Title

Variable impact of geochemical gradients on the functional potential of bacteria, archaea, and phages from the permanently stratified Lac Pavin

Permalink
https://escholarship.org/uc/item/6nd6v4q2

Journal
Microbiome, 11(1)

ISSN
2049-2618

Authors
Jaffe, Alexander L
Bardot, Corinne
Le Jeune, Anne-Hélène
et al.

Publication Date
2023

DOI
10.1186/s40168-022-01416-7

Peer reviewed

RESEARCH

Open Access

Variable impact of geochemical gradients on the functional potential of bacteria, archaea, and phages from the permanently stratified Lac Pavin

Alexander L. Jaffe[1], Corinne Bardot[2], Anne-Hélène Le Jeune[2], Jett Liu[1], Jonathan Colombet[2], Fanny Perrière[2], Hermine Billard[2], Cindy J. Castelle[3], Anne-Catherine Lehours[2] and Jillian F. Banfield[3,4,5,6*]

Abstract

Background Permanently stratified lakes contain diverse microbial communities that vary with depth and so serve as useful models for studying the relationships between microbial community structure and geochemistry. Recent work has shown that these lakes can also harbor numerous bacteria and archaea from novel lineages, including those from the Candidate Phyla Radiation (CPR). However, the extent to which geochemical stratification differentially impacts carbon metabolism and overall genetic potential in CPR bacteria compared to other organisms is not well defined.

Results Here, we determine the distribution of microbial lineages along an oxygen gradient in Lac Pavin, a deep, stratified lake in central France, and examine the influence of this gradient on their metabolism. Genome-based analyses revealed an enrichment of distinct C1 and CO_2 fixation pathways in the oxic lake interface and anoxic zone/sediments, suggesting that oxygen likely plays a role in structuring metabolic strategies in non-CPR bacteria and archaea. Notably, we find that the oxidation of methane and its byproducts is largely spatially separated from methane production, which is mediated by diverse communities of sediment methanogens that vary on the centimeter scale. In contrast, we detected evidence for RuBisCO throughout the water column and sediments, including form II/III and form III-related enzymes encoded by CPR bacteria in the water column and DPANN archaea in the sediments. On the whole, though, CPR bacteria and phages did not show strong signals of gene content differentiation by depth, despite the fact that distinct species groups populate different lake and sediment compartments.

Conclusions Overall, our analyses suggest that environmental gradients in Lac Pavin select for capacities of CPR bacteria and phages to a lesser extent than for other bacteria and archaea. This may be due to the fact that selection in the former groups is indirect and depends primarily on host characteristics.

Keywords CPR bacteria, Meromictic lake, C1 metabolism, Bacteriophages, RuBisCO, Methane cycle

*Correspondence:
Jillian F. Banfield
jbanfield@berkeley.edu
[1] Department of Plant and Microbial Biology, University of California, Berkeley, CA, USA
[2] Laboratoire Microorganismes: Génome et Environnement (LMGE), UMR CNRS 6023, Université Clermont-Auvergne, F-63000 Clermont-Ferrand, France
[3] Innovative Genomics Institute, University of California, Berkeley, CA, USA
[4] Department of Earth and Planetary Science, University of California, Berkeley, CA, USA
[5] Department of Environmental Science, Policy, and Management, University of California, Berkeley, CA, USA
[6] Chan Zuckerberg Biohub, San Francisco, CA, USA

Background

Meromictic lakes are permanently stratified aquatic ecosystems with a global distribution [1]. Despite variation in size and morphology, these lakes generally have several common physical features, including the presence of two major layers—a shallower, oxygenated layer called the mixolimnion, and a deeper, generally anoxic zone called the monimolimnion [2]. These layers are separated by an oxycline of variable steepness that occurs in the water column, forming an "interface" zone. A unique feature of meromictic lakes, compared to other similar bodies of water, is that mixing events between the two layers are rare, with seasonal mixing observed primarily within the oxygenated upper layer [2]. Based primarily on marker gene analyses, these compartments are known to host distinct and diverse microbial communities that co-vary with gradients in oxygen, light, and sulfur [3, 4], and thus can serve as model systems for probing the relationships between community structure, metabolism, and geochemistry.

Archaea and bacteria with the capacity to utilize one-carbon (C1) molecules as carbon and energy sources are widespread across the planet and are important players in many biogeochemical processes. As such, they occupy an array of diverse anaerobic and aerobic niches and perform unique metabolisms that enable the utilization of carbon dioxide (CO_2), carbon monoxide (CO), formate (HCOOH), methanol (CH_3OH), and methane (CH_4) for growth, including in meromictic lakes. However, the spatial distribution and phylogenetic affiliation of pathways associated with carbon compound transformation over the water column and sediments of these lakes is still emerging. Previous work has also shown that these ecosystems can host organisms from divergent lineages without representatives in pure culture, including the candidate phyla radiation (CPR) bacteria and a novel sibling lineage of the cyanobacteria (*Ca.* Tanganyikabacteria) [5–7]. CPR bacteria are typically ultra-small celled and are predicted to live as obligate episymbionts or parasites of other microorganisms [8–10]. Despite their ubiquity in the microbiomes of diverse aquatic ecosystems, the ecological roles of CPR bacteria and the extent to which their gene content is shaped by environmental factors remain poorly understood. These questions could be addressed in part using genome-resolved metagenomics in systems with strong geochemical gradients.

Here, we examine microbial community composition and function in Lac Pavin, a meromictic lake in central France with a strong oxygen gradient, high methane concentrations in the anoxic water column and sediments, and very low methane concentrations in the upper water layer. We performed genome-resolved metagenomics to obtain hundreds of draft and high-quality genomes for members of diverse lineages of bacteria, archaea, and their viruses (phages) and use these genomes to examine changes in C1 metabolic potential across the lake's depth and oxygen gradient, including methane production and consumption. Together, our analyses shed new light on the diversity and metabolism of Lac Pavin's microbial communities and how the lake's geochemical conditions differentially shape the gene inventories of CPR bacteria, other bacteria and archaea, and phages.

Materials and methods

Sample collection and measurements of oxygen

Samples from the shallow oxygen maximum, lake interface, anoxic zone, and shallow sediments (17 samples total) were collected in mid to late October each year from 2017 to 2019. Prior to sampling, oxygen concentration and saturation were measured through the water column with a SDOT dissolved oxygen data logger (NKE Instrumentation). All water column samples were collected using a 12-l Niskin bottle. Between 3 and 5 bottles were taken for each water column sample (36-60 liters total), and all the water from one depth was pooled and filtered by tangential flow filtration (0.2-μm cartridge) to yield concentrate. The 12-m and 57-m samples from 2019 were pre-filtered before tangential flow filtration with a 20-μm filtration tissue to exclude larger eukaryotic organisms. Filtrate from the 2018 70-m sample was subsequently re-filtered using a 30-kilodalton cartridge to yield a small size fraction concentrate (sample 70S, < 0.2μdDa). Filtrate from each depth was then split into fractions, and each of which was subjected to centrifugation at 24,500g (30 min, 4°C). Supernatant was removed, and pellets from each fraction from the same samples were resuspended in TE buffer (Tris 10mM; EDTA 1mM; pH8) and pooled. The pooled samples were centrifuged one final time as described above, and pellets were stored at −20°C until DNA extraction.

In 2019, a sediment core was taken and 5 strata were collected: 3 to 5 cm (SED1), 10 to 13 cm (SED2), 20 to 24 cm (SED3), 32 to 36 cm (SED4), and 54 to 58 cm (SED5). Each stratum was homogenized and stored at −20°C for DNA extraction. All sample metadata is listed in Table S1.

DNA extraction and sequencing

DNA was extracted according to the method described by [11], with slight modifications. Specifically, each sample pellet was defrosted and incubated in TE buffer (pH 8). Cells were lysed with lysozyme (Sigma, final concentration 20 mg/ml) for 30 min at 37°C and then incubated for 1 h at 37°C with sodium dodecyl sulphate (SDS, final concentration 0.5%), proteinase K (Sigma, final concentration 0.1 mg/ml), and RNaseA (Sigma, final concentration 0.8 µg/ml). Cell debris, proteins, and humic substances were removed by adding 100 µl of a solution of cetyltrimethylammonium bromide (CTAB)/NaCl (10%, w/v, and 0.7 M, respectively) and 80 µl of NaCl 5 M for 10 min at 65°C. CTAB-protein/polysaccharide complexes were extracted with 1 volume of phenol/chloroform/isoamyl alcohol (24:24:1) after centrifugation (10,000g, 4°C, 30 min). The aqueous phase was re-extracted with 1 volume of chloroform/isoamyl alcohol (24:1) after centrifugation (10,000g, 4°C, 30 min). Nucleic acids (aqueous phase) were precipitated with 0.6 volume of ice-cold isopropanol overnight and centrifuged (10,000g, 4°C, 30 min). Pellets were washed with 70% ethanol and centrifuged (10,000g, 4°C, 30 min). Pellets (nucleic acid) were air-dried and dissolved overnight in 50 µl of H_2O. DNA was quantified using the Qubit® 2 .0 fluorometer (Life Technologies) and stored at −20°C until processing.

250mg of sediment from each stratum was defrosted and centrifuged for 30 s at 10,000g to remove the supernatant. For the lysis step, samples were heated to 70°C and then vortexed for 10 min. DNA was then extracted using the Qiagen DNEasy Power soil kit following manufacturer instructions.

Library preparation and metagenomic sequencing were performed at the QB3 (University of California, Berkeley) Functional Genomics Laboratory/Vincent J. Coates Genomics Sequencing Laboratory. Libraries were sequenced with 150 base-pair, paired-end reads on either a Illumina HiSeq 4000 platform (2017 and 2018 samples) or a NovaSeq S4 platform (2019 samples) with a target depth of 15–25 gigabasepairs per sample.

Metagenomic assembly, coverage calculation, gene prediction, and preliminary annotation

Metagenomic assembly was performed following the procedures of He et al. 2021 [12]. Briefly, sequencing reads were rid of Illumina adapters using BBTools and were trimmed using Sickle [13] (default thresholds). Quality-filtered reads were then assembled using MEGAHIT (v. 1.2.9 , default parameters) [14] and scaffolded using IDBA-UD [15]. Reads were mapped back to the entire assembly for each sample using bowtie2 [16] to compute scaffold coverage values. Assembled scaffolds were subsetted to those ≥1000 bp for gene prediction and binning. Genes were predicted using Prodigal (*meta* mode) [17], and predicted proteins were annotated using USEARCH against the KEGG (v. 74.0), UniRef (release-2015_10), and UniProt (release-2015_05) databases.

Genome binning, refinement, taxonomic assignment, and relative abundance calculation

Genome binning was performed using both a manual and an automated approach. For the manual approach, scaffold information was loaded into ggKbase (ggkbase.berkeley.edu) and scaffolds were binned on the basis of coverage, GC content, taxonomic affiliation, and inventories of bacterial "single copy" genes. For the automated approach, reads were cross-mapped against every other assembly within each biome type using bowtie2. Sediment and water column samples were binned separately. Coverage tables were generated using the jgi_summarize_bam_contig_depths script (bitbucket.org/berkeleylab/metabat/src/master) and passed to MetaBAT2 (minimum contig size 1500 bp) [18] for automated bin generation. Bins derived from the manual and automated approaches were reconciled with DAS Tool [19] to create the best merged set of bins for downstream analysis.

To prepare for the bin refinement step, preliminary taxonomic classifications were performed using GTDB-Tk [20]. Bins associated with the Candidate Phyla Radiation bacteria or DPANN archaea were separated and profiled for reduced sets of marker genes sensitive to lineage-specific losses in these groups [21]. Completeness and redundancy were calculated as the percentage of marker genes present and duplicated, respectively. These quality metrics were combined with those for all other, non-CPR, and non-DPANN bins estimated by CheckM [22]. Bins were then filtered to those with ≥70 completeness and de-replicated at 95% average nucleotide identity (ANI) using dRep to create a secondary set for manual curation. All quality-filtered bins were loaded into Anvi'o [23] and visualized individually using the *anvi-refine* command. Bins were refined by removing sets of scaffolds with aberrant coverage profiles across all cross-mapped samples. Completeness and redundancy metrics were also considered. Refined bins were then re-assessed for quality (as above) and additionally profiled with GUNC [24] as an alternative metric of contamination.

Draft genomes from the CPR bacteria and DPANN archaea were classified by extracting 16 (or 14, in the case of DPANN) phylogenetically informative, syntenic ribosomal proteins [25] from each genome (where present) and combining them with comprehensive set of reference sequences from previous studies [26, 27]. Each marker protein was aligned individually using MAFFT [28] and trimmed using BMGE (*-m BLOSUM30*) [29]. All individual protein alignments were concatenated to generate a supermatrix. Any bins for which less than 7 of the 16 marker proteins were present were discarded. Finally, a maximum-likelihood phylogenetic tree was inferred using IQTREE (*-m TEST --b 1000*) [30]. Taxonomic assignments of newly recovered draft genomes from

Lac Pavin were manually assigned using the tree topology and taxonomy of proximal reference sequences. For all other bacteria and archaea, phylum-level taxonomy was assigned using a combination of GTDB-Tk results, protein annotations from ggKbase, and manual BLAST searches of the S3 or L6 ribosomal proteins.

Finally, reads from all samples were re-mapped to the refined bin set using bowtie2. Bin coverage was analyzed using CoverM (https://github.com/wwood/CoverM) with a 95% identity threshold for read mapping and 50% breadth threshold (fraction of genome covered by reads). Relative coverage values for each genome were computed by dividing its mean coverage by the total and summed coverage of all genomes present within a given sample. Read counts from this step were also used to compute the percentage of sequencing reads mapping to the total genome set (Table S1).

Marker gene-based analyses of community composition

Predicted proteins from all assemblies were probed using a custom Hidden Markov Model (HMM) describing ribosomal protein S3 sequences (rps3) from diverse bacteria and archaea and thresholded using an HMM alignment score of 40 [31]. Recovered rps3 sequences from across samples were subsequently clustered into “species groups” using usearch -cluster_fast (*-sort length -id 0.99 -maxrejects 0 -maxaccepts 0*) [32]. For each species group, the longest rps3 protein sequence and the longest scaffold from any species group member were chosen as representatives [31]. Representative protein sequences fragmented into multiple pieces by assembly errors were removed by selecting the longer partial sequence, and all representative sequences smaller than 1 standard deviation below the mean sequence length were removed.

Taxonomic affiliation was assigned at the phylum level using three separate pieces of evidence: the taxonomy of the bin containing the representative scaffold (where relevant), the taxonomy of the top UniRef100 BLAST hit to the representative protein sequence, and the consensus taxonomy of the representative scaffold. Consensus scaffold taxonomy was computed by comparing each predicted protein against UniRef100 using a custom DIAMOND database (diamond blastp) [33]. Hits were filtered to those with 70% or greater coverage of the query sequence and an *e* value less than or equal to $1\text{x}10^{-10}$, as done previously [34], and the most common phylum-level designation was taken. If the same phylum-level designation was predicted by any two pieces of evidence, it was assigned as the final taxonomy; otherwise, scaffold taxonomy was manually curated using BLAST searches and annotation information.

Finally, relative abundance of species groups was assessed by mapping all raw sequencing reads to the combined scaffold set using Bowtie2 (default parameters). Coverage was calculated using CoverM with a stringent mapping threshold (99% read identity) and a minimum coverage breadth of 50% of representative scaffolds. Relative coverage of each species group was calculated by dividing its mean coverage value by the total coverage of all species groups in a given sample. To examine overall similarity in community composition across samples, the relative coverage metrics computed above were subsequently used to generate a Bray-Curtis distance matrix (skbio’s diversity package in Python) and subjected to a principal coordinates analysis (implemented in skbio’s stats module).

Analysis of community-wide carbon metabolism

Proteins were re-predicted for all refined bins using Prodigal (*single* mode). All proteins that were not from Candidate Phyla Radiation bacteria were subjected to annotation with kofamscan [35]. Protein annotations were filtered first by *e* value (≤1e−6), and the highest scoring hit per protein was taken. Annotations were searched for a curated, previously published set of hidden Markov models (HMMs) representing key marker genes for pathways involved in C1 metabolism, carbon fixation, and methane metabolism (Table S4) [36]. We next manually adjusted score thresholds for each model using a previously reported approach [26]. This approach draws on the distribution of HMM-protein bitscores, phylogenetic relationships among proteins, and the relationship between phylogeny and bitscore to account for partial or divergent sequences and functionally dissimilar but related enzyme forms. All manually set HMM thresholds can be found in Table S4. Putative methane monooxygenases (above threshold hits to pmoA/K10944) were combined with reference sequences from the broader copper membrane monooxygenase family [37] in a phylogenetic tree (Fig. S1) and manually confirmed to associate with known methane monooxygenases rather than evolutionarily related, but functionally distinct ammonia monooxygenases (amo). Generally, a pathway was considered present in a genome if at least one marker gene was found, as implemented in [36]. To reduce the incidence of false positives for a subset of pathways (formate oxidation, aerobic CO oxidation, and the Wood Ljungdahl pathway), we set a more stringent threshold, requiring that at least 2, 3, and 2, respectively, of marker genes to be present. Form I and II RuBisCO (Calvin Cycle) were identified as described below.

We next examined the distribution of each pathway as a function of the relative coverage of genomes encoding it. Per-genome relative coverage values computed above were merged with pathway presence information and summed by sample for each pathway (Fig. 2a). To examine the same metric by phylogenetic group (Fig. S2) relative coverage values of genomes with a given pathway were aggregated by phylum-level designation, then averaged within groups of samples (interface, anoxic zone, etc). Phyla which did not attain ≥ 1% mean relative coverage in any sample group were removed. Relative coverage values for potential methane producers (taxa containing the *mcr* operon or taxonomically affiliated with known methanogenic lineages) were visualized on a per-genome basis (Fig. 2b).

Analysis of RuBisCO diversity and abundance

To determine the diversity of RuBisCO genes across the water column and sediments, all metagenomic assemblies were probed with a set of Hidden Markov Models (HMMs) describing the form I, II, II, II/II, and III RuBisCO, respectively [21]. Stringent score cutoffs originally published with the models were applied, and the best hit was selected for each individual protein. For the form III HMM, the score threshold was lowered to better account for divergent form III-a RuBisCO found in methanogens [38]. A custom threshold of 380.3 was determined by scoring of form III-a reference sequences from a previous publication [39]. Scaffolds encoding above-threshold hits were gathered and clustered across samples using dRep at 95% ANI (*cluster -p 20 -pa 0.80 -sa 0.95*) [40]. A non-redundant, representative set of scaffolds encoding RuBisCO was created by selecting the set of cluster centroids that were binned (i.e., belonged to a draft genome) and/or were the longest. Reads from all samples were re-mapped to this unique set and mean coverage per sample was computed using CoverM, as described above. Only scaffolds with ≥50% breadth were considered "present" in a given sample and retained for diversity analysis.

Next, putative taxonomy was assigned to all scaffolds encoding RuBisCO using a combination of bin taxonomy (for binned fragments), phylogenetic affiliation of co-encoded proteins, and manual curation. Sequence fragments clearly originating from eukaryotic organisms, such as algae and diatoms, were removed. Classification of newly recovered RuBisCO sequences was obtained by placing them within a reference RuBisCO large subunit protein phylogeny [39]. Finally, the relative coverage of each representative scaffold encoding RuBisCO within each sample was computed by dividing mean coverage by the total coverage across all scaffolds. Relative coverage was first summed by phylum-level taxonomy of the encoding scaffold and then averaged across samples within the same lake compartment.

CPR gene content analyses

Gene content of CPR bacteria was analyzed using a two-step de novo protein clustering pipeline [41]. First, proteins were clustered into "subfamilies" and overlapping subfamilies were merged (–coverage 0.75). For sequences in each protein cluster, the most common KEGG annotation was determined alongside the frequency of this annotation. To determine statistically enriched functions by lake compartment, we drew on a previously developed computational pipeline that tests the frequency of occurrence of a function/protein cluster within genomes from a given lake compartment versus those from outside of that compartment [42]. To determine the degree of enrichment, a ratio describing the frequency of occurrence within a given compartment relative to the frequency of occurrence outside that compartment was computed, and statistical significance assigned using the Fisher's exact test (the fisher_exact function in scipy.stats, alternative= "two-sided"). In cases where no genomes outside of a given ecosystem compartment encoded the function of interest, the function was labeled "exclusive" and no ratio was computed. All comparisons were corrected for false discovery rate using the multipletests function in statsmodels.stats.multitest ("fdr_bh"). Only protein families with more than 5 member sequences were considered.

Phage-related analyses

A preliminary set of putative viral scaffolds was constructed by subjecting all assemblies (including the post 0.2 µm "70S" sample) to viral prediction using VIBRANT [43] (default parameters) and VirSorter2 [44] (*--include-groups dsDNAphage,ssDNA --min-length 5000 --min-score 0.5*). All potential viral scaffolds ≥50 kb (including potential provirus scaffolds) were subsequently passed to CheckV [45] (*end_to_end* program), and results were used to further refine the set. Specifically, if a given scaffold was found to have either zero "host genes" (as determined by CheckV) and 5 or more "viral genes", or, alternatively, ≥1 "host genes" but a greater number of "viral genes" and no "provirus" call, the scaffold was assigned likely virus status. If a given scaffold had 5 times or more the number of host genes as viral genes, and was not considered a provirus, it was discarded. Alternatively, if such a scaffold was considered a provirus by CheckV, it was assigned likely provirus status. If a given scaffold had fewer than 5 times the number of viral genes than host genes, or no viral genes but less than 4 host genes, it was flagged for manual curation regardless of provirus status. Manual curation was performed by examining taxonomic and functional affiliations of proteins using a combination of databases as well as BLAST.

To further refine the set of putative viral genomes, we programmatically scanned for evidence of eukaryotic viruses. First, all viral proteins were predicted using Prodigal (*-p meta*) [17] and compared against UniRef100 as described above for rps3-bearing scaffolds. Taxonomic information for best, above-threshold hits for each protein was collected, and those scaffolds with ≥25% of annotated proteins affiliated with eukaryotes or eukaryotic viruses were discarded. The remaining set of putative phage were clustered at 95% ANI using dRep (*-pa 0.80 -sa 0.95 --clusterAlg single*) [40] and cluster representatives were chosen on the basis of scaffold length. Phage scaffolds were considered circularized if marked 100% complete by CheckV (via "high confidence" direct tandem repeats).

Putative phage scaffolds were integrated with those published previously from a subset of Lac Pavin metagenomes [46, 47], many of which were manually curated to completion. Distribution of the integrated set of phages was determined by stringent re-mapping (95% identity) of sequencing reads to the combined scaffold set as described above for bacterial and archaeal genomes. Host predictions were made using previously reported methodology [34] that combines CRISPR-Cas targeting, taxonomic profiling of phage genes, and, for predicted prophages, taxonomic profiling of surrounding genomic regions. Where host assignment methods were conflicting, assignment based on CRISPR-Cas targeting was prioritized, followed by taxonomic affiliation of regions surrounding prophages, where relevant.

Finally, we re-predicted proteins for the refined set of phage genomes using Prodigal (*-m single*). To account for the possibility of alternatively-coded phages, predictions were made in genetic codes 11, 15, and 4, and coding density was compared. If an "alternate" code (15 or 4) displayed an increase in coding density ≥10% relative to standard (11) code for a given phage [48], protein predictions in that code were retained. Otherwise, protein predictions in standard code were retained. Phage metabolic potential with depth was then analyzed as described above for CPR bacteria. Recovered phage *pmoC* genes were combined with those recovered from bacterial draft genomes, aligned using MAFFT and trimAl [49], and incorporated into a maximum-likelihood tree using IQTree with reference phage/host pmoC [50]. Predicted proteins were also used to compare newly reported phages from Lac Pavin (viral_region_length ≥ 50 kbp, Table S9) to those from reference databases/previous publications [46, 50, 51] using vContact2 [52]. The network was created using default parameters with the "ProkaryoticViralRefSeq201-Merged" reference database (*--pcs-mode MCL --vcs-mode ClusterONE*). Clusters reported by vContact2 were further classified as those containing only reference sequences, both Lac Pavin and reference sequences, and only Lac Pavin sequences. The network was visualized using Cytoscape v3.9.1 [53] in a prefuse force directed layout with a modified spring coefficient of $1x10^{-6}$. One large subnetwork containing multiple clusters was manually extracted from the broader network for the purposes of visualization.

Results

Diversity of microbial communities across Lac Pavin's water column and sediments

To examine microbial communities in Lac Pavin, we sampled the water column and shallow sediments of Lac Pavin annually from 2017 to 2019 (Table S1). The first year of sampling focused on the lake's oxycline (40–60 m), whereas subsequent sampling in the second and third years primarily targeted the permanently anoxic monimolimnion (~60–90 m) and the shallow sediments at depths between 3 and 58 centimeters (Fig. 1bc, Table S1). We also took one sample from the shallow oxygen maximum (12 m) as a point of comparison. Water column samples were filtered to the 0.2–20-μm size fraction, and whole-community genomic DNA was extracted and sequenced (the "Materials and methods" section). In addition, one post 0.2-μm fraction was used to extend genomic sampling to capture ultra-small cells and phages.

First, we surveyed assembled metagenomic scaffolds for ribosomal protein S3 (rpS3) sequences and assigned gene sequences to "species groups" (the "Materials and methods" section). Coverage of these rpS3 marker genes was then used to examine microbial community similarity across sampling depths/years. A principal coordinates analysis (PCoA) suggested that samples from the same lake compartment, e.g., the samples from the anoxic zone, showed highest similarity to each other across years and were distinct from samples from other compartments, consistent with the permanently stratified nature of the lake (Fig. 1c). Based on community composition, the shallow water column (12 m) most closely resembled samples from the lake interface (Fig. 1c), despite its relatively high oxygen concentration (Fig. 1b).

We investigated the specific phylogenetic groups responsible for driving dissimilarity between microbial communities in the lake, assigning rpS3-defined species groups to phylum-level lineages. Microbial communities from the water column interface were highly enriched in members of the Actinobacteria relative to other samples, consistent with findings from other lakes with pronounced oxyclines [7, 54] (Fig. 1d, Table S2). The shallow regions of the lake also displayed higher abundances of Cyanobacteria (Table S2) and Planctomycetes compared to the anoxic zone of the water column, where members of the Proteobacteria, Bacteroidetes, and Omnitrophota were particularly abundant (Fig. 1d, Table S2). Members of the Verrucomicrobia were distributed across the water column, but only minimally

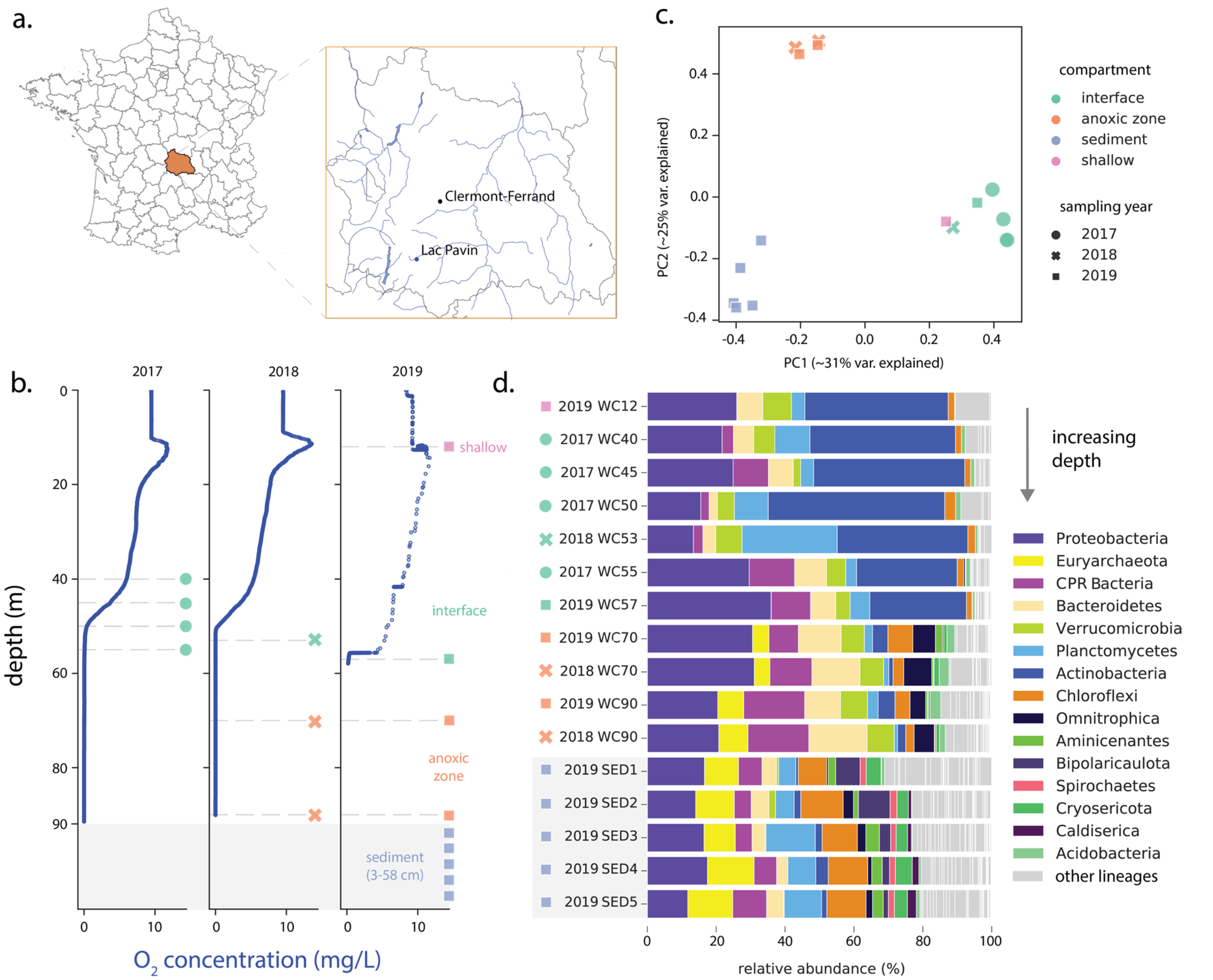

Fig. 1 Physical characteristics and microbial communities in Lac Pavin. **a** Map showing the location of the meromictic Lac Pavin in the Auvergne region of central France. **b** Oxygen profiles across the lake water column during each of the three yearly samplings. Depths sampled on a given year are indicated with a dashed gray line, and colored shapes match those in panels **c** and **d**. **c** Principal coordinates analysis of microbial community composition at the "species group" level. **d** Microbial community composition in Lac Pavin at the phylum level, based on relative coverage of representative scaffolds encoding the ribosomal protein S3. Samples are ordered from shallow to deep, including the five sediment samples collected in 2019. **N.B.** For legibility, only the 15 lineages with the highest median abundance across all samples are shown, all others are indicated in gray as "other". Abbreviations: PC, principal coordinate; var, variance; CPR, Candidate Phyla Radiation, WC, water column; SED, sediment

detected in the shallow sediments. These shallow sediments were themselves enriched for members of the Chloroflexi, Cryosericota, Caldiserica, Bipolaricaulota (candidate division OP1), DPANN archaea, and a variety of methanogens, some of which were also found in the anoxic zone (Fig. 1d, Table S2). Surprisingly, the shallowest sediment horizon (3–5 cm) also harbored levels of Cyanobacteria comparable to those in the shallower water column, despite the lack of light (Table S2). One of the three Cyanobacteria species found in the sediment at the times of sampling was also found in the photic zone, suggesting dispersal to the sediment via vertical settling. Notably, members of the Candidate Phyla Radiation were found across the lake's water column and shallow sediments, but were barely detected in the region of the oxygen maximum (Fig. 1d).

Recovery of draft quality bacterial and archaeal genomes from Lac Pavin

To more closely examine the phylogenetics and metabolic potential of key community members, including CPR bacteria, we generated metagenomic bins by grouping assembled scaffolds into draft genomes with both manual and automated approaches (the "Materials and methods"

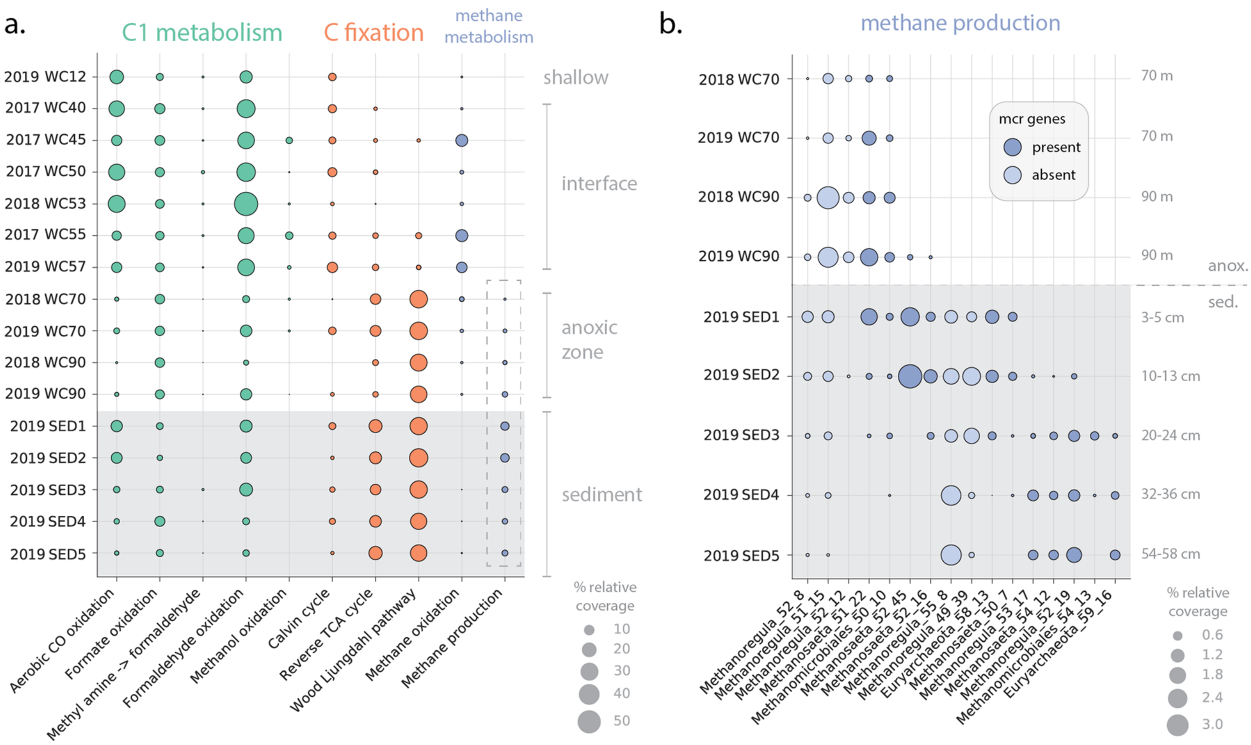

Fig. 2 Genetic potential for select carbon metabolism across the water column and sediments of Lac Pavin. **a** Relative abundance of genomes encoding C1 pathways across metagenomic samples. Bubble size indicates the combined coverage of all organisms encoding a given pathway in each sample normalized by the total coverage of all genomes in the quality-filtered, non-redundant set. The dashed box around methane production indicates the data points expanded in part b), showing the relative abundance (again, as indicated by percent of total genome coverage) for organisms from methanogenic lineages, with and without *mcr* genes. Abbreviations: CO, carbon monoxide; TCA, tricarboxylic acid cycle; anox., anoxic zone; sed., sediment; mcr, methyl-coenzyme M reductase

section). The resulting bins were de-replicated at the species level (95% ANI) and manually curated using Anvi'o and ggKbase tools (the "Materials and methods" section). Refined bins were then filtered again for quality (≥70% completeness) to yield a final set of 738 non-redundant, draft-quality, or better genomes for downstream analysis (Table S3) that accounted for ~11 to ~51% of sequencing reads (mean ~30.7%, Table S1). Taxonomy assignment via multiple methods indicated that this set included genomes from most major groups identified by marker gene analysis, including 106 draft genomes from the CPR bacteria, 574 from other bacteria, and 58 from archaea (Table S3).

Contributions of bacteria and archaea to C1 metabolism over an oxygen gradient

We searched the draft genomes for marker genes for pathways involved in C1 metabolism, carbon fixation, and methane metabolism (Table S4, the "Materials and methods" section) [36]. Using these results, we assigned the presence or absence of metabolic functions (pathways) across the set of non-CPR bacteria and archaea and explored their relative abundance in each lake compartment (CPR function and gene content is discussed below). Genomic analyses suggest that the capacity for aerobic carbon monoxide (CO) oxidation is most prevalent at the shallow oxygen maximum and in the interface zone of the lake ecosystem, consistent with the presence of oxygen in both of these lake compartments (Fig 2a). In these regions, CO oxidation was dominated by members of the Actinobacteria (Fig. S2), which represented between 3.5 and 18.6% of the microbial community (Table S5). Surprisingly, we also found some evidence for the capacity for aerobic CO oxidation in the anoxic sediments (Fig 2a), primarily in a single genome from the candidate phylum Bipolaricaulota (candidate division OP1; Fig. S2). However, the genes encoding CO oxidation were encoded on a very short contig (~2200 bp), and thus, their taxonomic affiliation could not be confidently assigned.

Biological methane oxidation is thought to act as the major "sink" for methane in Lac Pavin [55]. We thus examined the distribution of genes associated with this process, finding that genomes containing components of the methane monooxygenase operon were most abundant in the interface, where in combination they reached a maximum of 13.8% of the microbial community (sample LacPavin_0818_WC55, Fig. 2a, Table S5). Recovered *pmoC* genes were primarily affiliated with the known methane-oxidizing lineage *Methylococcales,* including the genus *Methylobacter*. Some methane oxidizers were present at low abundance in the anoxic zone, suggesting that they may be performing methane oxidation under microaerophilic conditions in this compartment [55, 56]. On the contrary, potential methane oxidizers were essentially undetectable in the sediments, where genetic potential for methane production was most concentrated (Fig. 2a). Additionally, we did not detect ANME (anaerobic methanotrophic) archaea capable of anaerobic methane oxidation in either the anoxic zone or the sediments. Thus, overall, we report that methane oxidation is largely spatially separated from methane production in Lac Pavin, consistent with previous amplicon-based studies [56]. A subset of potential methane oxidizers encoded genes for methanol oxidation, which had a parallel distribution through the water column (Fig. 2a, Table S5), suggesting that these bacteria may also be capable of catalyzing the oxidation of methane to formaldehyde.

Microorganisms with the genetic capacity to oxidize formaldehyde were also highly abundant in the lake interface, reaching a maximum relative abundance of ~51% in one water column sample from 53 m. Potential for formaldehyde oxidation in the interface was dominated by members of the *Actinobacteria*, *Planctomycetes*, and *Proteobacteria*, whereas in the sediments *Planctomycetes* and several methanogenic archaea were the greatest contributors. These findings are consistent with the prior discovery of divergent formaldehyde-activating enzymes (*fae*) in *Planctomycetes* [57] as well as the association of *fae* with methanogenesis [58]. The genetic potential for oxidation of formate, a breakdown product of formaldehyde, was more evenly distributed across the water column (Fig. 2a).

Genes encoding the central complex involved in methane production (methyl-coenzyme reductase, alpha, and beta subunits, MCR) are primarily localized to the lake sediments, but were also detected in the anoxic zone (Fig. 2). We recovered eleven high-quality, non-redundant archaeal genomes containing the MCR genes that were affiliated with the lineages *Euryarchaeota* (order *Methanomassiliicoccales)* as well as *Methanosaeta* and *Methanomicrobiales* (including the genus *Methanoregula*), which are commonly detected in freshwater lakes [59]. Read mapping showed that these organisms represent between ~0.05 and 3.5% of the microbial community in each sample but are differentially distributed across the deeper water column and sediments (Fig. 2b). Recovered *mcr* genes in the deep water column (70 m) and at the water-sediment interface (90 m) were almost entirely attributable to two organisms (one *Methanosaeta* and one *Methanomicrobiales* organism). The potential methane producers in the sediment were more diverse at the genus level, particularly in the shallower sediments (3–24 cm) (Fig. 2b). *Mcr* genes in the deeper sediments (32–58 cm) were primarily attributable to *Methanoregula, Methanosaeta*, and *Euryarchaeota* genomes which were rare or non-existent in the shallower sediments as well as the water column (Fig. 2b). Five widespread genomes affiliated with the *Methanoregula* were also recovered by our binning efforts; however, they did not contain the *mcr* gene cluster, potentially due to lower genome completeness for some members (ranging from 76.48 to 92.4%) (Table S3, Fig. 2b).

We also examined the genetic potential for carbon fixation across the water column and sediments. The Calvin Cycle (form I and II RuBisCO) was present throughout the water column, but was most prevalent in the shallower lake zones (Fig. 2a), primarily among Cyanobacteria, Actinobacteria, and Proteobacteria (Fig. S2). In contrast, we found that the genetic potential for CO_2 fixation via the Wood-Ljungdahl (W-L) pathway was nearly absent in oxygenated parts of the lake (Fig. 2a) but was common among Chloroflexi, Proteobacteria, and Euryarchaeota in the anoxic water column and sediments, which together comprised ~25–31% of the community (Fig. S2, Table S5). To a lesser extent, micro-organisms with the genetic potential for the reverse TCA cycle were also associated with the lake's anoxic zone and sediments, with the exception of samples taken from 90 m at the anoxic/sediment interface (Fig. 2a), where this capacity was rare. We found no evidence for alternative carbon fixation pathways involving 3-hydroxypropionate or 4-hydroxybutyrate among Lac Pavin organisms.

Together, these findings point to depth/oxygen as a structuring force for not only microbial community composition in Lac Pavin, but also for certain metabolic capacities related to carbon cycling that display clear evidence of spatial stratification.

CPR bacteria with the metabolic potential to incorporate CO_2 via RuBisCO

Previous work based on 16S rRNA amplicon-based approaches showed that Lac Pavin hosts diverse communities of bacteria from groups now recognized to be part of the CPR and showed that these communities vary with depth [5]. Given prior research demonstrating the diversity of RuBisCO enzymes among these organisms [39, 60], we hypothesized that some CPR in Lac Pavin may play a role in carbon cycling via CO_2 incorporation into biomass. We identified form II/III RuBisCO in members of the Absconditabacteria and the Komeilibacteria (Parcubacteria radiation) and form III-c in another

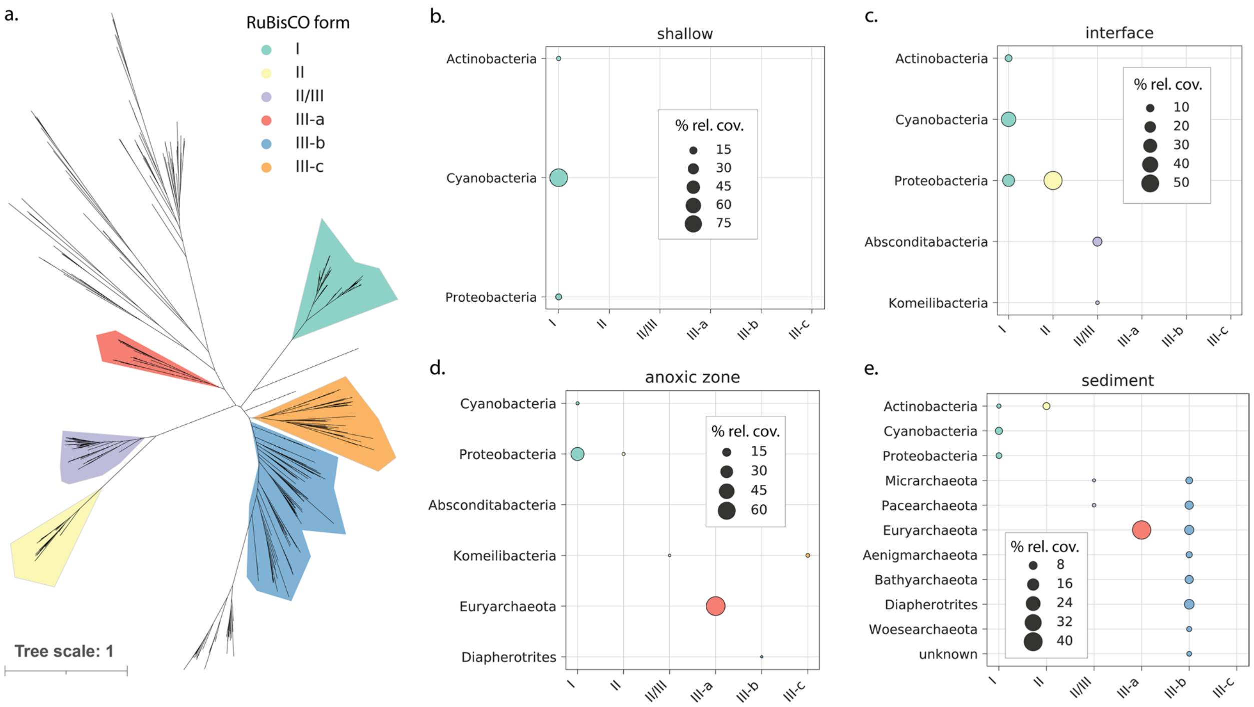

Fig. 3 Diversity and distribution of RuBisCO in Lac Pavin. **a** Unrooted phylogenetic tree showing relationships among major clades within the RuBisCO superfamily. **b**–**e** Distribution of community metabolic potential for carbon fixation/CO_2 incorporation via RuBisCO in each lake compartment. Bubble size represents the percentage of coverage for all scaffolds with RuBisCO attributable to a given lineage/RuBisCO pair. This percent coverage is averaged across all samples within a lake compartment

Komeilibacteria (Fig. 3), a lineage in which form III-c has not previously been observed. Both form II/III and all form III RuBisCO likely function by incorporating carbon dioxide with a free nucleotide to form glyceraldehyde-3-phosphate, a central glycolytic intermediate [61].

We quantified the relative abundance of CPR RuBisCOs in comparison to other RuBisCO genes encoded by other bacteria and archaea, including those found both in genomic bins as well as in unbinned scaffolds. In the interface, form II/III from the Absconditabacteria composed on average 13.6% (ranging from ~10 to 17%) of RuBisCO gene copies and form II/III from the Komeilibacteria contributed only ~2.2% of gene copies on average (Fig. 3c, Table S6). Similarly, in the anoxic zone, Komeilibacteria accounted for an average of 3.88% of RuBisCO gene copies, reaching a maximum of 7.77% just above the sediment-water interface (90-m sample) (Table S6, Fig. 3d). However, community potential for CO_2 incorporation via RuBisCO within this compartment (and the shallow sediments) was dominated by archaea bearing the form III-a and III-b enzymes (Fig. 3 d, e, Table S6). The most abundant genes were encoded by various methanogens within the Euryarchaeota and DPANN lineages, including the Pacearchaeota and Diapherotrites (Fig. 3 d, e).

Diversity and distribution of CPR bacteria in Lac Pavin

We examined the broader phylogenetic diversity of CPR bacteria in Lac Pavin using genome sequences from which we extracted and concatenated 16 syntenic, ribosomal proteins. This analysis revealed that CPR bacteria from Lac Pavin are highly diverse and include members from nearly all groups currently recognized within the CPR. Beyond the Absconditabacteria and Komeilibacteria mentioned above, we also identified organisms from the Saccharibacteria, Peregrinibacteria, and Microgenomates and recently delineated clades of the Parcubacteria [26], each of which each contains numerous sub-lineages (Fig. 4a).

By mapping metagenomic reads back to the draft CPR genomes, we determined the abundance of individual CPR organisms across the lake's oxygen gradient. While CPR bacteria were detectable throughout the water column and shallow sediments, they were most abundant in the lower interface and anoxic zone (55–90-m depth) (Fig. 1d). Intriguingly, different species from the same lineages populated each zone, e.g., different species of Berkelbacteria were uniquely associated with the interface, anoxic zone, and sediments (Fig. 4b, Table S7). In contrast, some lineages were only detected in a single compartment: for example, members of the

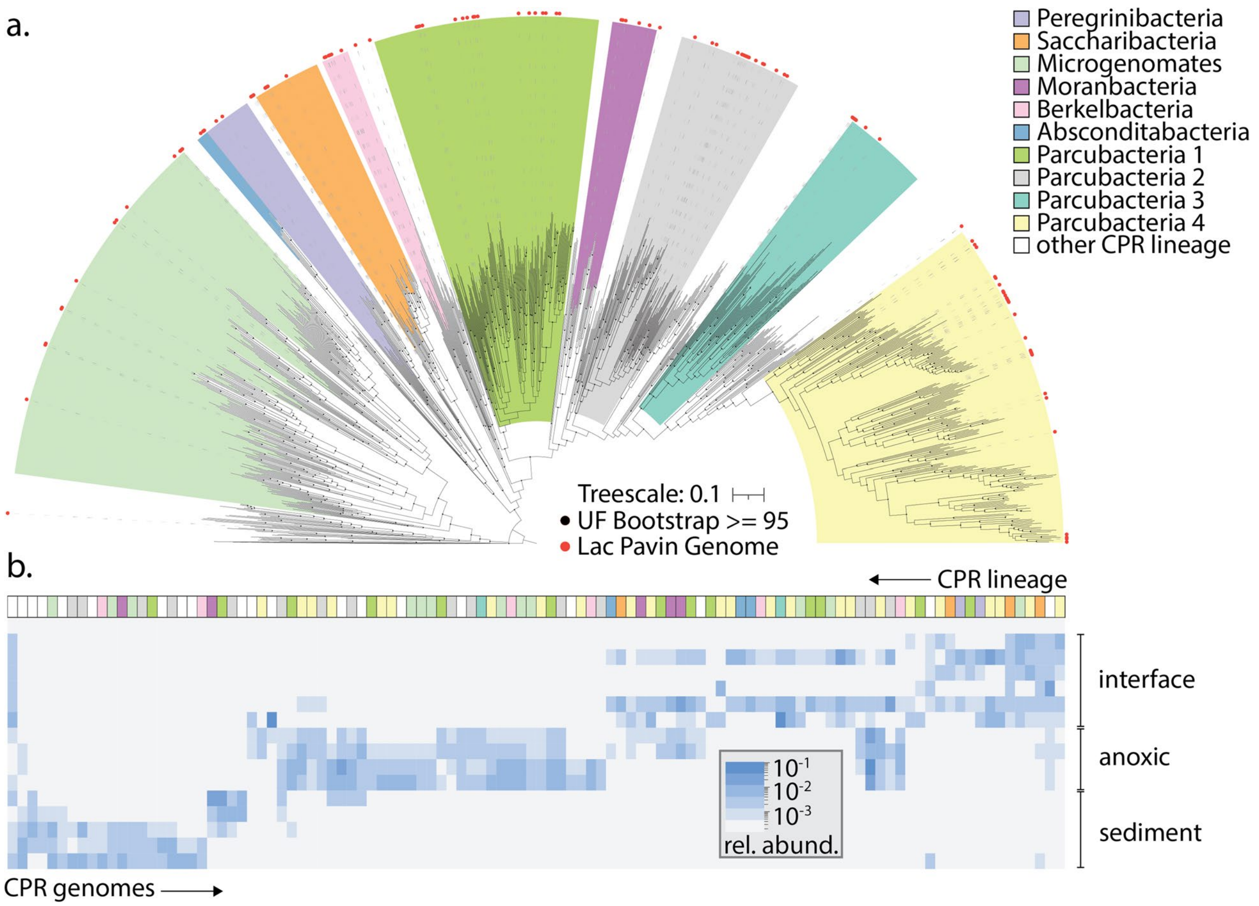

Fig. 4 Phylogenomics and abundance distribution of CPR bacteria in Lac Pavin. **a** Phylogenetic placement of CPR bacteria from Lac Pavin (red circles) using a set of reference genomes from across habitats. Major phylogenetic groupings are indicated; however, note that not all are equivalent in terms of rank. The scale bar represents the average number of substitutions per site. **b** Relative abundance (coverage) of individual CPR genomes across samples. Lineage assignment (matching the color scheme from panel **a**) is indicated at the top of panel **b**.

Saccharibacteria were mostly restricted to the lake interface (Fig. 4b, Table S7).

Broad patterns of gene content in CPR bacteria

To complement our analyses of RuBisCO in CPR, we examined the extent to which entire predicted proteomes of CPR bacteria vary throughout Lac Pavin's water column. Using a protein clustering approach that is agnostic to gene function [41], we resolved ~1500 protein families (≥ 5 proteins in size) among all CPR bacteria and visualized their distribution within the draft genome set. Clustering based on protein family presence/absence revealed a large "module" (or group of protein families) that were found in nearly all CPR genomes (Fig. S3), paralleling results from a prior analysis of CPR proteomes [41]. As expected, this "core" module contained key functions such as transcription, translation, and cell division. Outside of this core module, most other protein families were patchily distributed (Fig. S3).

While there was some evidence for groups of protein families that are specific to a small number of CPR genomes (narrow, horizontal strips in Fig. S3), no large modules were associated with organisms from the interface, anoxic zone, sediment, or combinations thereof. Detailed statistical analysis revealed only 22 protein families moderately to highly enriched in a specific compartment (enrichment ratio ≥3, $p \leq 0.05$, two-sided Fisher's exact test corrected for multiple comparisons). Most of these families (~82%) were enriched in CPR bacteria from the interface and included protein families that were involved in the metabolism of malate (K00027), zinc transport (K09815), and metabolism of certain amino acids (K00928) (Table S8). Also notable was a family confidently annotated as an Fe-Mn superoxide dismutase (K04564) that was moderately enriched (enrichment ratio ~2.7) in CPR bacteria from the interface relative to those from the anoxic zone and sediments (Table S8). Overall, these results suggest that while organisms from the

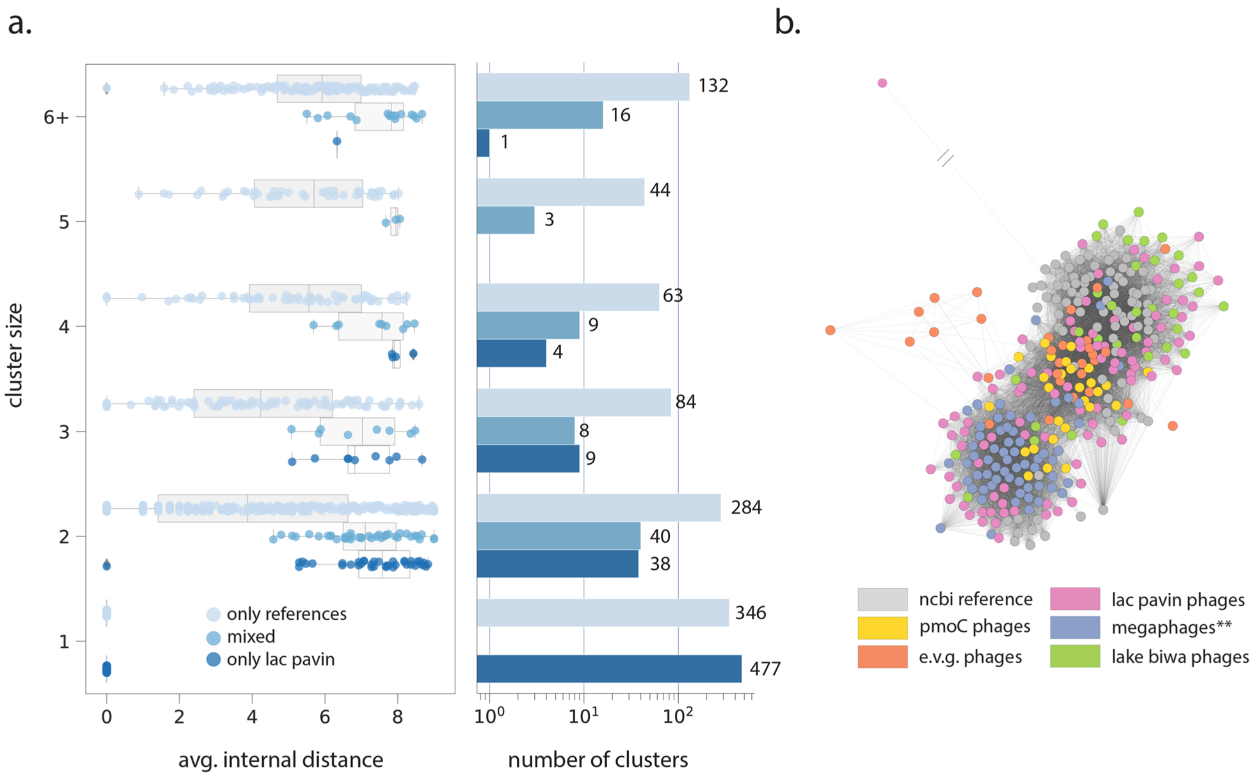

Fig. 5 Diversity of phages (≥50-kb genome size) in Lac Pavin. **a** Cluster composition, sequence diversity (described by average pairwise distance among nodes within the cluster), and size for a network analysis describing the relationship of Lac Pavin phages to known references based on protein content. "Mixed" clusters are those containing both reference and Lac Pavin phage genomes. **b** Subset of phage network analysis displaying interconnectedness between Lac Pavin phages and those from various reference databases. Shorter edges between nodes (phages) indicate a higher degree of similarity based on protein content. **Megaphages from diverse environments (Al-Shayeb, Sachdeva, et al., 2020)

interface may have slight differences in their metabolic capacities and stress responses compared to those from the anoxic zone and sediments, in general proteomes of CPR bacteria across the water column are not highly differentiated from one another.

Diversity and characteristics of recovered phage genomes

Viruses of bacteria (bacteriophages or phages) are important components of the "microbial loop" in the shallow regions of Lac Pavin and its associated sediments, where they are known to exhibit significant morphological and genomic diversity [62–64]. Here, we leveraged our metagenomic data to extend analyses of the diversity, distribution, and metabolism of Lac Pavin phages that were retained by our sampling/filtration methods (the "Materials and methods" section). We refined a preliminary set of 17,151 potential phage scaffolds to yield a final set of 767 non-redundant scaffolds (focusing on those larger phages with ≥50-kb genome size) that could be confidently identified as phages or bacterial/archaeal genome fragments with integrated prophages (the "Materials and methods" section, Table S9). Approximately, 7.7% of phage scaffolds were circularized and thus likely represent complete genome sequences (Table S9). We merged these newly reported phage scaffolds with those from our previous studies of the same environment [46, 47], yielding a total set of ~800 non-redundant phage scaffolds that accounted for ~0.19 to ~3.69% of sequencing reads (mean ~1.25%, Table S1). We then subjected the total genome set to a network analysis based on protein content.

This analysis revealed that many phages from Lac Pavin fall into broad clusters with known phages from diverse environments, suggesting they are part of well-sampled clades defined by protein content ("mixed" clusters, Fig. 5a). Of particular interest were Lac Pavin phage scaffolds that clustered with megaphages from diverse environments [46], phages encoding the S21 ribosomal protein from freshwater environments [47], and phages from another large freshwater lake in Japan, Lake Biwa [51] (Fig. 5b). We also recovered 52 phage clusters that contained only newly reported sequences from Lac Pavin (Fig. 5a). These clusters were generally small (5 or fewer related sequences) and contained

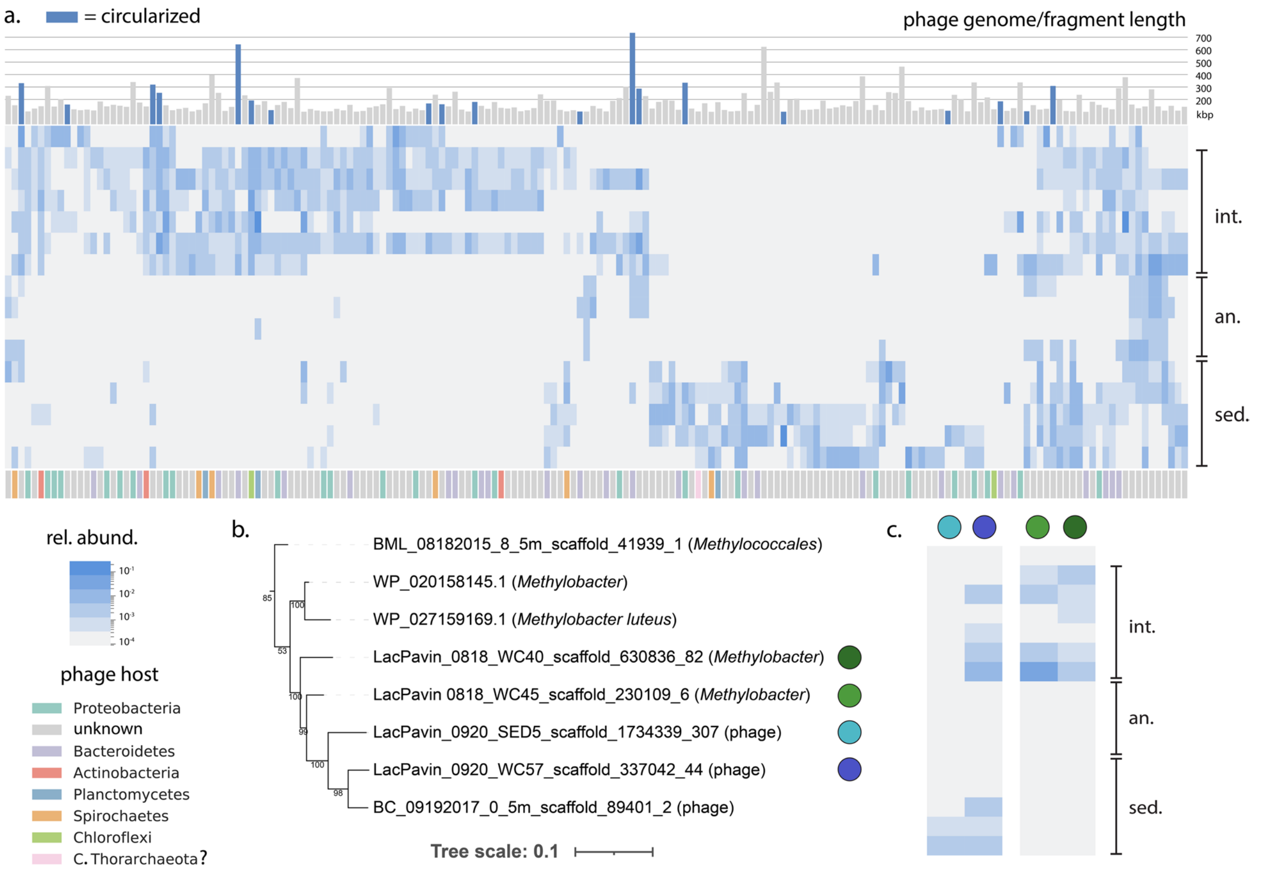

Fig. 6 Distribution and metabolism of phages in Lac Pavin. **a** Relative abundance of large phages (≥100 kbp) based on read mapping. Barplots above the heatmap represent fragment/genome size in kbp). Bars in blue represent phages that are likely to be circularized. The color strip below the heatmap indicates host assignment based on CRISPR-Cas targeting or taxonomic profiling of phage genes. **b** *pmoC* subtree depicting the evolutionary relationships between phage proteins and sequences from potential *Methylobacter* hosts. The full protein tree is available as a supplementary data file. **c** Relative abundance mapping of phages encoding *pmoC* (left) and potential hosts (right) through the water column and shallow sediments. Abbreviations: int., interface; an., anoxic zone; sed., sediments; rel., relative

similar or higher sequence diversity (as measured by pairwise distances) than clusters of the same size that included both Lac Pavin and reference sequences ("mixed") or only reference sequences (Fig. 5a). However, the majority of Lac Pavin phage scaffolds (n=477) did not cluster with any other phages (cluster size = 1, Fig. 5a). These results suggest that Lac Pavin viromes may also harbor considerable genomic novelty in addition to members of broadly distributed, well-sampled viral groups.

Phage distribution and metabolism

Motivated by the recent recognition of the importance of large phages across ecosystems, including Lac Pavin [46], we further analyzed the distribution of phages with genomes ≥100 kbp in length across the lake compartments using read mapping. Our results indicate that large phages are present throughout the water column and shallow sediments (Fig. 6a, Table S10). While many phages were restricted to certain compartments—e.g., the interface or sediments, we also found evidence for phages that were present in multiple zones (Fig. 6a, Table S10). The three largest megaphages, attaining ~735 kb (the largest phage genome yet reconstructed) [46], 642 kb, and 623 kb in length were present in the interface and shallow sediment, the interface, and sediment, respectively (Fig. 6a). The largest two of these phages were predicted to infect members of the Bacteroidetes (Table S9). Other phages predicted to infect Proteobacteria, Planctomycetes, and Spirochaetes bacteria were detected through the water column and sediments. Surprisingly, one phage from the sediment was predicted to infect a member of the Candidatus Thorarchaeota based on taxonomic profiling of its predicted proteins (the "Material and methods" section). However, given that we did not recover strong evidence for the host archaeon itself in the sediments, the strength of this phage-host association is unclear.

The extent to which environmental factors determine the gene content of phages is unclear. We tested whether any protein clusters derived from our phage genomes displayed signatures of enrichment in particular compartments of the lake. Paralleling our findings for CPR bacteria, we did not observe strong evidence for gene content differentiation by depth, with only two protein families statistically enriched in phages from the interface and anoxic zones, respectively ($p \leq 0.05$, two-sided Fisher's exact test corrected for multiple comparisons) (Table S11). While not statistically enriched in any compartment, we also found two phages encoding a *pmoC* gene. These genes have been previously reported in diverse phages from freshwater lakes [50] potentially implicating them in Lac Pavin's cycle of methane production (primarily in the sediments) and methane oxidation (primarily in the water column interface). While we were not able to obtain host assignments for these phages by CRISPR-Cas targeting or by taxonomic profiling, we observed that the phage *pmoC* sequences were closely related to those from *Methylobacter* genomes resolved from the Lac Pavin water column interface (Fig. 6b, Table S12), suggesting them as possible hosts. Surprisingly, relative abundance mapping showed that both phages harboring a *pmoC* gene were found in the deeper sediments, where potential *Methylobacter* hosts were not detected (Fig. 6c), suggesting either an alternate host in these samples or burial and preservation of phage particles in the sediments. Supporting the second explanation, we note many interface-associated phages are absent in the anoxic zone but present in the sediment (Fig. 6a). One of the *pmoC* phages was also found in the water column interface alongside *Methylobacter*, consistent with a possible host relationship (Fig. 6c).

Discussion

Despite differences in morphology and origin, meromictic lakes generally have similar physical characteristics that shape their microbial communities [65]. Our results show that like other similar lakes, the availability of oxygen across the water column and shallow sediments of Lac Pavin can be linked to changes in both microbial community composition and the metabolic capacities of its members. This oxygen gradient, alongside gradients of dissolved organic/inorganic carbon, sulfur, and iron, which also vary substantially with depth [66–69], likely play an important role in determining distributions of key bacterial and archaeal lineages in this ecosystem. Here, we provide high-quality genomic data for many of these lineages, including abundant members of the *Actinobacteria*, which dominate in the shallow water column of Pavin and other similar lakes [70, 71]. On the other hand, we found very little if any evidence for the photosynthetic purple/green sulfur bacteria that frequently occur in the oxycline of meromictic lakes, sometimes at extremely high densities [72–74]. While the absence of these bacteria may be due to relatively low light intensities in Lac Pavin's oxycline compared to those in other lakes, there is sufficient light to support some members of the *Cyanobacteria* in this zone (Table S2). Another factor may be the low abundance of sulfide in this zone of the water column.

Microbial metabolism is a major driver of biogeochemical cycling within lake ecosystems. We examined the extent to which changes in community composition across depth in Lac Pavin are accompanied by changes in metabolic potential, with a specific focus on a subset of C1 carbon metabolic pathways. Our analyses show that oxygen (and, likely, linked gradients of other compounds) impacts metabolic potential, in part by selecting for bacteria and archaea with certain C1 and carbon fixation pathways in different zones of the lake. In particular, the Wood-Ljungdahl and reverse TCA cycle showed strong signals of stratification and were essentially restricted to the lake's anoxic zone, consistent with findings from another stratified lake [75], although that lake is ~15 times shallower. Thus, potential for autotrophy via these pathways may vary predictably with geochemical conditions in a scale-independent manner. We found only minimal evidence for anaerobic methane oxidation (e.g., by ANME archaea) and thus expand upon prior evidence for spatial segregation of methane production and consumption in this lake. At a finer scale, we leveraged our genome-based approach to determine that the species composition of methane producers changes on the scale of centimeters within the shallow sediments (Fig. 2b). This observation hints at the possibility of niche partitioning among related methanogenic lineages, a trend which may continue into the deeper sediments not sampled in this study [76].

A prior study used amplicon sequencing to document CPR in Lac Pavin [5]. Here, we build upon those results by providing high-quality draft genomes for over 100 CPR bacteria, showing that Lac Pavin harbors a high abundance of CPR bacteria with representatives from virtually all of the major lineages. Broadly speaking, gene content in CPR bacteria, even within specific lineages, can be highly variable [26, 41] raising the important question as to what governs gene content in these organisms. One possibility is that ecosystem characteristics, like local geochemistry, drive the distribution of capacities in CPR bacteria. On the other hand, metabolic capacities of the host cell(s), and/or the spectrum of resources that are exchanged between them, may determine which pathways are present in a given CPR organism.

Here, we found that CPR bacteria display relatively little evidence for metabolic differentiation across depth in Lac Pavin, despite large variations in geochemical conditions. This suggests that CPR gene content is likely primarily determined by the metabolic capacities of their hosts, rather than geochemistry—for example, the rhodopsin found in some Lac Pavin Saccharibacteria (CPR) that may derive the essential cofactor retinal from their Actinobacteria hosts [77]. In Lac Pavin, this appears to stand in contrast with patterns of gene content observed for other, non-CPR organisms, which are likely shaped more heavily by the availability of oxygen, light, and other key nutrients. Synthesizing these observations, we postulate that geochemistry selects indirectly for CPR bacterial types via selecting for hosts with metabolisms that utilize the spectrum of resources provided across the water column.

One exception to the weak linkage between geochemistry and CPR metabolic traits is the moderate enrichment of superoxide dismutases in CPR bacteria from the partially aerobic interface relative to those from the anoxic water column and sediments. Proteins from this family may be involved in radical detoxification at the lake interface where oxygen levels are relatively high compared to deeper compartments (Fig. 1b). Other analyses have shown that superoxide dismutases are rare in CPR from anoxic human oral microbiomes [42], yet they do occur in organisms from oxic groundwaters [78]. Taken together, these findings suggest that this protein family may be one of the few components of CPR genomic repertoires that may correlate with environmental conditions across ecosystems.

Despite their reduced metabolic platforms, CPR bacteria probably still impact biogeochemistry throughout lakes. Along with genes that shape carbon, sulfur, and nitrogen cycles [79–81], RuBisCO is encoded by some CPR bacteria and could be used to fix carbon dioxide at some rate [39, 60, 61]. Our analyses of RuBisCO in Lac Pavin show that some CPR bacteria carry form II and III variants in the water column, but RuBisCO genes were not found in sediment-associated CPR. There, DPANN and other archaea tend to carry form III-related RuBisCO, suggesting spatial segregation of the organisms whose genomes encode this enzyme form. Given their wide environmental prevalence and the relatively little biochemical research on form III RuBisCOs [82], additional work is required to determine their rates of carbon dioxide incorporation. With this information it will be possible to quantify the potential contributions of form III RuBisCOs to carbon fixation in situ, including by CPR bacteria.

Phages, including the largest megaphages yet reported, were distributed throughout the water column and sediments in Lac Pavin, but like CPR bacteria, their metabolic capacities did not vary significantly by depth (Fig. 6, Table S10-11). This observation is consistent with the notion that gene content in phages is primarily selected for by interactions with their microbial hosts. This selection may extend to recently described metabolic genes in phages that may augment host metabolism of the same or similar compounds [83–86], like the phage *pmoC* genes described above (Fig. 6bc). Taken together, our findings raise the intriguing possibility that selection on CPR gene content may resemble that on phages more so than that shaping those of other bacteria, in part due to their shared strong reliance on hosts. However, there is currently no strong evidence for recent lateral transfer between CPR bacteria and known hosts validated by laboratory co-cultivation [87], as is known for phages. This suggests the evolutionary processes that shape gene selection in CPR bacteria and phages probably differ.

A final important feature of our study is the concurrent investigation of microbial communities in sediments in tandem with those in the overlying lake water column. In addition to finding functional partitioning likely driven by geochemical conditions, we also detected members of the Cyanobacteria, *pmoC* phages without their expected methanotrophic hosts, and phages from the interface zone that were present in the sediment. It is perhaps unsurprising that sediments can accumulate cells via settling; however, our findings raise the possibility that phage particles are preserved there and thus are capable of contributing to long term carbon storage.

Conclusions

Overall, our analyses provide insights into the factors shaping the gene content of CPR bacteria, non-CPR bacteria, archaea, and phages through a stratified lake, revealing the differential impact of geochemistry on their genetic repertoires. These findings are more broadly relevant to understanding the ecological and evolutionary constraints upon CPR genomes across the many environments in which they are found.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40168-022-01416-7.

Additional file 1: Figure S1. Protein tree describing the evolutionary relationships between putative methane monooxygenases from Lac Pavin microbial genomes and reference sequences from the broader copper membrane monooxygenase superfamily (Diamond et al. 2022). Abbreviations: AMO, ammonia monooxygenase; PMO, particulate methane monooxygenase; MMO, methane monooxygenase; HMO, hydrocarbon monooxygenase. **Figure S2.** Genetic potential for select carbon metabolism across compartments Lac Pavin by phylogeny. a) Relative abundance of genomes encoding various pathways across metagenomic samples. Bubble size indicates the combined coverage of all organisms of a given phylum-level designation that encode a given pathway in each sample, normalized by the total coverage of all genomes in the quality-filtered, non-redundant set. This percent coverage is averaged across all samples within a lake compartment. **Figure S3.** Proteome content in the CPR bacteria by depth a) Heatmap describing the presence/absence of ~1,500 protein families (≥5 sequence members) across 106 genomes of CPR bacteria. Darker patches indicate presence, while lighter patches indicate absence. Protein families are hierarchically clustered by their distribution pattern across genomes. CPR genomes are organized by their distribution across the lake water column and sediments. In panel at right, darker patches signify that a given genome was present/detectable in a given compartment, whereas a lighter patch indicates absence/below detection threshold. Dashed lines distinguish sets of genome with similar distribution patterns, e.g. the bottom represents CPR bacteria detected only in the sediments.

Additional file 2: Table S1. Characteristics of metagenomic samples analyzed in this study. Perc_reads_* columns indicate the percentage of reads that map stringently (95% identity) to the dereplicated archaeal/bacterial or phage genome set, respectively. **Table S2.** Phylogenetic affiliation and relative abundance of ribosomal protein S3 (rps3) 'species groups' used to define community composition. **Table S3.** Characteristics of bacterial and archaeal genomes analyzed in this study. **Table S4.** Metabolic HMMs involved in carbon metabolism and manually-assigned score thresholds. **Table S5.** Presence of metabolic HMMs/pathways across the bacterial/archaeal genome set, and their abundance by sample percentage_relative_coverage refers to the percent of total coverage of all genomes in the quality-filtered, non-redundant set within a sample attributable to a given genome. **Table S6.** Phylogenetic affiliation, form affiliation, and relative abundance of RuBisCO genes across samples percentage_relative_coverage refers to the percent of all coverage to RuBisCO-containing scaffolds within a sample attributable to a given scaffold. **Table S7.** Relative abundance (percentage relative coverage of all genomes in the non-redundant set) of CPR bacteria genomes across samples. **Table S8.** Metrics of enrichment for CPR protein families by lake compartment. **Table S9.** Characteristics of phage genomes analyzed in this study. **Table S10.** Relative abundance (percentage relative coverage of all genomes in the non-redundant set) of phage genomes across samples. **Table S11.** Metrics of enrichment for phage protein families by lake compartment. **Table S12.** Characteristics of *Methylobacter*/phage pmoC sequences analyzed in this study.

Acknowledgements
We thank Yue Clare Lou, Spencer Diamond, Shufei Lei, Adair Borges, Marie Schoelmerich, Lin-Xing Chen, Alex Thomas, and Rohan Sachdeva for helpful discussions and informatic support. We also thank Guillaume Borrel and François Enault for their comments on the manuscript, and Lily Law, Christopher Lefevre, and Didier Jézéquel for their assistance with sampling logistics. We also thank the Innovative Genomics Institute at UC Berkeley, Y. Justin Choi, Shana McDevitt, and QB3 Genomics at UC Berkeley.

Authors' contributions
A.L.J, C.J.C., A.C.L., and J.F.B. designed the project. A.L.J., J.L., and J.F.B. performed the bioinformatic analyses and genome curation. A.L.J, C.J.C., C.B., H.B., J.C., F.P., A.H.L.J., and A.C.L. performed the sample collection. C.B., H.B., J.C., F.P., A.H.L.J., and A.C.L. performed the sample preparation and DNA extractions. A.L.J. and J.F.B. wrote the manuscript. All authors made comments on the manuscript. The authors read and approved the final manuscrpt.

Funding
Funding was provided by the Berkeley Fellowship to A.L.J. and Moore Foundation Grant 71785 to J.F.B. Sequencing was partially supported by the NIH S10 OD018174 Instrumentation Grant.

Availability of data and materials
Read data and draft genomes from this study are available through NCBI at PRJNA852163. Genome accession information for bacterial/archaeal genomes is also listed in Table S3. Custom code for the described analyses are available on GitHub (https://github.com/alexanderjaffe/lac-pavin-metagenomics). All supplementary figures, tables, data files, and genomes (including phage/viral genomes) are also available through Zenodo (https://doi.org/10.5281/zenodo.6857590).

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
Jillian Banfield is a founder of Metagenomi. The other authors declare that they have no competing interests.

Received: 20 July 2022 Accepted: 7 November 2022
Published online: 25 January 2023

References
1. Hall KJ, Northcote TG. Meromictic lakes. Encyclopedia of lakes and reservoirs: Springer Netherlands Dordrecht; 2012. p. 519–24.
2. Boehrer B, von Rohden C, Schultze M. Physical features of meromictic lakes: stratification and circulation. In: Gulati RD, Zadereev ES, Degermendzhi AG, editors. Ecology of Meromictic Lakes. Cham: Springer International Publishing; 2017. p. 15–34.
3. Andrei A-Ş, Robeson MS 2nd, Baricz A, Coman C, Muntean V, Ionescu A, et al. Contrasting taxonomic stratification of microbial communities in two hypersaline meromictic lakes. ISME J. 2015;9:2642–56.
4. Lehours A-C, Evans P, Bardot C, Joblin K, Gérard F. Phylogenetic diversity of archaea and bacteria in the anoxic zone of a meromictic lake (Lake Pavin, France). Appl Environ Microbiol. 2007;73:2016–9.
5. Borrel G, Lehours A-C, Bardot C, Bailly X, Fonty G. Members of candidate divisions OP11, OD1 and SR1 are widespread along the water column of the meromictic Lake Pavin (France). Arch Microbiol. 2010;192:559–67.
6. Peura S, Eiler A, Bertilsson S, Nykänen H, Tiirola M, Jones RI. Distinct and diverse anaerobic bacterial communities in boreal lakes dominated by candidate division OD1. ISME J. 2012;6:1640–52.
7. Tran PQ, Bachand SC, McIntyre PB, Kraemer BM, Vadeboncoeur Y, Kimirei IA, et al. Depth-discrete metagenomics reveals the roles of microbes in biogeochemical cycling in the tropical freshwater Lake Tanganyika. ISME J. 2021;15:1971–86.
8. Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, et al. Unusual biology across a group comprising more than 15% of domain Bacteria. Nature. 2015;523:208–11.
9. Castelle CJ, Banfield JF. Major new microbial groups expand diversity and alter our understanding of the tree of life. Cell. 2018;172:1181–97.
10. Moreira D, Zivanovic Y, López-Archilla AI, Iniesto M, López-García P. Reductive evolution and unique predatory mode in the CPR bacterium Vampirococcus lugosii. Nat Commun. 2021;12:2454.
11. Langenheder S, Kisand V, Wikner J, Tranvik LJ. Salinity as a structuring factor for the composition and performance of bacterioplankton degrading riverine DOC. FEMS Microbiol Ecol. 2003;45:189–202.
12. He C, Keren R, Whittaker ML, Farag IF, Doudna JA, Cate JHD, et al. Genome-resolved metagenomics reveals site-specific diversity of episymbiotic CPR bacteria and DPANN archaea in groundwater ecosystems. Nat Microbiol. 2021. https://doi.org/10.1038/s41564-020-00840-5.
13. Joshi NA, Fass JN. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. 2011. Available at https://github.com/najoshi/sickle.
14. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015;31:1674–6.
15. Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics. 2012;28:1420–8.

16. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
17. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119.
18. Kang D, Li F, Kirton ES, Thomas A, Egan RS, An H, et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies [Internet]. https://doi.org/10.7287/peerj.preprints.27522v1.
19. Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG, et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nat Microbiol. 2018;3:836–43.
20. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics. 2019. https://doi.org/10.1093/bioinformatics/btz848.
21. Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, Probst AJ, et al. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. Nat Commun. 2016;7:13219.
22. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25:1043–55.
23. Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, et al. Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ. 2015;3:e1319.
24. Orakov A, Fullam A, Coelho LP, Khedkar S, Szklarczyk D, Mende DR, et al. GUNC: detection of chimerism and contamination in prokaryotic genomes. Genome Biol. 2021;22:178.
25. Hug LA, Castelle CJ, Wrighton KC, Thomas BC, Sharon I, Frischkorn KR, et al. Community genomic analyses constrain the distribution of metabolic traits across the Chloroflexi phylum and indicate roles in sediment carbon cycling. Microbiome. 2013;1:22.
26. Jaffe AL, Castelle CJ, Matheus Carnevali PB, Gribaldo S, Banfield JF. The rise of diversity in metabolic platforms across the Candidate Phyla Radiation [Internet]. BMC Biol. 2020. https://doi.org/10.1186/s12915-020-00804-5.
27. Castelle CJ, Wrighton KC, Thomas BC, Hug LA, Brown CT, Wilkins MJ, et al. Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. Curr Biol. 2015;25:690–701.
28. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80.
29. Criscuolo A, Gribaldo S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol Biol. 2010;10:210.
30. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol. 2020;37:1530–4.
31. Diamond S, Andeer PF, Li Z, Crits-Christoph A, Burstein D, Anantharaman K, et al. Mediterranean grassland soil C–N compound turnover is dependent on rainfall and depth, and is mediated by genomically divergent microorganisms. Nat Microbiol. 2019;4:1356–1367.
32. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26:2460–1.
33. Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods. 2021;18:366–8.
34. Jaffe AL, Fuster M, Schoelmerich MC, Chen L-X, Colombet J, Billard H, et al. Long-term incubation of lake water enables genomic sampling of consortia involving planctomycetes and candidate phyla radiation bacteria. mSystems. 2022:e0022322.
35. Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, Goto S, et al. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. Bioinformatics. 2020;36:2251–2.
36. Zhou Z, Tran PQ, Breister AM, Liu Y, Kieft K, Cowley ES, et al. METABOLIC: high-throughput profiling of microbial genomes for functional traits, metabolism, biogeochemistry, and community-scale functional networks. Microbiome. 2022;10:33.
37. Diamond S, Lavy A, Crits-Christoph A, Matheus Carnevali PB, Sharrar A, Williams KH, et al. Soils and sediments host Thermoplasmata archaea encoding novel copper membrane monooxygenases (CuMMOs). ISME J. 2022;16:1348–62.
38. Kono T, Mehrotra S, Endo C, Kizu N, Matusda M, Kimura H, et al. A RuBisCO-mediated carbon metabolic pathway in methanogenic archaea. Nat Commun. 2017;8:14007.
39. Jaffe AL, Castelle CJ, Dupont CL, Banfield JF. Lateral gene transfer shapes the distribution of RuBisCO among candidate phyla radiation bacteria and DPANN archaea. Mol Biol Evol. 2019;36:435–46.
40. Olm MR, Brown CT, Brooks B, Banfield JF. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J. 2017;11:2864–8.
41. Méheust R, Burstein D, Castelle CJ, Banfield JF. The distinction of CPR bacteria from other bacteria based on protein family content. Nat Commun. 2019;10:4173.
42. Jaffe AL, Thomas AD, He C, Keren R, Valentin-Alvarado LE, Munk P, et al. Patterns of gene content and co-occurrence constrain the evolutionary path toward animal association in candidate phyla radiation bacteria. MBio. 2021;12:e0052121.
43. 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. 2020;8:90.
44. Guo J, Bolduc B, Zayed AA, Varsani A, Dominguez-Huerta G, Delmont TO, et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses [Internet]. Microbiome. 2021. https://doi.org/10.1186/s40168-020-00990-y.
45. Nayfach S, Camargo AP, Schulz F, Eloe-Fadrosh E, Roux S, Kyrpides NC. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nat Biotechnol. 2021;39:578–85.
46. Al-Shayeb B, Sachdeva R, Chen L-X, Ward F, Munk P, Devoto A, et al. Clades of huge phages from across Earth's ecosystems. Nature. 2020;578:425–31.
47. Chen L-X, Jaffe AL, Borges AL, Penev PI, Nelson TC, Warren LA, et al. Phage-encoded ribosomal protein S21 expression is linked to late-stage phage replication. ISME Commun. 2022;2:1–10.
48. Borges AL, Lou YC, Sachdeva R, Al-Shayeb B, Penev PI, Jaffe AL, et al. Widespread stop-codon recoding in bacteriophages may regulate translation of lytic genes. Nat Microbiol. 2022;7:918–27.
49. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25:1972–3.
50. Chen L-X, Méheust R, Crits-Christoph A, McMahon KD, Nelson TC, Slater GF, et al. Large freshwater phages with the potential to augment aerobic methane oxidation. Nat Microbiol. 2020;5:1504–15.
51. Okazaki Y, Nishimura Y, Yoshida T, Ogata H, Nakano S-I. Genome-resolved viral and cellular metagenomes revealed potential key virus-host interactions in a deep freshwater lake. Environ Microbiol. 2019;21:4740–54.
52. Bin JH, Bolduc B, Zablocki O, Kuhn JH, Roux S, Adriaenssens EM, Brister JR, Kropinski AM, Krupovic M, Lavigne R, Turner D. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. Nat Biotechnol. 2019;37(6):632–9.
53. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13:2498–504.
54. İnceoğlu Ö, Llirós M, Crowe SA, García-Armisen T, Morana C, Darchambeau F, et al. Vertical distribution of functional potential and active microbial communities in meromictic Lake Kivu. Microb Ecol. 2015;70:596–611.
55. Lopes F, Viollier E, Thiam A, Michard G, Abril G, Groleau A, et al. Biogeochemical modelling of anaerobic vs. aerobic methane oxidation in a meromictic crater lake (Lake Pavin, France). Appl Geochem. 2011;26:1919–32.
56. Biderre-Petit C, Jézéquel D, Dugat-Bony E, Lopes F, Kuever J, Borrel G, et al. Identification of microbial communities involved in the methane cycle of a freshwater meromictic lake. FEMS Microbiol Ecol. 2011;77:533–45.
57. Chistoserdova L, Jenkins C, Kalyuzhnaya MG, Marx CJ, Lapidus A, Vorholt JA, et al. The enigmatic planctomycetes may hold a key to the origins of methanogenesis and methylotrophy. Mol Biol Evol. 2004;21:1234–41.
58. Goenrich M, Thauer RK, Yurimoto H, Kato N. Formaldehyde activating enzyme (Fae) and hexulose-6-phosphate synthase (Hps) in Methanosarcina barkeri: a possible function in ribose-5-phosphate biosynthesis. Arch Microbiol. 2005;184:41–8.
59. Borrel G, Jézéquel D, Biderre-Petit C, Morel-Desrosiers N, Morel J-P, Peyret P, et al. Production and consumption of methane in freshwater lake ecosystems. Res Microbiol. 2011;162:832–47.

60. Wrighton KC, Castelle CJ, Varaljay VA, Satagopan S, Brown CT, Wilkins MJ, et al. RubisCO of a nucleoside pathway known from Archaea is found in diverse uncultivated phyla in bacteria. ISME J. 2016;10:2702–14.
61. Sato T, Atomi H, Imanaka T. Archaeal type III RuBisCOs function in a pathway for AMP metabolism. Science. 2007;315:1003–6.
62. Bettarel Y, Sime-Ngando T, Amblard C, Carrias J-F, Sargos D, Garabétian F, et al. The functional importance of bacteriophages in the microbial loop of an oligomesotrophic lake over a diel cycle. Ann Limnol Int J Limnol. 2002;38:263–9.
63. Borrel G, Colombet J, Robin A, Lehours A-C, Prangishvili D, Sime-Ngando T. Unexpected and novel putative viruses in the sediments of a deep-dark permanently anoxic freshwater habitat. ISME J. 2012;6:2119–27.
64. Colombet J, Robin A, Sime-Ngando T. Genotypic, size and morphological diversity of virioplankton in a deep oligomesotrophic freshwater lake (Lac Pavin, France). J Environ Sci. 2017;53:48–59.
65. Baatar B, Chiang P-W, Rogozin DY, Wu Y-T, Tseng C-H, Yang C-Y, et al. Bacterial communities of three saline Meromictic Lakes in Central Asia. PLoS One. 2016;11:e0150847.
66. Michard G, Viollier E, Jézéquel D, Sarazin G. Geochemical study of a crater lake: Pavin Lake, France—Identification, location and quantification of the chemical reactions in the lake. Chem Geol. 1994;115:103–15.
67. Assayag N, Jézéquel D, Ader M, Viollier E, Michard G, Prévot F, et al. Hydrological budget, carbon sources and biogeochemical processes in Lac Pavin (France): constraints from δ18O of water and δ13C of dissolved inorganic carbon. Appl Geochem. 2008;23:2800–16.
68. Sime-Ngando T, Boivin P, Chapron E, Jezequel D, Meybeck M. Lake Pavin: History, geology, biogeochemistry, and sedimentology of a deep meromictic maar lake: Springer; 2016.
69. Tisserand D, Guédron S, Viollier E, Jézéquel D, Rigaud S, Campillo S, et al. Mercury, organic matter, iron, and sulfur co-cycling in a ferruginous meromictic lake. Appl Geochem. 2022;146:105463.
70. Hugoni M, Vellet A, Debroas D. Unique and highly variable bacterial communities inhabiting the surface microlayer of an oligotrophic lake. Aquat Microb Ecol. 2017;79:115–25.
71. Ghai R, McMahon KD, Rodriguez-Valera F. Breaking a paradigm: cosmopolitan and abundant freshwater actinobacteria are low GC. Environ Microbiol Rep. 2012;4:29–35.
72. Hamilton TL, Bovee RJ, Thiel V, Sattin SR, Mohr W, Schaperdoth I, et al. Coupled reductive and oxidative sulfur cycling in the phototrophic plate of a meromictic lake. Geobiology. 2014;12:451–68.
73. Meyer KM, Macalady JL, Fulton JM, Kump LR, Schaperdoth I, Freeman KH. Carotenoid biomarkers as an imperfect reflection of the anoxygenic phototrophic community in meromictic Fayetteville Green Lake. Geobiology. 2011;9:321–9.
74. Panwar P, Allen MA, Williams TJ, Hancock AM, Brazendale S, Bevington J, et al. Influence of the polar light cycle on seasonal dynamics of an Antarctic lake microbial community. Microbiome. 2020;8:116.
75. Peura S, Buck M, Aalto SL, Morales SE, Nykänen H, Eiler A. Novel autotrophic organisms contribute significantly to the internal carbon cycling potential of a Boreal Lake. MBio. 2018:9. Available from:. https://doi.org/10.1128/mBio.00916-18.
76. Borrel G, Lehours A-C, Crouzet O, Jézéquel D, Rockne K, Kulczak A, et al. Stratification of Archaea in the deep sediments of a freshwater meromictic lake: vertical shift from methanogenic to uncultured archaeal lineages. PLoS One. 2012;7:e43346.
77. Jaffe AL, Konno M, Kawasaki Y, Kataoka C, Béjà O, Kandori H, et al. Saccharibacteria harness light energy using type-1 rhodopsins that may rely on retinal sourced from microbial hosts. ISME J. 2022;16:2056–9.
78. Chaudhari NM, Overholt WA, Figueroa-Gonzalez PA, Taubert M, Bornemann TLV, Probst AJ, et al. The economical lifestyle of CPR bacteria in groundwater allows little preference for environmental drivers [Internet]. bioRxiv. 2021:2021.07.28.454184 Available from: https://www.biorxiv.org/content/10.1101/2021.07.28.454184v1.abstract. [Cited 2021 Nov 12].
79. Danczak RE, Johnston MD, Kenah C, Slattery M, Wrighton KC, Wilkins MJ. Members of the Candidate Phyla Radiation are functionally differentiated by carbon- and nitrogen-cycling capabilities. Microbiome. 2017;5:112.
80. Wrighton KC, Thomas BC, Sharon I, Miller CS, Castelle CJ, VerBerkmoes NC, et al. Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. Science. 2012;337:1661–5.
81. Castelle CJ, Brown CT, Anantharaman K, Probst AJ, Huang RH, Banfield JF. Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. Nat Rev Microbiol. 2018;16:629–45.
82. Flamholz AI, Prywes N, Moran U, Davidi D, Bar-On YM, Oltrogge LM, et al. Revisiting trade-offs between Rubisco kinetic parameters. Biochemistry. 2019;58:3365–76.
83. Hurwitz BL, Hallam SJ, Sullivan MB. Metabolic reprogramming by viruses in the sunlit and dark ocean. Genome Biol. 2013;14:R123.
84. Thompson LR, Zeng Q, Kelly L, Huang KH, Singer AU, Stubbe J, et al. Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. Proc Natl Acad Sci U S A. 2011;108:E757–64.
85. Kieft K, Zhou Z, Anderson RE, Buchan A, Campbell BJ, Hallam SJ, et al. Ecology of inorganic sulfur auxiliary metabolism in widespread bacteriophages. Nat Commun. 2021;12:3503.
86. Gazitúa MC, Vik DR, Roux S, Gregory AC, Bolduc B, Widner B, et al. Potential virus-mediated nitrogen cycling in oxygen-depleted oceanic waters. ISME J. 2021;15:981–98.
87. McLean JS, Bor B, Kerns KA, Liu Q, To TT, Solden L, et al. Acquisition and adaptation of ultra-small parasitic reduced genome bacteria to mammalian hosts. Cell Rep. 2020;32:107939.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.