



Adaption mechanism and ecological role of CPR bacteria in brackish-saline groundwater



Jiawen Wang^{1,2,3,4}, Haohui Zhong^{1,2,4}, Qian Chen^{1,2}✉ & Jinren Ni^{1,2}

Candidate phyla radiation (CPR) constitutes a substantial fraction of bacterial diversity, yet their survival strategies and biogeochemical roles in brackish-saline groundwater remain unknown. By reconstructing 399 CPR metagenome-assembled genomes (MAGs) and 2007 non-CPR MAGs, we found that CPR, affiliated with 44 previously proposed phyla and 8 putative novel phyla, played crucial roles in maintaining the microbial stability and complexity in groundwater. Metabolic reconstructions revealed that CPR participated in diverse processes, including carbon, nitrogen, and sulfur cycles. Adaption of CPR to high-salinity conditions could be attributed to abundant genes associated with heat shock proteins, osmoprotectants, and sulfur reduction, as well as their cooperation with Co-CPR (non-CPR bacteria co-occurred with CPR) for metabolic support and resource exchange. Our study enhanced the understanding of CPR biodiversity in high-salinity groundwater, highlighting the collaborative roles of self-adaptive CPR bacteria and their reciprocal partners in coping with salinity stress, maintaining ecological stability, and mediating biogeochemical cycling.

The candidate phyla radiation (CPR), as a recently discovered bacterial lineage, constitutes a substantial fraction of all bacterial diversity^{1,2}. CPR organisms are widely distributed, but our current knowledge is limited due to the challenges associated with obtaining pure isolate cultures³. Recent advancements in cultivation-free methods, such as genome-resolved metagenomics and single-cell sequencing, have greatly expanded the diversity of this radiation^{4–6}. CPR members were initially recognized as a monophyletic group within the bacterial domain and were classified under the superphylum Pateschibacteria according to the GTDB Genome Taxonomy Database⁷. Recent studies have further subdivided CPR bacteria into a minimum of 80 potentially phylum-level lineages to achieve improved resolution and establish distinct names for each lineage^{8–10}. Growing metagenomic explorations have extended the CPR members in diverse habitats, such as drinking water¹¹, sediments¹², soils¹³, and the human oral cavity³, and CPR organisms are known to be dominant in groundwater environments⁵.

Detangling the diversity and function of CPR memberships is crucial for understanding the biogeochemical roles of CPR lineages in groundwater ecosystems. Previous research has indicated that CPR can be differentiated based on specific host preferences or hydrogeochemical parameters¹⁴. For instance, *Ca. Uhrbacteria* could establish commensal or mutualistic relationships with *Ignavibacteria* and *Betaproteobacteria* in agriculturally

impacted groundwater¹⁰, whereas *Patescibacteria* might exhibit preferred associations with *Omnitrophota*, *Bacteroidota*, and *Nitrospirota* in oxic and anoxic groundwaters¹⁵. Numerous metagenomic studies have demonstrated that the relative abundance and dominance of CPR members can be influenced by environmental fluctuations, such as acetate stimulation in groundwater adjacent to the Colorado River¹⁶ and shifts between reduction and oxidation conditions in groundwater in the Ohio Department¹¹. Importantly, different CPR members have been reported to potentially participate in various nutrient cycling processes, including carbon transformation, nitrogen denitrification, and sulfate reduction^{10,11,17}. However, comprehensive information regarding the composition structure and metabolic functions of CPR bacteria in coping with brackish-saline groundwater remains largely unknown.

The majority of CPR members exhibit small genomes (~0.5–1.5 Mbp) with limited biosynthetic capabilities^{10,18}. Some studies have indicated that CPR bacteria can evolve specific functions while retaining essential functions for growth and reproduction to adapt to groundwater environments⁵. Insights from metabolic reconstructions suggest that CPR bacteria may play essential roles in consuming carbon compounds derived from plants and bacteria but rely on extracting energy and essential biomolecules (e.g. amino acids, vitamins, and nucleotides) from potential host cells^{9,19,20}. These findings suggest that CPR bacteria possibly exhibit host-dependent lifestyles in

¹Key Laboratory of Water and Sediment Sciences, Ministry of Education, 100871 Beijing, P. R. China. ²State Environmental Protection Key Laboratory of All Material Fluxes in River Ecosystems, 100871 Beijing, P. R. China. ³School of Human Settlements and Civil Engineering, Xi'an Jiaotong University, 710049 Xi'an, Shaanxi Province, P.R. China. ⁴These authors contributed equally: Jiawen Wang, Haohui Zhong. ✉e-mail: qianchen@pku.edu.cn

various environments, enabling them to obtain nutrients and energy sources from their hosts^{29,21}. While it has been shown that the overall bacterial community can vary in compositions and metabolic capacities across salinity gradients in groundwater systems²²⁻²⁴, the interactions between host-dependent CPR bacteria and their potential bacterial hosts, as well as the metabolic roles of CPR bacteria and their host partners may play in brackish-saline groundwater, remain poorly characterized.

Here, we successfully recovered 399 medium-high quality CPR genomes from 66 brackish-saline groundwater samples collected from the Aquifer beneath Tang-he Wastewater Reservoir (ATWR) (Supplementary Fig. 1 in Supplementary File 1). Based on previous studies on CPR bacteria^{9,10}, we categorized these 399 CPR genomes into 44 previously proposed phyla and 8 candidate novel phyla. We have provided the first integrated biogeography pattern of CPR members and unveiled the crucial contributions of CPR bacteria to maintaining the stability of the microbial community in brackish-saline groundwater. Furthermore, we have predicted the metabolic capacities and ecological roles of CPR lineages by characterizing possible metabolic interconnections with their co-occurring organisms. These findings highlighted the essential roles played by CPR bacteria and their co-varying partners in coping with environmental stressors and driving elemental cycles within complex groundwater systems.

Methods

Groundwater sampling and hydrochemistry measurement

A total of 66 groundwater samples were collected in wet and dry seasons from 39 newly constructed monitoring wells of the Aquifer beneath Tang-he Wastewater Reservoir (ATWR) ($38^{\circ}47'43''$ – $38^{\circ}48'58''$ N, $115^{\circ}40'11''$ – $115^{\circ}50'33''$ E), Hebei Province, China (Supplementary Fig. 1). The detailed sampling sites were provided in Supplementary Table 1, and the surface geology of the study area was described in our previous work²⁵. Groundwater in this area was mainly derived from atmosphere precipitation and predominated by the Na-Cl-SO₄ type^{23,25}. All the samples were collected after purging the well volume and flushing the system for more than 30 min. Physicochemical parameters including pH, conductivity (COND), water temperature (T), and dissolved oxygen (DO) were measured in situ. To collect the biomass, 3000–12,000 L of groundwater was pumped and filtered through 0.01- μ m hollow fiber membranes (Toray, Japan). Filters were transported to laboratories with dry ice and stored at -80° C until processed. Meanwhile, the 10 L well-mixed groundwater samples were collected and stored at 4° C for measuring physicochemical parameters, including total dissolved solids (TDS), chemical oxygen demand (COD), ammonium nitrogen (NH₄⁺-N), nitrate nitrogen (NO₃⁻-N), total nitrogen (TN), chloride (Cl⁻), fluorine (F⁻), sulfate (SO₄²⁻), and total organic carbon (TOC), according to the Environmental Quality Standards for Surface Water (GB3838-2002) recommended by the Ministry of Ecology and Environment of China^{25,26}. The sulfur isotopic composition ($\delta^{34}\text{S}$ -SO₄) of groundwater was analyzed on a Thermo Scientific Delta V Plus IRMS with a Flash elemental analyzer after the pretreatment by saturating BaCl₂, precipitated as barite, and mixed with excess V₂O₅ for in-line combustion²⁵. Metals including sodium (Na), magnesium (Mg), calcium (Ca), potassium (K), aluminum (Al), iron (Fe), titanium (Ti), manganese (Mn), zinc (Zn), barium (Ba), copper (Cu), lead (Pb), arsenic (As), nickel (Ni), tin (Sn), lithium (Li), chromium (Cr), selenium (Se), cobalt (Co), cadmium (Cd), antimony (Sb), beryllium (Be), thallium (Tl), and molybdenum (Mo) were determined using ICP-MS (X Series II, Thermo Fisher Scientific, USA) and ICP-OES (Prodigy, Leeman, USA), as described in our previous studies^{25,27}.

DNA extraction and sequencing

The biomass on the filters was captured using ultra-sonication and centrifugation. Genomic DNA was extracted using the FastDNA Spin Kit Soil (MP Biomedicals, USA) following the manufacturer's protocol. Genomic DNA was quantified using a NanoDrop Spectrophotometer (NanoDrop Technologies Inc., Wilmington, DE, USA). The metagenomic sequencing was performed on an Illumina Hiseq 3000/4000 platform (Majorbio Company in Shanghai, China), resulting in about 30 Gbp raw reads per sample.

Metagenomic assembly and binning

SeqPrep v1.2 (<https://github.com/jstjohn/SeqPrep>) was used to remove Illumina adapters. Raw reads were trimmed using Sickle v1.33 with the default quality threshold of 20 (-q20 -l 50) (<https://github.com/najoshi/sickle>; default parameters). Further quality filter was carried out using fastp v0.21.0²⁸ with parameters: '-l 140 -t 1 -c' to remove reads shorter than 140 bp, trim the last base of every read, and correct bases in overlapping regions of paired-end reads. Clean reads were assembled by megahit v1.2.9 with k-mer parameters: '--k-min 19 --k-max 139 --k-step 10'²⁹. Assembled contigs over 200 bp were used for the downstream binning analysis by MetaBAT2 v2.12.1 with default parameters³⁰.

Taxonomic assignments and abundance calculation for metagenome-assembled genomes (MAGs)

After binning, the completeness and contamination of MAGs were evaluated using CheckM v1.1.2³¹, and the medium-high quality MAGs (completeness > 70%, contamination < 10%) were kept for further analysis^{5,10}. MAG quality was further assessed using CheckM2 v1.0.2³², which can provide accurate genome quality predictions specifically for CPR genomes. The results indicated that the majority of our CPR genomes exhibited completeness levels exceeding 90% (Supplementary Table 2). The preliminary taxonomic assignments for MAGs were performed using GTDB-Tk v1.3.0⁷. The redundant MAGs were removed by dRep v3.2.0³³ through the 'dereplicate' workflow³⁴. We also conducted digital DNA-DNA hybridization tests³⁵ through the online service GGDC (<https://ggdc.dsmz.de/ggdc.php>)³⁶ to further validate the robustness of the dereplication process performed by dRep (Supplementary Table 3).

To estimate the relative abundance of these MAGs, clean reads from each sample were first mapped to all non-redundant MAGs using BBMap v38.86 with parameters "minid = 0.95". Then, the sequencing depth of each contig in each sample was calculated by the script 'jgi_summarize_bam_contig_depths' from MetaBAT2 v2.12.1³⁰ with default parameters (identity threshold = 97%). Subsequently, the sequencing depth of each MAG in each sample was calculated by the weighted arithmetic mean of the sequencing depths of all its contigs (weighting by the length of each contig). Finally, the relative abundance of a certain MAG in each sample was calculated as 'sequencing depth of the MAG in this sample/total sequencing depths of all MAGs in this sample'.

CPR identification and a straightforward cluster analysis

A total of 399 MAGs annotated as "Patescibacteria" in the GTDB results were classified as CPR. According to previous studies of CPR bacteria^{9,10}, 16 ribosomal proteins (L2, L3, L4, L5, L6, L14, L15, L16, L18, L22, L24, S3, S8, S10, S17, and S19) were chosen as phylogenetic markers. First, we retained 381 CPR MAGs that contained at least 8 single-copy genes from the 16 ribosomal protein genes for further phylogenetic analysis. Reference genomes were obtained from previously published studies to serve as the backbones^{9,10}. The genes encoding the ribosomal proteins in MAGs and reference CPR genomes were extracted and searched using hmmsearch (HMMER v3.3³⁷). Any multiple-copy ribosomal protein gene in MAGs was excluded. Amino acid sequences corresponding to each ribosomal protein were individually aligned using MAFFT v7.471³⁸ with default parameters, followed by trimming and concatenation. Absent marker genes during concatenation were filled with gaps, as shown in Supplementary Table 4 of Supplementary File 2. The phylogeny of CPR MAGs was constructed using IQ-TREE v2.1.3³⁹ with ModelFinder assigning the best-fit ML model (LG + F + R10) according to Bayesian information criterion (BIC), supported by 1000 ultrafast bootstrap replicates. Second, the remaining 18 CPR MAGs, which had fewer than 8 ribosomal proteins, were classified by comparing their average nucleotide identities to those of their relatives. Finally, the 399 CPR genomes obtained in this study were affiliated with 44 previously proposed phyla^{9,10} and 8 candidate novel phyla. Additional phylogenetic analyses with other bacterial phyla as referential backbone (a selected subset of GTDB representative genomes from each bacterial phyla) were also conducted using a concatenation of 120 bacterial markers based

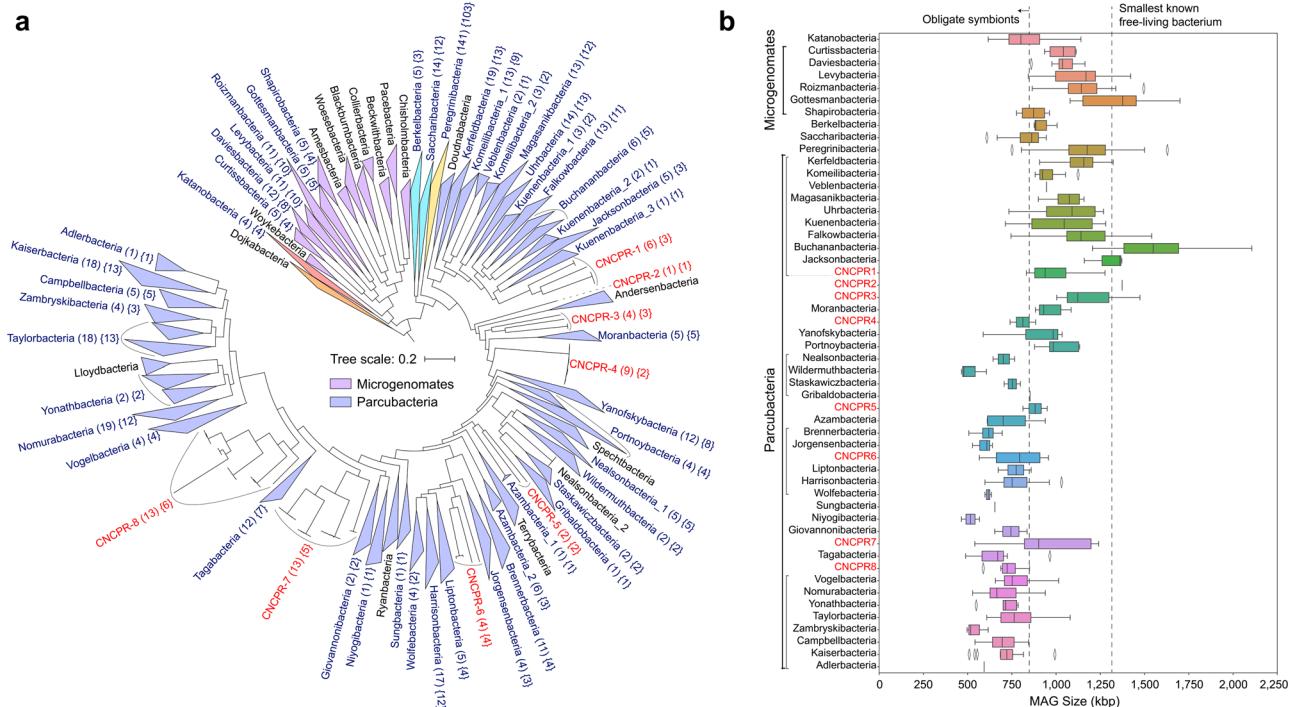


Fig. 1 | Phylogenetic tree and genome size of candidate phyla radiation (CPR) bacteria. **a** The maximum-likelihood tree was inferred from the concatenation of 16 ribosomal proteins and spans a dereplicated set of 381 strain-representative ATWR-MAGs (from a redundant set of 525 ATWR-MAGs) and 283 publicly available reference genomes. The eight phyla colored in red are candidate novel CPR

(CNCPR) affiliating to the superphylum Parcubacteria. The numbers in rounded brackets “()” and curly brackets “{}” indicate the numbers of ATWR-MAGs before and after the strain-level de-replication (>99% ANI), respectively. **b** The size distribution of the obtained CPR MAGs from the ATWR.

on 5036 informative amino-acid sites under the best-fitting model LG + R10. These 120 markers were selected and the concatenating process was conducted both by GTDB-tk⁷.

Functional annotation and metabolic prediction

Open reading frames of MAGs were predicted using prodigal v2.6.3⁴⁰ with the ‘-p meta’ parameter, and annotated using eggNOG-mapper v2.0.0⁴¹ against the eggNOG database 5.0 with parameters ‘-m diamond --seed_ortholog_evalue 1e-5’. Ambiguous annotations or fusion genes were checked and adjusted manually by comparing the results of eggNOG and the online alignment results of the conserved domain database⁴². Meanwhile, the intron prediction was also conducted for each MAG. All the 16S rRNA genes in MAGs were predicted by barrnap from Prokka⁴³ with default parameters and later manually curated. Then these 16S rRNA genes were searched against the online Rfam⁴⁴ database to predict the introns inside these genes. Protein coding regions in these introns were also searched according to the abovementioned eggNOG annotation results.

The metabolisms of these CPR bacteria were predicted by examining the ko numbers in eggNOG annotation results. These ko numbers were mapped to the pathways of the Kyoto Encyclopedia of Genes and Genomes (KEGG) database and then identified for key genes related to biogeochemical cycles (carbon, sulfur, nitrogen, and hydrogen), energy conservation, essential biomolecule syntheses, and other pathways of interests. Pathway completeness was calculated based on the percentage of ko numbers presenting in each KEGG module (pathway). The metabolic reconstructions of MAGs were manually conducted by collectively summarizing the core pathways in the MAGs. The crucial metabolic pathways, ko numbers, and gene names used for the downstream analysis were provided in the Supplementary Table 5.

Statistical analysis

Non-metric multidimensional scaling (NMDS) and analysis of similarity statistics (ANOSIM) were conducted to explore the differences in CPR

members among the prior sampling groups using the *vegan* package in R⁴⁵. One-way analysis of variance (one-way ANOVA) was also performed to test the significant differences of specific CPR at the phylum level among groups. The co-occurrence network between bacteria was visualized using Gephi⁴⁶, based on statistically strong ($|r| > 0.80$) and significant (FDR-adjusted $P < 0.01$) Spearman’s correlations.

Driving forces on CPR members were explored based on the following methods. First, distance-decay patterns of microbial communities (Bray–Curtis distance matrices) along geographical and environmental distances were investigated using the Mantel test based on Spearman’s correlations with *vegan*. Second, random forest analysis was carried out to identify the importance of each environmental variable in structuring CPR members²⁶. The significance of the models and cross-validated R^2 values were determined using the A3 package (1000 permutations). Similarly, the significance of each predictor on CPR members was assessed based on the increase in the mean square error (IncMSE) using the *rfPermute* package. Afterward relationships among seasonal factors, geographical factors, Co-CPR, and CPR were explored by a partial least-squares path model (PLS-PM) using the *vegan* package.

Results

Recovery of 399 CPR and 2007 non-CPR bacterial genomes

A total of 2406 medium-high quality non-redundant MAGs (metagenome-assembled genomes) (completeness > 70%, contamination < 10%, dereplicated at 99% average nucleotide identity) were reconstructed from 66 groundwater samples. The genome dataset included 399 CPR genomes and 2007 non-CPR bacterial genomes according to the preliminary taxonomy assignment by GTDB-tk⁷ (Supplementary Tables 2 and 6 in Supplementary File 2). Utilizing a phylogenetic analysis with 16 ribosomal protein genes as the marker set¹⁰, we further categorized the 399 CPR MAGs into 44 previously proposed CPR phyla^{9,10} and identified 8 candidate novel CPR branches (Fig. 1, alignments provided in Supplementary File 3, and detailed statistics of each protein gene provided in Supplementary Table 4). To

solidify the phylogenetic placement of these 8 candidate novel CPR, we constructed an additional phylogenetic tree using 120 marker genes⁴⁷ recommended by GTDB-Tk (Supplementary Fig. 2, alignments provided in Supplementary File 4), which consistently support their classification as novel lineages. We provisionally designated these new phylum-level lineages as CNCPR1-8 (Fig. 1 and Supplementary Table 2). Further phylogenetic analysis, incorporating referential genomes from every bacterial order in GTDB (Supplementary Fig. 3, alignments provided in Supplementary File 5), supported the notion that CPR bacteria form a monophyletic group^{2,8,13,48}. Notably, the 399 CPR MAGs exhibited diminutive genome sizes, with an average value of 868.6 ± 232.2 kbp (Fig. 1b). The median sizes of most CPR MAGs were smaller than the smallest free-living *Pelagibacter* (1.3 Mbp)⁴⁹, potentially indicating a symbiotic lifestyle for these CPR bacteria as previously reported⁹.

We conducted a comparative analysis of the co-occurrence networks for the entire bacterial community and the subset excluding CPR bacteria, focusing on strong (Spearman $|r| > 0.8$) and significant (P -value < 0.01) correlations. We found that the entire bacterial network exhibited significantly higher node, edge, and degree, but lower average path length, graph density, clustering coefficient, and betweenness centralization (Supplementary Fig. 4a), compared to the co-occurrence network of non-CPR bacteria. This suggested that CPR bacteria played a crucial role in maintaining the stability and complexity of the overall bacterial communities in the face of environmental disturbances^{50,51}. Notably, CPR bacteria, particularly those affiliated with Sungbacteria, Wildermuthbacteria, Nealonbacteria, Daviesbacteria, Yanofskybacteria, and Uhrbacteria, harbored the highest average connection degree in the whole bacterial co-occurrence network, reinforcing their significant roles in maintaining the community stability and complexity (Supplementary Fig. 4b). Compared with the non-CPR bacterial network, the co-occurrence network between CPR and non-CPR displayed higher degree, graph density, and betweenness centralization despite lower nodes and edges (Supplementary Fig. 4a), implying that the complex network structure may contribute to enhancing the resilience of CPR and Co-CPR (non-CPR bacteria co-occurred with CPR) in their adaption to environmental stresses^{50,51}.

We further revealed positive associations between 99 CPR MAGs and 255 non-CPR bacterial MAGs (Fig. 2). Among these CPR members, Daviesbacteria, Nealonbacteria, Yanofskybacteria, Wildermuthbacteria, CNCPR8, Sungbacteria, and Levybacteria displayed the highest connection degrees, with counts of 171, 102, 99, 97, 65, 56, and 48, respectively. Non-CPR bacterial phyla such as Omnitrophota (degree: 430), Chloroflexota (333), Desulfobacterota (151), Planctomycetota (26), Nitrospirota (23), Acidobacteriota (21), Bacteroidota (14), and Methylomirabilota (14) exhibited strong co-occurrence pattern with the CPR members. Specifically, CPR taxa belonging to Daviesbacteria and Nealonbacteria tended to co-exist with microbes affiliated with Omnitrophota, Chloroflexota, and Nealonbacteria, while CNCPR8 exhibited co-occurrence patterns with Omnitrophota, Desulfobacterota, and Chloroflexota. These bacteria co-occurred with CPR (Co-CPR) could provide important guidance for identifying putative hosts of CPR in saline groundwater environments¹⁰.

Variations in relative abundances of CPR bacteria along the hydrochemical gradient

Among the retrieved 52 phylum-level CPR (Fig. 3a), Daviesbacteria, Peregrinibacteria, CNCPR4, and Harrisonbacteria showed notably high relative abundances. These phyla appear to be dominant CPR phyla within this groundwater environment, accounting for 3.02%, 2.58%, 1.60%, and 0.99% of the whole microbial communities, respectively. We further categorized the 66 groundwater samples into three groups based on their salinity levels. Group1 was characterized by the lowest levels (e.g., TDS: 93.85–956.31 mg/L, SO_4^{2-} : 70.26–579.65 mg/L, Na: 88.85–386.58 mg/L), while samples in Group2 displayed moderate salinity (e.g., TDS: 173.72–4120.69 mg/L, SO_4^{2-} : 92.75–2761.22 mg/L, Na: 154.73–1168.64 mg/L), and those in Group3 exhibited the highest salinity (e.g., TDS: 2108.60–9529.62 mg/L, SO_4^{2-} : 1273.50–5724.38 mg/L, Na: 779.25–2023.03 mg/L) (Supplementary

Figs. 5–7). It was observed that the relative abundance and composition of CPR bacteria varied significantly along the three hydrochemical gradients (Fig. 3b and Supplementary Fig. 8, ANOSIM $r = 0.478$, $P = 0.001$), while a non-significant relation was observed between wet and dry seasons (ANOSIM $P > 0.05$).

The taxa responsible for the differentiation in CPR members among the three groups were further identified (Fig. 3c and Supplementary Fig. 9). In the low-salinity Group1, three CPR candidate phyla, including Taylorbacteria (relative abundance: 1.36% of the whole microbial communities), Niyogibacteria (0.64%), and CNCPR1 (0.21%) displayed significantly higher relative abundance than other sampling groups. In the moderate-salinity Group2, five phyla including Azambacteria (0.35%), Daviesbacteria (4.18%), and Gottesmanbacteria (0.29%) showed the highest proportion. In the highest-salinity Group3, CPR members associated with Komeilibacteria (1.56%), Uhrbacteria (1.11%), Roizmanbacteria (1.05%), Berkelbacteria (0.71%), Campbellbacteria (0.42%), Brennerbacteria (0.26%), CNCPR3 (0.25%), Gribaldobacteria (0.18%), Buchananbacteria (0.16%), and Katanobacteria (0.14%) were markedly predominant.

Driving forces for variations in relative abundances of CPR

We observed that the relative abundances of CPR bacteria may be more vulnerable to environmental changes than those of non-CPR bacteria in brackish-saline groundwater. Firstly, the relative abundances of CPR bacteria displayed stronger distance–decay relationships along environmental distance (Supplementary Fig. 10b, partial Mantel $r = 0.545$, $P = 0.001$) than geographic distance (Supplementary Fig. 10a, partial Mantel $r = 0.357$, $P = 0.001$), implying that environmental selection rather than dispersal limitation had a more significant impact on regulating CPR structure from the perspective of the relative abundance⁵². The relative abundances of non-CPR community, including Co-CPR bacteria and residual non-CPR members (Non-Co-CPR), appeared to be less affected by environmental distance compared to the CPR community (Supplementary Fig. 10c–f). The random forest analysis further revealed that environmental parameters, such as Co, Na, Cl^- , TDS, and SO_4^{2-} , explained about 72.1% of the variations in relative abundances of CPR bacteria, which was higher than those observed for Co-CPR (55.3%) and Non-Co-CPR bacteria (58.4%) (Fig. 4a–c). Secondly, the PLS-PM analysis confirmed that environmental factors had a more substantial impact on the relative abundances of CPR bacteria (Fig. 4d–f path coefficient $\lambda = 0.354$) than Co-CPR members ($\lambda = 0.291$). Thirdly, the relative abundances of CPR phyla predominantly distributed in high-salinity Group3 were positively associated with Na^+ , SO_4^{2-} , Co^{2+} , Ni^{2+} , and Cl^- (Supplementary Fig. 11), potentially reflecting significant niche differentiation of CPR due to the environmental heterogeneity in saline groundwater. Additionally, the PLS-PM analysis also indicated that the relative abundances of Co-CPR bacteria had a significant direct effect ($\lambda = 0.491$) on the relative abundances of the CPR population, highlighting their indispensable role in shaping the variations in relative abundances of CPR.

Metabolic potentials of CPR bacteria in groundwater

Compared with non-CPR bacteria (Supplementary Fig. 12 and Supplementary Table 7), the 399 CPR non-redundant MAGs exhibited limited metabolic capabilities in terms of biosynthesis, biogeochemical cycles, and energy metabolisms (Fig. 5, Supplementary Fig. 13, and pathway completeness in Supplementary Table 8). The most frequently observed fully complete metabolisms in CPR genomes included pyrimidine ribonucleotide biosynthesis (UMP to UDP/UTP/CDP/CTP without the de novo part: 54.89% of genomes), de novo purine biosynthesis (de novo to inosine monophosphate: 32.05%; inosine monophosphate to ATP/ADP: 32.58%) and oxidative or non-oxidative PPP (oxidative: 17.54%; non-oxidative: 20.80%). Glycolysis and gluconeogenesis were seldom fully present in these genomes (complete glycolysis/gluconeogenesis only in 0.75%/1.75% of genomes), but a near-complete glycolysis or gluconeogenesis pathway with more than 75% completeness was common (presenting in 44.61%/57.89% of genomes), suggesting that CPR bacteria might be involved in carbon scavenging and carbon degradation¹⁰. CPR bacteria also contained genes

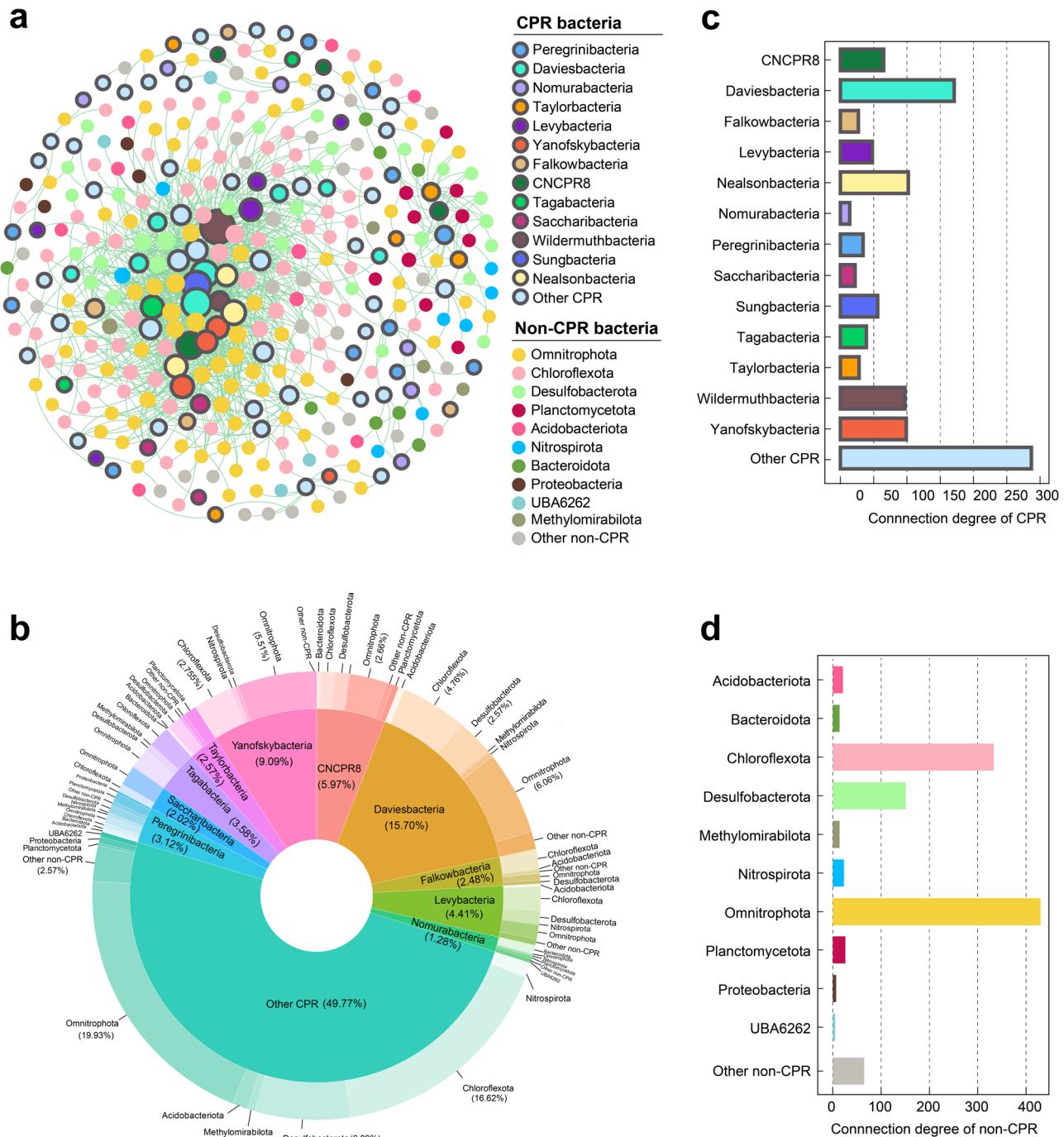


Fig. 2 | Co-occurrence pattern between CPR and non-CPR bacteria. **a** Networks displaying significant correlations (Spearman's $r > 0.8$, $P < 0.01$) between CPR and non-CPR bacteria. Each node represents one microbial species, and the node size is proportional to the number of connection degrees. **b** Connection relationships

between CPR and non-CPR bacteria. The inner and outer circles, respectively, indicate CPR and non-CPR bacteria. The bar charts showing the connection degree of CPR (c) and non-CPR bacteria (d).

encoding enzymes capable of degrading complex carbohydrates such as cellulose and starch, highlighting their contribution to carbon cycling in the study area⁹.

We identified some distinctive metabolic potentials of CPR bacteria in the brackish-saline groundwater (Fig. 5). A considerable proportion of CPR MAGs encoded genes associated with nitrite reduction (*nirK/nirS* in 23.13% CPR MAGs, Fig. 5), reflecting the potential involvement of CPR bacteria in the nitrogen cycle, although complete nitrogen reduction has not been observed in CPR members. Genes responsible for reducing sulfur compound, particularly converting sulfate to sulfite (*cysH/aprA* in 4.95% CPR MAGs), were also identified in specific CPR genomes, indicating the

probable involvement of CPR bacteria in the process of sulfur reduction within this sulfate-rich groundwater. Hydrogenases, especially [NiFe] group 3 hydrogenases, were moderately distributed in CPR bacteria (71.42%), possibly indicating the potential for producing or consuming hydrogen of these MAGs.

Besides the above metabolic potentials, stress-related genes were also identified in CPR genomes (Fig. 5). For example, heat shock protein genes (especially *hspX*, a member of the Hsp90 family) were found in CPR bacteria, which may act as molecular chaperones to facilitate proper protein folding and regulate protein homeostasis under high-salinity conditions⁵³. The presence of osmoprotectant genes encoding trehalose, inositol, and

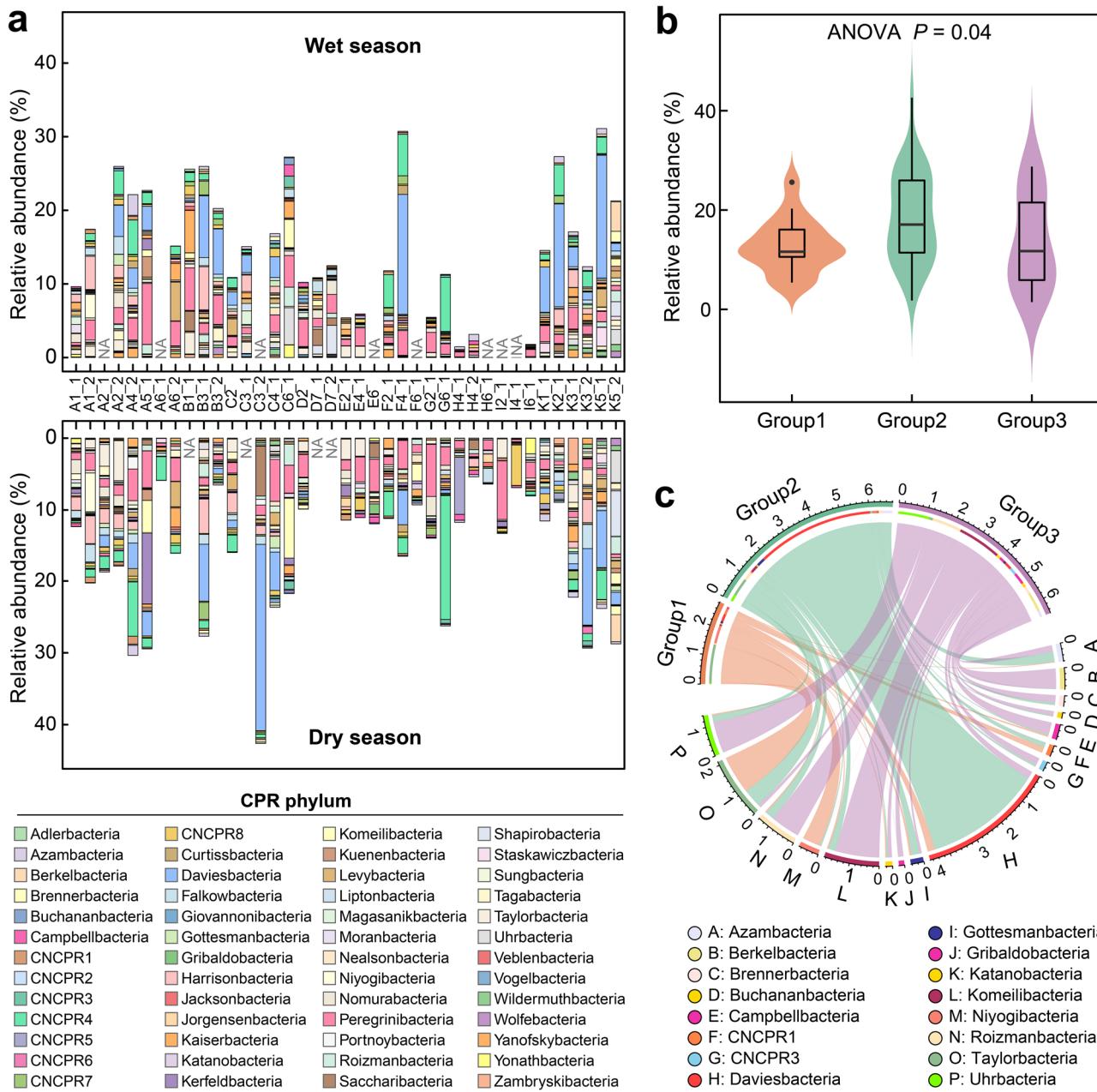


Fig. 3 | Spatiotemporal distribution of CPR in the groundwater. **a** Stacked bar chart of CPR phyla from all sampling stations in both wet and dry seasons. **b** Relative abundance of CPR members obtained from the three groundwater groups.

“Group1”, “Group2”, and “Group3” indicate the groundwater samples categorized into the lowest, moderate, and highest salinity levels, respectively. **c** Circular visualization of the CPR phyla differentially distributed among the three groups.

proline in some CPR genomes can provide these lineages with an advantage to adapt to the osmotic pressure caused by high-sulfate concentrations⁵⁴. We also identified an HTH-type transcriptional regulator gene (*plcR*) in several CPR genomes that may function as quorum sensing mechanisms of bacteria to control cell density, express virulence factors, or participate in host-symbiont communication⁵⁵. The presence of these genes in CPR suggests that they likely play crucial roles in the survival of CPR bacteria in brackish-saline groundwater. To validate this hypothesis, we conducted a comparative analysis of stress-related genes between CPR genomes from the present study and publicly available references related to lakes⁵⁶, acid mine drainage⁵⁷, and groundwater^{2,10,16,58} (Supplementary Table 9). Notably, genes associated with heat shock protein (*hspX*) and osmoprotectants (*ostA*, *tps*, *betB*, *gldA*, *proC*, *glnA*, and *ggsS*) were significantly more abundant in this high-salinity environment compared to other environments, indicating their crucial role in the remarkable adaptability of CPR bacteria to high-salinity groundwater.

Additional distinctive features of CPR bacteria were further explored. One of the most important traits was the presence of introns, which set CPR apart from other non-CPR bacteria². Among the 399 CPR MAGs, 73.4% were predicted to harbor 16S rRNA genes, and the absence of 16S rRNA in the rest, 26.6%, was possibly attributed to genome incompleteness. Of those containing 16S rRNA genes, 15.4% contained a total of 56 introns (Supplementary Table 10), indicating the widespread intron distribution of CPR bacteria. Many of these introns encode RNA-based self-splicing ribozymes (Group I catalytic introns) and/or amino-acid-based LAGLIDADG homing endonucleases, suggesting potential roles in gene expression regulation or ribosomal function. Another notable feature of CPR bacteria is the scarcity of CRISPR-Cas systems in their genomes, with only 1.4% of publicly available CPR MAGs reported to possess CRISPR arrays⁵⁹. Likewise, only three CPR MAGs in the ATWR were identified with genes encoding cas3, cas5, or cas7, indicating the absence of CPR CRISPR-Cas systems in this aquifer. Consistent with a previous study⁵⁷, we identified low gene occurrences and even

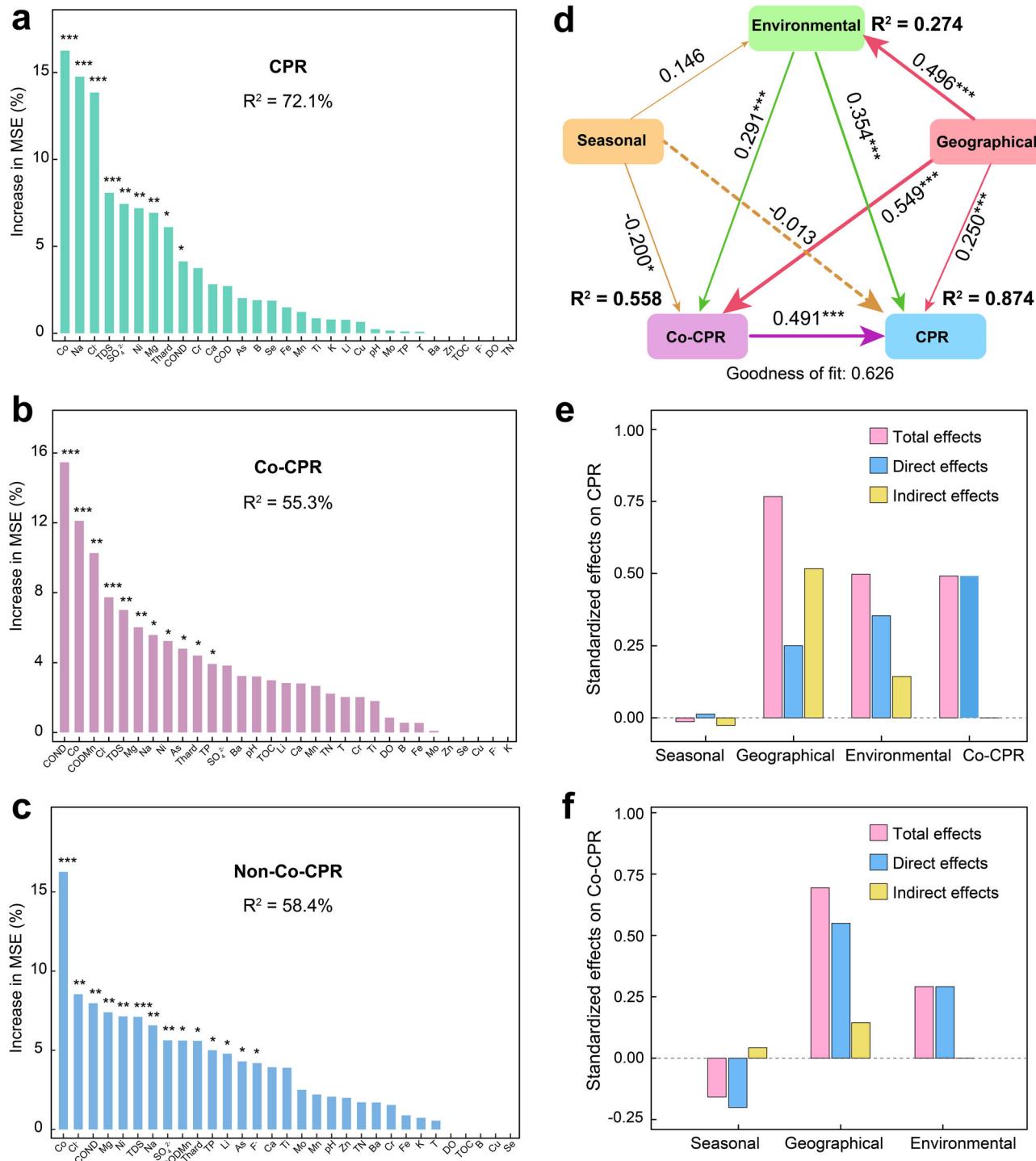


Fig. 4 | Driving forces of CPR and non-CPR bacteria. Random forest (RF) indicating the potential environmental drivers of variations in the CPR community (a), non-CPR bacteria that co-occurred with CPR (Co-CPR) (b), and non-CPR bacteria that not co-occurred with CPR (Non-Co-CPR) (c). PLS-PM describing the relationships between environmental factors, seasonal groups, geographical factors,

non-CPR bacteria that co-occurred with CPR and CPR (d). Numbers adjacent to each arrow denote partial correlation coefficients (significance codes: *** $p \leq 0.001$; ** $p \leq 0.01$; * $p \leq 0.05$). R^2 values display the proportion of variance explained for each factor. The bar-chart showing the standardized total effect of each factor on the CPR (e) and Co-CPR (f).

gene losses in CPR genomes, particularly in biosynthesis genes responsible for essential amino acids, fatty acids, and nucleotides (Supplementary Table 11). This finding underscores the significance of genomic streamlining in the evolution of CPR genomes, which may facilitate the minimization of cellular complexity and ultimately result in a decrease in genome size⁶⁰. Meanwhile, the absence of CPR CRISPR-Cas systems could also be the consequence of genomic streamlining for high metabolic efficiency. Interestingly, the lack of

CRISPR-Cas systems could position CPR bacteria as “viral decoys” for their hosts, thereby potentially protecting the hosts and reducing their viral load⁵⁹.

On the basis of the above functional prediction, we observed that many metabolic genes of CPR MAGs displayed the highest abundances in the high-salinity Group3, suggesting that these metabolic potentials may contribute to the environmental adaptation of CPR bacteria to high-salinity groundwater (Supplementary Table 12 and Fig. 6). The relevant functional

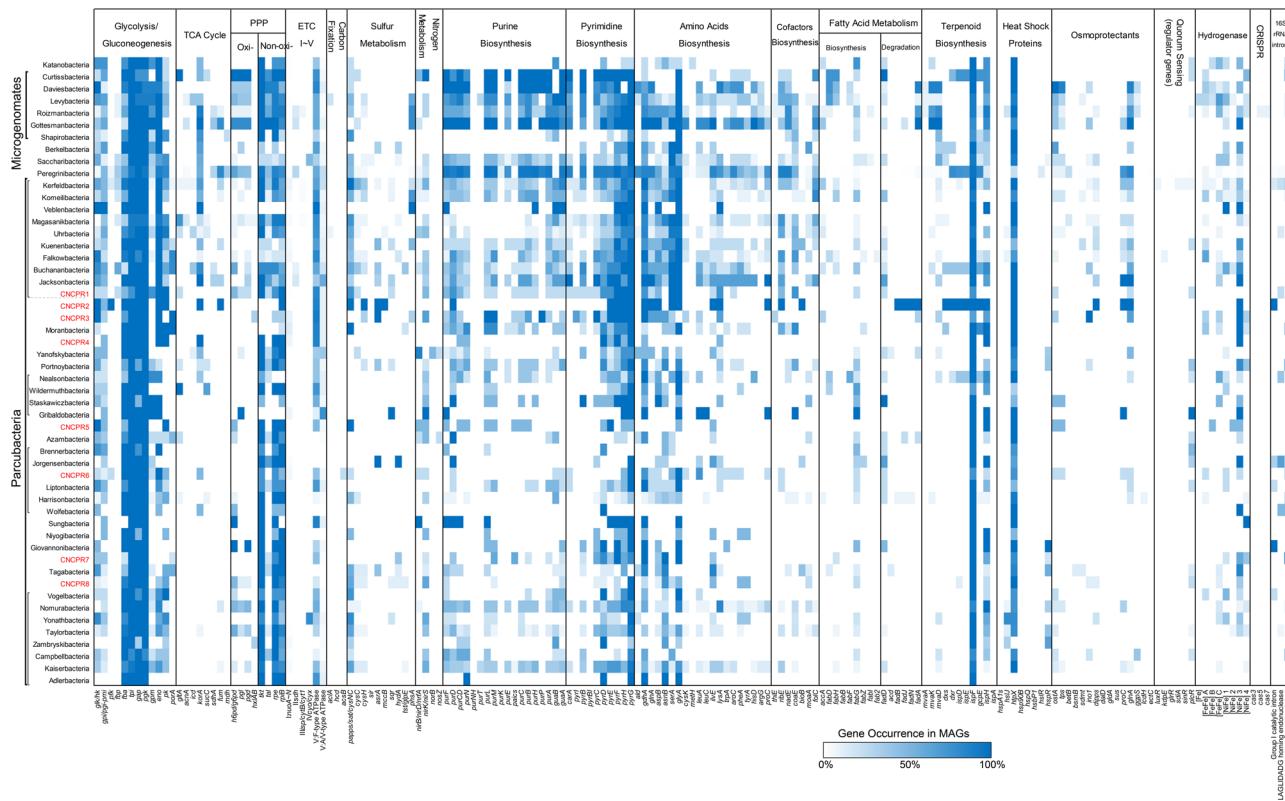


Fig. 5 | Metabolic heatmap inferred for CPR phyla. The occurrence of genes is calculated based on the number of MAGs regardless of their abundance. Acetyl-CoA acetyl-coenzyme A, a pivotal intermediate in multiple carbon pathways; TCA cycle tricarboxylic acid cycle; PPP pentose phosphate pathway; ETC electron transport chain.

genes were mainly related with amino acids biosynthesis (*serA*, *ilvE*, *leuA*, *cysK*, and *metH*), fatty acid metabolism (*fabGZ* and *fadBDN*), electron transport chain (*nuoFG* and *atpBCDFI*), glycolysis/gluconeogenesis (*gpi/pgi-pmi*, *fbp*, *glk/hk*, and *gap*), nitrogen metabolism (*nirK/nirS*, *nirB/nirD*, *nrfA*, and *nosZ*), and sulfur metabolism (*phsA* and *asrA*). Particularly, CPR genomes exhibited the highest gene abundances associated with sulfate reduction steps (the production or consumption of HS^- : $\text{SO}_3^{2-} \rightarrow \text{HS}^-$, $\text{S}_2\text{O}_3^{2-} \rightarrow \text{HS}^-$, and $\text{S}_{(n)} \rightarrow \text{HS}^-$) in the highest-salinity Group3 (0.74–2.90%) compared to the medium-salinity Group2 (0.28–1.51%) and the lowest-salinity Group1 (0.35–0.66%) (Supplementary Figs. 14 and 15), indicating an elevated contribution of CPR bacteria to sulfate reduction processes in high-salinity condition. This observation reflected a significant selective pressure for sulfate reduction on CPR exerted by the highest sulfate concentration in Group3 (Fig. 4 and S6). Pressure-regulating genes of CPR MAGs, including heat shock protein genes (*htpX*) and osmoprotectant transporter genes (*lcdH* and *sdmt*), also showed the highest relative abundances in Group3 than other groups (Supplementary Table 12 and Fig. 6).

Potential reciprocal relationships between CPR and Co-CPR

We observed that several metabolic genes of Co-CPR bacteria, such as *fbp*, *htpX*, *lcdH*, *pyrI*, and *asrA*, exhibited the lowest relative abundance in Group3, contrasting with their highest abundance in CPR bacteria within the same Group3 (Fig. 6). Conversely, certain genes of CPR MAGs showed the lowest abundance in Group3 but were markedly abundant in Co-CPR members. Examples included genes *nuoHN* (electron transport chain), *rpe* (pentose phosphate pathway), *guaB* (purine biosynthesis), *purn* (purine biosynthesis), *pyre* (purine biosynthesis), and *mdh* (TCA cycle). Given the symbiotic lifestyle of CPR bacteria, these findings potentially implied a potential metabolic synergy between CPR and Co-CPR bacteria, indicating the intricate interplay within the microbial community in salinity-stressed groundwater ecosystems.

Based on the crucial genes for the key steps of biogeochemical cycling (Supplementary Table 5), metabolic reconstructions of CPR and non-CPR

bacteria revealed the auxiliary contributions of CPR bacteria to carbon, nitrogen, and sulfur cycling (Fig. 7 and Supplementary Fig. 14). Compared with Non-Co-CPR bacteria, both CPR and Co-CPR played modest roles in biogeochemical cycling, but Co-CPR showed a higher abundance of genes in specific pathways. These included glycolysis [99.42% (Co-CPR) vs. 88.81% (Non-Co-CPR)], phosphofructokinase gene *pfk*, de novo lipogenesis (96.57% vs. 89.26%, S-malonyltransferase gene *fabD*), nitrite reduction (81.13% vs. 79.48%, both assimilatory and dissimilatory nitrite reductase *nirA*, *nirB* and *nrfA*), assimilatory sulfate/sulfite reduction (90.50%/89.46% vs. 79.37%/83.06%, phosphoadenosine phosphosulfate reductase and sulfite reductase genes *cysH*, *cysJ* and *sir*), and carbon fixation (81.34% vs. 51.15%, mostly through Wood-Ljungdahl pathway as indicated by the key acetyl-CoA synthase gene *acsB*) (Fig. 7, Supplementary Fig. 14, and Supplementary Table 7). These pathways might represent predicted hotspots of interactions between CPR and Co-CPR (Supplementary Fig. 16). Furthermore, Co-CPR bacteria also exhibited high functional potentials associated with the electron transport chain, pentose phosphate pathway, purine biosynthesis, and TCA cycle under high-salinity conditions, which could further help the co-occurred CPR to participate in the corresponding pathways. Additionally, the higher relative abundance of heat shock protein and osmoprotectant genes in Co-CPR compared to Non-Co-CPR may contribute to the adaptation of CPR through putative metabolites exchanges between CPR members and their co-existed partners.

Discussion

Drawing insights from metagenomic binning, our study restructured 399 CPR MAGs and expanded the CPR diversity by affiliating them with 44 previously proposed phyla and 8 new phyla in brackish-saline groundwater. We observed pronounced distribution patterns in the relative abundances of CPR bacteria along the environmental heterogeneity. Particularly, CPR bacteria not only played essential roles in regulating the stability and complexity of microbial community, but also exhibited high metabolic potentials associated with sulfur reduction, heat shock proteins,

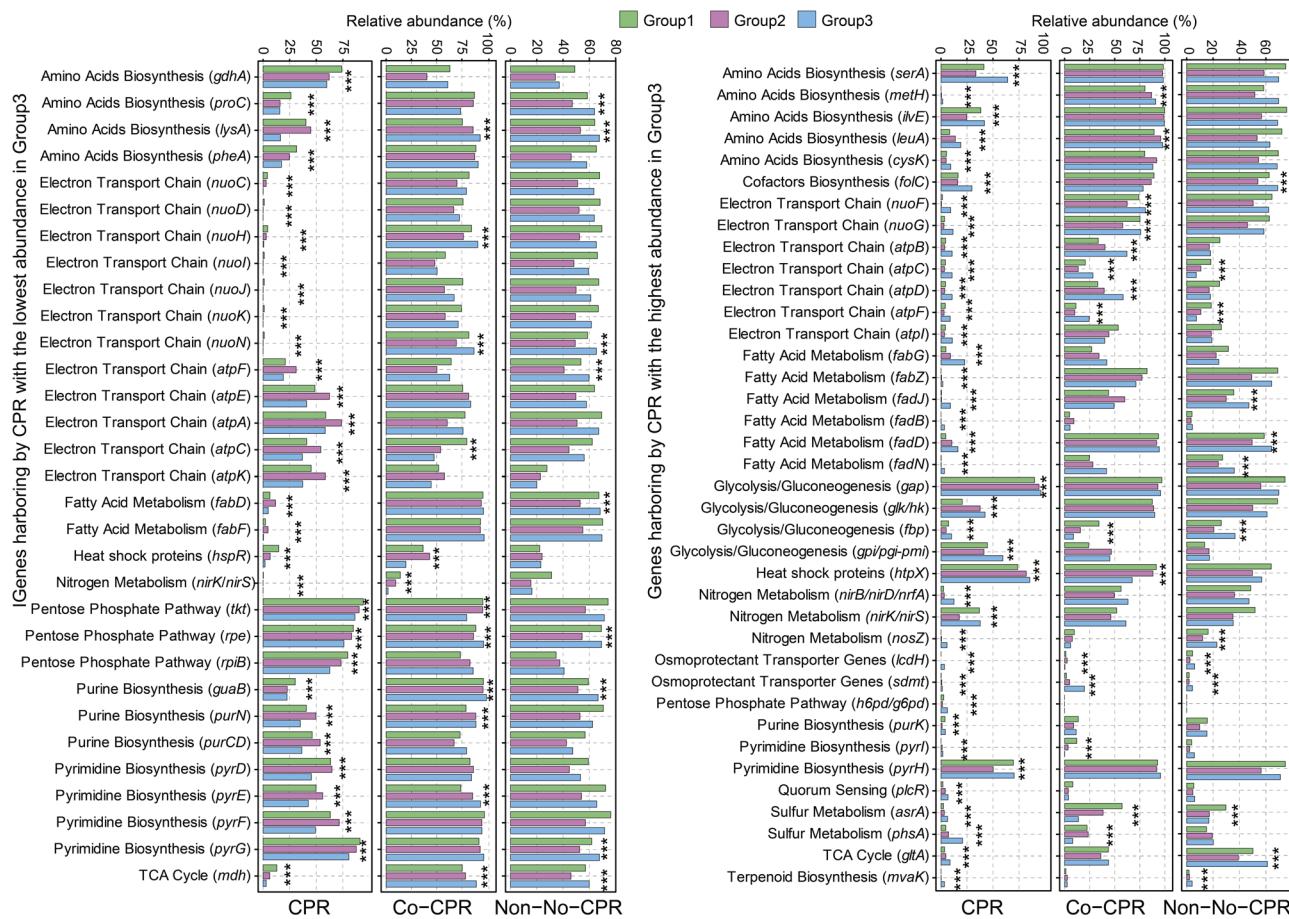


Fig. 6 | Genes harboring by CPR significantly showed the highest or lowest abundance in Group3 based on one-way ANOVA analysis. “Group1”, “Group2”, and “Group3” indicate the groundwater samples categorized into the lowest, moderate, and highest salinity levels, respectively.

osmoprotectants, and fatty acid metabolism, enabling them to effectively adapt to the high-salinity environmental stress.

Our findings revealed that the relative abundance of CPR members was susceptible to variations in groundwater hydrochemistry, particularly salinity characterized by Na^+ and SO_4^{2-} . Salinity emerged as a paramount driver, shaping the differentiation in relative abundances and ecological niches of CPR species^{61,62}. With increasing salinity levels, CPR bacteria experienced heightened selective pressure. Adaptive CPR members exhibited a higher relative abundance of pressure-regulating genes, such as osmoprotectant genes, enabling them to effectively regulate osmotic pressure under high-salinity conditions^{53,54,63}. Although the salinity levels in the studied brackish-saline groundwater were not as extreme as those inhabited by extreme halophilic Euryarchaeota and halotolerant ammonia-oxidizing archaea, insights from previous research also shed light on the role of osmoprotectants in coping with high salt concentrations^{64,65}. CPR bacteria also harbored abundant sulfate-reduction genes and fatty acid metabolism genes, facilitating energy and nutrient acquisition in sulfur-rich environments⁶⁶. Moreover, environmental heterogeneity not only directly influenced the structure of CPR members⁵ but also indirectly drove variations in the CPR community through Co-CPR bacteria (Fig. 4d–f), which might serve as potential hosts of CPR. Therefore, it could be concluded that the joint influence of environmental heterogeneity and Co-CPR bacteria drove the dynamics of CPR in groundwater, attributed to their limited metabolic capabilities and host-dependent lifestyles^{10,12,67–69}. However, it is essential to acknowledge that relative abundance in metagenomic analyses can be influenced by species dominance and sequencing depth. Thus, the dynamics of CPR bacteria across various environmental conditions should be further confirmed based on absolute abundance measurements in future studies.

CPR bacteria were found to be of great significance in maintaining the stability and complexity of the whole bacterial communities against external environmental perturbations (Supplementary Fig. 4). Furthermore, we observed that non-CPR bacteria belonging to Chloroflexota, Desulfovibrionia, and Planctomycetes could be potential hosts for CPR species from Daviesbacteria, Wildermuthbacteria, Taylorbacteria, and CNCPR8, from the perspective of co-occurrence network. Although the identified potential CPR partners need to be further verified by alternative approaches, such as cryogenic transmission electron microscopy imaging, culture-based method, and single-cell genomic analyses^{10,70}, these findings expanded the spectrum of potential partners for CPR members and provided significant insights into understanding complex symbiotic associations in underground ecosystems. Notably, the strongest links with CPR bacteria were observed in the phyla of *Omnitrophota*, *Chloroflexota*, and *Desulfovibrionia*. The latter two were highly related to sulfate reduction, suggesting the potential relationships between CPR and sulfating reduction bacteria in this sulfate-rich environment. These findings suggested that CPR bacteria might actively or passively select different reciprocal partners in response to changing environmental conditions, thereby enhancing their survival and the resilience of the whole bacterial community in stressful groundwater.

Furthermore, the epi-symbiotic lifestyle common among CPR members, facilitated by their attachment to host membranes via pili and the need to acquire essential biomolecules, suggests a potential mechanism for the joint adaptation of CPR and Co-CPR bacteria to high-salinity environments^{10,56,71,72}. We have developed a concept model to illustrate how CPR bacteria cope with environmental stressors (Fig. 8). On one hand, CPR bacteria exhibited a unique genetic makeup, encoding many fragmented genes across multiple incomplete pathways. They participated in crucial metabolic processes and regulated responses to salinity stress

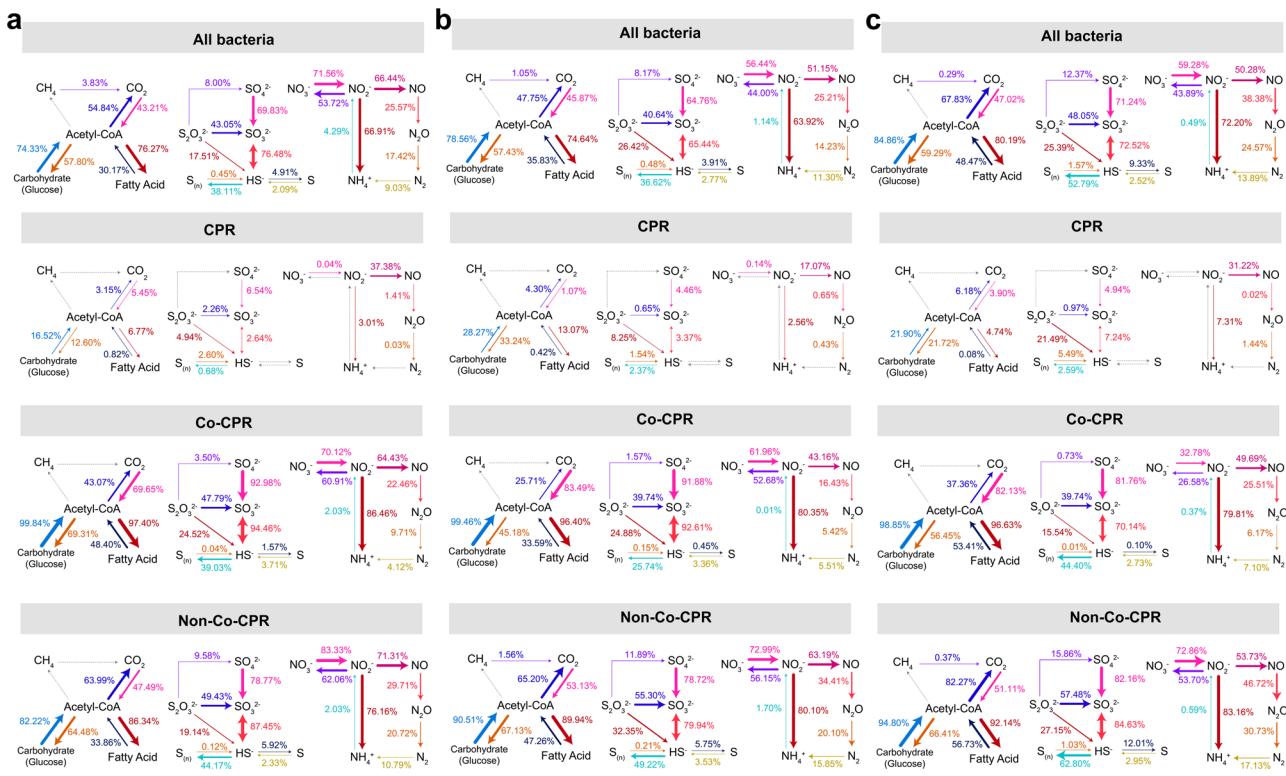


Fig. 7 | The abundance of genes related to sulfur, nitrogen, and carbon cycles for all bacteria, CPR, Co-CPR, and Non-Co-CPR. Detailed information on these genes in the lowest-salinity Group 1 (a), the moderate-salinity Group 2 (b), and the highest-salinity Group 3 (c). The percentage near each arrow indicates the MAG capacity for

catalyzing metabolic steps. The percentage was calculated by (the sequencing depth of genomes that possess the genes for certain steps/total sequencing depths of the selected genomes).

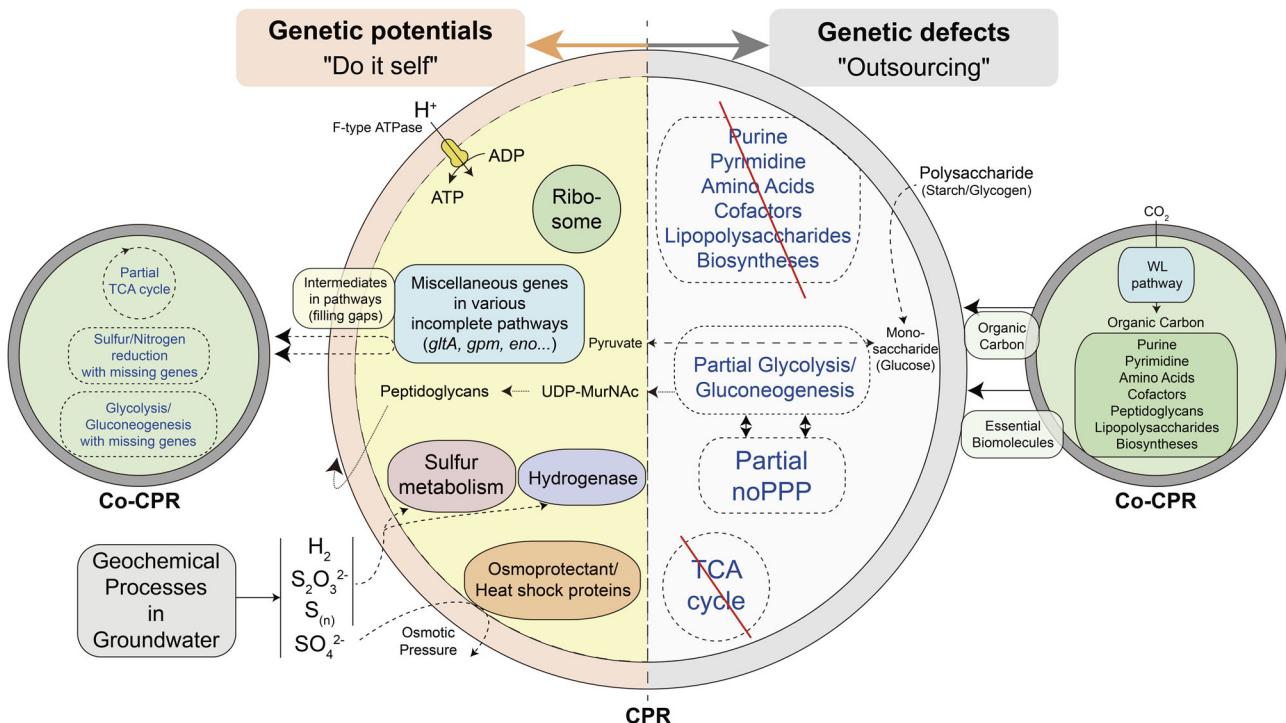


Fig. 8 | Proposed diagram of metabolic capacities and incapabilities in CPR, as well as their interactions with Co-CPR. Key intermediates might be exchanged between these organisms to complete the partial pathways. noPPP non-oxidative

pentose phosphate pathway, TCA cycle tricarboxylic acid cycle, WL pathway Wood-Ljungdahl pathway.

(“do itself”). In detail, the relative abundance of sulfate reduction genes in CPR genomes positively correlates with sulfate levels in groundwater, potentially promoting their adaptation to sulfate-rich environments. Higher relative abundances related to fatty acid metabolism may provide CPR with energy and nutrients to tolerate environmental stress⁶⁶. Genes encoding osmoprotectant or heat shock protein assist CPR bacteria in osmoregulation, maintaining positive cell turgor and growth groundwater with varying salinity levels^{53,54}. On the other hand, the limited biosynthetic potential of CPR bacteria suggests a high dependency on other organisms for resources (as shown in “outsourcing”)⁹. Co-CPR bacteria may serve as a promising source of essential biomolecules for CPR members since they predominantly encode pathways for autotrophic carbon fixation. These pathways could produce organic carbon and synthesize complex biomolecules that were lacking in CPR bacteria. In turn, CPR bacteria potentially facilitate Co-CPR by bypassing several missing steps in certain near-complete pathways, such as the partial TCA cycle (*gltA*, citrate synthase) or glycolysis/gluconeogenesis (*gpm*, phosphoglycerate mutase, 3-phosphoglycerate ↔ 2-phosphoglycerate) (Supplementary Fig. 16). This could be supported by the gene occurrence patterns observed for CPR and Co-CPR bacteria (Fig. 7, Supplementary Figs. 12 and 14). There are also chances for CPR horizontally exchanging genes with the putative hosts⁷² granting these CPR members to evolve more specialized niches with their hosts in groundwater once certain key genes are transferred. The abundant quorum sensing gene (*plcR*) in several CPR genomes could stimulate the metabolic interactions between CPR and relevant co-existing partners to adapt to the salinity stress⁵⁵.

Overall, potential collaborations between CPR bacteria and Co-CPR members might play essential roles in driving the transformations of primary elements such as carbon, nitrogen, and sulfur in complex groundwater ecosystems. Although it is difficult to determine whether the co-occurred partners of CPR are real hosts, these findings highlighted the combined contribution of CPR and their potential reciprocal partners to environmental adaptation, ecological stability, and biogeochemical cycling, substantially expanding our understanding of the ecological roles of CPR in groundwater^{20,73,74}.

Overall, we depicted the integrated biogeography patterns and metabolic adaptations of CPR microorganisms within brackish-saline groundwater ecosystems, by capturing 399 non-redundant CPR genomes spanning 44 previously proposed phyla and 8 potential novel phyla. Compared with non-CPR bacteria, environmental heterogeneity significantly influenced the spatial distribution and the niche differentiation of CPR memberships. CPR bacteria, as key components regulating the stability and complexity of microbial community, exhibited high relative abundances of functional genes associated with sulfate reduction, heat shock proteins, osmoprotectants, and fatty acid metabolism, potentially enabling them to cope with the high-salinity environmental stressors. More importantly, we disclosed the vital roles of co-occurred partners of CPR bacteria, which not only dominated key metabolic pathways such as the TCA cycle, gluconeogenesis/glycolysis process, sulfur metabolism, and nitrogen metabolism but also potentially engaged in metabolic interactions with CPR bacteria, facilitating their adaptation to high-salinity conditions. Together, this study elucidated the remarkable adaptation of environment-sensitive CPR to high-salinity conditions and highlighted the significance of close cooperation between CPR and their co-existing potential partners in mediating ecological stability and driving biogeochemical cycles within dynamic groundwater environments.

Data availability

Complete datasets supporting the findings of this article are available in the NCBI BioProject (PRJNA763045). All CPR and non-CPR MAGs sequences in this present article are available in <https://doi.org/10.17632/wb6hncs2b7.1>.

Received: 25 October 2023; Accepted: 19 November 2024;
Published online: 30 November 2024

References

- Rinke, C. et al. Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437 (2013).
- Brown, C. T. et al. Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**, 208–211 (2015).
- He, X. et al. Cultivation of a human-associated TM7 phylotype reveals a reduced genome and epibiotic parasitic lifestyle. *Proc. Natl Acad. Sci. USA* **112**, 244–249 (2015).
- Parks, D. H. et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**, 1533–1542 (2017).
- Tian, R. et al. Small and mighty: adaptation of superphylum Patescibacteria to groundwater environment drives their genome simplicity. *Microbiome* **8**, 51 (2020).
- Hug, L. A. et al. A new view of the tree of life. *Nat. Microbiol.* **1**, 16048 (2016).
- Chaumeil, P. A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database. *Bioinformatics* **36**, 1925–1927 (2020).
- Castelle, C. J. & Banfield, J. F. Major new microbial groups expand diversity and alter our understanding of the tree of life. *Cell* **172**, 1181–1197 (2018).
- Castelle, C. J. et al. Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. *Nat. Rev. Microbiol.* **16**, 629–645 (2018).
- He, C. et al. Genome-resolved metagenomics reveals site-specific diversity of episymbiotic CPR bacteria and DPANN archaea in groundwater ecosystems. *Nat. Microbiol.* **6**, 354–365 (2021).
- Danczak, R. E. et al. Members of the candidate phyla radiation are functionally differentiated by carbon- and nitrogen-cycling capabilities. *Microbiome* **5**, 112 (2017).
- Vavourakis, C. D. et al. A metagenomics roadmap to the uncultured genome diversity in hypersaline soda lake sediments. *Microbiome* **6**, 168 (2018).
- Méheust, R., Burstein, D., Castelle, C. J. & Banfield, J. F. The distinction of CPR bacteria from other bacteria based on protein family content. *Nat. Commun.* **10**, 4173 (2019).
- Martiny, J. B. H. et al. Microbial biogeography: putting microorganisms on the map. *Nat. Rev. Microbiol.* **4**, 102–112 (2006).
- Chaudhari, N. M. et al. The economical lifestyle of CPR bacteria in groundwater allows little preference for environmental drivers. *Environ. Microbiome* **16**, 1–18 (2021).
- Anantharaman, K. et al. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat. Commun.* **7**, 1–11 (2016).
- Vigneron, A. et al. Ultra-small and abundant: candidate phyla radiation bacteria are potential catalysts of carbon transformation in a thermokarst lake ecosystem. *Limnol. Oceanogr. Lett.* **5**, 212–220 (2020).
- Luef, B. et al. Diverse uncultivated ultra-small bacterial cells in groundwater. *Nat. Commun.* **6**, 6372 (2015).
- Wrighton, K. C. et al. Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science* **337**, 1661–1665 (2012).
- Wrighton, K. C. et al. Metabolic interdependencies between phylogenetically novel fermenters and respiratory organisms in an unconfined aquifer. *ISME J.* **8**, 1452–1463 (2014).
- Nelson, W. C. & Stegen, J. C. The reduced genomes of Parcubacteria (OD1) contain signatures of a symbiotic lifestyle. *Front. Microbiol.* **6**, 713 (2015).
- Zhang, X., Qi, L., Li, W., Hu, B. X. & Dai, Z. Bacterial community variations with salinity in the saltwater-intruded estuarine aquifer. *Sci. Total Environ.* **755**, 142423 (2021).
- Liu, S. et al. Different spatiotemporal dynamics, ecological drivers and assembly processes of bacterial, archaeal and fungal communities in brackish-saline groundwater. *Water Res.* **214**, 118193 (2022).

24. Sang, S. et al. Diversity and predictive metabolic pathways of the prokaryotic microbial community along a groundwater salinity gradient of the Pearl River Delta, China. *Sci. Rep.* **8**, 1–11 (2018).

25. Li, J., Chen, Q., Wang, T., Wang, H. & Ni, J. Hydrochemistry and nutrients determined the distribution of greenhouse gases in saline groundwater. *Environ. Pollut.* **286**, 117383 (2021).

26. Wang, J. et al. Impacts of municipal wastewater treatment plant discharge on microbial community structure and function of the receiving river in Northwest Tibetan Plateau. *J. Hazard. Mater.* **423**, 127170 (2022).

27. Wang, J., Liu, T., Sun, W. & Chen, Q. Bioavailable metal(loid)s and physicochemical features co-mediating microbial communities at combined metal(loid) pollution sites. *Chemosphere* **260**, 127619 (2020).

28. Chen, S., Zhou, Y., Chen, Y. & Gu, J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).

29. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).

30. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).

31. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).

32. Chklovski, A., Parks, D. H., Woodcroft, B. J. & Tyson, G. W. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat. Methods* **20**, 1203–1212 (2023).

33. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. DRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).

34. Kim, M., Oh, H. S., Park, S. C. & Chun, J. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *Int. J. Syst. Evol. Microbiol.* **64**, 346–351 (2014).

35. Auch, A. F., von Jan, M., Klenk, H. P. & Göker, M. Digital DNA–DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. *Stand. Genom. Sci.* **2**, 117–134 (2010).

36. Meier-Kolthoff, J. P., Carbasse, J. S., Peinado-Olarte, R. L. & Göker, M. TYGS and LPSN: a database tandem for fast and reliable genome-based classification and nomenclature of prokaryotes. *Nucleic Acids Res.* **50**, D801–D807 (2022).

37. Johnson, L. S., Eddy, S. R. & Portugaly, E. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinform.* **11**, 431 (2010).

38. Katoh, K., Misawa, K., Kuma, K. I. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).

39. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

40. Hyatt, D. et al. Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* **11**, 119 (2010).

41. Huerta-Cepas, J. et al. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).

42. Marchler-Bauer, A. et al. CDD: NCBI's conserved domain database. *Nucleic Acids Res.* **43**, D222–D226 (2015).

43. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).

44. Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. & Eddy, S. R. Rfam: An RNA family database. *Nucleic Acids Res.* **31**, 439–441 (2003).

45. Oksanen, J. et al. vegan: Community Ecology Package. *R Package Version 2.4-0* <https://CRAN.R-project.org/package=vegan> (2016).

46. Bastian, M., Heymann, S. & Jacomy, M. Gephi: an open source software for exploring and manipulating networks. *ICWSM* **8**, 361–362 (2009).

47. Parks, D. H. et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996 (2018).

48. Coleman, G. A. et al. A rooted phylogeny resolves early bacterial evolution. *Science* **372**, eabe0511 (2021).

49. Giovannoni, S. J. et al. Genetics: genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309**, 1242–1245 (2005).

50. Wang, J. et al. Insights into the responses of fungal taxonomy and function to different metal(loid) contamination levels. *Sci. Total Environ.* **877**, 162931 (2023).

51. Yang, S. et al. Multiple metal(loid) contamination reshaped the structure and function of soil archaeal community. *J. Hazard. Mater.* **436**, 129186 (2022).

52. Wang, X. et al. Habitat-specific patterns and drivers of bacterial β -diversity in China's drylands. *ISME J.* **11**, 1345–1358 (2017).

53. Sakoh, M., Ito, K. & Akiyama, Y. Proteolytic activity of HtpX, a membrane-bound and stress-controlled protease from *Escherichia coli*. *J. Biol. Chem.* **280**, 33305–33310 (2005).

54. Sleator, R. D. & Hill, C. Bacterial osmoadaptation: the role of osmolytes in bacterial stress and virulence. *FEMS Microbiol. Rev.* **26**, 49–71 (2002).

55. Declerck, N. et al. Structure of PlcR: insights into virulence regulation and evolution of quorum sensing in Gram-positive bacteria. *Proc. Natl. Acad. Sci. USA* **104**, 18490–18495 (2007).

56. Chiriac, M. C. et al. Ecogenomics sheds light on diverse lifestyle strategies in freshwater CPR. *Microbiome* **10**, 1–21 (2022).

57. Lemos, L. N. et al. Genomic signatures and co-occurrence patterns of the ultra-small Saccharimonadia (phylum CPR/Patescibacteria) suggest a symbiotic lifestyle. *Mol. Ecol.* **28**, 4259–4271 (2019).

58. Probst, A. J. et al. Differential depth distribution of microbial function and putative symbionts through sediment-hosted aquifers in the deep terrestrial subsurface. *Nat. Microbiol.* **3**, 328–336 (2018).

59. Burstein, D. et al. Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. *Nat. Commun.* **7**, 10613 (2016).

60. Hessen, D. O., Jeyasingh, P. D., Neiman, M. & Weider, L. J. Genome streamlining and the elemental costs of growth. *Trends Ecol. Evol.* **25**, 75–80 (2010).

61. Zhou, J. et al. Stochasticity, succession, and environmental perturbations in a fluidic ecosystem. *Proc. Natl. Acad. Sci. USA* **111**, E836–E845 (2014).

62. Hanson, C. A., Fuhrman, J. A., Horner-Devine, M. C. & Martiny, J. B. H. Beyond biogeographic patterns: processes shaping the microbial landscape. *Nat. Rev. Microbiol.* **10**, 497–506 (2012).

63. Gagnon, J., Beauregard-tousignant, S., Marcil, J. & Lazar, C. S. Deep isolated aquifer brines harbor atypical halophilic microbial communities in Quebec, Canada. *Genes (Basel)* **14**, 1529 (2023).

64. Becker, E. A. et al. Phylogenetically driven sequencing of extremely halophilic archaea reveals strategies for static and dynamic osmo-response. *PLoS Genet.* **10**, e1004784 (2014).

65. Ngugi, D. K. et al. Comparative genomics reveals adaptations of a halotolerant thaumarchaeon in the interfaces of brine pools in the Red Sea. *ISME J.* **9**, 396–411 (2015).

66. Couvillion, S. P. et al. Rapid remodeling of the soil lipidome in response to a drying–rewetting event. *Microbiome* **11**, 34 (2023).

67. Nicolas, A. M. et al. Soil candidate phyla radiation bacteria encode components of aerobic metabolism and co-occur with nanoarchaea in the rare biosphere of rhizosphere grassland communities. *mSystems* **6**, e01205–e01220 (2021).

68. Wiegand, S., Dam, H. T., Riba, J., Vollmers, J. & Kaster, A. K. Printing microbial dark matter: using single cell dispensing and genomics to

investigate the patescibacteria/candidate phyla radiation. *Front. Microbiol.* **12**, 635506 (2021).

69. Lynch, M. D. J. & Neufeld, J. D. Ecology and exploration of the rare biosphere. *Nat. Rev. Microbiol.* **13**, 217–229 (2015).

70. Zhang, G. et al. Metagenomic and network analyses decipher profiles and co-occurrence patterns of antibiotic resistome and bacterial taxa in the reclaimed wastewater distribution system. *J. Hazard. Mater.* **400**, 123170 (2020).

71. Moreira, D., Zivanovic, Y., López-Archilla, A. I., Iniesto, M. & López-García, P. Reductive evolution and unique predatory mode in the CPR bacterium *Vampirococcus lugosii*. *Nat. Commun.* **12**, 2454 (2021).

72. Wang, Y. et al. Genome-centric metagenomics reveals the host-driven dynamics and ecological role of CPR bacteria in an activated sludge system. *Microbiome* **11**, 56 (2023).

73. Castelle, C. J., Brown, C. T., Thomas, B. C., Williams, K. H. & Banfield, J. F. Unusual respiratory capacity and nitrogen metabolism in a Parcubacterium (OD1) of the candidate phyla radiation. *Sci. Rep.* **7**, 40101 (2017).

74. León-Zayas, R. et al. The metabolic potential of the single cell genomes obtained from the Challenger Deep, Mariana Trench within the candidate superphylum Parcubacteria (OD1). *Environ. Microbiol.* **19**, 2769–2784 (2017).

Acknowledgements

This work was funded by the National Natural Science Foundation of China (Nos. U2240205, 52070002, 51925901, and 52209078) and the China Postdoctoral Science Foundation (2023T160020).

Author contributions

J.W.W. and H.H.Z. conceived the study, conducted data analysis, and wrote the manuscript. J.R.N. contributed new ideas and information. Q.C. supervised the whole process. All authors have read and approved the final manuscript for publication.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41522-024-00615-4>.

Correspondence and requests for materials should be addressed to Qian Chen.

Reprints and permissions information is available at <http://www.nature.com/reprints>

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

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024