Microbial diversity of Atlantic Rainforest ponds assessed by nanopore sequencing

Samir V. F. Atum[1,2,3], Douglas M. M. Soares[4], Isaias Santos[5], Grant A. Johnson[5], Adão H. R. Domingos[5], Etelvino J. H. Bechara[1,2], João C. Setubal[2], Lara Urban[6], Cassius V. Stevani[1,2,7,*], Renato S. Freire[1,7,*]

[1] *Departamento de Química Fundamental, Instituto de Química, Universidade de São Paulo, SP, Brazil*

[2] *Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, São Paulo, Brazil*

[3] *Helmholtz AI Institute, Helmholtz Pioneer Campus, Munich, Germany*

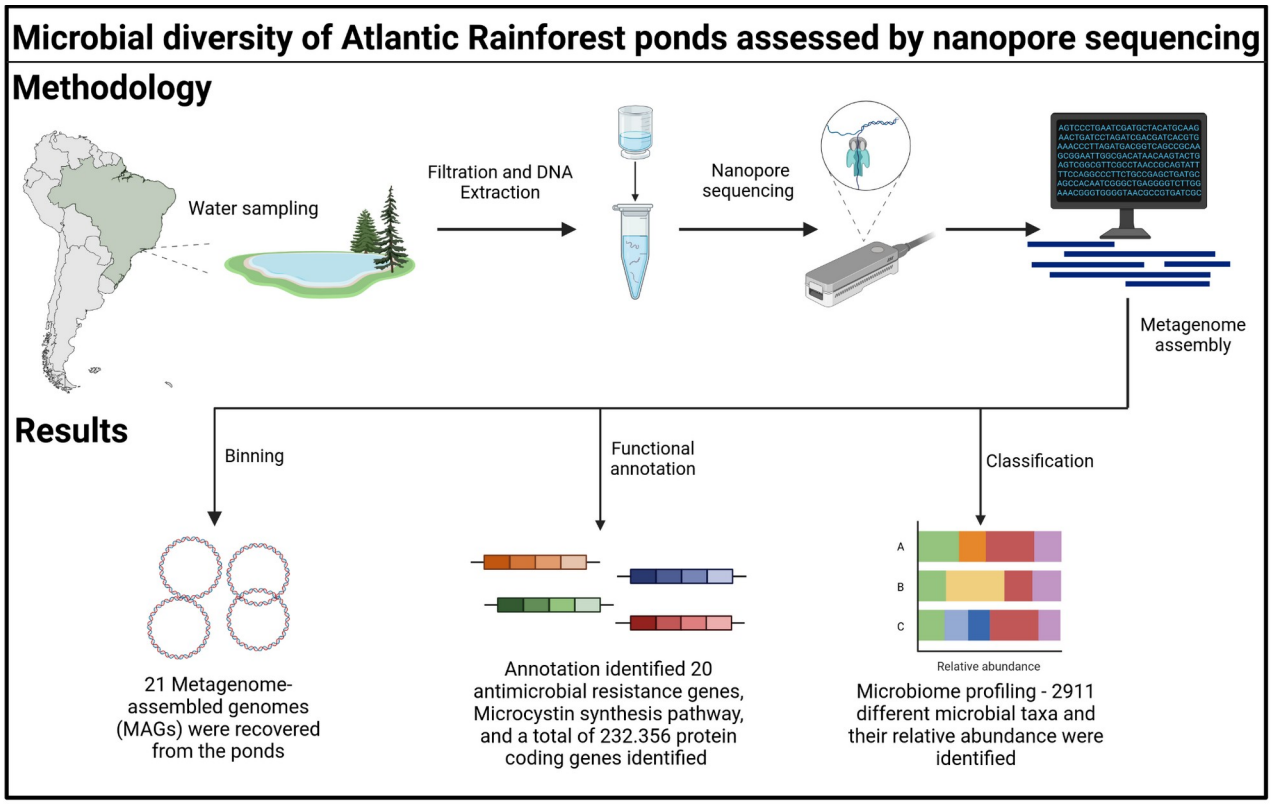
[4] *Department of Bioprocess Engineering and Biotechnology, School of Pharmaceutical Sciences, São Paulo State University, Araraquara, SP, Brazil*

[5] *IPBio - Instituto de Pesquisas da Biodiversidade, Iporanga, SP, Brazil*

[6] *Vetsuisse Faculty, University of Zurich, Switzerland*

[7] *Research Centre for Greenhouse Gas Innovation (RCGI-POLI-USP), University of São Paulo, São Paulo, SP Brazil*

GRAPHICAL ABSTRACT



ABSTRACT

Despite their limited size, ponds are ecologically important habitats that harbor rich microbial communities and play key roles in global biodiversity and biogeochemical cycling. In this study, we conducted a metagenomic survey of microbial diversity in three ponds - Vermelha, Grande, and Furnas - situated within the Atlantic Rainforest biome in Brazil. Vermelha and Grande are natural, remote, and minimally impacted by human activity, whereas Furnas is an artificial pond with greater human accessibility and long-term anthropogenic influence. Using a long-read nanopore shotgun metagenomics approach, we sequenced DNA extracted from pond water, assembled metagenomes, and performed taxonomic classification and functional annotation. From these datasets, 21 metagenome-assembled genomes (MAGs) were recovered. Our results reveal pond-specific microbial community compositions and functional profiles. Although a direct temporal comparison would require synchronized sampling, the current data offer compelling snapshots of microbial diversity. Vermelha was characterized by high cyanobacterial diversity and exclusive Planctomycetes, while Furnas displayed broader taxonomic diversity, and more notable Actinobacteria and Patescibacteria. Viral communities, including dsDNA bacteriophages and eukaryotic viruses, also varied across sites, with Prasinovirus consistently detected in all ponds. Functional annotation identified putative antibiotic resistance genes (ARGs), most of them in Furnas, potentially reflecting human impact. Besides ARGs, cyanotoxin biosynthesis genes— especially those linked to microcystin production— were predominantly found in Vermelha. Our findings underscore the ecological distinctiveness of each pond and demonstrate the utility of nanopore-based metagenomics for investigating microbial biodiversity in understudied freshwater ecosystems. This work provides knowledge on microbial community structure and anthropogenic influence in tropical pond environments, with implications for conservation and biotechnological exploration.

1. Introduction

Ponds hold significant value as microorganism reservoirs, fulfilling crucial roles in ecosystems and human activities. Despite their typical size of less than 50,000 m^2; collectively, these water bodies globally surpass lakes in total area (Chopyk et al., 2020; Downing, 2010). With distinct ecological importance, ponds play a substantial role in the global carbon cycle and overall biodiversity (Biggs et al., 2017). They nurture diverse macro- and microspecies, while also fostering intricate microbial communities crucial for aquatic ecosystem health. However, their significance is often overshadowed by larger aquatic systems, despite being more sensitive to environmental and human influences due to their shallow nature and limited size. This sensitivity

can lead to issues such as algal blooms and contamination, impacting water quality and ecosystem health (Chopyk et al., 2020). Therefore, comprehending and addressing pond microbial dynamics is vital to maintain their ecological integrity and services. Additionally, human-induced alterations in landscape dynamics, including urbanization, agricultural practices, and hydrological changes, pose threats to pond ecosystems, including nutrient loading, contamination, habitat loss, and species encroachment, with potential cascading effects on entire ecosystems, and negatively impacting plant and animal diversity, including unique and endangered species like amphibians, dragonflies, and aquatic plants that rely on ponds as refuges (Oertli et al., 2005). Studying pond microbial biodiversity can inform conservation and restoration efforts, benefiting both microbial communities and dependent species. Hence, mitigating human-induced impacts on pond ecosystems is crucial for preserving ecological function and the diversity of inhabitants.

Among the currently available tools to assess biodiversity, metagenomic approaches consist of a reliable and affordable way to identify species at a molecular level. Many metagenomic studies are based on the sequencing of DNA barcodes, such as the 16S rRNA gene and the Internal Transcribed Spacer (ITS), widely used for molecular taxonomy. While this approach is useful for exploring the diversity and microbial dynamics in environmental samples, it often falls short in revealing the potential genetic resources of the organisms and achieving high taxonomic resolution, especially when only sequencing a few variable regions of these marker genes. Additionally, its effectiveness is dependent on the completeness of databases, particularly when not using de novo assemblies.

While short-read platforms such as Illumina are renowned for high accuracy, coverage, and data throughput, they require a high initial cost - and short-reads result in highly fragmented assemblies. In this scenario, nanopore long-read sequencers are an affordable alternative to produce more contiguous assemblies. The long read technology also overcomes the uneven coverage of reads as a result of GC bias, which can be introduced during the amplification and sequencing steps of Illumina sequencing (Chen et al., 2013). Consequently, the better reads classification results in a powerful tool to assess the relative abundance of taxa (Meslier et al., 2022). Considering the necessity of *de novo* assemblies for unexplored water resources, long-read sequencing approaches offer a reliable, affordable, and informative way to investigate the microbial diversity of the ponds.

The Upper Ribeira Tourist State Park (Parque Estadual Turístico do Alto Ribeira - PETAR) is a conservation area located in the southern part of São Paulo state, and it plays a crucial role for the preservation of the Atlantic Rainforest biome. Grande pond is localized in the state park boundaries and is accessible only by 2-h hiking from Bairro da Serra (Iporanga, SP) through dense vegetation. In the case of Vermelha pond (Apiaí, SP) it is located on a wilderness plateau accessible

by helicopter or a 10-h challenging hike (Fig. 1). These difficulties to reach them limit their access and utilization by humans. As a result, they remain less affected by direct human impact stemming from productive or recreational activities. Although the ponds are remote, they have already been the subject of a geological study by (Saia, 2006), who landed in Vermelha with a helicopter and analyzed the sediments of both ponds, estimating their ages at approximately 4,500 and 1,000 years, respectively. Additionally, the authors proposed that the area has been covered by vegetation for at least *ca.* 16,000 years, and the climate has remained stable and humid for the past 4,500 years (Saia, 2006). These ponds have therefore been evolving naturally for thousands of years with minimal human intervention, making them ideal subjects for microbiomics. In contrast, Furnas pond was artificially built by a mining company probably in the 1950s and is easily accessible.

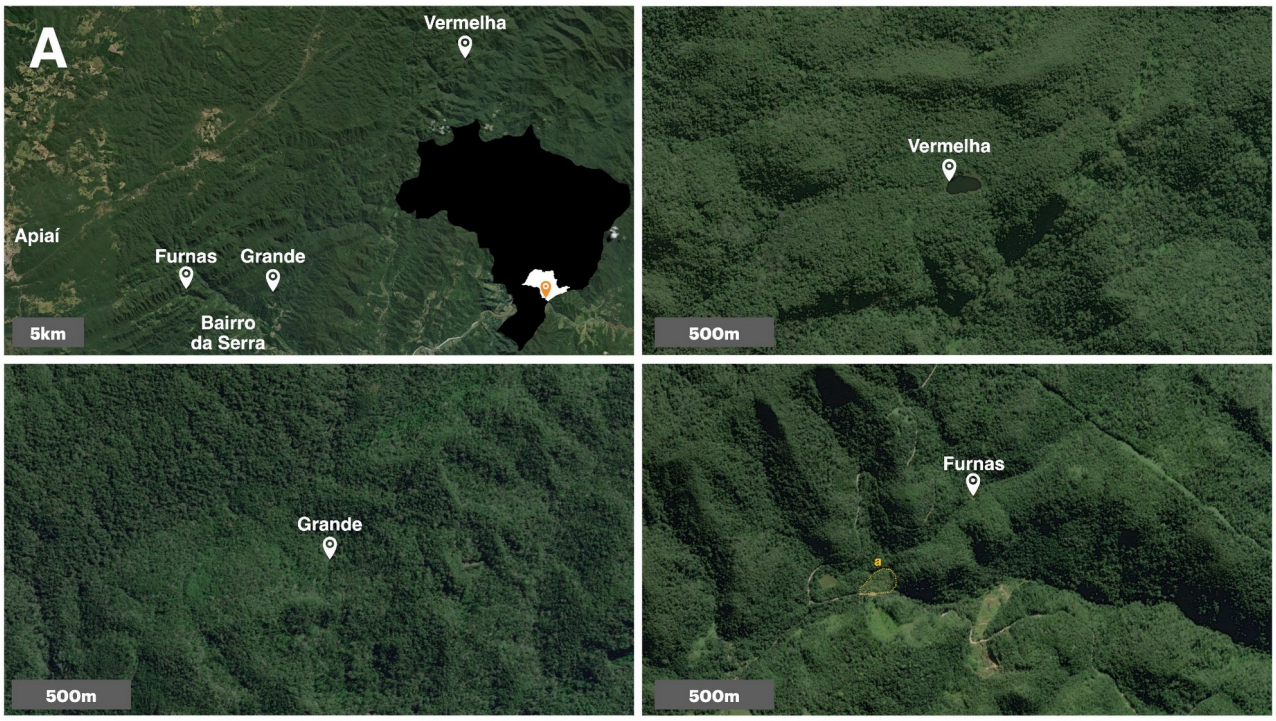


Figure 1. (A) Location and distance of the ponds (Vermelha, Grande and Furnas) to Bairro da Serra, Iporanga, SP, Brazil (inset). Note that Furnas Pond is located around 200m of a dirty road. In the past it supplied water for worker's houses (a). (B) Site photos of Vermelha, Grande and Furnas ponds, respectively.

Beyond enhancing our understanding of biodiversity in these aquatic ecosystems, the data generated in this study offer valuable insights for identifying potential bioindicators of ecosystem health and for uncovering novel genes encoding enzymes with promising biotechnological applications—such as those involved in bioplastic production and bioremediation. By illuminating the unique microbial profiles of ponds within the Atlantic Rainforest, this work lays the groundwork for future studies on the microbiomes of this biome, which are essential for assessing the impacts of human activity and for prospecting genes, organisms, and biochemical pathways of interest. Ultimately, these findings can inform the development of sustainable technologies and evidence-based conservation strategies for this rich and threatened ecosystem.

2. Materials and methods

2.1. Sample collection

The estimated dimensions of Vermelha Pond (VER) are 70 × 100 m (Saia, 2006), with sampling conducted at coordinates 24º23'19"S, 48º31'50"W on May 10, 2021. Furnas Pond (FUR), measuring approximately 50 × 57 m, and Grande Pond (GDE), the smallest at 10 × 30 m, were both sampled on August 18, 2021, at coordinates 24º32'6"S, 48º39'43"W and 24º32'4"S, 48º39'44"W, respectively. The region encompassing PETAR receives an average annual rainfall of approximately 1,800 mm and has a subtropical climate, characterized by warm, humid summers and cooler, drier winters. The rainy season, spanning from October to April, accounts for roughly 80% of the total annual precipitation, with average temperatures ranging from 15°C to 30°C. For each pond, 1 L of water was collected from five different points within the euphotic zone at a depth of 0.5 m. Samples were stored at 4°C and filtered within 24 hours of collection

2.2. Sample filtering, DNA isolation and nanopore sequencing

Water samples were vacuum filtered using membranes with 0.22 µm pore size. Following filtration, each membrane was carefully removed with sterile tweezers and transferred into a PowerWater DNA Bead Tube (QIAGEN). Total genomic DNA was extracted using the DNeasy PowerWater Kit (QIAGEN), with an additional purification step performed using DNeasy columns (QIAGEN) to eliminate potential contaminants. DNA concentration and purity were assessed with a NanoPhotometer N80 (IMPLEN) and quantified using a Qubit fluorometer (Life Technologies). Sequencing libraries were prepared from 450 ng of genomic DNA using the SQK-RAD004 Rapid Sequencing Kit (Oxford Nanopore Technologies, ONT). Sequencing was carried out on a MinION device (ONT) equipped with R10.3 FLO-MIN111 flow cells. Sequencing progress and performance were monitored using the MinKNOW software.

2.3. Bioinformatic analysis

Basecalling was performed using Guppy v.5.0.7 with the high-accuracy configuration model r10.3_450_bps_hac. Metagenomic assemblies were generated using metaFlye v.2.9 (Kolmogorov et al., 2019). To remove potential contaminants, contigs were screened against the NCBI UniVec database. Taxonomic classification was carried out with Kraken2 v.2.1.1 (Lu et al., 2022; Wood et al., 2019), using the "PlusPF" database, comprising of RefSeq sequences for the human genome, prokaryotes, viruses, protozoa, and fungi - downloaded on January 7, 2024. Taxon abundance was estimated by mapping reads back to the classified contigs and counting the number of reads assigned to each taxon. The script used for this step is available at: https://github.com/Sam-Tuna/CCCC. Functional gene annotation was performed using the JGI

IMG/M pipeline (Chen et al., 2023), and the resulting annotated metagenomes, along with associated metadata, have been deposited in the JGI GOLD database (Mukherjee et al., 2023). Screening for possible antimicrobial resistance genes was carried out by DeepARG (Arango-Argoty et al., 2018). Binning of assembled contigs was conducted using metaBAT2 (Kang et al., 2019), CONCOCT (Alneberg et al., 2013), and MaxBin2 (Wu et al., 2016) via the metaWRAP pipeline v.1.3 (Uritskiy et al., 2018), specifically the BINNING, BIN_REFINEMENT, and REASSEMBLE_BINS modules. The pipeline was adapted for long-read data: Minimap2 v.2.17 (Li, 2018) replaced bwa_mem (Li, 2013) for read alignment, and Flye v.2.9 was used instead of SPAdes (Bankevich et al., 2012) for bin reassembly. MAG completeness and contamination were evaluated via CheckM v.1.2 (Parks et al., 2015), and good quality MAGs (>50% completeness, <10% contamination) were classified with GTDB-Tk v.1.7 (Chaumeil et al., 2022) using the r202 database. Samtools v.1.1 (Danecek et al., 2021) was employed for the processing of multiple alignment files.

2.4. Data availability

The raw reads obtained in this project have been deposited in the NCBI Sequence Read Archive under the accession SRR15663634, SRR33480271, and SRR33480272. The assemblies and annotations are available under the JGI GOLD project IDs Gp0614634, Gp0596633 and Gp0569555.

3. Results and discussion

3.1. DNA sequencing parameters and taxonomic profiling

Metagenome information from the Vermelha (VER), Grande (GDE), and Furnas (FUR) ponds is given in Supplementary Table S1. All assemblies achieved an average coverage depth of approximately 9×, sufficient for downstream analyses. Among the three datasets, the VER sample yielded the lowest sequencing throughput but exhibited the highest read N50, reflecting longer average read lengths. Consequently, its assembly was the smallest in total length yet maintained a reasonable N50 value. In contrast, GDE produced the highest sequencing throughput and the largest assembly size, although it had the lowest read and assembly N50 values. The FUR sample provided the most balanced results: while its throughput did not surpass GDE, and its read N50 was lower than VER, the resulting assembly had the highest contig N50 among all three ponds (Table S1), indicating improved continuity and assembly quality.

Across all ponds, the microbial communities were predominantly composed of bacteria, with Pseudomonadota (formerly Proteobacteria) as the most abundant phylum (Fig. 2). Although this study provides only temporal snapshots of the microbial diversity—seasonal fluctuations are expected in aquatic microbiomes—distinct compositional differences were observed between the ponds. The VER sample stood out for its elevated relative abundance of Cyanobacteria and Planctomycetes compared to the other sites. Planctomycetes are often associated with cyanobacteria and diatoms—photosynthetic protists that were also detected in this sample (Kaboré et al., 2020). These differences may be partially explained by seasonal variation: VER was sampled at the beginning of the dry season, in May, whereas FUR and GDE were sampled later in the same season, in August, when the photoperiod was approximately 20 min shorter. A shorter photoperiod can lead to reduced photosynthetic activity across all ponds; however, FUR still exhibited a notable proportion of cyanobacteria, suggesting pond-specific responses to environmental cues.

At the genus level, microbial profiles diverged further. Only three genera—*Streptomyces*, *Escherichia*, and *Pseudomonas*—were consistently detected across all samples at comparable relative abundances. The ubiquity of these genera is notable: *Streptomyces* is renowned for its production of antimicrobial secondary metabolites (De Simeis and Serra, 2021), while *Escherichia* and *Pseudomonas* include species with pathogenic potential. The VER pond was dominated by two genus of cyanobacteria (*Microcystis* and *Leptolyngbya*) and two Planctomycetes (*Fuerstiella* and *Gimesia*), all of which were virtually absent from the other two ponds, indicating habitat specificity. The FUR Pond displayed the highest genus-level diversity. Dominant taxa included *Limnohabitans*, a relatively understudied genus of Burkholderiales previously isolated from another pond in São Paulo state (Hahn et al., 2010); *Planktophila*, an Actinobacterium; and *Prochlorococcus*, a globally distributed cyanobacterium notably absent from VER and GDE (Biller et al., 2015). Other genera found in FUR included *Flavobacterium* (Bacteroidota), *Acinetobacter* (Gammaproteobacteria), and *Opitutus* (Verrucomicrobia), the latter represented by only one described species, *Opitutus terrae*. Finally, the microbial community in GDE pond was more evenly distributed among the most abundant genera. *Rhizobium*, whose species are generally plant root-associated, was the most abundant taxon, followed by *Streptomyces*, *Escherichia*, *Nostoc* (a cyanobacterium also found in FUR but absent in VER), *Asticcacaulis* (a stalked Proteobacterium), and a set of other genera including *Phenylobacterium*, *Bradyrhizobium*, *Curvibacter*, *Rhodoferax*, and *Novosphingobium*.

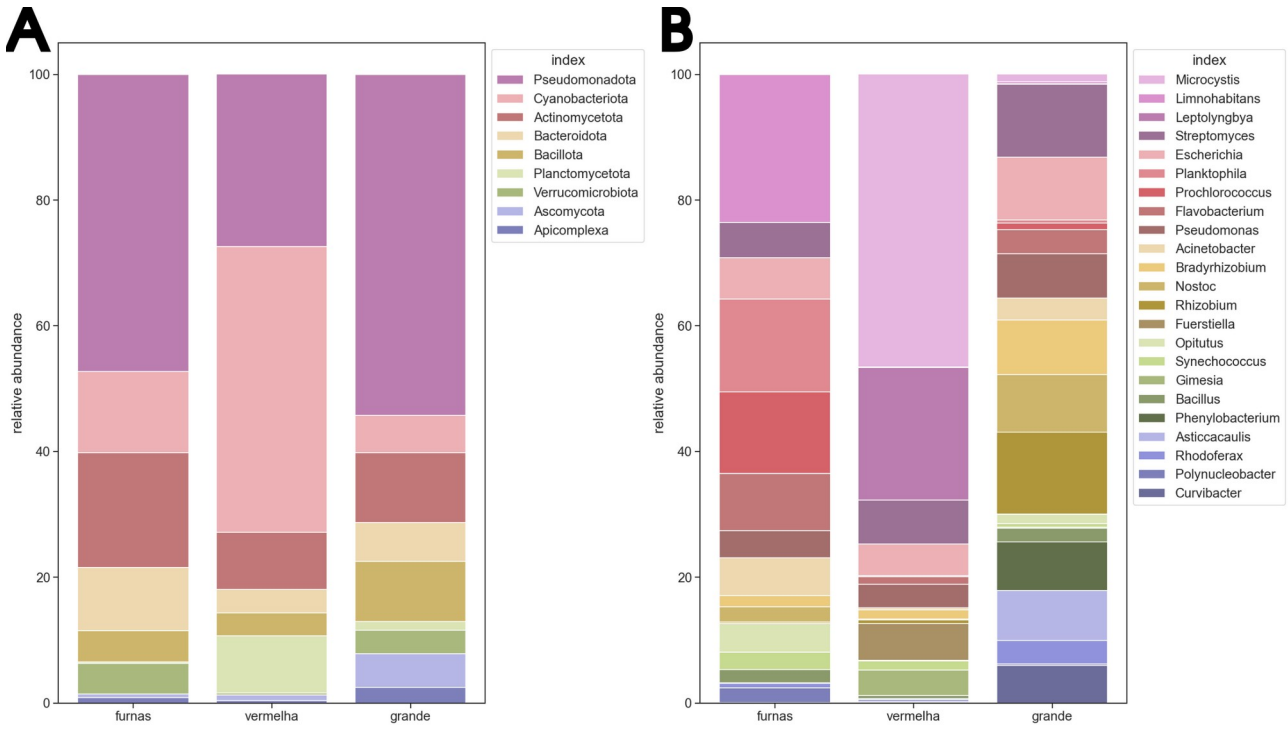


Figure 2. (A) Stacked bar graph showing the 10 most abundant phyla on average, across each pond. (B) Stacked bar graphs showing the 23 most abundant microbial genera. In both graphs the data was normalized and scaled within each sample so they would sum up to 100. The full genus level classification graph is in SI (Figure S1) and the full classification results are in Supplementary File S1.

Although the primary focus of this study was not the targeted collection of viral particles, a substantial number of viral DNA sequences were recovered from the metagenomic data. Given the sequencing strategy employed, only double-stranded DNA (dsDNA) viruses were detected. Most of these viral sequences were classified as belonging to *Uroviricota* (primarily bacteriophages) and *Nucleocytoviricota* (eukaryotic viruses). Despite their relatively low representation in the total microbial communities—comprising 0.26% of relative abundance in GDE, 0.66% in VER, and 1.28% in FUR—these viral assemblages revealed notable differences in composition across the three ponds.

Viral genus-level diversity varied markedly between sites, with most genera detected being pond-specific. GDE exhibited the highest number of operational taxonomic units (OTUs) and the most even distribution among viral taxa, suggesting a more diverse and balanced viral community. In contrast, FUR was predominantly dominated by *Prasinovirus*, while VER showed a higher abundance of *Nodensvirus*, *Mimivirus*, and *Sokavirus*. Notably, *Prasinovirus* was the only viral genus consistently found across all three ponds, albeit with striking differences in relative

abundance—it was highly prevalent in FUR and markedly less abundant in VER. As *Prasinovirus* is known to infect unicellular green algae (prasinophytes), its widespread presence serves as an indirect indicator of algal prevalence in these ecosystems.

In addition to the viral component, archaeal DNA was also detected, albeit in low abundance. Archaea accounted for approximately 0.1% of the total community in VER and FUR, and 0.5% in GDE. These archaeal sequences were primarily affiliated with the phylum *Euryarchaeota*, as well as members of the TACK superphylum, particularly *Thermoproteota* and *Nitrososphaerota*, indicating a minor but taxonomically diverse archaeal presence within these pond microbiomes.

3.2. Metagenome-Assembled Genomes (MAGs)

Although most metagenomic reads were taxonomically assigned to Pseudomonadota, the MAGs encompassed a broader phylogenetic spectrum, including members of Dependentiae, Cyanobacteria, Chloroflexota, Chlamydiota, Planctomycetota, Actinobacteriota, Verrucomicrobiota, Patescibacteria, Eisenbacteria, and Bacteroidota (Table 1). Among the three sites, FUR exhibited the greatest MAG diversity, with several phyla—Actinobacteriota, Bacteroidota, Patescibacteria, and Eisenbacteria—recovered exclusively from this pond. VER also showed unique MAGs, particularly those classified within Dependentiae, Chlamydiota, and Planctomycetota, while GDE was the least diverse, yielding only four MAGs (despite having the highest number of reads): three from Pseudomonadota and one from Verrucomicrobiota. No MAGs from the same taxonomic order were found in more than one site. This absence of shared orders among MAGs highlights the ecological distinctiveness of each pond and suggests microhabitat-driven divergence of microbial populations. Moreover, the number of MAGs recovered per pond was positively correlated with metagenome assembly quality, as measured by N50 values (Figure S2). FUR, which had the highest assembly N50, yielded the greatest number of MAGs, followed by VER and then GDE. This underscores the impact of sequencing and assembly quality on genome recovery from complex microbial communities.

Table 1

MAGs recovered from each pond.[a]

Pond	bin	Completeness (%)	Contamination (%)	GC (%)	Taxonomy[b]	N50 (Mbp)	Size (Mbp)
VER	8	94.3	0.634	0.529	Nodosilinea (g)	3.79	3.79
	2	93.4	0.235	0.445	CAIUCS01 sp903897715 (s)	3.28	3.28

	6	84.6	0	0.412	Rhabdochlamydiaceae (f)	0.36	1.61
	3	76.2	7.345	0.364	Rickettsiaceae (f)	0.11	1.49
	7	75.2	7.001	0.578	SXXK01(g)	0.64	8.89
	5	63.9	5.818	0.516	CAIRDJ01 sp903877335 (s)	0.07	3.45
	4	56.8	5.007	0.43	*Microcystis novacekii* (s)	0.04	5.28
	1	51.4	1.626	0.354	CAIQNR01 (f)	0.63	0.86
FUR	11	97.6	0.25	0.34	Ferruginibacter (g)	2.22	2.63
	5	94.0	0.28	0.5	Limnohabitans (g)	2.78	2.79
	13	86.1	0	0.42	AAA044-D11 (g)	0.99	1.34
	4	83.4	2.27	0.35	Sediminibacterium (g)	0.27	2.35
	2	80.3	6.03	0.39	CAIXMG01 (g)	0.50	0.77
	6	80.0	0	0.64	BAIKAL-G1 (g)	0.20	3.56
	8	69.0	1.92	0.45	Polynucleobacter (g)	0.11	2.46
	10	67.7	3.38	0.56	UBA967 (g)	0.12	3.15
	3	67.6	9.56	0.69	CADEFY01 (g)	0.11	2.71
	12	62.1	4.22	0.44	Nanopelagicaceae (f)	0.08	1.20
	1	60.9	0.85	0.64	RAAP-2 (g)	0.08	1.70
	9	58.6	0	0.44	Planktophila (g)	0.11	1.22
	7	56.9	3.04	0.52	UBA965 (g)	0.62	3.44
GDE	2	94.5	0.324	484	Caulobacteraceae (f)	2.54	2.54
	3	89.6	0.71	530	Rhizobiaceae (f)	2.14	4.26
	1	87.1	3.625	638	Phenylobacterium (g)	2.21	3.91
	4	60.1	0.198	605	UBA6821 (g)	0.48	2.66

[a] Entire Genome Taxonomy Database (GTDB) taxonomy is available in Supplementary File S2. [b] Family (f), genus (g) and species (s).

3.3. Annotations

Functional annotation of the assembled metagenomes revealed a wide array of gene functions across all three ponds (File S3). Most identified genes were associated with fundamental cellular processes such as protein and nucleic acid metabolism, cell wall biosynthesis, membrane receptor signaling, and multidrug efflux mechanisms. Putative antimicrobial resistance genes (ARGs) were identified in all three ponds. FUR exhibited the highest number of ARGs, with a total of 13 genes, followed by GDE with 5, and VER with 2. The elevated number of resistance genes in FUR—an artificial pond with greater human accessibility—suggests a potential anthropogenic

influence, possibly due to environmental contamination with antibiotic residues transported via rainwater runoff or surface flow from surrounding areas.

It is important to note, however, that antimicrobial resistance genes can also be naturally occurring, independent of human impact. For example, members of the genus *Streptomyces*—known producers of antibiotics —were detected in all three ponds and may partially be the reason ARGs were identified in these more remote sites. In addition to confirmed ARGs, several more general resistance-associated genes were found across all samples, including those encoding ABC-type multidrug transporters and AcrAB-TolC efflux pump components. The widespread occurrence of these resistance elements, even in relatively undisturbed environments, emphasizes the complexity of resistance dynamics in natural microbial communities and reinforces the importance of prudent antibiotic use and disposal practices.

Beyond ARGs, the annotation also uncovered genes of environmental and biotechnological interest. These included genes involved in cyanotoxin production, bacteriocin synthesis, and polyhydroxyalkanoate (PHA) biosynthesis. In VER metagenome, the complete microcystin biosynthetic gene cluster was identified, distributed across 70 contigs. Additionally, seven contigs from the GDE assembly also contained microcystin-related genes. Regarding bacteriocins, two contigs encoding the bacteriocin *curvaticin FS47* were identified—one in VER and one in FUR. Genes involved in PHA biosynthesis were less prevalent, with only a single *phaC* gene detected, located in the VER metagenome. These findings point to the biotechnological potential harbored within these microbial communities, particularly in the context of natural product biosynthesis and sustainable biopolymer production.

Nevertheless, a substantial proportion of predicted proteins remained unclassified due to limitations in current annotation databases, which are largely biased toward clinical and temperate-region model organisms. Specifically, approximately 18% of predicted proteins in FUR dataset, 21% in GDE, and 23% in VER could not be functionally annotated. This underlines the ongoing challenge of studying microbial communities in underexplored environments and highlights the need to expand reference databases with genomes and proteins from diverse, ecologically relevant sources.

4. Conclusion

Most microbial DNA sequences recovered from the pond samples were of bacterial origin. While some phylum-level differences were observed—most notably the elevated presence of *Cyanobacteria* in Vermelha—a genus-level comparison more effectively captured the distinctiveness and complexity of each microbial community, revealing highly diverse and unique

assemblages across the three ponds. Most of the recovered MAGs corresponded to uncultured or previously uncharacterized taxa, and their phylogenetic diversity significantly exceeded that observed in the read-based taxonomic classifications. This discrepancy underscores the novelty of the sampled environments and highlights the limitations of current reference databases. Notably, the most abundant taxon in each pond was represented by a MAG: *Microcystis novacekii* in VER, a *Limnohabitans* MAG in FUR, and a MAG classified only to the family *Rhizobiaceae* in GDE. Except for *M. novacekii*, which is a known and cultured cyanobacterial species, the other MAGs remain hypothetical—assembled from sequence data without corresponding cultured isolates—illustrating the challenges in verifying the existence of these organisms without cultivation-based validation.

Interestingly, several MAGs were classified to taxa that were not highly abundant in the initial classification results, suggesting that even low-abundance organisms may achieve sufficient sequencing coverage for genome reconstruction. However, many of these MAGs could not be confidently assigned to known species, further emphasizing the extent to which microbial diversity in novel or underexplored environments remains uncharacterized due to gaps in existing databases.

This database incompleteness also extends to functional annotation. On average, approximately 20% of predicted proteins in each pond's metagenome could not be assigned any known function. Nonetheless, several functionally relevant genes were identified. The complete biosynthetic gene cluster responsible for microcystin production was recovered from the VER metagenome, along with an additional copy of the *mycC* gene in the GDE sample. Furthermore, a gene associated with polyhydroxyalkanoate (PHA) synthesis was found in VER, and the bacteriocin gene *curvaticin FS47* was identified in both VER and FUR ponds.

Regarding antimicrobial resistance, all three ponds contained a total of 20 resistance genes, as identified by the deepARG tool. These genes were most concentrated in FUR, the pond with the highest degree of human activity and accessibility. This pattern suggests a potential anthropogenic influence on the dissemination of resistance genes.

It is important to note that existing genomic and functional databases are heavily biased toward microorganisms from North America and Europe. As such, the relatively remote and understudied South American ecosystems represented in this study are likely underrepresented in reference datasets. These findings underscore the importance of conducting exploratory metagenomic studies in poorly sampled biomes and reinforce the notion that, while metagenomics is a powerful tool for characterizing microbial diversity, it remains dependent on cultivated isolates and continued advances in microbial genomics to fully unlock the functional and taxonomic complexity of environmental microbiomes.

Credit authorship contribution statement

SVFA: Bioinformatic analysis, graphs and statistics. DMMS: experimental design, sample preparation and sequencing. DMMS, IS, GAJ, AHRD, CVS and RSF: sample collection. SVFA, DMMS, EJHB, JCS, LU, CVS and RSF: experimental design and draft manuscript preparation. All authors contributed to the article and approved the submitted version.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, 2017/22501-2 CVS, 2022/03548-6 RSF, 2019/12605-0 DMMS, 2021/10577-0 JCS), The Brazilian National Council for Scientific and Technological Development (CNPq, 303525/2021-5 CVS, 301623/2019-8 EJHB, 305961/2019-5 JCS), the Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES, 88887.514039/2020-00 and 88887.836021/2023-00 SVFA) and Helmholz Information & Data Science Academy (HiDA) to SVFA). We also gratefully acknowledge the support of the RCGI – Research Centre for Greenhouse Gas Innovation, hosted by the University of São Paulo (USP), sponsored by FAPESP and Shell Brasil under grants 2014/50279-4 and 2020/15230-5 (for CVS and RSF), and the strategic importance of the support given by ANP (Brazil's National Oil, Natural Gas and Biofuels Agency) through the R&D levy regulation. Graphical abstract created in BioRender. Urban, L. (2025) https://BioRender.com/pv6772p.

References

Alneberg, J., Bjarnason, B.S., Bruijn, I. de, Schirmer, M., Quick, J., Ijaz, U.Z., Loman, N.J., Andersson, A.F., Quince, C., 2013. CONCOCT: Clustering cONtigs on COverage and ComposiTion. https://doi.org/10.48550/arXiv.1312.4038

Arango-Argoty, G., Garner, E., Pruden, A., Heath, L.S., Vikesland, P., Zhang, L., 2018. DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. Microbiome 6, 23. https://doi.org/10.1186/s40168-018-0401-z

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. Journal of Computational Biology 19, 455–477. https://doi.org/10.1089/cmb.2012.0021

Biggs, J., von Fumetti, S., Kelly-Quinn, M., 2017. The importance of small waterbodies for biodiversity and ecosystem services: implications for policy makers. Hydrobiologia 793, 3–39. https://doi.org/10.1007/s10750-016-3007-0

Biller, S.J., Berube, P.M., Lindell, D., Chisholm, S.W., 2015. Prochlorococcus: the structure and function of collective diversity. Nat Rev Microbiol 13, 13–27. https://doi.org/10.1038/nrmicro3378

Chaumeil, P.-A., Mussig, A.J., Hugenholtz, P., Parks, D.H., 2022. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. Bioinformatics 38, 5315–5316. https://doi.org/10.1093/bioinformatics/btac672

Chen, I.-M.A., Chu, K., Palaniappan, K., Ratner, A., Huang, J., Huntemann, M., Hajek, P., Ritter, S.J., Webb, C., Wu, D., Varghese, N.J., Reddy, T.B.K., Mukherjee, S., Ovchinnikova, G., Nolan, M., Seshadri, R., Roux, S., Visel, A., Woyke, T., Eloe-Fadrosh, E.A., Kyrpides, N.C., Ivanova, N.N., 2023. The IMG/M data management and analysis system v.7: content updates and new features. Nucleic Acids Research 51, D723–D732. https://doi.org/10.1093/nar/gkac976

Chen, Y.-C., Liu, T., Yu, C.-H., Chiang, T.-Y., Hwang, C.-C., 2013. Effects of GC Bias in Next-Generation-Sequencing Data on De Novo Genome Assembly. PLOS ONE 8, e62856. https://doi.org/10.1371/journal.pone.0062856

Chopyk, J., Nasko, D.J., Allard, S., Bui, A., Pop, M., Mongodin, E.F., Sapkota, A.R., 2020. Seasonal dynamics in taxonomy and function within bacterial and viral metagenomic assemblages recovered from a freshwater agricultural pond. Environmental Microbiome 15, 18. https://doi.org/10.1186/s40793-020-00365-8

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., Li, H., 2021. Twelve years of SAMtools and BCFtools. GigaScience 10, giab008. https://doi.org/10.1093/gigascience/giab008

De Simeis, D., Serra, S., 2021. Actinomycetes: A Never-Ending Source of Bioactive Compounds—An Overview on Antibiotics Production. Antibiotics (Basel) 10, 483. https://doi.org/10.3390/antibiotics10050483

Downing, J.A., 2010. Emerging global role of small lakes and ponds : little things mean a lot. Limnética 29, 0009–0024.

Hahn, M.W., Kasalický, V., Jezbera, J., Brandt, U., Šimek, K., 2010. Limnohabitans australis sp. nov., isolated from a freshwater pond, and emended description of the genus Limnohabitans. Int J Syst Evol Microbiol 60, 2946–2950. https://doi.org/10.1099/ijs.0.022384-0

Kaboré, O.D., Godreuil, S., Drancourt, M., 2020. Planctomycetes as Host-Associated Bacteria: A Perspective That Holds Promise for Their Future Isolations, by Mimicking Their Native Environmental Niches in Clinical Microbiology Laboratories. Front Cell Infect Microbiol 10, 519301. https://doi.org/10.3389/fcimb.2020.519301

Kang, D.D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., Wang, Z., 2019. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ 7, e7359. https://doi.org/10.7717/peerj.7359

Kolmogorov, M., Yuan, J., Lin, Y., Pevzner, P.A., 2019. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol 37, 540–546. https://doi.org/10.1038/s41587-019-0072-8

Li, H., 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H., 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM [WWW Document]. arXiv.org. URL https://arxiv.org/abs/1303.3997v2 (accessed 3.12.25).

Lu, J., Rincon, N., Wood, D.E., Breitwieser, F.P., Pockrandt, C., Langmead, B., Salzberg, S.L., Steinegger, M., 2022. Metagenome analysis using the Kraken software suite. Nat Protoc 17, 2815–2839. https://doi.org/10.1038/s41596-022-00738-y

Meslier, V., Quinquis, B., Da Silva, K., Plaza Oñate, F., Pons, N., Roume, H., Podar, M., Almeida, M., 2022. Benchmarking second and third-generation sequencing platforms for microbial metagenomics. Sci Data 9, 694. https://doi.org/10.1038/s41597-022-01762-z

Mukherjee, S., Stamatis, D., Li, C.T., Ovchinnikova, G., Bertsch, J., Sundaramurthi, J.C., Kandimalla, M., Nicolopoulos, P.A., Favognano, A., Chen, I.-M.A., Kyrpides, N.C., Reddy, T.B.K., 2023. Twenty-five years of Genomes OnLine Database (GOLD): data updates and new features in v.9. Nucleic Acids Research 51, D957–D963. https://doi.org/10.1093/nar/gkac974

Oertli, B., Biggs, J., Céréghino, R., Grillas, P., Joly, P., Lachavanne, J.-B., 2005. Conservation and monitoring of pond biodiversity: introduction. Aquatic Conservation: Marine and Freshwater Ecosystems 15, 535–540. https://doi.org/10.1002/aqc.752

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

Saia, S.E.M.G., 2006. Reconstrução paleoambiental (vegetação e clima) no Quaternário tardio com base em estudo multi/interdisciplinar no Vale do Ribeira (sul do Estado de São Paulo) (text). Universidade de São Paulo. https://doi.org/10.11606/T.64.2006.tde-04092006-151920

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

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

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