Supplementary Information

A comprehensive genomic resource for biosynthetic discovery in saline lakes

Yu Wang[1,2,15], Runzhi Cao[1,15], Xuejiao Qiao[1,15], Yujun Wu[1], Zhiguang Qiu[1], Chun-Ang Lian[1], Yan Yan[3], Fan Wang[4], Bing Li[5], Baowei Chen[6], Yanping Mao[7], Liguan Li[8], Lei Gao[9], Jianyu Jiao[9], Wenjun Li[9], Shengwei Hou[10], Liang Yang[11], Meng Li[12], Nan Xu[1], Huapeng Qin[1], Jinren Ni[1], Tong Zhang[13],Wei-Qin Zhuang[14], Ke Yu[1*]

1 Eco-environment and Resource Efficiency Research Laboratory, School of Environment and Energy, Shenzhen Graduate School, Peking University, Shenzhen, PR China

2 College of Life Sciences, Inner Mongolia University, Hohhot, PR China

3 State Key Laboratory of Deep Earth Processes and Resources, Guangzhou Institute of Geochemistry, Chinese Academy of Sciences, Guangzhou, PR China

4 School of Ecology, Sun Yat-sen University, Shenzhen 518000, PR China

5 Shenzhen International Graduate School, Tsinghua University, Shenzhen, PR China.

6 Guangdong Provincial Key Laboratory of Marine Resources and Coastal Engineering, School of Marine Sciences, Sun Yat-sen University, Zhuhai, PR China.

7 College of Chemistry and Environmental Engineering, Shenzhen University, Shenzhen, PR China

8 Department of Science and Environmental Studies, The Education University of Hong Kong, Taipo, Hong Kong Special Administrative Region of China, PR China

9 State Key Laboratory of Biocontrol, Guangdong Provincial Key Laboratory of Plant Stress Biology and Innovation Center for Evolutionary Synthetic Biology, School of Life Sciences, Sun Yat-Sen University, Guangzhou, PR China

10 Shenzhen Key Laboratory of Marine Archaea Geo-Omics, Department of Ocean Science and Department of Ocean Science & Engineering, Southern University of Science and Technology, Shenzhen, PR China

11School of Brewing Engineering, Moutai Institute, Renhuai, PR China

12 Archaeal Biology Centre, Synthetic Biology Research Center, Shenzhen Key Laboratory of Marine Microbiome Engineering, Key Laboratory of Marine Microbiome Engineering of Guangdong Higher Education Institutes, Institute for Advanced Study, Shenzhen University, Shenzhen, PR China

13 Department of Civil Engineering, University of Hong Kong, Hong Kong, PR China

14 Department of Civil and Environmental Engineering, Faculty of Engineering, University of Auckland, Auckland, New Zealand

15 These authors contributed equally: Yu Wang, Runzhi Cao, Xuejiao Qiao.

* Correspondence should be addressed to Ke Yu (email: yuke.sz@pku.edu.cn).

Supplementary Figures

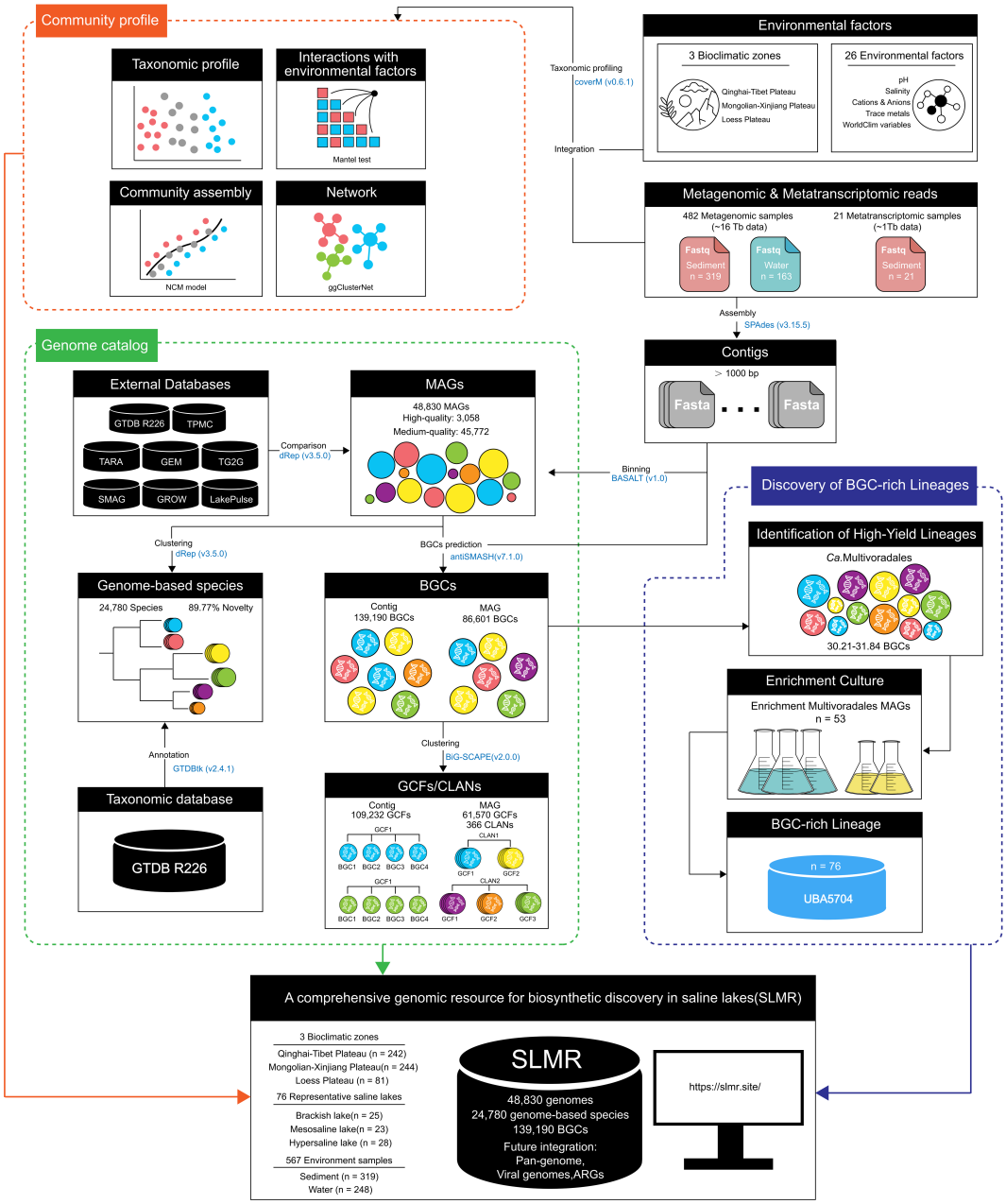


Fig. S1 | Overview of the construction and analysis framework of the Saline Lake Microbiome Resource. Quality-controlled, high-throughput metagenomic sequencing reads from saline lake samples across three bioclimatic regions are used for: 1) taxonomic profiling of microbial communities and association analyses with environmental factors; 2) assembly into metagenomic contigs for the construction of a metagenome-assembled genome catalog, followed by taxonomic annotation and comparison with external genome databases; 3) prediction of BGCs and clustering into

GCFs and GCCs; and 4) identification of BGC-rich lineages, enrichment of representative Candidatus *Multivoradales* MAGs and proposal of the candidate family UBA5704. BGC, biosynthetic gene cluster; GCF, gene cluster family; GLAN, gene cluster clan; MAG, metagenome-assembled genome.

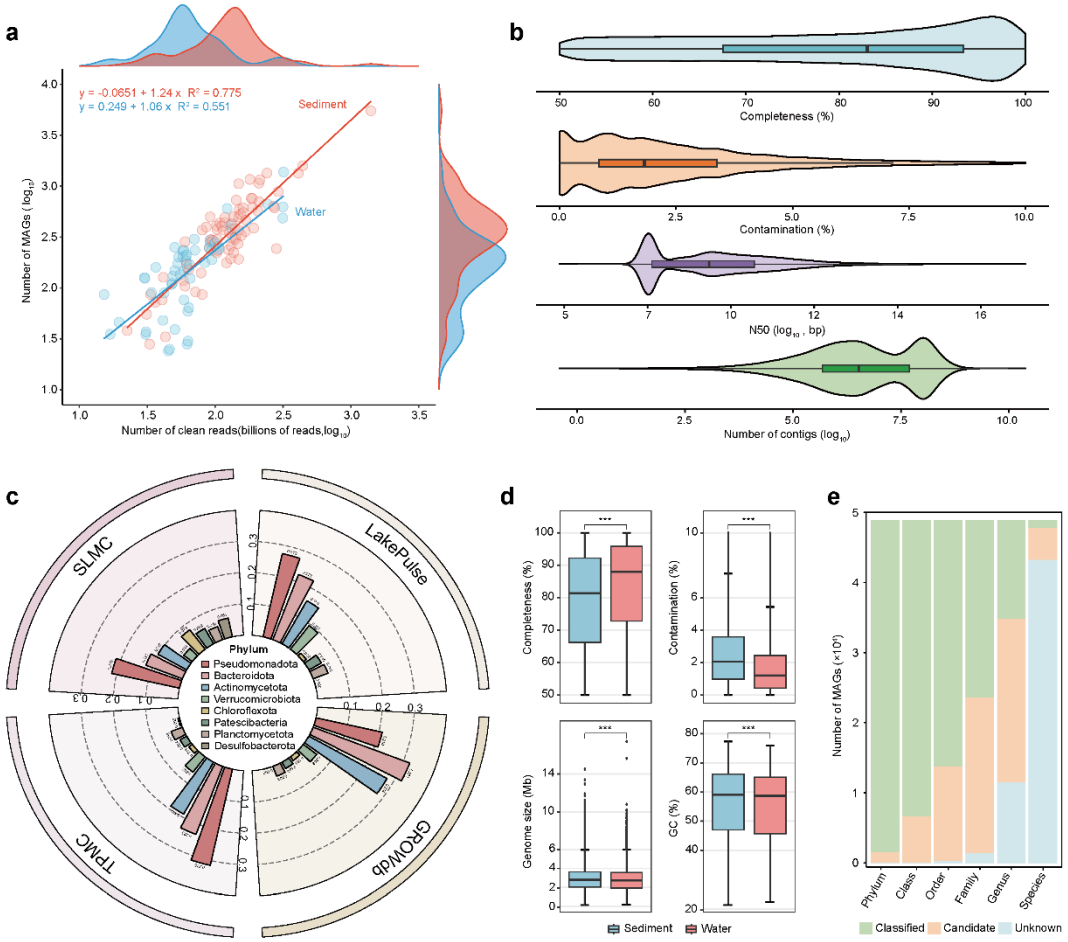


Fig. S2 | Sequencing depth, genome quality and taxonomic annotation of SLMR MAGs.

a, Relationship between sequencing depth and the number of reconstructed MAGs in sediment and water metagenomes. Marginal density plots show the distributions of sequencing depth and MAG recovery in sediment and water samples. Linear regression lines are shown separately for sediment and water samples. **b,** Quality profiles of reconstructed MAGs, including completeness, contamination, N50 and the number of contigs. **c,** Comparison of the dominant phylum-level composition of SLMR MAGs with other aquatic microbiome genome catalogs, including LakePulse, GROWdb and TPMC. **d,** Comparison of genomic features between sediment- and water-derived MAGs, including completeness, contamination, genome size and GC content. Statistical significance of pairwise differences was calculated using two-sided Wilcoxon rank-sum tests with false discovery rate correction for multiple comparisons (*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$). **e,** Taxonomic annotation status of SLMR MAGs

across different taxonomic ranks. Stacked bars show the numbers of MAGs classified as known, candidate or unknown at the phylum, class, order, family, genus and species levels. MAG, metagenome-assembled genome.



Fig. S3 | Differential abundance patterns of selected BGC classes across habitats, salinity gradients and bioclimatic regions.

a, Comparison of selected BGC classes between sediment and water samples. **b,** Abundance patterns of selected BGC classes among brackish, mesosaline and hypersaline lakes within each bioclimatic region and habitat. Samples were grouped by the LP, MXP and QTP regions, and analysed separately for sediment and water samples. **c,** Abundance patterns of selected BGC classes among bioclimatic regions within each salinity category and habitat. Samples were grouped by brackish, mesosaline and hypersaline lakes, and analysed separately for sediment and water samples. Statistical significance of pairwise differences was assessed using two-sided Wilcoxon rank-sum tests with FDR correction for multiple comparisons (*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$). BGC, biosynthetic gene cluster; NRPS, non-ribosomal peptide synthetase; PKS, polyketide synthase; RiPP, ribosomally synthesized and post-translationally modified peptide.

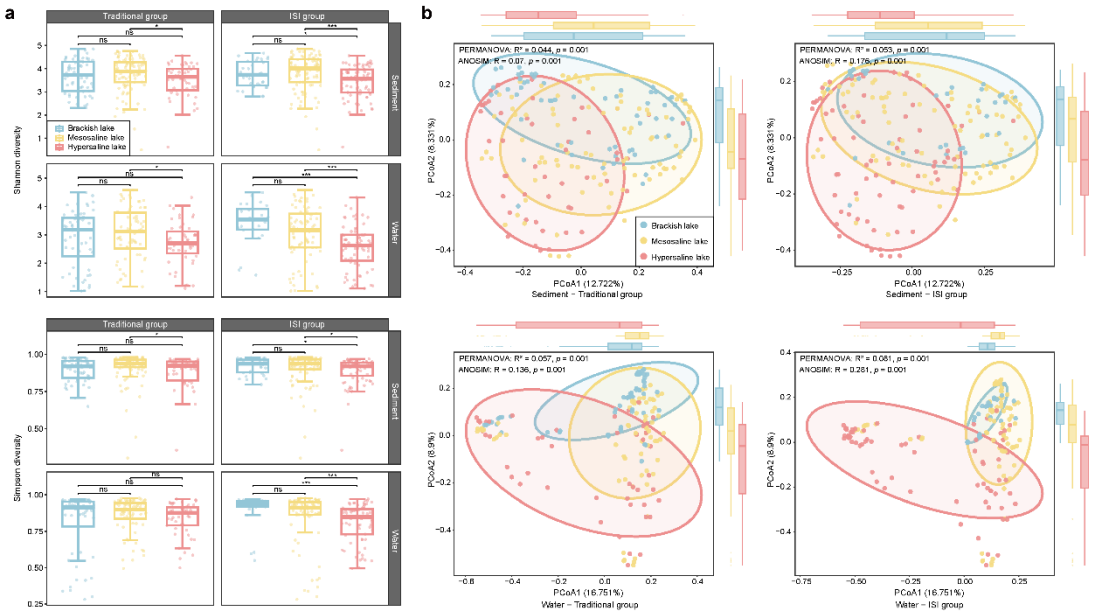


Fig. S4 | Comparison of traditional and ISI-based salinity grouping for microbial diversity analyses.

a, Shannon and Simpson diversity of microbial profiles in sediment and water samples under traditional salinity grouping and ISI-based grouping. Samples were classified into brackish, mesosaline and hypersaline lake. **b,** PCoA visualizing microbial community variation under traditional and ISI-based salinity groupings, based on Bray–Curtis dissimilarities. Statistical significance of compositional differences among salinity categories was assessed by PERMANOVA and ANOSIM (*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$; R and P values shown). ISI, integrated salinity index; PCoA, principal coordinate analysis; PERMANOVA, permutational multivariate analysis of variance; ANOSIM, analysis of similarities.

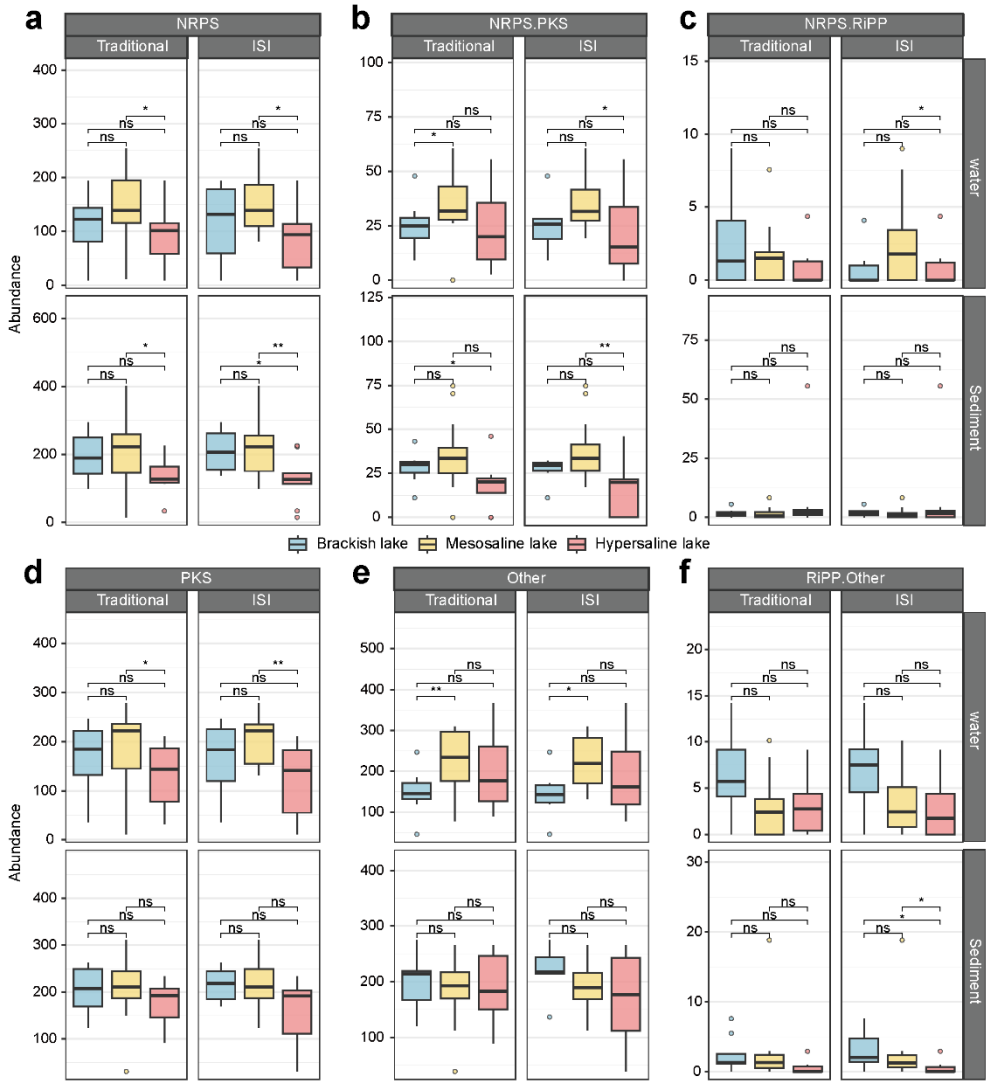


Fig. S5 | Comparison of BGC abundance patterns between traditional and ISI-based salinity grouping.

a–f, Abundance patterns of representative BGC classes under traditional salinity grouping and ISI-based grouping, including NRPS (**a**), NRPS.PKS (**b**), NRPS.RiPP (**c**), PKS (**d**), Other (**e**) and RiPP.Other (**f**) clusters. Samples were classified into brackish, mesosaline and hypersaline lakes and analysed separately for water and sediment habitats. Statistical significance of pairwise differences was assessed using two-sided Wilcoxon rank-sum tests with FDR correction for multiple comparisons (*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$). BGC, biosynthetic gene cluster; ISI, integrated salinity index; NRPS, non-ribosomal peptide synthetase; PKS, polyketide synthase; RiPP, ribosomally synthesized and post-translationally modified peptide.

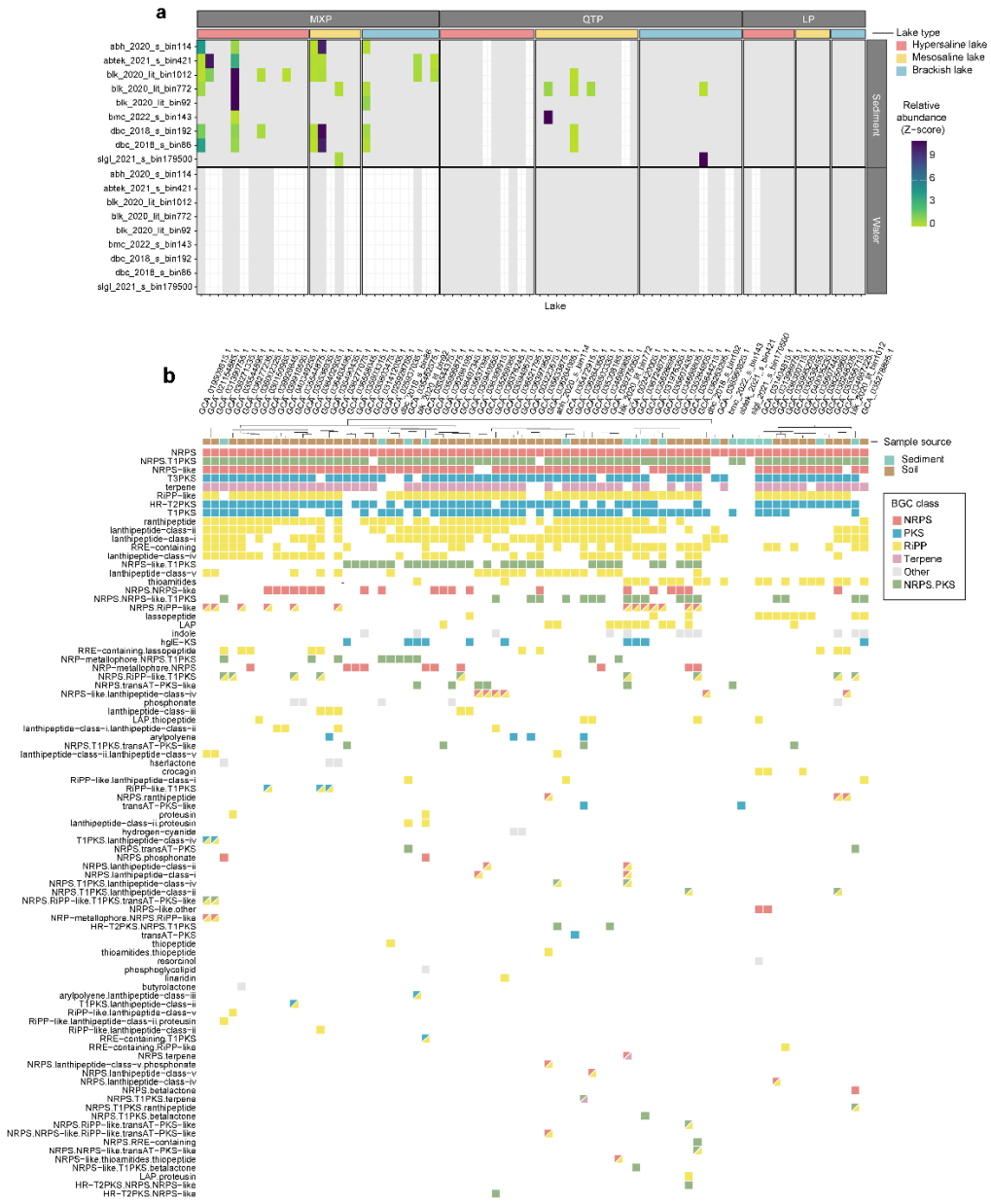


Fig. S6 | Biosynthetic profiles and abundance patterns of UBA5704 genomes.

a, Lake-level relative abundance distribution of UBA5704 genomes across saline lake samples. Heatmap colours indicate Z-score-transformed relative abundance. **b,** BGC class composition of an expanded UBA5704 genome set, including genomes from SLMR, GTDB and NCBI. BGC, biosynthetic gene cluster; MAG, metagenome-assembled genome; NRPS, non-ribosomal peptide synthetase; PKS, polyketide synthase; RiPP, ribosomally synthesized and post-translationally modified peptide; SLMR, saline lake microbiome resource; GTDB, genome taxonomy database; NCBI, national center for biotechnology information.

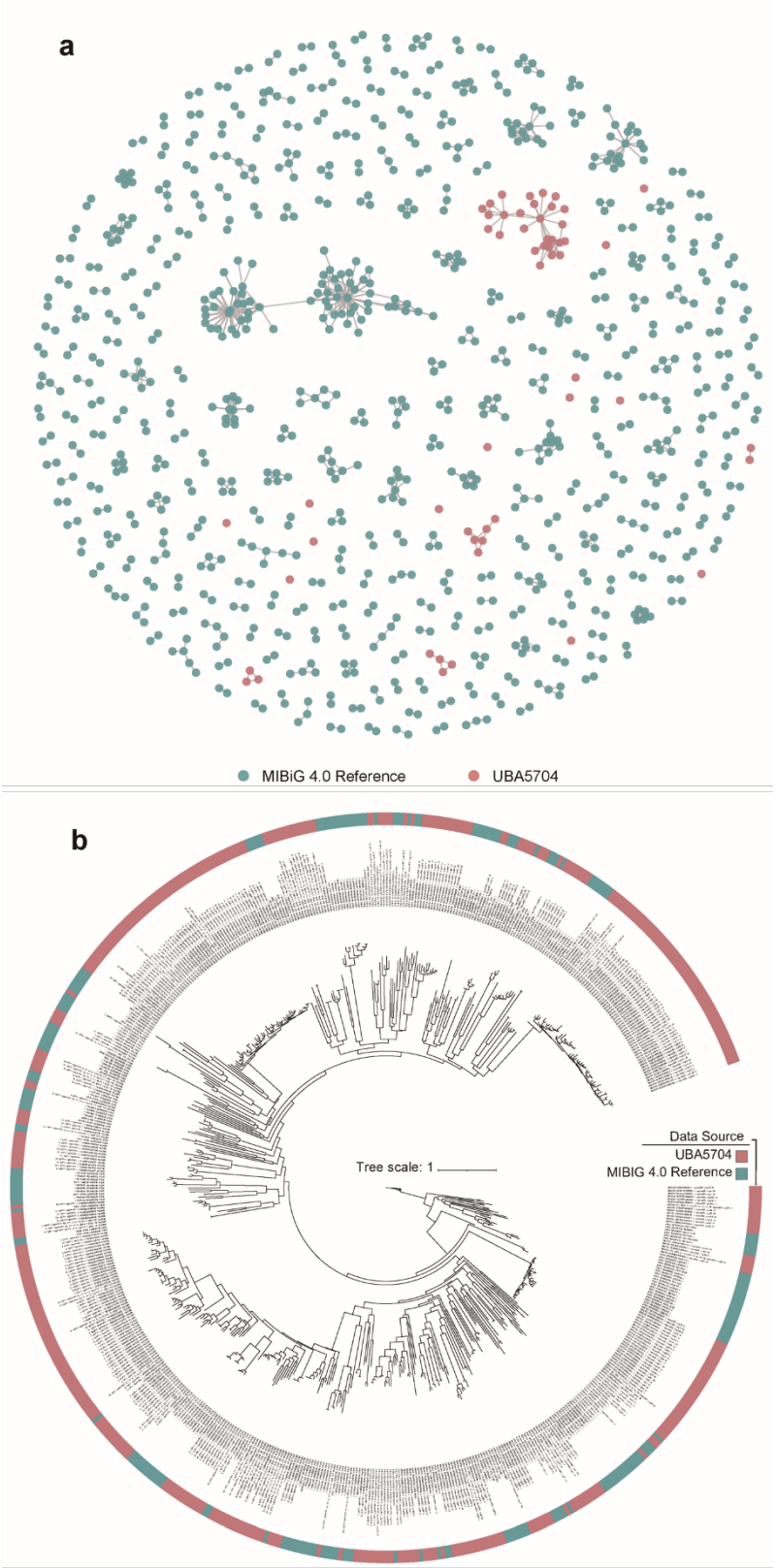


Fig. S7 | Similarity network and phylogenetic placement of ranthipeptide-related core enzymes in UBA5704 genomes.

a, Similarity network of ranthipeptide-like BGCs from UBA5704 genomes and MIBiG 4.0 reference BGCs. A total of 54 UBA5704 ranthipeptide-like BGCs were clustered

together with MIBiG 4.0 reference BGCs at a cutoff of 0.3. No UBA5704 ranthipeptide-like BGC was connected to any MIBiG 4.0 reference BGC at this threshold. **b,** Phylogenetic tree of radical rSAM core enzymes from UBA5704 genomes and MIBiG 4.0 reference BGCs. The tree includes 397 high-confidence rSAM core enzymes from 76 UBA5704 genomes and 192 reference rSAM enzymes from MIBiG 4.0. The outer ring indicates the data source, distinguishing UBA5704 sequences from MIBiG 4.0 reference sequences. BGC, biosynthetic gene cluster; rSAM, radical S-adenosylmethionine; MIBiG, minimum information about a biosynthetic gene cluster.

Supplementary Text

1. Robustness analysis using water salinity

Traditional salinity assessments typically rely exclusively on measurements of the overlying water column, which may fail to capture the long-term regulatory mechanisms governing salt cycling within saline lake ecosystems[1]. In these systems, the water and sediment constitute a coupled hydrological compartment, wherein solutes are actively exchanged across the water – sediment interface via diffusion, advection, and bioturbation[2]. Sediment functions not only as a passive repository but also as a critical buffer, or "memory bank" of salts. Through diagenetic processes and diffusive fluxes, it continuously releases or sequesters salts, exerting a profound influence on the instantaneous salinity of the overlying water[1]. Consequently, accurately characterizing the salinity environment that shapes the microbial community structure and function requires the consideration of both compartments. While water column salinity is highly susceptible to short-term hydrological dynamics, such as seasonal variations and precipitation inputs, sediment salinity records the long-term physicochemical evolution of the lake.

In light of these coupled water – sediment dynamics and salt cycling mechanisms, we introduce an Integrated Salinity Index (ISI), a novel metric that synthesizes both water and sediment salinity measurements to construct a more ecologically meaningful salinity gradient. This new index not only overcomes the partiality of traditional assessments but also accurately reflects the overall salt storage and exchange status of saline lakes, thereby providing a robust quantitative foundation for comparative microbiome analyses across diverse bioclimatic zones.

To assess whether the core findings of this study are dependent on the definition of the ISI, we conducted a parallel analysis using water salinity alone as the grouping criterion. To maintain consistency with the ISI-based classification (which excluded lakes with ISI $\leq$ 1 g/L), we restricted this analysis to lakes with water salinity > 1 g/L. Water salinity values were used directly to classify lakes into the same three salinity categories (brackish, mososaline, and hypersaline) based on the established thresholds (1 – 5, 5 – 50, and >50 g/L)[3–5]. The specific analysis is as follows:

1.1 α-diversity patterns

Under the water salinity-based classification, α-diversity patterns varied between water and sediment samples. In both habitats, hypersaline lakes exhibited the lowest diversity. Shannon diversity was significantly higher in mesosaline lakes than in hypersaline lakes in both water and sediment ($P < 0.05$), whereas Simpson diversity showed a similar pattern only in sediment samples, with no significant difference observed in water samples (Fig. S4a). Under the ISI-based classification, the patterns were more consistent across habitats. Both brackish and mesosaline lakes exhibited significantly higher Shannon and Simpson diversity compared to hypersaline lakes in both water and sediment samples ($P < 0.05$), indicating a clearer separation of diversity patterns across the salinity gradient (Fig. S4a).

1.2 β-diversity patterns

In both water and sediment samples, the ISI-based classification consistently explained a greater proportion of variation in microbial community composition compared to the water salinity-based classification (Fig. S4b). PERMANOVA analysis results showed higher R^2 values under the ISI-based scheme than traditional group in both habitats (sediment: 0.053 vs. 0.044; water: 0.081 vs. 0.057; all $P < 0.001$), indicating that this integrated index better captured salinity-driven differentiation of microbial communities.

Collectively, these robustness analyses confirm that the core ecological patterns are not artifacts of the specific salinity index employed. Moreover, the ISI-based classification outperformed the traditional water salinity-based approach in discriminating microbial community structure across salinity gradients, supporting its utility as a more comprehensive measure of the overall salinity environment in saline lake ecosystems.

2. Detailed comparison of biosynthetic gene cluster distribution across bioclimatic regions and salinity gradients

To provide a comprehensive view of the biosynthetic potential of saline lake microbiomes, we systematically compared the distribution of biosynthetic gene clusters

(BGCs) across different bioclimatic regions and salinity gradients.

2.1 Comparison of BGC abundance across salinity gradients

Our main-text focused on five predominant BGC types, including PKS, NRPS, RiPPs, terpenes, and NRPS/PKS hybrids (Fig. 3e and S3b), while other low-abundance BGC types with significant differences were presented in Fig. S3b. Specifically, the abundances of PKS and NRPS gene clusters in the QTP region were significantly higher in both water and sediment of brackish or/and mesosaline lakes than in hypersaline lakes (all $P < 0.05$). In the LP region, NRPS, PKS, and NRPS/PKS hybrids were more prevalent in the water of brackish and mesosaline lakes compared to hypersaline lakes (all $P < 0.05$); however, RiPPs and terpene gene clusters were significantly more abundant in the water of hypersaline lakes than in brackish and/or mesosaline lakes ($P < 0.05$). In the MXP region, no significant differences were detected in water samples for any BGC type. In sediment samples, NRPS abundance was significantly higher in brackish lakes than in hypersaline lakes, whereas terpene clusters were significantly more abundant in hypersaline lakes than in mesosaline or brackish lakes (all $P < 0.05$). This relatively simple pattern suggests that salinity exerts a limited effect on BGC composition in this mid-altitude temperate arid region.

2.2 Effects of bioclimatic factors on BGC differentiation

To assess the influence of bioclimatic factors, we further stratified the analysis by salinity group and compared BGC abundance across the three bioclimatic regions (Fig. 3e and Fig. S3c). In brackish lake ecosystems, with the exception of the PKS gene cluster, which exhibited a significantly higher prevalence in sediments from the QTP region compared to the MXP region ($P < 0.05$), all other significantly differentiated BGCs were of low-abundance types and were primarily enriched in sedimentary environments at LP region (all $P < 0.05$). In mesosaline and hypersaline lakes, significantly differentiated BGCs were predominantly derived from water samples, with only one low-abundance BGC type showing significant differences in hypersaline sediments. Among these, NRPS and NRPS/PKS hybrid gene clusters exhibited significant enrichment in the MXP and/or QTP regions (all $P < 0.05$). Notably, in hypersaline lakes, all significantly differentiated BGCs showed significantly higher

abundance in the MXP region than in the LP region (all $P < 0.05$). These findings indicate that in brackish lakes, the LP region harbors a greater diversity of low-abundance BGCs, whereas in mesosaline and hypersaline lakes, the MXP and QTP regions possess more high-abundance BGCs.

Reference

1. Bischoff, J. L., Fitts, J. P. & Fitzpatrick, J. A. Responses of sediment geochemistry to climate change in owens lake sediment: an 800-k.y. record of saline/fresh cycles in core OL-92. **317**, 37–47 (1997).
2. *Saline Lakes*. (Springer Netherlands, Dordrecht, 1990). doi:10.1007/978-94-009-0603-7.
3. Song, C. *et al.* Widespread declines in water salinity of the endorheic tibetan plateau lakes. *Environ. Res. Commun.* **4**, 091002 (2022).
4. Deng, M. *et al.* A non-optically active lake salinity dataset by satellite remote sensing. *Sci. Data* **12**, 1324 (2025).
5. Zheng, M. & Liu, X. Hydrochemistry of salt lakes of the qinghai-tibet plateau, china. *Aquat. Geochem.* **15**, 293–320 (2009).