Ecogenomics sheds light on diverse lifestyle strategies in freshwater CPR

Maria-Cecilia Chiriac (cecilia.chiriac@icbcluj.ro)
Biology Centre of the Academy of Sciences of the Czech Republic
Paul-Adrian Bulzu
Biology Centre of the Academy of Sciences of the Czech Republic
Adrian-Stefan Andrei
University of Zurich
Yusuke Okazaki
Kyoto University
Shin-ichi Nakano
Kyoto University
Markus Haber
Biology Centre of the Academy of Sciences of the Czech Republic
Vinicius Silva Kavagutti
Biology Centre of the Academy of Sciences of the Czech Republic
Paul Layoun
Biology Centre of the Academy of Sciences of the Czech Republic
Rohit Ghai
Biology Centre of the Academy of Sciences of the Czech Republic
Michaela M. Salcher
Biology Centre of the Academy of Sciences of the Czech Republic

Research Article

Keywords: Patescibacteria, CPR, freshwater lakes, metagenomics, genome reduction, metabolism, lifestyle, CARD-FISH

Posted Date: March 23rd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-776685/v3

Version of Record: A version of this preprint was published at Microbiome on June 4th, 2022. See the published version at https://doi.org/10.1186/s40168-022-01274-3.

Abstract

Background. The increased use of metagenomics and single-cell genomics led to the discovery of organisms from phyla with no cultivated representatives and proposed new microbial lineages such as the candidate phyla radiation (CPR, or Patescibacteria). These bacteria have peculiar ribosomal structures, reduced metabolic capacities, small genome and cell sizes, and a general host-associated lifestyle was proposed for the radiation. So far, most CPR genomes were obtained from groundwaters, however, their diversity, abundance, and role in surface freshwaters is largely unexplored. Here we attempt to close these knowledge gaps by deep metagenomic sequencing of 119 samples of 17 different freshwater lakes located in Europe and Asia. Moreover, we applied Fluorescence *in situ* Hybridization followed by Catalyzed Reporter Deposition (CARD-FISH) for a first visualization of distinct CPR lineages in freshwater samples.

Results. A total of 174 dereplicated metagenome-assembled genomes (MAGs) of diverse CPR lineages were recovered from the investigated lakes, with a higher prevalence from hypolimnion samples (162 MAGs). They have reduced genomes (median size 1 Mbp) and were generally found in low abundances (0.02 – 14.36 coverage/Gb) and with estimated slow replication rates. The analysis of genomic traits and CARD-FISH results showed that the radiation is an eclectic group in terms of metabolic capabilities and potential lifestyles, ranging from what appear to be free-living lineages to host- or particle-associated groups. Although some complexes of the electron transport chain were present in the CPR MAGs, together with ion-pumping rhodopsins and heliorhodopsins, we believe that they most probably adopt a fermentative metabolism. Terminal oxidases might function in O_2 scavenging, while heliorhodopsins could be involved in mitigation against oxidative stress.

Conclusions. A high diversity of CPR MAGs was recovered, and distinct CPR lineages did not seem to be limited to lakes with specific trophic states. Their reduced metabolic capacities resemble the ones described for genomes in groundwater and animal-associated samples, apart from Gracilibacteria that possesses more complete metabolic pathways. Even though this radiation is mostly host-associated, we also observed organisms from different clades (ABY1, Paceibacteria, Saccharimonadia) that appear to be unattached to any other organisms or were associated with 'lake snow' particles (ABY1, Gracilibacteria), suggesting a broad range of potential life-strategies in this phylum.

Background

Patescibacteria, also known as Candidate Phyla Radiation (CPR) is a bacterial phylum [1] with peculiar ribosomal structures [2], reduced metabolic capacities, and small genome and cell sizes [3]. While the majority of this phylum remains uncultivated, some bacteria of the Saccharimonadia group (formally known as TM7) were successfully cultivated alongside their actinobacterial hosts [4–6] and 'Ca. Vampirococcus lugosii' and 'Ca. Absconditicoccus praedator', members of the Gracilibacteria class, were cultivated along their Gammaproteobacterial hosts [7, 8]. Another CPR bacterium ('Ca. Sonnebornia yantaiensis') was shown to be an endosymbiont of the protist *Paramecium bursaria* [9]. Together, these

observations lead to a general assumption of an unusual host-associated, symbiotic or parasitic lifestyle for CPRs [3, 10, 11]. At this point, most CPR genomes have been recovered from groundwater samples [11–14], with only a few studies focusing on CPRs in soils [15, 16], marine systems [17, 18], boreal peatland ecosystem [19] and freshwater lakes [20, 21] although they have been frequently reported in 16S rRNA amplicon studies [22, 23]. Because of this gap in knowledge, our study aims to present the first comprehensive investigation of CPRs in freshwater lakes on a wide scale, evaluating aspects that are still largely unknown, such as their abundance, distribution, and metabolic capacities. Additionally, we wanted to understand if freshwater CPR genomes have peculiar characteristics and whether a general host-associated lifestyle is common in the whole radiation.

Studies on CPR in freshwater lakes showed that they represent about 3–4.5% in the bathypelagic of Lake Baikal, ~ 2.5% in Lake Tanganyika and between 12–13% in a permafrost thaw lake in the 0.22 µm fraction [20, 21, 24]. Their ecological role is little understood, but they appear to have enzymatic resistance to O_2 [20, 21], while being able to ferment acetate and pyruvate. Additionally, metagenome assembled genomes (MAGs) from the permafrost lake encoded many carbohydrate active enzymes, which probably have a key role in transforming complex organic matter present in freshwater lakes [21].

Previous studies hypothesized that some CPR lineages could have the capacity for independent survival [3, 10, 25], but no direct evidence exists at the moment. In order to investigate if a universal lifestyle is adopted by the whole radiation, the association of different CPRs lineages with other organisms in freshwater lakes was analyzed using multiple CARD-FISH probes. In addition, we used the representative genomes from GTDB (Genome Taxonomy Database) [1] together with the MAGs obtained in this study to compare the genome characteristics of CPRs as a whole group with known parasitic/symbiotic and free-living freshwater bacteria with the scope of finding where on this spectrum different CPR groups fit. For instance, we considered genome sizes, the number of genes, coding densities, GC content or the number of pseudogenes, aspects that evolve in different directions in host-associated *vs.* free-living bacteria when genome streamlining occurs [26]. The completeness of metabolic pathways and the presence of secretion systems was also checked in each genome, as the lack of the ability to synthesize amino acids, nucleotides, lipids, cofactors, or to generate ATP could indicate a dependency upon other microbes or their local environments, while some secretion systems may imply a direct interaction with a host (e.g. Type III, IV, VI and VII secretion systems) [27].

Materials And Methods

Sample collection, DNA extraction and metagenomic sequencing

Freshwater samples (n = 119) were collected from 17 lakes located in Europe and Asia (Supplementary Figure S1, Supplementary Table S1). Details about the sampling procedures, DNA extraction and sequencing were previously published for Lakes Baikal [20], Biwa [28], Medard, Zurich and Constance [29] as well as for Jiřická pond and the Římov Reservoir [30]. The Ikeda samples were collected after 5 µm

prefiltration on 0.22 µm Sterivex cartridge until the filter got clogged, which occurred after 2–5 L lake water were filtered. Then DNA was extracted by PowerSoil DNA Isolation Kit (MoBio Laboratories). For the other 10 lakes, approximately 20 L of water were collected from the epi- and hypolimnion (Supplementary Table S1). The water was sequentially filtered through a 20 µm mesh plankton net to remove larger organisms, followed by 5 µm, and 0.22 µm polyethersulfone membrane filters (Millipore, Merck, Darmstadt, DE) until they got completely clogged. These filters were immersed in DNA/RNA Shield (Zymo Research, Irvine, CA, USA) and stored at -80°C until later use. The 0.22 µm filters were cut with sterile scissors into small pieces, followed by DNA purification using the ZR Soil Microbe DNA MiniPrepTM kit (Zymo Research, Irvine, CA, USA) according to the manufacturer's instructions. Shotgun metagenomic sequencing (2 x 151 bp) was performed using the Illumina Novaseq 6000 machine or NextSeq 500 platform.

Data preprocessing, metagenomic assembly and binning

BBMap project tools (https://github.com/BioInfoTools/BBMap/) were used to preprocess the raw data [31]. In brief, the bbduk.sh script was used to remove poor quality reads (qtrim = rl trimq = 18), the phiX and p-Fosil2 control reads (k = 21 ref = vectorfile ordered cardinality) and the Illumina adaptors (k = 21 ref = adapterfile ordered cardinality). Preprocessed reads were assembled *de novo* with MEGAHIT v1.1.4-2 [32] using default parameters and the following selection of k-mers: 29, 49, 69, 89, 109, 119, 129, 149. Only contigs ≥ 3kbp were further used for binning. Quality filtered reads were mapped on the contigs to obtain mean base coverage for each contig, and hybrid binning (tetranucleotide frequencies and coverage data) was performed using MetaBAT2 with default parameters [33]. Gene prediction for all bins was performed with Prodigal v2.6.3 [34] and the taxonomy of the genes was determined by screening each gene against the genes in the GTDB r89 [1] and UniProt release 2020-02 with MMseq2 [35] (blast criteria for each protein: evalue 1e-3, similarity 10%, coverage 10%, bitscore 50) and retrieving the taxonomy of the best hit gene. Contigs that had less than 30% of the genes assigned to the dominant taxonomic class for their bin were considered contaminants and were eliminated. Viral sequences were predicted with both VirSorter [36] and Vibrant [37] tools, and contigs with > 25% of genes of viral origin were discarded. Completeness of CPR bins was assessed using a set of 43 single-copy genes (SCGs) (Supplementary Table S2) [38]. CheckM v1.0.18 [1] was run using these 43 SCGs in order to estimate bin completeness, contamination and strain heterogeneity. Bins with > 40% completeness (based on the 43 SCGs set) and < 5% contamination were selected for further analysis (282 bins). The set of high-quality bins obtained from each lake (Supplementary Table S3) was dereplicated using dRep (average nucleotide identity (ANI) > 99%) [39], resulting in 174 representative bins (bins with highest dRep score, a metric that considers genome sizes, levels of completeness and contamination, strain-heterogeneity, N50 as well as how similar each genome is to all other genomes in their cluster) (Supplementary Table S4). These bins were classified with GTDB-Tk v1.3.0 [40] toolkit (https://ecogenomics.github.io/GTDBTk/) based on the GTDB r89. SSU rRNA gene sequences were identified in a subset of 20 million reads for each metagenome using ublast [41] and SSU-ALIGN [42] with the RDP release 11 database [43] clustered at 90% identity. The SILVA SSU database RefNR99 138 was used for taxonomic assignment [44].

Genome annotation

Proteins were predicted with Prodigal v2.6.3, while rRNA and tRNA coding sequences were predicted using rna_hmm3 [45] and tRNAscan-SE [46] respectively. Protein annotations were performed with an in-house pipeline that uses hmmsearch [47] against collections of COG [48], TIGRFAM [49], Pfam [50] hidden Markov models (HMMs) and KOALA algorithm against a non-redundant KEGG GENES database [51]. For a protein to be considered a hit a minimum coverage of 50% for its whole length as well as the HMM model was mandatory, using an e-value threshold of 1e-3. Protein domains were annotated using InterProScan [52] with default parameters. Metabolic pathways were inferred from KEGG [53] and were manually examined for completeness in all MAGs (282 non-dereplicated, high quality bins) (Supplementary Table S5). Carbohydrate-active enzymes (CAZy) were searched in the MAGs using hmmscan [47] and the dbCAN CAZyme domain HMM database v10 (release date 17. 08. 2021) [54].

Fragment recruitment, growth rate and doubling time estimation

All the 119 metagenomes from the 17 lakes were used for fragment recruitment (Supplementary Table 6). CPR MAGs were screened for rRNA gene sequences with barrnap (http://www.vicbioinformatics.com/software.barrnap.shtml) and these sequences were masked to avoid biases. From each metagenome, 20 million quality filtered reads were mapped against our MAGs using RazerS 3 (–no-gaps, –max-hits 1000000) [55]. The obtained number of hits was used to compute coverage per Gbp values, offering normalized abundances comparable for different MAGs and metagenomes (Supplementary Table S6). The rate of bacterial replication was estimated using the GRiD multiplex module with default options [56]. As recommended by the developers, only the forward reads from each metagenome were used for mapping and the GRiD refined values were subjected to further analysis. Doubling time for each genome was estimated with the default parameters (excluding temperature option fit) using gRodon [57].

Bacterial lifestyle assessment

RefSeq release 81 reference genomes [58] were annotated and manually inspected for the presence of possible symbionts, parasites or commensals (for simplicity only the term symbionts will be further used). The description in the literature for each bacterium was the main criteria for including a genome in the symbionts category. Other features such as genome size, low GC%, or reduced coding density were also considered [26]. Based on this criteria, 254 genomes were classified as symbionts. A similar manual approach was performed to obtain a free-living freshwater bacteria database, checking the literature, and downloading genomes from NCBI. The final collection contained 359 free-living freshwater bacteria. Basic genome statistics, as well as functional annotations were used to compare CPRs with the free-living freshwater and symbiotic bacteria (Supplementary Tables S7-S10).

Phylogenomic analysis

The genomes of all representative CPR species were retrieved from GTDB r89 (1031 genomes in total). The set of 43 SCGs (Supplementary Table S2) [38] was used to build a phylogenetic tree that included 1203 CPR bins and 13 other bacterial genomes as outgroup (affiliated to Deferribacteres, Fusobacteria, Spirochaetota, Aquificae and Epsilonproteobacteria). Individual markers were aligned with PRANK (-protein + F) (Loytynoja 2014), trimmed with BMGE (-m BLOSUM62 -t AA -g 0.5 -b 5) [59] and concatenated. A maximum-likelihood tree was generated with IQ-TREE (-bb 1000, -alrt 1000) [60] with ultrafast bootstrapping and the LG + R10 evolutionary model suggested as the best model for the dataset by ModelFinder [61]. A minimum of 21 markers were required for a genome to be retained in the tree. Markers present in less than half of the bins were removed from the final multiple alignment. Based on these criteria, the phylogenetic tree was generated using 38 markers and 1196 genomes.

Phylogenetic tree reconstruction for terminal oxidases

Gene prediction was done for all GTDB representative genomes using Prodigal v2.6.3, and the three/four subunits (COG0843, COG1622, COG1845, COG3125) of heme/copper-type cytochrome/quinol oxidase (HCO) were extracted with hmmsearch [47], followed by the generation of individual blast databases for these proteins. A local blast search was run for identifying the closest 1000 sequences of each HCO subunit from our freshwater bins. For every subunit, the BLAST results were dereplicated using MMseqs2 (easy-cluster workflow) with a minimum sequence identity of 90%. As subunit I is conserved in all HCO [62], we added sequences belonging to all phyla to this tree. To reduce the number of proteins in the alignment and tree, HCO subunit I sequences from GTDB were clustered at 70% identity (15682 initial sequences reduced to 1051). Representative sequences of this clustering step, together with those obtained through BLAST were aligned with MAFFT v7.055b [63], and the multiple alignment was trimmed with BMGE (-m BLOSUM62 -t AA -g 0.5). A Maximum-likelihood (ML) tree was generated with IQ-TREE (-bb 1000, -alrt 1000) using mtZOA + F + R10 as the best-fit evolutionary model [60, 64, 65]. As the other two/three HCO subunits are not conserved, the ML trees were generated using only the clustered top 1000 blast hits (blast e-value 1e-10; best evolutionary models: COG1622 – LG + R6, COG1845 – LG + F + R5, COG3125 – mtInv + F + R6). The same approach as for HCO subunits II-IV was used for cytochrome bd-type oxidase subunits I (COG1271) and II (COG1294). Top 1000 blast hits were clustered by MMseq2 at 90% sequence similarity and were used to generate maximum likelihood trees (blast e-value 1e-10; best evolutionary model for both subunits was LG + F + R9).

Phylogenetic analysis of rhodopsins

All GTDB representative CPR genomes, together with our 282 freshwater bins and recently published CPR MAGs [66] were scanned for the presence of rhodopsins. First, proteins were predicted using Prodigal v2.6.3 and then we used hmmsearch to find significant hits to rhodopsin HMMs, specifically Pfam HMMs corresponding to bac_rhodopsin and heliorhodopsin. All hits with more than 150 amino acids and P-values < 1e-2 were selected and blasted using MMseqs against a rhodopsin database that included all known rhodopsin sequences in UniProt [67] and GTDB. The top 50 hits for each query sequence were aligned with MAFFT (–localpair –maxiterate 1000), and the multiple sequence alignment was analyzed

using Polyphobius for transmembrane helix predictions [68]. Each candidate protein was checked for the number of helices (7 for canonical rhodopsins), orientation of the protein (Type I rhodopsins vs. heliorhodopsins), and the motifs for retinal binding in transmembrane helix 7 (DxxxK for Type I rhodopsins vs. SxxxK for heliorhodopsin). A number of 511 putative rhodopsin sequences were aligned with PASTA [69] and the phylogenetic tree was generated using IQ-TREE 2 with 1000 ultrafast bootstrap replicates and the LG + F + G4 evolutionary model [70].

Probe design and fluorescence in situ hybridization followed by catalyzed reporter deposition (CARD-FISH)

Oligonucleotide probes targeting the 16S rRNA gene of different CPR lineages were designed as previously described [71]. Briefly, all CPR MAGs were screened for 16S rRNA gene sequences with barrnap (http://www.vicbioinformatics.com/software.barrnap.shtml), were aligned with the SINA web aligner [72], imported into ARB [73] and added to the reference tree of the SILVA SSU database RefNR99 138 [44] with Maximum Parsimony. Randomized Axelerated Maximum Likelihood subtrees (RAxML, 100 bootstraps, GTR-GAMMA model [74]) for different CPR lineages including 240 16S rRNA gene sequences from MAGs and closely related reference sequences from the database were computed after manual improvements of alignments. A final RAxML tree (100 bootstraps, GTR-GAMMA model) was constructed for the most promising candidates for probe design selected by high values in fragment recruitment in metagenomes and the availability of corresponding samples for FISH (Supplementary Figure S2). Probes were designed with the tools Probe_Design and Probe_Match in ARB and all candidate probes were checked in silico for specificity and coverage in ARB and online using the TestProbe function of SILVA [44]. Competitor probes were designed for two probes that target out-group sequences by allowing 1–2 mismatches [75]. Formamide concentrations for CARD-FISH as well as mismatch and competitor analyses were estimated online with the tool MathFish [76].

For preparation of CARD-FISH samples, lake water was immediately fixed with formaldehyde (2% final concentration) at room temperature for 2 h or at 4˚C overnight. Between 5–10 ml of water (5 ml for eutrophic, 9–10 ml for oligotrophic lakes) were slowly filtered (100 mbar maximum pressure) onto 47 mm diameter 0.2 µm pore-sized polycarbonate filters (Millipore, Merck, Darmstadt, DE) and washed twice by filtering 5 ml of Milli-Q® water. Filters were air dried and stored at 20°C before being embedded in 0.1% agarose. Enzymatic pretreatment was done using lysozyme (10 mg/ml of lysozyme, 50 mM EDTA, and 0.1 M Tris-HCl, 60 min, 37°C) and achromopeptidase (60 U, 1 mM NaCl, 1 mM Tris-HCl, 30 min, 37°C), followed by an inactivation step using 0.01 M HCl for 10 min at room temperature. CARD-FISH was performed as described previously using fluorescein-labeled tyramines [77]. The newly designed probes (Supplementary Table S11, Supplementary Figure S2) were tested with a gradient of formamide concentrations guided by MathFish predictions to achieve optimal hybridization conditions. Negative controls for unspecific binding of fluorescein and cellular peroxidases were done using the nonspecific probe NON338 [78] and the CARD reaction only, i.e., FISH was done without adding a probe to the hybridization buffer. A double hybridization with a general bacterial probe was carried out for two CPR lineages (SacA-77 and Pgri-121) that are also targeted by probe EUB I-III [79] while all other CPR lineages

have > 1 mismatch with this probe. Double hybridization was done as previously described [80, 81] by using tyramides labeled with Alexa546 (probe EUBI-III) and fluorescein (probes SacA-77 and Pgri-121). All filters were counterstained with 4',6-diamidino-2-phenylindole (DAPI) and analyzed by epifluorescence microscopy (Zeiss Imager.Z2, Carl Zeiss, Oberkochen, DE) with a colibri LED light system and filter sets for DAPI (LED module 385 nm; filter set 49; Excitation 365; beam splitter [farb teiler, FT] 395; Emission BP [Em BP] 445/50), fluorescein (LED module 475 nm; filter set 38 HE; Excitation band pass [Ex. BP] 470/40; FT 495; Em BP 525/50) and autofluorescence (LED module 567 nm; filter set 62 HE; Ex BP 370/40, 474/28, 585/35; FT 395 + 495 + 610; Em TBP 425 + 527 + long pass LP 615). In case of double hybridization, we additionally used the filter set for DsRed (LED module 567 nm; filter set 43; Excitation 545; Emission 572; FT 570). Multi-channel z-stack micrographs (9–11 z-stacks with 100 nm offset for each channel) of CARD-FISH-stained cells were recorded with an Axiocam 506 (Carl Zeiss, Oberkochen, DE) and merged to one image by orthogonal projection using the software ZEN 2.6 (Carl Zeiss, Oberkochen, DE). Cell dimensions (length and width) of individual CARD-FISH stained cells were measured on the DAPI channel with the software NIS – Elements AR 4.6 using the Annotation and Measurements tool.

Results

Diversity of CPRs in freshwater lakes and their general genome characteristics

A total of 282 CPR MAGs (> 40% completeness, < 5% contamination) from 8 classes were assembled from 119 freshwater metagenomes collected from 17 lakes. Their estimated genome sizes ranged from ~ 0.5 to 2.5 Mbp (median: 1.02 Mbp, Fig. 1; assembly length ~ 0.2–1.5 Mbp, median ~ 0.63 Mbp, Supplementary Table S3). When compared to organisms with known lifestyles from RefSeq r81 (Fig. 1, Supplementary Table S7), CPRs have genome sizes and numbers of genes comparable to those of obligate intra- or extracellular symbionts/parasites. Nevertheless, their coding density (median: 89.47%; range: 76–95%) and GC content (median: 42.04%; range: 24–63%) in general resemble free-living organisms or facultative intra- or extracellular symbionts/parasites.

A phylogeny was generated using 38 single copy genes (SCGs) [11] for 1012 representative CPR genomes retrieved from GTDB r89 together with 171 dereplicated freshwater genomes obtained in this study (Fig. 2, Supplementary Table S4). In the phylogenetic tree, no clear grouping was observed based on either the genome size, isolation source or the trophic state of the lake. Metagenomic fragment recruitment was used to estimate the abundance of different CPR MAGs in 119 freshwater metagenomes. Coverage per Gbp values for each bin in their own metagenome varied between 0.02 and 14.36. The highest coverage was recovered for MAGs obtained in the hypolimnion of Lake Ikeda (~ 2–8.1) and the epilimnion of Lake Zurich (6.7–14.3) (Supplementary Table S6). Based on 16S rRNA gene data, Lake Ikeda and Jiricka pond had the highest percentage of CPRs (maximum values for these lakes were 10.89 and 21.05%, respectively), especially of sequences affiliated to the classes Paceibacteria,

ABY1 and Microgenomatia (Supplementary Figure S3). Gracilibacteria and Saccharimonadia appear to be only minor components of the lake communities (~ 1% abundance) according to the 16S rRNA gene data. Nevertheless, these quantitative results might underestimate the true abundance of CPRs in our samples as some free or unattached small cells could have passed through the 0.22 µm membrane filters.

Fragment recruitment analyses suggested MAGs were generally specific to the lake of origin, although a few exceptions to this rule were observed. Almost identical genomes from the hypolimnion of Lake Thun and both epi- and hypolimnion of Lake Traunsee were recovered in the hypolimnion of Lake Maggiore (ANI value of 98.56–98.83%), lakes located at relatively short distances (~ 400 km between Lake Traunsee and Maggiore and ~ 100 km between Lake Thun and Maggiore), while MAGs assembled from Lake Baikal were also found in Lake Biwa (ANI value of 99.53%, ~ 3000 km distance between lakes) (Supplementary Table S6, S12).

Genome replication rates in freshwater CPRs, based on ori/ter values provided by GRiD, varied between 1–1.35 (Supplementary Table S13, Fig. 3A), indicating slow growth or even stagnation at the time of sampling. Doubling time estimates for CPRs, symbionts and free-living bacteria followed a binomial distribution (Fig. 3B), with the main peak in case of CPRs and free-living microbes around 4h, whereas symbionts were predicted to replicate slower (median ~ 7.5h).

Assessment of different lifestyles in the CPR group

Natural abundances and visualization of most CPR lineages has remained elusive till now. To amend this, we designed eight FISH probes targeting different CPR lineages from four classes: ABY1 (2 probes), Paceibacteria (3 probes), Gracilibacteria (2 probes) and Saccharimonadia (1 probe) (Fig. 2, Supplementary Table S11, Supplementary Figure S2). This approach enabled the visualization of distinct CPR groups and brought into light new evidence about their potential life strategies (Fig. 4; Supplementary Figures S4-S11). A double hybridization with probe EUB I-III [79, 80] was possible only for 2 probes (SacA-77 and Pgri-121) as all other CPR lineages have > 1 mismatch to this general bacterial probe. Both double hybridizations resulted in a staining with both fluorochromes (Supplementary Figures S12-13), proving that the targeted CPR lineages are indeed bacteria. Negative controls using a nonspecific probe (NON338 [78]) and the CARD reaction without probe resulted in low, unspecific background signals, but no obvious staining of cells (Supplementary Figures S14-15). The very low abundances of individual CPR lineages targeted by our probes (Supplementary Figure S3) did not allow a precise quantification, however, multiple images were recorded, and cells could be sized (between 20 to 71 cells per probe, Supplementary Table S11). CPRs were generally small in size (0.36–0.70 µm length, 0.30–0.59 µm width, Supplementary Table S11, Supplementary Figure S16), but in the same range as genome-streamlined free-living freshwater microbes like '*Ca*. Nanopelagicales' (0.33–0.50 µm length, 0.24–0.30 µm width [82]) or '*Ca*. Fonsibacter' (0.38 µm length, 0.27 µm width [83]). However, the observed cell sizes could be partially a result of filter size (0.2 µm) used for this approach, as smaller cells might have passed through.

ABY1: While being similar in terms of metabolic potential (data not shown), the two targeted ABY1 families of the same order (SG8-24) showed slightly different lifestyle preferences (Fig. 4a, b; Supplementary Figures S4, S5). Members of candidate family UBA9934, targeted by probe ABY1b-1343 (Fig. 2, Supplementary Figure S2, S5) were found exclusively unattached to other cells, while members of family GWF2-40-263 (probe ABY1a-193, (Supplementary Figure S4) targeting several MAGs from this study and one MAG by Anantharaman et al. [11] were either free-living or attached to so-called 'lake snow', aggregates of living or decomposing microorganisms kept together by extracellular polymeric substances, and an important source of organic matter [84, 85]. Cell sizes for the 2 families were very similar, averaging 0.52 ± 0.12 µm in length and 0.46 ± 0.12 µm in width for GWF2-40-263 and 0.52 ± 0.14 µm by 0.45 ± 0.14 µm in case of UBA9934 (Supplementary Table S11, Supplementary Figure S16).

Paceibacteria: Representatives of the candidate family UBA11359 (probe ZE-1429) were identified as very small cocci (average size 0.35 ± 0.08 µm by 0.30 ± 0.06 µm), consistently associated with other larger prokaryotes (Fig. 4l, m, Supplementary Figure S11). Members of the genus GWA1-54-10 (order UBA9983_A; family UBA2163; aka. '*Ca*. Alderbacteria') were visualized with two CARD-FISH probes (Adl1-132 and Adl2-134, Supplementary Table S11) and they were shown to be also small (average cell sizes 0.67 ± 0.2 µm by 0.59 ± 0.18 µm for Adl1-132 and 0.37 ± 0.1 µm by 0.31 ± 0.09 µm for Adl2-134). All microbes targeted by probe Adl1-132 appear to be free-living (Fig. 4c, d, Supplementary Figure S6). Only one free-living representative was observed with probe Adl2-134 (Supplementary Figure S7A, B), while all other cells targeted by this probe were associated with hosts with up to 10 small GWA1-54-10 cells surrounding a large prokaryotic cell (Fig. 4e-g).

Gracilibacteria: Gracilibacteria family LOWO2-01-FULL-3 (order UBA1369), targeted by probe Pgri-124 (average cell sizes 0.48 ± 0.13 µm by 0.43 ± 0.12 µm), was associated with various small prokaryotes and picocyanobacteria (Fig. 4j, Supplementary Figure S9) and therefore was not limited to a definite host. Another Gracilibacteria family (2-02-FULL-48-14) of the same order targeted by probe Pgri-99 (Supplementary Figure S8) had cell sizes of 0.49 ± 0.12 µm by 0.45 ± 0.12 µm. They were likewise associated with small prokaryotes and cyanobacteria (Fig. 4h) but were also found in close vicinity to 'lake snow' particles (Fig. 4i).

Saccharimonadia: Members of the family UBA10212, targeted by probe SacA-77, were observed as diplococci (Fig. 4k, Supplementary Figure S10). Although their genomes are highly reduced and they lack important metabolic pathways for survival (Supplementary Table S5), they were not always associated with other cells (Supplementary Figure S4). They were observed either as elongated (on average 0.70 ± 0.07 µm by 0.41 ± 0.05 µm) or approximately round (average sizes 0.49 ± 0.08 µm by 0.43 ± 0.08 µm) cells.

Metabolic capabilities in freshwater CPR groups

In terms of average gene composition among metabolic pathways, Gracilibacteria form a cluster with free-living bacteria, while all other CPR groups possess a much more depleted metabolic repertoire and are similar to known symbionts (Supplementary Figure S17). Gracilibacteria that were visualized by

CARD-FISH (order UBA1369, formerly known as Perigrinibacteria) encode the core 3-carbon compound module of glycolysis, the genes for nucleotide sugar biosynthesis, pentose phosphate pathway (PPP), parts of the Calvin cycle and the biosynthesis of phosphoribosyl diphosphate (PRPP) that is required to produce both purines and pyrimidines (Fig. 5). The capacity for beta-oxidation of fatty-acids and the pyruvate carboxylase are missing in all Gracilibacteria genomes along with the genes required for the synthesis of cofactors (biotin and thiamine) involved in these functions, biotin and thiamine. The genes for NAD^+ and THF biosynthesis are present, two cofactors involved in nucleotide synthesis. The pathways necessary to produce riboflavin, FMN and FAD were present in both Gracilibacteria orders. The same was true for pantothenate, although the enzymes required for its conversion to acetyl-CoA were not encoded in Gracilibacteria but were present in some Microgenomatia and Paceibacteria MAGs. In Gracilibacteria, acetyl-CoA can be obtained from pyruvate through the activity of pyruvate:ferredoxin oxidoreductase. Even though NADH dehydrogenase and the F-type ATPase are encoded by Gracilibacteria, together with a putative proton-pumping rhodopsin (Supplementary Table S15), terminal oxidases were not identified.

MAGs belonging to ABY1, Gracilibacteria, Microgenomatia, Paceibacteria and Saccharimonadia usually encode the part of the reductive pentose phosphate cycle (Calvin cycle) responsible for converting glyceraldehyde 3-phosphate (G3P) to ribulose-5P (Supplementary Table S5, Supplementary Figure S18) which is necessary for nucleotide biosynthesis. As appears common for CPRs [3, 10, 86], the Embden-Meyerhof glycolysis pathway is also rarely complete in our freshwater MAGs with only a few exceptions in Paceibacteria. In general, 6-phosphofructokinase and glucokinase/hexokinase are missing, with only the core module involving 3-carbon compounds being completely encoded in all groups. Glycolysis can still be achieved through a metabolic loop involving the pentose phosphate pathway [38] that is encoded in all classes (Supplementary Table S5). All genes encoding the enzymes involved in gluconeogenesis were present in freshwater Paceibacteria, but not in the same MAGs and therefore it is difficult to conclude if the pathway is complete. On the other hand, phosphoenolpyruvate carboxykinase, and sometimes fructose-1,6-bisphosphatase, showed patchy distributions or were missing completely in the other classes making gluconeogenesis impossible. As previously reported, enzymes involved in the Krebs cycle are not encoded in freshwater CPRs, indicating a fermentative lifestyle [25, 87].

As previously described [10, 12], the capacity for fermentation is widespread in CPRs, highlighted by the prevalence of lactate dehydrogenase in ~ 50% of the freshwater MAGs assembled in this study, with the vast majority of ABY1 and Gracilibacteria possessing this gene (Fig. 6). Less common (< 30% of genomes) are the Zn-dependent and short chain alcohol dehydrogenases (ADH) that catalyze the interconversion between acetaldehyde and ethanol. Pyruvate decarboxylase, the enzyme converting pyruvate directly into acetaldehyde was not identified in our freshwater MAGs, therefore the alcoholic fermentations could potentially occur with an additional step. Firstly, pyruvate would be converted into acetyl-CoA by pyruvate:ferredoxin oxidoreductase, then it would be further processed by acetaldehyde dehydrogenase (ALDH). The last step involves the conversion of acetaldehyde into ethanol by alcohol dehydrogenase (ADH). Both lactic and alcoholic fermentations are coupled with the oxidation of

NAD(P)H to NAD(P)+, providing a steady supply of NAD + for powering glycolysis and the slow generation of ATP in the absence of an ETC. Although acetate kinase is common in Gracilibacteria, and to some extent in Paceibacteria and Saccharimonadia, how the acetyl-phosphate is obtained from acetyl-CoA in the last two classes is not clear, as the gene for phosphate acetyltransferase (pta) was observed only in Gracilibacteria. Nevertheless, if acetyl-P becomes available, these groups seem to be able to generate acetate by coupling this reaction with substrate level phosphorylation (Fig. 6).

All freshwater CPRs were found to encode carbohydrate-active enzymes (CAZy) with an average of 15 genes per genome (Supplementary Table S16), therefore in lower amounts than previously reported in a thermokarst lake [21]. By far the most encountered enzymes belonged to the GT4 (involved in rhamnose degradation, found in > 95% of genomes) and GT2 (polysaccharide conversion, present in > 92.5% of genomes) families, representing ~ 35% and ~ 23% from the total number of identified CAZy. Enzymes for the degradation of chitin, cellulose and mannose were also identified, but in low abundances, ranging from 1.1–2.8% of total CAZy (Supplementary Table S16).

Though carbohydrate anabolism is limited, Microgenomatia and ABY1 (Supplementary Figure S18), and as mentioned before Gracilibacteria, can perform nucleotide sugar biosynthesis. The production of dTDP-L-rhamnose, a cell envelope component, is conserved in the classes ABY1, Kazan-3B-28, Microgenomatia, Gracilibacteria, Paceibacteria and Saccharimonadia, while the biosynthesis of ADP-L-glycero-D-manno-heptose is encoded only in Paceibacteria (Fig. 5, Supplementary Figure S18, Supplementary Table S5). In case of our Gracilibacteria freshwater MAGs, C5 isoprenoid biosynthesis is usually performed through a mevalonate pathway typical for eukaryotes/bacteria [3, 88], with only one exception, the UBA1369 order, that uses the common methylerythritol phosphate pathway for bacteria. Other classes seem able to perform only C10-C20 isoprenoid biosynthesis through a typical bacterial pathway (Fig. 5, Supplementary Table S5, Supplementary Figure S18). No ability whatsoever for the synthesis or beta-oxidation of fatty acids was detected in any group, thus the way in which CPRs produce their cellular membranes remains enigmatic.

It was previously hypothesized that CPR might enable phage infection as a somewhat risky source of nucleotides [3]. By analyzing over 1300 high quality CPR genomes, CRISPR-Cas systems were detected in 1.68–6.36% (mean = 3.89%) of the MAGs in classes with > 50 genome representatives (for a detailed report about phage defense mechanisms see Additional File 1 and Supplementary Table S14), which puts them in a low range for bacteria [89]. An alternative source for nucleotides is the uptake of free DNA from the environment. We identified competence-related DNA transformation transporters (ComEA/ComEC) in most of our MAGs, providing a stable mechanism for DNA uptake [7, 90]. Moreover, inosine monophosphate (IMP) can be synthesized from phosphoribosyl diphosphate by Gracilibacteria, Microgenomatia and Paceibacteria. Further processing of IMP to ATP occurs in ABY1, Gracilibacteria, Microgenomatia and Paceibacteria, but the synthesis of GTP is not completely encoded in Microgenomatia which lack guanylate kinase. Microgenomatia MAGs seem to possess all genes for uridine monophosphate (UMP) biosynthesis, but not the other CPR classes. In contrast, the ability to turn UMP into CTP is widespread in most CPR groups, as well as its conversion into TTP in ABY1,

Gracilibacteria and Paceibacteria. Metabolic pathways for the synthesis of amino acids are usually absent, with some exceptions. For example, some MAGs in Microgenomatia and Paceibacteria encode the genes necessary to produce serine, while threonine can be apparently synthesized by ABY1 (Supplementary Figure S18), and possibly by Gracilibacteria and Paceibacteria. Genes for biosynthesis of lysine, proline and tryptophan are present in Paceibacteria and Gracilibacteria, the latter also encoding pathways for other aromatic and branched-chain amino acids (Fig. 5). Histidine degradation to glutamate is encoded only in Paceibacteria. Although freshwater CPRs of the ABY1 class encode restricted biosynthesis pathways, its members seem to possess numerous importers for tyrosine, branched-chain amino acids, multiple sugars, 3-phenyl propionate, polysaccharides, as well as transporters for ions, such as Fe^{2+}, Mn^{2+}, Mg^{2+}, Zn^{2+}, and nitrogen oxides. They are also able to export heavy metals, polysaccharides, and numerous antibiotics (Fig. 4, Supplementary Figure S18). Even though type II secretion systems (T2SS) were reported in CPRs [7, 90], we were able to detect only the presence of Type II/IV secretion ATPase GspE together with the inner membrane platform protein GspL [91]. The remaining T2SS components were not detected, but general secretion (Sec) and twin-arginine-translocation Tat systems were found in our freshwater CPRs.

Subunits of electron transport chain in freshwater CPRs

Regarding the genes involved in generating an electron transport chain (ETC), subunits of the NADH dehydrogenase (complex 1) are present in > 30% of the freshwater MAGs, being common in ABY1, Gracilibacteria, Paceibacteria and Saccharimonadia. Also, three MAGs affiliated to Saccharimonadia (sampled from the oxygenated hypolimnions of lake Most and Římov, 50m and 30m depth) and four belonging to Paceibacteria (recovered from the oxygenated hypolimnion of lake Thun and Ikeda, 180m and 100m depth) encode all subunits of cytochrome o oxidase (HCO), indicating a putative capacity for oxygenic respiration (complex IV). Additionally, five Paceibacteria MAGs recovered from the oxygenated hypolimnion of lake Thun and Maggiore (180m and 300m depth, respectively) seem to have a functional cytochrome bd-type quinol oxidase.

The phylogeny of HCO subunit I, which included 1439 representative sequences, showed that CPRs probably obtained this gene horizontally from Proteobacteria (Supplementary Figure S19), as previously proposed [92]. The closest group to CPRs in this tree belongs to Gammaproteobacteria, more specifically the orders Legionellales and Thiotrichales. Most of these organisms are facultative or obligate intracellular parasites, which might imply that the association with a host facilitated the HGT. Unexpectedly, the order Parachlamydiales of Verrucomicrobia, comprising mainly endosymbionts of free-living amoebae [93], appears to have obtained this subunit from Saccharimonadia. In our freshwater MAGs, HCO subunits are adjacent, forming an operon that was likely acquired horizontally at one time point. The phylogeny of the other HCO subunits follows the same evolutionary pattern as subunit I, with CPR sequences diverging from Proteobacteria, and Parachlamydiales sequences radiating from within Saccharimonadia (Supplementary Figure S19). The ML trees generated for cytochrome bd-type oxidases (Supplementary Figure S20) show that both subunits follow the same evolutionary pattern, in which the

genes in CPRs appear to be transferred from a proteobacterial source, forming a cluster together with cyanobacterial sequences.

Rhodopsins occurrence in CPRs

A total of 1326 CPR genomes (1032 GTDB representative genomes, 282 freshwater CPR assembled in this study, 12 MAGs analyzed by Jaffe et al. [94]) were screened for the presence of rhodopsins. We were able to detect 115 rhodopsin sequences in 86 genomes, out of which 17 were predicted to be proton-pumping rhodopsins while the rest had a reverse orientation (N-terminal in the inside, C-terminal in the outside of the membrane [95]) and were therefore inferred to be heliorhodopsins (HeRs). Both proton-pumping rhodopsins and HeRs were identified predominantly in Saccharimonadia. HeRs had 80 occurrences in Saccharimonadia, 7 in Dojkabacteria, 3 in ABY1 and less in other classes, while we identified 15 proton-pumping rhodopsin sequences in Saccharimonadia, 1 in Gracilibacteria and another 1 in Paceibacteria (Supplementary Table S15). While checking for the conserved lysine in the transmembrane helix 7 that is required for retinal binding in HeRs, we observed that the most common motifs were SLVAK, SLIAK and SFVAK, the interchangeable amino acids belonging to the same group of compounds with hydrophobic side chains (Supplementary Table S15).

Discussion

Genome reduction in CPRs

One of the features of CPRs that caught the attention of the scientific community was their reduced genome sizes, cell sizes and metabolic capacities [10]. Though the mechanisms of genome reduction are different for free-living vs. symbiotic bacteria, genome reduction comes with a dependency upon other organisms: either a host, a co-symbiont, or in the establishment of a consortium with other microbes [26]. In free-living organisms, genome reduction could be an advantage especially in oligotrophic environments such as the pelagial of lakes, as it lowers the energy requirement for survival and reproduction [26]. In symbionts, gene loss occurs as a consequence of a protected and stable environment, rich in nutrients required for growth [96]. Our freshwater CPR genome sizes resembled the values reported previously for free-living streamlined bacteria [97], symbionts [98], and other CPRs [10, 38]. The higher coding density observed in CPRs (~ 76–95%, median 89%) compared to symbionts (54–95%, median ~ 76%) (Fig. 1, Supplementary Table S8, S10) indicates a reduced number of pseudogenes or recent gene loss, one of the traits for early stages of symbiosis and parasitism [98]. However, the observed coding densities in CPRs (Fig. 1) are still below the observed coding-densities in free-living streamlined microbes where values of > 95% are common - e.g., in freshwater ‘Ca. Nanopelagicales’ [82], or ‘Ca. Methylopumilus’ [99], suggesting that not all Patescibacteria are in the final stages of genome streamlining [3]. In some free-living and host associated bacteria, the reduction in GC% content was associated with a loss of 6-O-methylguanine-DNA methyltransferase (adaB gene), among other DNA repair genes [26]. The adaB gene is still common in our Paceibacteria and Gracilibacteria bins, but not in the other groups (data not shown). This is probably not sufficient to explain the variability in genome size

and GC% values, but their drop might offer a selective advantage in phosphorus and nitrogen depleted environments such as oligotrophic and ultra-oligotrophic lakes [97, 100]. Thus, general genomic features suggest that CPRs combine characteristics of both symbiotic and free-living bacteria.

CPRs occurrence and diversity in freshwater lakes

In the phylogenetic tree, some loose clusters were formed for MAGs belonging to Paceibacteria, ABY1 and Saccharimonadia that seem to prefer more eutrophic lakes. However, they were closely related to genomes isolated from groundwater or ultra-oligotrophic lakes (Fig. 2), suggesting that they might be more dependent on their host or co-occurring microbes than the environmental conditions [26]. Most of these MAGs belonged to the Paceibacteria class, as was also the case for Lake Baikal [20], Lake Alinen Mustajärvi [101] and a permafrost thaw lake [21], followed by Saccharimonadia and ABY1 (Supplementary Table S3). Paceibacteria, ABY1 and Microgenomatia were found in relatively high abundance in their lake of origin according to the 16S rRNA gene abundance data, while Gracilibacteria and Saccharimonadia were observed in lower proportions (~ 1% abundance) (Supplementary Figure S3). Even at lower abundances, CPR members might still be meaningful contributors to the ecosystem as the importance of the rare microbial biosphere in nutrient cycling and organic matter breakdown is starting to be recognized [21, 102]. Genome replication results derived from GRiD (1–1.35, Fig. 3A) were similar to the values obtained for Patescibacteria in groundwaters [56], which could imply that they replicate only in certain conditions, for example when they manage to get attached to a host cell [103]. Doubling time estimation by gRodon is just tentative (Fig. 3B), as it performs poorly for slow growing organisms or for those with atypical effective population sizes, such as parasites [57]. Although CPRs have a clear peak at ~ 4h, implying that at least a proportion of them are able to duplicate relatively fast in optimal conditions, the median is approximately at 6h, which simply means that the doubling time cannot be predicted accurately and that the organisms are probably slow growers [57].

CPR observations in natural environments and their interaction with hosts

Though a symbiotic lifestyle was previously assumed for this bacterial radiation, several studies hypothesized about the capacity for independent survival at least in some groups [3, 10, 13, 25]. Our CARD-FISH results indicate that a wide variety of lifestyles could exist (free-living, attached to lake snow, host-associated) among different CPR lineages (Fig. 4, Supplementary Figures S4-S11). For example, in a single genus affiliated to Paceibacteria, we observed both free-living and host-associated cells (FISH probes Adl1-132 and Adl2-134, Supplementary Figure S6, S7A, S7B). Another peculiar observation was a group of free-living Saccharimonadia (TM7) because previous work showed them invariably in association with other bacteria [5, 6]. It is possible though that the free-living cells are facultative epibionts in search of a suitable host or they were loosely attached to host cells and got separated during sample preparation.

The relationship between parasites/symbionts and their host or the environment is carefully modulated through protein secretion. In case an association is formed between two species, secretion systems must include mechanisms to translocate secreted proteins (effectors and toxins) across the plasma membrane

of the host [104]. The inner membrane platform proteins of the T2SS (GspM, GspF, GspC) were not identified, which implies that they are either too divergent or absent in Patescibacteria (Supplementary Table S5). Typically, the pseudopilus in Gram-negative bacteria is composed of five pseudopilins (GspG-K), out of which CPRs seem to encode only GspG and two of the minor pseudopilins (GspH and GspI). Furthermore, none of the outer-membrane complex proteins were identified, such as GspD (forming the secretin pore) and GspS, which questions the functional ability of T2SS. Patescibacteria were predicted to be Gram-positive [105], therefore it is quite peculiar for them to encode parts of the inner membrane and periplasmic T2SS. General secretion (Sec) and twin-arginine-translocation Tat systems are universal in bacteria, both being used to export proteins in an unfolded state [104]. All CPR classes seem to have members that encode the complete Sec membrane complex (SecD-G, SecY) as well as the cytoplasmic SecA, which hydrolyses ATP to drive proteins translocation (Supplementary Table S5) and PrsA, a membrane-associated lipoprotein that was proven to be essential for protein secretion in *Bacillus subtilis* [106, 107]. Additionally, YidC protein is also common, mediating the membrane insertion of Sec proteins [108]. Tat translocation was identified in Microgenomatia, Paceibacteria and Saccharimonadia, being composed of TatA (the major pore-forming subunit) and TatC (the subunit involved in recognizing the targeted proteins for secretion) proteins. Bacterial secretion systems known to directly interact with the host cell (Type III, IV, VI and VII secretion systems) were not detected, even though at least some CPRs are clearly associated with Actinobacteria [4, 5, 109] and Proteobacteria [7, 90]. Also, as it was apparent from our work, members of two Gracilibacteria families associated with cyanobacteria, although they did not seem to be restricted to them as hosts (Fig. 4, Supplementary Fig. 8, 9). Therefore, the mode of interaction between symbiotic/parasitic CPRs and their hosts is still unknown, though some electron microscopy evidence points towards interaction through pili-like structures [3, 66].

Energy and carbon sources in freshwater CPRs

The absence of respiration was reported to be a common feature of CPRs, making them reliant on fermentation for energy conservation [2]. Indeed, the complete set of five complex structures similar to the mitochondrial ETC was never present in any of our MAGs. Nevertheless, subunits of different respiratory complexes (complex I - NADH dehydrogenase, complex IV – terminal oxidases) are present in many freshwater CPRs and a complete F-type ATPase is usually the norm (Supplementary Table S5). The encoded oxidases form an operon in freshwater CPRs and were probably obtained as a result of a HGT event. These enzymes differ in their affinity for oxygen; compared to HCOs, cytochrome bd oxidases have a higher affinity and may allow cells to respire O_2 even when its concentration is very low [110]. Even though we found MAGs from the same lake and lake layers encoding different terminal oxidases, it was observed that cytochrome bd-type oxidase was more often encountered at lower O_2 concentrations (Supplementary Figure S21). Interestingly, a group of phylogenetically closely related Saccharimonadia MAGs encoded only the F1 subunit of ATPase, lacking any other genes for the synthesis of respiratory oxygen reductases, which probably restrict them to fermentation. The almost universal presence of ATPase in CPRs together with some parts of the ETC, and the absence of any genes for respiratory oxygen reductases in the absence of functional ATPase, makes sense only in case CPRs can generate a

proton motive force (PMF). It was hypothesized that the PMF might be generated using yet undescribed protein systems or unusual combinations of proteins (e.g., ferredoxins, rubredoxins) [7, 90], maybe even by some rhodopsins (see the discussion below; Supplementary Table S15). The presence of terminal oxidases in some CPRs is somewhat striking as it implies an oxygenic metabolism. Although it cannot be certain, it was proposed that the lack of other ETC components indicates an O_2 scavenging function rather than that of energy production [25].

Interestingly, at least some Gracilibacteria seem to be able to directly import ATP due to the presence of the ATP/ADP translocase, a common transporter primarily in obligate intracellular bacteria and plant plastids [111, 112]. These transporters mediate the import of host ATP across the bacterial cell membrane, which otherwise would not be permeable for such a large and charged compound [111]. This might indicate that at least in some episymbiotic/parasitic CPR groups, the host itself or the local environment could be the direct source of ATP required for survival.

The presence of different CAZy in CPRs suggest for the capacity of degradation of both simple and complex carbon substrates, contributing to the turnover of organic matter. The capacity for rhamnose degradation is by far the most common of all (Supplementary Table S16). Rhamnose is a widespread constituent of cell walls of plants and algae, and it can become an abundant sugar in lake ecosystems, especially in the profundal zone, making it a stable monosaccharide source [113]. CAZy involved in chitin, cellulose and mannose degradation were also encoded by our freshwater CPRs, indicating for a role in decomposition and fermentation of organic matter. In any case, the capacity for fermentation is widespread in all freshwater CPR groups, allowing a slow but steady supply of ATP and the regeneration of oxidized $NAD(P)^+$. Fermentation products, such as lactate, alcohol and acetate could be secreted in the environment, supporting both aerobic and anaerobic microorganisms [10], and accomplishing another key ecological role in freshwaters. Genes required for glycerolipids (GT28) and polysaccharide conversion (GT2) were also identified, representing maybe important mechanisms for obtaining these compounds in the absence of *de novo* biosynthetic pathways [21].

Putative rhodopsins roles in CPRs

Type I rhodopsin [114] can generate a PMF or ion gradient, which could be used for energy production. As some Saccharimonadia MAGs obtained from freshwater lake samples in Sweden and Finland, as well as from glacial surface ice in Greenland [94] encoded genes annotated with low confidence as bacteriorhodopsin, it was hypothesized that they might encode protein pumping rhodopsins (Supplementary Table S15). Bins that were investigated by Jaffe et al. [94] form a freshwater cluster in our rhodopsin phylogeny in between Sensory Rhodopsin I (SRI) and II (SRII) (Fig. 7), but we were unable to find any of the proteins involved in the signal transduction from SRI/II in the near genome context [115, 116] (Supplementary Figure S22). Therefore, these could indeed be ion-pumping rhodopsins, indicating a photoheterotrophic lifestyle in some Saccharimonadia groups (Supplementary Table 15).

Interestingly, the same freshwater Saccharimonadia MAGs encode also heliorhodopsins (HeRs), a recently discovered group of rhodopsins with a peculiar, reversed orientation (cytoplasmic N-terminus, extracellular C-terminus) [95]. HeRs were also found in other Saccharimonadia, as well as in ABY1, Microgenomatia, Dojkabacteria, Kazan-3B-28, UBA1384, CPR2 and CPR3 groups (Supplementary Table 15). CPR HeRs appear to form new clades in the phylogenetic tree (Fig. 7). With the use of strand-specific transcriptomics and genomic context analysis, Bulzu et al. [117] proposed that HeRs have a critical role in protecting especially monoderms against light-induced oxidative stress through the increased transcription of glyoxylases, glutaredoxins, peroxiredoxins and catalases. While glyoxylases were present in only few freshwater CPR bins, the other enzymes are encountered in almost all MAGs. Hence, in the context of genome streamlining, protection against oxidative stress is a feature highly conserved in CPRs and at least in some cases it might be linked with the HeRs function as a sensory rhodopsin.

Conclusion

This study presents new insights about CPRs diversity, distribution, and physiology in an under-explored habitat and on a large scale, namely 119 metagenomic samples of 17 freshwater lakes from central Europe and Asia. Though in low abundance and with low replication rates, various CPR MAGs were consistently found in these lakes, with no apparent preference for the trophic state of the biome. By employing CARD-FISH, we were able to visualize several CPR lineages for the first time and the evidence suggests the existence of different life strategies in this radiation. We found likely free-living or dispersing forms of several CPRs lineages, CPRs attached to ‘lake snow’ particles and host-associated CPRs. Freshwater CPRs showed reduced metabolic capacities, similar to groundwater ones, with the exception of Gracilibacteria which seem to possess more complete metabolic pathways. Even though some MAGs encode parts of the ETC, the organisms are most probably fermenters, providing lactate and acetate for other microorganisms in the ecosystem. The presence of heliorhodopsins together with oxidative stress mitigating enzymes in CPRs indicates that protection against oxidative stress is still a conserved feature in these reduced genomes. Overall, this study brought forward new information about Patescibacteria, a still enigmatic bacterial group, and helped us expand the picture about their occurrence and life strategies to freshwater lakes.

Abbreviations

ADH - alcohol dehydrogenase

ADP – Adenosine diphosphate

ALDH - acetaldehyde dehydrogenase

ANI – Average Nucleotide Identity

ATP – Adenosine triphosphate

CARD-FISH - Fluorescence in situ Hybridization followed by Catalyzed Reporter Deposition

CPR - Candidate Phyla Radiation

CTP – Cytidine triphosphate

DAPI - 4',6-diamidino-2-phenylindole

dTDP – Thymidine diphosphate

ETC – Electron transport chain

FAD - Flavin adenine dinucleotide

FISH - Fluorescence in situ Hybridization

FMN - Flavin mononucleotide

G3P - Glyceraldehyde 3-phosphate

GTDB - Genome Taxonomy Database

HCO - Heme/copper-type Cytochrome/quinol Oxidase

HeR – Heliorhodopsin

HMM - Hidden Markov Model

IMP – Inosine monophosphate

MAG - Metagenome-Assembled Genome

ML - Maximum-likelihood

NAD(P)H - Nicotinamide adenine dinucleotide phosphate

NAD^+/NADH - Nicotinamide adenine dinucleotide

PMF - Proton Motive Force

PPP - Pentose Phosphate Pathway

PRPP - Phosphoribosyl diphosphate

RAxML - Randomized Axelerated Maximum Likelihood

SCG – Single Copy Genes

SSU rRNA – Small SubUnit ribosomal Ribonucleic Acid

T3SS – Type three secretion system

T4SS - Type four secretion system

T6SS - Type six secretion system

T7SS - Type seven secretion system

THF – Thetrahydrofolate

TTP – Tymidine triphosphate

UMP -Uridine monophosphate

* Abbreviations used in figures are explained in figures legends

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Sequence data generated in this study have been deposited in the European Nucleotide Archive (ENA) at EMBL-EBI under project accession numbers PRJEB35640 and PRJEB35770. The genomic data that support the findings of this paper are available in FigShare (link: https://figshare.com/s/7f5c78f4949068e5492b). All other relevant data supporting the findings of this study are available within the paper and its supplementary information files.

Competing interests

The authors declare no competing interests.

Funding

M.-C. C. was supported by the Program for the Support of Perspective Human Resources (PPLZ), Czech Academy of Sciences (Grant No. L200961953) and by the research grant 21-21990S (Grant Agency of the Czech Republic). P.-A.B. and R.G. were supported by the research grant 20-12496X (Grant Agency of the Czech Republic). V.S.K. was supported by research grants 20-12496X (Grant Agency of the Czech

Republic), 116/2019/P (Grant Agency of the University of South Bohemia in České Budějovice, 2019-2021) and 21-21990S (Grant Agency of the Czech Republic). A.-Ş.A. was supported by Ambizione grant PZ00P3_193240 (Swiss National Science Foundation). M.M.S., P.L. and M.H. were supported by the research grant 19-23469S (Grant Agency of the Czech Republic). P.L. was additionally supported by the research grant 022/2019/P (Grant Agency of the University of South Bohemia in České Budějovice, 2019-2021) and M.H. received additional support by research grants 310030_185108 (Swiss National Science Foundation) and 21-21990S (Grant Agency of the Czech Republic). Y.O. was supported by research grant 18J00300 (JSPS KAKENHI) and by The Kyoto University Foundation Overseas' Research Fellowship. S.N. was supported by research grant 17K19289 (JSPS KAKENHI).

Author's contributions

M.-C.C., R.G. and M.M.S. designed the study. M-C.C. analyzed the data and wrote the manuscript. P.-A.B. and R.G. and helped with data analysis. M.M.S., M.H., P.-A.B., P.L., V.S.K., A.-S.A., Y.K. and S.-i.N. performed the sampling. M.M.S designed the CARD-FISH probes, M.-C.C. and P.L. performed CARD-FISH. P.-A.B., A.-S.A., Y.O., M.H., R.G., and M.M.S. helped with data interpretation. All authors commented on and approved the manuscript.

Acknowledgements

P. Znachor, P. Rychtecky and J. Nedoma are acknowledged for help in sampling of Římov Reservoir, T. Posch and E. Loher for help in sampling of Lake Zurich and Greifensee. The team of K. Řeháková is acknowledged for help in sampling of lakes Medard, Most and Milada and P. Porcal for providing samples of Jiřická pond. C. Callieri, F. Leporelli, T. Shabarova and V. Lanta are acknowledged for help in sampling of Lake Maggiore and Lake Lugano. B. Sonntag is acknowledged for providing lab space for sample processing of the Austrian lakes. S. Dirren and the crew of the research vessel 'Kormoran' are thanked for help in sampling of Lake Constance and the Laboratory for Water and Soil Protection of the Canton of Bern (GBL), Switzerland, for help in sampling Lake Thun. We thank Y. Hodoki and I. Mukherjee for assistance in Lake Ikeda sampling. Sampling in Lake Biwa was performed using the research vessel 'Hasu' operated by Captain Goda and Vice-Captain Akatsuka, Center for Ecological Research, Kyoto University.

References

1. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil PA, Hugenholtz P: **A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life**. Nat Biotechnol 2018, **36**(10):996–1004.
2. Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, Wilkins MJ, Wrighton KC, Williams KH, Banfield JF: **Unusual biology across a group comprising more than 15% of domain Bacteria**. Nature 2015, **523**(7559):208–211.

3. Castelle CJ, Banfield JF: **Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life**. Cell 2018, **172**(6):1181–1197.
4. Bor B, Collins AJ, Murugkar PP, Balasubramanian S, To TT, Hendrickson EL, Bedree JK, Bidlack FB, Johnston CD, Shi W *et al*: **Insights Obtained by Culturing Saccharibacteria With Their Bacterial Hosts**. J Dent Res 2020, **99**(6):685–694.
5. Cross KL, Campbell JH, Balachandran M, Campbell AG, Cooper SJ, Griffen A, Heaton M, Joshi S, Klingeman D, Leys E *et al*: **Targeted isolation and cultivation of uncultivated bacteria by reverse genomics**. Nat Biotechnol 2019, **37**(11):1314–1321.
6. He X, McLean JS, Edlund A, Yooseph S, Hall AP, Liu SY, Dorrestein PC, Esquenazi E, Hunter RC, Cheng G *et al*: **Cultivation of a human-associated TM7 phylotype reveals a reduced genome and epibiotic parasitic lifestyle**. Proc Natl Acad Sci U S A 2015, **112**(1):244–249.
7. Moreira D, Zivanovic Y, Lopez-Archilla AI, Iniesto M, Lopez-Garcia P: **Reductive evolution and unique predatory mode in the CPR bacterium Vampirococcus lugosii**. Nat Commun 2021, **12**(1):2454.
8. Yakimov MM, Merkel AY, Gaisin VA, Pilhofer M, Messina E, Hallsworth JE, Klyukina AA, Tikhonova EN, Gorlenko VM: **Cultivation of a vampire: 'Candidatus Absconditicoccus praedator'**. Environ Microbiol 2021.
9. Gong J, Qing Y, Guo X, Warren A: **"Candidatus Sonnebornia yantaiensis", a member of candidate division OD1, as intracellular bacteria of the ciliated protist Paramecium bursaria (Ciliophora, Oligohymenophorea)**. Syst Appl Microbiol 2014, **37**(1):35–41.
10. 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**(10):629–645.
11. Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, Probst AJ, Thomas BC, Singh A, Wilkins MJ, Karaoz U *et al*: **Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system**. Nat Commun 2016, **7**:13219.
12. 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**(1):112.
13. Probst AJ, Castelle CJ, Singh A, Brown CT, Anantharaman K, Sharon I, Hug LA, Burstein D, Emerson JB, Thomas BC *et al*: **Genomic resolution of a cold subsurface aquifer community provides metabolic insights for novel microbes adapted to high CO2 concentrations**. Environ Microbiol 2017, **19**(2):459–474.
14. Probst AJ, Ladd B, Jarett JK, Geller-McGrath DE, Sieber CMK, Emerson JB, Anantharaman K, Thomas BC, Malmstrom RR, Stieglmeier M *et al*: **Differential depth distribution of microbial function and putative symbionts through sediment-hosted aquifers in the deep terrestrial subsurface**. Nat Microbiol 2018, **3**(3):328–336.
15. Starr EP, Shi S, Blazewicz SJ, Probst AJ, Herman DJ, Firestone MK, Banfield JF: **Stable isotope informed genome-resolved metagenomics reveals that Saccharibacteria utilize microbially-processed**

plant-derived carbon. Microbiome 2018, **6**(1):122.

16. Nicolas AM, Jaffe AL, Nuccio EE, Taga ME, Firestone MK, Banfield JF: **Unexpected diversity of CPR bacteria and nanoarchaea in the rare biosphere of rhizosphere-associated grassland soil**. *BioRxiv* 2020.
17. Lannes R, Cavaud L, Lopez P, Bapteste E: **Marine Ultrasmall Prokaryotes Likely Affect the Cycling of Carbon, Methane, Nitrogen, and Sulfur**. Genome Biol Evol 2021, **13**(1).
18. Orsi WD, Richards TA, Francis WR: **Predicted microbial secretomes and their target substrates in marine sediment**. Nat Microbiol 2018, **3**(1):32–37.
19. Woodcroft BJ, Singleton CM, Boyd JA, Evans PN, Emerson JB, Zayed AAF, Hoelzle RD, Lamberton TO, McCalley CK, Hodgkins SB *et al*: **Genome-centric view of carbon processing in thawing permafrost**. Nature 2018, **560**(7716):49–54.
20. Cabello-Yeves PJ, Zemskaya TI, Zakharenko AS, Sakirko MV, Ivanov VG, Ghai R, Rodriguez-Valera F: **Microbiome of the deep Lake Baikal, a unique oxic bathypelagic habitat**. Limnology and Oceanography 2019, **65**(7):1471–1488.
21. Vigneron A, Cruaud P, Langlois V, Lovejoy C, Culley AI, Vincent WF: **Ultra-small and abundant: Candidate phyla radiation bacteria are potential catalysts of carbon transformation in a thermokarst lake ecosystem**. Limnology and Oceanography Letters 2019, **5**(2):212–220.
22. Herrmann M, Wegner CE, Taubert M, Geesink P, Lehmann K, Yan L, Lehmann R, Totsche KU, Kusel K: **Predominance of Cand. Patescibacteria in Groundwater Is Caused by Their Preferential Mobilization From Soils and Flourishing Under Oligotrophic Conditions**. Front Microbiol 2019, **10**:1407.
23. Baricz A, Chiriac CM, Andrei AS, Bulzu PA, Levei EA, Cadar O, Battes KP, Cimpean M, Senila M, Cristea A *et al*: **Spatio-temporal insights into microbiology of the freshwater-to-hypersaline, oxic-hypoxic-euxinic waters of Ursu Lake**. Environ Microbiol 2021, **23**(7):3523–3540.
24. Tran PQ, Bachand SC, McIntyre PB, Kraemer BM, Vadeboncoeur Y, Kimirei IA, Tamatamah R, McMahon KD, Anantharaman K: **Depth-discrete metagenomics reveals the roles of microbes in biogeochemical cycling in the tropical freshwater Lake Tanganyika**. ISME J 2021, **15**(7):1971–1986.
25. Castelle CJ, Brown CT, Thomas BC, Williams KH, Banfield JF: **Unusual respiratory capacity and nitrogen metabolism in a Parcubacterium (OD1) of the Candidate Phyla Radiation**. Sci Rep 2017, **7**:40101.
26. Martinez-Cano DJ, Reyes-Prieto M, Martinez-Romero E, Partida-Martinez LP, Latorre A, Moya A, Delaye L: **Evolution of small prokaryotic genomes**. Front Microbiol 2014, **5**:742.
27. Green ER, Mecsas J: **Bacterial Secretion Systems: An Overview**. Microbiol Spectr 2016, **4**(1).
28. Okazaki Y, Nishimura Y, Yoshida T, Ogata H, Nakano SI: **Genome-resolved viral and cellular metagenomes revealed potential key virus-host interactions in a deep freshwater lake**. Environ Microbiol 2019, **21**(12):4740–4754.
29. Mukherjee I, Salcher MM, Andrei AS, Kavagutti VS, Shabarova T, Grujcic V, Haber M, Layoun P, Hodoki Y, Nakano SI *et al*: **A freshwater radiation of diplonemids**. Environ Microbiol 2020.

30. Kavagutti VS, Andrei AS, Mehrshad M, Salcher MM, Ghai R: **Phage-centric ecological interactions in aquatic ecosystems revealed through ultra-deep metagenomics**. Microbiome 2019, **7**(1):135.
31. Bushnell B, Rood J, Singer E: **BBMerge - Accurate paired shotgun read merging via overlap**. PLoS One 2017, **12**(10):e0185056.
32. Li D, Liu CM, Luo R, Sadakane K, Lam TW: **MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph**. Bioinformatics 2015, **31**(10):1674–1676.
33. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, Wang Z: **MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies**. PeerJ 2019, **7**:e7359.
34. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ: **Prodigal: prokaryotic gene recognition and translation initiation site identification**. BMC Biointomatics 2010, **11**:119.
35. Steinegger M, Soding J: **MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets**. Nat Biotechnol 2017, **35**(11):1026–1028.
36. Roux S, Enault F, Hurwitz BL, Sullivan MB: **VirSorter: mining viral signal from microbial genomic data**. PeerJ 2015, **3**:e985.
37. 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**(1):90.
38. Anantharaman K, Brown CT, Burstein D, Castelle CJ, Probst AJ, Thomas BC, Williams KH, Banfield JF: **Analysis of five complete genome sequences for members of the class Peribacteria in the recently recognized Peregrinibacteria bacterial phylum**. PeerJ 2016, **4**:e1607.
39. 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**(12):2864–2868.
40. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH: **GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database**. *Bioinformatics* 2019.
41. Edgar RC: **Search and clustering orders of magnitude faster than BLAST**. Bioinformatics 2010, **26**(19):2460–2461.
42. Nawrocki EP: **Structural RNA homology search and alignment using covariance models**. *PhD thesis.* Washington University in St. Luis; 2009.
43. Maidak BL, Olsen GJ, Larsen N, Overbeek R, M.J. M, Woese CR: **The RDP (Ribosomal Database Project)**. Nucleic Acids Res 1996, **25**(1):109–110.
44. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glockner FO: **The SILVA ribosomal RNA gene database project: improved data processing and web-based tools**. Nucleic Acids Res 2013, **41**(Database issue):D590-596.
45. Huang Y, Li W, Finn PW, Perkins DL: **Ribosomal RNA identification in metagenomic and metatranscriptomic datasets**. In: *Handbook of Molecular Microbial Ecology, Metagenomics and*

Complementary Approaches. Edited by de Bruijn FJ, vol. 1, 1 edn: John Wiley & Sons; 2011: 387–391.

46. Lowe TM, Eddy WR: **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence**. Nucleic Acids Res 1997, **25**(5):955–964.
47. Finn RD, Clements J, Eddy SR: **HMMER web server: interactive sequence similarity searching**. Nucleic Acids Res 2011, **39**(Web Server issue):W29-37.
48. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV: **The COG database: new developments in phylogenetic classification of proteins from complete genomes**. Nucleic Acids Res 2001, **29**(1):22–28.
49. Haft DH, Loftus BJ, Richardson DL, Yang F, Eisen JA, Paulsen IT, White O: **TIGRFAMs: a protein family resource for the functional identification of proteins**. Nucleic Acids Res 2001, **29**(1):41–43.
50. Mistry J, Bateman A, Finn RD: **Predicting active site residue annotations in the Pfam database**. BMC Bioinformatics 2007, **8**:298.
51. Kanehisa M, Sato Y, Morishima K: **BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences**. J Mol Biol 2016, **428**(4):726–731.
52. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G *et al*: **InterProScan 5: genome-scale protein function classification**. Bioinformatics 2014, **30**(9):1236–1240.
53. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes**. Nucleic Acids Res 2000, **28**(1):27–30.
54. Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y: **dbCAN: a web resource for automated carbohydrate-active enzyme annotation**. Nucleic Acids Res 2012, **40**(Web Server issue):W445-451.
55. Weese D, Holtgrewe M, Reinert K: **RazerS 3: faster, fully sensitive read mapping**. Bioinformatics 2012, **28**(20):2592–2599.
56. Emiola A, Oh J: **High throughput in situ metagenomic measurement of bacterial replication at ultra-low sequencing coverage**. Nat Commun 2018, **9**(1):4956.
57. Weissman JL, Hou S, Fuhrman JA: **Estimating maximal microbial growth rates from cultures, metagenomes, and single cells via codon usage patterns**. Proc Natl Acad Sci U S A 2021, **118**(12).
58. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D *et al*: **Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation**. Nucleic Acids Res 2016, **44**(D1):D733-745.
59. 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.
60. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ: **IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies**. Mol Biol Evol 2015, **32**(1):268–274.
61. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS: **ModelFinder: fast model selection for accurate phylogenetic estimates**. Nat Methods 2017, **14**(6):587–589.

62. Sousa FL, Alves RJ, Pereira-Leal JB, Teixeira M, Pereira MM: **A bioinformatics classifier and database for heme-copper oxygen reductases**. PLoS One 2011, **6**(4):e19117.
63. Katoh K, Standley DM: **MAFFT multiple sequence alignment software version 7: improvements in performance and usability**. Mol Biol Evol 2013, **30**(4):772–780.
64. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS: **UFBoot2: Improving the Ultrafast Bootstrap Approximation**. Mol Biol Evol 2018, **35**(2):518–522.
65. Naser-Khdour S, Minh BQ, Zhang W, Stone EA, Lanfear R: **The Prevalence and Impact of Model Violations in Phylogenetic Analysis**. Genome Biol Evol 2019, **11**(12):3341–3352.
66. Jaffe AL, Castelle CJ, Matheus Carnevali PB, Gribaldo S, Banfield JF: **The rise of diversity in metabolic platforms across the Candidate Phyla Radiation**. BMC Biol 2020, **18**(1):69.
67. UniProt C: **UniProt: the universal protein knowledgebase in 2021**. Nucleic Acids Res 2021, **49**(D1):D480-D489.
68. Kall L, Krogh A, Sonnhammer EL: **An HMM posterior decoder for sequence feature prediction that includes homology information**. Bioinformatics 2005, **21 Suppl 1**:i251-257.
69. Mirarab S, Nguyen N, Guo S, Wang LS, Kim J, Warnow T: **PASTA: Ultra-Large Multiple Sequence Alignment for Nucleotide and Amino-Acid Sequences**. J Comput Biol 2015, **22**(5):377–386.
70. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R: **IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era**. Mol Biol Evol 2020, **37**(5):1530–1534.
71. Salcher MM, Andrei AS, Bulzu PA, Keresztes ZG, Banciu HL, Ghai R: **Visualization of Lokiarchaeia and Heimdallarchaeia -Asgardarchaeota- by Fluorescence In Situ Hybridization and Catalyzed Reporter Deposition**. *mSphere* 2020, **5**(4):e00686-00620.
72. Pruesse E, Peplies J, Glockner FO: **SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes**. Bioinformatics 2012, **28**(14):1823–1829.
73. Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar, Buchner A, Lai T, Steppi S, Jobb G *et al*: **ARB: a software environment for sequence data**. Nucleic Acids Res 2004, **32**(4):1363–1371.
74. Stamatakis A, Ludwig T, Meier H: **RAxML-II: a program for sequential, parallel and distributed inference of large phylogenetic trees**. Concurrency and Computation: Practice and Experience 2005, **17**(14):1705–1723.
75. Fuchs BM, Glockner FO, Wulf J, Amann R: **Unlabeled Helper Oligonucleotides Increase the In Situ Accessibility to 16S rRNA of Fluorescently Labeled Oligonucleotide Probes**. Appl Environ Microbiol 2000, **66**(8):3603–3607.
76. Yilmaz LS, Parnerkar S, Noguera DR: **mathFISH, a web tool that uses thermodynamics-based mathematical models for in silico evaluation of oligonucleotide probes for fluorescence in situ hybridization**. Appl Environ Microbiol 2011, **77**(3):1118–1122.
77. Pernthaler A, Pernthaler J, Amann R: **Fluorescence in situ hybridization and catalyzed reporter deposition for the identification of marine bacteria**. Appl Environ Microbiol 2002, **68**(6):3094–3101.

78. Wallner G, Amann R, Beisker W: **Optimizing fluorescent in situ hybridization with rRNA-targeted oligonucleotide probes for flow cytometric identification of microorganisms**. Cytometry 1993, **14**:136–143.
79. Daims H, Brühl A, Amann R, Schleifer K-H, Wagner M: **The Domain-specific Probe EUB338 is Insufficient for the Detection of all Bacteria: Development and Evaluation of a more Comprehensive Probe Set**. Systematic and Applied Microbiology 1999, **22**(3):434–444.
80. Sekar R, Pernthaler A, Pernthaler J, Warnecke F, Posch T, Amann R: **An improved protocol for quantification of freshwater Actinobacteria by fluorescence in situ hybridization**. Appl Environ Microbiol 2003, **69**(5):2928–2935.
81. Shabarova T, Kasalicky V, Simek K, Nedoma J, Znachor P, Posch T, Pernthaler J, Salcher MM: **Distribution and ecological preferences of the freshwater lineage LimA (genus Limnohabitans) revealed by a new double hybridization approach**. Environ Microbiol 2017, **19**(3):1296–1309.
82. Neuenschwander SM, Ghai R, Pernthaler J, Salcher MM: **Microdiversification in genome-streamlined ubiquitous freshwater Actinobacteria**. ISME J 2018, **12**(1):185–198.
83. Salcher MM, Pernthaler J, Posch T: **Seasonal bloom dynamics and ecophysiology of the freshwater sister clade of SAR11 bacteria 'that rule the waves' (LD12)**. ISME J 2011, **5**(8):1242–1252.
84. Decho AW, Gutierrez T: **Microbial Extracellular Polymeric Substances (EPSs) in Ocean Systems**. Front Microbiol 2017, **8**:922.
85. Grossart HP, Simon M: **Significance of limnetic organic aggregates (lake snow) for the sinking flux of particulate organic matter in a large lake**. Aquatic Microbial Ecology 1998, **15**:115–125.
86. Wrighton KC, Thomas BC, Sharon I, Miller CS, Castelle CJ, VerBerkmoes NC, Wilkins MJ, Hettich RL, Lipton MS, Williams KH *et al*: **Fermentation, Hydrogen, and Sulfur Metabolism in Multiple Uncultivated Bacterial Phyla**. Science 2012, **337**:1661–1665.
87. Lemos LN, Medeiros JD, Dini-Andreote F, Fernandes GR, Varani AM, Oliveira G, Pylro VS: **Genomic signatures and co-occurrence patterns of the ultra-small Saccharimonadia (phylum CPR/Patescibacteria) suggest a symbiotic lifestyle**. Mol Ecol 2019, **28**(18):4259–4271.
88. Hoshino Y, Gaucher EA: **On the Origin of Isoprenoid Biosynthesis**. Mol Biol Evol 2018, **35**(9):2185–2197.
89. Doron S, Melamed S, Ofir G, Leavitt A, Lopatina A, Keren M, Amitai G, Sorek R: **Systematic discovery of antiphage defense systems in the microbial pangenome**. Science 2018, **359**(6379).
90. Moreira D, Zivanovic Y, López-Archilla AI, Iniesto M, López-García P: **Reductive evolution and unique infection and feeding mode in the CPR predatory bacterium Vampirococcus lugosii**. *bioRxiv* 2020.
91. Korotkov KV, Sandkvist M, Hol WG: **The type II secretion system: biogenesis, molecular architecture and mechanism**. Nat Rev Microbiol 2012, **10**(5):336–351.
92. Beam JP, Becraft ED, Brown JM, Schulz F, Jarett JK, Bezuidt O, Poulton NJ, Clark K, Dunfield PF, Ravin NV *et al*: **Ancestral Absence of Electron Transport Chains in Patescibacteria and DPANN**. Frontiers in Microbiology 2020, **11**.

93. Horn M, Collingro A, Schmitz-Esser S, Beier CL, Purkhold U, Fartmann B, Brandt P, Nyakatura GJ, Droege M, Frishman D *et al*: **Illuminating the evolutionary history of Chlamydiae**. Science 2004, **304**(5671):728–730.
94. Jaffe AL, He C, Keren R, Valentin-Alvarado LE, Munk P, Bouma-Gregson K, Farag IF, Amano Y, Sachdeva R, West PT *et al*: **Patterns of gene content and co-occurrence constrain the evolutionary path toward animal association in CPR bacteria**. *bioRxiv* 2021.
95. Pushkarev A, Inoue K, Larom S, Flores-Uribe J, Singh M, Konno M, Tomida S, Ito S, Nakamura R, Tsunoda SP *et al*: **A distinct abundant group of microbial rhodopsins discovered using functional metagenomics**. Nature 2018, **558**(7711):595–599.
96. Morowitz HJ: **Beginnings of Cellular Life: Metabolism Recapitulates Biogenesis**. New Haven: Yale University Press; 1993.
97. Giovannoni SJ, Cameron Thrash J, Temperton B: **Implications of streamlining theory for microbial ecology**. ISME J 2014, **8**(8):1553–1565.
98. McCutcheon JP, Moran NA: **Extreme genome reduction in symbiotic bacteria**. Nat Rev Microbiol 2011, **10**(1):13–26.
99. Salcher MM, Schaefle D, Kaspar M, Neuenschwander SM, Ghai R: **Evolution in action: habitat transition from sediment to the pelagial leads to genome streamlining in Methylophilaceae**. ISME J 2019, **13**(11):2764–2777.
100. Wetzel RG: **The Phosphorus Cycle**. In: *Limnology Lake and River Ecosystems*. 3rd edn: Academic Press; 2001: 239–288.
101. Peura S, Eiler A, Bertilsson S, Nykanen H, Tiirola M, Jones RI: **Distinct and diverse anaerobic bacterial communities in boreal lakes dominated by candidate division OD1**. ISME J 2012, **6**(9):1640–1652.
102. Jousset A, Bienhold C, Chatzinotas A, Gallien L, Gobet A, Kurm V, Kusel K, Rillig MC, Rivett DW, Salles JF *et al*: **Where less may be more: how the rare biosphere pulls ecosystems strings**. ISME J 2017, **11**(4):853–862.
103. He C, Keren R, Whittaker ML, Farag IF, Doudna JA, Cate JHD, Banfield JF: **Genome-resolved metagenomics reveals site-specific diversity of episymbiotic CPR bacteria and DPANN archaea in groundwater ecosystems**. Nat Microbiol 2021.
104. Tseng TT, Tyler BM, Setubal JC: **Protein secretion systems in bacterial-host associations, and their description in the Gene Ontology**. BMC Microbiol 2009, **9 Suppl 1**:S2.
105. Meheust R, Burstein D, Castelle CJ, Banfield JF: **The distinction of CPR bacteria from other bacteria based on protein family content**. Nat Commun 2019, **10**(1):4173.
106. Schneewind O, Missiakas DM: **Protein secretion and surface display in Gram-positive bacteria**. Philos Trans R Soc Lond B Biol Sci 2012, **367**(1592):1123–1139.
107. Kontinen VP, Sarvas M: **The PrsA lipoprotein is essential for protein secretion in Bacillus subtilis and sets a limit for high-level secretion**. Molecular Microbiology 1993, **8**(4):727–737.

108. Kuhn A, Stuart R, Henry R, Dalbey RE: **The Alb3/Oxa1/YidC protein family: membrane-localized chaperones facilitating membrane protein insertion?** Trends Cell Biol 2003, **13**(10):510–516.
109. Murugkar PP, Collins AJ, Chen T, Dewhirst FE: **Isolation and cultivation of candidate phyla radiation Saccharibacteria (TM7) bacteria in coculture with bacterial hosts**. J Oral Microbiol 2020, **12**(1):1814666.
110. White D: **The Physiology and Biochemistry of Prokaryotes**, Third edition edn. New York: Oxford University Press; 2007.
111. Schmitz-Esser S, Linka N, Collingro A, Beier CL, Neuhaus HE, Wagner M, Horn M: **ATP/ADP translocases: a common feature of obligate intracellular amoebal symbionts related to Chlamydiae and Rickettsiae**. J Bacteriol 2004, **186**(3):683–691.
112. Greub G, Raoult D: **History of the ADP/ATP-translocase-encoding gene, a parasitism gene transferred from a Chlamydiales ancestor to plants 1 billion years ago**. Appl Environ Microbiol 2003, **69**(9):5530–5535.
113. Grassle F, Plugge C, Franchini P, Schink B, Schleheck D, Muller N: **Pelorhabdus rhamnosifermentans gen. nov., sp. nov., a strictly anaerobic rhamnose degrader from freshwater lake sediment**. Syst Appl Microbiol 2021, **44**(4):126225.
114. Govorunova EG, Sineshchekov OA, Li H, Spudich JL: **Microbial Rhodopsins: Diversity, Mechanisms, and Optogenetic Applications**. Annu Rev Biochem 2017, **86**:845–872.
115. Inoue K, Tsukamoto T, Sudo Y: **Molecular and evolutionary aspects of microbial sensory rhodopsins**. Biochim Biophys Acta 2014, **1837**(5):562–577.
116. Gordeliy VI, Labahn J, Moukhametzianov R, Efremov R, Granzin J, Schlesinger R, Buldt G, Savopol T, Scheidig AJ, Klare JP *et al*: **Molecular basis of transmembrane signalling by sensory rhodopsin II - transducer complex**. Nature 2002, **419**:484–487.
117. Bulzu PA, Kavagutti VS, Chiriac MC, Vavourakis CD, Inoue K, Kandori H, Andrei AS, Ghai R: **Heliorhodopsin evolution is driven by photosensory promiscuity in monoderms**. *bioRxiv* 2021.

Figures

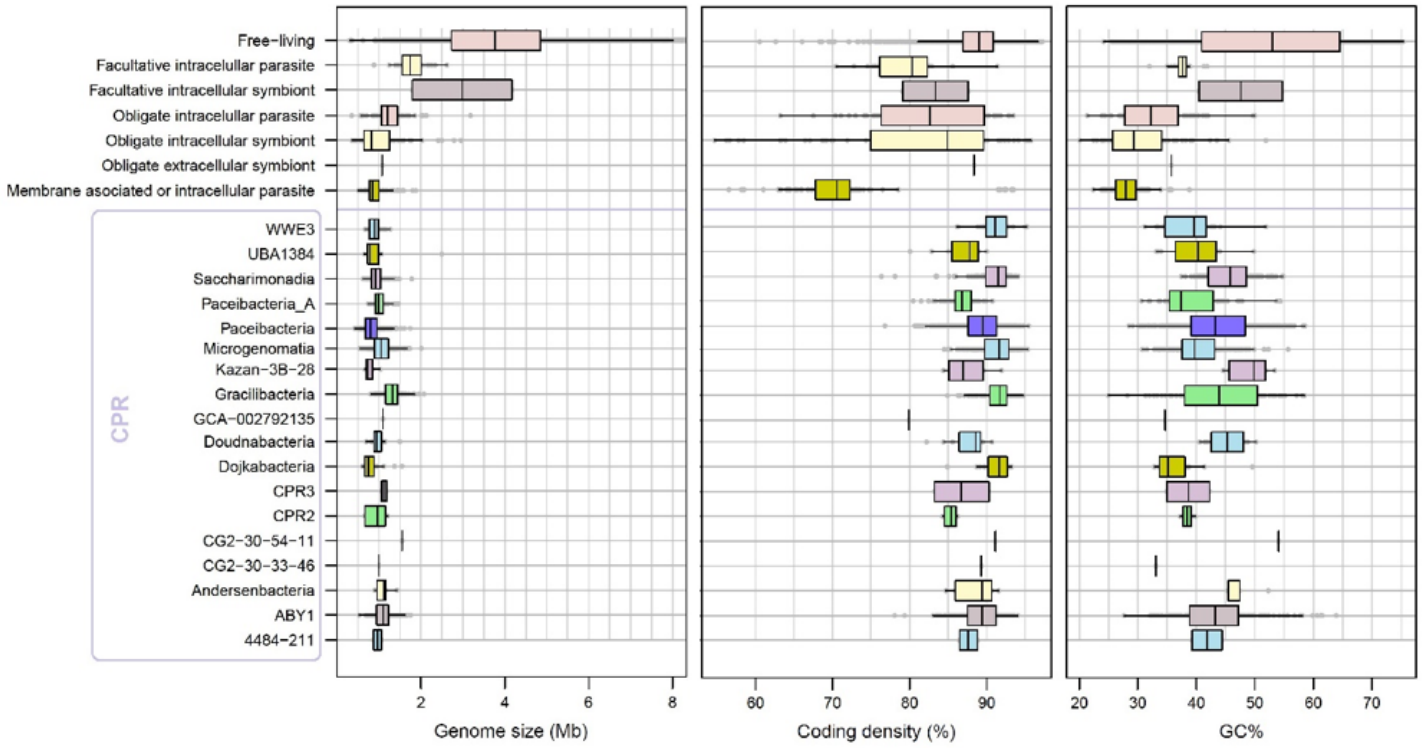

Figure 1

General genome characteristics for CPR classes compared to free-living bacteria and known parasites or symbionts. All representative CPR genomes from GTDB r89 were selected for this purpose together to the 282 MAGs assembled in this study. RefSeq 81 database was manually curated, and the genomes were classified in different life-strategies categories.

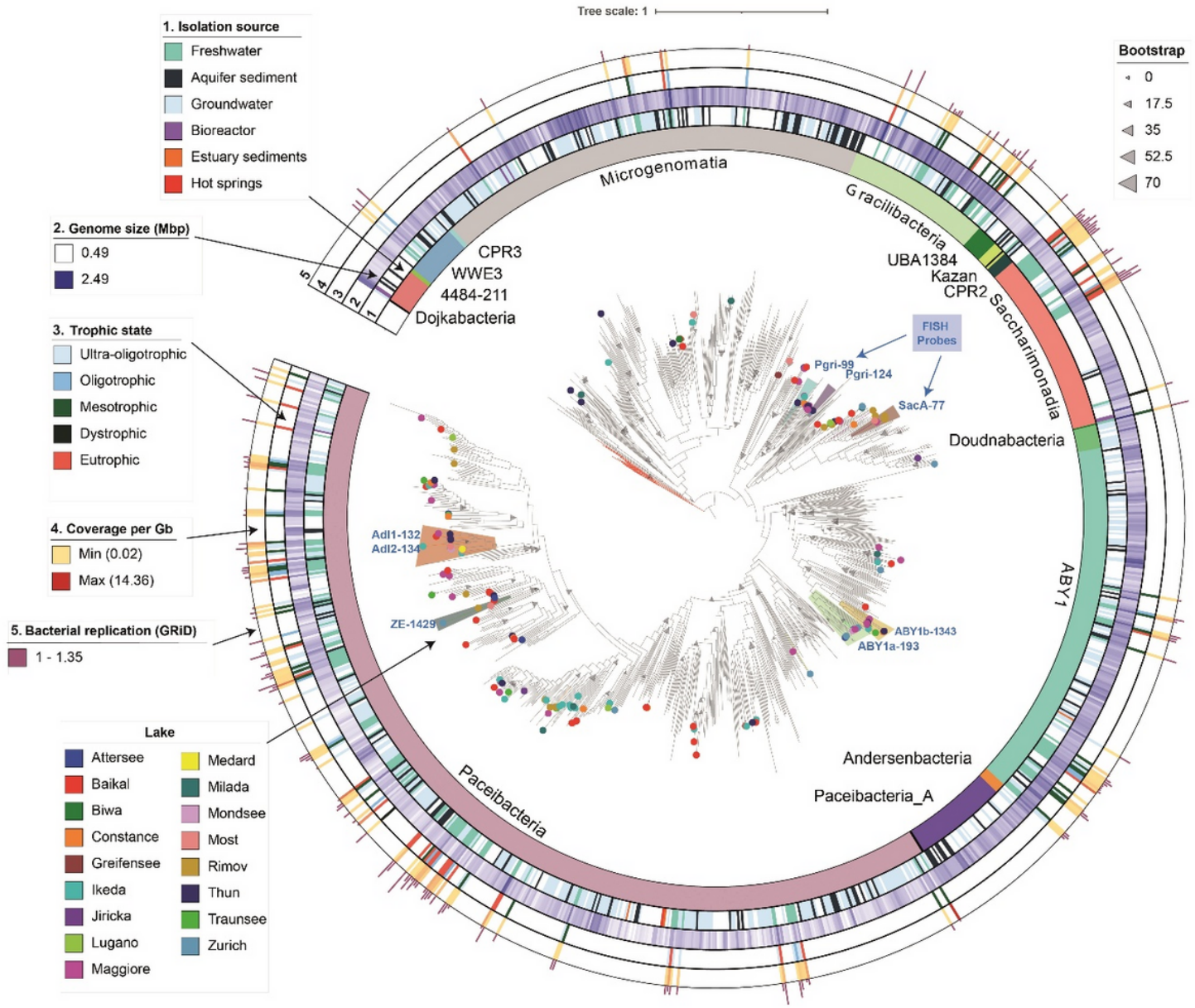

Figure 2

Maximum likelihood (LG+R10, general matrix and FreeRate model with 10 categories for amino acid substitution; 1000 ultrafast bootstraps) phylogeny for the CPR radiation based on 38 concatenated SCGs (Supplementary Tab. 2). The lake origin of each freshwater MAG obtained from this study is marked by a dot with a different color at the end of the tips. CPR classes are shown with distinct colors in the inner circle. Following annotations starting from the inner circles represent 1) the isolation source of the genome; 2) the estimated genome size; 3) the trophic state of the lake; 4) the relative abundance of each MAG in metagenomic read recruitment expressed as coverage per Gb of metagenome; 5) the GRiD values for estimation of bacterial replication rates. Probe targets for CARD-FISH visualization are indicated by different colors and the name of probes marked with light blue.

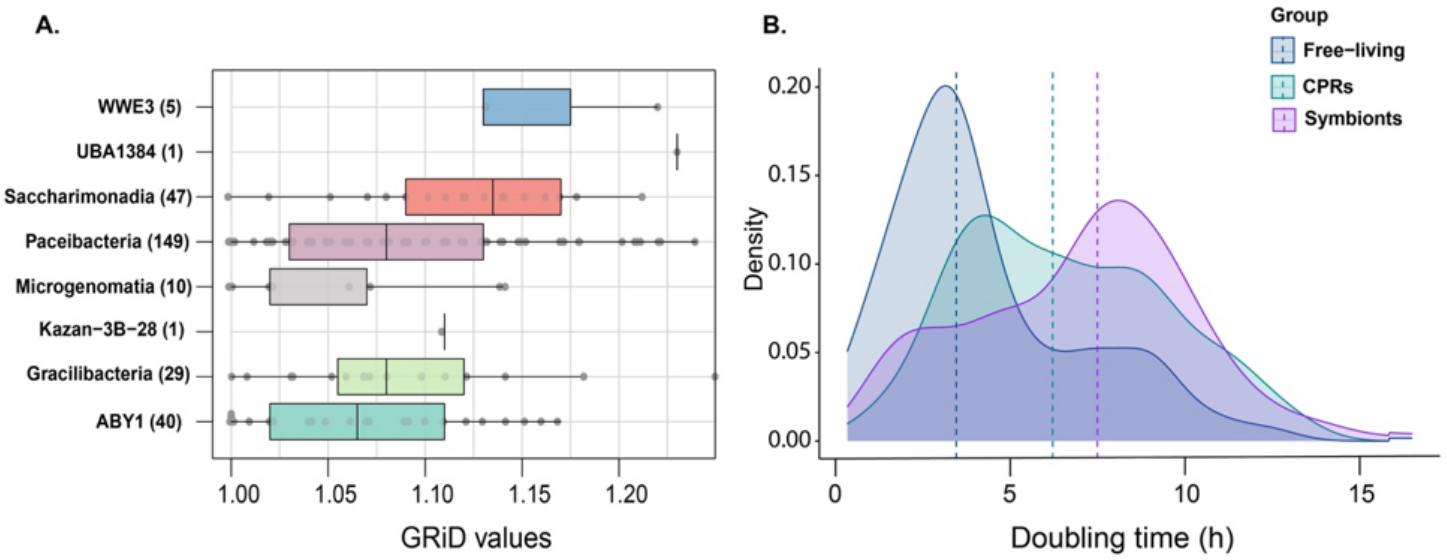

Figure 3

A. Boxplot of GRiD values for freshwater CPRs according to their classes. B. Growth rate estimations for freshwater CPRs obtained in this study and for our collection of free-living organisms' genomes and symbionts. Doubling time was predicted using gRodon and the median value for each group is represented by a vertical line.

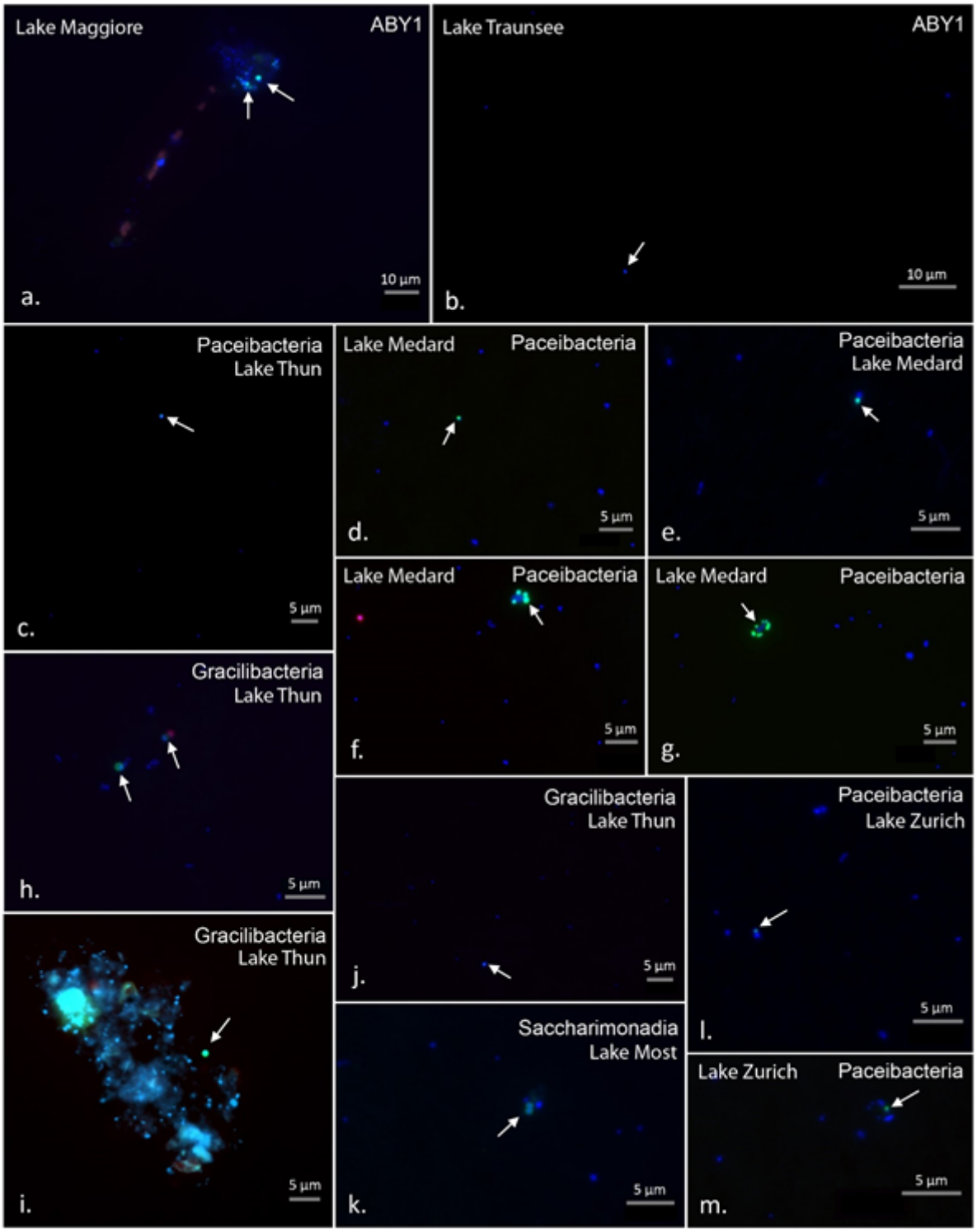

Figure 4

CARD-FISH imaging of different CPR clades. The panels show an overlap of the probe (green), DAPI (blue) and autofluorescence (red) signals. a, b) ABY1 members from the GWF2-40-263 and UBA9934 families stained with 2 distinct probes (probe ABY1a-193 and ABY1b-1343). c-f) Paceibacteria proposed genus GWA1-54-10 visualized with 2 probes (adl1-132 and adl2-134). h-i) Bacteria affiliated to Gracilibacteria proposed genus 2-02-FULL-48-14 are observed using the Pgri-99 probe. j) Organisms of the family level group LOWO2-01-FULL-3 (Gracilibacteria) stained with the Pgri-124 probe. k) Members of

the Saccharimonadia uncultivated family UBA10212 were observed using the SacA-77 probe. l, m) Paceibacteria family level group UBA11359 was detected at the surface of other prokaryotes using the ZE-1429 probe.

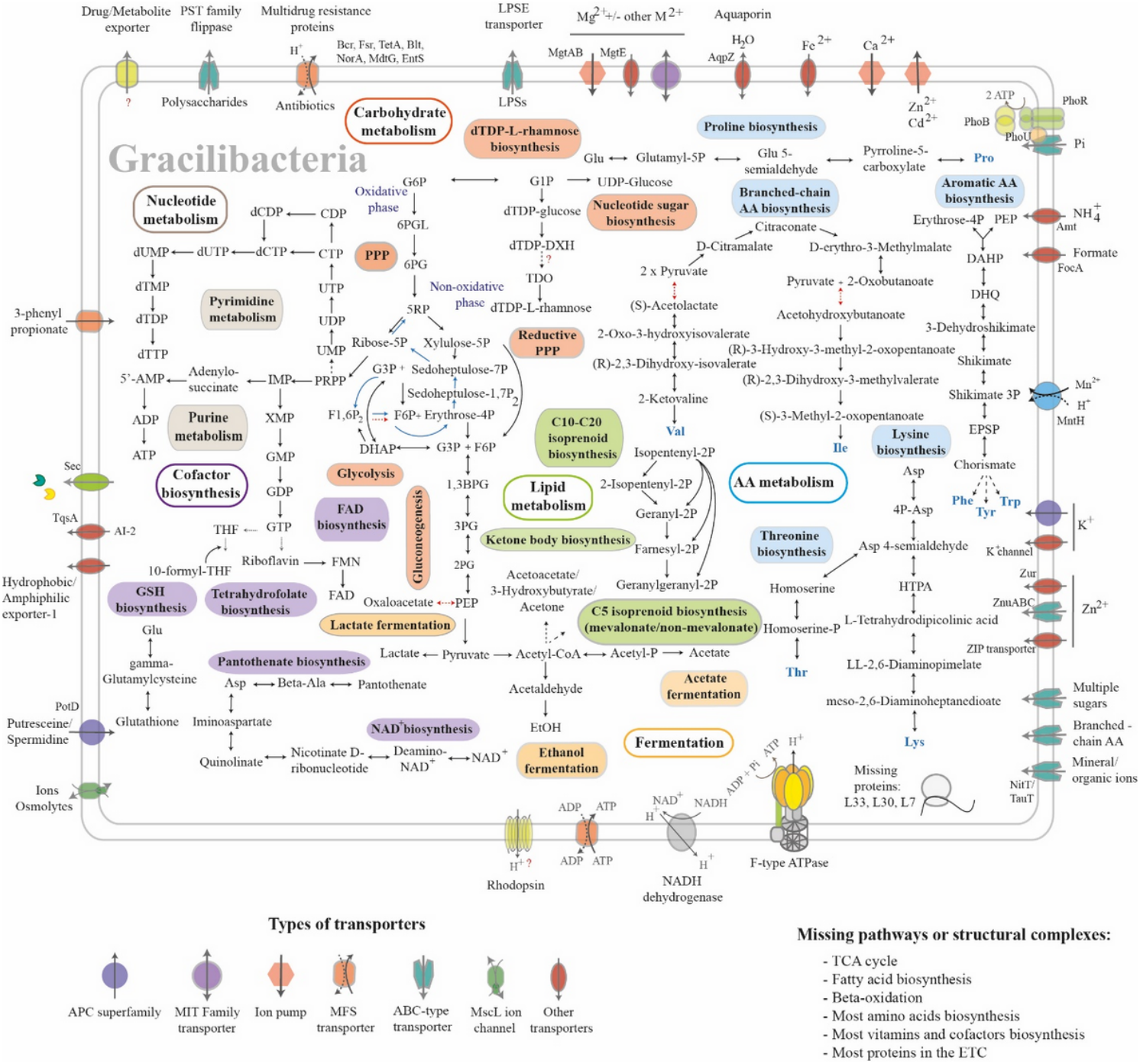

Figure 5

Metabolic map reconstruction for the Gracilibacteria class. **Abbreviations for transporters**: ABC – ATP-binding cassette, APC – Amino acid-polyamine-organocation, MIT – metal ion transporter, MFS – major facilitator superfamily, MscL – large conductance mechanosensitive ion channel, PSTE - polysaccharide transporter. **Abbreviations for compounds:** 1,3BPG – 1,3-bisphosphoglycerate, 2PG – 2-phosphoglycerate, 3PG – 3-phosphoglycerate, 5RP – ribulose 5-phosphate, 6PG – 6-phosphogluconate, 6PGL – 6-

phoshoglucono lactone, DAHP - 2-Dehydro-3-deoxy-D-arabino-heptonate 7-phosphate, DHAP – dihydroxyacetone phosphate, DHQ – 3-Dehydroquinate, dTDP – deoxythymidine diphosphate, dTDP-DXH – dTDP-6-deoxy-D-xylo-4-hexulose, EPSP - 5-enolpyruvylshikimate-3-phosphate, F1,6P2 – fructose 1,6-bisphosphate, F6P – fructose 6-phosphate, FAD – flavin adenine dinucleotide, FMN – flavin mononucleotide, G3P – glyceraldehyde 3-phosphate, G6P – glucose 6-phosphate, HTPA - (2S,4S)-4-Hydroxy-2,3,4,5-tetrahydrodipicolinic acidIMP - inosine monophosphate, LPS – lipopolysaccharides, NAD^+ - nicotinamide adenine dinucleotide, P – phosphate, PEP – phosphoenolpyruvate, PPi – pyrophosphate, PPRP - phosphoribosyl pyrophosphate, TDO –dTDP-4-oxo-L-rhamnose, THF – tetrahydrofolate, TXN – thioredoxin, TXN-S-S - thioredoxin disulfide, XMP - xanthosine monophosphate. Pathways or structural complexes: ETC – Electric transport chain, TCA – Tricarboxylic acid cycle.

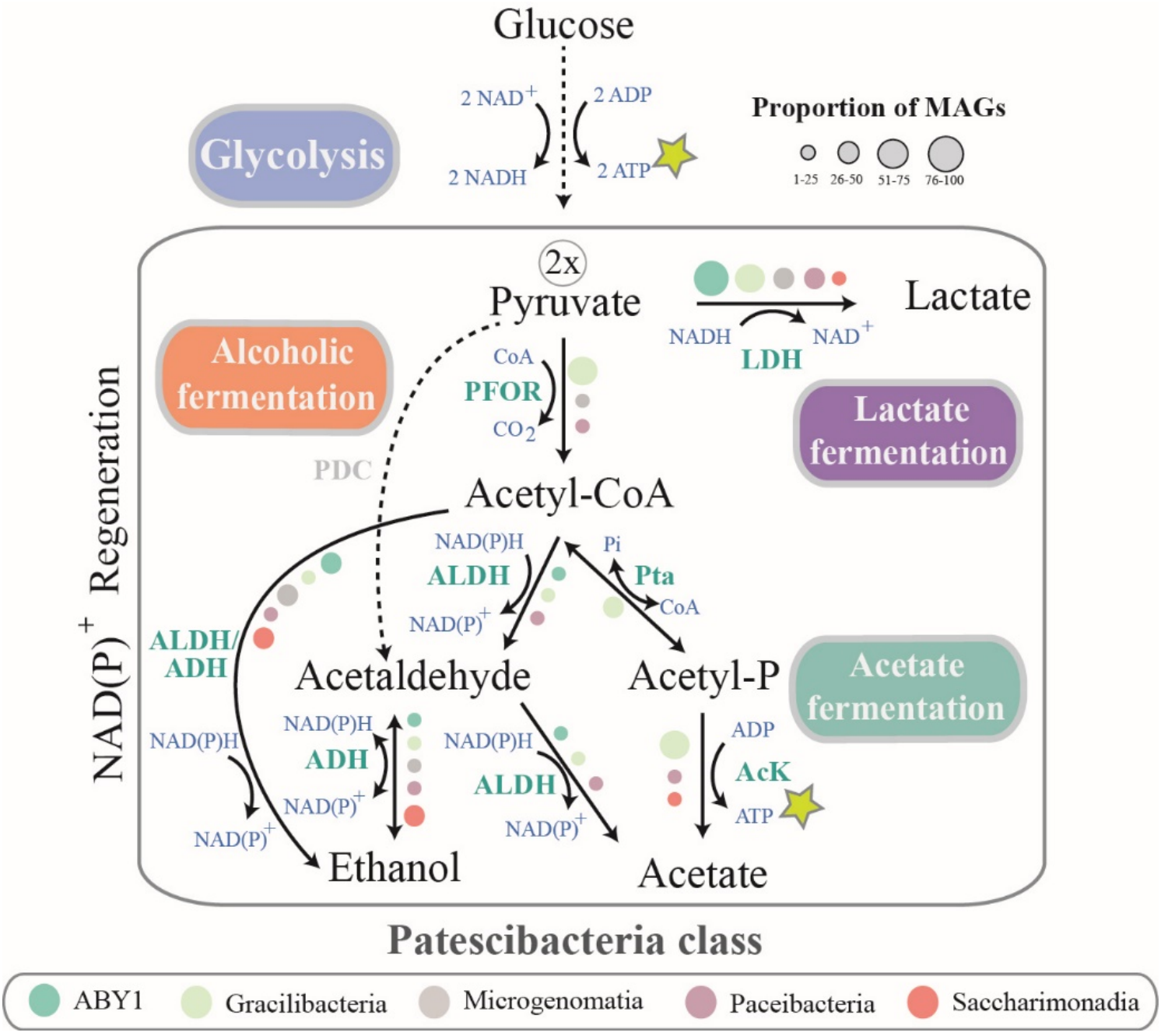

Figure 6

Capacity to perform certain types of fermentation in freshwater CPR. Colors mark different classes and the proportions of MAGs in each of them encoding specific enzymes is depicted by symbol size.
Abbreviations for enzymes: AcK – acetate kinase, ADH – alcohol dehydrogenase, ALDH – aldehyde dehydrogenase, LDH – lactate dehydrogenase, PFOR - pyruvate:ferredoxin oxidoreductase, Pta – phosphotransacetylase.

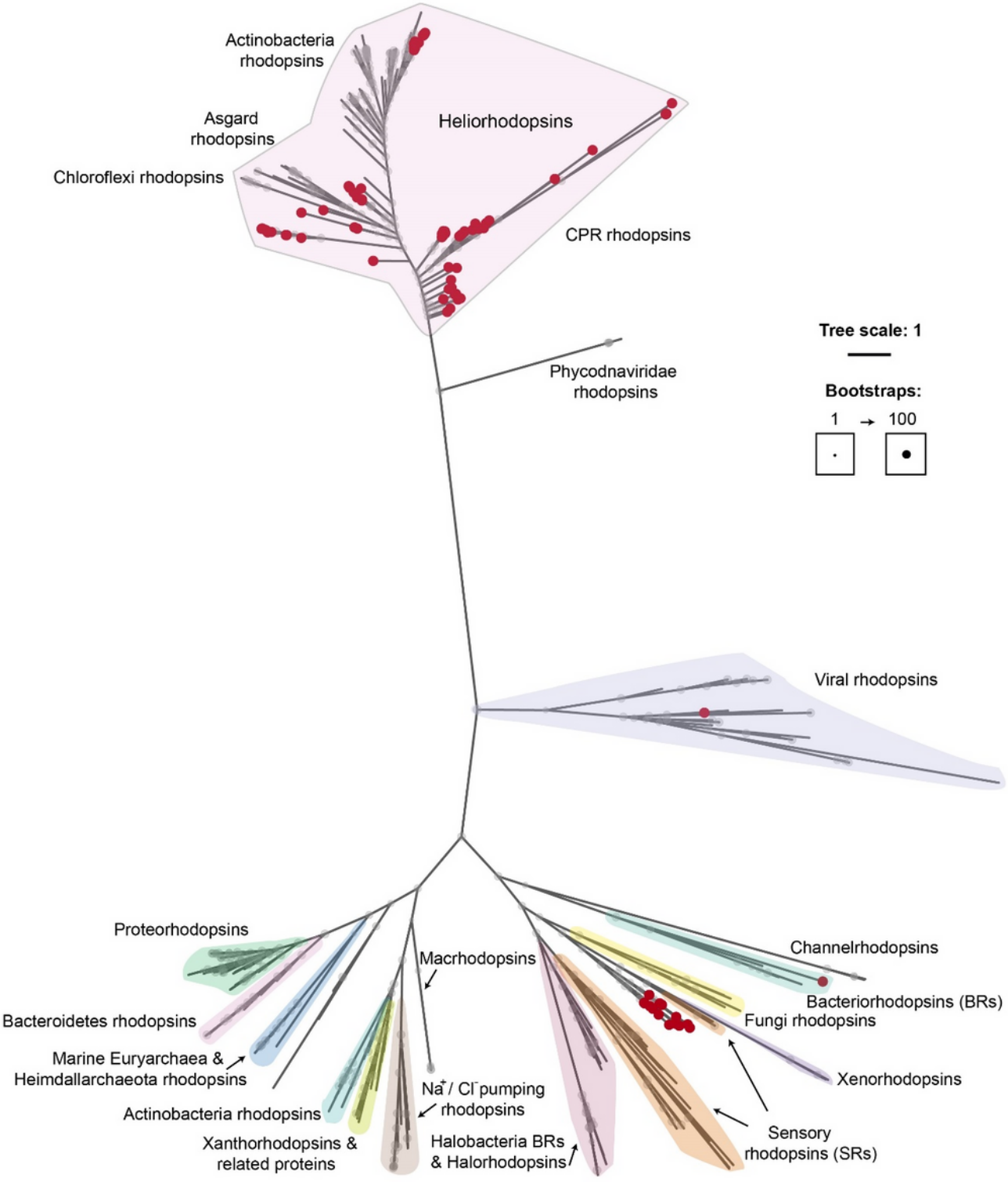

Figure 7

Maximum likelihood (LG+F+G4, general matrix with empirical codon frequencies counted from data and discrete Gamma model with 4 rate categories) phylogeny for rhodopsins. CPR rhodopsins are marked with a red dot at the end of the tips. A total of 511 sequences were used to generate the tree (alignment length - 792), including a database of 392 proteins from known rhodopsin families.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalFile1.doc
- SupplementaryFigures.doc
- Supplementarytables.xlsx