Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The Saline Lake Microbiome Resource: a genome-resolved atlas for biosynthetic discovery

Ke Yu

yuke.sz@pku.edu.cn

Peking University https://orcid.org/0000-0001-5039-6056

Yu Wang

Shenzhen Graduate School, Peking University

Runzhi Cao

Shenzhen Graduate School, Peking University https://orcid.org/0009-0008-1838-1901

Xuejiao Qiao

Shenzhen Graduate School, Peking University

Yujun Wu

Shenzhen Graduate School, Peking University

Zhiguang Qiu

Shenzhen Graduate School, Peking University

Chun-Ang Lian

Peking University

Yan Yan

Guangzhou Institute of Geochemistry, Chinese Academy of Sciences

Fan Wang

School of Ecology, Sun Yat-sen University

Bing Li

Tsinghua University https://orcid.org/0000-0002-7161-0477

Baowei Chen

Sun Yat-Sen University https://orcid.org/0000-0002-5230-8264

Yanping Mao

Shenzhen University

Liguan Li

Department of Science and Environmental Studies, The Education University of Hong Kong
https://orcid.org/0000-0003-4349-0532

Lei Gao

School of Life Sciences, Sun Yat-Sen University

Jian-Yu Jiao

School of Life Sciences, Sun Yat-sen University

Wen-Jun Li
Xinjiang Institute of Ecology and Geography https://orcid.org/0000-0002-1233-736X
Shengwei Hou
Southern University of Science and Technology https://orcid.org/0000-0002-4474-7443
Liang Yang
School of Brewing Engineering, Moutai Institute https://orcid.org/0000-0002-2025-2067
Meng Li
Shenzhen University https://orcid.org/0000-0001-8675-0758
Nan Xu
Shenzhen Graduate School, Peking University https://orcid.org/0000-0001-9737-396X
Huapeng Qin
Peking University Shenzhen Graduate School https://orcid.org/0000-0002-5132-5329
Jinren Ni
Peking University https://orcid.org/0000-0002-9114-8347
Tong Zhang
The University of Hong Kong https://orcid.org/0000-0003-1148-4322
Wei-Qin Zhuang
The University of Auckland https://orcid.org/0000-0001-9600-5225

Article

Keywords:

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11286303/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

The Saline Lake Microbiome Resource: a genome-resolved atlas for biosynthetic discovery

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).

Abstract

Saline lakes harbor unique microbial diversity and biosynthetic potential, yet remain largely unexplored genomically. We constructed the Saline Lake Microbiome Resource (SLMR), a genome-resolved atlas built from 567 water and sediment samples (paired by lake) from 76 lakes spanning salinities of 0.29–595.84 g/L and pH 6.15–10.38, covering brackish to hypersaline conditions. The resource comprises 48,830 metagenome-assembled genomes and ~140,000 predicted biosynthetic gene clusters (BGCs), most showing substantial divergence from characterized reference clusters. Its 24,780 species representatives span 151 phyla; 94.69% lack species-level annotation and 89.77% are absent from existing genome catalogs. Water and sediment harbor distinct, complementary microbial and biosynthetic profiles along the salinity gradient, with biosynthetic richness comparable to marine and coral microbiomes. Catalog mining identified the poorly characterized, BGC-rich order *Multivoradales*, and enrichment cultivation yielded 53 additional MAGs. Its candidate family UBA5704 has an exceptionally large biosynthetic repertoire, exemplified by a complete, novel 21.4-kb ranthipeptide cluster resolved by structural modeling. SLMR provides an open resource for biosynthetic discovery in extreme environments.

As one of the most distinctive and widely distributed aquatic ecosystems on Earth, saline lakes exhibit pronounced environmental heterogeneity and unique ecological functions[1]. Unlike freshwater, saline lakes are predominantly located in inland arid or semiarid regions. Shaped by their unique geographic locations, distinctive climatic conditions, and complex hydrological processes, different saline lakes exhibit pronounced heterogeneity in environmental conditions[2,3]. This multidimensional environmental heterogeneity imposes diverse selective pressures that drive microbial niche differentiation and adaptive evolution. In response to environmental selection pressures primarily dominated by salinity, which span a broad gradient from brackish to hypersaline conditions, and other concomitant environmental stresses, the microbial communities in saline lakes have evolved highly specialized physiological and metabolic mechanisms[4–6]. These mechanisms not only provide critical insights into microbial

environmental adaptation and metabolic diversity but also endow saline lake microorganisms and their metabolites with unique biotechnological potential.

In recent years, the potential of saline lake microorganisms for novel natural product discovery has become increasingly evident, particularly in the biomedical field. For instance, microbiome studies of sediments from the Great Salt Lake in the United States and Chitu Lake in Ethiopia have identified compounds with antitumor and antimicrobial activities, as well as numerous uncharacterized novel biosynthetic gene clusters[7,8]. Notably, Liu et al. reported that *Actinomycetes* isolated from saline lake sediments in the northern Tibetan Plateau exhibited broad-spectrum antimicrobial activity, and subsequent isolation from a highly active strain yielded echinomycin and calcimycin—two compounds with both antitumor and antimicrobial properties[9]. These findings demonstrate that the functional diversity of microorganisms inhabiting extreme environments such as saline lakes represents a valuable resource for novel natural product discovery and microbial drug development. However, research on lake microbiomes remains in its infancy compared to marine and human-associated systems. This gap is most pronounced for extreme environments such as saline lakes, where the discovery of novel species and natural products remains nascent. The vast uncultivated microbial "dark matter" in these ecosystems, along with its hidden metabolic and biosynthetic potential, urgently requires systematic exploration via metagenomics and synthetic biology.

Metagenomic binning enables the cultivation-independent reconstruction of numerous bacterial and archaeal draft genomes by clustering microbial sequences from metagenomic sequencing data. These metagenome-assembled genomes (MAGs), combined with strain-resolved gene annotation, functional profiling, comparative genomics, and phylogenetic analysis, allow researchers to gain deep insights into the metabolic potential, nutritional interaction mechanisms, and ecological adaptation strategies of uncultivated microorganisms in diverse environments[10,11]. This approach has been successfully applied to a range of extreme environments, including global oceans[12,13], aquatic ecosystems on the Tibetan Plateau[14], rivers[15], and glaciers[16], resulting in the construction of microbial genome catalogs that provide critical insights into the evolutionary relationships and metabolic features of previously uncultured bacteria and archaea. In lake ecosystems, recent studies have begun to employ this technology to explore microbial diversity. For example, Garner et al. constructed a freshwater

lake microbial genome catalog and identified trophic state as a key driver of microbial community structure[17]. Zhang et al., working in Tibetan Plateau lakes spanning a gradient from freshwater to low-salinity conditions, demonstrated the pivotal role of salinity in shaping microbial community assembly[18]. However, existing studies suffer from two notable limitations. First, research on Chinese lakes has largely been confined to single geographic regions (e.g., the Qinghai–Tibet Plateau), with insufficient coverage of lake types spanning distinct bioclimatic regions and salinity gradients. Second, the lack of concurrent surveys of both water and sediment habitats has hindered a comprehensive understanding of the ecological adaptation strategies of saline lake microorganisms. Addressing these gaps will facilitate a deeper elucidation of the diversity distribution patterns and biosynthetic potential of saline lake microbiomes, and provide a scientific foundation for the biotechnological exploration of related microbial resources.

To address these knowledge gaps, we conducted the first systematic study of saline lake microbiomes across three major bioclimatic regions in China and along salinity gradients. Representative saline lakes were selected from the Qinghai-Tibet Plateau (alpine cold arid zone, hereafter "QTP"), the Mongolian-Xinjiang Plateau (temperate arid zone, "MXP"), and the Loess Plateau (temperate semiarid to subhumid zone, "LP"). Water and sediment samples were systematically collected and subjected to multi-omics analyses, including physicochemical parameter measurements (such as water and sediment salinity, pH, and ionic composition, and notably with an integrated salinity index introduced to characterize the overall salinity of the lake system), metagenomic and metatranscriptomic data. By integrating sample physicochemical parameters, MAGs, microbial community composition, biosynthetic gene cluster (BGC) profiles, we constructed the first saline lake microbiome resource spanning diverse bioclimatic regions across extensive salinity and pH gradients, designated as the Saline Lake Microbiome Resource (SLMR) (Fig. S1). This resource not only fills a critical gap in large-scale genomic studies of saline lake ecosystems but also provides a foundational dataset for exploring the adaptation mechanisms and metabolic potential of extremophilic microorganisms. To promote open access and in-depth exploration of these resources, we developed the first publicly accessible interactive online platform for saline lake microbiomes. The platform systematically integrates all SLMR data with corresponding visualization figures,

offering researchers a comprehensive view from community structure to functional potential. In the future, the platform will continue to incorporate global multi-source data, with the goal of evolving into a dynamically expanding global saline lake microbiome resource center covering major bioclimatic regions worldwide. This initiative aims to provide an open data foundation and scientific support for research in microbial ecology, bioprospecting, and synthetic biology, facilitating the translation of saline lake microbial resources from fundamental research to biotechnological applications.

Results

Construction of a publicly accessible SLMR. To establish the Saline Lake Microbiome Resource (SLMR) that comprehensively covers the saline lake ecosystem in China, we gathered data from 76 lakes over a span of four years. These lakes were geographically dispersed across diverse bioclimatic regions, including 34 lakes from the QTP region, 28 lakes from the MXP region, and 14 lakes from the LP region. The altitude ranges from -154 meters to 4,718 meters, and the geographical coordinates fall between 81.28°E and 111.07°E, as well as between 29.13°N and 47.30°N. This dataset encompasses nearly all the distribution zones of saline lakes in China (Fig. 1a, 1d and Table S1). A grand total of 567 collected environmental specimens were gathered across diverse saline lakes, including 319 sediment samples and 248 water samples (Fig. 1b). While sampling in select lakes of the MXP region only yielded sediment records, the vast majority of study sites achieved water–sediment pairing, a dual-phase sampling strategy that ensures comprehensive geochemical and ecological insight. The sampling design system of this study comprehensively covered multiple environmental dimensions, including different media (water/sediments), altitude gradients, salinity and pH ranges, and a time span from 2018 to 2022. Therefore, this represents the first large-scale, paired investigation of the microbiomes of water and sediment within the saline lake ecosystem.

While not all individual samples were profiled, comprehensive physicochemical data (including salinity and pH) were obtained for both water and sediment from each lake, spanning a broad salinity range of 0.29–595.84 g/L and a pH range of 6.15–10.38 (Fig. 1c). This wide coverage encompasses brackish, mesosaline, and hypersaline conditions, making SLMR a

valuable resource for saline lake research. Even more impressively, 255 samples additionally contained a full suite of major anions like fluoride (F^-), chloride (Cl^-), nitrite (NO_2^-), bromide (Br^-), nitrate (NO_3^-), sulfate (SO_4^{2-}), and phosphate (PO_4^{3-}), several cations (K^+, Al^{3+}, Li^+), and a diverse array of trace metal elements such as copper (Cu), iron (Fe), manganese (Mn), molybdenum (Mo), nickel (Ni), and zinc (Zn) (Table S1). In addition, some sediment samples also contained total organic carbon (TOC) and organic matter content. In terms of omics data, a total of 482 metagenomic samples (approximately 16 Tb of data) and 21 metatranscriptomic samples (around 1 Tb of data) were obtained (Table S1). In conclusion, SLMR not only integrates multi-omics data and comprehensive environmental variables but also serves as a data repository for microbial communities and their functions, underpinning microbial ecological research in saline lake ecosystems.

To promote the extensive utilization of microbial community resources in saline lake ecosystems, all the raw sequencing data of metagenomes and metatranscriptomes generated in this study have been deposited in the public databases of the National Center for Biotechnology Information (NCBI) and the China National GeneBank Database (CNGBdb), and were freely accessible to researchers worldwide. Meanwhile, we have developed an open-access, interactive online platform (https://slmr.site/). This platform systematically integrates sample environmental physicochemical data, MAGs along with their multi-level taxonomic annotations, and BGCs. Users can browse, search for, and download data via the graphical interface of this platform. The SLMR platform serves as a crucial resource for the in-depth exploration of the microbiology of saline lakes. In the future, we will persistently expand and update its content. Specifically, we plan to incorporate functional annotation datasets, including the pan-genome of saline lake ecosystems, viral genomes, and antibiotic resistance genes (ARGs). This will gradually transform the platform into a sustainable repository that integrates and unifies global saline lake multi-omics data, thereby offering fundamental support for comparative microbial ecological research across regions and habitats.

~50,000 metagenome-assembled genomes recovered from diverse saline lake ecosystems. We undertook a comprehensive effort to catalog the microbial genomes present in these saline lake ecosystems across China, covering a wide variety of water and sediment samples. From

482 saline lake metagenomes (288 sediment and 194 water), metagenomic assembly and binning yielded a total of 48,830 MAGs, enabling the construction of a remarkably comprehensive Saline Lake Microbial Catalog (SLMC) (Fig. 2a, S2a, Table S1 and S2). This catalog comprises 11,902 MAGs from water samples and 36,928 MAGs from sediment samples. All MAGs meet or exceed the medium-quality MIMAG standards (completeness ≥ 50%, contamination < 10%)[19], with a mean completeness of 79.95% ± 14.82% and a mean contamination of 2.34% ± 2.05% (Fig. S2b). Among these MAGs, 39,740 exhibited a quality score surpassing 50. Moreover, 3,058 of these MAGs were classified as high-quality, characterized by the presence of the 23S, 16S, and 5S rRNA genes alongside at least 18 tRNAs, showcasing their remarkable genomic integrity and functional potential (Fig. 2b and Table S2). Notably, sediment-derived MAGs exhibited significantly higher contamination (water: 1.73 ± 1.80%, sediment: 2.53 ± 2.08%) and lower completeness (water: 83.36 ± 14.47%, sediment: 78.85 ± 14.77%) compared to those from water samples ($P < 0.001$) (Fig. S2d). This disparity likely stems from the fact that sediments tend to host richer and more intricate microbial ecosystems than water, thereby increasing the likelihood of genomic contamination during the recovery process. Furthermore, both genome size (water: 2.89 ± 1.42 Mb, sediment: 2.94 ± 1.36 Mb) and GC content (water: average 55.58%, ranging from 22.72% to 75.84%, sediment: average 56.57%, ranging from 21.72% to 77.32%) of these MAGs differed significantly between water and sediment samples ($P < 0.001$) (Fig. S2d), further reflecting the divergence of microbial communities between these two ecological niches.

To evaluate the novelty of the SLMC, we clustered 48,830 MAGs into 24,780 species-representative genomes (hereafter "rMAGs") based on genomic similarity (Table S2). Comparative analysis was conducted against reference genomes from the Genome Taxonomy Database (GTDB release R226)[20,21], Genomes from Earth's Microbiomes (GEM catalog)[22], Tibetan Plateau Microbial Catalog (TPMC)[14], Tibetan Glacier Genome Catalog (TG2G)[16], TARA ocean[23], soil microbial assembled genome (SMAG catalog)[24], and Genome Resolved Open Watersheds (GROW catalog)[15]. Strikingly, 94.51%, 98.79%, 95.52%, 99.82%, 99.98%, 98.89%, and 99.86% of the SLMC genomes exhibited an aligned fraction below 30% to the aforementioned reference catalogs, respectively (Fig. 2c and Table S3). Moreover, a total of 22,244 rMAGs (89.77%) could not be matched to any genome in these datasets, underscoring

the exceptional novelty of the SLMC genomes. Given that these reference datasets cover diverse aquatic and terrestrial ecosystems, such as oceans, rivers, deep subsurface environments, and soils, the observed high level of novelty suggests that saline lake ecosystems may contain unique evolutionary pathways and adaptive mechanisms, with microbial diversity and functional potential being collectively shaped by saline lake-specific conditions such as high salinity[25], semi-enclosed hydrology[26], and water–sediment coupling[27]. To further assess the novelty of SLMC, we compared 24,780 rMAGs with the LakePulse MAG catalog, a comprehensive genomic reference for freshwater lake microbiomes[17]. Only 57 rMAGs from SLMC matched genomes in the LakePulse catalog (Fig. 2c and Table S3), further underscoring the distinctiveness of saline lake microbiomes in terms of microbial composition and metabolic profiles, setting them apart from other ecosystems, including freshwater environments. Additionally, the addition of SLMC expanded the known genome repertoires of GTDB, GEM, and TPMC by 20.71% (23,419 novel genomes compared to 113,104 references), 106.23% (24,481 novel genomes compared to 23,045 references), and 220.73% (23,669 novel genomes compared to 10,723 references), respectively. Overall, SLMC contributes to completing the global microbiome reservoir and deepens our understanding of microbial diversity within these distinctive saline lake environments.

Based on taxonomic annotations from the GTDB (release R226)[20,21], the SLMC encompasses 151 known phyla, representing a striking 89.3% of all currently recognized GTDB phyla, underscoring the substantial genomic diversity and richness of saline lake microbiomes. To unveil their deep phylogenetic architecture, we reconstructed a comprehensive, genome-informed phylogenetic tree for the 24,780 SLMC species and specifically highlighted the top 15 phyla of the total MAGs (Fig. 2d). Notably, saline lake ecosystems exhibit a distinct microbial community structure that is significantly different from those in other aquatic ecosystems. Specifically, in bacterial MAGs, *Pseudomonadota*, *Bacteroidota*, *Actinomycetota*, and *Verrucomicrobiota* predominate in freshwater (74.90%)[17], rivers (86.10%)[15], and the aquatic ecosystem of Tibetan Plateau (80.00%)[14](Fig. S2c). However, these four phyla account for only 45.88% of the microbial community in saline lakes. In contrast, we successfully recovered a greater number of MAGs belonging to phyla that are typically less abundant in other aquatic ecosystems, such as *Chloroflexota* (7.58%),

Patescibacteria (6.73%), *Planctomycetota* (5.06%), and *Desulfobacterota* (4.56%), which consistent with previous observations in saline lake microbiomes[28–30]. These observations suggest that microbial communities in saline lake ecosystems exhibit higher diversity and a more evenly distributed structure compared to other aquatic environments. The archaeal MAGs were predominantly comprised of *Halobacteriota* (3.59%) and *Nanoarchaeota* (1.68%), aligning with findings from previous studies on Tibetan Plateau water ecosystems[14]. Beyond providing genomic resources for these ecologically characterized taxa, our analysis revealed that a considerable number of SLMC rMAGs remain unclassified by GTDB across various taxonomic ranks: phyla (n=4), class (n=35), order (n=186), family (n=1,005, 4.06%), genus (n=7,127, 28.76%), and species (n=23,464, 94.69%) (Table S2 and Fig. S2e). Moreover, a large proportion of MAGs were assigned to lineages currently designated by alphanumeric or numeric identifiers (e.g., SLHW01, 0-14-0-80-60-11) at the phylum (n=333, 1.34%), class (n=2,249, 9.08%), order (n=6,289, 25.38%), family (n=11,143, 44.97%), and genus (n=10,965, 44.25%) levels. These unclassified MAGs likely represent potential novel lineages, substantially expanding our understanding of the genomic diversity of saline lake microbiomes.

Biosynthetic capacity in saline lake microbiome across bioclimatic regions and salinity. The biosynthetic potential of microbiome inhabiting the unique ecological niches of saline lakes, and their significant promise as a vast, untapped source of novel natural products, remains largely unexplored. Consequently, drawing upon more than 9.5 million contigs (exceeding 5 kb) assembled from a collection of 482 metagenomic datasets, we conducted a comprehensive and systematic prediction of the biosynthetic gene clusters (BGCs) embedded within these sequences. A total of 139,190 BGCs were identified, with 103,269 derived from sediments and 35,921 from water samples (Fig. 3a and Table S4). Notably, the abundance of BGCs in sediments was significantly higher than that in water (coverage, abundance/Gb) ($P < 0.05$), unveiling a striking disparity in biosynthetic potential (Fig. 3a). Then, we examined the influence of environmental factors on the abundance of BGCs in water and sediments. Our findings revealed that spatial processes (represented by PCNM2) and the mean temperature of the wettest quarter emerged as crucial environmental drivers, exerting statistically significant influences on the abundance of BGCs across both water and sediment microbiomes (Fig. 3b).

Furthermore, altitude and pH were identified as additional critical modulators, shaping BGC richness specifically within the water with striking significance.

All BGCs were categorized into 26 types, of which ribosomally synthesized and posttranslationally modified peptides (RiPPs), terpene, polyketide synthase (PKS), non-ribosomal peptide synthetase (NRPS), and polyketide-nonribosomal peptide combinations (PKS/NRPS hybrids) were the most dominant types of BGCs in both sediment and water microbiome (Fig. 3c). Among these 26 BGC types, six were notably enriched in sediment environments, especially the NRPS and RiPP gene clusters (all $P < 0.001$) (Fig. 3d and S3a). In contrast, only the terpene gene cluster exhibited a significantly higher enrichment in aquatic environments ($P < 0.001$) (Fig. 3d). This spatial heterogeneity reflects the adaptive selection of microbial metabolic strategies in response to distinct habitat conditions (water vs. sediment), highlighting the habitat-dependent functional differentiation of BGCs within saline lake ecosystems.

To characterize the salinity environment of the studied lakes, we first examined the relationship between water and sediment salinity. For lakes with both water and sediment samples (n = 52), water salinity and sediment salinity exhibited a strong positive correlation (Spearman's $\rho = 0.73$, $P < 0.01$). Given the coupled nature of water and sediment habitats in saline lakes, where microorganisms in both compartments experience a shared salinity regime, we used the Integrated Salinity Index (ISI) defined in the Methods as a comprehensive measure of the overall salinity environment. The strong correlation between the two salinity measures supports the feasibility of combining them into a single integrated index. Based on ISI values, the lakes were classified into three groups: brackish (n = 25), mesosaline (n = 23), and hypersaline (n = 28) (Fig. 2a and Table S5). This classification allowed us to effectively capture salinity-driven differences in microbial community structure and functional potential across saline lakes. Notably, the major ecological patterns observed with ISI were robust to the definition of the salinity index. A parallel analysis using water salinity alone confirmed the reliability of the ISI-based classification, while further demonstrating that ISI provided a stronger differentiation of microbial community structure across salinity gradients (Fig. S4 and Table S6). Additional detailed comparisons are provided in the Supplementary Information.

Subsequently, we systematically compared the biosynthetic potential of microbial

communities across bioclimatic regions and salinity gradients, focusing on five predominant BGC types (PKS, NRPS, RiPPs, terpenes, and NRPS/PKS hybrids) (Fig. 3e and S3b). The ISI-based classification revealed more pronounced differences across salinity groups than the traditional water-salinity grouping, supporting the utility of ISI for linking salinity to biosynthetic potential (Fig. S5). Specifically, across the QTP and LP regions, PKS, NRPS, and NRPS/PKS hybrid gene clusters were consistently enriched in brackish and mesosaline lakes, a pattern that persisted across both water and sediment as well as low- and high-altitude environments, highlighting their broad environmental adaptability and consistent biogeographical occurrence. In the MXP region, the response pattern was relatively simple, with only limited salinity-driven variation in BGC composition detected in sediments. Bioclimatic factors also exerted differential effects on BGC differentiation across salinity backgrounds (Fig. 3e and Fig. S3c). In brackish lakes, low-abundance BGCs were primarily enriched in the LP region, whereas in mesosaline and hypersaline lakes, high-abundance BGCs were more prevalent in the MXP and QTP regions, with NRPS and NRPS/PKS hybrids showing significant enrichment in these areas. Collectively, these results highlight the selective role of bioclimatic factors such as altitude, temperature, and geographic location on specific BGC types, underscoring the importance of integrating both local salinity conditions and broader environmental contexts when assessing the biosynthetic potential of microbial communities in saline lakes. Additional detailed comparisons are provided in the Supplementary Information.

We clustered all BGCs into 109,232 non-redundant gene cluster families (GCFs), with RiPPs, PKS, and other types all exhibiting high diversity. Among these, we examined the distribution of 12 relatively abundant natural products across saline lake ecosystems, categorized by potential function into three groups: environmental adaptation, molecular recognition and regulation, and antibiotics and antioxidants. Some substances (e.g., arylpolygene, betalactone, ectoine, and hserlactone) were widely distributed across both water and sediment, suggesting roles in microbial adaptation to saline lake conditions (Fig. 3e). Others showed distinct habitat preferences: hglE-KS and ranthipeptide were enriched in sedimentary microbiomes, with ranthipeptide primarily in mid- to high-altitude (MXP and QTP regions) brackish lakes, while phosphonates were concentrated in mid-altitude hypersaline

sediments. Together, these patterns suggest that interactions between salinity and bioclimatic context drive spatial differentiation of biosynthetic pathways, with sediments serving as a primary reservoir for habitat-specific secondary metabolic potential in saline lakes ecosystems.

Discovery of novel and undocumented BGC-rich lineages. To tackle the taxonomy of BGCs hosts, we carried out comprehensive BGC predictions across the 48,830 MAGs, ultimately identifying a total of 86,579 BGCs (Fig. 4a and Table S7). Next, we aimed to investigate the richness and the degree of novelty of the biosynthetic potential in the SLMC. To achieve this, we first clustered the 86,579 BGCs into 61,533 GCFs, which are designed to offer a more precise grouping for the BGCs that are predicted to encode similar natural products. A comparative analysis of BGCs from saline lake microbiomes (61,533 GCFs from 24,780 species, 2.48 GCF per species) with those from ocean microbiomes (6,907 GCFs from 8,304 species, 0.83 GCF per species), recently characterized coral microbiomes (6,612 GCFs from 4,224 species, 1.57 GCF per species), and Tibetan Plateau aquatic microbiomes (18,414 GCFs from 10,723 species, 1.72 GCF per species) revealed that saline lake microbial communities exhibited the highest biosynthetic richness, underscoring their substantial biosynthetic potential (Fig. 4b). When comparing the GCFs detected in saline lake microbiomes with those in the OMD[12], RMD[31], TPMC[14], and biochemically characterized (MIBiG 4.0) database[32], we found that 93.36% of the GCFs were newly identified in this study (Fig. 4b). This high novelty is likely driven by both the exceptional phylogenetic breadth of saline lake microbiomes (spanning 151 phyla, 89.3% of GTDB) and the limited prior investigation of BGCs in these unique environments. Additionally, we selected the longest BGC sequence for each species from each GCF and compared them against the BiG-SLiCE database[33], which contains 1.2 million known BGCs. BGCs with BiG-SLiCE distances greater than 0.4 were considered novel. In the SLMC, we identified a total of 56,289 (79.18%) novel BGCs (Table S8). Among these, 6,212 BGCs exhibited BiG-SLiCE distances greater than 1, indicating substantial divergence from known BGC families. All predominant BGC types had more than 75% novel BGCs, including RiPP (n = 14,395, 77.62%), terpenes (n = 12,502, 78.60%), PKS (n = 11,232, 83.80%), NRPS (n =9,069, 85.57%), and NRPS/PKS hybrid (n = 1,829, 81.83%) gene clusters. These findings indicate that the microbial communities in the saline lake ecosystem harbor

extensive, diverse, and previously unrecognized biosynthetic potential, which is expected to substantially enrich the current global BGCs repository.

To broaden our comprehension of the biosynthetic capabilities harbored within the microbiome of saline lakes, we sought to uncover novel phylogenetic clades that exhibit a heightened enrichment of BGCs. Our study revealed that the prevalence of these BGC-rich MAGs (representatives with ≥ 15 BGCs) was significantly higher in sediment samples compared to those in the water ($P < 0.001$). Several BGC-rich clades that are either well known for their biosynthetic potential were identified in our datasets, such as *Actinomycetota* (*Micromonosporaceae*)[34], *Planctomycetota*[35], *Pseudomonadota*[36], and *Myxococcota* (*Myxococcaceae*, *Polyangiales* and *Enhygromyxa*)[37–39] (Table S9 and Fig. 4a). Notably, the BGC-rich lineage *Eremiobacterota*, previously identified in marine environments[12], was also recovered in saline lakes, where three MAGs belonging to an uncharacterized family within the order *Xenobiales* were identified.

Among the BGC-rich MAGs identified in this study, members of the phylum *Acidobacteriota* exhibited the highest biosynthetic potential, with an average of 31.73 BGCs per genome. Notably, the majority of these high-potential MAGs (48 out of 59) belonged to the order *Multivoradales*, a lineage that remains largely uncultivated and has been rarely documented in previous studies (Table S9). These *Multivoradales* MAGs averaged 29.94 BGCs per genome, with one MAG harboring as many as 87 BGCs. The BGCs identified in these genomes were compositionally diverse, predominantly featuring NRPS and NRPS/PKS hybrid gene clusters, with several MAGs also containing relatively high numbers of terpene and RiPP gene clusters (Fig. 4c). To facilitate future efforts in natural product synthesis and to further explore the biosynthetic potential of this lineage, we attempted to recover a substantial number of *Multivoradales* genomes through enrichment cultures. By applying various enrichment strategies to representative lake samples, we successfully obtained 53 *Multivoradales* MAGs (Table S10). Among these, BGCs were identified in 49 MAGs, with 12 classified as BGC-rich. Collectively, these results further underscore the substantial biosynthetic potential of this lineage.

UBA5704 reveals extraordinary biosynthetic potential and novel ranthipeptide clusters.

To resolve the phylogenomic framework of the order *Multivoradales*, we integrated 178 MAGs from our metagenomes, 53 from enrichment cultures, and 293 public genomes, resulting in a dataset of 524 medium- to high-quality genomes (Table S11). A lineage belonging to a potential new family (UBA5704) exhibited exceptionally high biosynthetic potential (Fig. 5a). Comprising 76 MAGs, the UBA5704 family exhibits an average of 40.95 BGCs per genome, which is comparable to the reported average of the well-established BGC-rich genus *Streptomyces* (29.76–44)[40,41]. Furthermore, over 92% (70 out of 76) of the genomes were BGC-rich, highlighting the extensive and consistent biosynthetic potential of this lineage. This pattern is even more pronounced in our nine MAGs from saline lakes, which have an average of 42 BGCs per genome. Notably, our stringent <5-kb filtering may underestimate BGC counts (our main-text standard), whereas *Streptomyces* studies typically use unfiltered complete genomes[42,43]. Interestingly, all 76 UBA5704 members originate from soil (59 MAGs) and sediment (17 MAGs). While *Streptomyces* is a well-known BGC-rich lineage predominantly inhabiting soil[44,45], our study reveals UBA5704 represents another highly prolific lineage derived from both soil and sediment environments, indicating that the biosynthetic potential of these habitats is not solely attributable to a few well-known taxa. In saline lake ecosystems, UBA5704 is predominantly distributed across all salinity types in the MXP bioclimatic region (Fig. S6a).

The Candidatus UBA5704 family harbours diverse BGCs, including NRPS, PKS, terpenes, and various RiPPs (e.g., the emerging ranthipeptide and the well-studied lanthipeptide) gene clusters (Fig. S6b). Of note, although ranthipeptide gene clusters are rarely annotated in public databases, they are broadly distributed across members of this family, suggesting that this lineage may represent a rich and previously unrecognized source of ranthipeptide diversity. To systematically evaluate their novelty, we analysed the predicted clusters from four perspectives: global architecture, radical S-adenosylmethionine (rSAM) core enzymes, precursor peptides and evolutionary relationships. BiG-SLiCE distance calculations showed that all predicted ranthipeptide clusters from Candidatus UBA5704 were markedly distant from known entries in MIBiG 4.0 (distance range 0.51–1.06, mean 0.82; distances >0.4 are considered indicative of novelty) (Table S12). Sequence similarity networks (SSN) further revealed negligible connections to known references (Fig. S7a). At the functional-component

level, more than 89 % of the rSAM enzymes exhibited <45 % sequence identity to known counterparts (Table S12), and no significant matches for the precursor peptides were detected in MIBiG 4.0. Phylogenetic analysis of rSAM enzymes placed the vast majority of UBA5704 sequences into several branches distinct from known clusters, substantially expanding the sequence diversity of this family (Fig. S7b). These findings indicate that the overall architecture of these gene clusters is highly novel and that they are likely to encode structurally unprecedented natural products.

To elucidate the physical interaction networks among key components of the ranthipeptide gene cluster, we selected a complete representative BGC (abh_2020_s_bin114_r3_i4_rsl_9, 21.39 kb) from the genome of abh_2020_s_bin114 in saline lake datasets (a member of Candidatus UBA5704) for structural prediction analysis. Its predicted gene arrangement confirmed the presence and colocalisation of core catalytic domains: an rSAM enzyme (ctg9_39), a putative RiPP precursor peptide (ctg9_40), a cognate RiPP recognition element PqqD (RRE, ctg9_38), and accessory genes including an O-antigen ligase Wzy_C (ctg9_34) and a tetratricopeptide repeat protein (TPR_2, ctg9_35), flanked at both ends by genes annotated as transport- or regulation-related (Fig. 5b). The RRE and the precursor peptide lie immediately adjacent to the rSAM enzyme, forming a spatially contiguous module for substrate recognition and modification. Using AlphaFold 3[46], we co-folded the rSAM enzyme, the precursor peptide, and the RRE, yielding a high-confidence ternary complex (ipTM = 0.83, pTM = 0.86; PAE plot showing clear inter-chain interaction signals) (Fig. 5c), suggesting a stable relative conformation among the three components. In this model, PyMOL[47] structural analysis revealed the RRE domain forms close contacts with the N-terminal region of the candidate precursor peptide—specifically, RRE Asp76–precursor Tyr21 (2.6 Å), RRE Glu79–precursor Arg23 (2.7 Å), and RRE Glu34–precursor Arg29 (2.5 Å). These short-range interactions are compatible with hydrogen-bonding or salt-bridge geometries, indicating that the RRE engages the N-terminal leader region for substrate loading. This binding mode recapitulates the canonical RiPP biosynthetic model, wherein an RRE recognizes the leader peptide and assists in positioning the substrate core region for the modifying enzyme. Notably, rSAM Cys251 and precursor Cys52 are positioned approximately 3.1 Å and 3.8 Å from the Fe–S center, respectively, placing the C-terminal Cys52 in close proximity to the catalytic site

and thereby providing structural support for its role as a potential substrate of the rSAM enzyme. Co-folding of the membrane-associated proteins Wzy_C and TPR_2 also produced a high-confidence model (ipTM = 0.90, pTM = 0.85), with the PAE plot supporting stable relative positioning between the two chains (Fig. 5d). This results also revealed multiple charge-complementary close-range interactions across a broad contact surface between the two chains, including Asp368–Arg135 (2.5 Å), Asp143–Arg24 (2.8 Å), Glu139–Arg28 (2.5 Å), and Lys71–Glu89 (3.1 Å). These interactions provide strong computational evidence for the formation of a direct Wzy_C–TPR_2 complex and suggest that TPR_2 may be involved in the recruitment, positioning, or stabilization of nearby proteins.

Taken together, the abh_2020_s_bin114_r3_i4_rsl_9 cluster probably operates in a modular biosynthetic mode: RRE-mediated recognition and loading of the precursor peptide, rSAM enzyme-catalysed core peptide modification, and participation of membrane-associated proteins in subsequent localisation or transport. This model is primarily based on structural predictions and awaits experimental validation through mutagenesis, in vitro enzymatic assays and LC-MS/MS product identification. Nonetheless, these highly novel ranthipeptide BGCs from Candidatus UBA5704 represent ideal targets for heterologous expression and the discovery of new natural products.

Discussion

Over the course of four years, we collected 567 water and sediment samples from 76 representative saline lakes across three major bioclimatic regions (QTP, MXP, and LP regions) in China and constructed the SLMR, the first comprehensive microbiome database systematically covering saline lake ecosystems across China. This resource integrates metagenomic and metatranscriptomic data, accompanied by comprehensive physicochemical parameters, enabling the first large-scale paired survey of water and sediment microbiomes in saline lakes. An interactive online platform developed alongside SLMR integrates environmental data, MAGs and their taxonomic annotations, and BGCs, providing a critical data infrastructure for comparative microbial ecology across regions and habitats.

SLMR reveals exceptionally high taxonomic novelty and ecological distinctiveness of saline lake microbiomes. Their dominant phylum-level composition differs markedly from that

of other aquatic ecosystems (such as freshwater and river), whereas taxa typically less abundant in other aquatic environments are significantly enriched in saline lakes, reflecting a more evenly distributed community structure and unique ecological adaptation strategies. Correspondingly, saline lake microbiomes also harbour substantial biosynthetic potential. Different BGC types exhibit marked habitat preferences between sediments and water samples, reflecting adaptive divergence in microbial metabolic strategies. By introducing an Integrated Salinity Index (ISI), we found that along the salinity gradient, specific BGC types (e.g., PKS and NRPS) are consistently enriched in brackish and mesosaline lakes, a pattern that holds across different geographic regions. Notably, among BGC-rich microbial lineages, the previously uncharacterized order *Multivoradales* exhibits the highest biosynthetic potential. Within this order, the candidate family UBA5704 carries, on average, a large number of BGCs, with some members containing far more BGCs than known high-yielding taxa that *Streptomyces*[42,43]. This family harbours a diverse array of BGCs, among which the ranthipeptide gene clusters are particularly notable for their high novelty in both sequence and architecture, providing promising leads for natural product discovery.

In summary, the establishment of SLMR systematically reveals the unique value of saline lake microbiomes in terms of taxonomic diversity and biosynthetic potential. Looking forward, SLMR will continue to integrate multi-omics data and functional annotations, with the goal of evolving into an open-access platform for global saline lake microbiome research, thereby facilitating the deep exploration and utilization of microbial resources in extreme environments.

References

1. Saccò, M. *et al.* Salt to conserve: a review on the ecology and preservation of hypersaline ecosystems. *Biol. Rev.* **96**, 2828–2850 (2021).
2. Hammer, U.T. *Saline Lake Ecosystems of the World*. (Springer Science & Business Media, 1986).
3. Williams, W.D. Environmental threats to salt lakes and the likely status of inland saline ecosystems in 2025. *Environ. Conserv.* **29**, 154–167 (2002).
4. Meuser, J.E. *et al.* Contrasting patterns of community assembly in the stratified water column of great salt lake, utah. *Microb. Ecol.* **66**, 268–280 (2013).

5. Gunde-Cimerman, N. *et al.* Strategies of adaptation of microorganisms of the three domains of life to high salt concentrations. *FEMS Microbiol Rev* **42**, 353–375 (2018).
6. Casamayor, E.O. *et al.* Microbial biodiversity in saline shallow lakes of the monegros desert, spain. *FEMS Microbiol. Ecol.* **85**, 503–518 (2013).
7. Bring Horvath, E.R. *et al.* Bacterial diversity and chemical ecology of natural product–producing bacteria from great salt lake sediment. *ISME Commun.* **4**, ycae029 (2024).
8. Bekele, G.K. *et al.* Unveiling microbial secondary metabolite biosynthetic gene clusters from alkaline soda lake chitu, in the ethiopian rift valley. *Electron. J. Biotechnol.* **77**, 48–58 (2025).
9. Liu, S.-W. *et al.* Bioprospecting of actinobacterial diversity and antibacterial secondary metabolites from the sediments of four saline lakes on the northern Tibetan Plateau. *Microorganisms* **11**, 2475 (2023).
10. 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).
11. Anantharaman, K. *et al.* Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat. Commun.* **7**, 13219 (2016).
12. Paoli, L. *et al.* Biosynthetic potential of the global ocean microbiome. *Nature* **607**, 111–118 (2022).
13. Chen, J. *et al.* Global marine microbial diversity and its potential in bioprospecting. *Nature* **633**, 371–379 (2024).
14. Cheng, M. *et al.* A genome and gene catalog of the aquatic microbiomes of the Tibetan Plateau. *Nat. Commun.* **15**, 1438 (2024).
15. Borton, M.A. *et al.* A functional microbiome catalogue crowdsourced from north American rivers. *Nature* **637**, 103–112 (2025).
16. Liu, Y. *et al.* A genome and gene catalog of glacier microbiomes. *Nat. Biotechnol.* **40**, 1341–1348 (2022).
17. Garner, R.E. *et al.* A genome catalogue of lake bacterial diversity and its drivers at continental scale. *Nat. Microbiol.* **8**, 1920–1934 (2023).
18. Zhang, Z.F. *et al.* A deep metagenomic atlas of qinghai-xizang plateau lakes reveals their microbial diversity and salinity adaptation mechanisms. *Cell Rep.* **44**, 116483 (2025).
19. Bowers, R.M. *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).
20. Parks, D.H. *et al.* A complete domain-to-species taxonomy for bacteria and archaea. *Nat. Biotechnol.*

38, 1079–1086 (2020).

21. Parks, D.H. *et al.* GTDB release 10: a complete and systematic taxonomy for 715230 bacterial and 17245 archaeal genomes. *Nucleic Acids Res.* **54**, D743–D754 (2026).
22. Nayfach, S. *et al.* A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* **39**, 499–509 (2021).
23. Sunagawa, S. *et al.* Tara oceans: towards global ocean ecosystems biology. *Nat. Rev. Microbiol.* **18**, 428–445 (2020).
24. Ma, B. *et al.* A genomic catalogue of soil microbiomes boosts mining of biodiversity and genetic resources. *Nat. Commun.* **14**, 7318 (2023).
25. Jurdzinski, K.T. *et al.* Large-scale phylogenomics of aquatic bacteria reveal molecular mechanisms for adaptation to salinity. *Sci. Adv.* **9**, eadg2059 (2023).
26. Oren, A. Novel insights into the diversity of halophilic microorganisms and their functioning in hypersaline ecosystems. *npj Biodivers.* **3**, 18 (2024).
27. Jiang, H. *et al.* Microbial diversity in water and sediment of lake chaka, an athalassohaline lake in northwestern China. *Appl. Environ. Microbiol.* **72**, 3832–3845 (2006).
28. Liu, Y.-H. *et al.* Sediment prokaryotic microbial community and potential biogeochemical cycle from saline lakes shaped by habitat. *Microbiol. Res.* **270**, 127342 (2023).
29. Xamxidin, M. *et al.* Metagenomics-assembled genomes reveal microbial metabolic adaptation to athalassohaline environment, the case lake barkol, china. *Front. Microbiol.* **16**, (2025).
30. Xin, G. *et al.* Characterization of bacterial community dynamics dominated by salinity in lakes of the inner mongolian plateau, china. *Front. Microbiol.* **15**, (2024).
31. Wiederkehr, F. *et al.* Coral microbiomes as reservoirs of unknown genomic and biosynthetic diversity. *Nature* **652**, 686–693 (2026).
32. Zdouc, M.M. *et al.* MIBiG 4.0: advancing biosynthetic gene cluster curation through global collaboration. *Nucleic Acids Res.* **53**, D678–D690 (2025).
33. Draisma, A. *et al.* BiG-SCAPE 2.0 and BiG-SLiCE 2.0: scalable, accurate and interactive sequence clustering of metabolic gene clusters. *Nat. Commun.* **17**, 2000 (2026).
34. Alas, I. *et al.* Micromonosporaceae biosynthetic gene cluster diversity highlights the need for broad-spectrum investigations. *Microb. Genomics* **10**, 001167 (2024).
35. Wiegand, S. *et al.* Cultivation and functional characterization of 79 planctomycetes uncovers their unique biology. *Nat. Microbiol.* **5**, 126–140 (2020).

573 36. Kumar, A. *et al.* The prominence of biosynthetic gene clusters in the halophilic bacteria. *Geomicrobiol.*
574 *J.* **42**, 581–589 (2025).
575 37. Gavriilidou, A. *et al.* Compendium of specialized metabolite biosynthetic diversity encoded in bacterial
576 genomes. *Nat. Microbiol.* **7**, 726–735 (2022).
577 38. Richy, E. *et al.* Long-read sequencing sheds light on key bacteria contributing to deadwood
578 decomposition processes. *Environmental Microbiome* **19**, 99 (2024).
579 39. Amiri Moghaddam, J. *et al.* Analysis of the genome and metabolome of marine myxobacteria reveals
580 high potential for biosynthesis of novel specialized metabolites. *Sci. Rep.* **8**, 16600 (2018).
581 40. Caicedo-Montoya, C. *et al.* Pan-genome of the genus streptomyces and prioritization of biosynthetic
582 gene clusters with potential to produce antibiotic compounds. *Front. Microbiol.* **12**, (2021).
583 41. Otani, H. *et al.* Comparative and pangenomic analysis of the genus streptomyces. *Sci. Rep.* **12**, 18909
584 (2022).
585 42. Mohite, O.S. *et al.* Pangenome mining of the streptomyces genus redefines species' biosynthetic
586 potential. *Genome Biol.* **26**, 9 (2025).
587 43. Belknap, K.C. *et al.* Genome mining of biosynthetic and chemotherapeutic gene clusters in streptomyces
588 bacteria. *Sci. Rep.* **10**, 2003 (2020).
589 44. Shepherdson, E.M. *et al. Streptomyces* behavior and competition in the natural environment. *Curr. Opin.*
590 *Microbiol.* **71**, 102257 (2023).
591 45. Deng, Z. *et al.* Multidimensional insights into the biodiversity of streptomyces in soils of China: a pilot
592 study. *Microbiol. Spectrum* **13**, e01692-24 (2025).
593 46. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*
594 **630**, 493–500 (2024).
595 47. Schrödinger, LLC. The PyMOL Molecular Graphics System. (2025).

Figure legends

Fig. 1 | Geographic distribution and sampling design of the Saline Lake Microbiome Resource.

a, Geographic distribution of saline lake sampling sites across China. Sampling sites covered three major saline lake regions, including the Northwest Salt Lake Zone, Qinghai-Tibet Plateau Salt Lake Zone, and Eastern Scattered Salt Lake Zone. The map is overlaid with elevation gradients, and dashed boundaries indicate the distribution range of saline lake zones. Blue circles represent samples with paired metagenomic and environmental-factor data, whereas red triangles indicate samples with both metagenomic, environmental-factor and metatranscriptomic data. Site abbreviations denote individual lakes or sampling locations. **b,** Number of water and sediment samples collected from the three bioclimatic plateau regions, including the Qinghai–Tibet Plateau (QTP), Mongolian–Xinjiang Plateau (MXP), and Loess Plateau (LP). In total, 567 samples were collected, comprising 319 sediment samples and 248 water samples. **c,** Distribution of pH and salinity across samples from QTP, MXP and LP. Colours indicate bioclimatic regions, symbols distinguish sediment and water samples. **d,** Elevation profiles of sampled saline lakes across QTP, MXP and LP, arranged in descending order of elevation. The highest and lowest sampling elevations are represented by Nam Co Lake (4,718 m) and Aydingkol Lake (−154 m), respectively.

Fig. 2 | Construction and phylogenomic characterization of the SLMR genome catalog. **a,** The number of metagenomic samples used for genome reconstruction across three bioclimatic regions and salinity gradient. The 482 metagenomes included samples from the MXP, QTP, and LP, spanning brackish, mesosaline and hypersaline lakes. Lower circles and side bars indicate the numbers of water and sediment metagenomes in each region. **b,** Overview of metagenome assembly, binning and MAG quality assessment. A total of 48,830 MAGs were reconstructed from 482 metagenomes after filtering for completeness ≥50% and contamination <10%. MAG quality was further evaluated using the quality score defined as completeness − 5 × contamination. High-quality MAGs were defined as genomes with completeness >90%, contamination <5%, the presence of 23S, 16S and 5S rRNA genes, and at least 18 tRNAs. **c,** Dereplication and novelty assessment of SLMR MAGs. The 48,830 MAGs were clustered into 24,780 rMAGs using 95% ANI and 30% aligned fraction thresholds. Representative genomes were compared against multiple reference genome catalogs, including TPMC, GEM, TARA, TG2G, GTDB, GROWdb, LakePulse and SMAG. **d,** Phylogenomic distribution of SLMR representative genomes mapped onto the GTDB R226 reference tree. Inner sectors indicate major bacterial and archaeal phyla, while outer rings show habitat type, bioclimatic region, salinity category, MAG abundance, BGC abundance and mapping status against GTDB, GEM and TPMC. MAG, metagenome-assembled genome; rMAG, representative metagenome-assembled genome; ANI, average nucleotide identity; BGC, biosynthetic gene cluster.

Fig. 3 | Habitat-specific distribution and environmental associations of contig-derived biosynthetic gene clusters in saline lake microbiomes.

a, Distribution of biosynthetic gene clusters (BGCs) predicted from metagenomic contigs in sediment and water samples. The lower bar indicates the total number and proportion of BGCs detected in each habitat, and the violin-boxplot shows BGC abundance normalized as the number of BGCs per gigabase of assembled metagenomic contigs. **b,** Associations between environmental variables, spatial factors and contig-derived BGC composition in sediment and water habitats. Pairwise correlations among environmental and spatial variables are shown by the heatmap, and Mantel-test links indicate the strength and significance of correlations between these variables and BGC profiles in sediment and water samples. **c,** Shared and habitat-specific BGC classes between sediment and water samples. The inset shows the relative distribution of major and hybrid BGC classes in sediment and water. **d,** Habitat-specific abundance patterns of representative BGC classes, including NRPS, RiPP and terpene clusters. **e,** Heatmap showing the distribution patterns of major BGC classes and selected metabolite-associated BGC types across samples from different habitats, plateau regions and lake salinity categories. Statistical significance of pairwise differences in b and e was calculated using two-sided Wilcoxon rank-sum tests with false discovery rate correction for multiple comparisons (*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$). BGC, biosynthetic gene cluster; NRPS, non-ribosomal peptide synthetase; RiPP, ribosomally synthesized and post-translationally modified peptide; PKS, polyketide synthase; BIO, WorldClim bioclimatic variables; PCNM, principal coordinates of neighbour matrices.

Fig. 4 | Diversity and novelty of biosynthetic gene clusters in the SLMR genome catalog.

a, Representation of CLAN-level BGC clustering results. The inner tree shows hierarchical clustering based on BGC distances among CLANs. From inner to outer layers, tracks indicate the taxonomic composition of each CLAN based on ln-transformed phylum frequencies, BGC class composition, CLAN distance, CLAN novelty and BGC count. Circle sizes represent the frequencies of different BGC classes within each CLAN. The outer layers indicate, for each CLAN, the number of BGCs, prevalence across samples and the minimum distance to reference BGCs from MIBiG 4.0. CLANs with larger distances to reference BGCs represent more divergent and potentially novel biosynthetic groups. **b,** Comparison of biosynthetic gene cluster family (GCF) richness and overlap between saline lake and other microbiome or natural product resources. BGCs from saline lake were merged with BGC datasets from the Tibetan Plateau Microbial Catalog (TPMC), Ocean Microbiomics Database (OMD), Reef Microbiomics Database (RMD) and MIBiG 4.0. Cross-database BGC clustering was performed using Clust-O-Matic with MMseqs2 at a clustering threshold of T = 0.5 to identify GCFs shared among datasets (main panel). In addition, BGCs from saline lake alone were clustered separately using BiG-SLiCE software to assess the diversity and richness of saline lake-specific GCFs (inset bar charts). **c,** BGC class composition of 47 BGC-rich *Multivoradales* MAGs. BGC, biosynthetic gene cluster; CLAN, gene cluster clan; GCF, gene cluster family; MAG, metagenome-assembled genome; rMAG, representative metagenome-assembled genome; SLMR, Saline Lake Microbiome Resource; TPMC, Tibetan Plateau Microbial Catalog; OMD, Ocean Microbiomics Database; RMD, Reef Microbiomics Database; MIBiG, Minimum Information about a Biosynthetic Gene Cluster; NRPS, non-ribosomal peptide synthetase; PKS, polyketide synthase; RiPP, ribosomally synthesized and post-translationally modified peptide.

Fig. 5 | Phylogenomic distribution of *Multivoradales* and structural analysis of a complete representative ranthipeptide-like biosynthetic gene cluster from its member family UBA5704.

a, Phylogenomic distribution of *Multivoradales*. The circular tree shows SLMR genomes and reference genomes, with outer rings indicating data source, bioclimatic region, lake salinity category, sample source, genome size and BGC count. The outermost family-level annotation highlights BGC-rich lineages assigned to UBA5704, Candidatus *Multivoradaceae* and other families. **b,** Gene organization of a complete representative ranthipeptide-like BGC from the UBA5704 family. Predicted genes and domains include PqqD/RRE, Radical_SAM/SPASM and a putative precursor peptide. **c,** AlphaFold 3-predicted ternary complex of the radical SAM enzyme (ctg9_39), putative precursor peptide (ctg9_40) and RiPP recognition element (ctg9_38). Insets highlight the putative Fe-S cluster region and the RRE-mediated recognition interface with the precursor peptide leader region. **d,** Predicted complex between the Wzy_C-containing protein (ctg9_34) and the TPR-containing protein (ctg9_35). The enlarged region shows the major predicted interaction interface, with putative polar contacts or salt bridges indicated by yellow dashed lines. BGC, biosynthetic gene cluster; GTDB, Genome Taxonomy Database; RRE, RiPP recognition element; RiPP, ribosomally synthesized and post-translationally modified peptide; rSAM, radical S-adenosylmethionine.

Methods

Sample collection. Lake water and sediment samples were collected during the summer-to-autumn transition period each year from 2018 to 2022. A total of 76 saline lakes were surveyed across three plateau regions of China: 34 on the Qinghai–Tibet Plateau (encompassing Qinghai Province and the Xizang Autonomous Region), 28 on the Mongolian–Xinjiang Plateau (including the Xinjiang Uygur Autonomous Region and Inner Mongolia Autonomous Region), and 14 on the Loess Plateau (covering Shanxi Province). In total, 567 samples were collected, comprising 319 sediment samples and 248 water samples (details of sample information in Table S1).

At each lake, three to five sampling sites were established haphazardly, with a minimum distance of 500 m between any two sites to ensure spatial representativeness and minimize pseudoreplication. At each sampling site, 5 L of surface water was collected from a depth of 10 – 50 cm using a sterile water sampler. The collected water was immediately pre-filtered through qualitative filter paper (Whatman, GE Healthcare, NY, USA) to remove aquatic plants, coarse particulates, and other debris. The filtrate was subsequently passed through 0.22 μm polyethersulfone (PES) membranes (Millipore, Billerica, MA, USA) via vacuum filtration to retain microbial biomass. Each membrane was then folded, wrapped in sterile aluminum foil to protect from light, and immediately placed into a sterile vial containing silica gel desiccants for on-site dehydration. Sediment samples (0 – 20 cm depth) were collected using a sterile spatula. To account for small-scale spatial heterogeneity, five subsamples were taken randomly within a 1 m^2 area and thoroughly homogenized to form a single composite sample per site. Visible debris, including stones, plant residues, and macroinvertebrates, were removed aseptically using sterile forceps. Each composite sample was then partitioned into three aliquots for distinct downstream analyses: (i) approximately 10 – 20 g of sediment was transferred into sterile 2 mL cryogenic tubes, immediately flash-frozen in liquid nitrogen, and subsequently stored at –80℃ for DNA extraction; (ii) approximately 100 g of sediment was placed in a sterile zip-lock bag and stored at –20℃ for physicochemical analyses; and (iii) approximately 100 g of sediment was collected in another sterile zip-lock bag and stored at –20℃ for subsequent enrichment cultivation experiments. All sediment samples, together with the membrane filters

collected from water samples, were transported to the laboratory on dry ice and maintained at the designated temperatures until further processing. Throughout the sampling process, all equipment and tools were sterilized between sites with 75% ethanol followed by DNA-free water rinses to prevent cross-contamination.

Measurement of environmental parameters. At each lake, geographic coordinates (longitude, latitude) and altitude were recorded on-site using a handheld GPS device. A comprehensive set of bioclimatic variables for each sampling location was extracted from the WorldClim database (https://www.worldclim.org), including annual mean temperature, isothermality, max temperature of warmest month, mean diurnal range, mean temperature of coldest quarter, mean temperature of driest quarter, mean temperature of warmest quarter, mean temperature of wettest quarter, min temperature of coldest month, temperature annual range, temperature seasonality, precipitation of coldest quarter, precipitation of driest month, precipitation of driest quarter, precipitation of warmest quarter, precipitation of wettest month, precipitation of wettest quarter, and precipitation seasonality.

For physicochemical characterization, sediment samples were first freeze-dried using a laboratory freeze dryer (SCIENTZ-10N, SCIENTZ, Ningbo, China), then homogenized, ground, and passed through a 10 mm sieve to remove coarse particles. Approximately 2 g of the freeze-dried, sieved sediment was suspended in 20 mL of sterile deionized water in a 50 mL Falcon tube and vigorously shaken to ensure thorough mixing. The suspension was centrifuged at 5,000 × g for 5 min, and the resulting supernatant was collected for subsequent analyses. For water samples, no pre-treatment was applied prior to measurement. The pH and salinity of both sediment supernatants and water samples were determined using a portable multi-parameter water quality analyzer (Multi3430, WTW, Germany). Concentrations of major cations (K^+, Al^{3+}, Li^+) and anions (F^-, Cl^-, NO_2^-, Br^-, NO_3^-, SO_4^{2-}, PO_4^{3-}) were quantified by ion chromatography (SH-CC-3L, SHINE, Qingdao, China). Trace metal elements, including Cu, Fe, Mn, Mo, Ni, and Zn, were also analyzed using the same ion chromatography system. Total organic carbon (TOC) and organic matter content in sediments were determined using an Elementar vario TOC select analyzer (Langenselbold, Germany).

DNA extraction and metagenomic sequencing. Microbial DNA was extracted from either filtered membranes or sediment samples with the DNeasy PowerSoil Pro Kit (Qiagen, Hilden, Germany) as per the manufacturer's protocol. DNA purity and concentration were initially assessed using a NanoDrop 2000 Spectrophotometer (Thermo Scientific, MA, USA), and integrity was verified by 1.0% agarose gel electrophoresis. For precise quantification prior to library construction, DNA concentrations were measured with the Qubit 4.0 fluorometer (ThermoFisher Scientific, MA, USA). Paired-end (2 × 150 bp) metagenomic libraries were prepared and sequenced on the Illumina HiSeq 2500 platform at Novogene Bioinformatics Technology Co., Ltd. (Tianjin, China).

Metagenomic assembly and binning. For each metagenomic dataset, raw sequencing reads were trimmed and filtered to remove Illumina adapters, low-quality bases, and duplicate sequences using fastp (v0.20.0)[48] with default parameters. The resulting clean reads from each sample were then assembled individually using metaSPAdes (v3.15.5)[49]. To exploit differential coverage information and enhance the recovery of shared genomes across similar habitats, reads originating from the same lake environment were pooled by merging forward and reverse files separately and subsequently co-assembled with metaSPAdes. Both single-sample and co-assembly-derived contigs longer than 1,000 bp were subjected to binning using BASALT (v1.0)[50] with default settings to recover metagenome-assembled genomes (MAGs). The completeness and contamination levels of each MAG were estimated via CheckM (v1.0.2)[51] employing the "lineage_wf" workflow and default parameters. To further assess genomic integrity, the presence of ribosomal RNA genes (5S, 16S, and 23S rRNA) and transfer RNA genes were detected using Barrnap (v0.9)[52] and tRNAscan-SE (v2.0.9)[53], respectively. Following the Minimum Information about a Metagenome-Assembled Genome (MIMAG) criteria[19], only MAGs classified as medium or high quality were retained for downstream analyses. Specifically, medium-quality MAGs were defined as those with completeness $\geqslant 50\%$ and contamination $< 10\%$, whereas high-quality MAGs required completeness $> 90\%$, contamination $\leq 5\%$, and the presence of all three rRNA genes alongside at least 18 tRNA genes. A quality score was subsequently assigned to each retained genome as completeness – (5 × contamination).

Species-level genome clustering of MAGs and comparison with other ecosystems. The complete set of 48,830 MAGs was clustered using dRep (v3.5.0)[54] based on whole-genome average nucleotide identity (ANI) with a threshold of 95% ANI and >30% aligned fraction, applying parameters of -nc 0.3 and -sa 0.95. This approach follows established practices in microbiome studies across diverse ecosystems. When multiple MAGs were assigned to the same genomospecies, a representative genome was selected according to the highest quality score, serving as a proxy for species membership in the final MAG catalogue and subsequent analyses.

To assess the phylogenetic novelty of the recovered MAGs, the representative genomes were compared against multiple publicly available genome collections using the "compare" module of dRep (v3.5.0). These reference datasets included: 113,104 representative genomes from the Genome Taxonomy Database (GTDB release R226)[20,21]; 23,045 environmentally derived MAGs from the Genomes from Earth's Microbiomes (GEM)[22] project, filtered for water and sediment origins via keyword queries; 10,723 MAGs from the Tibetan Plateau Microbial Catalog (TPMC)[14]; 968 MAGs from the Tibetan Glacier Genome catalog (TG2G)[16]; 957 MAGs from the Tara Oceans expedition[23]; 21,076 MAGs from the Soil Microbial Assembled Genome (SMAG) catalog; 2,093 MAGs from the Genome Resolved Open Watersheds (GROW) dataset; and 1,008 MAGs from the LakePulse MAG catalogue. Based on comparative analyses, MAGs exhibiting ANI values below 0.95 and alignment coverage below 0.3 relative to any reference genome were classified as putative novel species.

Taxonomic classification and phylogenetic analysis. Taxonomic assignment of all MAGs was performed using the GTDB-Tk classify_wf workflow (v2.4.1)[55] with default parameters. MAGs lacking a taxonomic assignment at one or more ranks—namely species, genus, family, or higher—were designated as putative novel lineages at the corresponding levels. For phylogenetic inference, 24,780 representative MAGs and placed into the GTDB-Tk reference tree using maximum-likelihood-based placement implemented in pplacer. The resulting phylogenetic tree was visualized and annotated using the R packages ggtree (v3.10.1)[56] and ggtreeExtra (v1.12.0)[57] in R (v4.3.3).

Integrated Salinity Index (ISI) and salinity grouping. To avoid the potential bias inherent in defining lake ecological niches based solely on water salinity, we introduced sediment porewater salinity as a complementary parameter. For each lake, we defined an Integrated Salinity Index (ISI) as the arithmetic mean of water salinity and sediment porewater salinity:

$$\mathrm{ISI} = \frac{S_w + S_s}{2}$$

Where S_w represents water salinity (g/L) and S_s represents sediment porewater salinity (g/L).

The use of equal weighting (arithmetic mean) was adopted to eliminate potential bias arising from the substantial variation in lake volume and sediment thickness across the studied lakes. This equal-weight approach mitigates the volume effect that could otherwise skew the integrated value toward the larger compartment, thereby enhancing comparability among lakes of different scales[58]. This rationale is consistent with previous studies examining total salt load or salt balance in saline lake systems, where water and sediment are often treated as two major salt pools of comparable importance.

Based on ISI values, the studied lakes were categorized into three salinity groups: brackish (1 g/L < ISI ≤ 5 g/L), mesosaline (5 g/L < ISI ≤ 50 g/L), and hypersaline (ISI > 50 g/L)[59,60]. This classification follows established salinity thresholds for inland saline waters while accounting for the combined influence of water and sediment compartments[61].

Robustness analysis of ISI grouping. To assess whether the core findings of this study are dependent on the definition of the Integrated Salinity Index (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 ≤ 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, mesosaline, and hypersaline) based on the established thresholds (1–5, 5–50, and >50 g/L).

We compared the performance of the two grouping approaches in capturing microbial community patterns using α-diversity (Shannon and Simpson indices), β-diversity (principal coordinate analysis based PERMANOVA), and BGCs abundance, separately for water and

sediment samples. Alpha diversity of microbial communities was calculated at the genus level using the Shannon index. Beta diversity was assessed by principal coordinate analysis (PCoA) based on Bray-Curtis distance matrices, and statistical differences in community structure between groups were rigorously tested using PERMANOVA (with 999 permutations)[62] and ANOSIM[63]. These analyses were selected as they represent the primary descriptors of microbial community structure and functional potential, which are central to the main findings of this study.

Prediction, clustering, and environmental correlation of BGCs at the contigs-level. To mitigate the interference of sequence fragmentation on the prediction of multi-domain enzymes, sequences shorter than 5 kb were filtered out prior to biosynthetic gene cluster (BGC) prediction. Open reading frames were predicted using Prodigal (v2.6.3)[64] in meta mode, followed by BGC identification and classification using antiSMASH (v7.1.0)[65] with the --genefinding-tool prodigal-m parameter. For the antiSMASH outputs, all identified BGCs were precisely mapped to 26 major secondary metabolite classes based on their core biosynthetic enzymes using BiG-SCAPE (v2.0.0)[33,66] with the options --classify category --alignment-mode global --extend-strategy greedy, and subsequently clustered into gene cluster families (GCFs).

To avoid redundancy and multicollinearity among environmental variables, variance inflation factors (VIF) were calculated for each factor using the R package car (v3.1.3). Variables with VIF > 10 were sequentially removed via stepwise regression, retaining a set of core independent factors. Spearman's rank correlation analysis was then performed separately for water and sediment samples to evaluate monotonic correlations between the BGC count of each sample and each retained factor. To control false discovery rates, raw p-values were corrected using the Benjamini–Hochberg (FDR) method, and only corrected p-values < 0.05 were considered significant. To assess how BGC abundance responds to environmental gradients, Mantel tests were performed within each habitat (water and sediment) using the R package linkET (v0.0.7.4)[67].

BGC prediction, clustering, and novelty assessment at the MAGs-level. BGC prediction

(with filtering of sequences <5 kb) and GCF clustering for MAGs were performed following the same procedure as for contigs. To construct a higher-order evolutionary network of MAG-derived BGCs, we further grouped GCFs into gene cluster clans (CLANs) using the BIRCH algorithm with a distance threshold of 0.7[68]. To avoid statistical bias in subsequent quantitative analyses (including taxonomic and functional composition of CLANs/GCFs, and distance calculations between GCFs and reference databases), we applied a strict dereplication step to the 86,579 BGCs identified from MAGs. For each representative BGC, we calculated its distance to reference databases using BiG-SLiCE (v2.0.0)[33,69], which applies a cosine-like distance metric with L2 normalization to ensure balanced and accurate comparisons across diverse BGC classes. The representative distance of a given GCF or CLAN was then defined as the average of the minimum distances of all BGCs within that GCF or CLAN, respectively. For novelty classification, a GCF was considered novel if its representative distance exceeded 0.2, indicating substantial divergence from known reference sequences. For CLANs, a threshold of 0.4 (twice the GCF threshold) was applied to effectively identify and capture novel groups with distant relationships to reference databases. To compare the BGCs from saline lake microbiomes with those from ocean[12], coral[31], Tibetan Plateau aquatic ecosystems[14], and the MIBiG 4.0 database[32], we clustered all BGCs using clust-o-matic with a threshold T of 0.5[37]. The resulting GCFs were enumerated for each dataset, and an UpSet plot[70] was employed to illustrate both the overlaps and the unique BGCs across the datasets.

Enrichment of the order *Multivoradales*. A 40 g aliquot of complex sediment sample was mixed with 160 mL of sterile original environmental water that had been filtered through a 0.05 μm membrane. The mixture was placed in a 250 mL sterilized glass bottle and shaken at 300 rpm and 16°C for 30 min. Subsequently, the suspension was centrifuged at 600 g and 16°C for 1 min, and the supernatant was collected. The supernatant was then supplemented with basal medium, antibiotics, and lysozyme to final concentrations of 200 μg/mL for antibiotics and 3 g/L for lysozyme. The culture was incubated under anaerobic conditions for 4 days, with the headspace gas mixture of N_2:CO_2 (8:2, v/v), at 16°C with shaking at 300 rpm. The basal medium contained Tryptone, Peptone, casein hydrolysate, milk powder, amino acids, glucose, sodium pyruvate, sodium bicarbonate, ammonium chloride, reducing agents, sodium lactate,

and a vitamin solution. One liter of the vitamin solution contained: biotin 4.9 mg, folic acid 8.8 mg, pyridoxine hydrochloride 4.1 mg, thiamine hydrochloride 6.7 mg, riboflavin 7.5 mg, nicotinic acid 2.4 mg, calcium pantothenate 9.5 mg, vitamin B12 0.1 mg, p-aminobenzoic acid 2.7 mg, and lipoic acid 4.1 mg. The antibiotics were selected from one or more of the following: oxytetracycline, erythromycin, kanamycin, streptomycin, vancomycin, ampicillin, ofloxacin, nisin, chloramphenicol, rifampicin, gentamicin sulfate, tetracycline, apramycin sulfate, fosfomycin phenylethylamine, and penicillin, at a concentration of 200 μg/mL.

After enrichment, the culture was filtered through a 0.05 μm membrane to collect microbial cells retained on the filter. Total DNA was extracted from the collected cells and subjected to metagenomic sequencing. The sequencing data were quality-controlled using fastp, assembled with metaSPAdes (v3.15.5)[49] in meta-sensitive mode, and binned using BASALT (v1.0)[50]. The resulting bins were assessed for quality with CheckM (v1.0.2)[51], and only those with completeness ≥ 50% and contamination < 10% were retained. Taxonomic classification was performed using GTDB-Tk (v2.4.1)[55] to identify genomes belonging to the order *Multivoradales*.

Phylogenomics of *Multivoradales*. To comprehensively understand the phylogenomic background of the order *Multivoradales*, we retrieved all publicly available *Multivoradales* genomes from NCBI (As of April 12, 2026) and GTDB (release 226), yielding 296 genomes. After quality assessment, three genomes that did not meet the MIMAG criteria[19] (completeness < 50% or contamination ≥ 10%) were excluded, resulting in 293 median- and high-quality public genomes. Together with 178 genomes recovered from our saline lake metagenomes and 53 genomes from enrichment cultivation, a total of 524 genomes were used for phylogenomic tree construction.

Representative genomes from the phylum *Acidobacteriota* (to which *Multivoradales* belongs) were included as the ingroup, while genomes from *Patescibacteria*, *Chloroflexota*, *Bacillota_A*, *Actinomycetota*, and *Planctomycetota* were selected as outgroups to root the tree. Phylogenetic analysis was performed based on 120 bacterial single-copy conserved marker genes (bac120) extracted using GTDB-Tk. The resulting amino acid multiple sequence alignments were trimmed with trimAl (v1.5.rev0)[71] to remove low-quality columns containing

more than 50% gaps. Maximum-likelihood phylogenetic analysis was conducted on the trimmed concatenated amino acid alignment using IQ-TREE 2 (v2.3.6)[72] with 1,000 ultrafast bootstrap replicates (-bb 1000).

Novelty assessment and phylogenetic analysis of ranthipeptide-like BGCs in UBA5704. To systematically evaluate the diversity and novelty of ranthipeptide-like BGCs in the UBA5704 lineage, we first retrieved all candidate BGCs predicted as ranthipeptide or ranthipeptide-like from the antiSMASH (v7.1.0)[65] annotation results. Novelty assessment was then conducted at two levels: overall gene cluster architecture and protein sequence similarity. At the architecture level, candidate BGCs from UBA5704 were combined with reference BGCs from the MIBiG 4.0 database[32] and clustered using BiG-SLiCE (v2.0.0)[33] to generate a sequence similarity network (SSN) with a distance threshold of 0.3. In this network, nodes represent BGCs and edges indicate detectable similarity between two BGCs under the specified threshold. The connectivity between UBA5704 candidate BGCs and MIBiG reference BGCs was used to assess the degree of architectural resemblance to known clusters. At the protein sequence level, a local MIBiG 4.0 protein database was constructed, and protein sequences from UBA5704 ranthipeptide-like BGCs were used as queries for BLASTP (v2.12.0)[73] searches.

Given that ranthipeptides belong to the rSAM-RiPP class of natural products, in which the radical SAM (rSAM) enzyme typically serves as the core modifying enzyme, we further performed phylogenetic analysis using rSAM enzymes as key marker genes. Protein sequences from UBA5704 candidate BGCs were scanned with HMMER (v3.4)[74] against the rSAM domain profile (PF04055) to obtain high-confidence rSAM enzyme sequences. The same HMMER search procedure was applied to protein sequences from the MIBiG 4.0 reference database to retrieve reference rSAM enzymes. UBA5704-derived rSAM sequences and MIBiG reference sequences were combined, aligned using MAFFT (v7.471)[75], and subjected to maximum-likelihood phylogenetic tree inference using IQ-TREE 2 (v2.3.6).The resulting phylogenetic tree was subsequently visualized and annotated using the Interactive Tree of Life (iTOL v7)[76].

Structural prediction of a complete representative ranthipeptide-like BGC. To investigate potential spatial interactions among key components of a complete representative ranthipeptide-like BGC (abh_2020_s_bin114_r3_i4_rsl_9) from the genome (abh_2020_s_bin114), we performed multimeric structure prediction using AlphaFold 3. Based on antiSMASH (v7.1.0) annotation results, the rSAM modifying enzyme, the putative precursor peptide, and the RiPP recognition element (RRE) protein were selected as core targets. Their protein sequences were submitted to AlphaFold 3[46] for co-folding prediction. Model confidence was evaluated using ipTM, pTM, and predicted aligned error (PAE) matrices, with ipTM serving to assess inter-chain interaction confidence and PAE plots used to evaluate stable predicted spatial contacts between different chains. The predicted models were visualized and analyzed in PyMOL (v3.1.8)[47], focusing on spatial contacts between the RRE and the precursor peptide, the relative positioning of the precursor peptide relative to the rSAM enzyme active site, and inter-residue distances at potential interfaces.

To further explore potential functional associations among adjacent auxiliary proteins, we performed pairwise AlphaFold 3 co-folding predictions on membrane- and scaffold-associated modules. The primary pair analyzed was the Wzy_C-containing protein and the TPR-containing protein. For comparison, the TPR protein was also co-folded with RRE, ABC transporter-associated proteins, and glycosyltransferase-like modules. Interaction confidence for each pair was assessed using ipTM and PAE matrices. For pairs exhibiting high ipTM and PAE profiles supporting stable inter-chain contacts, predicted interfaces were further inspected in PyMOL, and potential polar contacts or salt bridges were measured. Differences in ipTM and PAE signals across pairs were used to infer relative interaction specificity.

Data availability

The raw sequencing data generated in this study have been deposited in the NCBI Sequence Read Archive under BioProject accession PRJNA1494100 (review link: https://dataview.ncbi.nlm.nih.gov/object/PRJNA1494100?reviewer=uei47651e3ik7sgqcr7fjhbele) and in the China National GeneBank Sequence Archive (CNSA) of the China National GeneBank DataBase (CNGBdb) under accession CNP0009831. These records are not yet publicly accessible. The metagenome-assembled genomes have been deposited in Zenodo

(https://zenodo.org/records/22827828), as have the 86,579 MAG-derived and 139,190 contig-derived biosynthetic gene clusters (https://zenodo.org/records/22859136). Files in both Zenodo records are currently under embargo. The processed resources generated in this study, including sample metadata, metagenome-assembled genomes, non-redundant gene sequences, biosynthetic gene clusters and their associated annotations, are available through the SLMR portal (https://slmr.site/).

The reference genome set included 113,104 representative genomes from GTDB release R226[20,21], 23,045 environmental genomes from GEM[22], 10,723 genomes from TPMC[14], 968 genomes from TG2G[16], 957 genomes from the Tara Oceans dataset[23], 2,093 genomes from GROWdb[15], 1,008 genomes from the LakePulse MAG catalogue[17] and 21,076 genomes from SMAG[24]. These datasets are available at https://gtdb.ecogenomic.org/, https://portal.nersc.gov/GEM/genomes/, https://download.cncb.ac.cn/bigd/TPMC/TPMC_genome_catalog/, https://www.biosino.org/node/analysis/detail/OEZ00008946, https://figshare.com/articles/dataset/TARA-NON-REDUNDANT-MAGs/4902923, https://doi.org/10.5281/zenodo.11193259, https://www.ebi.ac.uk/ena/browser/view/PRJEB29238 and https://doi.org/10.5281/zenodo.8223844, respectively. Additional reef microbiome data from the Reef Microbiomics Database are available at https://doi.org/10.5281/zenodo.10182966. The MIBiG and BiG-FAM databases used for biosynthetic gene cluster comparisons are available at https://mibig.secondarymetabolites.org/ and https://bigfam.bioinformatics.nl/, respectively.

References

48. Chen, S. *et al.* fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
49. Nurk, S. *et al.* metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
50. Qiu, Z. *et al.* BASALT refines binning from metagenomic data and increases resolution of genome-resolved metagenomic analysis. *Nat. Commun.* **15**, 2179 (2024).
51. Parks, D.H. *et al.* CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).

52. Seemann, T. tseemann/barrnap. (2026).
53. Chan, P.P. *et al.* tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res.* **49**, 9077–9096 (2021).
54. Olm, M.R. *et al.* 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).
55. Chaumeil, P.-A. *et al.* GTDB-tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* **38**, 5315–5316 (2022).
56. Xu, S. *et al.* Ggtree: a serialized data object for visualization of a phylogenetic tree and annotation data. *iMeta* **1**, e56 (2022).
57. Xu, S. *et al.* ggtreeExtra: compact visualization of richly annotated phylogenetic data. *Mol. Biol. Evol.* **38**, 4039–4042 (2021).
58. Shope, C.L. *et al.* Calculating salt loads to great salt lake and the associated uncertainties for water year 2013; updating a 48year old standard. *Sci. Total Environ.* **536**, 391–405 (2015).
59. Song, C. *et al.* Widespread declines in water salinity of the endorheic tibetan plateau lakes. *Environ. Res. Commun.* **4**, 091002 (2022).
60. Deng, M. *et al.* A non-optically active lake salinity dataset by satellite remote sensing. *Sci. Data* **12**, 1324 (2025).
61. Stumm, W. *et al. Aquatic Chemistry: Chemical Equilibria and Rates in Natural Waters*. (John Wiley & Sons, 2013).
62. Anderson, M.J. Permutational multivariate analysis of variance (PERMANOVA). in *Wiley StatsRef: statistics reference online* (eds Kenett, R.S., Longford, N.T., Piegorsch, W.W. & Ruggeri, F.) 1–15 (Wiley, 2017). doi:10.1002/9781118445112.stat07841.
63. Clarke, K.R. Non-parametric multivariate analyses of changes in community structure. *Aust. J. Ecol.* **18**, 117–143 (1993).
64. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinf.* **11**, 119 (2010).
65. Blin, K. *et al.* antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. *Nucleic Acids Res.* **51**, W46–W50 (2023).
66. Navarro-Muñoz, J.C. *et al.* A computational framework to explore large-scale biosynthetic diversity. *Nat. Chem. Biol.* **16**, 60–68 (2020).

67. Huang, Houyun. linkET: Everything is Linkable. (2021).
68. Zhang, T. *et al.* BIRCH: an efficient data clustering method for very large databases. *ACM SIGMOD Rec.* **25**, 103–114 (1996).
69. Kautsar, S.A. *et al.* BiG-SLiCE: a highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters. *GigaScience* **10**, giaa154 (2021).
70. Conway, J.R. *et al.* UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**, 2938–2940 (2017).
71. Capella-Gutiérrez, S. *et al.* trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
72. 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).
73. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinf.* **10**, 421 (2009).
74. Eddy, S.R. Accelerated profile HMM searches. *PLOS Comput. Biol.* **7**, e1002195 (2011).
75. Katoh, K. *et al.* MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
76. Letunic, I. *et al.* Interactive tree of life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Res.* **52**, W78–W82 (2024).

Acknowledgements

This work was supported by National Natural Science Foundation of China (NSFC) (grant no. 32470697, 32602315, 42477127, 42530701); the Shenzhen Knowledge Innovation Program Basic Research Project (grant no. JCYJ20250604175827037, JCYJ20241202125905008).

Author contributions

K.Y. conceived the study, acquired funding, and revised the manuscript. Y.W. developed the overall research framework and drafted the manuscript. Y.W., R.C., and X.Q. performed data organization and analyses. J.Y.W. and C.L. developed the SLMR web portal. Z.Q., Y.Y., F.W., B.L., B.C., Y.M., L.L., and L.G. collected samples and determined environmental factors. J.J., W.L., S.H., L.Y., M.L., N.X., H.Q., J.N., T.Z., and W.Z. interpreted results and reviewed the manuscript. All authors contributed to editing the text and approved the final version.

Competing interests

All authors declare no competing interests.

Supplementary Files

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

- SupplementaryTables.xlsx
- SupplementaryInformation.pdf
- nrreportingsummary.pdf
- nrsoftwarepolicy.pdf
- BGCmanuscriptcode.zip