Metagenomicc Profiling of Drinking Water Microbiomes: Insights into Microbial Diversity and Antimicrobial Resistance

Vikas Kumar[1†], Shradha Sharma[2,3†], Kaomud Tyagi[1], Barathi Lenin[3,4], Aarti Ravindran[2,3,4], Karthik Raman[3,4*], Inderjeet Tyagi[1,*]

[1]*Centre for DNA Taxonomy, Molecular Systematics Division, Zoological Survey of India, Kolkata 700053, West Bengal, India*

[2]*Department of Biotechnology, Bhupat and Jyoti Mehta School of Biosciences, Indian Institute of Technology Madras, Chennai, 600036, Tamil Nadu, India*

[3]*Centre for Integrative Biology and Systems mEdicine (IBSE), Wadhwani School of Data Science and AI, Indian Institute of Technology Madras, Chennai, 600036, Tamil Nadu, India*

[4]*Department of Data Science and AI, Wadhwani School of Data Science and AI, Indian Institute of Technology Madras, Chennai, 600036, Tamil Nadu, India*

†These authors contributed equally to this work

** Corresponding authors*

Abstract

Monitoring microbial components in drinking water is as essential as tracking its chemical composition. While traditional culture-based methods can offer valuable insights into the morphology, behavior, and phenotype of organisms, their scope is limited to culturable species. With the advent of high-throughput sequencing, we can now detect a wider range of microbes in any ecosystem, gaining deeper insights into their functional potential and metabolic capabilities more efficiently. In this study, we employed metagenomic analysis to gain a comprehensive understanding of the microbiome of drinking water supplied through public distribution systems in an Indian city. Our findings identified bacteria from the phyla Pseudomonadota, Planctomycetota, Bacteroidota, and Actinomycetota, consistent with previous drinking water microbiome studies. We observed that the composition and diversity of the microbial community were influenced by the temperature of the water samples. No significant correlation between microbial abundance and metal concentration in the sample was detected. Additionally, we assessed the distribution of antibiotic resistance genes (ARGs), finding widespread resistance to aminoglycosides, tetracyclines, and macrolides across samples.

However, the risk associated with these ARGs was out of the scope of this study. Overall, this research provides a preliminary analysis of the drinking water microbiome, offering a foundation for future management strategies.

Keywords: Metagenomic analysis; Antimicrobial resistance genes; Clean drinking water; Diversity analysis; Urban water management

1.Introduction

The provision of safe drinking water to all is a fundamental objective of public health agencies of nations worldwide. The global monitoring of drinking water, sanitation and hygiene, collectively known as WASH, is a critical focal point of the World Health Organization (WHO)(WHO, 2022). Among the 17 Sustainable Development Goals (SDG) adopted by the United Nations in 2015, SDG 6 was specially formulated to make water universally accessible and sustainable. While SDG 3, created to promote good health and well-being, is closely intertwined with water quality and equitable sanitation, it is evident that water is the core connecting element across all SDGs. Consequently, the achievement of these goals depends on the effective implementation and continuous monitoring of Drinking Water Treatment Processes (DWTP) and supply systems (Alcamo, 2019; Bhaduri et al., 2016).

The design of DWTPs involves managing the physical, chemical, and biological contaminants present in untreated water. Briefly, physical processes such as sedimentation and filtration facilitate the removal of particulate matter. Chemical treatments eliminate pollutants such as dissolved harmful substances like Polycyclic Aromatic Hydrocarbons (PAHs), Per- and PolyFluoroalkyl substances (PFAs), heavy metals like lead and mercury, and synthetic dyes like malachite green and azo dyes. Biological contaminants comprise organic debris and a host of organisms, typically bacteria, viruses, protozoans, fungi and algae (Tyagi et al., 2024b). Methods such as adsorption, disinfection, filtration, flocculation, chemical treatment and UV irradiation are commonly used to effectively remove pathogenic microorganisms that cause waterborne diseases (Crittenden et al., 2012; Obayomi et al., 2024). However, the microbiological component of water consists of more than just the live organisms, making it essential to address and neutralize associated elements that adversely impact human health.

Both free-living and biofilm-forming microbes may contain antibiotic resistance genes (ARG) leading to be known as reservoirs of antibiotic-resistant bacteria (ARB). Horizontal gene transfer (HGT) is known to play a critical role in the dissemination of AMR in drinking water microbiomes (Dias et al., 2020). HGT can occur through various mechanisms like transformation, transduction, and conjugation, leading to the spread of ARGs even in the absence of live pathogens. Moreover, the microbial load also comprises elements such as spores, endotoxins, exotoxins, mycotoxins, and phycotoxins . The toxins are metabolites that are

produced by certain microorganisms and can contribute to long-term health risks if ingested (Tyagi et al., 2024a).

Ensuring the availability of safe and clean drinking water is a top priority for India. It actively collaborates with international organizations to develop Water Quality Indices (WQIs) and establish water quality standards. In India, the Bureau of Indian Standards (BIS) regulates drinking water quality, with the latest amendment to the standards made in 2015 (Bureau of Indian Standards, 2012; Chidiac et al., 2023). Despite substantial improvements in water infrastructure, such as those driven by Swachh Bharat Mission and Jal Jeevan mission (Ministry of Jal Shakti), gaps remain in understanding the microbial components of treated drinking water. Notably, metagenomic surveillance and analysis have not been carried out to monitor the quality of drinking water in India. Our study is the first of its kind to carry out metagenomic sequencing of samples collected from drinking water outlets in a metropolitan city to understand the microbial component of treated water.

2. Materials and Methods

2.1. Collection of water samples and DNA extraction

Drinking water samples (municipal tap water) of approximately ≈10 litres each in triplicate (*n*=63) were collected from 21 locations selected based on a grid map approach across a metropolitan city (Precise location masked) located in India during the year March 2022. All the samples were obtained using high-density polyethylene (HDPE)-autoclaved sterile bottles marked as W1-W21 and were kept at 4 ^{o}C till transportation (3 h) to the laboratory. The drinking water samples were then filtered immediately using 0.22 µm MCE membrane (Merck Life Science Pvt. Ltd., Bangalore) with the help of vacuum pump, and the biomass-containing membranes were placed into DNA extraction tubes and stored at -20 ^{o}C till DNA extraction. The extraction of DNA from drinking water samples (n=63) was carried out using DNeasy PowerWater DNA isolation kit (Qiagen Cat No.: 14900- 100-NF) following the standard manufacturer protocol and pooled into 21 drinking water samples for sequencing. The pooling was carried out due to limited resources available for this study. The extracted DNA was qualitatively and quantitatively checked using agarose gel electrophoresis (Cell BioScience Alphalmager MINI) and a Qubit® Fluorometer (Invitrogen, USA), respectively, as per our previous studies (Tyagi et al. 2023). In addition to the drinking water samples, a negative control was filtered as mentioned above; however, we did not get the amplifiable DNA for sequencing from this sample.

2.2. Library preparation, Whole genome sequencing (WGS) and preprocessing of sequencing reads

Using a Covaris ultrasonic disruptor, a 1 μg genomic DNA sample is randomly fragmented into segments of about 350 bp to build the library. End repair, A-tail addition, sequencing adapter ligation, purification, and PCR amplification are the final steps in library construction. Following library building, AATI analysis is used to evaluate the size and integrity of the inserted pieces and the library as a whole. The quality of the library was confirmed by accurately quantifying its effective library concentration using Q-PCR, which was greater than 3 nM. Once the library passed the quality check, it was combined with other libraries based on their effective concentrations and the data output targets. Then, PE150 sequencing was performed. Further, the Illumina sequencing platform's (Novaseq 6000 with 150*2 chemistry) raw data was preprocessed using Fastp (Chen, 2023) to produce clean data for further analysis. Adapters were removed and the paired-end reads were discarded in the following situations 1) when there are more than 10% unknown nucleotides in each read; 2) when there are more than 50% low-quality nucleotides (base quality less than 5) in either read. The sequences generated were submitted to NCBI under the bio project id: PRJNA1008921 with accession number SRR25754722 to SRR25754742.

2.3. Metagenomic assembly, binning and quality control

To remove human contamination from the adapter-trimmed reads, BMTagger v3.101 was employed to align the reads against the hg38 reference genome. Read pairs where only one read aligned to the human genome are also excluded. Quality assessments were conducted pre- and post-filtering using FASTQC v0.12.1 (Brown et al., 2017). The filtered reads were assembled using the MEGAHIT v1.1.1 (Li et al., 2016) assembler within the MetaWRAP pipeline (Uritskiy et al., 2018). MEGAHIT employs a multiple k-mer-based strategy, where smaller k-mers fill gaps in low-coverage regions, while larger k-mers resolve repeat sequences. This approach leads to a highly accurate assembly. We utilized the default k-mers of lengths 21, 29, 39, 59, 79, 99, 119, and 141. Assembly was performed on a per-sample basis, and contigs shorter than 1,000 bp were removed from the analysis.

The assembled contigs were then binned using three algorithms—CONCOCT v0.4.0 (Alneberg et al., 2014), MaxBin2 v2.2.4 (Wu et al., 2016), and MetaBAT2 v2.9.1 (Kang et al., 2015)—each run independently with default parameters. These algorithms leverage tetranucleotide frequencies, coverage information, and sequence composition to achieve precise binning. The consensus approach integrates the strengths of each algorithm, maximizing the capture of genomic diversity while minimizing potential errors that could arise from using a single method.

Each of these algorithms produced separate bins, which were then consolidated using the bin refinement module of MetaWRAP. Subsequently, the refined bins were subjected to re-assembly

using SPades v3.10.1 (Bankevich et al., 2012). CheckM v1.0.7 (Parks et al., 2015) was run on the final re-assembled bins to prepare a report for completeness and contamination. Metagenome-assembled genomes MAGs with over 90% completion and less than 5% contamination were selected for further analysis. For the estimation of the abundance of each bin, we used the Quant_bins module from the MetaWRAP pipeline with default parameters.

2.4. Diversity and differential abundance analysis

After removing the human reads, the filtered reads were assigned with taxonomy using KRAKEN2 v2.1.3 (Wood et al., 2019) with the Kraken nt database (downloaded in May 2023). Species level taxonomy was used for subsequent analysis. The abundance estimation was done using the BRACKEN v2.9 (Lu et al., 2017). The abundance table, taxa table, and metadata were integrated into a single object using the *phyloseq* package in R v4.3.3 for further diversity analysis (McMurdie and Holmes, 2013).

Beta-diversity was evaluated using the Principal Coordinate Analysis (PCoA) technique. The analysis was conducted with the *ordinate* function from the *phyloseq* package. Temperatures were categorized into two groups: Low (25.61 $\pm$ 0.11) and High (31.55 $\pm$ 0.24). The homogeneity of group members was assessed using the *betadisp* function from the vegan package. To test whether dispersions were significantly different between groups, ANOVA was performed on the *betadisp* results. Additionally, ANOSIM was used to evaluate the similarity in composition within and between these two groups. Adonis2 was used to measure the percentage of variation in the microbiome due to temperature changes. Boruta was used to identify features (taxa at the species level) which are significantly different at high vs low temperatures (Miron B. Kursa, 2010).

Shannon and observed alpha diversity indices were measured using the *estimate_richness* function of *phyloseq*. Significance testing was done using the Wilcoxon test between temperature groups and the Kruskal–Wallis test between zones.

2.5. Taxonomy and functional classification

The MAGs generated were classified using the classic workflow (classify_wf) from GTDB-Tk v2.3.2 (Chaumeil et al., 2020). The reference database used for the same was R214. For generating a phylogenetic tree, we used GToTree v1.8.,4 which is a Hidden Markov Model (HMM) based command line tool which builds a tree using a set of single copy marker genes (Lee, 2019). The default cutoff is the presence of at least 50% of the marker genes. However, since some MAGs did not have 50% of the marker genes, we lowered this cutoff to 20%. The output tree from GToTree was passed to IQ-TREE for bootstrapping for 1000 iterations, and the output from IQ-TREE was visualized using iTOL v7(Nguyen et al., 2015),(Letunic and Bork, 2021). We employed PROKKA v1.14.,6 which employs Prodigal for the gene prediction (Seemann, 2014). The output files

generated by PROKKA were subsequently used for identifying Clusters of Orthologous Genes (COGs) using cogclassifier v1.0.5 (https://pypi.org/project/cogclassifier/).

2.6. Analysis of antibiotic resistance genes and virulence factors

Identification of antimicrobial resistance genes was performed using two approaches. First, using sequencing reads, the RGI (Resistance Gene Identifier) tool's BWT (Burrows-Wheeler Transform) feature in RGI v6.0.3 with the default *kma* aligner was employed (https://github.com/arpcard/rgi/blob/master/docs/rgi_bwt.rst) (Clausen et al., 2018). The sequencing reads were mapped to the CARD (Comprehensive Antibiotic Resistance Database) v3.2.7 (Alcock et al., 2023). To ensure robust gene selection, results were filtered to include only those genes with at least 100 mapped reads and an average percentage coverage of 80% or greater. When a particular read mapped to multiple reference sequences, the median reference sequence was used to ensure a more accurate representation. Calculation of RPKM for genes meeting these criteria was done for normalization across samples.

$$\text{Reads per million kilobase (RPKM)} = \frac{Reads\ mapped\ to\ gene}{Gene\ length\ (Kbp) * Total\ mapped\ reads\ in\ millions}$$

In some cases, the reads obtained were large enough to not only get the gene but also the whole open reading frame (ORF). In these cases, the ARG prediction was done by running the MAGs obtained against the CARD database using the RGI *main* feature (https://github.com/arpcard/rgi/blob/master/docs/rgi_main.rst). ORFs with 60% identity with CARD references were further selected for analysis. Furthermore, potential virulence factor genes (VFGs) in MAGs were identified using BLASTp v2.12.0 against the VFDB database (core protein dataset), with a query coverage threshold of 80%, an identity cutoff of 50% and an e-value cutoff of 1e-5 (Liu et al., 2022).

2.7. Analysis of water quality and heavy metals

To investigate the water quality, analysis of heavy metals such as magnesium (Mg) manganese (Mn), copper (Cu), arsenic (As), cadmium (Cd), mercury (Hg) and lead (Pb) along with pH and temperature was carried out using AAS ElementAS Model AAS4129 and Hanna 9829 multiparameter respectively. The findings were compared with the guidelines as prescribed by the WHO, 2022 and IS 10500:2012 in guidelines for guidelines for drinking water quality (**Table 1**).

3. Results

3.1. Water quality (heavy metals, pH and temperature)

The pH and temperature of drinking water samples (W1–W21) ranged from 6.8 to 7.5 and 25.4°C to 32°C, respectively. The pH levels were within the acceptable limits defined by IS:10500:2012 for potable water. Additionally, toxic heavy metals with known carcinogenic potential—arsenic (As), cadmium (Cd), lead (Pb), copper (Cu), and mercury (Hg)—were detected at concentrations

of 0.56–3.05, 0–0.11, 0–1.72, 0.25–11.6, and 0–0.11 µg/ml (ppb), respectively. These values comply with the drinking water quality standards set by WHO and IS:10500:2012 (**Table 1**). Moreover, essential minerals such as magnesium (Mg), manganese (Mn), iron (Fe), and zinc (Zn) were found to be within the permissible limits outlined by these standards. Thus, it can be inferred that in terms of physicochemical contaminants, the drinking water obtained from the metropolitan city falls under the criteria "Safe for drinking purposes".

Table. 1 Heavy metal analysis and their comparison with permissible limits as prescribed by the WHO, 2022 and IS: 10500:2012.

№	**Samples**	**Metals/parameters**	**Observed value range (ppb)**	**Standard permissible limit (WHO, 2022) (ppb)**	**Standard permissible limit (IS:10500:2012) (ppb)**
1.	W1-W21	Arsenic (As)	0.56-3.05	10	NP*
2.	W1-W21	Cadmium (Cd)	0-0.11	3	3
3.	W1-W21	Copper (Cu)	0.25-11.6	2000	50
4.	W1-W21	Lead (Pb)	0-1.72	10	10
5.	W1-W21	Manganese (Mn)	0.56-105.19	80	100
6.	W1-W21	Mercury (Hg)	0-0.11	6	1
7.	W1-W21	Magnesium (Mg)	4391.08-28792.6	NP*	30000
8.	W1-W21	Zinc (Zn)	0-306.99	NP*	5000
9.	W1-W21	Iron (Fe)	0-277.42	NP*	300
10.	W1-W21	pH	6.8-7.5	NP*	6.5-8.5
11.	W1-W21	Temperature ^{o}C	25.4-32	NP*	NP*

*NP: Not provided

3.2. Diversity of drinking water metagenome

The taxonomy assignment of the filtered reads showed that, on average, 59.19 ± 6.34% of the reads in a sample were successfully assigned a taxonomy, with 40.85 ± 6.35% classified at the species level (**Supplementary_Table_1**). Only reads assigned to Bacteria, Archaea, Fungi, and Viruses were included in the diversity analyses. Bacteria dominated the microbial composition,

comprising 98.41% of the identified taxa, while the other domains accounted for 0.38%, 0.99%, and 0.21%, respectively.

We attempted to identify core taxa present across all samples. By calculating species prevalence at a read cutoff of over 10,000, five species were found to have 100% prevalence: *Afipia carboxidovorans, Klebsiella pneumoniae, Pseudomonas aeruginosa, Sphingopyxis macrogoltabida,* and *Variovorax paradoxus*. When the read cutoff was reduced to 1,000, the number of species with 100% prevalence increased to over 500.

3.3. Temperature as a driver of microbial diversity

Principal Coordinates Analysis (PCoA) revealed distinct clustering of samples from the low- and high-temperature groups (**Figure 1**). The sample sizes were balanced, with 10 samples from the lower temperature group (~25°C) and 11 samples from the higher temperature group (~31°C). Homogeneity of variances was assessed using Beta Dispersion with the *betadisper* function, and ANOVA confirmed significant variance in dispersion between groups (p = 0.002413), indicating differences in the variability of microbial community composition across temperature conditions. Additionally, ANOSIM indicated greater similarity within temperature groups compared to between groups (p = 0.001, R = 0.4513) (**Supplementary_Fig._1**). PERMANOVA, performed using the Adonis2 package, identified temperature as a significant contributor to microbial diversity, with an R^2 value of 0.4265 and p = 0.001 (**Figure 1(b)**).

Alpha diversity analysis, using both the Shannon index (which measures richness and evenness) and the observed species index (which estimates richness), was measured to evaluate microbial richness and evenness across the zones. No significant differences were observed, which aligns with expectations since the water supplied to the zones underwent uniform treatment. Although the PCoA plot shows a separation between the North and East zones, their species richness and evenness were not significantly different. In contrast, the North and West zones were distinctly separated in the PCoA plot, with a significant difference in alpha diversity (p = 0.024) (**Supplememtary_Fig._2**).

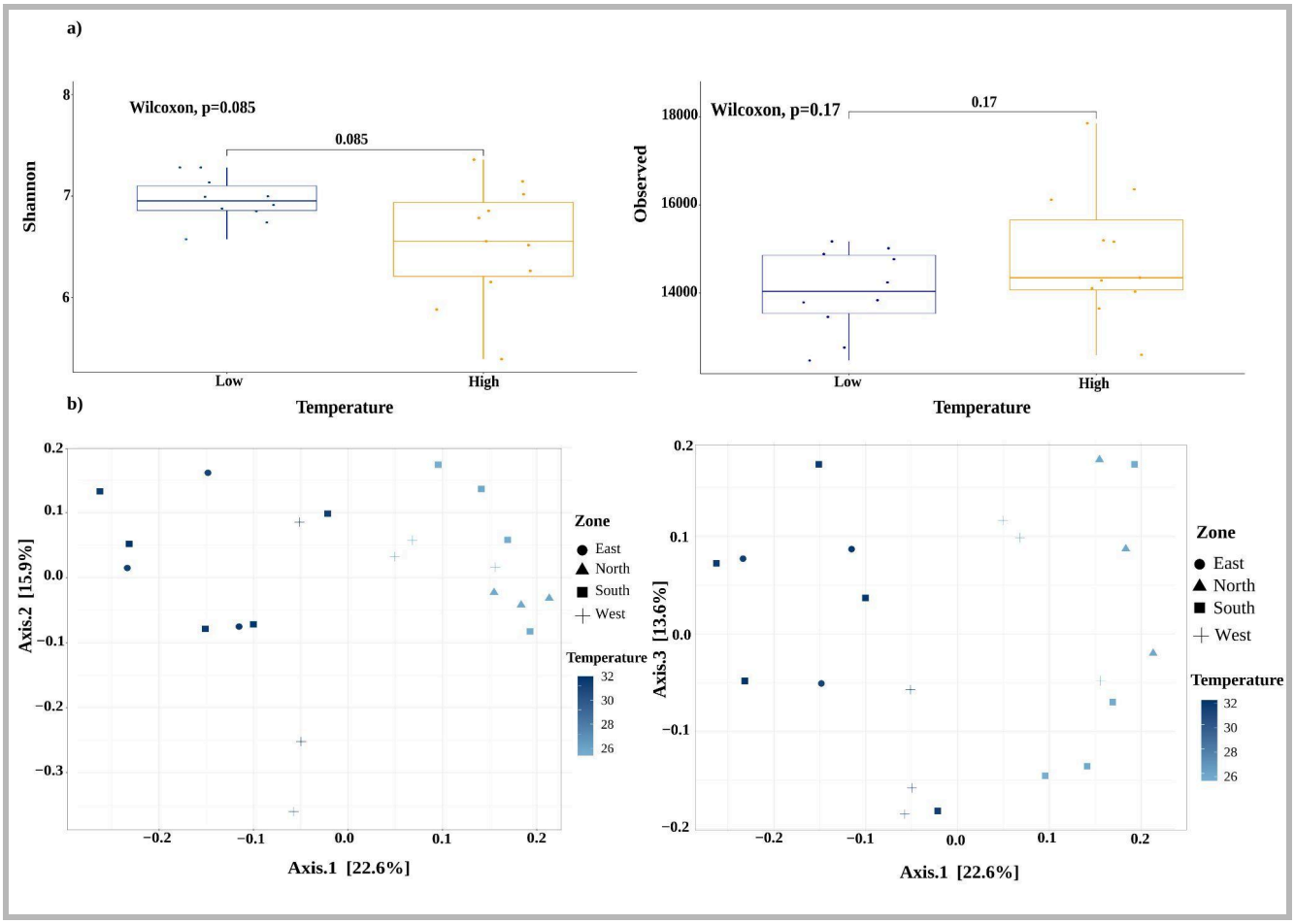

Figure 1: Effect of temperature on the diversity of the microbial community in the samples a) Alpha diversity was estimated between high and low temperature groups using the Shannon and Observed diversity indices. The *p*-values for Shannon (0.085) and Observed (0.17) suggest that there is no significant difference in species richness or evenness between the two groups. b) Beta diversity comparisons were visualized using Principal Coordinates Analysis (PCoA). These plots reveal that, while species richness and evenness do not differ significantly between temperature groups, temperature plays a crucial role in determining the species composition of the samples.

To identify taxa that are differentially abundant under different temperature conditions, we applied Boruta analysis, a random forest-based feature selection method. Out of the 58 differentially abundant species identified by Boruta, species with a ten fold difference in abundance between the two temperature categories are presented in **Figure 2**.

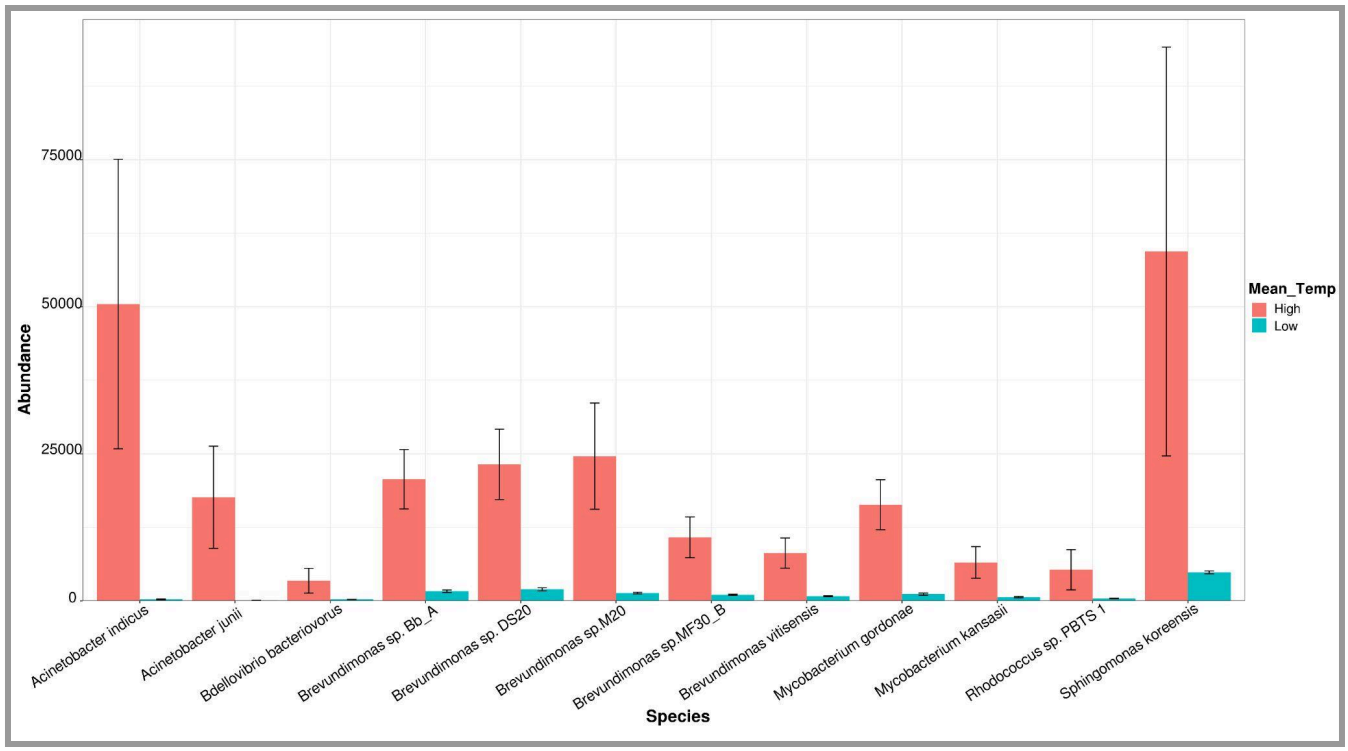

Figure 2: Difference in species abundance identified by Boruta analysis. The differential abundance analysis using Boruta identified taxa with varying abundances in high and low-temperature conditions. The plot shows the mean abundance of species for both temperature groups.

3.4. Community structure of drinking water

After applying a stringent cut-off of over 90% completion and less than 5% contamination, 358 MAGs (all Bacteria) were obtained that were used for further analysis. We chose not to run dREP so as to retain information on the sample origin and the genomic content of individual strains for COG analysis. Taxonomy assignment of the MAGs revealed that Pseudomonadota was the most dominant phylum, present in 68.44 ± 14.35% of the samples across zones. This was followed by Planctomycetota at 10.74 ± 8.68%, Bacteroidota at 6.35 ± 6.28%, and Actinomycetota at 6.92 ± 6.80%. While Pseudomonadota was the dominant phylum in all zones, Actinomycetota was not observed in the north zone at all. For a deeper understanding of the structure of the microbial composition, quantification of the MAGs was done to find the most abundant groups. Sphingomonadaceae was the most abundant family in the community. The abundance of different families is shown in **Supplementary_Fig._3(a).**

Propidium MonoAzide (PMA) treatment is used to remove DNA from dead cells, thereby ensuring that metagenomic analysis is carried out only on living cells. Since our samples were not PMA treated, we performed GRiD analysis to identify genomes with higher replication rates in the community. GRiD scores indicate dividing bacterial cells by computing the ratio of origin of replication sites to that of termination sites. Among the families analyzed, Burkholderiaceae_B exhibited the highest growth rate. Notably, members of both Sphingomonadaceae and Burkholderiaceae_B families demonstrated both high growth rates and substantial abundance within the community. (**Figure 4, Supplementary_Fig._3(b))**.

A broader understanding of microbes present in the drinking water was a major objective of this study. To achieve this, we annotated Open Reading Frames (ORFs) to identify coding sequences (CDS). A total of 969,369 distinct Clusters of Orthologous Genes (COGs) were identified, belonging to 25 functional categories. To ensure both specificity and sensitivity, we applied a 30% identity cutoff, which refined the number of COGs to 661,322. A comparative analysis of these COGs across all zones revealed that the number of genes associated with different functional categories did not differ significantly between the zones. The abundance of COG categories is represented in **Supplementary_Fig._4**. Although the concentrations of all metals in the samples were within permissible limits, the concentration of iron in one of the samples was quite different from the mean values. Our analysis identified three COGs that are associated with iron metabolism: COG0735 (Fe^{2+} or Zn^{2+} uptake regulation protein Fur/Zur), COG1914 (Mn^{2+} or Fe^{2+} transporter, NRAMP family), and COG0672 (High-affinity Fe^{2+}/Pb^{2+} permease). Interestingly, the presence of COG0735 was also high in the Sphingomonadaceae family and samples at high temperature (**Figure 5, Supplementary_Fig._4**). A complete presence/absence list of all COGs identified in the MAGs has been presented in **Supplementary_Table_2.**

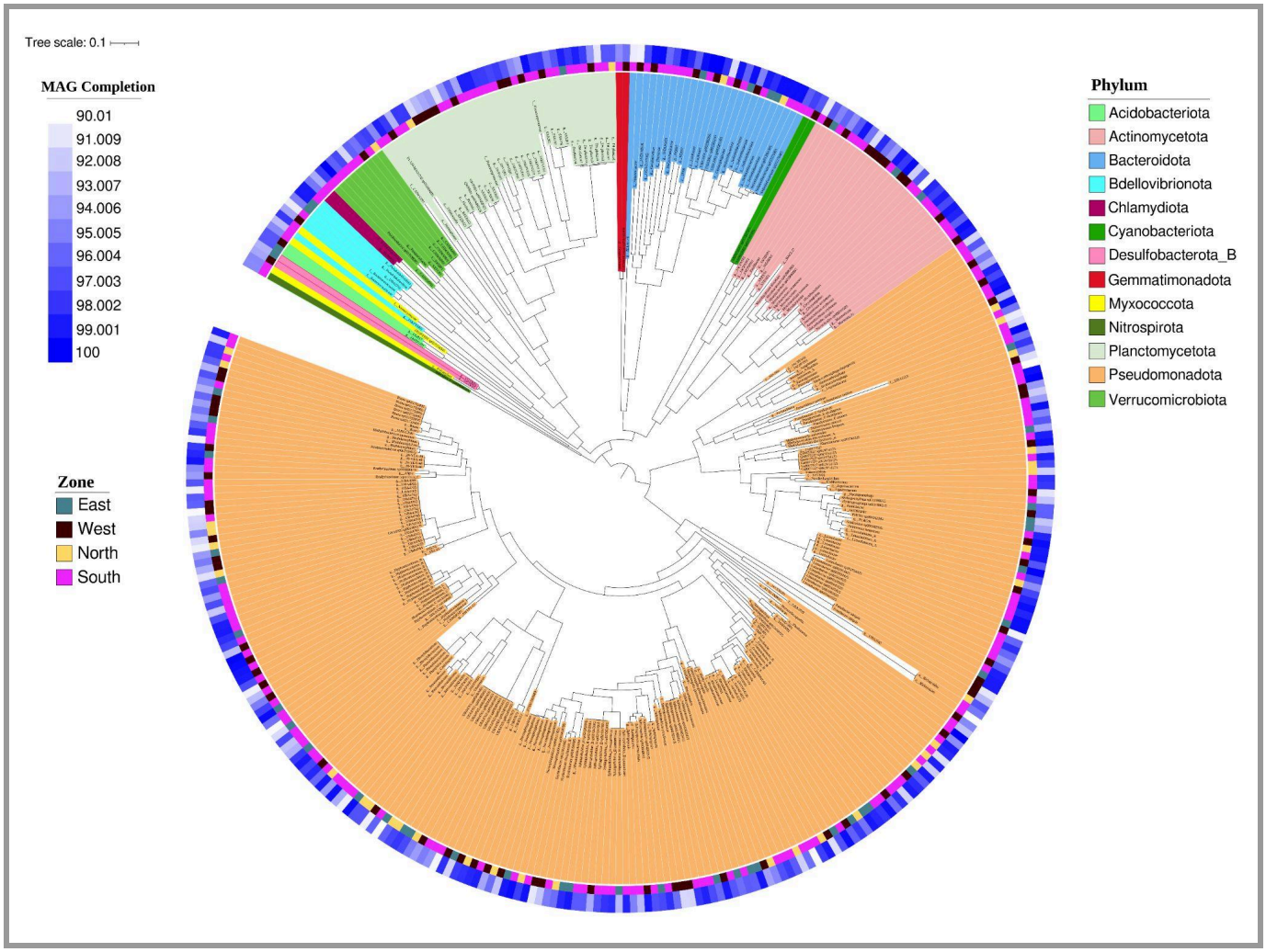

Figure 3: Phylogenetic tree of the MAGs. Phylum-level classification of the MAGs obtained from the samples, showing their completion levels and the respective zones from which the samples were collected.

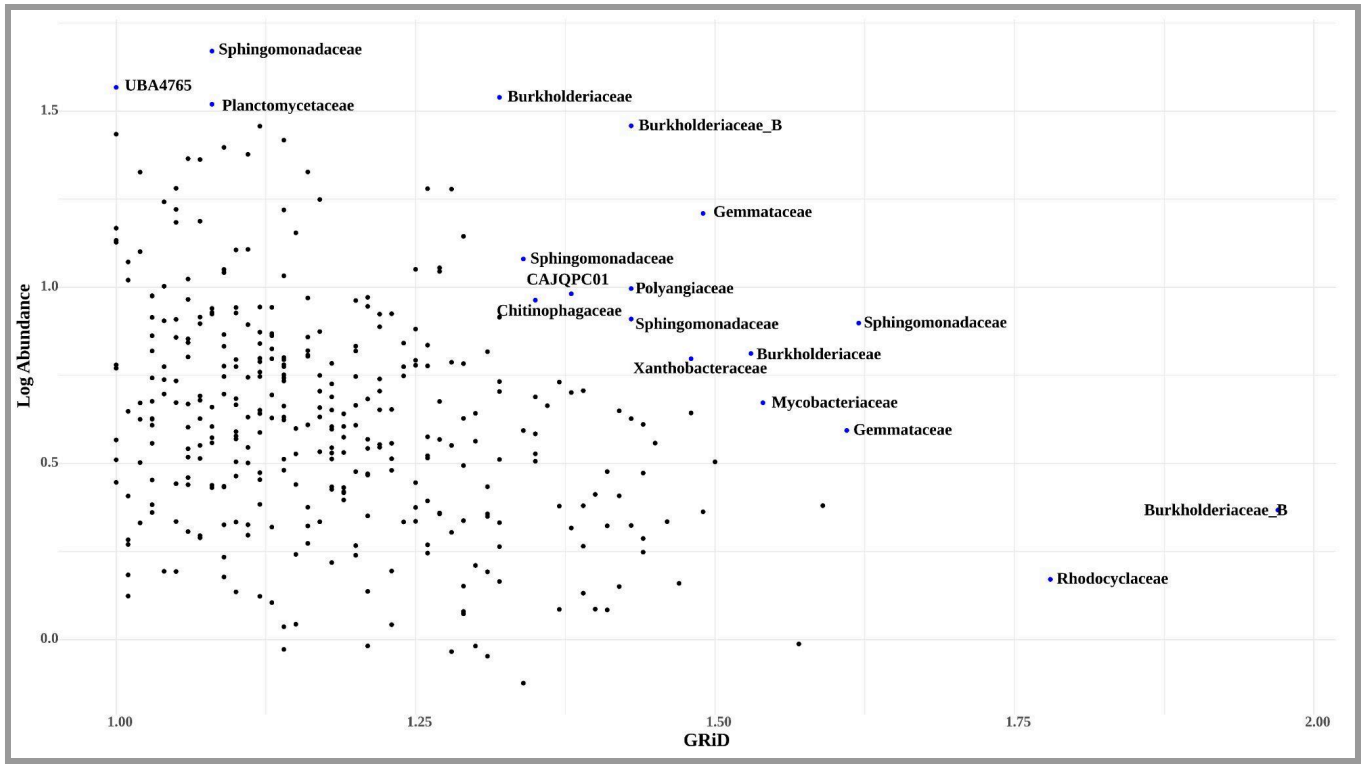

Figure 4: Replication rate of families identified in the samples. Scatter plot depicting bacterial families with a GRiD score greater than 1.3 (indicating actively dividing bacteria) and a log-transformed abundance exceeding 0.75. Only families meeting these thresholds are labelled.

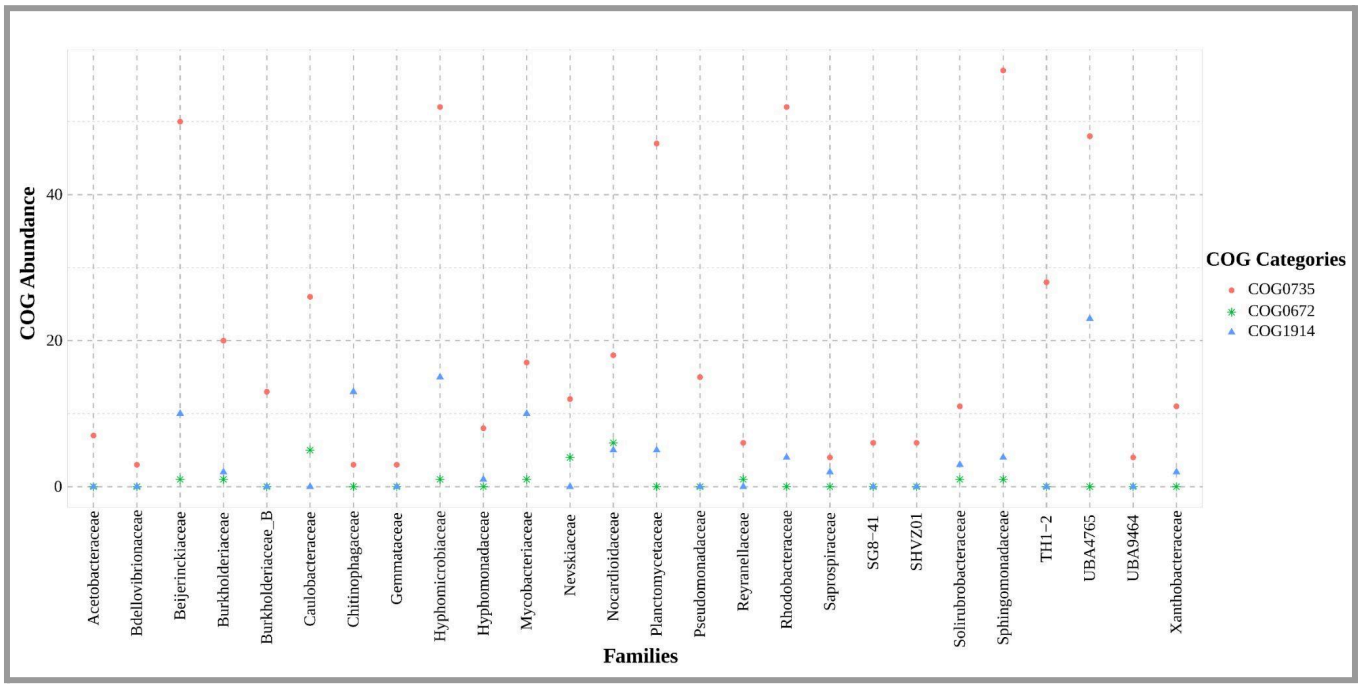

Figure 5: Distribution of COGs involved in iron metabolism among the Bacterial families. COG0735 (Fe^{2+} or Zn^{2+} uptake regulation protein Fur/Zur), COG1914 (Mn^{2+} or Fe^{2+} transporter, NRAMP family), and COG0672 (High-affinity Fe^{2+}/Pb^{2+} permease). Only families with more than 10% relative abundance are shown in the figure.

3.5. Distribution of ARGs in the drinking water

The characterization of Antimicrobial Resistance Genes (ARGs) from the metagenomic reads was done using *RGI-bwt*, following the filtration criteria described in the **Methods** section. Upon applying these filters, we identified 82 unique ARGs spanning across 43 different AMR gene families. To quantify the abundance of antibiotic resistance genes and compare them across samples, normalization was done by calculating RPKM (reads per million kilobase) for each gene identified. Notably, genes associated with resistance to aminoglycoside, elfamycin, fluoroquinolone, macrolide, and tetracycline drug classes were consistently detected across all samples and zones (**Figure 6)**. Among all drug classes, aminoglycoside and macrolide antibiotics had the highest diversity of ARGs, with 22 and 25 unique genes, respectively.

Multidrug-resistant genes accounted for 14% of the total genes identified. The *Pseudomonas aeruginosa* CpxR gene, detected exclusively in the W18 sample, showed resistance to 15 drug classes, the highest among all identified genes. Overall, the W18 sample contained resistance genes for 28 out of 29 drug classes identified. Resistance to drug classes aminocoumarin, monobactam and sulfonamide were exclusively found in this sample as well. **Supplementary_Table_2** provides a detailed list of the drug classes, associated resistance genes, and their corresponding RPKM values.

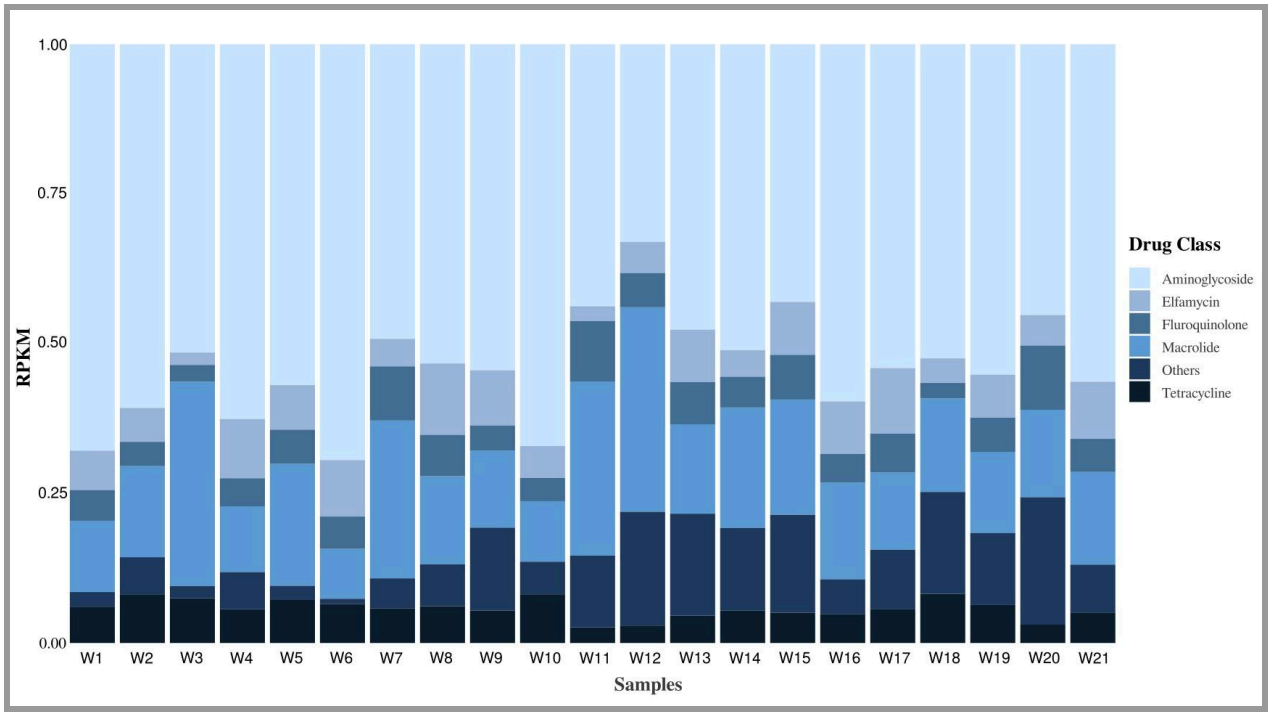

Figure 6: Stacked bar plot showing the abundance of antimicrobial resistance genes (ARGs) conferring resistance to different drug classes across samples. The x-axis represents individual samples, while the y-axis indicates the normalized abundance of ARGs, expressed in RPKM (Reads Per Kilobase of transcript per Million mapped reads). Each stack corresponds to a specific drug class, with the height proportional to the total abundance of ARGs identified for that class. Only the top five drug classes for which ARGs are identified in all samples are shown here.

The mechanisms by which the genes in this study confer resistance include antibiotic efflux, antibiotic inactivation, target replacement, target alteration, and resistance by absence. While the majority of genes conferred resistance through target alteration, most multidrug resistance genes conferred resistance primarily via the efflux mechanism.

To better understand the antibiotic resistance genes (ARGs) in the metagenome-assembled genomes (MAGs), we used the *rgi-main* tool with the CARD database. The *adeF* gene, conferring resistance to fluoroquinolone and tetracycline was the most prevalent ARG in the samples, detected across 15 bacterial families, with the highest representation in the Pseudomonadaceae Family. Focusing on multidrug-resistant genes, *YajC* was exclusively identified in Pseudomonadaceae species. It exhibited resistance to ten drug classes, including fluoroquinolones, cephalosporins, glycylcyclines, penams, tetracyclines, oxazolidinones,

glycopeptides, rifamycins, phenicols, disinfectant agents and antiseptics. In addition, the *Pseudomonas aeruginosa* soxR gene identified was resistant to 8 drug classes and was exclusively found in Xanthobacteraceae **(Figure 7, 8).** Among the drug classes, fluoroquinolones had the highest prevalence of resistance-associated genes in the MAGs. A detailed list of the genes identified in MAGs, along with their mechanism of action, and drug classes to which they confer resistance is presented in **Supplementary_Table_2.** Overall, the majority of ARGs were found in the Pseudomonadaceae and Mycobacteriaceae families. The distribution of ARGs across bacterial families is shown in **Figure 8**.

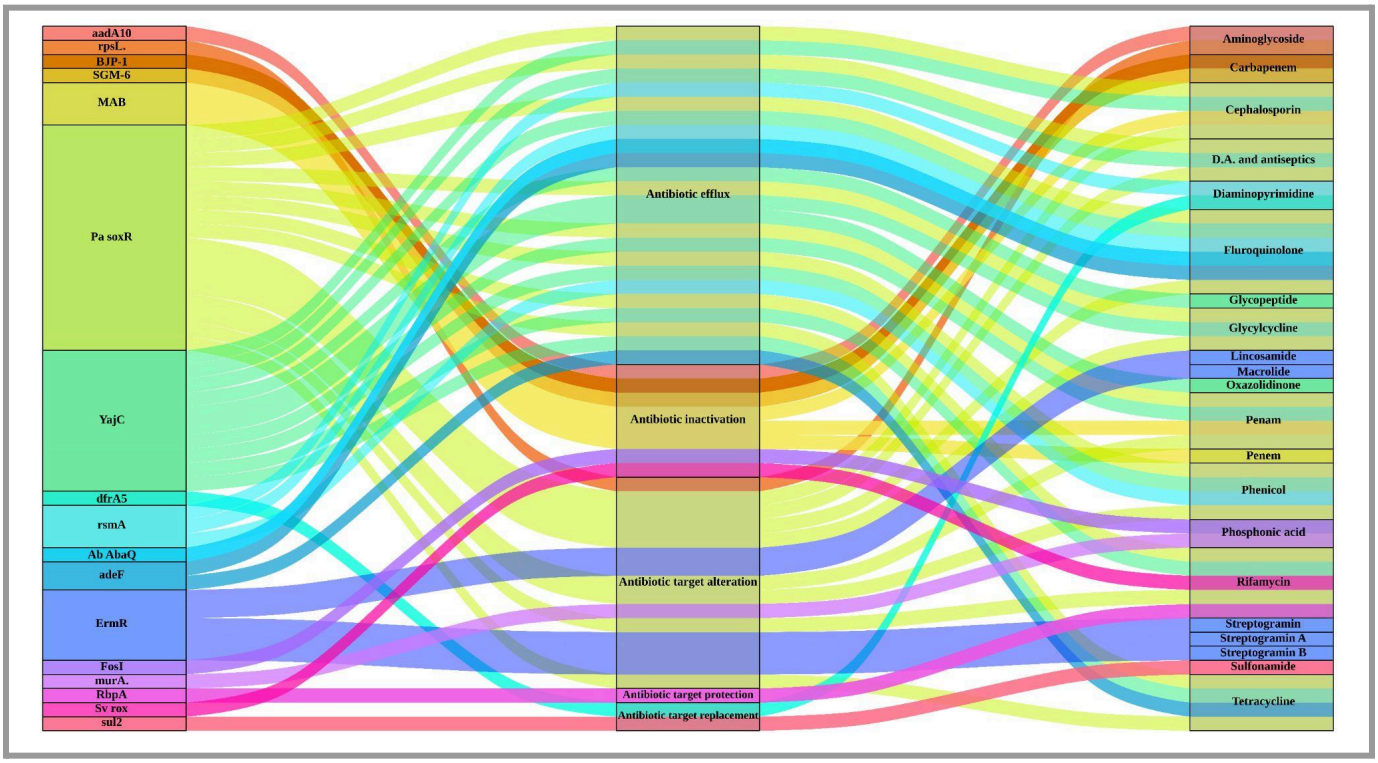

Figure 7: Genes detected in the MAGs, their mechanism of action and the drug classes to which they are resistant. Amongst all ARGs represented here, Pa soxR, MAB, YajC, rsmA, adeF and ErmR are multidrug resistant. The abbreviations used in the figure are as follows: Pa soxR: *Pseudomonas aeruginosa* soxR, Ab AbaQ: *Acinetobacter baumann*ii AbaQ, murA: *Mycobacterium tuberculosis* intrinsic murA conferring resistance to fosfomycin, rpsL: *Mycobacterium tuberculosis* rpsL mutations conferring resistance to streptomycin, D.A.: Disinfecting agents, Sv: *Streptomyces venezuela* rox.

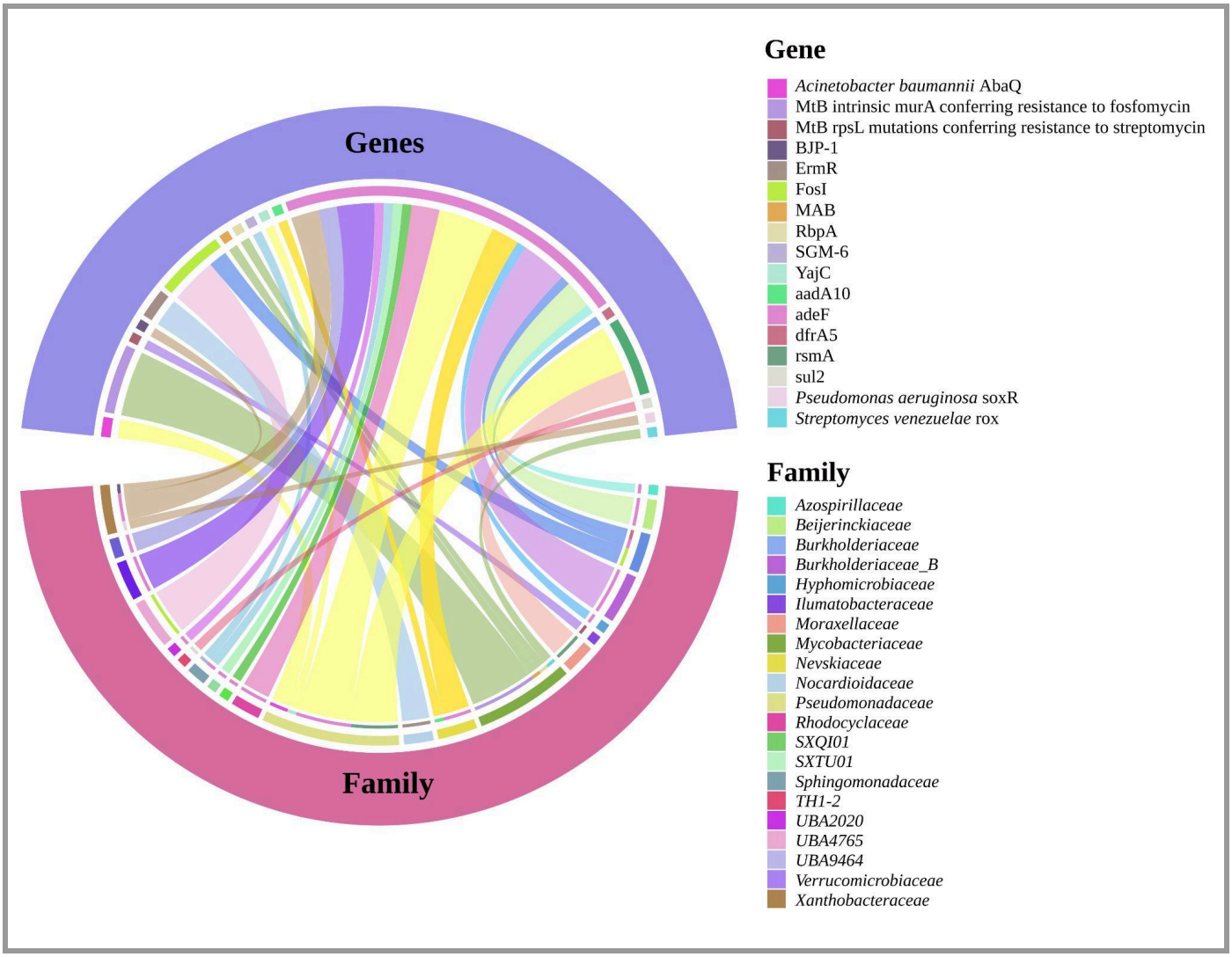

Figure 8: Chord diagram representing the distribution of genes identified in the MAGs across different Families. In the chord diagram, the gene sectors are displayed from left to right in the same order as they appear in the legend. Similarly, the Family sectors are displayed from right to left, maintaining the order specified in the legend.

3.6. Distribution of virulence factors across samples

The analysis of virulence mechanisms in the MAGs revealed that the Sphingomonadaceae family contained the highest number of virulence factors (**Supplementary_Fig._5**). This result is consistent with expectations, as the Sphingomonadaceae family is the most abundant in our samples. A sample-wise analysis of the virulence factors indicated that sample W18 exhibited significantly higher virulence factors (**Supplementary_Table_2**). In line with these findings, the

AMR analysis also revealed that W18 contained antimicrobial resistance (AMR) profiles resistant to the majority of drug classes.

4. Discussion

Periodic monitoring of drinking water supplied through the public water distribution system, post- treatment is essential. This ensures that quality is maintained and that public health is not adversely impacted due to persisting abiotic or biotic contaminants. The emergence of metagenomics offers a comprehensive and holistic understanding of the microbial community that will help in developing effective management and mitigation strategies to ensure safe and high-quality drinking water (Mahajna et al., 2022).

In the present study, we sampled water from 21 locations across four zones of an Indian city. To our understanding, this is the first metagenomic study of drinking water from India. In our analysis, the most abundant phyla were Pseudomonadota (Proteobacteria), Planctomycetota, Bacteroidota, and Actinomycetota. This finding is consistent with the other reports that have been published previously (Sudarshan et al., 2024).

Diversity analyses showed that temperature was the main factor influencing microbial composition, followed by the zone to which a sample belonged, to a much lesser extent. The impact of metal concentration on microbial diversity could not be assessed due to the narrow range of values for some metals and the limited sample size, which made it hard to draw conclusions. Additionally, for each metal, there was at least one outlier that would have affected the analysis, so we focused our study on temperature and zone. The levels were also within permissible limits, rendering further analysis unnecessary.

Among the microbial families identified in our samples, Sphingomonadaceae was the most dominant, with 45 MAGs classified under this family. Members of this family have been identified in various ecosystems including soil, corals, clinical samples as well as drinking water (Molina et al., 2024). They are also known to be strictly aerobic chemoautotrophs and play a crucial role in initiating biofilm formation, as well as contributing significantly to biofilm maturation (de Vries et al., 2019; Vaz-Moreira et al., 2011). Additionally, the higher prevalence of the *Fur* gene within this family highlights intriguing avenues for future research. The Fur gene is known to regulate cellular iron homeostasis, with the iron-Fur complex becoming activated in the presence of elevated ferrous iron levels, thereby repressing iron metabolism (Fujita et al., 2020). However, no correlation was found between the abundance of the Sphingomonadaceae family and iron concentrations in our study. Additionally, although there was a trend suggesting higher levels of COG0735 in high-temperature groups, the relationship remains inconclusive due to the limited sample size.

Apart from members of the Sphingomonadaceae family, members of the Rhodobacteraceae and Beijerinckiaceae families, despite being half as abundant as Sphingomonadaceae, showed a significant presence of COG0735 (**Figure 5)**. The higher abundance of nutritional/metabolic virulence factors in Rhodobacteraceae and Beijerinckiaceae highlights their potential roles in competitive nutrient acquisition and survival under trace metal-limited conditions (**Supplementary_Fig._5**). Rhodobacteraceae, in particular, are known for their dependence on trace metals, which supports their diverse metabolic pathways and ecological strategies (Kong et al., 2024). The profiles of virulence factors and COGs in these two families raise important questions about the role of iron in their potential synergistic interactions or competition, which could be explored in future studies.

The analysis of antibiotic resistance genes revealed the presence of the *ermR* gene in both metagenomic reads and MAGs. The sole carrier of this gene is *Aeromicrobium erythreum*, which was also binned from our samples. The gene was found in samples W11, W18, and W19 (**Supplementary_Table_2**). This gene, previously known as *ermA*, has been reported to confer resistance to macrolide-lincosamide-streptogramin (MLS) antibiotics, which is consistent with our findings. MAGs for *Aeromicrobium erythreum* were also binned from these three samples: W18_bin.26.orig, W19_bin.11.orig, and W11_bin.8.orig (**Supplementary_Table_1**). To our knowledge, this bacterium has not been previously reported in drinking water (Brikun et al., 2004; Miller, 1991). It is to be noted that the presence of a gene is not an indication of its expression. Expression of a gene is dependent on environmental conditions and regulatory mechanisms.

Conclusion

A comprehensive understanding of the microbial community of the drinking water is crucial for ensuring its quality and safety. In this study, we utilized metagenomics for an in-depth analysis of the microbial diversity of drinking water supplied through public distribution systems of an Indian city. Our analysis revealed that among all the physicochemical parameters measured, temperature was the most significant factor influencing the composition of the microbial community. The influence of other factors could not be assessed because of the limited sample size, which is insufficient to draw any statistically significant patterns. Additionally, since propidium monoazide (PMA) treatment was not performed on the samples, we could not determine whether the microbes identified were live or dead. Our analysis also identified antimicrobial resistance genes (ARGs). While the presence of ARGs does not directly indicate a risk, it underscores the importance of future research focused on horizontal gene transfer and risk assessment of these ARGs. Overall, while this study has enforced previously reported observations of drinking water microbiome, it has also yielded new insights about its

composition. It enhances our understanding of the drinking water ecosystem and underscores the importance of monitoring to protect public health.

Acknowledgements

V.K., K.T. and I.T. are thankful to the Director, Zoological Survey of India and the RAMC committee for funding the project titled "Drinking Water Microbiome: Monitoring of the Water Quality and Associated Health Implications." S.S. acknowledges the half-time research assistantship from the Ministry of Education, Government of India. B.L. acknowledges support from the Centre for Integrative Biology and Systems mEdicine (IBSE). A.R. acknowledges support from the project RB22231279BTHINT008481.

Data Availability Statement: The sequences generated were submitted to NCBI under the bio project ID: PRJNA1008921 with accession number SRR25754722 to SRR25754742.

References

1. Alcamo, J., 2019. Water quality and its interlinkages with the Sustainable Development Goals. Curr. Opin. Environ. Sustain., Environmental Change Assessment 36, 126–140. https://doi.org/10.1016/j.cosust.2018.11.005
2. Alcock, B.P., Huynh, W., Chalil, R., Smith, K.W., Raphenya, A.R., Wlodarski, M.A., Edalatmand, A., Petkau, A., Syed, S.A., Tsang, K.K., Baker, S.J.C., Dave, M., McCarthy, M.C., Mukiri, K.M., Nasir, J.A., Golbon, B., Imtiaz, H., Jiang, X., Kaur, K., Kwong, M., Liang, Z.C., Niu, K.C., Shan, P., Yang, J.Y.J., Gray, K.L., Hoad, G.R., Jia, B., Bhando, T., Carfrae, L.A., Farha, M.A., French, S., Gordzevich, R., Rachwalski, K., Tu, M.M., Bordeleau, E., Dooley, D., Griffiths, E., Zubyk, H.L., Brown, E.D., Maguire, F., Beiko, R.G., Hsiao, W.W.L., Brinkman, F.S.L., Van Domselaar, G., McArthur, A.G., 2023. CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. Nucleic Acids Res. 51, D690–D699. https://doi.org/10.1093/nar/gkac920
3. Alneberg, J., Bjarnason, B.S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U.Z., Lahti, L., Loman, N.J., Andersson, A.F., Quince, C., 2014. Binning metagenomic contigs by coverage and composition. Nat. Methods 11, 1144–1146. https://doi.org/10.1038/nmeth.3103
4. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A., Pevzner, P.A., 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J. Comput. Biol. 19, 455–477. https://doi.org/10.1089/cmb.2012.0021
5. Bhaduri, A., Bogardi, J., Siddiqi, A., Voigt, H., Vörösmarty, C., Pahl-Wostl, C., Bunn, S.E., Shrivastava, P., Lawford, R., Foster, S., Kremer, H., Renaud, F.G., Bruns, A., Osuna, V.R., 2016. Achieving Sustainable Development Goals from a Water Perspective. Front. Environ. Sci. 4. https://doi.org/10.3389/fenvs.2016.00064
6. Brikun, I.A., Reeves, A.R., Cernota, W.H., Luu, M.B., Weber, J.M., 2004. The

erythromycin biosynthetic gene cluster of Aeromicrobium erythreum. J. Ind. Microbiol. Biotechnol. 31, 335–344. https://doi.org/10.1007/s10295-004-0154-5
7. Brown, J., Pirrung, M., McCue, L.A., 2017. FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. Bioinformatics 33, 3137–3139. https://doi.org/10.1093/bioinformatics/btx373
8. Bureau of Indian Standards, 2012. Indian Standard DRINKING WATER — SPECIFICATION (Second Revision).
9. Chaumeil, P.-A., Mussig, A.J., Hugenholtz, P., Parks, D.H., 2020. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics 36, 1925–1927. https://doi.org/10.1093/bioinformatics/btz848
10. Chen, S., 2023. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. iMeta 2, e107. https://doi.org/10.1002/imt2.107
11. Chidiac, S., El Najjar, P., Ouaini, N., El Rayess, Y., El Azzi, D., 2023. A comprehensive review of water quality indices (WQIs): history, models, attempts and perspectives. Rev. Environ. Sci. Biotechnol. 22, 349–395. https://doi.org/10.1007/s11157-023-09650-7
12. Clausen, P.T.L.C., Aarestrup, F.M., Lund, O., 2018. Rapid and precise alignment of raw reads against redundant databases with KMA. BMC Bioinformatics 19, 307. https://doi.org/10.1186/s12859-018-2336-6
13. Crittenden, J.C., Trussell, R.R., Hand, D.W., Howe, K.J., Tchobanoglous, G., 2012. MWH's Water Treatment: Principles and Design. John Wiley & Sons.
14. de Vries, H.J., Beyer, F., Jarzembowska, M., Lipińska, J., van den Brink, P., Zwijnenburg, A., Timmers, P.H.A., Stams, A.J.M., Plugge, C.M., 2019. Isolation and characterization of Sphingomonadaceae from fouled membranes. NPJ Biofilms Microbiomes 5, 6. https://doi.org/10.1038/s41522-018-0074-1
15. Dias, M.F., da Rocha Fernandes, G., Cristina de Paiva, M., Christina de Matos Salim, A., Santos, A.B., Amaral Nascimento, A.M., 2020. Exploring the resistome, virulome and microbiome of drinking water in environmental and clinical settings. Water Res. 174, 115630. https://doi.org/10.1016/j.watres.2020.115630
16. Fujita, M., Sakumoto, T., Tanatani, K., Yu, H., Mori, K., Kamimura, N., Masai, E., 2020. Iron acquisition system of Sphingobium sp. strain SYK-6, a degrader of lignin-derived aromatic compounds. Sci. Rep. 10, 12177. https://doi.org/10.1038/s41598-020-68984-2
17. Kang, D.D., Froula, J., Egan, R., Wang, Z., 2015. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ 3, e1165. https://doi.org/10.7717/peerj.1165
18. Kong, Y., Zhang, R., Blain, S., Obernosterer, I., 2024. Seasonal dynamics in microbial trace metals transporters during phytoplankton blooms in the Southern Ocean. Environ. Microbiol. 26, e16695. https://doi.org/10.1111/1462-2920.16695
19. Lee, M.D., 2019. GToTree: a user-friendly workflow for phylogenomics. Bioinformatics 35, 4162–4164. https://doi.org/10.1093/bioinformatics/btz188
20. Letunic, I., Bork, P., 2021. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res. 49, W293–W296. https://doi.org/10.1093/nar/gkab301

21. Li, D., Luo, R., Liu, C.-M., Leung, C.-M., Ting, H.-F., Sadakane, K., Yamashita, H., Lam, T.-W., 2016. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. Methods, Pan-omics analysis of biological data 102, 3–11. https://doi.org/10.1016/j.ymeth.2016.02.020
22. Liu, B., Zheng, D., Zhou, S., Chen, L., Yang, J., 2022. VFDB 2022: a general classification scheme for bacterial virulence factors. Nucleic Acids Res. 50, D912–D917. https://doi.org/10.1093/nar/gkab1107
23. Lu, J., Breitwieser, F.P., Thielen, P., Salzberg, S.L., 2017. Bracken: estimating species abundance in metagenomics data. PeerJ Comput. Sci. 3, e104. https://doi.org/10.7717/peerj-cs.104
24. Mahajna, A., Dinkla, I.J.T., Euverink, G.J.W., Keesman, K.J., Jayawardhana, B., 2022. Clean and Safe Drinking Water Systems via Metagenomics Data and Artificial Intelligence: State-of-the-Art and Future Perspective. Front. Microbiol. 13. https://doi.org/10.3389/fmicb.2022.832452
25. McMurdie, P.J., Holmes, S., 2013. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLOS ONE 8, e61217. https://doi.org/10.1371/journal.pone.0061217
26. Miller, E.S., 1991. Cloning vectors, mutagenesis, and gene disruption (ermR) for the erythromycin-producing bacterium Aeromicrobium erythreum. Appl. Environ. Microbiol. 57, 2758–2761. https://doi.org/10.1128/aem.57.9.2758-2761.1991
27. Miron B. Kursa, W.R.R., 2010. Feature Selection with the Boruta Package. J. Stat. Softw. 36, 1–13. https://doi.org/10.18637/jss.v036.i11
28. Molina, C.A., Quiroz-Moreno, C., Jarrín-V., P., Díaz, M., Yugsi, E., Pérez-Galarza, J., Baldeón-Rojas, L., 2024. Bacterial community assessment of drinking water and downstream distribution systems in highland localities of Ecuador. J. Water Health 22, 536–549. https://doi.org/10.2166/wh.2024.290
29. Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., Minh, B.Q., 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol. Biol. Evol. 32, 268–274. https://doi.org/10.1093/molbev/msu300
30. Obayomi, O.V., Olawoyin, D.C., Oguntimehin, O., Mustapha, L.S., Kolade, S.O., Oladoye, P.O., Oh, S., Obayomi, K.S., 2024. Exploring emerging water treatment technologies for the removal of microbial pathogens. Curr. Res. Biotechnol. 8, 100252. https://doi.org/10.1016/j.crbiot.2024.100252
31. Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., Tyson, G.W., 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 25, 1043–1055. https://doi.org/10.1101/gr.186072.114
32. Seemann, T., 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30, 2068–2069. https://doi.org/10.1093/bioinformatics/btu153
33. Sudarshan, A.S., Dai, Z., Gabrielli, M., Oosthuizen-Vosloo, S., Konstantinidis, K.T., Pinto, A.J., 2024. New Drinking Water Genome Catalog Identifies a Globally Distributed Bacterial Genus Adapted to Disinfected Drinking Water Systems. Environ. Sci. Technol. 58, 16475–16487. https://doi.org/10.1021/acs.est.4c05086
34. Tyagi, I., Kumar, V., Tyagi, K., 2024a. Chapter 2 - Water pollution—sources and

health implications of the environmental contaminants on the aquatic ecosystem and humans: approach toward sustainable development goals, in: Dehghani, M.H., Karri, R.R., Tyagi, I., Scholz, M. (Eds.), Water, The Environment, and the Sustainable Development Goals. Elsevier, pp. 35–66. https://doi.org/10.1016/B978-0-443-15354-9.00008-6

35. Tyagi, I., Tyagi, K., Kumar, V., Ghosh, A.K., 2024b. Chapter 6 - Health implications of pathogens on aquatic fauna and humans: Waterborne diseases and their surveillance, in: Tyagi, I., Kumar, V., Karri, R.R. (Eds.), Environmental Metagenomics, Water Quality and Suggested Remediation Measures of Polluted Waters. Elsevier, pp. 61–76. https://doi.org/10.1016/B978-0-443-13659-7.00024-2
36. Uritskiy, G.V., DiRuggiero, J., Taylor, J., 2018. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. Microbiome 6, 158. https://doi.org/10.1186/s40168-018-0541-1
37. Vaz-Moreira, I., Nunes, O.C., Manaia, C.M., 2011. Diversity and Antibiotic Resistance Patterns of Sphingomonadaceae Isolates from Drinking Water. Appl. Environ. Microbiol. 77, 5697–5706. https://doi.org/10.1128/AEM.00579-11
38. WHO, 2022. WHO Global water, sanitation and hygiene: Annual report 2021 (No. ISBN: 978-92-4-005725-8). Geneva: World Health Organization.
39. Wood, D.E., Lu, J., Langmead, B., 2019. Improved metagenomic analysis with Kraken 2. Genome Biol. 20, 257. https://doi.org/10.1186/s13059-019-1891-0
40. Wu, Y.-W., Simmons, B.A., Singer, S.W., 2016. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics 32, 605–607. https://doi.org/10.1093/bioinformatics/btv638