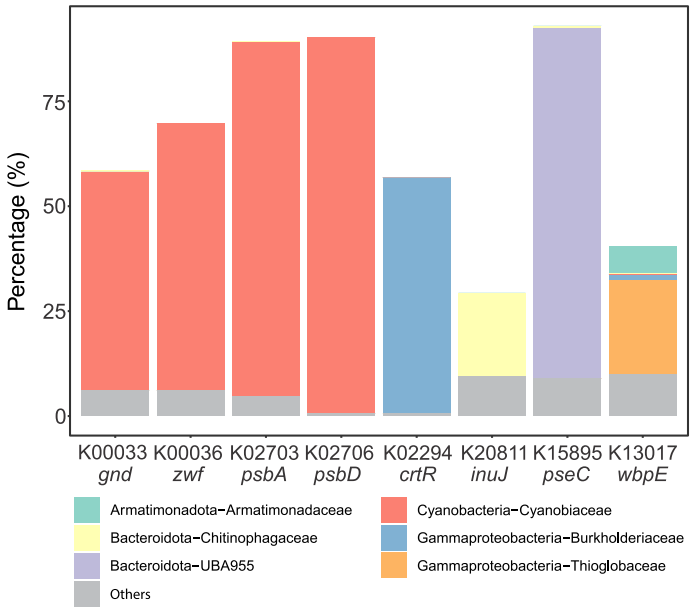
Extended Data Fig. 2 | AMG metabolisms and functions. a AMGs involved in photosynthesis, **b** AMGs involved in methane and related metabolism, **c** AMGs involved in nitrogen metabolism, **d** AMGs involved in CO_2 fixation, **e** AMGs involved in sulfur and related metabolism, **f** AMGs involved in nucleotide metabolism, **g** AMGs involved in amino sugar and nucleotide sugar metabolism, **h** AMGs involved in pyruvate metabolism, **i** AMGs involved in porphyrin, haem, and cobalamin metabolism, **j** AMGs involved in folate biosynthesis, **k** AMGs involved in nicotinate and nicotinamide metabolism, **l** AMGs involved in riboflavin metabolism, **m** AMGs assigned to AMG clusters with other important functions (distributed in >300 metagenomes), **n** AMGs assigned to clusters with other important functions. Gene symbols, the corresponding enzyme name, and CAZy ID for genes in n were depicted in brown together with the occurrence and abundance values (labeled as "occurrence|abundance"; occurrence, the number of metagenomes in which an AMG cluster can be found; abundance, the mean normalized abundance of AMG carrying viruses in the metagenomes in which this AMG can be found). Dotted arrows indicate steps that are not encoded by AMGs. Detailed information on each AMG cluster can be found in Supplementary Table S6.

https://doi.org/10.1038/s41564-024-01876-7



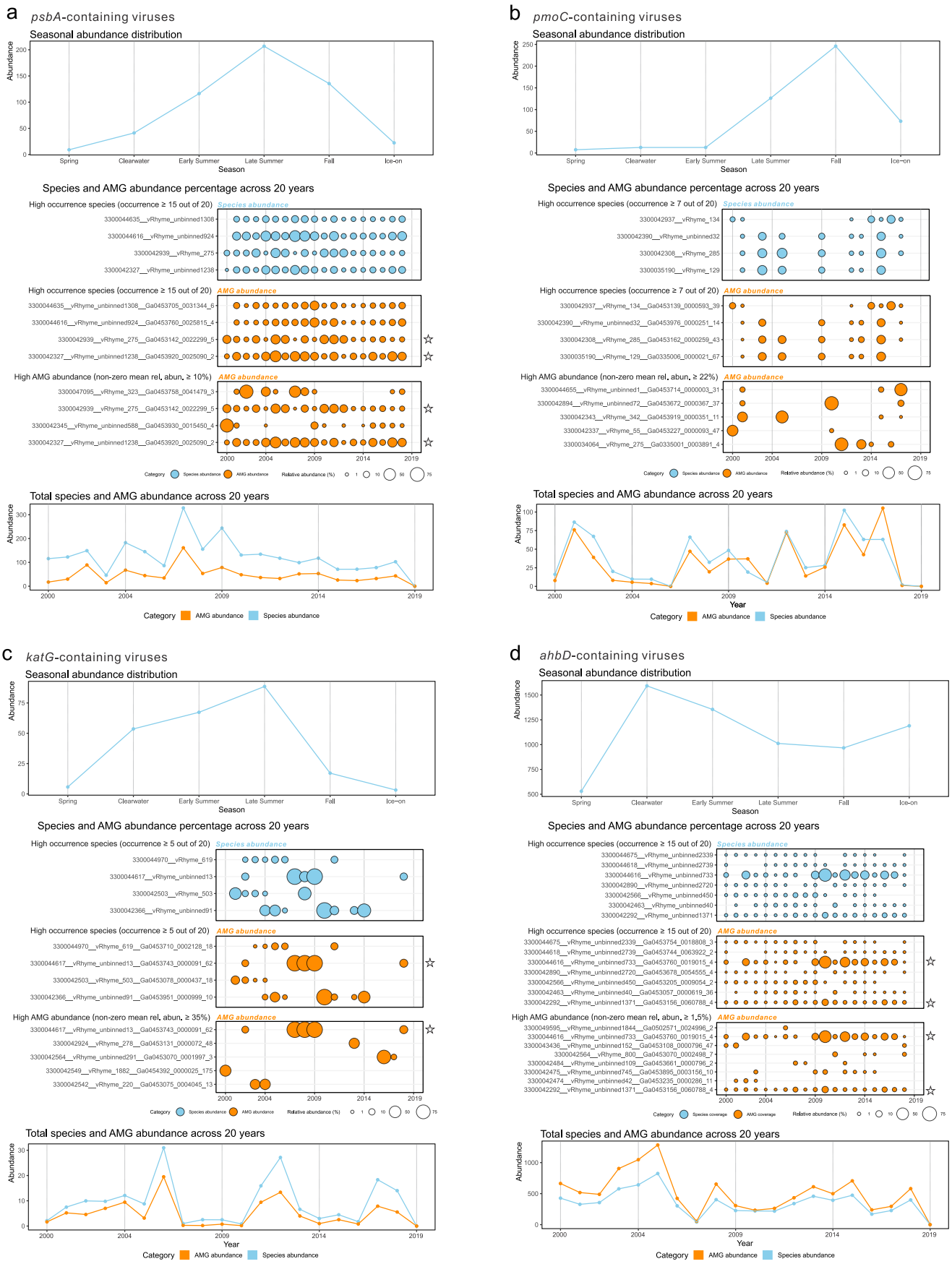
Extended Data Fig. 3 | AMG cluster variation in species and high occurrence AMG cluster distribution across different seasons. a AMG cluster variation in viral species. The left bar plot represents the AMG cluster presence ratio pattern among all AMG cluster and species combinations. The x-axis indicates the size category of species and the number of AMG cluster and species combinations. The y-axis indicates the fractions of four quartiles of AMG cluster presence ratios. The right scatter plot represents the AMG cluster count fraction (the percentage of one AMG cluster being encountered among all AMG clusters within a species) to the mean AMG cluster presence ratio (the percentage that one AMG cluster appears among all members within a species) across all species. This scatter plot used the AMG cluster and species combinations of the 1st quartile (75-100%) of AMG cluster presence ratio category (the highest presence ratio) with the species size in the 4th quartile (the largest species size), which was shown as the connection by dash lines. High occurrence AMG clusters (distributed > 400 metagenomes) were colored red, and other AMG clusters were colored green. **b** Seasonal distribution of high occurrence AMG clusters (distributed > 400 metagenomes)) across metagenomes. The percentage indicates the AMG cluster containing metagenome number over the total metagenome number in each season.

https://doi.org/10.1038/s41564-024-01876-7



Extended Data Fig. 4 | The taxonomic distribution (classified to the family level) of predicted hosts for eight AMG cluster-containing viruses with low Simpson indices. Unclassified hosts were not depicted and low abundance families (with abundance < 5% in all eight AMG clusters) were integrated into a group named "Others".

https://doi.org/10.1038/s41564-024-01876-7



Extended Data Fig. 5 | See next page for caption.

https://doi.org/10.1038/s41564-024-01876-7

Extended Data Fig. 5 | Species and AMG abundance across the time-series. The seasonal abundance distribution, species and AMG abundance percentage, and total species and AMG abundance across 20 years for *psbA*- (**a**), *pmoC*- (**b**), *katG*- (**c**), *ahbD*-containing (**d**) viruses are summarized. In each subpanel, high occurrence species were picked according to the occurrence across 20 years, high abundance AMGs were picked according to the non-zero mean relative abundance across 20 years, and the abundance for each year was represented by the season with the highest/second to the highest species abundance in each year (Late Summer for *psbA*, Fall for *pmoC*, Late Summer for *katG*, and Early Summer for *ahbD*). Species and AMGs were colored in blue and orange, respectively. Star-labeled AMGs indicate the overlap of the high occurrence species and high abundance AMG in subpanels **a**, **c**, **d**. The abundance values (for both species and AMGs) were normalized by 100 M reads/metagenome. For *psbA*- and *ahbD*-containing viruses, only species with ≥ 20 occurrences out of 471 metagenomes were included in the analysis; for *pmoC*- and *katG*-containing viruses, only species with ≥ 5 occurrences out of 471 metagenomes were included in the analysis. The species and AMG abundance percentage calculation was based on the total occurrence-filtered viral species.

nature portfolio

Corresponding author(s): Karthik Anantharaman

Last updated by author(s): May 17, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Codes used in this project are available at the following GitHub repository: https://github.com/AnantharamanLab/TYMEFLIES_Viral. Software used in this study include VIBRANT v1.2.1, IMGAP v5.0.20, metaSPADES v3..14.1, vRhyme v1.0.0, CheckV v0.8.1, DIAMOND v0.9.14.115, dRep v3.2.2, NCBI RefSeq (2023-01-13 release), geNomad v1.5.1, iPHoP v1.2.0, MetaPop v0.0.60, Bowtie2 v2.4.5
Data analysis	Codes used in this project are available at the following GitHub repository: https://github.com/AnantharamanLab/TYMEFLIES_Viral

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The metagenomic datasets (including assemblies and raw reads) are all available under JGI Proposal ID 504350 at the platform of the Integrated Microbial Genomes & Microbiomes system (IMG/M: https://img.jgi.doe.gov/m/). The retrieved viral genomes were deposited in NCBI Bioproject PRJNA1130067. At the same time, the

TYMEFLIES viral genomes and related properties, including annotations for viral proteins, taxonomic classification, host prediction, and virus clustering results, were deposited in the following address: https://figshare.com/articles/dataset/TYMEFLIES_vMAGs_and_related_properties/24915750. The raw environmental parameter spreadsheets are available in the Environmental Data Initiative (EDI, https://edirepository.org/) database.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Here, we leveraged time-series metagenomes collected over 20 years (2000-2019; the "TYMEFLIES" (Twenty Years of Metagenomes Exploring Freshwater Lake Interannual Eco/evo Shifts) metagenome project) to study freshwater viral diversity, ecology, and their association with metabolism and their hosts.
Research sample	In this study, we analyzed a total of 471 metagenome samples that were collected over twenty years. For each year, we divided the samples into six seasons. In this study, we analyzed a total of 471 metagenome samples. For each year, we divided the samples into six seasons. These seasons—ice-on, spring, clearwater, early summer, late summer, and fall—were defined by environmental data and most accurately represent microbial phenology.
Sampling strategy	In this study, 471 water filter samples were collected from a pelagic integrated 12-meter zone in Lake Mendota, Madison, WI, USA (GPS: 43.0995, -89.4045). Lake Mendota is a eutrophic freshwater lake located in Madison, WI (size: 39.4 km2; average depth: 12.8 m; pH: 8.5) and an important component within the North Temperate Lakes Long Term Ecological Research project (NTL-LTER) started in 1981. The samples were collected over several timepoints across different seasons each year, and the total sample period spanned 20 years (2000-2019). For each sample date, an approximately 250 mL integrated water sample was collected by filtering through a 0.2 µm pore size polyethersulfone Supor filter (Pall Corporation, Port Washington, NY). Filters were stored at -80°C for long-term storage. For omics sequencing, DNA extraction was conducted by using the FastDNA Spin Kit (MP Biomedicals, Burlingame, CA) with minor modifications.
Data collection	Environmental parameters were acquired from the sampling station at Lake Mendota (GPS: 43.0988, -89.4054) through the NTL-LTER program (https://lter.limnology.wisc.edu/) and are available through the Environmental Data Initiative (EDI, https://edirepository.org/). Extracted DNA from 471 samples was submitted to the Department of Energy Joint Genome Institute (DOE JGI) (Walnut Creek, CA) for metagenomic sequencing. Illumina regular fragments with ~300 bp length were made for metagenome library construction; afterward, the high-throughput sequencing was conducted using the Illumina NovaSeq S4 (Illumina, San Diego, CA), yielding paired-end reads of 150 bp each and approximately 1.5×108 reads (accounting for both ends) per sample.
Timing and spatial scale	Details of all samples collected are provided in Table S1. All samples were acquired from the sampling station at Lake Mendota (GPS: 43.0988, -89.4054).
Data exclusions	N/A
Reproducibility	N/A
Randomization	N/A

Blinding: N/A

Did the study involve field work? ☒ Yes ☐ No

Field work, collection and transport

Field conditions: All parameters are available in Tables S1 and S13. They are also available through through the NTL-LTER program (https://lter.limnology.wisc.edu/) and the Environmental Data Initiative (EDI, https://edirepository.org/).

Location: Lake Mendota (GPS: 43.0988, -89.4054)

Access & import/export: All samples and environmental parameters were collected as part of the North Temperate Lakes Long Term Ecological Research project (NTL-LTER) started in 1981.

Disturbance: No disturbances were caused by this study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks: N/A

Novel plant genotypes: N/A

Authentication: N/A