




Reconstructing Draft Genomes Using Genome Resolved Metagenomics Reveal Arsenic Metabolizing Genes and Secondary Metabolites in Fresh Water Lake in Eastern India

Bioinformatics and Biology Insights
Volume 15: 1–19
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/11779322211025332


Samrat Ghosh^{1,2} , Aditya Narayan Sarangi¹, Mayuri Mukherjee^{1,2},
Deeksha Singh^{1,2}, Madduluri Madhavi¹ and Sucheta Tripathy^{1,2} 

¹Computational Genomics Lab, Structural Biology and Bioinformatics Division, CSIR Indian Institute of Chemical Biology, Kolkata, India. ²Academy of Scientific and Innovative Research (AcSIR), Ghaziabad, India.

ABSTRACT: Rabindra Sarovar lake is an artificial freshwater lake in the arsenic infested eastern region of India. In this study, using the genome resolved metagenomics approach; we have deciphered the taxonomic diversity as well as the functional insights of the gene pools specific to this region. Initially, a total of 113 Metagenome Assembled Genomes (MAGs) were recovered from the two predominant seasons, that is, rainy (n=50) and winter (n=63). After bin refinement and de-replication, 27 MAGs (18 from Winter season and 9 from Rainy season) were reconstructed. These MAGs were either of high-quality (n=10) or of medium quality (n=17) that was determined based on genome completeness and contamination. These 27 MAGs spanning across 6 bacterial phyla and the most predominant ones were Proteobacteria, Bacteroidetes, and Cyanobacteria regardless of the season. Functional annotation across the MAGs suggested the existence of all known types of arsenic resistance and metabolism genes. Besides, important secondary metabolites such as zoocin_A, prochlorosin, and microcin were also abundantly present in these genomes. The metagenomic study of this lake provides the first insights into the microbiome composition and functional classification of the gene pools in two predominant seasons. The presence of arsenic metabolism and resistance genes in the recovered genomes is a sign of adaptation of the microbes to the arsenic contamination in this region. The presence of secondary metabolite genes in the lake microbiome has several implications including the potential use of these for the pharmaceutical industry.

KEYWORDS: NRPS (Non-ribosomal Peptide Synthases), PKS (Polyketide Synthases), freshwater, genome-resolved metagenomics, Rabindra Sarovar lake, MAGs (Metagenome assembled genomes)

RECEIVED: February 17, 2021. **ACCEPTED:** May 24, 2021.

TYPE: Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: ST would like to acknowledge the Department of Biotechnology (DBT), Govt. of India—Ramalingaswamy fellowship for supporting this work. SG, MM1, and DS would like to acknowledge DBT, CSIR (Council of Scientific and Industrial Research), and ICMR (Indian Council of Medical Research) for fellowship to pursue Ph.D. program. Sequencing was done using funds

from NIPER (National Institute of Pharmaceutical Education and Research), Kolkata, Govt. of India.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Sucheta Tripathy, Computational Genomics Lab, Structural Biology and Bioinformatics Division, CSIR Indian Institute of Chemical Biology, 4 Raja S.C. Mullick Road, Kolkata 700032, India. Emails: tsucheta@gmail.com; tsucheta@iicb.res.in

Introduction

Freshwater lakes are commonly found in low lying areas that are fed from rivers, streams, and overflow from the enclosing sites.¹ Compared to other aquatic environments such as moving water and oceans, freshwater lakes provide a different environment for microbes.² Microbes in such environments carry out several important processes which include sequestering inorganic compounds, decomposing organic matter, mineralizing nitrogen, and so on.³ However, climate change and human activities affect lake ecosystems especially the biogeochemical cycle.⁴ Bacteria are highly sensitive to environmental changes,⁵ therefore change in the environment leads to the change in microbial eco-dynamics.⁴

Rabindra Sarovar lake is an artificial freshwater lake situated at south Kolkata (22.5121°N and 88.3637°E) in the Indian state of West Bengal. The lake area is about 192 acres, of which 73 acres are the water body (Figure 1). It is encircled by Southern Avenue to the north, the Kolkata sub-urban Railway track to the south, Dhakuria to the east, and Shyama Prasad Mukherjee road to the west. This area is in the humid tropical climate zone with a maximum temperature of 40°C

and a minimum temperature of 10°C. Due to its environmental and ecological significance, the status of the lake was upgraded to a national lake. However, because of serious anthropogenic pressures, the lake is witnessing rapid degradation of its ecosystem. In the recent past, unrestricted growth of several aquatic, semi-aquatic species of grasses and sedges are observed in the lake area.⁶ In addition to these, Cyanobacterial blooms are also formed regularly.⁶ It is known that 25% to 75% of Cyanobacterial blooms are harmful, especially in freshwater.^{7,8}

In freshwater lakes, Proteobacteria is found to be the most diverse group of microbes, controlling the water quality and pollution.^{9,10} Members of Phylum Bacteroidetes also colonize in freshwater environments to a large extent. Environmental Bacteroidetes are considered to be specialized in the degradation of polysaccharides and proteins in the biosphere.¹¹ Interestingly, the class Flavobacteria from the Bacteroidetes group contain opportunistic human pathogens, infecting hosts with a low immune system.¹¹ Like Proteobacteria, microbes of the Bacteroidetes group play a crucial role in controlling the water quality.¹⁰



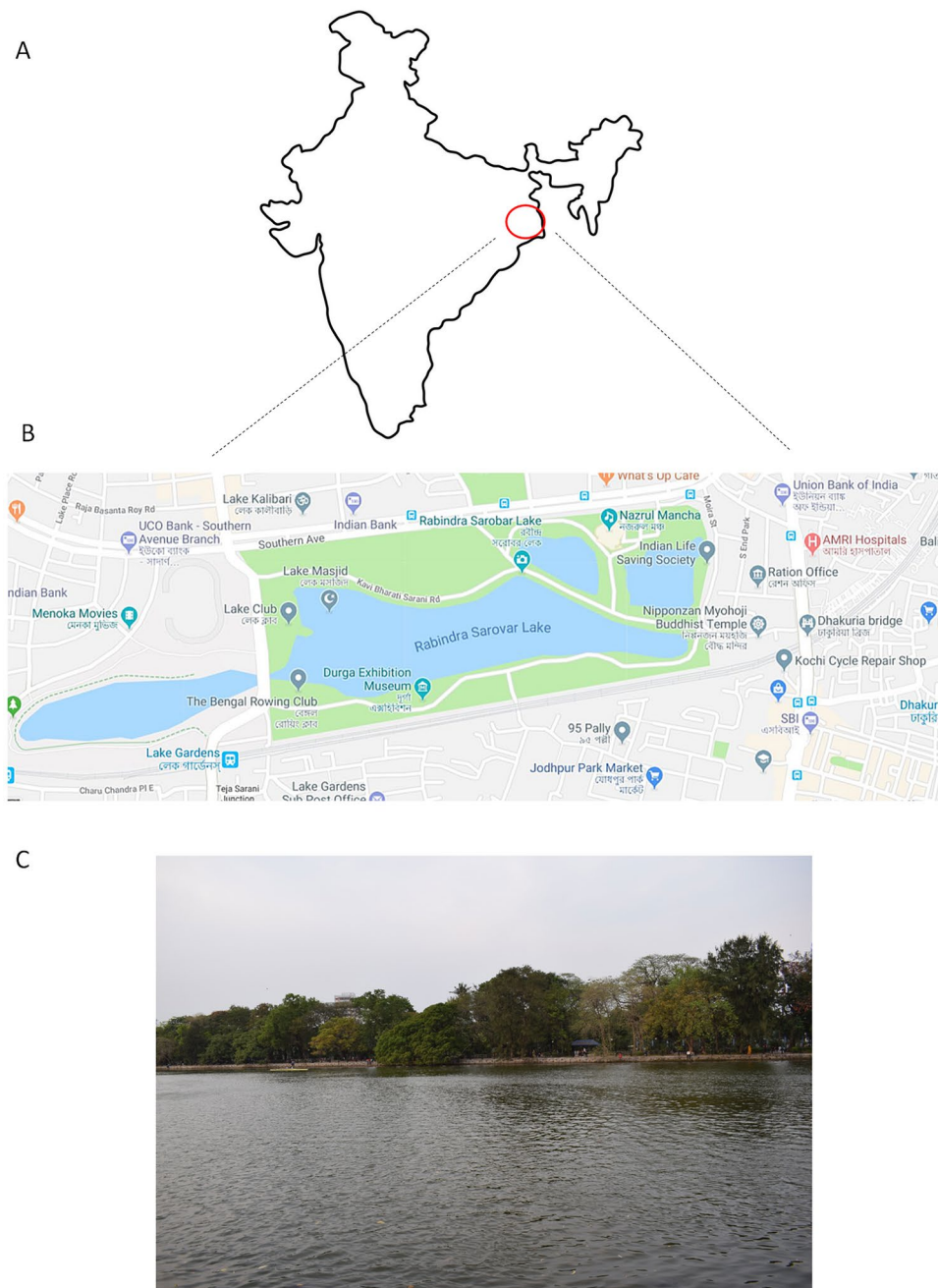


Figure 1. Sampling site of this project including (A) a broad view, (B) google map view and (C) photographic view of the Rabindra Sarovar lake, an artificial freshwater lake located in south Kolkata, West Bengal, India. The photo was taken by Dr. Abhishek Das (ORCID: 0000-0002-1640-1154).

Microbes of aquatic ecosystems possess a broad array of metabolic genes that play a key role in maintaining the biogeochemical cycles.¹² These metabolic genes need to be characterized for a better understanding of their roles in biogeochemical cycles.¹³ For example, the function of arsenic resistance and metabolism-related genes are well-characterized that have substantial effects on biogeochemistry.¹⁴ Arsenic resistant microbes are reportedly widespread in the environment and are major contaminants of the aquatic ecosystems worldwide.¹⁵ Organisms resistant to arsenic have reported from the sites even with low arsenic concentrations (<7 ppm).¹⁶

Lake microbial communities are also known for fertile sources of secondary metabolites such as non-ribosomal peptide synthases (NRPS) and polyketide synthases (PKS). NRPS and PKS are two families of modular mega-synthases, used to treat cancer, infectious diseases, inflammation, and many other diseases.¹⁷ Both mega-synthases families producing peptides (using aminoacyl monomers) and polyketides (using acyl-coA monomers) are present in this ecosystem and are found in many taxonomical groups vary from bacteria (Proteobacteria, Cyanobacteria, Actinobacteria) to fungi.^{18,19} Secondary metabolites such as bacteriocins are often seen as part of an elaborate chemical defense system²⁰

produced in all major lineages of bacteria.²¹ These are obtained from short ribosomally produced precursor proteins that comprised of a conserved N-terminal leader sequence and a C-terminal core peptide. The bacteriocins are carefully analyzed in this study since they are powerful potential alternatives to antibiotics and food preservatives.^{22–24}

Our hypothesis is: in a highly urbanized area where arsenic and other heavy metal contamination is a common occurrence, the microbial composition and dynamics must be reflected in terms of metabolism of such ingredients by the microbial population. Furthermore, seasonal changes of these microbial composition is a perfect indicator of changed microbial dynamics due to physiological and environmental conditions. In this study, we have discussed the genomic features of the microbial genomes reconstructed from the lake ecosystem using a genome-resolved metagenomics approach. Screening of arsenic metabolism and resistance genes among the microbes help to understand their impact on local biogeochemistry and levels of contamination. Our analysis reveals strong presence of well-characterised arsenic resistance and metabolism-related genes among the metagenome assembled genomes (MAGs), and is a possible sign of arsenic stabilization and neutralization by the identified microbes in this region. Genes involved in secondary metabolite production could be explored for their industrial applications and potential use for environmental bio-indicators. Using the current approach, we have identified several important classes of secondary metabolites including NRPS, PKS, and bacteriocins across the MAGs, indicating the lake is a rich source of natural products.

Materials and Methods

Sampling, DNA isolation, and sequencing

Two liters of surface water samples were collected from the middle of the lake and were transported to the lab for further processing. Collections were made at 6 months interval, that is, during Winter season (January 2015) and Rainy season (July 2015). Simultaneously, the temperature and pH of lake water were also measured.

Water filtration was carried out using the Millipore filtration unit attached to a vacuum pump. Each of the water samples (sample volume of 2 L) was immediately filtered through a 1.2 µm pore size membrane filter to remove pollutants. The filtrate was then passed through a 0.2 µm pore size membrane filter to capture the cells on the filter. The filter membranes were cut into pieces and transferred to a 15 ml falcon tube containing 1 ml lysis buffer. The lysis buffer containing tubes were stored at -20°C over the weekend for extraction of DNA. DNA was extracted following the protocol from Oh et al.²⁵ Briefly, cells were treated with lysozyme and incubated at 37°C for 30 minutes. Sodium dodecyl sulfate (SDS) was added and then the sample was incubated at 55°C in a water bath for 4 hours. Phenol chloroform extraction was carried out and the upper phase carrying DNA was

precipitated with 100% cold ethanol. Precipitated DNA was stored in -20°C for overnight. Subsequently, the DNA pellet was washed with 75% ethanol. The pellet was air-dried and dissolved in 20 µl autoclaved Milli-Q water.

Approx 3 µg of genomic DNA was sheared using Covaris tubes (#SN001887) to obtain 300 base-pairs to 500 base-pairs fragment sizes. The size distribution was checked by running an aliquot of the sample on High Sensitivity Bioanalyzer Chip (Agilent, #5067-4626). Sample cleanup was done using HighPrep PCR clean up system (Magbio#AC-60050). A series of enzymatic reactions were carried out using the NEXTflex DNA Sequencing kit that involves overhang addition and adaptor ligation. After ligation, ~300–600bp fragments were size selected on a 2% agarose gel and cleaned using the MinElute column (QIAGEN, #28604 or 28606). The library thus prepared was quantified using Qubit fluorometer (Invitrogen, Cat # Q32854) and quality validated by running an aliquot on High Sensitivity Bioanalyzer Chip (Agilent, #5067-4626). The QC qualified metagenomic DNA library sample was sequenced (150bp paired-end) using the Illumina Miseq platform. The whole procedure (Library preparation & Sequencing) was performed at the Next-Generation Sequencing (NGS) facility, Genotypic Technology Pvt Ltd, Karnataka, India.

Quality control of reads

Quality of raw paired-end reads obtained from the two samples (Winter and Rainy) were checked using FastQC.²⁶ Adapters, poor-quality reads ($Q < 30$), and reads containing ambiguous bases were discarded using fastp (v.0.19.1)²⁷ with default parameters; except for `—qualified_quality_phred` and `—n_base_limit` parameters where they were set at 30 and 0 respectively.

Statistical analysis of metagenome data sets to check seasonal effects on lake microbes

After the initial quality check, unassembled reads were uploaded into the MG-RAST server²⁸ for taxonomic and functional profiling. To check seasonal variation in lake microbes at a taxonomic and functional level, statistical analysis was performed using the MG-RAST server²⁸ followed by STAMP analysis (Statistical Analysis of Metagenomic Profiles) v 2.1.3.²⁹ Statistically significant taxonomic and functional differences between Winter and Rainy season samples were calculated using the Fisher exact test with Benjamini-Hochberg FDR correction. The Newcombe–Wilson test was performed to calculate confidence intervals (CIs).

Read assembly, binning and evaluation

Simultaneously metaSPAdes assembler (v3.12)^{30,31} was used for the assembly of Q30 metagenomic reads. The Binning of the assembled reads was done using MetaBAT (v.2.12.1).³² During both assembly and binning, all the parameters of the

Table 1. Pre- and post-QC read counts and assembly size after assembly with metaSPAdes.

SAMPLE NAME	RAW-READS (FILE SIZE IN GB)	Q30-READS (FILE SIZE IN GB)	ASSEMBLED-READS (FILE SIZE IN GB)
SRR8209851/WINTER	53,902,467 × 2 (38)	37659495 × 2 (24)	2230279 scaffolds (1.2)
SRR8209852/RAINY	39523695 × 2 (30)	24195845 × 2 (15.2)	2738321 scaffolds (1.2)

tools were kept as default. CheckM (v.1.0.11)³³ was used to estimate the genome completeness and contamination of the recovered genome bins.

Bin refinement, de-replication, evaluation, coverage calculation, taxonomic assignment, 16S gene detection, and functional annotation

RefineM (v.0.0.3)³⁴ was used to improve the completeness of each bin. Duplicate bins were removed using dRep.³⁵ The de-replicated refined bins were further evaluated using CheckM (v.1.0.11)³³ and selected as per stated standards.³⁶ The coverage of each bin was calculated using the coverage and profile function of CheckM (v.1.0.11).³³

Taxonomy was assigned to the final improved bins using CheckM (v.1.0.11),³³ MiGA,³⁷ and GTDBtk v.0.2.1³⁸ tools. After the taxonomy assignment, 16S genes of each de-replicated genomes were detected using the *ssu_finder* function of CheckM (v.1.0.11).³³

The taxonomically assigned refined bins were annotated using both Prokka v1.13³⁹ and RAST (Rapid Annotation using Subsystem Technology)⁴⁰ software. COG annotations were carried out with the STRING v10.5 database⁴¹ (only top hits were taken). Pathways were annotated using GhostKOALA.⁴²

Average nucleotide identity and average amino acid identity calculation of the refined bins

All pair-wise ANI (Average Nucleotide Identity) and AAI (Average Amino Acid Identity) values of the selected bins were calculated with “pyani”⁴³ and “comparem”⁴⁴ python packages respectively. Finally, results were visualized with R package “pheatmap” using R studio.⁴⁵

Construction of phylogenomic tree to infer taxonomy of the refined bins

The taxonomy of the refined bins (n = 27) was further confirmed by placing them in a phylogenomic context. For this purpose, we retrieved 827 representative complete bacterial genomes from the RefSeq. The phylogenomic tree was constructed by UBCG v 3.0⁴⁶ pipeline where all the default parameters were used. The constructed tree was edited using iTOLv4.⁴⁷

Comparison of lake water sample bins

We compared the genome bins among themselves to get a deeper insight into their composition. The profiling of

carbohydrate activating enzyme families was done using the stand-alone dbCAN⁴⁸ against the CAZy database.⁴⁸ In-addition, bacteriocins, CRISPRs, prophages, and ISs, were screened using BAGEL4, a web-based resource,⁴⁹ CRISPRfinder stand-alone tool,⁵⁰ PHASTER webserver⁵¹ and ISSaga web platform,⁵² respectively. The presence of arsenic resistance genes among the genome bins was confirmed by using blastp⁵³ against already known publicly available arsenic resistance databases.¹⁴ Transporter related genes were determined using the TCDB (The Transporter Classification Database) database.⁵⁴ Secondary metabolites were identified using Anti-SMASH⁵⁵ (detection strictness parameter kept default, that is, “relaxed” and extra features parameter kept “all on”) and NAPDOS.⁵⁶

Data availability

Merged paired-end reads with annotations are publicly available in the MG-RAST server under the following IDs: mgm4875225.3 (Rainy season), mgm4875226.3 (Winter season). Assembled data, recovered genome bins (n = 27) along with their annotations (protein files) as well as phylogenetic trees are available at <https://ndownloader.figureshare.com/articles/8058902/versions/6>. Apart from the 27 refined bins, the rest of the bins (n = 86) are available at <https://ndownloader.figureshare.com/articles/8796320/versions/3>. Due to the poor quality of these bins (n = 86), they are not used for further downstream analysis and are not discussed in this paper. Besides, the raw fastq files were deposited in the NCBI SRA database with accession numbers SRR8209851 and SRR8209852.

Results

Environmental conditions of sampling site and sequencing of metagenome samples

Sampling site environmental conditions (temperature and pH) differences were small. During sample collection, the temperature of the lake was 21°C (Winter season) and 31°C (Rainy season) whereas pH was 8.5 (Winter season) and 8 (Rainy season).

Whole-genome shotgun sequencing of Winter and Rainy season samples (microbial DNA of surface water) using Illumina produced 53,902,467 × 2 and 39,523,695 × 2 reads respectively. After initial quality check (QC), the high quality (Q30) reads (with read length 151 bp) from the Winter season (37659495 × 2) and Rainy season (24195845 × 2) were obtained (Table 1).

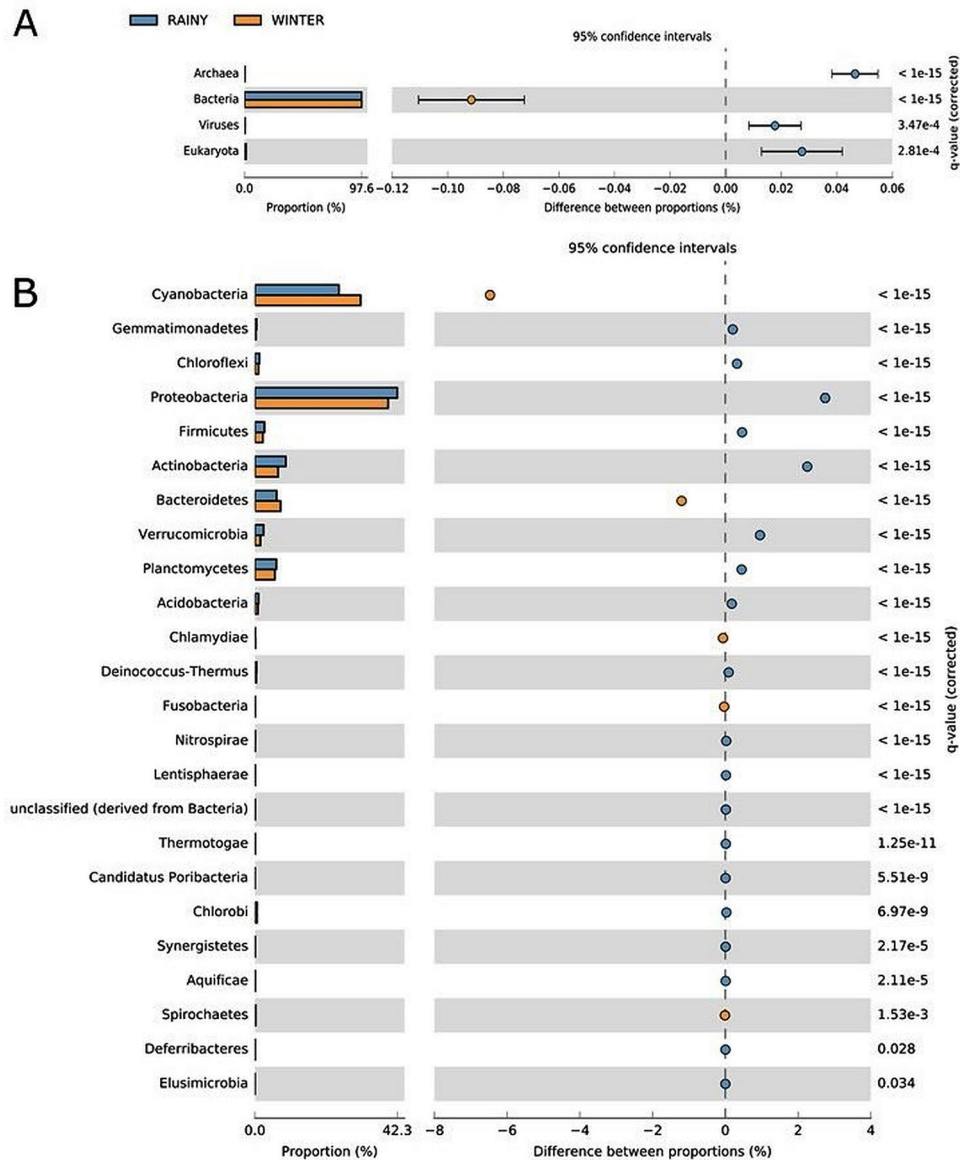


Figure 2. Taxonomic comparison of Winter season (yellow) and Rainy season (blue) microbial communities by Refseq (A) at the domain level and (B) phylum level using STAMP. STAMP indicates Statistical Analysis of Metagenomic Profiles.

Comparison between the winter and rainy season samples suggests a very subtle difference in microbial composition

To get a comparative account of seasonal variation in lake microbial community at the taxonomic and functional level, high quality (Q30) reads from Rainy and Winter season were uploaded to the MG-RAST server. Later the outputs from the MG-RAST platform were statistically validated using STAMP software. Bacteria appeared to be the most dominant kingdom across both the seasons (Figure 2A). Among the bacterial phyla, Proteobacteria was found to be the phylum with the highest proportion detected in both the seasons followed by Cyanobacteria, Actinobacteria, Bacteroidetes, Planctomycetes, Verrucomicrobia, and Firmicutes (Figure 2B). Phylum Proteobacteria, Actinobacteria, Planctomycetes, Verrucomicrobia and Firmicutes were significantly enriched

in the Rainy season while phylum Cyanobacteria and Bacteroidetes were enriched in the Winter season (Figure 2B). Pathogenic bacteria such as *Bordetella*, *Microcystis*, *Burkholderia*, *Mycobacterium*, *Aeromonas*, *Vibrio*, and *Escherichia* were found in both the seasons. (Supplementary Table S1).

Functional annotation at level 1 revealed that carbohydrates, amino acids, and derivatives related genes were highly dominant in the lake microbial communities and these genes were slightly higher in the Rainy season than in the Winter season (Figure 3A). Interestingly, mobile genetic elements such as phage, prophage, and plasmids were found to be enriched in the Rainy season compared to the Winter season (Figure 3A). Annotation at level 2 and level 3 also shown that there was an enrichment of mobile genetic elements in the Rainy season (Supplementary Table S1). In addition to mobile genetic

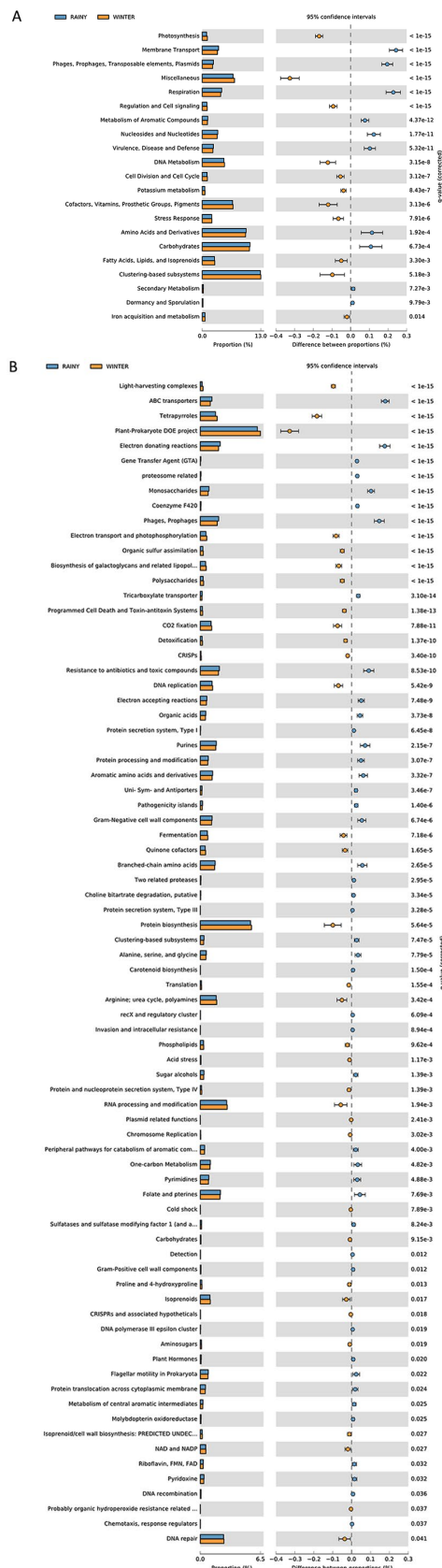


Figure 3. Functional group comparison of Winter season (yellow) and Rainy season (blue) microbial communities by SEED subsystems (A) at level 1 and (B) level 2 using STAMP. STAMP indicates Statistical Analysis of Metagenomic Profiles.

elements, antibiotic resistance genes were also found enriched in Rainy season (Figure 3B and **Supplementary Table S1**).

Bacteria remains the predominant phyla in MAGs with arsenic methylating genes in abundance

Assembly of Q30 reads generated a total of 2,230,279 and 2,738,321 scaffolds for Winter and Rainy season samples respectively (Table 1).

To reconstruct quality single genomes from assembled metagenome, we used the MetaBAT2 binning tool which produced a total of 113 genome bins (a.k.a Metagenome Assembled Genomes or MAGs). Among the 113 bins, 63 bins were from the Winter season and the remaining 50 bins were from Rainy season samples (**Supplementary Table S2: CheckM of all bins**). After successful refinement, de-replication, and assessment, out of 113 bins, 27 bins (18 from Winter and 9 from Rainy season) were selected according to the defined genome standards (high-quality draft [$> 90\%$ complete, $< 5\%$ contamination], medium-quality draft [$> 50\%$ complete, $< 10\%$ contamination] or low-quality draft [$< 50\%$ complete, $< 10\%$ contamination]).³⁶ Based on these standards, 10 MAGs were classified as high quality and 17 were categorized as a medium quality draft (Table 2). Other genome quality assessment parameters such as N50, L50, and GC% were calculated for these MAGs and were given in Table 2 and **Supplementary Table S3: Quast**. Also, information regarding computed coverage of reconstructed bins and bins with identified 16S rRNA can be found in **Supplementary Table S3: Coverage of all bins, Refined bins coverage, and SSU information**.

Taxonomic ranks of all the 27 MAGs were determined using CheckM, MiGA, and GTDB-tk platform. MAGs were taxonomically inferred as bacteria spanning across 7 phyla, that is, Proteobacteria, Cyanobacteria, Bacteroidetes, Verrucomicrobia, Planctomycetes, Gemmatimonadetes, and Chlamydia (Table 3 and **Supplementary Table S2: Refined bins taxonomy**). Among these 27 MAGs, 10 represent Proteobacteria, 8 Bacteroidetes, 4 Cyanobacteria, 2 Planctomycetes, and 1 each Verrucomicrobia, Gemmatimonadetes, and Chlamydia (Table 3). We further investigated whether MAGs of the same phylum were part of the same species or not. This was confirmed by average nucleotide identity (ANI) and average amino-acid identity (AAI) analysis. ANI and AAI values of bin_winter_40_filtered *vs* bin_rainy_3_filtered as well as bin_winter_42_filtered *vs* bin_rainy_48_filtered were found greater than 95% (**Supplementary Table S4 and Supplementary Figure S1**). ANI and AAI values above 95 % are considered as a threshold for assigning genomes to the same species.^{57,58}

Initial annotation of each of the MAGs was carried out using Prokka. All predicted genes of 27 bins were cataloged in **Supplementary Table S5**. Protein coding genes involved in the arsenic biogeochemical cycle were identified. The

Table 2. Characteristics of recovered genome bins based on CheckM and QUAST analysis.

BIN NAME	COMPLETENESS (%)	CONTAMINATION (%)	DRAFT GENOME QUALITY	N50	L50	NO. OF SCAFFOLDS	SIZE (MB)	GC %
bin_winter_42_filtered	99.15	0.45	High	81344	9	40	2.25	53.71
bin_winter_49_filtered	92.61	0	High	23814	37	189	3.10	51.38
bin_winter_31_filtered	92.35	0.55	High	31412	21	128	2.44	57.68
bin_winter_40_filtered	91.37	1.08	High	20066	40	181	2.59	68.92
bin_winter_6_filtered	90.8	0.47	High	13054	86	348	3.49	41.95
bin_winter_53_filtered	90.46	3.94	High	20859	58	295	3.91	46.1
bin_winter_34_filtered	90.32	1.08	High	14315	64	264	2.86	54.28
bin_winter_45_filtered	81.69	1.11	Medium	9566	106	427	3.23	43.38
bin_winter_59_filtered	80.41	3.04	Medium	7842	69	256	1.64	44.27
bin_winter_55_filtered	76.53	4.54	Medium	6044	245	817	4.47	52.08
bin_winter_33_filtered	68.38	0.6	Medium	6163	118	391	2.19	40.34
bin_winter_46_filtered	64.42	1.15	Medium	12520	60	207	2.22	68.56
bin_winter_7_filtered	64.2	0.16	Medium	7041	56	196	1.14	59.48
bin_winter_4_filtered	64	0.08	Medium	5225	82	255	1.26	35.58
bin_winter_22_filtered	63.79	1.72	Medium	71625	9	39	1.98	71.05
bin_winter_1_filtered	60.08	0.48	Medium	5052	127	401	1.89	40.18
bin_winter_23_filtered	54.9	3.15	Medium	4098	145	416	1.70	41.51
bin_winter_24_filtered	51.27	0	Medium	4576	162	494	2.20	43.15
bin_rainy_49_filtered	93.6	4.48	High	11199	68	289	2.46	69.22
bin_rainy_34_filtered	91.85	0.41	High	9550	158	579	4.39	56.22
bin_rainy_48_filtered	91.52	0.88	High	8316	78	276	1.88	53.96
bin_rainy_3_filtered	88.81	0.16	Medium	13245	57	255	2.49	68.96
bin_rainy_23_filtered	78.89	2.1	Medium	7634	57	212	1.45	47.54
bin_rainy_9_filtered	57.05	0	Medium	12329	64	450	2.37	72.97
bin_rainy_27_filtered	54	0.48	Medium	4717	74	232	1.07	57.54
bin_rainy_35_filtered	53.45	0	Medium	13596	63	242	2.66	70.77
bin_rainy_7_filtered	50.85	1.08	Medium	4694	146	450	2.03	66.25

annotations were confirmed by running a blast against publicly available known arsenic-related genes database.¹⁴ Generally, nine genes, that is, *acr3*, *aioA*, *arrA*, *arsB*, *arsC_glut*, *arsC_thio*, *arsD*, *arsM*, and *arxA* are considered as important players for arsenic resistance and metabolism in microbes.^{59,14} Out of these nine arsenic linked genes eight genes were found to be present across the recovered MAGs/bins ($n=27$) and only missing gene was *arsD* (**Supplementary Table S6**). With the help of the existing knowledge on arsenic resistance and metabolism-related genes functions in microbes,^{14,59} we made a hypothetical map of arsenic metabolism (Figure 4).

Importantly, irrespective of the bin quality, higher occurrences of *arsM* gene were predicted in the genome bins (Figure 5). Arsenic detoxification genes (*acr3*, *arsB*, *arsC_glut*, *arsC_thio*) were less abundant than arsenic metabolism genes (*aioA*, *arrA*, *arsM*, and *arxA*) (Figure 5).

Annotation of transporter genes was carried out using data from TCDB (Transporter Classification Database) and cataloged in **supplementary Table S7**. The functional COG and RAST categories distribution were illustrated in heatmaps (Figures 6 and 7) and respective values were provided in table **Supplementary Table S8**.

Table 3. Taxonomy of recovered genome bins based on CheckM, MiGA, and GTDB-tk analysis.

BIN NAME	CHECKM TAXONOMY		MIGA TAXONOMY		GTDB-TK TAXONOMY	
	ASSIGNED TAXON	TAXON LEVEL	ASSIGNED TAXON	TAXON LEVEL	ASSIGNED TAXON	TAXON LEVEL
bin_winter_42_filtered	Betaproteobacteria	Class	Burkholderiaceae	Family	Burkholderiaceae	Family
bin_winter_49_filtered	Bacteroidetes	Phylum	Flavobacteriia	Class	Chitinophagales	Order
bin_winter_31_filtered	Bacteria	Kingdom	Bacteroidetes	Phylum	Kapabacteriaceae	Family
bin_winter_40_filtered	Alphaproteobacteria	Class	Caulobacteraceae	Family	Phenylobacterium	Genus
bin_winter_6_filtered	Cyanobacteria	Phylum	Cyanobacteria	Phylum	Pseudanabaena	Genus
bin_winter_53_filtered	Cytophagales	Order	Cytophagia	Class	Cyclobacteriaceae	Family
bin_winter_34_filtered	Bacteria	Kingdom	Flavobacteria	Class	Flavobacteriales	Order
bin_winter_45_filtered	Bacteria	Kingdom	Verrucomicrobia	Phylum	Kapabacteriales	Order
bin_winter_59_filtered	Bacteria	Kingdom	Chlamydia	Class	Parachlamydiales	Order
bin_winter_55_filtered	Bacteria	Kingdom	Planctomycetes	Class	Pirellulaceae	Family
bin_winter_33_filtered	Bacteroidetes	Phylum	Chitinophagaceae	Family	Chitinophagaceae	Family
bin_winter_46_filtered	Bacteria	Kingdom	Bacteria	Kingdom	Pirellulales	Order
bin_winter_7_filtered	Proteobacteria	Phylum	Betaproteobacteria	Class	Burkholderiaceae	Family
bin_winter_4_filtered	Bacteria	Kingdom	Betaproteobacteria	Class	Caenarcarniphilales	Order
bin_winter_22_filtered	Bacteria	Kingdom	Rhodanobacteraceae	Family	Ahniellaceae	Family
bin_winter_1_filtered	Bacteroidetes	Phylum	Sphingobacteriia	Class	Bacteroidia	Class
bin_winter_23_filtered	Bacteroidetes	Phylum	Chitinophagaceae	Family	Chitinophagales	Order
bin_winter_24_filtered	Cyanobacteria	Phylum	Synechococcales	Order	Microcystis	Genus
bin_rainy_49_filtered	Cyanobacteria	Phylum	Synechococcaceae	Family	Cyanobiaceae	Family
bin_rainy_34_filtered	Cyanobacteria	Phylum	Cyanobacteria	Phylum	Prochlorotrichaceae	Family
bin_rainy_48_filtered	Betaproteobacteria	Class	Burkholderiaceae	Family	Burkholderiaceae	Family
bin_rainy_3_filtered	Alphaproteobacteria	Class	Caulobacteraceae	Family	Phenylobacterium	Genus
bin_rainy_23_filtered	Burkholderiales	Order	Burkholderiaceae	Family	Polynucleobacter	Genus
bin_rainy_9_filtered	Bacteria	Kingdom	Gemmatimonadetes	Class	Gemmatimonadaceae	Family
bin_rainy_27_filtered	Betaproteobacteria	Class	Betaproteobacteria	Class	Burkholderiaceae	Family
bin_rainy_35_filtered	Bacteria	Kingdom	Acetobacteraceae	Family	Roseomonas	Genus
bin_rainy_7_filtered	Bacteria	Kingdom	Flavobacteria	Class	Flavobacteriales	Order

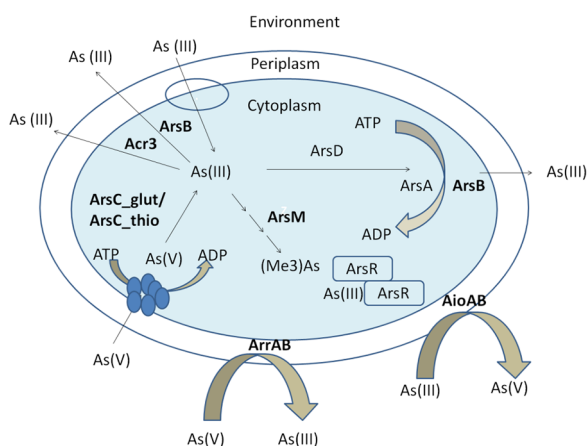


Figure 4. A schematic overview of the arsenic metabolism and resistant related genes common pathways present in lake microbes. The presence of the genes is shown in bold black.

Pathway annotation using the Ghost-KOALA platform identified 80 complete KEGG pathway modules from all the 27 combined MAGs. Completeness of a pathway is demonstrated by including all essential components (gene blocks) to complete a metabolic cycle. Among the 80 modules, central carbohydrate metabolism ($n=13$), ATP synthesis ($n=8$), cofactor-vitamin metabolism ($n=6$), and carbon fixation ($n=6$) related modules were predominant (**Supplementary Table S9**). Central carbohydrate metabolism pathway module M00005 (PRPP biosynthesis, ribose 5P= \Rightarrow PRPP) and fatty acid metabolism module M00086 (beta-oxidation, acyl-CoA synthesis) were found to be present in most of the MAGs ($n=20$) (**Supplementary Table S9**). Complete pathway modules of sulfur metabolism such as M00595 (Thiosulfate oxidation by SOX complex, thiosulfate= \Rightarrow sulfate) and M00176 (Assimilatory sulfate reduction, sulfate= \Rightarrow H₂S) were found in the lake (**Supplementary Table S9**). Secondary metabolism-related pathway modules, that is, M00096 (C5 isoprenoid biosynthesis, non-mevalonate pathway), M00364 (C10-C20 isoprenoid biosynthesis, bacteria), and M00793 (dTDP-L-rhamnose biosynthesis) were also identified (**Supplementary Table S9**). The abundance of ABC transporter and two-component system-related genes under the “environmental information processing” category was found to be high in all bins ($n=27$) (**Supplementary Table S9**). Importantly, under “xenobiotics biodegradation and metabolism” category benzoate degradation linked genes abundance was comparatively high (**Supplementary Table S9**).

Phylogenomic analysis placed metagenome-assembled genomes into six bacterial phyla

Phylogenomic analysis placed MAGs into the six bacterial phyla (Figure 8). Among the recovered MAGs ($n=27$) nine were placed in the Proteobacteria (bin_rainy_3_filtered, bin_rainy_35_filtered, bin_winter_40_filtered, bin_rainy_

23_filtered, bin_rainy_27_filtered, bin_rainy_48_filtered, bin_winter_7_filtered, bin_winter_42_filtered and bin_winter_22_filtered); nine in Bacteroidetes (bin_winter_1_filtered, bin_winter_23_filtered, bin_winter_31_filtered, bin_winter_33_filtered, bin_winter_34_filtered, bin_winter_45_filtered, bin_winter_49_filtered, bin_winter_53_filtered and bin_rainy_7_filtered); five in Cyanobacteria (bin_rainy_49_filtered, bin_rainy_34_filtered, bin_winter_24_filtered, bin_winter_6_filtered and bin_winter_4_filtered); two in Planctomycetes (bin_winter_46_filtered and bin_winter_55_filtered) and one each in Verrucomicrobiota (bin_winter_59_filtered) and in Gemmatimonadetes (bin_rainy_9_filtered) (Figure 8). Within the Proteobacteria, phylogenomic analysis placed one MAG in gamma-Proteobacteria (bin_winter_22_filtered), three in alpha-Proteobacteria (bin_rainy_3_filtered, bin_rainy_35_filtered and bin_winter_40_filtered) and five in beta-Proteobacteria (bin_rainy_23_filtered, bin_rainy_27_filtered, bin_rainy_48_filtered, bin_winter_7_filtered and bin_winter_42_filtered) (Figure 8). Importantly, all these five MAGs of beta-Proteobacteria group were part of Burkholderiaceae family (Table 3). Further analysis with MiGA suggested that there were three ambiguities e.g.; bin_winter_59_filtered was assigned to Chlamydia (**Supplementary Table S2: Refined bins taxonomy**); bin_winter_45_filtered was assigned to Verrucomicrobia (**Supplementary Table S2: Refined bins taxonomy**); and bin_winter_4_filtered was assigned to Proteobacteria/beta-Proteobacteria. Phylogenomic analysis placed these three bins under phyla Verrucomicrobia, Bacteroidetes and Cyanobacteria respectively. In case of identification with GTDB-tk one ambiguity found, that is, bin_winter_59_filtered was assigned to Parachlamydiales whereas phylogenomically it comes under phylum Verrucomicrobiota (**Supplementary Table S2: Refined bins taxonomy**). The complete tree in Newick format was deposited to Figshare (Details in data availability section)

Glycobiome profile of the recovered MAGs

We identified 72 **glycoside hydrolases** (GHs), 42 **glycosyl transferases** (GTs), 13 **polysaccharide lyases** (PLs), 28 **carbohydrate-binding modules** (CBMs), 12 **carbohydrate esterases** (CEs), 7 **auxiliary activities** (AAs) families and 3 other families among the recovered genome bins ($n=27$). Throughout the bins, GH23 family among the glycosyl hydrolases (25 out of 72 GHs) families and GT4 family among the glycosyltransferases (27 out of 42 GTs) families predicted to be predominant. Except for bin_winter_46_filtered (Pirellulales) and bin_winter_59_filtered (Parachlamydiales), the GH23 family enzyme was identified in all the bins. On the other hand, GT4 family enzyme was found in all the recovered bins. The total number of all the carbohydrate activating enzyme families present in each bin varied from 12 (present in bin_winter_7_filtered) to 65 (present in bin_winter_53_filtered) (**Supplementary Table S10**).

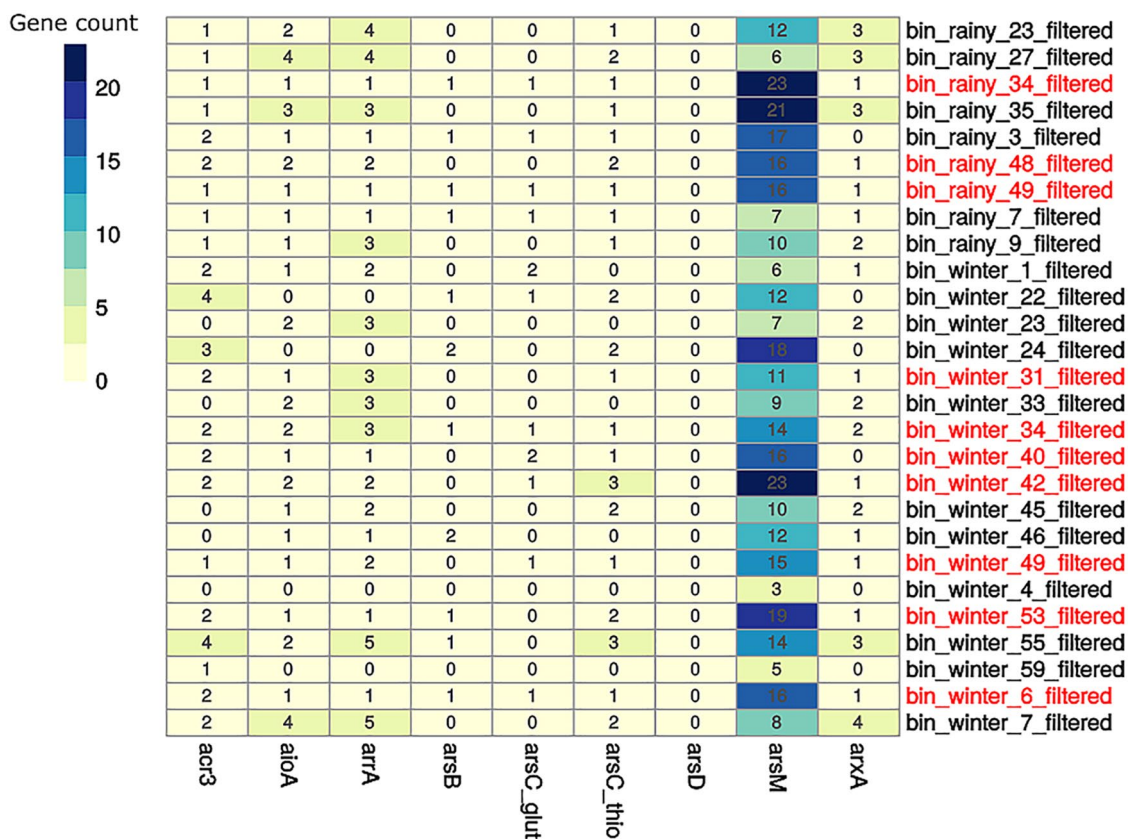


Figure 5. Distribution of arsenic resistance and metabolism-related genes across 27 MAGs. MAGs indicates Metagenome Assembled Genomes.

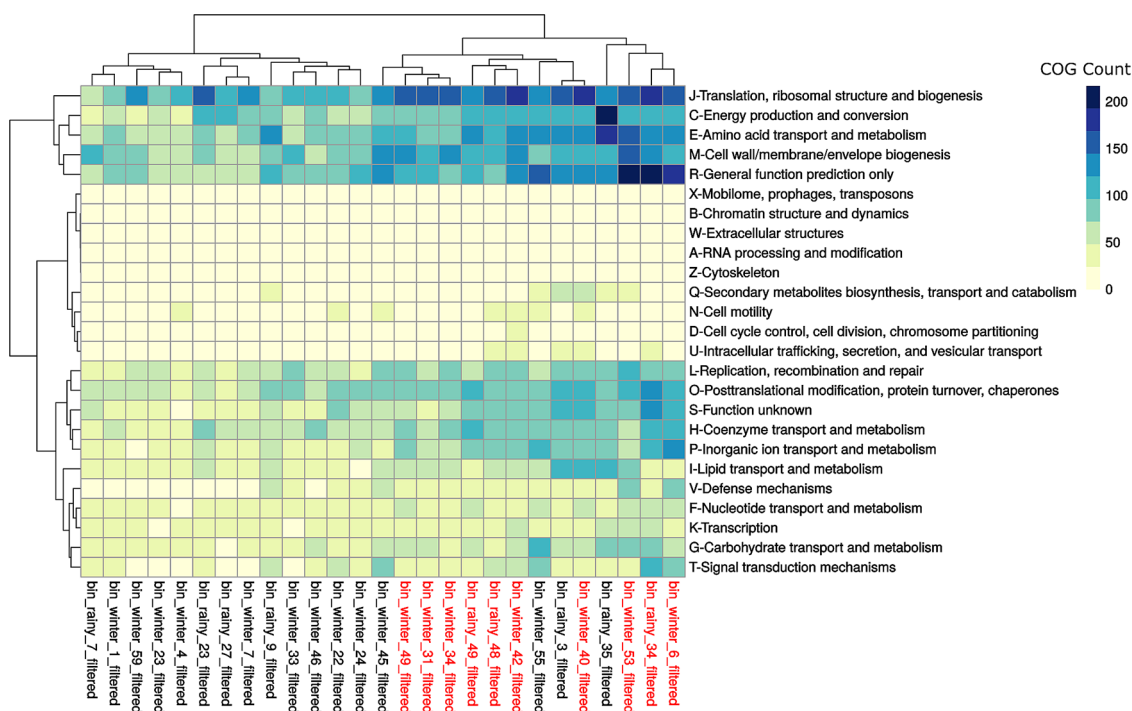


Figure 6. COG frequency heat map of recovered genome bins based on hierarchical clustering. The horizontal axis shows COG categories and the vertical axis represents bins.



Figure 7. Heat map based on hierarchical clustering illustrates the distribution of the RAST categories. The horizontal axis shows RAST categories and the vertical axis represents bins. RAST indicates Rapid Annotation using Subsystem Technology.

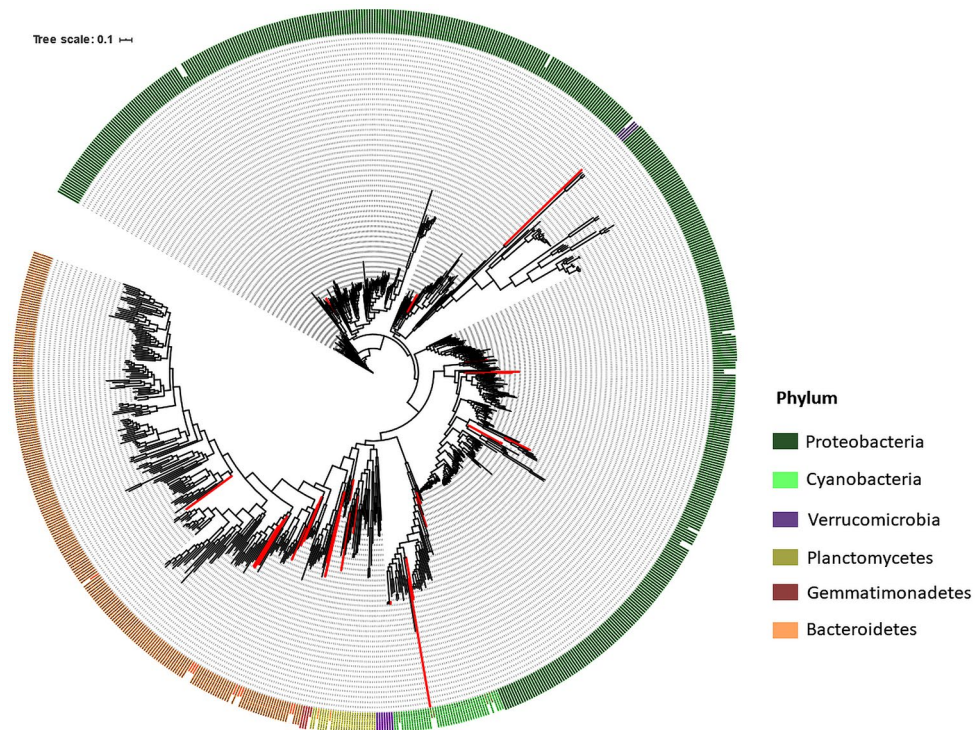


Figure 8. A phylogenomic tree based on the alignment of 92 concatenated marker genes to infer the taxonomic position of 27 reconstructed MAGs. Branches representing MAGs are marked as red. MAGs indicates Metagenome Assembled Genomes.

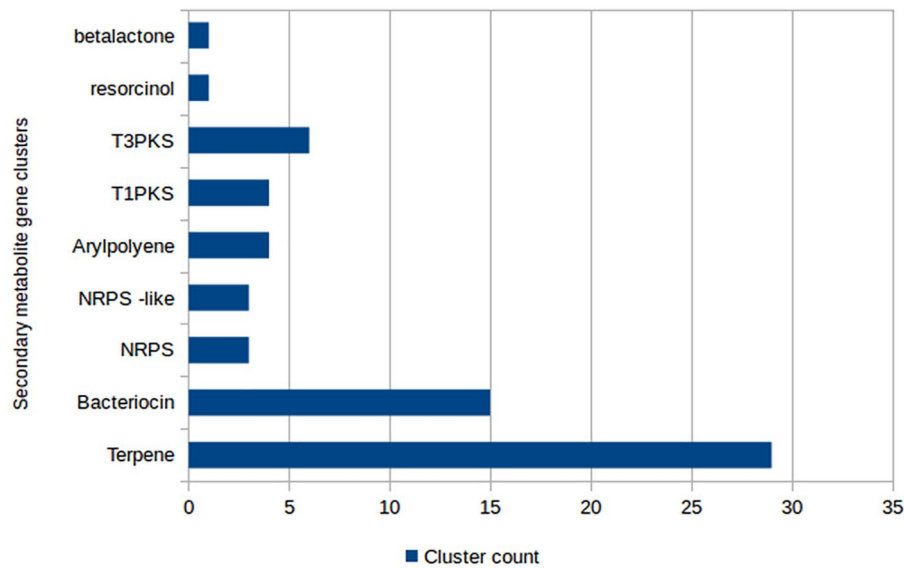


Figure 9. Abundance of secondary metabolite gene cluster types among the recovered genome bins (n=27) obtained with Anti-SMASH.

Genomics cluster of secondary metabolites screening indicates its abundance in the MAGs

We analyzed the MAGs using NAPDOS. A total of 53 ketide synthase (KS) domains were obtained and most of them were identified as the FAS (fatty acid synthase) domain. Interestingly, condensation (C) domains were completely missing from most of the recovered bins (n=24) except for bin_rainy_7_filtered (Flavobacteriales), bin_rainy_34_filtered (Prochlorotrichaceae) and bin_winter_24_filtered (*Microcystis*). C domain associated pathway products of these three bins were identified as lychenicin (bin_rainy_7_filtered), Ca-dependent antibiotic (bin_rainy_filtered_34), syringomycin (bin_rainy_filtered_34 and bin_winter_24_filtered) and microcystin (bin_winter_24_filtered). We also observed that both KS and C domains were absent in bin_winter_4_filtered. The highest number (n=7) of KS domains were detected in bin_winter_55_filtered (**Supplementary Table S11**). With the help of Anti-SMASH, 66 secondary metabolite gene clusters were identified and these gene clusters were classified as terpene, NRPS-like, bacteriocins, NRPS, arylpolyene, t1pks, t3pks, resorcinol and beta lactone (Figures 9 and 10). The most abundant cluster class belonged to terpene and bacteriocins were the second most abundant class (Figure 9). Out of 66 gene clusters, 9 clusters belonged to bin_rainy_49_filtered and 6 clusters were from bin_winter_45_filtered. Notably, Anti-SMASH analysis revealed these gene clusters were absent in bin_rainy_3_filtered, bin_winter_1_filtered, bin_winter_23_filtered, and bin_winter_40_filtered (**Supplementary Table S12**). Since bacteriocins can be analyzed best using BAGEL4, we carried out an analysis that further revealed that there were several classes of bacteriocins such as zoocin_A, prochlorosin, and microcin. A total of 39 bacteriocins were predicted. Out of which, the maximum number of bacteriocins were identified

in bin_rainy_49_filtered (n=8) and bin_winter_55_filtered (n=5) (Figure 11 and **Supplementary Table S13**). Bacteriocins were completely missing in bin_rainy_3_filtered (*Phenylobacterium*), bin_rainy_7_filtered (Flavobacteriales), bin_rainy_27_filtered (Burkholderiaceae), bin_winter_7_filtered (Burkholderiaceae), bin_winter_23_filtered (Chitinophagaceae), bin_winter_31_filtered (Kapabacteriaceae), bin_winter_40_filtered (*Phenylobacterium*) and bin_winter_59_filtered (Parachlamydiales).

Various genomic features derived from the comparative genome centric analysis of MAGs

We extended our analysis to phage elements that may contribute to bacterial genome plasticity. Among the 27 genome bins (or MAGs), 16 bins (bin_rainy_7_filtered, bin_rainy_23_filtered, bin_rainy_27_filtered, bin_rainy_48_filtered, bin_winter_1_filtered, bin_winter_4_filtered, bin_winter_6_filtered, bin_winter_7_filtered, bin_winter_22_filtered, bin_winter_24_filtered, bin_winter_31_filtered, bin_winter_33_filtered, bin_winter_34_filtered, bin_winter_45_filtered, bin_winter_46_filtered and bin_winter_49_filtered) were devoid of phage elements. Bins containing phage elements were found to be incomplete. The highest number of phage elements (n=2) were detected in bin_winter_53_filtered and the rest of the bins contained only a single phage element (Figure 12 and **Supplementary Table S14**). In addition, IS elements were predicted across these genome bins using the latest version of ISSaga. A total of 134 IS elements were found across the genome bins. The higher number of IS elements were identified in bin_winter_6_filtered (n=30) followed by bin_winter_42_filtered (n=13) and bin_winter_40_filtered (n=10). (Figure 12 and **Supplementary Table S15**).

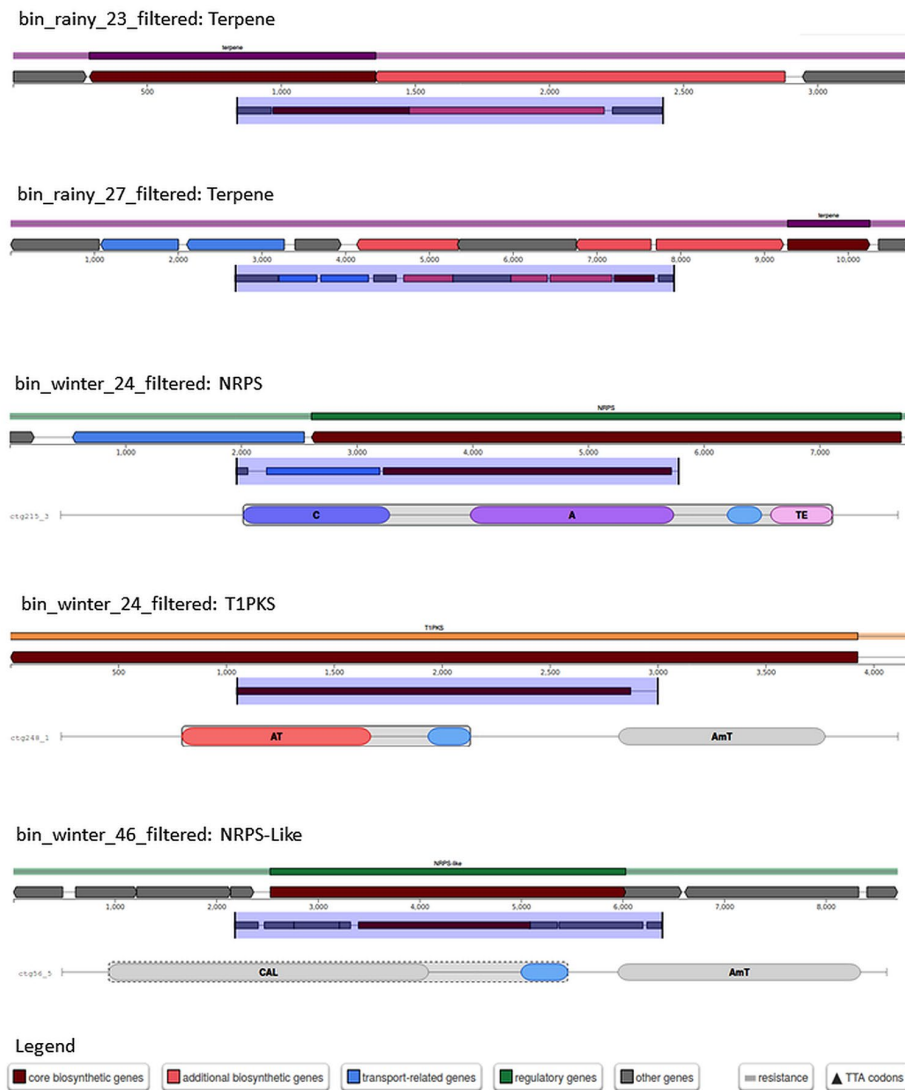


Figure 10. Bins with secondary metabolite genes having more than 75% similarity with the biosynthetic gene clusters encoding secondary metabolites (BGC) database.

A total of 102 CRISPR sites were detected among these bins ($n=27$). Out of which, the highest number ($n=27$) was found in *bin_rainy_34_filtered*. In-addition, CRISPR sites were found totally absent in *bin_rainy_48_filtered*, *bin_rainy_35_filtered*, *bin_rainy_27_filtered*, *bin_rainy_23_filtered*, *bin_winter_42_filtered* and *bin_winter_7_filtered* (Figure 12 and **Supplementary Table S16**).

Discussion

In freshwater ecosystems, bio-resources are largely unexplored, especially in the eastern part of India. Freshwater ecosystems are a treasure trove for the extraction of pharmaceutical and biotechnologically important bioactive compounds.⁶⁰ Freshwater bacteria are also known to monitor the water quality and stabilizing the biogeochemical cycle.³ In this study, the analysis of MAGS (Metagenome Assembled Genomes) provides new insights into the structure and composition of indigenous microbial communities within this anthropogenically influenced artificial lake ecosystem. The preliminary objective

was to catalog the differences in lake microbes during the Winter and Rainy seasons where there was a substantial amount of microbial growth observed in the lake region. However, our detailed analysis revealed that the microbial diversity of the lake during the aforesaid two seasons was minimal (Figure 2B). Across the seasons, the majority of the bacterial members were found to be Proteobacteria, Cyanobacteria, Actinobacteria, Bacteroidetes, Planctomycetes, Verrucomicrobia, and Firmicutes (Figure 2B). The phylum Proteobacteria is a group of Gram-negative bacteria, most of them are identified as agriculturally, industrially, and medically relevant.³ But few of them are known as opportunistic pathogens, for example, *Roseomonas*. The genome of this microbe was recovered as *bin_rainy_35_filtered* (Table 3). Among the recovered genome bins ($n=27$), *bin_winter_42_filtered*, *bin_winter_7_filtered*, *bin_rainy_48_filtered*, *bin_rainy_27* and *bin_rainy_23_filtered* were identified as members of “Burkholderiaceae” family (Table 3). The family Burkholderiaceae belongs to the order Burkholderiales under

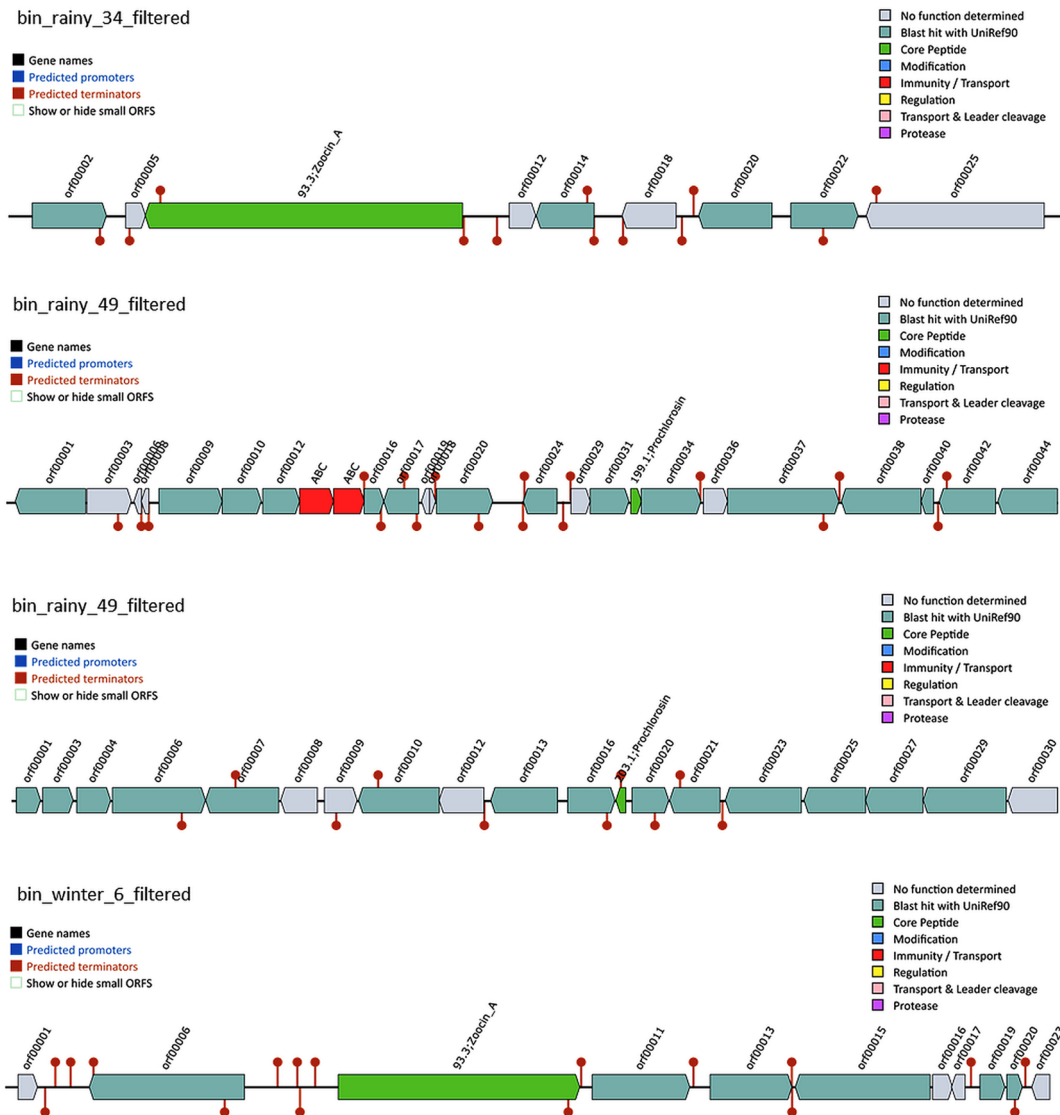


Figure 11. Bacteriocin (Prochlorosin, Zoocin, and Microcin) gene clusters of the recovered genome bins.

the class Betaproteobacteria. Bacteria of this family can be found in various ecological niches (e.g. coral, oceans, freshwater, plant-association).⁶¹ It has demonstrated that species from this group especially those are associated with plants can utilize a large number of aromatic compounds as energy and carbon sources, and some have a considerable biotechnological potential to reduce the concentration and toxicity of chemical pollutants from the environment.⁶² In-addition several reports mentioned that the genus *Polynucleobacter* from Burkholderiaceae family form symbiotic relationships with plants are found in freshwater^{63,64} and possess genes related to the degradation of pesticides used in agriculture.⁶⁵ In our study, MAG (bin_rainy_23_filtered) from the Burkholderiaceae family identified as genus *Polynucleobacter*.

Cyanobacteria, a phylum of bacteria found to be dominant at the lake system (Figure 2B), are popular for secondary metabolite production.⁶⁶ Since Cyanobacteria are known to produce secondary metabolites of biological significance,

therefore this group could prove to be a source of a plethora of industrially relevant metabolites. Taxonomic analysis revealed that out of the 27 recovered genome bins, 4 belonged to group Cyanobacteria. Among these 4 Cyanobacterial bins, 2 were from Rainy season and the remaining 2 were from the Winter season (Table 2). Out of these 2 Winter season bins, one bin identified as genus *Microcystis*. It is noticeable that STAMP analysis also validates the presence of *Microcystis* in the lake (Supplementary Table S1). In freshwater lakes due to eutrophication, several Cyanobacterial blooms are formed and *Microcystis* play an important role in these bloom formation. Importantly, *Microcystis* bloom release toxic secondary metabolites such as microcystins that causes serious health issues to humans and other organisms.^{67,68}

Eight out of the 27 MAGs belonged to “Bacteroidetes” (bin_rainy_7_filtered, bin_winter_1_filtered, bin_winter_23_filtered, bin_winter_31_filtered, bin_winter_33_filtered, bin_winter_34_filtered, bin_winter_49_filtered and

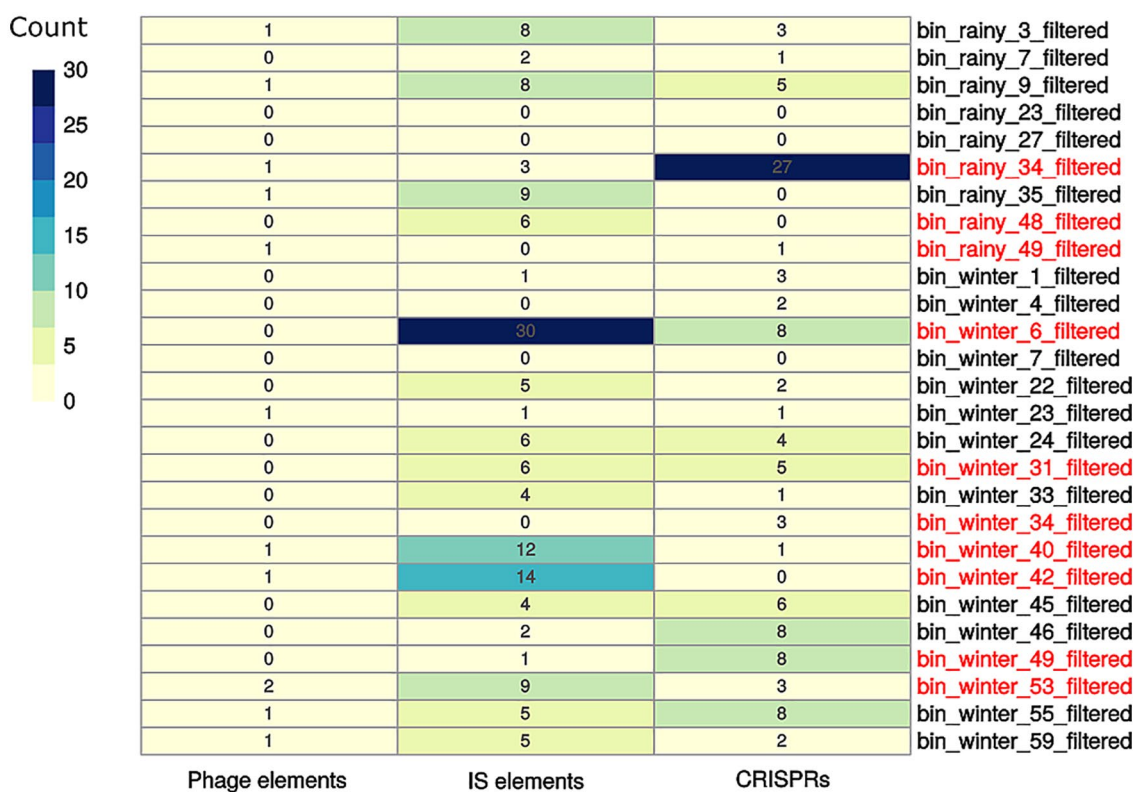


Figure 12. Distribution of PHAGE elements, CRISPRs and IS elements identified from the recovered genome bins.

bin_winter_53_filtered) (Table 3). This makes us wonder whether the lake was exposed to the human intestinal wastes. Considering their anaerobic nature, occurrence of these bacteria in the aquatic environment may indicate recent contamination.^{69,70} On the contrary, Bacteroidetes from normal bacterial communities of diverse habitats plays a beneficial role in the degradation or recycling of organic matter like complex polysaccharides and proteins.¹¹

Glycobiome analysis revealed that all the recovered Bacteroidetes encoded GH23 family of enzymes and this family as per CAZy database was identified as chitinase and peptidoglycan lyase (Supplementary Table S10). The presence of chitinase and peptidoglycan lyases is an indication that other competing organisms such as fungi and bacteria are selectively eliminated.

Although Actinobacterial genomes are not part of the MAGs, STAMP analysis using unassembled reads confirmed the presence of this group. However, this phylum was in low abundance compared to other phyla such as Proteobacteria, Cyanobacteria, and Bacteroidetes (Figure 2B and Supplementary Table S1). One possible reason why this group is in low abundance maybe because of the predominance of Cyanobacteria in the lake region. Cyanobacterial blooms under circumstances of inorganic nutrient availability and high organic matter may compete out highly streamlined Actinobacterial cells. As a result of competition, the population shift is quite obvious and which could also affect the functional capacity of these ecosystems. The phylum Actinobacteria is the

most prospective group among all living bacteria for the exploration of bioactive compounds such as antimicrobials, antiparasitics, antitumor agents, anticancer agents, and enzymes.⁶⁰ It is reported that 45% of all the documented bioactive compounds of microbial origin are produced by Actinobacteria. Therefore, freshwater Actinobacteria have the potential to become standards of ecological freshwater quality.⁷¹

DNA metabolism, photosynthesis, respiration, amino acid derivatives, carbohydrates, and membrane transport like proteins were found in lake microbes (Figure 3A and Supplementary Table S1). These genes are essential for any living system to sustain in the environment. Central carbohydrate metabolism, ATP synthesis, and carbon fixation related pathway modules were dominant throughout the recovered bins (n=27) (Supplementary Table S9). However, the presence of a few functional genes is a distinct signature of this particular lake ecosystem.

Sulfate-reducing bacteria can change the arsenic concentration of a lake by producing hydrogen sulfide, which can further re-precipitate arsenic.⁷² Complete pathway modules related to sulfur metabolism such as M00595 (Thiosulfate oxidation by SOX complex, thiosulfate=>sulfate) and M00176 (Assimilatory sulfate reduction, sulfate=>H₂S) were found in the lake (Supplementary Table S9). The presence of sulfur oxidizers and reducers in this lake marks a very significant game-changer in reducing arsenic toxicity in this lake ecosystem.

Low levels of metal ions that occur naturally in aquatic systems due to slow leaching of soil and rocks have no harmful

impact on aquatic biota.⁷³ But excessive metal ions in waters across many parts of the globe are reported which are primarily originated from agricultural, industrial, municipal, and other anthropogenic activity related waste. The area of West Bengal has reported to be heavily contaminated with arsenic which can cause serious health problems such as keratosis, skin cancer, liver fibrosis, chronic respiratory disease, etc.⁷⁴ In the lake dataset, mobile genetic elements (MGEs) such as plasmids and transposable elements were found and significantly enriched in Rainy season which may enter this environment through urban runoff (Figure 3A and Supplementary Table S1). These elements are one of the important ingredients for manufacturing arsenic resistant microbes.¹⁵ Therefore the possible occurrence of resistance genes related to arsenic (As) contamination in lake microbes can be investigated with genome resolved metagenomics approach.⁷⁵⁻⁷⁸ Importantly, arsenic (As) resistant and metabolism-related genes were found to be present in the recovered genome bins of the lake sample (Figure 5 and Supplementary Table S6); This is an indication that the mechanism to stabilize the Arsenic toxicity is already active in the lake microbes. The predominant forms of As in environment are inorganic arsenate (As [V]) and arsenite (As [III]). The Arsenite (trivalent form) is considered to be more toxic and mobile. Thus, redox transformations of As are the keys to predict their fate in the environment. To date, the bioremediation potential of microorganisms in arsenic-rich environments is mainly based on the ability to resist or metabolize arsenic through oxido-reduction reactions.⁷⁹

ArsB and Acr3 are arsenic biotransformations linked genes that encode arsenite (As [III]) pumps to get rid of reduced arsenic (As [III]) from the cell.⁸⁰ Microorganisms evolved in the earlier anoxic environments, the presence of high concentrations of reduced As (III)⁸¹ is observed in their cell. Majority of microbes have developed efflux systems to remove As (III) from their cells.⁸⁰ Hence, almost every extant microbe is equipped with As (III) permease encoding genes e.g.; ArsB or Acr3.⁷² Some organisms evolved to have anaerobic respiratory pathway genes that utilize As (III) as an electron donor.⁸² In the present study, this type of arsenic cycling is hypothesized to be present (Figure 4).

With increasing, atmospheric oxygen concentrations, As (III) is oxidized to As (V), also a toxic compound that can enter the cells of most microbes via the phosphate uptake systems.⁸⁰ In consequence, to survive with such toxic environments, organisms need to find ways. Independent evolution of the ArsC gene encoding As (V) reductases system is one such way.⁷² Arsenic methylation gene ArsM was found exclusively in the lake with higher abundance (Figure 5). This gene is largely responsible for methylating the arsenic species and making it less toxic for the organisms. Arsenic methylation is another strategy to get rid of toxic As (V).⁷² In some microbes, this coupled with ATP hydrolysis is an energy-dependent strategy to actively pump out As (III) from the cell.^{72,83} ArsB

gene was identified in the MAGs that can bind to ArsA and use membrane potential to pump out As (III). The presence of a whole set of Arsenic metabolizing and detoxifying genes is a powerful indicator that the microbes are stabilizing this region against Arsenic contamination.

The up-surge of genome data research on natural products has significantly influenced in finding new lead structures⁸⁴ and discovering new precursors.⁸⁵ Ribosomally synthesized and post-translationally modified peptides (RiPPs) are an important class of natural products that are unveiled by the genome sequencing.⁸⁶ These compounds are synthesized in all three domains of life and hold immense structural diversity. Bacteriocins belong to a class of ribosomally synthesized and post-translationally modified antimicrobial peptides that can inhibit or kill other bacterial strains but will not harm themselves.²² This class of compounds are viable alternatives to conventional antibiotics and can be utilized as narrow-spectrum antibiotics.²³ Using a metagenomics approach very few studies have carried out to screen bacteriocin genes e.g. the fermented food microbiome⁸⁷⁻⁸⁹ and host-associated microbiomes.⁹⁰ Our study identified several classes of bacteriocins such as zoocin_A, prochlorosin and microcin using advanced metagenomic approach i.e; genome-resolved metagenomics (Figure 11 and Supplementary Figure S2). Among these class of bacteriocins, the biological role of prochlorosin is presently elusive, but they are considered to be functional, as they are found in several *Prochlorococcus* strains, and their biosynthetic genes are transcribed in response to environmental changes.⁹¹ It has also reported that, prochlorosin leader peptides are unique in that they have sequence homology with the Nif11 proteins.⁸⁵ The exact role of the Nif11 proteins is not understood yet, but they are believed to play a role in nitrogen fixation.⁹²

Unlike prochlorosin, the biological role of microcins and zoocin_A are well studied. Many reports suggested that microcins hold crucial antimicrobial properties. These antimicrobial properties of microcins make them excellent candidates as next-generation antibiotics for their application in humans as well as in veterinary medicine. It has also evident that microcins could be used as a targeted strategy to treat Enterobacterial colitis and other infectious diseases.⁹³

Application of zoocin_A to an artificial plaque significantly and selectively minimized the numbers of *Streptococcus* mutants, leading to the suggestion that zoocin A could be useful as an anti-carcinogenic agent.⁹⁴ The presence of zoocin_A, prochlorosin, microcin, and other bacteriocins in the recovered bins suggests that the microbes can be potentially exploited for therapeutic purposes.

Apart from the bacteriocins, our study also focused on other secondary metabolite classes such as NRPS and PKS. Non-ribosomal peptides and polyketides are key natural products and potential drug candidates.⁹⁵ The most dominant pathway products for the KS domain among the recovered genome bins were fatty acid synthesis. In-addition, other KS domain pathway

products assigned to epothilone, chlorothricin, kirromycin, 5-alkenyl-3, 3 (2h)-furanone, nystatin, spinosad, and c-1027 were also found in few numbers across these bins (Supplementary Table S11). Important C domain pathway products such as syringomycin, microcystin, and calcium-dependent antibiotic were identified in bin_winter_24_filtered and bin_rainy_34_filtered. Moreover, NRPS and NRPS-like clusters were also identified in bin_winter_24_filtered and bin_rainy_34_filtered respectively (Supplementary Table S12). Interestingly, all these bins carried polyketides (C domain), NRPS, and NRPS-like clusters belonged to phylum Cyanobacteria. Recent studies demonstrate there are rich sources of non-ribosomal peptides and polyketides in lineages of Cyanobacteria,⁹⁶ Myxobacteria,⁹⁷ Streptomyces⁹⁸ and Pseudomonas.⁹⁹ Natural products like Non-ribosomal peptide and polyketide hold exceptionally diverse structures and can be cyclic, linear, or have branched structures.¹⁰⁰ Moreover, they can be further re-engineered to generate complex products with exotic chemical structures and biological activities.¹⁰¹ These products show a wide range of biological activity and many of them are clinically important anti-parasitic, anti-fungal, anti-microbial, anti-tumor, and immunosuppressive agents.¹⁰²⁻¹⁰⁴

Conclusion

Analyzing lake metagenomes with Illumina reads using a genome-resolved metagenomic approach provides comprehensive information about the taxonomic diversity and potential commercially important metabolites available. The seasonal variation of lake microbiome was not significantly different since the temperature variation was not that drastic during Winter and Rainy season. Taxonomic analysis of the recovered genomes (n = 27) identified that the predominant families were Bacteroidetes, Cyanobacteria, and Proteobacteria. Some of these bacteria are well known for their pathogenic activity. Functional annotation of each recovered genomes revealed the diversity as well as the distribution of arsenic resistance and metabolism-related genes. The presence of arsenic resistance and metabolism-related genes across the genome bins is a possible indicator of arsenic stabilization and neutralization by the ensuing microbes in this region. Recovery of important classes of secondary metabolites including NRPS, PKS, and bacteriocins, indicates that the lake is a rich source of natural products that can be further explored for their applications in the pharmaceutical industry. The identification of secondary metabolite producing arsenic resistant microbes can be used in bioremediation.⁷⁹ Therefore, the present study leading to the conclusion that recovered microbes has immense potential in bio-prospecting.



Acknowledgements

The author would like to acknowledge Computational Genomics Lab members of CSIR-IICB for valuable discussion.

Author Contributions

ST conceived the project. SG and ST designed the project. SG, ANS, MM1, DS, and MM2 performed data analysis. SG and ST wrote the manuscript.

ORCID iDs

Samrat Ghosh  <https://orcid.org/0000-0001-6462-1197>
Sucheta Tripathy  <https://orcid.org/0000-0003-0611-8088>

Supplemental material

Supplemental material for this article is available online.

REFERENCES

- Bhateria R, Jain D. Water quality assessment of lake water: a review. *Sustain Water Resour. Manag.* 2016;2:161-173.
- Zhang K, Yang X, Kattel G, Lin Q, Shen J. Freshwater lake ecosystem shift caused by social-economic transitions in Yangtze River Basin over the past century. *Sci Rep.* 2018;8:17146.
- Newton RJ, Jones SE, Eiler A, McMahon KD, Bertilsson S. A guide to the natural history of freshwater lake bacteria. *Microbiol Mol Biol Rev.* 2011;75:14-49.
- Ma Y, Li J, Wu J, et al. Bacterial and fungal community composition and functional activity associated with lake wetland water level gradients. *Sci Rep.* 2018;8:760.
- Williamson CESJ, Vincent WF, Smol JP. Lakes and reservoirs as sentinels, integrators, and regulators of climate change. *Limnol Oceanography.* 2009;54:2273-2282.
- Basu SKZP. Rabindra Sarovar: the last oasis in the city of Kolkata, India. https://www.researchgate.net/publication/277559105_Rabindra_Sarovar_The_last_oasis_in_the_city_of_Kolkata_India. Updated 2015.
- Blahova L, Babica P, Adamovsky O, Kohoutek J, Marsalek B, Blaha L. Analyses of cyanobacterial toxins (microcystins, cylindrospermopsin) in the reservoirs of the Czech Republic and evaluation of health risks. *Environ Chem Lett.* 2008;6:223-227.
- Blahova L, Babica P, Marsalkova E, Marsalek B, Blaha L. Concentrations and seasonal trends of extracellular microcystins in freshwaters of the Czech Republic results of the national monitoring program. *Clean-Soil Air Water.* 2007;35:348-354.
- Yannarell A, Kent A. Bacteria, distribution and community structure. <https://experts.illinois.edu/en/publications/bacteria-distribution-and-community-structure>. Updated 2009.
- Kurilkina MI, Zakharova YR, Galachyants YP, et al. Bacterial community composition in the water column of the deepest freshwater Lake Baikal as determined by next-generation sequencing. *FEMS Microbiol Ecol.* 2016;92:fiw094.
- Thomas F, Hehemann JH, Rebuffet E, Czjzek M, Michel G. Environmental and gut bacteroidetes: the food connection. *Front Microbiol.* 2011;2:93.
- Azam F, Worden AZ. Oceanography. Microbes, molecules, and marine ecosystems. *Science.* 2004;303:1622-1624.
- Boon E, Meehan CJ, Whidden C, Wong DH, Langille MG, Beiko RG. Interactions in the microbiome: communities of organisms and communities of genes. *FEMS Microbiol Rev.* 2014;38:90-118.
- Dunivin TK, Yeh SY, Shade A. A global survey of arsenic-related genes in soil microbiomes. *BMC Biology.* 2019;17:45.
- Andres J, Bertin PN. The microbial genomics of arsenic. *FEMS Microbiol Rev.* 2016;40:299-322.
- Dunivin TK, Miller J, Shade A. Taxonomically-linked growth phenotypes during arsenic stress among arsenic resistant bacteria isolated from soils overlying the Centralia coal seam fire. *PLoS ONE.* 2018;13:e0191893.
- Newman DJ, Cragg GM. Natural products as sources of new drugs from 1981 to 2014. *J Natural Prod.* 2016;79:629-661.
- Gokhale RS, Sankaranarayanan R, Mohanty D. Versatility of polyketide synthases in generating metabolic diversity. *Curr Opin Struct Biol.* 2007;17:736-743.
- Koglin A, Walsh CT. Structural insights into nonribosomal peptide enzymatic assembly lines. *Nat Prod Rep.* 2009;26:987-1000.
- Wang H, Fewer DP, Sivonen K. Genome mining demonstrates the widespread occurrence of gene clusters encoding bacteriocins in cyanobacteria. *PLoS ONE.* 2011;6:e22384.
- Riley MA, Wertz JE. Bacteriocins: evolution, ecology, and application. *Annu Rev Microbiol.* 2002;56:117-137.

22. Yang SC, Lin CH, Sung CT, Fang JY. Antibacterial activities of bacteriocins: application in foods and pharmaceuticals. *Front Microbiol.* 2014;5:241.
23. Cotter PD, Ross RP, Hill C. Bacteriocins—a viable alternative to antibiotics? *Nat Rev Microbiol.* 2013;11:95-105.
24. Russell JB, Mantovani HC. The bacteriocins of ruminal bacteria and their potential as an alternative to antibiotics. *J Mol Microbiol Biotechnol.* 2002;4:347-355.
25. Oh S, Caro-Quintero A, Tsementzi D, et al. Metagenomic insights into the evolution, function, and complexity of the planktonic microbial community of Lake Lanier, a temperate freshwater ecosystem. *Appl Environ Microbiol.* 2011;77:6000-6011.
26. Babraham B. FastQC a quality control tool for high throughput sequence data. <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
27. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics.* 2018;34:i884-i890.
28. Glass EM, Wilkening J, Wilke A, Antonopoulos D, Meyer F. Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. *Cold Spring Harbor Protocols* 2010;2010:pdbprot5368.
29. Parks DH, Tyson GW, Hugenholtz P, Beiko RG. STAMP: statistical analysis of taxonomic and functional profiles. *Bioinformatics.* 2014;30:3123-3124.
30. Bankevich A, Nurk S, Antipov D, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol.* 2012;19:455-477.
31. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* 2017;27:824-834.
32. Kang DD, Froula J, Egan R, Wang Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ.* 2015;3:e1165.
33. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* 2015;25:1043-1055.
34. Parks DH, Rinke C, Chuvochina M, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol.* 2017;2:1533-1542.
35. Olm MR, Brown CT, Brooks B, Banfield JF. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* 2017;11:2864-2868.
36. Bowers RM, Kyrpides NC, Stepanauskas R, et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol.* 2017;35:725-731.
37. Rodriguez RL, Gunturu S, Harvey WT, et al. The Microbial Genomes Atlas (MiGA) webserver: taxonomic and gene diversity analysis of Archaea and Bacteria at the whole genome level. *Nucleic Acids Res.* 2018;46:W282-W288.
38. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics.* 2019;36:1925-1927.
39. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics.* 2014;30:2068-2069.
40. Overbeek R, Olson R, Pusch GD, et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res.* 2014;42:D206-D214.
41. Szklarczyk D, Franceschini A, Wyder S, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 2015;43:D447-D452.
42. Kanehisa M, Sato Y, Morishima K. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J Mol Biol.* 2016;428:726-731.
43. Pritchard LGR, Humphris S, Elphinstone JG, Toth IK. Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Anal Methods.* 2016;8:12-24.
44. Parks DH. CompareM: a software toolkit which supports performing large-scale comparative genomic analyses. <https://github.com/dparks1134/CompareM>. Updated 2014.
45. Team R. *Core. R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing; 2013.
46. Na SI, Kim YO, Yoon SH, Ha SM, Baek I, Chun J. UBCG: up-to-date bacterial core gene set and pipeline for phylogenomic tree reconstruction. *J Microbiol.* 2018;56:280-285.
47. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* 2019;47:W256-W259.
48. Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* 2012;40:W445-W451.
49. van Heel AJ, de Jong A, Song C, Viel JH, Kok J, Kuipers OP. BAGEL4: a user-friendly web server to thoroughly mine RiPPs and bacteriocins. *Nucleic Acids Res.* 2018;46:W278-W281.
50. Grissa I, Vergnaud G, Pourcel C. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics.* 2007;8:172.
51. Arndt D, Grant JR, Marcu A, et al. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res.* 2016;44:W16-W21.
52. Varani AM, Siguier P, Gourbeyre E, Charneau V, Chandler M. ISSaga is an ensemble of web-based methods for high throughput identification and semi-automatic annotation of insertion sequences in prokaryotic genomes. *Genome Biol.* 2011;12:R30.
53. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215:403-410.
54. Saier MH Jr, Tran CV, Barabote RD. TCDB: the Transporter Classification Database for membrane transport protein analyses and information. *Nucleic Acids Res.* 2006;34:D181-186.
55. Medema MH, Blin K, Cimermancic P, et al. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res.* 2011;39:W339-W346.
56. Ziemert N, Podell S, Penn K, Badger JH, Allen E, Jensen PR. The natural product domain seeker NaPDoS: a phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. *PLoS ONE.* 2012;7:e34064.
57. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol.* 2007;57:81-91.
58. Konstantinidis KT, Tiedje JM. Towards a genome-based taxonomy for prokaryotes. *J Bacteriol.* 2005;187:6258-6264.
59. Kurth D, Amadio A, Ordonez OF, Albarracin VH, Gartner W, Farias ME. Arsenic metabolism in high altitude modern stromatolites revealed by metagenomic analysis. *Sci Rep.* 2017;7:1024.
60. Zothanpuia, Passari AK, Leo VV, et al. Bioprospection of actinobacteria derived from freshwater sediments for their potential to produce antimicrobial compounds. *Microb Cell Fact.* 2018;17:68.
61. Shteinberg M, Kis-Papo T, Millar BC, et al. Survival dynamics of cystic fibrosis-related Gram-negative bacterial pathogens (*Pseudomonas aeruginosa* and *Burkholderia cenocepacia*) in Dead Sea and Atlantic Ocean waters. *J Water Health.* 2015;13:773-776.
62. Suárez-Moreno ZR, Caballero-Mellado J, Coutinho BG, Mendonça-Previato L, James EK, Venturi V. Common features of environmental and potentially beneficial plant-associated Burkholderia. *Microb Ecol.* 2012;63:249-266.
63. Hahn MW, Lang E, Brandt U, Wu QL, Scheuerl T. Emended description of the genus *Polynucleobacter* and the species *Polynucleobacter necessarius* and proposal of two subspecies, *P. necessarius* subsp. *necessarius* subsp. nov. and *P. necessarius* subsp. *asymbioticus* subsp. *Int J Syst Evol Microbiol.* 2009;59:2002-2009.
64. Hahn MW, Lang E, Brandt U, Lünsdorf H, Wu QL, Stackebrandt E. *Polynucleobacter cosmopolitanus* sp. nov., free-living planktonic bacteria inhabiting freshwater lakes and rivers. *Int J Syst Evol Microbiol.* 2010;60:166-173.
65. Fang H, Cai L, Yang Y, et al. Metagenomic analysis reveals potential biodegradation pathways of persistent pesticides in freshwater and marine sediments. *Sci Total Environ.* 2014;470-471:983-992.
66. Xue Y, He Q. Cyanobacteria as cell factories to produce plant secondary metabolites. *Front Bioeng Biotechnol.* 2015;3:57.
67. Falconer IR, Humpage AR. Cyanobacterial (blue-green algal) toxins in water supplies: Cylindrospermopsins. *Environ Toxicol.* 2006;21:299-304.
68. Xie M, Ren M, Yang C, et al. Metagenomic analysis reveals symbiotic relationship among bacteria in microcystis-dominated community. *Front Microbiol.* 2016;7:56.
69. Fiksdal L, Maki JS, LaCroix SJ, Staley JT. Survival and detection of *Bacteroides* spp., prospective indicator bacteria. *Appl Environ Microbiol.* 1985;49:148-150.
70. Wexler HM. Bacteroides: the good, the bad, and the nitty-gritty. *Clin Microbiol Rev.* 2007;20:593-621.
71. Ghai R, Mizuno CM, Picazo A, Camacho A, Rodriguez-Valera F. Key roles for freshwater Actinobacteria revealed by deep metagenomic sequencing. *Mol Ecol.* 2014;23:6073-6090.
72. Zhu YG, Xue XM, Kappler A, Rosen BP, Meharg AA. Linking genes to microbial biogeochemical cycling: lessons from arsenic. *Environ Sci Tech.* 2017;51:7326-7339.
73. Zhou Q, Zhang J, Fu J, Shi J, Jiang G. Biomonitoring: an appealing tool for assessment of metal pollution in the aquatic ecosystem. *Analytica Chimica Acta.* 2008;606:135-150.
74. Mazumder DN, Ghosh A, Majumdar KK, Ghosh N, Saha C, Mazumder RN. Arsenic contamination of ground water and its health impact on population of district of Nadia, West Bengal, India. *Indian J Community Med.* 2010;35:331-338.
75. Wang Y, Li P, Jiang Z, et al. Microbial community of high arsenic groundwater in agricultural irrigation area of Hetao Plain, Inner Mongolia. *Front Microbiol.* 2016;7:1917.
76. Luo J, Bai Y, Liang J, Qu J. Metagenomic approach reveals variation of microbes with arsenic and antimony metabolism genes from highly contaminated soil. *PLoS ONE.* 2014;9:e108185.
77. Ye H, Yang Z, Wu X, et al. Sediment biomarker, bacterial community characterization of high arsenic aquifers in Jiangnan Plain, China. *Sci Rep.* 2017;7:42037.

78. Leon CG, Moraga R, Valenzuela C, et al. Effect of the natural arsenic gradient on the diversity and arsenic resistance of bacterial communities of the sediments of Camarones River (Atacama Desert, Chile). *PLoS ONE*. 2018;13:e0195080.
79. Plewniak F, Crognale S, Rossetti S, Bertin PN. A genomic outlook on bioremediation: the case of arsenic removal. *Front Microbiol*. 2018;9:820.
80. Yan Y, Ding K, Yu X-W, Ye J, Xue X-M. Ability of periplasmic phosphate binding proteins from *Synechocystis* sp. PCC 6803 to discriminate phosphate against arsenate. *Water Air Soil Pollut*. 2017;228:148.
81. Oremland RSSC, Wolfe-Simon F, Stolz JF. Arsenic in the evolution of earth and extraterrestrial ecosystems. *Geomicrobiol J*. 2009;26:522-536.
82. Zhu YG, Yoshinaga M, Zhao FJ, Rosen BP. Earth abides arsenic biotransformations. *Ann Rev Earth Planet Sci*. 2014;42:443-467.
83. Lin YF, Walmsley AR, Rosen BP. An arsenic metallochaperone for an arsenic detoxification pump. *Proc Natl Acad Sci U S A*. 2006;103:15617-15622.
84. Bode HB, Muller R. The impact of bacterial genomics on natural product research. *Ang Chem*. 2005;44:6828-6846.
85. Haft DH, Basu MK, Mitchell DA. Expansion of ribosomally produced natural products: a nitrile hydratase- and Nif11-related precursor family. *BMC Biol*. 2010;8:70.
86. Arnison PG, Bibb MJ, Bierbaum G, et al. Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Nat Prod Rep*. 2013;30:108-160.
87. Więckowicz M, Schmidt M, Sip A, Grajek W. Development of a PCR-based assay for rapid detection of class IIa bacteriocin genes. *Lett Appl Microbiol*. 2011;52:281-289.
88. Illegheems K, Weckx S, De Vuyst L. Applying meta-pathway analyses through metagenomics to identify the functional properties of the major bacterial communities of a single spontaneous cocoa bean fermentation process sample. *Food Microbiol*. 2015;50:54-63.
89. Escobar-Zepeda A, Sanchez-Flores A, Quirasco Baruch M. Metagenomic analysis of a Mexican ripened cheese reveals a unique complex microbiota. *Food Microbiol*. 2016;57:116-127.
90. Zheng J, Gänzle MG, Lin XB, Ruan L, Sun M. Diversity and dynamics of bacteriocins from human microbiome. *Environ Microbiol*. 2015;17:2133-2143.
91. Li B, Sher D, Kelly L, et al. Catalytic promiscuity in the biosynthesis of cyclic peptide secondary metabolites in planktonic marine cyanobacteria. *Proc Natl Acad Sci U S A*. 2010;107:10430-10435.
92. Jacobson MR, Brigle KE, Bennett LT, et al. Physical and genetic map of the major nif gene cluster from *Azotobacter vinelandii*. *J Bacteriol*. 1989;171:1017-1027.
93. Sassone-Corsi M, Nuccio SP, Liu H, et al. Microcins mediate competition among Enterobacteriaceae in the inflamed gut. *Nature*. 2016;540:280-283.
94. Akesson M, Dufour M, Sloan GL, Simmonds RS. Targeting of streptococci by zoocin A. *FEMS Microbiol Lett*. 2007;270:155-161.
95. Walsh CT. Polyketide and nonribosomal peptide antibiotics: modularity and versatility. *Science*. 2004;303:1805-1810.
96. Welker M, von Döhren H. Cyanobacterial peptides—nature's own combinatorial biosynthesis. *FEMS Microbiol Rev*. 2006;30:530-563.
97. Wenzel SC, Müller R. Myxobacterial natural product assembly lines: fascinating examples of curious biochemistry. *Nat Prod Rep*. 2007;24:1211-1224.
98. Weber T, Welzel K, Pelzer S, Vente A, Wohlleben W. Exploiting the genetic potential of polyketide producing streptomycetes. *J Biotech*. 2003;106:221-232.
99. Gross H, Loper JE. Genomics of secondary metabolite production by *Pseudomonas* spp. *Nat Prod Rep*. 2009;26:1408-1446.
100. Kopp F, Marahiel MA. Where chemistry meets biology: the chemoenzymatic synthesis of nonribosomal peptides and polyketides. *Curr Opin Biotechnol*. 2007;18:513-520.
101. Walsh CT. The chemical versatility of natural-product assembly lines. *Acc Chem Res*. 2008;41:4-10.
102. Cane DE, Walsh CT, Khosla C. Harnessing the biosynthetic code: combinations, permutations, and mutations. *Science*. 1998;282:63-68.
103. Cane DE, Walsh CT. The parallel and convergent universes of polyketide synthases and nonribosomal peptide synthetases. *Chem Biol*. 1999;6:R319-R325.
104. Marahiel MA, Stachelhaus T, Mootz HD. Modular peptide synthetases involved in nonribosomal peptide synthesis. *Chem Rev*. 1997;97:2651-2674.