Faculty of Medicine

Systems biology

Department of Human and Medical Genetics

Ignas Juodžiūnas, second year

Master's thesis

Melsvabakterių sukcesijos poveikis mikroorganizmų bendrijos kaitai

Cyanobacteria Driven Dynamics of Microbial Communities

Supervisor Dr. Sigitas Šulčius

Head of the department or clinic Prof. Dr. Algirdas Utkus

Vilnius, 2026

Student's email ignas.juodziunas@mf.stud.vu.lt

CONTENTS

LIST OF ABBREVIATIONS

16S rRNA	16S ribosomal ribonucleic acid
AFA	*Aphanizomenon flos-aquae*
AMG	auxiliary metabolic gene
bp	base pairs
CRISPR	clustered regularly interspaced short palindromic repeats
DNA	deoxyribonucleic acid
DRAM-v	distilled and refined annotation of metabolism for viruses
Gbp	gigabase pairs
GTDB	Genome Taxonomy Database
HMM	hidden Markov model
HPC	high-performance computing
HQ	high-quality
iPHoP	integrated phage host prediction
kbp	kilobase pairs
LQ	low-quality
MAG	metagenome-assembled genome
Mbp	megabase pairs
MIUViG	minimum information about an uncultivated virus genome
MQ	medium-quality
N50	length at which contigs of that size or longer cover 50% of the assembly
ONT	Oxford Nanopore Technologies
rRNA	ribosomal ribonucleic acid
SPIEC-EASI	sparse inverse covariance estimation for ecological association inference
StARS	stability approach to regularization selection

INTRODUCTION

Cyanobacterial blooms are an increasingly frequent consequence of eutrophication in freshwater systems worldwide (Huisman *et al.*, 2018), yet the microbial and viral community dynamics they drive remain poorly resolved at the genomic level. Most metagenomic studies of bloom-associated communities have relied on short-read sequencing, which produces fragmented assemblies that limit genome recovery and obscure the complete genetic context needed to infer microbial interactions and virus-host linkages. Nanopore long-read sequencing addresses these limitations but lacks standardised, publicly available pipelines for freshwater metagenomics. Furthermore, the majority of long-read metagenomic studies have been conducted at single timepoints, leaving the temporal dynamics of bloom-associated microbial and viral communities largely uncharacterised. This thesis addresses these gaps by developing a reproducible nanopore-only metagenomics pipeline and applying it to a temporally resolved dataset capturing microbial and viral community dynamics during cyanobacterial bloom activity in a eutrophic freshwater pond.

Aims and Objectives

Aim:

This thesis aims to characterize the dynamics of microbial communities driven by cyanobacterial bloom activity in a eutrophic freshwater pond ecosystem, using a reproducible nanopore-only long-read metagenomics pipeline. Specific focus is placed on the recovery and taxonomic classification of metagenome-assembled genomes and viral sequences, the inference of bacterial co-occurrence networks and guild-level temporal dynamics, and a preliminary assessment of viral community composition and putative functional gene content.

Objectives:

1. Develop a reproducible nanopore-only bioinformatic pipeline for processing freshwater metagenomes, from raw reads through to genome assembly, quality assessment, and taxonomic classification.

2. Recover and identify metagenome-assembled genomes (MAGs) and viral genomes from freshwater pond samples collected during a cyanobacterial bloom, focusing on genomic features that short-read sequencing struggles to resolve.

3. Build microbial co-occurrence networks, track how bacterial community guilds change over time during bloom succession, and characterise the composition and putative functional gene content of the viral community.

4. Evaluate the ecological significance of the observed microbial community dynamics in the context of cyanobacterial bloom succession and assess the strengths and limitations of nanopore-only metagenomics for this type of analysis.

LAY DESCRIPTION

Freshwater ponds and lakes can experience rapid and dramatic changes in their microscopic inhabitants during summer months, when certain types of bacteria called cyanobacteria grow explosively to form blooms. These blooms affect water quality and aquatic life, and when they die off, they release large amounts of organic matter that feeds other microorganisms including bacteria and viruses. Understanding how these microscopic communities change over time and interact with each other is important for managing freshwater ecosystems.

This thesis used a DNA sequencing technology called nanopore sequencing to study the microorganisms living in a fish hatchery pond in Lithuania over the course of a cyanobacterial bloom event. By reading the genetic material collected from the pond water at ten different timepoints, it was possible to identify which bacteria and viruses were present, track how their communities changed as the bloom developed and declined, explore how different groups of bacteria interact with each other and viruses, and directly link these shifts to the rise and fall of the bloom using cell count data of the bloom-forming cyanobacterium *Aphanizomenon flos-aquae.*

LITERATURE REVIEW

This literature review examines how nanopore-only long-read metagenomics enables genome-resolved analysis of microbial and viral communities in freshwater ecosystems. The review begins with the methodological foundations of long-read metagenomics, covering assembly, polishing, and genome recovery, before addressing the ecological context of this study, covering cyanobacterial bloom dynamics and the freshwater bacterioplankton communities that respond to them. It then discusses approaches to inferring microbial interaction networks from metagenomic data and the role of viruses within these networks, concluding with current knowledge gaps and future directions relevant to the study system.

1. The Genomic Revolution: From Short Reads to Complete Genomes

Since the advent of high-throughput sequencing, short-read platforms such as Illumina have dominated metagenomic studies due to their high accuracy, throughput, and relatively low cost. However, these platforms face fundamental limitations when reconstructing genomes from complex microbial communities. Repetitive sequences, genomic rearrangements, and structurally complex regions such as CRISPR arrays, prophages, and biosynthetic gene clusters are difficult to resolve with short reads, and the resulting assemblies are often highly fragmented, consisting of hundreds or thousands of contigs per genome. This fragmentation complicates downstream analyses including gene prediction, functional annotation, and the identification of mobile genetic elements.

These limitations are particularly consequential when studying microbial interaction networks, where complete genomic context is essential. Identifying complete prophage sequences requires continuous assembly across both viral and host DNA, while mapping horizontal gene transfer events demands intact plasmid and transposon sequences. Fragmented assemblies also hinder accurate taxonomic classification, as 16S rRNA genes and other marker sequences may be split across multiple contigs or absent entirely.

Long-read sequencing technologies, particularly Oxford Nanopore (ONT) and PacBio platforms, have begun to address these shortcomings by generating reads ranging from several kilobases to megabases in length. These extended reads can span repetitive elements and resolve complex genomic structures that confound short-read assembly (Kim, Pongpanich and Porntaveetus, 2024),

and their ability to capture complete operons and gene clusters in single reads has substantially expanded the scope of genome-resolved metagenomics.

A key question in the field is whether short reads remain necessary or whether nanopore data can stand alone for metagenomic analysis. Liu et al. demonstrated that ONT-only metagenomics can produce high-quality results from activated sludge, recovering 275 MAGs with approximately 90% median completeness and N50 values around 735 kb, representing a 45-86x improvement in contiguity over short-read assemblies from the same samples (Liu *et al.*, 2022). A direct comparison across three sequencing strategies, Illumina-only, ONT-only, and hybrid, found that the hybrid approach recovered approximately twice as many MAGs as Illumina alone, while ONT-only assemblies produced the most contiguous individual genomes despite recovering fewer total MAGs (Overholt *et al.*, 2020). This contiguity advantage is particularly valuable for studies focused on complete metabolic pathways or mobile genetic elements, and ONT assemblies have also been shown to recover approximately four times more full-length 16S rRNA genes than Illumina, with direct implications for community profiling and phylogenetic analysis (Overholt *et al.*, 2020).

2. Technical Considerations for Nanopore Metagenomics

The choice of assembler has a substantial impact on the quality of metagenomic reconstructions from long-read data. metaFlye, an extension of the Flye assembler designed specifically for long-read metagenomics, addresses key challenges such as uneven coverage and intra-species strain heterogeneity using repeat graphs (Kolmogorov *et al.*, 2020). In both simulated and real datasets, metaFlye consistently produced assemblies with higher completeness and contiguity than other long-read assemblers, and its application to a sheep rumen microbiome enabled the recovery of 63 complete or near-complete bacterial genomes, many assembled as single contigs. A subsequent benchmarking study comparing Flye, Raven, and Redbean on ONT mock community data found that Flye was the most robust assembler, consistently yielding high completeness across taxa and successfully recovering plasmids that alternative tools missed (Buttler and Drown, 2022). More recently, nanoMDBG has emerged as a promising successor specifically optimised for high-accuracy ONT reads, reported to recover up to twice as many high-quality MAGs as metaFlye at a third of the computational cost (Benoit *et al.*, 2026), representing an important advance for future nanopore-only metagenomic workflows.

Despite their advantages in read length, nanopore reads historically exhibited higher error rates than short-read data, necessitating polishing steps to obtain accurate consensus sequences. The two most widely used tools for this purpose are Racon and Medaka. Racon corrects general sequencing errors through read-to-contig alignment, while Medaka applies neural network models trained on nanopore data to address platform-specific errors that alignment-based methods do not fully resolve. Studies evaluating polishing strategies have found that Medaka alone improves assembly completeness more than Racon alone, and that combining both tools produces higher-quality assemblies than individually (Lee *et al.*, 2021). With the introduction of R10.4 flow cell chemistry and improved basecalling models, per-read accuracy has improved substantially compared to earlier chemistries, and recent work has demonstrated that long-read-only pipelines using Flye and Medaka can recover near-finished bacterial genomes from metagenomes without short-read polishing (Sereika *et al.*, 2022).

Accurate quality assessment of recovered MAGs is essential for ensuring the reliability of downstream analyses. CheckM2 uses a machine learning approach to estimate genome completeness and contamination and has been shown to outperform earlier marker gene-based methods, particularly for novel and uncultivated lineages that are frequently encountered in environmental metagenomes (Chklovski *et al.*, 2023). For taxonomic classification, GTDB-Tk v2 assigns taxonomy based on placement within a phylogenetically standardised reference tree, and its substantially reduced memory requirements compared to earlier versions have made it more practical to deploy in large-scale metagenomic studies without compromising classification accuracy (Chaumeil *et al.*, 2022).

3. Cyanobacterial Bloom Dynamics and Associated Microbial Succession

Cyanobacterial blooms are recurrent phenomena in nutrient-enriched freshwater systems, driven primarily by elevated phosphorus and nitrogen concentrations, warm water temperatures, and stable thermal stratification. In eutrophic ponds and lakes, bloom-forming genera such as *Aphanizomenon*, *Microcystis*, and *Dolichospermum* can rapidly dominate the phytoplankton community during summer months, producing dense surface accumulations and releasing cyanotoxins that threaten aquatic biodiversity and water quality. Bloom dynamics are rarely simple or unimodal. Temporal succession within the cyanobacterial community itself has been documented across multiple systems, with early-season dominance by filamentous genera giving

way to toxin-producing species as conditions change, followed by a transition back to heterotrophic bacterioplankton dominance during senescence (Li *et al.*, 2020).

The organic matter released during cyanobacterial bloom senescence, including polysaccharides, proteins, pigment-protein complexes, and nucleic acids, represents a large and chemically diverse substrate pulse that drives predictable succession within the heterotrophic bacterial community. Members of the phylum Bacteroidota, particularly families Flavobacteriaceae and Sphingobacteriaceae, are consistently among the first and most abundant responders to bloom-derived organic matter, a pattern described as "blooming after the bloom" (Eiler and Bertilsson, 2007). These bacteria produce a wide repertoire of extracellular glycoside hydrolases, polysaccharide lyases, and proteases that enable enzymatic breakdown of complex biopolymers released during algal lysis. Recent work has demonstrated that this polysaccharide processing is highly selective, with certain Bacteroidota genera including *Flavobacterium*, *Sediminibacterium*, and *Luteolibacter* preferentially assimilating specific polysaccharide fractions in a manner consistent with niche partitioning during bloom degradation (Čačković *et al.*, 2025).

The temporal trajectory of decomposer populations typically follows the bloom peak with a characteristic lag, reflecting the time required for bacterial populations to detect and respond to the newly available substrates. Nitrogen-rich compounds released during early bloom senescence, such as proteins and photosynthetic pigments, are broken down more rapidly than structurally complex polysaccharides, suggesting that different bacterial groups process different components of the bloom-derived organic matter at different times. This cascade of resource utilisation drives a predictable community succession that can span several weeks, linking the rise and fall of cyanobacterial biomass directly to the composition and dynamics of the surrounding heterotrophic community.

4. Freshwater Bacterioplankton Ecology

Freshwater bacterioplankton communities are structured by a combination of resource availability, physical mixing, and biotic interactions, and their composition shifts substantially in response to seasonal and event-driven changes in organic matter supply. Several bacterial lineages have been identified as characteristic and recurring members of productive freshwater systems, and their ecological roles are increasingly well understood through culture-independent genomic approaches.

Polynucleobacter necessarius subspecies *asymbioticus* is among the most cosmopolitan and abundant freshwater bacteria globally, typically comprising 10-20% of total bacterioplankton in productive standing water bodies (Hahn *et al.*, 2012). Its ecological success has been attributed to its highly streamlined genome, rapid growth rate, and capacity to exploit dissolved organic carbon released by phytoplankton. Its abundance tends to peak during warmer, nutrient-rich periods, making it a characteristic member of bloom-associated communities.

The acI lineage of Actinomycetota, represented by genera including *Candidatus* Planktophila, is a ubiquitous component of freshwater bacterioplankton, typically accounting for a substantial proportion of total cell counts in mesotrophic and eutrophic systems. Genome streamlining within this lineage has resulted in auxotrophies for vitamins, amino acids, and reduced sulphur compounds, creating metabolic dependencies on co-occurring phytoplankton and other bacteria that supply these resources as exudates or lysis products (Neuenschwander *et al.*, 2018). Their abundance is closely coupled to phytoplankton productivity, and they frequently reach peak abundance during or immediately following bloom events.

Candidatus Nanopelagicales represents one of the most numerically dominant bacterioplankton lineages in many freshwater lakes globally. Members of this group are characterised by small cell volumes, reduced genome sizes, and metabolisms adapted to a free-living, low-nutrient lifestyle (Mondav *et al.*, 2020). Their abundance typically increases following bloom decline, when reduced particle load and nutrient concentrations favour their competitive strategy over that of larger, bloom-associated bacteria.

Verrucomicrobiota, particularly the genus *Luteolibacter*, are well-established polysaccharide degraders in freshwater systems. Metagenomic analyses have revealed high abundances of carbohydrate-active enzyme genes in their genomes, and experimental evidence has confirmed active uptake of algal-derived polysaccharides in lake systems (He *et al.*, 2017). Members of Comamonadaceae are commonly abundant in freshwater bacterioplankton and are capable of rapid growth under elevated nutrient conditions, degrading a wide range of organic compounds (Willems, 2014).

Together, these lineages form the ecological backbone of productive freshwater bacterioplankton communities, and their temporal dynamics during bloom events provide a framework for interpreting the co-occurrence networks and guild succession patterns explored in this study.

5. Microbial Interaction Networks: From Genomes to Ecology

Inferring microbial interaction networks from metagenomic data presents methodological challenges arising from the compositional nature of sequencing data and the high dimensionality of microbial community profiles. Standard correlation-based approaches are prone to producing spurious associations in compositional datasets, where the relative abundance of one taxon is mathematically constrained by the abundances of all others. SPIEC-EASI addresses this limitation through sparse inverse covariance estimation, which infers conditional dependencies between taxa rather than simple co-occurrence patterns (Kurtz *et al.*, 2015). In benchmarking analyses, this approach outperformed earlier correlation-based methods by recovering more accurate network structures, particularly in complex datasets such as the human gut microbiome.

Network-based approaches have been applied to investigate microbial and viral community dynamics in freshwater systems. One study used correlation-based network analysis to examine interactions between MAGs and viral genomes across freshwater mesocosm (experimentally controlled aquatic enclosures) samples subjected to multiple environmental stressors, finding that combined stressors such as warming, nutrient enrichment, and pesticide exposure produced simpler and more fragile networks with fewer connections and reduced overall robustness (Wang *et al.*, 2025). These findings demonstrate that environmental pressures can reshape virus-bacteria interaction networks in ways that are detectable through co-occurrence analysis.

Beyond co-abundance relationships, more mechanistic approaches to network inference exist, including genome-scale metabolic modelling tools such as iNAP 2.0 (Peng *et al.*, 2024) and functional annotation frameworks such as Microbetag (Zafeiropoulos *et al.*, 2025), which can identify potential cross-feeding relationships between community members. While such approaches were beyond the scope of this study, they represent a natural extension of the co-occurrence network framework developed here.

6. Viruses in Freshwater Microbial Networks

Long-read metagenomics has expanded understanding of the ecological roles of viruses in freshwater microbial communities, revealing that phages are not merely agents of cell lysis but active participants in nutrient cycling and metabolic reprogramming. Phages recovered from the Římov Reservoir carried genes associated with oxidative stress mitigation and enhanced host translation (Kavagutti *et al.*, 2019), while phages infecting nitrite-oxidising bacteria in Lake Baikal

carried auxiliary metabolic genes linked to carbon fixation and methanol metabolism, suggesting a capacity to augment host metabolic performance (Coutinho *et al.*, 2020). A global virome survey across 380 freshwater datasets found AMGs associated with carbohydrate and amino acid metabolism, photosynthesis, and quorum sensing interference to be widespread, indicating that viral metabolic reprogramming of host cells may be a broadly significant process in freshwater ecosystems (Elbehery and Deng, 2022).

Viral diversity in freshwater systems is high, with a substantial proportion of sequences lacking known homologs in current databases. In Lough Neagh, only approximately 15% of viral sequences could be classified (Skvortsov *et al.*, 2016), a pattern consistent with the high proportion of unclassified sequences reported in other freshwater virome studies. Phages infecting freshwater acI Actinobacteria have also been recovered from lake metagenomes, with some encoding toxin-associated genes possibly involved in predator defence (Ghai *et al.*, 2017), highlighting the diversity of ecological strategies present within freshwater viral communities.

Host prediction remains a significant challenge in freshwater viral ecology. CRISPR spacer matching is effective for hosts with functional CRISPR systems but fails to capture lytic phages that leave no genomic trace in the host. Machine learning tools such as iPHoP have improved host prediction by integrating multiple lines of evidence, though they continue to struggle with novel viral groups lacking database representation (Roux *et al.*, 2023). These limitations directly constrained the virus-host linkage analysis attempted in this study and identify deeper sequencing and improved MAG recovery as prerequisites for comprehensive virus-host network inference in freshwater systems.

7. Knowledge Gaps and Future Directions

Despite recent advances in long-read metagenomics, several methodological and ecological challenges remain. Nanopore sequencing, while substantially more accurate than earlier generations of the technology, still produces systematic errors in homopolymer regions that can complicate single-nucleotide variant detection and strain-level resolution, and polishing remains necessary even with high-accuracy chemistry (Sereika *et al.*, 2022). Continued improvements in basecalling algorithms will be necessary to fully exploit the resolution available from high-accuracy ONT chemistry for strain-resolved metagenomics.

Linking viruses to their bacterial hosts remains one of the most significant unresolved challenges in environmental viral ecology. CRISPR spacer matching provides reliable host assignments for

integrated phages in hosts with functional CRISPR systems, but is unable to identify hosts of lytic phages that leave no genomic signature. Prophage detection is similarly restricted to temperate viruses, leaving a large proportion of environmental viral diversity unassigned. Integrated approaches such as iPHoP, which combines multiple host prediction signals including sequence homology, CRISPR matches, and machine learning, represent progress toward addressing this gap, though performance remains limited for viral groups with no database representation (Roux *et al.*, 2023).

The functional significance of auxiliary metabolic genes identified in viral genomes remains poorly understood. While AMGs have been identified across a wide range of environmental viromes, experimental validation of their activity during infection and their quantitative contribution to host metabolism and ecosystem nutrient cycling remains scarce (Warwick-Dugdale *et al.*, 2019). Addressing this gap will require integrated approaches combining metagenomic surveys with transcriptomic data and targeted laboratory experiments.

Temporal and spatial sampling coverage in freshwater viral and microbial ecology studies remains limited. The majority of long-read metagenomic studies have been conducted at single or few timepoints and in a restricted range of environments, constraining understanding of how microbial interaction networks respond to dynamic ecological processes such as bloom succession (Wang *et al.*, 2025). For example, most virome characterisation studies in freshwater systems have been based on single sampling events, providing a snapshot of community composition rather than a dynamic view of how viral communities shift in response to changing host abundance and organic matter availability. Broader temporal coverage and standardised analytical frameworks, including adherence to reporting standards such as MIUViG (Roux *et al.*, 2019) and the adoption of portable, reproducible pipelines (Chang *et al.*, 2024), will be essential for improving comparability across studies and advancing the field toward a more comprehensive understanding of freshwater microbial ecology.

8. Summary

Nanopore-only metagenomics has emerged as a viable approach for recovering complete microbial and viral genomes from complex environmental samples without reliance on short-read data. When sequencing depth and polishing strategies are adequate, long-read-only pipelines can reconstruct high-quality MAGs, viral genomes, and mobile genetic elements with a level of contiguity that short-read assemblies rarely achieve. This capability is particularly valuable for

studying microbial interaction networks in freshwater systems, where complete genomic context is required to link viruses to their hosts and resolve metabolic interdependencies between community members.

As tools for assembly, polishing, and host prediction continue to improve, nanopore-only workflows are becoming increasingly practical for genome-resolved environmental metagenomics. Nevertheless, challenges remain in achieving sufficient sequencing depth for complex communities, standardising analytical approaches across studies, and validating the ecological relevance of inferred interaction networks through experimental approaches. The workflow developed in this study directly addresses several of these challenges by implementing a reproducible, publicly available nanopore-only pipeline for freshwater metagenomics, and by applying it to a temporally resolved dataset capturing microbial community dynamics during cyanobacterial bloom succession. In doing so, this work contributes to the methodological advancement of long-read metagenomics and to the ecological understanding of bloom-associated microbial communities in eutrophic freshwater systems.

METHODS

1. Sample collection and DNA extraction

Water samples were collected from Simnas Fish Hatchery Pond at approximately weekly intervals between 4 August and 8 September 2023, yielding ten sampling timepoints. DNA was extracted from filters using a commercial kit following the manufacturer's instructions. Sequence libraries were prepared using the Native Barcoding Kit 24 V14 (SQK-NBD114.24; Oxford Nanopore Technologies) with a minimum total DNA input of 400 ng, following the manufacturer's instructions. Sequencing was performed on a MinION Mk1B instrument using FLO-MIN114 flow cells (R10.4 chemistry; Oxford Nanopore Technologies) for 48 hours per run. Raw nanopore reads were basecalled using Guppy v6.5.1 with the dna_r10.4_e8.1_sup.cfg configuration, corresponding to the super-accuracy basecalling model.

2. Sample fractionation

Bacterial community DNA was collected by filtering pond water through 0.2 µm Sterivex or Supor filters. For viral community analysis, viruses were concentrated from the filtrate by iron chloride (FeCl3) flocculation and collected onto 1.0 µm polycarbonate filters following established protocols for environmental virome recovery.

3. AFA cell enumeration

AFA cell abundance was determined from 0.1 mL water samples preserved with 0.2 µm pre-filtered formalin solution (approximately 1% final concentration) and stored in the dark at +4°C until analysis. Cells were enumerated at each sampling timepoint by light microscopy (Nikon Eclipse Ci H550S) using a Nageotte counting chamber, with no fewer than 600 units examined per sample. AFA count data were provided by Dr. S. Šulčius (personal communication).

4. Read processing and quality control

Raw nanopore reads were trimmed using Porechop with the --discard_middle flag enabled to remove adapter sequences and chimeric reads containing internal adapters. While Porechop has since been deprecated by Oxford Nanopore Technologies in favour of adapter trimming integrated into the Dorado basecalling workflow, it remained appropriate for use with Guppy-basecalled data. No additional read-level filtering based on length or quality score was applied. Read quality was assessed using NanoPlot and FastQC, with reports aggregated using MultiQC.

5. Metagenomic assembly

Adapter-trimmed reads were assembled de novo using metaFlye in metagenome mode (--meta). The estimated metagenome size was set to 60 Mbp, based on the expected cumulative genome size of a moderately diverse freshwater bacterioplankton community. The assembly coverage parameter (--asm-coverage) was set to 40, following the recommended default for metagenomic assemblies. Assembly contiguity and completeness statistics were evaluated from the metaFlye assembly summary output.

6. Assembly polishing

Assembled contigs were polished in two sequential stages at the assembly level. First, trimmed reads were mapped back to the assembly using Minimap2 (Li, 2018) with the map-ont preset, and consensus errors were corrected using Racon. This mapping and polishing cycle was repeated for three rounds. Second, Medaka was applied for a single round of final consensus refinement using a neural network model automatically selected based on the sequencing chemistry.

7. Contig filtering prior to binning

Following assembly polishing, contigs were filtered prior to binning using a custom Python script. Per-contig coverage estimates were obtained by mapping trimmed reads back to the polished assembly using Minimap2 and calculating depth profiles using jgi_summarize_bam_contig_depths. Contigs were retained if they met the following criteria: minimum length of 1,500 bp, minimum mean coverage of 1×, and GC content between 0.20 and 0.80.

8. Genome binning and refinement

Metagenome-assembled genomes (MAGs) were reconstructed from filtered contigs using two complementary binning algorithms: MetaBAT2 and MaxBin2. MetaBAT2 clusters contigs based on sequence composition and differential coverage patterns, whereas MaxBin2 employs probabilistic models incorporating single-copy marker genes to assign contigs to bins. Binning was performed independently for each sample using filtered contigs and corresponding coverage profiles derived from read alignments summarised using jgi_summarize_bam_contig_depths.

The resulting bin sets from MetaBAT2 and MaxBin2 were consolidated using the MetaWRAP bin_refinement module (Uritskiy, DiRuggiero and Taylor, 2018), which selects the

best-supported genome reconstructions based on completeness and contamination metrics estimated by CheckM.

9. Bin-level polishing

Following genome binning and MetaWRAP refinement, each recovered bin was subjected to an additional round of polishing to improve per-genome consensus accuracy. Trimmed reads were mapped to each bin using Minimap2 with the map-ont preset, followed by one round of Racon consensus correction and one round of Medaka refinement. This per-bin polishing step was applied to all bins prior to quality assessment with CheckM2.

10. Genome quality assessment

The quality of reconstructed MAGs was assessed using CheckM2. Completeness and contamination were estimated for each polished MAG, and genomes were classified according to established quality standards (The Genome Standards Consortium *et al.*, 2017). High-quality MAGs were defined as those with ≥90% completeness and ≤5% contamination, medium-quality MAGs as those with ≥50% completeness and ≤10% contamination, and low-quality MAGs as those falling below these thresholds. Only high-quality and medium-quality MAGs were retained for subsequent taxonomic classification.

11. Taxonomic classification

Taxonomic classification of MAGs was performed using GTDB-Tk v2 with the Genome Taxonomy Database (GTDB) release 226. Only high-quality and medium-quality MAGs were included in this step. Classification was carried out using the classify_wf workflow, which assigns taxonomy based on placement of genomes within a standardised reference tree constructed from conserved marker genes. The --skip_ani_screen option was used to reduce computational overhead.

12. Bacterial community profiling

Taxonomic classification of quality-filtered reads from each sample was performed using Kraken2 against the full standard database (k2_standard_20260226), which includes bacterial, archaeal, viral, and human reference sequences (Wood, Lu and Langmead, 2019). Per-sample classification reports were parsed to extract family-level read counts, which were converted to relative abundances and compiled into a single abundance matrix across nine sampling timepoints.

13. Abundance table curation and filtering

Prior to network analysis, the family-level relative abundance table was manually curated to remove 26 bacterial families with no ecological relevance to the freshwater pond environment, including known human-associated and clinical taxa such as Enterobacteriaceae, Moraxellaceae, Staphylococcaceae, Lactobacillaceae, and Bifidobacteriaceae. The remaining families were re-normalised to 100%. A prevalence and abundance filter was subsequently applied, retaining only families present in at least 8 of 9 samples and with a mean relative abundance of ≥0.03% across the dataset, yielding a final matrix of 158 bacterial families across nine timepoints for network analysis.

14. Co-occurrence network construction

Co-occurrence network inference was performed using the SPIEC-EASI framework (Kurtz *et al.*, 2015) implemented in the SpiecEasi R package in R. Count data were derived by scaling the filtered relative abundance matrix by a factor of 1,000. The Meinshausen-Bühlmann neighbourhood selection method was applied with a minimum lambda ratio of 0.1, 30 lambda values, and 50 StARS stability repetitions (set.seed(42)). The resulting sparse inverse covariance matrix was used to construct an undirected network graph using the igraph R package (Csárdi *et al.*, 2025), with isolated nodes removed from the final network. Node-level metrics including mean relative abundance, degree, and normalised betweenness centrality were calculated for all retained nodes.

15. Guild identification

Community detection within the co-occurrence network was performed using the Leiden algorithm (Traag, Waltman and Van Eck, 2019) implemented in the cluster_leiden() function of the igraph R package, with modularity as the objective function and a resolution parameter of 1.0 (set.seed(42)). This yielded 12 community modules. Guild labels were assigned manually to nine of the twelve modules based on the dominant bacterial families within each module and their known ecological roles in freshwater systems as described in the literature. The three remaining modules consisted of taxonomically heterogeneous or low-abundance families and were not assigned specific guild labels. Network visualisation was performed using Cytoscape (Shannon *et al.*, 2003).

16. Viral sequence identification and quality assessment

Viral sequences were identified from per-sample metaFlye assemblies using VIBRANT (Kieft, Zhou and Anantharaman, 2020), which integrates protein similarity, HMM profiles, and genome quality metrics to distinguish viral from bacterial sequences. Quality assessment of recovered viral sequences was subsequently performed using CheckV (Nayfach *et al.*, 2021). Sequences were classified as high-quality (≥90% completeness), medium-quality (50–90% completeness), low-quality (<50% completeness), or undetermined based on CheckV output criteria.

17. Viral taxonomy and functional annotation

Taxonomic classification of viral sequences was performed using Kaiju (Menzel, Ng and Krogh, 2016) against the kaiju_db_viruses_2024-08-15 viral protein reference database. Functional annotation of predicted viral genes was performed using geNomad (Camargo *et al.*, 2024), which identifies structural, replication, and lysis-associated genes as well as putative auxiliary metabolic genes.

18. Computational environment

All bioinformatic analyses were executed on a high-performance computing cluster running a Linux-based operating system. Workflow execution was managed using Snakemake (Mölder *et al.*, 2021), with each pipeline step running within an isolated Conda environment. Network analysis and visualisation were performed locally in R. A summary of all software tools and versions used in this study is provided in Table 1.

Table 1 Versions of software tools used

Software tool	Version
metaFlye	2.9.6
Porechop	0.2.4
NanoPlot	1.46.2
Minimap2	2.30
Racon	1.5.0
Medaka	2.1.1
MetaBAT2	2.18
MaxBin2	2.2.7

MetaWRAP	1.3.2
CheckM2	1.1.0
GTDB-Tk	2.4.0
Snakemake	9.13.7
FastQC	0.12.1
MultiQC	1.25.1
Guppy	6.5.1
Kraken2	2.0.8-beta
VIBRANT	1.2.1
CheckV	1.0.3
Kaiju	1.10.1
geNomad	1.8.0
SpiecEasi	1.0.7
igraph	2.2.2
Cytoscape	3.10.4
R	4.4.3

RESULTS

1. Development of a reproducible nanopore-only metagenomics pipeline

A custom Snakemake-based metagenomics analysis pipeline was developed and applied for genome-resolved analysis of freshwater microbial communities of ten Simnas Fish Hatchery pond samples. The pipeline was run on an HPC using isolated Conda environments, producing MAGs with associated quality metrics and taxonomic assignments. The pipeline was made publicly accessible at https://github.com/Juodziukas/masters-thesis-pipeline.

2. Sequencing output and assembly statistics

Sequencing yielded between 7,234 and 116,789 reads per sample, with a total of 757,511 reads and 3.10 Gbp of sequence data across all samples (Table 2). Median read quality scores were consistent across timepoints, ranging from 15.6 to 16.0. The 08.09.2023 sample was a notable outlier, yielding only 7,234 reads and 0.05 Gbp compared to all other samples, resulting in a markedly reduced assembly of 83 contigs totaling 1.16 Mbp.

Table 2 Metagenomic assembly statistics per sample

Sample date	Reads (n)	Total bases (Gbp)	Mean read length (bp)	Median read quality	Contigs (n)	Total Assembly (Mbp)
04.08.2023	94,053	0.35	3,697	15.8	759	19.47
08.08.2023	77,684	0.30	3,864	15.9	1,119	22.53
14.08.2023	58,240	0.24	4,132	15.8	698	16.38
18.08.2023	91,210	0.33	3,594	15.7	946	22.01
22.08.2023	83,041	0.31	3,736	15.8	836	17.18
25.08.2023	116,789	0.50	4,283	15.8	1,375	34.53
29.08.2023	73,957	0.33	4,461	15.9	602	19.99
01.09.2023	88,463	0.40	4,574	15.9	705	28.80
05.09.2023	66,840	0.29	4,391	15.9	656	15.30
08.09.2023	7,234	0.05	6,572	16.0	83	1.16
Total	757,511	3.10	-	-	7,779	197.35

3. Genome binning and MAG recovery

Genome binning across ten samples yielded 45 bins in total, of which 13 met medium-quality criteria (≥50% completeness, ≤10% contamination) and 32 were classified as low quality (Table 3). No high-quality (≥90% completeness, ≤5% contamination) MAGs were recovered. Recovery varied considerably across timepoints, with the 25.08.2023 sample yielding the highest number of bins (n=8) and the 08.09.2023 sample yielding only one low-quality bin, consistent with its low sequencing output.

Table 3 Genome binning summary per sample

Sample date	Total bins	HQ	MQ	LQ
04.08.2023	2	0	2	0
08.08.2023	4	0	2	2
14.08.2023	2	0	1	1
18.08.2023	5	0	2	3
22.08.2023	6	0	0	6
25.08.2023	8	0	3	5
29.08.2023	7	0	1	6
01.09.2023	6	0	2	4
05.09.2023	4	0	0	4
08.09.2023	1	0	0	1
Total	45	0	13	32

The 13 recovered medium-quality MAGs represented four bacterial phyla: Bacteroidota, Pseudomonadota, Verrucomicrobiota, and Actinomycetota (Table 4). Bacteroidota was the most frequently recovered phylum, with MAGs affiliated with the genera *Flavobacterium*, UBA952, M0103, *Pedobacter*, *Sediminibacterium*, and CHPMRC01. Pseudomonadota MAGs were affiliated with the genera *Polynucleobacter* and Sphingorhabdus B, while Verrucomicrobiota was represented by *Luteolibacter* and Actinomycetota by *Planktophila*.

Table 4 MAGs recovered from Simnas Fish Hatchery pond samples

Sample date	Completeness (%)	Contamination (%)	Phylum	Genus
04.08.2023	57.1	0.8	Pseudomonadota	*Sphingorhabdus_B*
04.08.2023	71.9	1.2	Verrucomicrobiota	*Luteolibacter*
08.08.2023	54.5	2	Verrucomicrobiota	*Luteolibacter*
08.08.2023	76.1	3.2	Bacteroidota	*Flavobacterium*

14.08.2023	89.5	1.5	Bacteroidota	UBA952
18.08.2023	72.1	4.2	Bacteroidota	*Pedobacter*
18.08.2023	79.7	4.3	Bacteroidota	UBA952
25.08.2023	63.8	2	Bacteroidota	*Sediminibacterium*
25.08.2023	64.2	1.9	Bacteroidota	CHPMRC01
25.08.2023	55.1	5.1	Actinomycetota	*Planktophila*
29.08.2023	59	1.2	Pseudomonadota	*Polynucleobacter*
01.09.2023	60.8	5.4	Bacteroidota	M0103
01.09.2023	61.2	4.6	Pseudomonadota	*Polynucleobacter*

4. Bacterial community co-occurrence network and guild structure

SPIEC-EASI co-occurrence network analysis of 158 bacterial families across nine sampling timepoints identified 208 statistically supported associations. Leiden clustering of the resulting network resolved 12 communities, of which nine were ecologically interpretable and assigned guild labels based on dominant taxonomic composition (Fig. 1). The remaining three communities comprised minor or taxonomically diverse families and were not assigned to specific guilds. Node size reflects mean relative abundance across samples, with Comamonadaceae representing the most abundant node in the network. The nine ecologically interpretable guilds encompassed distinct functional groups including bloom-associated, decomposer, free-living pelagic, and particle-attached bacterial communities, reflecting the ecological roles present within the Simnas Fish Hatchery pond microbiome.

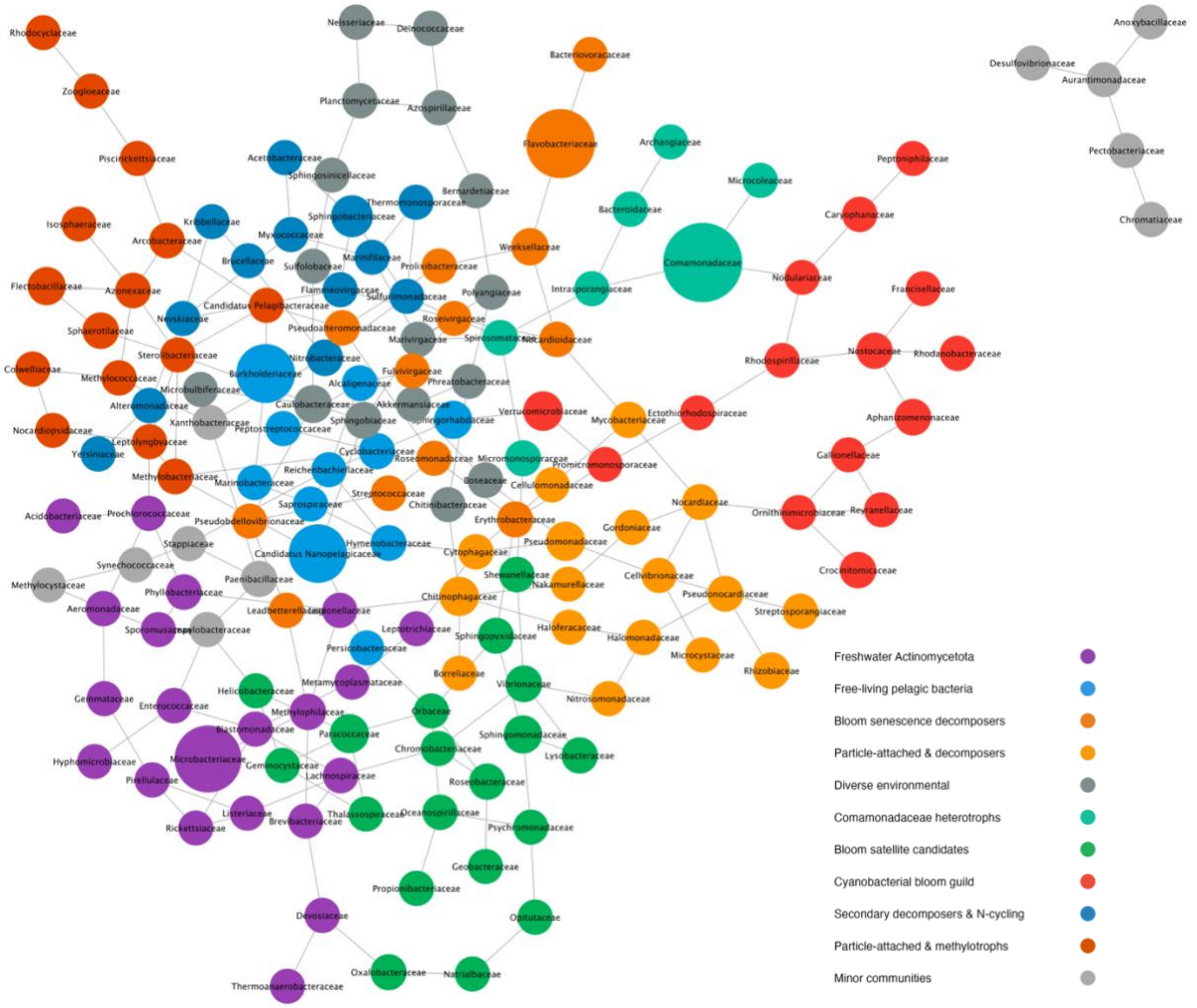

Fig. 1 Bacterial co-occurrence network inferred by SPIEC-EASI across nine Simnas Fish Hatchery pond samples. Node size reflects mean relative abundance. Colours indicate guild membership as determined by Leiden clustering.

5. Temporal dynamics of bacterial community guilds

Temporal dynamics of the nine bacterial guilds varied across the sampling period (Fig. 2). The Free-living pelagic bacteria guild started off as the most abundant one in early August, declining until mid-August and then increasing steadily up to being the dominating guild by early September. The Bloom senescence decomposer guild showed a marked increase in early August, followed by a drop in mid-August, afterwards declining at a steady rate until the beginning of September with a dip on September first. The Freshwater Actinomycetota guild showed a steady rise from beginning of August until end of August, peaking at the 25th , and then dropping on the 29th, before continuing to steadily decline until the 5th of September. The Comamonadaceae guild showed two notable peaks on the 14th of August and on the first of September with a notable low point on the 25th of August. The Cyanobacterial bloom guild shows two noticeable peaks at the

beginning of August and at the end of August with dips in the middle of August and on the start of September. The Secondary decomposers and N-cycling bacteria guild rose from the beginning of August before peaking on the 18th and then steadily declining until beginning of September. The Particle-attached and decomposers guild relative abundance fluctuated throughout the time, with dips in the middle of August and at the start of September and highs at the beginning of August and the end of August. The Bloom satellite candidates guild dropped dramatically at the beginning of August and then remained steady from middle of August until early September. The Particle-attached and methylotroph guild remained stable throughout the sampling period with a minor peak towards the end of August.

AFA cell counts measured across the sampling period were overlaid on the guild abundance profiles to contextualise community dynamics within the bloom. AFA abundance peaked at 1670 cells mL^{-1} on 8th August and declined sharply thereafter, falling to near-zero by 1st September. The Free-living pelagic bacteria guild showed a clear inverse relationship with AFA abundance, reaching its lowest point on 14th August before recovering steadily as AFA declined, eventually becoming the dominant guild by September. The Bloom senescence decomposer guild was low at the start of sampling and rose to its peak following the AFA peak, reaching maximum abundance around 8th August and declining afterwards. The Freshwater Actinomycetota guild showed a

similar delayed increase, rising steadily through August and peaking in late August as AFA abundance fell.

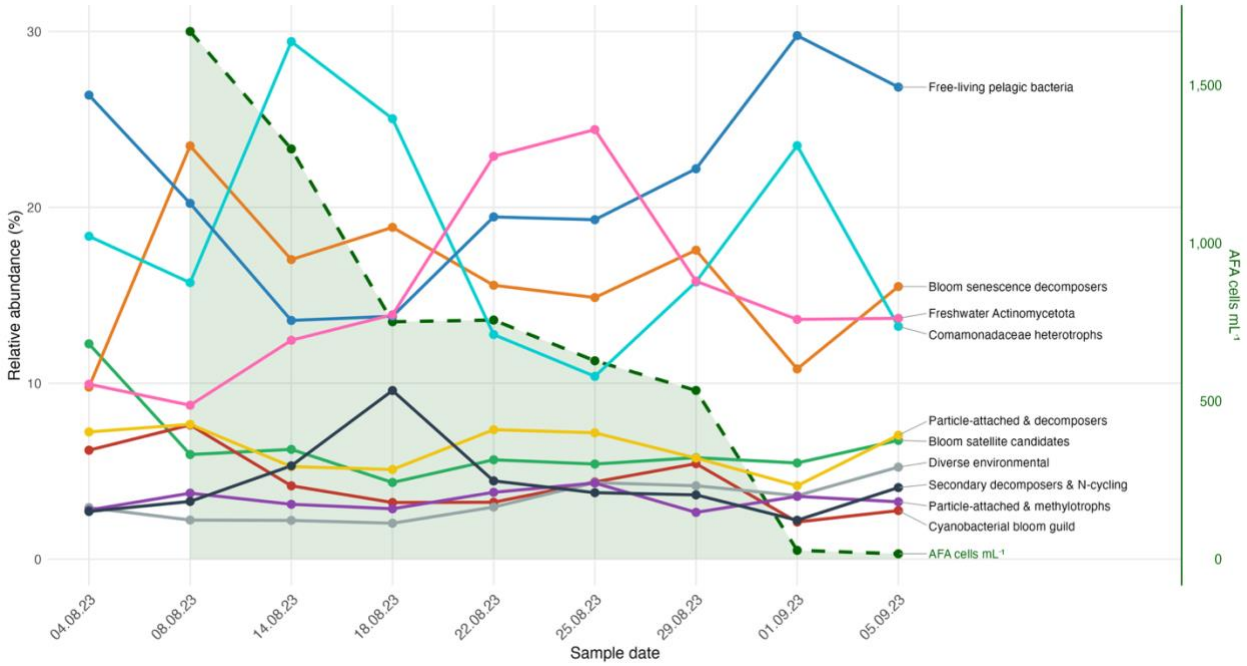

Fig. 2 Temporal dynamics of bacterial community guilds across nine Simnas Fish Hatchery pond samples. The dashed green line indicates Aphanizomenon flos-aquae cell counts (cells mL⁻¹, right axis).

Detailed family-level relative abundance composition of each guild across the sampling period is provided in Appendices.

6. Viral community recovery and quality assessment

Viral sequence identification across all individual sample assemblies using VIBRANT yielded 1,986 sequences, of which 1,974 were classified as lytic and 12 as lysogenic. Quality assessment using CheckV on the merged assembly identified 113 viral sequences, of which four were high-quality, ten medium-quality, 90 low-quality, and nine not-determined. One complete circular viral genome was recovered from the merged assembly (Table 5).

Table 5 Viral sequence recovery and quality assessment summary

Category	Count
Total viral sequences	1,986
Lytic	1,974
Lysogenic	12
Circular	1
High-quality	4
Medium-quality	10

Low-quality	90
Not-determined	9

7. Viral community composition and taxonomy

The classified viral community was dominated by the phylum Uroviricota (n=1,894), with Nucleocytoviricota representing a minor but notable component (n=15). At the family level, Autographiviridae was the most abundant classified family (n=137), followed by Kyanoviridae (n=67) and Zobellviridae (n=49) (Fig. 3). Additional families included Schitoviridae (n=22), Mesyanzhinovviridae (n=15), Peduoviridae (n=14), Stanwilliamsviridae (n=11), and Casjensviridae (n=11). Sequences assigned to Phycodnaviridae (n=6) and Mimiviridae (n=4) represented large DNA viruses associated with algal and protist hosts. A total of 1,163 sequences (58.6%) could not be assigned taxonomy at any rank.

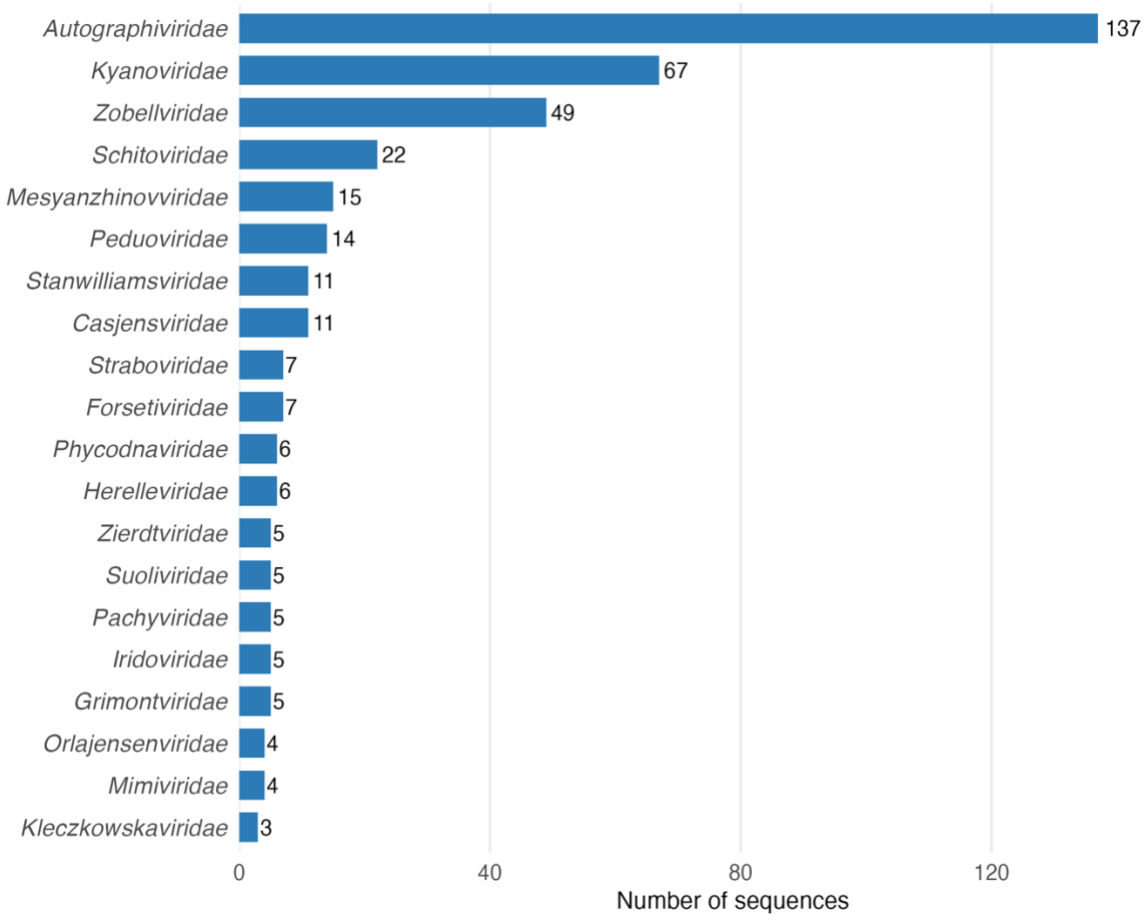

Fig. 3 Taxonomic composition of the viral community at family level.

8. Putative functional genes in viral genomes

Functional annotation of 31,930 predicted viral genes using geNomad assigned functions to 4,598 genes (14.4%). The most abundant annotated functions were lysis proteins (n=431), prophage-encoded structural proteins of the DUF1398 family (n=305), terminase large subunits (n=263 combined), head-to-tail connecting proteins (n=104), and portal proteins (n=95) (Fig. 4). Several genes with putative metabolic functions were also identified, including ribonucleoside-triphosphate reductase, flavin-dependent thymidylate synthase, glutaredoxin, and a CRISPR-associated Cas4 protein, representing putative auxiliary metabolic genes. Dedicated AMG analysis using DRAM-v was not performed within the scope of this study and represents a priority for future work.

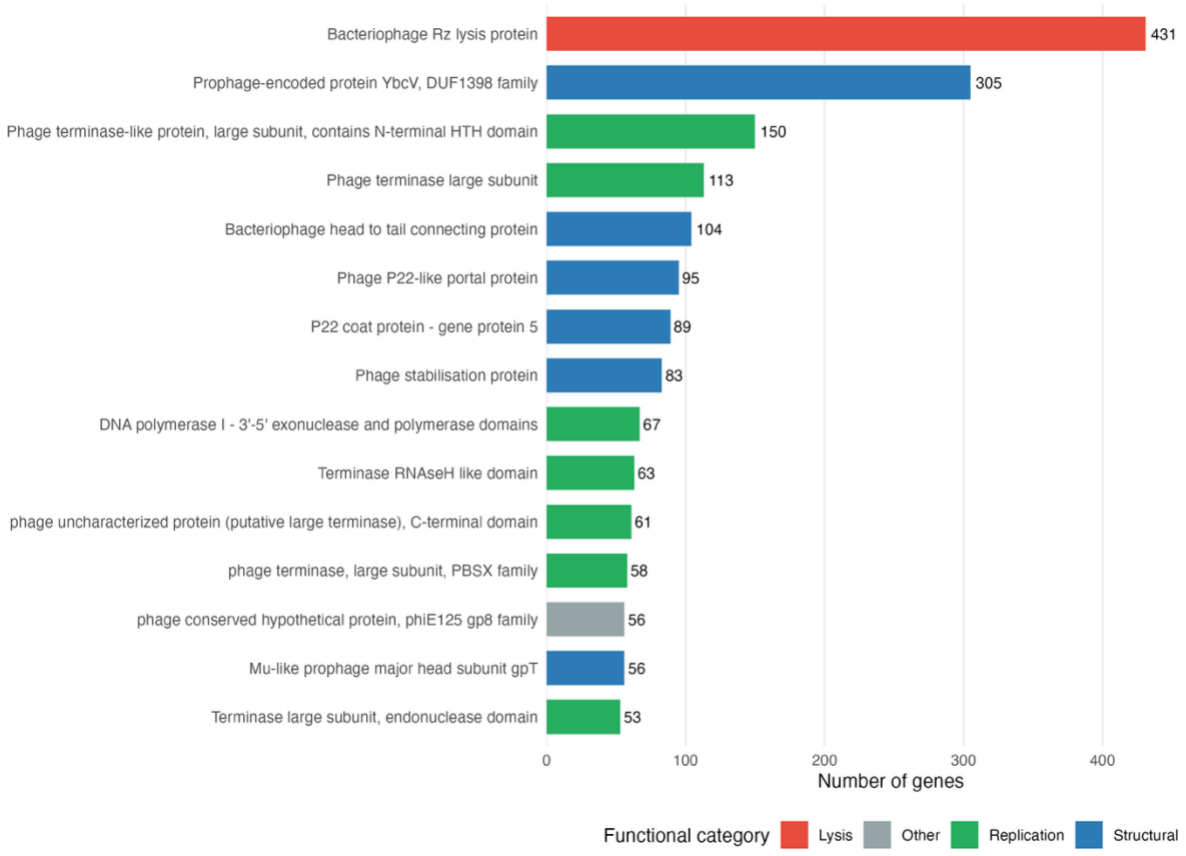

Fig. 4 Top 15 most abundant functionally annotated genes identified across viral genomes.

DISCUSSION

This study aimed to characterise the dynamics of microbial communities driven by cyanobacterial bloom activity in a eutrophic freshwater pond ecosystem, using a nanopore-only long-read metagenomics approach combining de novo genome assembly, MAG recovery, viral genome identification, co-occurrence network inference, and temporal community analysis. The results demonstrate that meaningful ecological signal can be extracted from nanopore-only data at modest sequencing depth, while simultaneously exposing the inherent challenges of applying long-read metagenomics to low-biomass, taxonomically diverse, and unevenly covered communities. The following sections discuss the pipeline's performance and reproducibility, the quality and taxonomic context of recovered MAGs, the composition and putative functional gene content of the viral community, and the ecological implications of the inferred bacterial co-occurrence networks and guild dynamics in the context of a cyanobacterial bloom system.

1. Pipeline development and reproducibility

A central deliverable of this thesis was the development of a reproducible, end-to-end Snakemake-based nanopore-only metagenomics pipeline, publicly available on GitHub. The pipeline was successfully deployed across ten samples on an HPC cluster, producing consistent outputs including assembled contigs, polished MAGs, quality metrics, and taxonomic assignments. The use of isolated Conda environments within the Snakemake workflow ensured computational reproducibility across different computing environments, an important consideration given the lack of standardised, portable pipelines in metagenomics (Chang *et al.*, 2024). Since its development, newer tools such as nanoMDBG have emerged as more performant alternatives to metaFlye for high-accuracy ONT data (Benoit *et al.*, 2026), and future versions of the pipeline would benefit from incorporating these advances alongside the transition from Guppy to the Dorado basecaller, which has superseded Guppy as Oxford Nanopore's recommended basecalling platform.

A notable methodological decision was to run genome binning on individual per-sample assemblies rather than a co-assembly. While co-assembly can increase contig depth and improve binning by pooling coverage signals across timepoints, it also risks conflating strain variants from different sampling dates and complicating the attribution of genomic features to specific timepoints (Olm *et al.*, 2017). The per-sample approach preserves temporal resolution and

minimises chimeric bin formation, at the cost of reduced sequencing depth available for binning within each sample. This temporal resolution was a deliberate priority of this study, as many freshwater metagenomics studies have been conducted at single or limited timepoints, constraining insight into the dynamic nature of microbial communities during ecologically significant events such as cyanobacterial blooms (Wang *et al.*, 2025). Given the modest read yields per sample (median ~80,000 reads, ~0.31 Gbp), this trade-off likely contributed to the limited number and quality of recovered MAGs, as discussed below.

2. MAG recovery and quality

Genome binning across ten samples yielded 45 bins in total, of which only 13 reached medium-quality (MQ) thresholds (≥50% completeness, ≤10% contamination) and none reached high-quality (HQ) criteria (≥90% completeness, ≤5% contamination). While the absence of HQ MAGs is a significant limitation, it is not unexpected given the modest sequencing depth per sample, as studies benchmarking nanopore-only metagenomics have generally found that depths of at least 5–10 Gbp per sample are needed to consistently recover HQ MAGs from complex communities (Liu *et al.*, 2022).

The primary driver of low MAG completeness in this study was insufficient sequencing depth. With a median of approximately 0.31 Gbp per sample, genomic coverage for most community members would have been far below what is required for reliable binning, particularly for lower-abundance taxa. The 08.09.2023 sample, which yielded only 0.05 Gbp and produced a single low-quality bin, represents an extreme outlier likely due to a technical issue during sequencing or library preparation rather than a true biological signal. Although community composition analyses were initially performed including this sample, its low sequencing yield introduced distortion in relative abundance calculations, and it was therefore excluded from all downstream ecological analyses. Assembly and binning statistics for this sample are retained in the results for completeness.

Beyond sequencing depth, the taxonomic diversity of the community itself posed an additional challenge. Freshwater pond ecosystems, particularly during bloom events, are known to harbour highly diverse bacterial assemblages with many low-abundance members, meaning that even moderately represented genomes may not accumulate sufficient coverage for complete assembly and reliable binning. The MetaWRAP refinement step, which consolidated bins from both MetaBAT2 and MaxBin2 and selected the best-supported reconstructions, partially mitigated bin

fragmentation, but could not compensate for the fundamental coverage limitations imposed by the sequencing depth available. Furthermore, sequencing was performed on the bacterial size-fraction of the pond water, which while advantageous for reducing eukaryotic contamination and focusing sequencing effort on the bacterial community, means that the full diversity of that fraction was competing for coverage within a limited sequencing depth budget, further reducing per-genome coverage for less abundant members.

Despite these constraints, the 13 MQ MAGs recovered represent ecologically meaningful taxa. Notably, the UBA952 MAG recovered from the 14.08.2023 sample achieved 89.5% completeness with only 1.5% contamination, falling just below the high-quality threshold, and should be considered near-HQ for practical purposes. Bacteroidota dominated the recovered set with six genera identified: *Flavobacterium*, *Pedobacter*, *Sediminibacterium*, and three uncultivated genera: UBA952, M0103, and CHPMRC01. Both UBA952 and M0103 belong to the family Crocinitomicaceae within Flavobacteriales, and their recovery at MQ across multiple timepoints suggests this uncultivated lineage may represent an ecologically recurring component of the bloom-associated Bacteroidota community. CHPMRC01 is particularly novel, belonging to an undescribed family (JADJOU01) within Chitinophagales, and was recovered at MQ from the 25.08.2023 sample. The recovery of *Flavobacterium* and *Sediminibacterium* MAGs is ecologically noteworthy, as members of these genera are frequently detected during and after bloom events in eutrophic freshwater systems, where they are thought to contribute to bloom senescence through hydrolysis of extracellular polymeric substances (Čačković *et al.*, 2025). The recovery of *Polynucleobacter* from two separate timepoints is consistent with its status as one of the most abundant and cosmopolitan freshwater bacterial genera globally (Hahn *et al.*, 2012), and its appearance in late August/early September samples aligns with its known preference for warmer, nutrient-rich conditions. Similarly, *Planktophila*, a member of the freshwater acI Actinomycetota lineage, is a characteristic oligotrophic freshwater taxon whose highly streamlined genome is thought to create metabolic dependencies on co-occurring organisms, including potentially *Polynucleobacter*(Neuenschwander *et al.*, 2018).

The recovery of *Luteolibacter* (Verrucomicrobiota) from two early August samples, but not later timepoints, is an interesting temporal signal. Verrucomicrobiota are well-established polysaccharide degraders in freshwater systems, with genomic analyses revealing high abundances of glycoside hydrolase genes and experimental evidence for active polysaccharide uptake (He *et*

al., 2017). *Luteolibacter* has been shown to demonstrate active polysaccharide processing in freshwater lakes, with its abundance driven by dissolved organic carbon availability (Čačković *et al.*, 2025). Its early-season presence in the Simnas pond may therefore reflect the availability of specific polysaccharide substrates at the onset of sampling, perhaps associated with early bloom initiation or phytoplankton-derived exudates, with the subsequent absence from later timepoints suggesting substrate depletion or competitive exclusion as the community shifted toward the bloom-senescence decomposer guild.

Improving MAG recovery in future work would require deeper sequencing per sample (ideally ≥5 Gbp) which would increase per-genome coverage and provide the differential coverage signal that binning algorithms rely on heavily (Liu *et al.*, 2022). Although this study already employed R10.4 chemistry with super-accuracy Guppy basecalling, a further improvement in consensus accuracy could be achieved by switching to the R10.4.1 flow cell in combination with the Dorado basecaller, which has superseded Guppy as Oxford Nanopore's recommended basecalling platform and has been shown to yield near-finished bacterial genomes from complex metagenomes without requiring short-read polishing (Sereika *et al.*, 2022). The integration of a co-assembly step across temporally adjacent samples could also increase MAG recovery rates by pooling sequencing depth and strengthening the coverage signals available to binning algorithms, without sacrificing temporal resolution. Per-sample read mapping back to the co-assembly would still allow community composition and guild dynamics to be tracked across timepoints.

3. Microbial interaction networks and temporal dynamics in the context of bloom ecology

To explore how bacterial communities were structured and changed across the bloom period, co-occurrence network analysis was applied to bacterial family abundance profiles across nine sampling timepoints. Leiden community detection identified nine distinct guilds, which are groups of bacterial families that consistently co-occurred across the sampling period, suggesting shared ecological roles or environmental preferences. The composition and temporal dynamics of these guilds are discussed below in the context of cyanobacterial bloom dynamics in eutrophic freshwater systems.

The Cyanobacterial bloom guild showed two distinct peaks, one in early August and a second at the end of August, with a dip in between. This pattern suggests the bloom underwent a partial decline and recovery, possibly driven by changes in nutrient availability, weather conditions, or viral lysis of cyanobacterial cells. This double-peak structure is important because each peak

would have released a pulse of organic matter into the water, driving the responses seen in the other guilds.

The most striking response came from the Bloom senescence decomposer guild, dominated by Flavobacteriaceae and Sphingobacteriaceae. This guild rose sharply in early August, peaked around 18th August, then gradually declined through September, lagging slightly behind the cyanobacterial peaks. This pattern, sometimes called "blooming after the bloom", has been observed in multiple studies of eutrophic lakes and reflects the role of these bacteria as the primary degraders of organic material that is released when cyanobacterial cells die and break apart (Eiler and Bertilsson, 2007). The gradual decline of this guild through September likely reflects the progressive exhaustion of bloom-derived organic matter as the season ended.

The Free-living pelagic bacteria guild declined through the height of the bloom in August before recovering strongly through September, eventually becoming the dominant guild by the end of the sampling period. This shift could reflect the gradual transition from bloom conditions toward a clearer, lower-nutrient water column as the bloom declined, which are conditions that favour small, streamlined free-living bacteria over bloom-associated specialists. The dominance of *Candidatus* Nanopelagicaceae within this guild is consistent with Nanopelagicales being among the most abundant bacterioplankton lineages in freshwater lakes globally, with their streamlined genomes and small cell size reflecting adaptation to a free-living lifestyle (Mondav *et al.*, 2020).

The Comamonadaceae guild showed two abundance peaks in mid-August and early September, with a notable dip in between. Members of this family are commonly abundant in freshwater lakes and can grow rapidly under elevated nutrient conditions (Willems, 2014). Their episodic peaks may reflect the transient availability of organic substrates released during bloom senescence events, though without metabolite profiling data it is difficult to determine the precise drivers of these fluctuations. Their high connectivity within the co-occurrence network means they may interact broadly with other community members, but whether this reflects real ecological dependencies or shared environmental responses requires further experimental investigation.

The Freshwater Actinomycetota guild increased steadily through August, peaking at the end of the month before declining sharply into September. This timing coincides with the second cyanobacterial bloom peak, which fits with the known reliance of acI Actinobacteria on organic carbon released by cyanobacteria and phytoplankton (Neuenschwander *et al.*, 2018). Interestingly, the guild did not respond to the first bloom peak in early August, which makes the relationship

less straightforward. This could reflect differences in the type of organic matter released during each bloom peak, or other environmental conditions not captured in this study. Dissolved organic carbon profiling, pigment analysis to characterise bloom composition at each timepoint, or metatranscriptomic approaches to assess active gene expression would all help clarify the drivers of this pattern in future work.

The Secondary decomposers and N-cycling bacteria guild peaked on 18th August before gradually declining. Its early peak relative to the Bloom senescence decomposer guild suggests that different components of the bloom-derived organic matter were broken down at different times — with nitrogen-rich compounds such as proteins and pigments being processed first, followed by more complex polysaccharides later in the season. This kind of temporal substrate partitioning has been observed in other bloom systems and likely reflects the different enzymatic capabilities of the bacteria involved (Eiler and Bertilsson, 2007).

Taken together, the guild dynamics suggest a community that responds in a structured way to the availability of organic matter released during cyanobacterial bloom events. It should be noted that the guild names used throughout this section were assigned based on the dominant taxa identified within each network module and should be treated as interpretive labels rather than strict functional definitions. The guilds themselves were defined by Leiden community detection, meaning the bacterial families within each guild co-occurred more strongly with each other than with families in other guilds, but given the large number of families included in the analysis, each guild likely still contains members with diverse and not fully characterised ecological roles. Additionally, while SPIEC-EASI co-occurrence network analysis is well suited to compositional microbiome data as it reduces spurious correlations, co-occurrence patterns reflect statistical associations rather than confirmed ecological interactions (Kurtz *et al.*, 2015). The nine-timepoint dataset, while providing useful temporal resolution, is relatively small for robust network inference, and the guild structure and interaction patterns identified here should therefore be treated as hypotheses to be tested in future work rather than definitive conclusions.

To place the observed guild dynamics within the context of the phytoplankton bloom, AFA cell counts were overlaid on the guild abundance profiles. AFA abundance peaked at 1670 cells mL^{-1} on 8th August and declined sharply thereafter, providing a reference timeline against which bacterial community responses could be assessed. Verrucomicrobiaceae, a family well known for polysaccharide degradation in freshwater systems, was most abundant at the bloom peak in early

August and declined alongside AFA through the sampling period, suggesting that this family was closely associated with living bloom biomass rather than with senescence-derived organic matter (Čačković *et al.*, 2025). Sphingobacteriaceae, by contrast, peaked in mid-August following the AFA maximum and then gradually declined, consistent with its established role as a primary responder to bloom-derived organic matter released during senescence (Eiler and Bertilsson, 2007). The Free-living pelagic bacteria guild, dominated by Candidatus Nanopelagicaceae, showed the clearest inverse relationship with AFA abundance, recovering strongly as the bloom collapsed and eventually becoming the dominant guild by September, reflecting the transition toward post-bloom conditions that favour streamlined, oligotrophic lifestyles (Mondav *et al.*, 2020). Taken together, these patterns suggest that AFA bloom dynamics were a primary structuring force for the bacterial community across the sampling period, with different families responding to distinct phases of the bloom lifecycle.

4. Viral community composition and host-interaction potential

The viral community identified in this study adds an important but only partially explored dimension to the bloom dynamics picture. Prior to viral analysis, a viral size fraction was isolated from the bacterial fraction through an additional filtration step, ensuring that the sequences recovered reflect genuine viral community members rather than bacterial contamination. The overwhelming dominance of lytic over lysogenic viral sequences suggests that active viral lysis is likely occurring in this system, which would contribute directly to the release of organic matter from bacterial and cyanobacterial cells and may therefore partly explain the guild succession patterns described above. The prevalence of Autographiviridae and Kyanoviridae points toward active phage predation targeting both heterotrophic bacteria and cyanobacteria, while the detection of Phycodnaviridae and Mimiviridae, though in low numbers, raises the possibility that eukaryotic phytoplankton may also be subject to viral lysis alongside the cyanobacterial bloom. Together these findings suggest that viruses may be contributing to bloom dynamics through multiple parallel infection pathways, though confirming this would require temporal resolution of viral community composition matched to the bacterial guild trajectories.

The putative auxiliary metabolic genes identified in several viral genomes suggest that some phages may be manipulating host cell metabolism during infection, which could have broader implications for nutrient cycling during bloom events (Warwick-Dugdale *et al.*, 2019). However,

a dedicated functional analysis was beyond the scope of this study and firm conclusions cannot be drawn from gene presence alone.

A key limitation of the viral analysis was the inability to link viral sequences to their bacterial hosts, which would have allowed a virus-host interaction network to be constructed alongside the bacterial co-occurrence network. This was constrained by two factors: the low number of MQ and HQ MAGs left few reference genomes available for CRISPR-based host prediction, and the relatively short mean read length limited the recovery of complete viral genomes needed for reliable homology-based approaches. Addressing these limitations through deeper sequencing and improved MAG recovery would open the door to host prediction tools such as iPHoP, which combined with the temporal guild framework developed here could provide a much more complete picture of virus-host dynamics during bloom events in future work (Roux *et al.*, 2023).

5. Strengths, limitations and future perspectives

This study presents several methodological and ecological contributions. The temporal sampling design, covering ten timepoints across an active bloom period, provides a dynamic view of community succession that is rarely achieved in freshwater metagenomics studies, and the publicly available Snakemake pipeline developed here offers a reproducible framework that other researchers can apply to their own nanopore datasets. However, the study is not without significant limitations. The most critical is insufficient sequencing depth, at a median of 0.31 Gbp per sample the dataset was substantially underpowered for MAG recovery, and the complete absence of high-quality MAGs restricts the functional and genomic conclusions that can be drawn. The absence of accompanying environmental metadata such as nutrient concentrations, temperature, and chlorophyll measurements means that the observed community dynamics cannot be directly linked to physicochemical drivers, leaving the guild temporal patterns correlative rather than mechanistic. The manual curation of the abundance table prior to network construction, while ecologically justified, introduces subjectivity, and the nine-timepoint dataset represents a relatively small sample size for robust co-occurrence inference. Finally, the inability to link viral sequences to their bacterial hosts meant that the viral and bacterial community analyses could not be integrated as originally intended, representing the most significant unmet objective of the study.

CONCLUSIONS

This thesis makes two concrete contributions to freshwater microbial ecology. First, it delivers a publicly available, reproducible nanopore-only metagenomics pipeline that can be directly applied to similar environmental datasets, addressing the lack of standardised long-read workflows in the field. Second, it provides one of the first applications of nanopore-only long-read metagenomics to temporally resolved bacterial guild dynamics during cyanobacterial bloom succession in a eutrophic freshwater pond, revealing structured and ecologically coherent community responses that would not have been detectable from single-timepoint sampling.

The co-occurrence network analysis demonstrated that guild-level temporal tracking is a practical and interpretable framework for studying bloom-associated microbial succession, capturing patterns consistent with known ecological roles of the dominant taxa. Overlaying AFA cell counts on the guild dynamics revealed that different bacterial families responded to distinct phases of the bloom, some tracking peak bloom biomass, others responding to organic matter released during senescence. Viral community characterisation revealed a predominantly lytic community with potential metabolic interactions with host cells, though the inability to link viral sequences to their bacterial hosts at the current sequencing depth meant that the bacterial and viral analyses could not be fully integrated, which is the most significant unmet objective of the study.

The results suggest that nanopore-only metagenomics at modest sequencing depth can yield ecologically meaningful signal from complex freshwater communities, but that substantially deeper sequencing, improved MAG recovery, and the inclusion of physicochemical environmental measurements are needed before genome-resolved virus-host interaction networks can be inferred from this type of system.

ACKNOWLEDGEMENTS

The author would like to thank supervisor Dr. Sigitas Šulčius, Senior Researcher at the Laboratory of Algology and Microbial Ecology, Nature Research Centre, for guidance and support throughout the preparation of this thesis. The author also thanks Dr. Gediminas Alzbutas for providing the viral analysis pipeline used in this study. Claude (Anthropic) was also used as a large language model assistant for text editing and grammatical corrections.

REFERENCES

Anthropic (2025) "Claude" [Large language model, used for text editing and grammatical corrections]. Available at: https://claude.ai

Benoit, G. *et al.* (2026) "High-quality metagenome assembly from nanopore reads with nanoMDBG," *Nature Communications*, 17(1), p. 3556. Available at: https://doi.org/10.1038/s41467-026-69760-y.

Buttler, J. and Drown, D.M. (2022) "Accuracy and Completeness of Long Read Metagenomic Assemblies," *Microorganisms*, 11(1), p. 96. Available at: https://doi.org/10.3390/microorganisms11010096.

Čačković, A. *et al.* (2025) "Selective heterotrophic bacteria can selfishly process polysaccharides in freshwater lakes," *Cell Reports*, 44(4), p. 115415. Available at: https://doi.org/10.1016/j.celrep.2025.115415.

Camargo, A.P. *et al.* (2024) "Identification of mobile genetic elements with geNomad," *Nature Biotechnology*, 42(8), pp. 1303–1312. Available at: https://doi.org/10.1038/s41587-023-01953-y.

Chang, W.-S. *et al.* (2024) "Improving the reporting of metagenomic virome-scale data," *Communications Biology*, 7(1), p. 1687. Available at: https://doi.org/10.1038/s42003-024-07212-3.

Chaumeil, P.-A. *et al.* (2022) "GTDB-Tk v2: memory friendly classification with the genome taxonomy database," *Bioinformatics*. Edited by K. Borgwardt, 38(23), pp. 5315–5316. Available at: https://doi.org/10.1093/bioinformatics/btac672.

Chklovski, A. *et al.* (2023) "CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning," *Nature Methods*, 20(8), pp. 1203–1212. Available at: https://doi.org/10.1038/s41592-023-01940-w.

Coutinho, F.H. *et al.* (2020) "New viral biogeochemical roles revealed through metagenomic analysis of Lake Baikal," *Microbiome*, 8(1), p. 163. Available at: https://doi.org/10.1186/s40168-020-00936-4.

Csárdi, G. *et al.* (2025) "igraph." Zenodo. Available at: https://doi.org/10.5281/ZENODO.3630268.

Eiler, A. and Bertilsson, S. (2007) "Flavobacteria blooms in four eutrophic lakes: linking population dynamics of freshwater bacterioplankton to resource availability," *Applied and Environmental Microbiology*, 73(11), pp. 3511–3518. Available at: https://doi.org/10.1128/AEM.02534-06.

Elbehery, A.H.A. and Deng, L. (2022) "Insights into the global freshwater virome," *Frontiers in Microbiology*, 13, p. 953500. Available at: https://doi.org/10.3389/fmicb.2022.953500.

Ghai, R. *et al.* (2017) "Metagenomic recovery of phage genomes of uncultured freshwater actinobacteria," *The ISME Journal*, 11(1), pp. 304–308. Available at: https://doi.org/10.1038/ismej.2016.110.

Hahn, M.W. *et al.* (2012) "The Passive Yet Successful Way of Planktonic Life: Genomic and Experimental Analysis of the Ecology of a Free-Living Polynucleobacter Population," *PLoS ONE*. Edited by J.H. Badger, 7(3), p. e32772. Available at: https://doi.org/10.1371/journal.pone.0032772.

He, S. *et al.* (2017) "Ecophysiology of Freshwater Verrucomicrobia Inferred from Metagenome-Assembled Genomes," *mSphere*, 2(5), pp. e00277-17. Available at: https://doi.org/10.1128/mSphere.00277-17.

Huisman, J. *et al.* (2018) "Cyanobacterial blooms," *Nature Reviews Microbiology*, 16(8), pp. 471–483. Available at: https://doi.org/10.1038/s41579-018-0040-1.

Kavagutti, V.S. *et al.* (2019) "Phage-centric ecological interactions in aquatic ecosystems revealed through ultra-deep metagenomics," *Microbiome*, 7(1), p. 135. Available at: https://doi.org/10.1186/s40168-019-0752-0.

Kieft, K., Zhou, Z. and Anantharaman, K. (2020) "VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences," *Microbiome*, 8(1), p. 90. Available at: https://doi.org/10.1186/s40168-020-00867-0.

Kim, C., Pongpanich, M. and Porntaveetus, T. (2024) "Unraveling metagenomics through long-read sequencing: a comprehensive review," *Journal of Translational Medicine*, 22(1), p. 111. Available at: https://doi.org/10.1186/s12967-024-04917-1.

Kolmogorov, M. *et al.* (2020) "metaFlye: scalable long-read metagenome assembly using repeat graphs," *Nature Methods*, 17(11), pp. 1103–1110. Available at: https://doi.org/10.1038/s41592-020-00971-x.

Kurtz, Z.D. *et al.* (2015) "Sparse and Compositionally Robust Inference of Microbial Ecological Networks," *PLOS Computational Biology*. Edited by C. Von Mering, 11(5), p. e1004226. Available at: https://doi.org/10.1371/journal.pcbi.1004226.

Lee, J.Y. *et al.* (2021) "Comparative evaluation of Nanopore polishing tools for microbial genome assembly and polishing strategies for downstream analysis," *Scientific Reports*, 11(1), p. 20740. Available at: https://doi.org/10.1038/s41598-021-00178-w.

Li, H. (2018) "Minimap2: pairwise alignment for nucleotide sequences," *Bioinformatics*. Edited by I. Birol, 34(18), pp. 3094–3100. Available at: https://doi.org/10.1093/bioinformatics/bty191.

Li, H. *et al.* (2020) "Microbial community successions and their dynamic functions during harmful cyanobacterial blooms in a freshwater lake," *Water Research*, 185, p. 116292. Available at: https://doi.org/10.1016/j.watres.2020.116292.

Liu, L. *et al.* (2022) "Nanopore long-read-only metagenomics enables complete and high-quality genome reconstruction from mock and complex metagenomes," *Microbiome*, 10(1), p. 209. Available at: https://doi.org/10.1186/s40168-022-01415-8.

Menzel, P., Ng, K.L. and Krogh, A. (2016) "Fast and sensitive taxonomic classification for metagenomics with Kaiju," *Nature Communications*, 7(1), p. 11257. Available at: https://doi.org/10.1038/ncomms11257.

Mölder, F. *et al.* (2021) "Sustainable data analysis with Snakemake," *F1000Research*, 10, p. 33. Available at: https://doi.org/10.12688/f1000research.29032.2.

Mondav, R. *et al.* (2020) "Streamlined and Abundant Bacterioplankton Thrive in Functional Cohorts," *mSystems*. Edited by R. Lal, 5(5), pp. e00316-20. Available at: https://doi.org/10.1128/mSystems.00316-20.

Nayfach, S. *et al.* (2021) "CheckV assesses the quality and completeness of metagenome-assembled viral genomes," *Nature Biotechnology*, 39(5), pp. 578–585. Available at: https://doi.org/10.1038/s41587-020-00774-7.

Neuenschwander, S.M. *et al.* (2018) "Microdiversification in genome-streamlined ubiquitous freshwater Actinobacteria," *The ISME journal*, 12(1), pp. 185–198. Available at: https://doi.org/10.1038/ismej.2017.156.

Olm, M.R. *et al.* (2017) "dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication," *The ISME Journal*, 11(12), pp. 2864–2868. Available at: https://doi.org/10.1038/ismej.2017.126.

Overholt, W.A. *et al.* (2020) "Inclusion of Oxford Nanopore long reads improves all microbial and viral metagenome-assembled genomes from a complex aquifer system," *Environmental Microbiology*, 22(9), pp. 4000–4013. Available at: https://doi.org/10.1111/1462-2920.15186.

Peng, X. *et al.* (2024) "iNAP 2.0: Harnessing metabolic complementarity in microbial network analysis," *iMeta*, 3(5), p. e235. Available at: https://doi.org/10.1002/imt2.235.

Roux, S. *et al.* (2019) "Minimum Information about an Uncultivated Virus Genome (MIUViG)," *Nature Biotechnology*, 37(1), pp. 29–37. Available at: https://doi.org/10.1038/nbt.4306.

Roux, S. *et al.* (2023) "iPHoP: An integrated machine learning framework to maximize host prediction for metagenome-derived viruses of archaea and bacteria," *PLOS Biology*. Edited by M. Arumugam, 21(4), p. e3002083. Available at: https://doi.org/10.1371/journal.pbio.3002083.

Sereika, M. *et al.* (2022) "Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing," *Nature Methods*, 19(7), pp. 823–826. Available at: https://doi.org/10.1038/s41592-022-01539-7.

Shannon, P. *et al.* (2003) "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks," *Genome Research*, 13(11), pp. 2498–2504. Available at: https://doi.org/10.1101/gr.1239303.

Skvortsov, T. *et al.* (2016) "Metagenomic Characterisation of the Viral Community of Lough Neagh, the Largest Freshwater Lake in Ireland," *PLOS ONE*. Edited by B.E. Dutilh, 11(2), p. e0150361. Available at: https://doi.org/10.1371/journal.pone.0150361.

The Genome Standards Consortium *et al.* (2017) "Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea," *Nature Biotechnology*, 35(8), pp. 725–731. Available at: https://doi.org/10.1038/nbt.3893.

Traag, V.A., Waltman, L. and Van Eck, N.J. (2019) "From Louvain to Leiden: guaranteeing well-connected communities," *Scientific Reports*, 9(1), p. 5233. Available at: https://doi.org/10.1038/s41598-019-41695-z.

Uritskiy, G.V., DiRuggiero, J. and Taylor, J. (2018) "MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis," *Microbiome*, 6(1), p. 158. Available at: https://doi.org/10.1186/s40168-018-0541-1.

Wang, T. *et al.* (2025) "Metagenomic analysis reveals how multiple stressors disrupt virus–host interactions in multi-trophic freshwater mesocosms," *Nature Communications*, 16(1), p. 7806. Available at: https://doi.org/10.1038/s41467-025-63162-2.

Warwick-Dugdale, J. *et al.* (2019) “Long-read viral metagenomics captures abundant and microdiverse viral populations and their niche-defining genomic islands,” *PeerJ*, 7, p. e6800. Available at: https://doi.org/10.7717/peerj.6800.

Willems, A. (2014) “The Family Comamonadaceae,” in E. Rosenberg et al. (eds.) *The Prokaryotes*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 777–851. Available at: https://doi.org/10.1007/978-3-642-30197-1_238.

Wood, D.E., Lu, J. and Langmead, B. (2019) “Improved metagenomic analysis with Kraken 2,” *Genome Biology*, 20(1), p. 257. Available at: https://doi.org/10.1186/s13059-019-1891-0.

Zafeiropoulos, H. *et al.* (2025) “microbetag: simplifying microbial network interpretation through annotation, enrichment tests, and metabolic complementarity analysis,” *Genome Biology*, 26(1), p. 292. Available at: https://doi.org/10.1186/s13059-025-03769-2.

SUMMARY

Vilnius University, Faculty of Medicine.

Systems Biology.

Ignas Juodžiūnas.

Cyanobacteria Driven Dynamics of Microbial Communities.

Supervisor Dr Sigitas Šulčius.

Vilnius, 2026.

Cyanobacterial blooms in eutrophic freshwater systems drive predictable succession in the surrounding microbial community, yet the temporal dynamics of these responses remain poorly characterised at the genome-resolved level. Standardised, reproducible pipelines for nanopore-only long-read metagenomics of freshwater communities are also lacking.

This thesis aimed to characterise cyanobacteria-driven microbial community dynamics in a eutrophic freshwater pond using a reproducible nanopore-only metagenomics approach, with objectives covering pipeline development, metagenome-assembled genome (MAG) and viral sequence recovery, co-occurrence network construction, guild-level temporal analysis, and viral community characterisation.

Water samples were collected from Simnas Fish Hatchery Pond at ten timepoints between August and September 2023. Bacterial and viral DNA fractions were sequenced on an Oxford Nanopore MinION using R10.4 chemistry. Assembly was performed with metaFlye, polished with Racon and Medaka, and binned using MetaBAT2 and MaxBin2 consolidated with MetaWRAP. MAG quality and taxonomy were assessed with CheckM2 and GTDB-Tk v2 respectively. Bacterial community profiles were generated with Kraken2, co-occurrence networks inferred with SPIEC-EASI, and communities detected using the Leiden algorithm. Viral sequences were identified with VIBRANT, quality-assessed with CheckV, and annotated with geNomad.

Sequencing yielded 757,511 reads and 3.10 Gbp of data. Thirteen medium-quality MAGs were recovered across four phyla, with Bacteroidota predominating. Nine ecologically interpretable bacterial guilds were identified, showing structured temporal dynamics consistent with bloom senescence succession. Contextualisation against AFA cell count data revealed distinct temporal responses among bacterial families, with Verrucomicrobiaceae tracking peak bloom biomass and Sphingobacteriaceae responding to senescence-derived organic matter. Viral analysis yielded

1,986 sequences, predominantly lytic, dominated by Autographiviridae and Kyanoviridae, with putative auxiliary metabolic genes identified in several genomes.

Nanopore-only metagenomics at modest sequencing depth can yield ecologically meaningful signal from complex freshwater communities. The publicly available pipeline and temporally resolved guild analysis represent the primary contributions of this work.

Keywords: nanopore metagenomics, cyanobacterial blooms, metagenome-assembled genomes, freshwater microbial communities, viral metagenomics

SUMMARY IN LITHUANIAN

Vilniaus Universitetas, Medicinos fakultetas.

Sistemų biologija.

Ignas Juodžiūnas.

Melsvabakterių sukcesijos poveikis mikroorganizmų bendrijos kaitai.

Darbo vadovas dr. Sigitas Šulčius.

Vilnius, 2026.

Cianobakterijų žydėjimas eutrofinėse gėlavandenėse sistemose lemia nuspėjamą aplinkinės mikrobų bendrijos sukcesiją, tačiau šio veiksnio įtaka laiko atžvilgiu lieka prastai ištirta. Taip pat trūksta standartizuotų, atkuriamų nanopore-only ilgų sekų metagenomikos duomenų apdorojimo grandinių skirtų gėlavandenėms bendrijoms.

Šio darbo tikslas buvo apibūdinti cianobakterijų sukeliamą mikrobinių bendrijų dinamiką eutrofiniame gėlavandeniniame tvenkinyje, naudojant atvirai pasiekiamą, lengvai atkartojamą, nanopore-only metagenominių duomenų apdorojimo grandinę. Darbo uždaviniai apėmė bioinformatinės analizės darbo eigos sukūrimą, metagenomo surinktų genomų (MAG) ir virusinių sekų išskyrimą, koegzistavimo tinklo sudarymą, gildijos lygio analizės atlikimą tam tikru laiko periodu ir virusinės bendrijos apibūdinimą.

Vandens mėginiai buvo surinkti Simno žuvivaisos tvenkinyje dešimtyje laiko taškų 2023 m. rugpjūčio-rugsėjo mėnesiais. Bakterinės ir virusinės DNR frakcijos buvo sekvenuotos naudojant Oxford Nanopore MinION su R10.4 chemija. Sekų surinkimas atliktas naudojant metaFlye, poliravimas Racon ir Medaka, o grupavimas MetaBAT2 ir MaxBin2, galiausiai kosnoliduota su MetaWRAP. MAG kokybė ir taksonomija įvertinta atitinkamai naudojant CheckM2 ir GTDB-Tk v2. Bakterinės bendrijos profiliai sukurti naudojant Kraken2, koegzistavimo tinklai nustatyti naudojant SPIEC-EASI, o bendruomenės aptiktos Leideno algoritmu. Virusinės sekos identifikuotos naudojant VIBRANT, kokybė įvertinta CheckV, o anotacija atlikta geNomad.

Sekvenavimo metu gauta 757 511 sekų ir 3,10 Gbp duomenų. Iš keturių tipų išskirti trylika vidutinės kokybės MAG, kuriuose vyravo Bacteroidota tipas. Identifikuotos devynios ekologiškai interpretuotinos bakterijų gildijos, pasižyminčios struktūrizuota laikine dinamika, atitinkančia žydėjimo senescencijos sukcesiją. Bakterijų gildijų palyginimas su *Aphanizomenon flos-aquae*

ląstelių skaičiaus duomenimis atskleide įdomias priklausomybes nuo žydėjimo reiškinio. Verrucomicrobiaceae šeima buvo gausiausia iškart po žydėjimo piko, o Sphingobacteriaceae vėliau, žydėjimui slūgstant ir organinei medžiagai sklaidantis. Virusinė analizė sugeneravo 1 986 sekas, daugiausia lizines, kuriose vyravo Autographiviridae ir Kyanoviridae šeimos. Kai kuriuose genomuose identifikuoti galimi pagalbiniai metaboliniai genai.

Nanopore-only metagenomika esant mažoms sekvenavimo apimtims gali suteikti ekologiškai reikšmingų įžvalgų iš sudėtingų gėlavandenių bendrijų. Pagrindiniai šio darbo indėliai yra viešai prieinama bioinformatinė darbo grandinė ir dešimties laiko taškų gildijų analizė.

Raktažodžiai: nanopore metagenomika, cianobakterijų žydėjimai, metagenome surinkti genomai, gėlavandenės mikrobinės bendrijos, virusų metagenomika.

APPENDICES

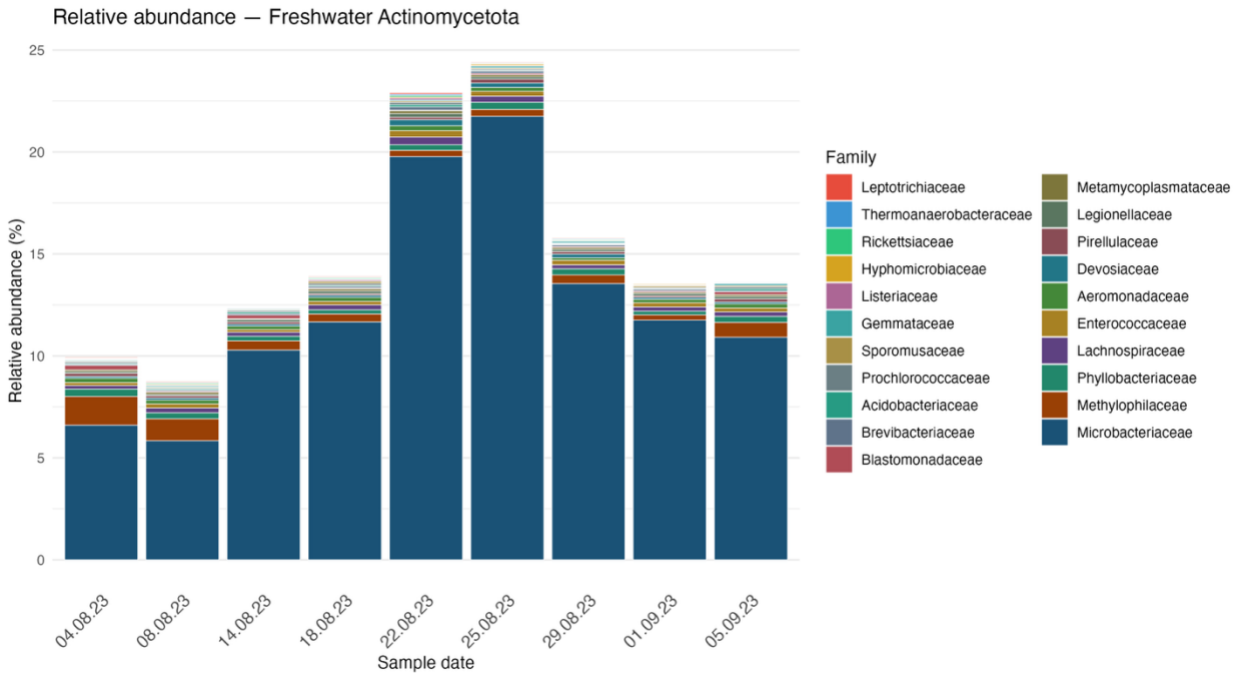

Fig. 5 Family-level relative abundance of the freshwater Actinomycetota guild across ten Simnas Fish Hatchery pond sampling timepoints

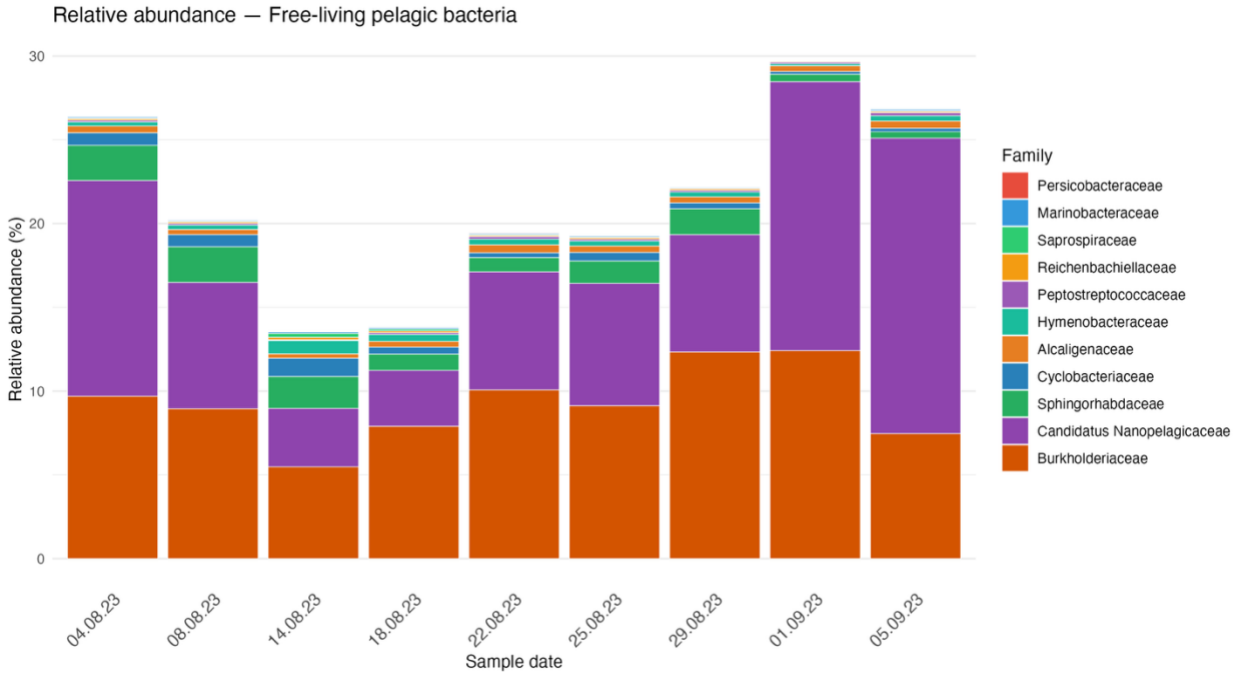

Fig. 6 Family-level relative abundance of the free-living pelagic bacteria guild across ten Simnas Fish Hatchery pond sampling timepoints.

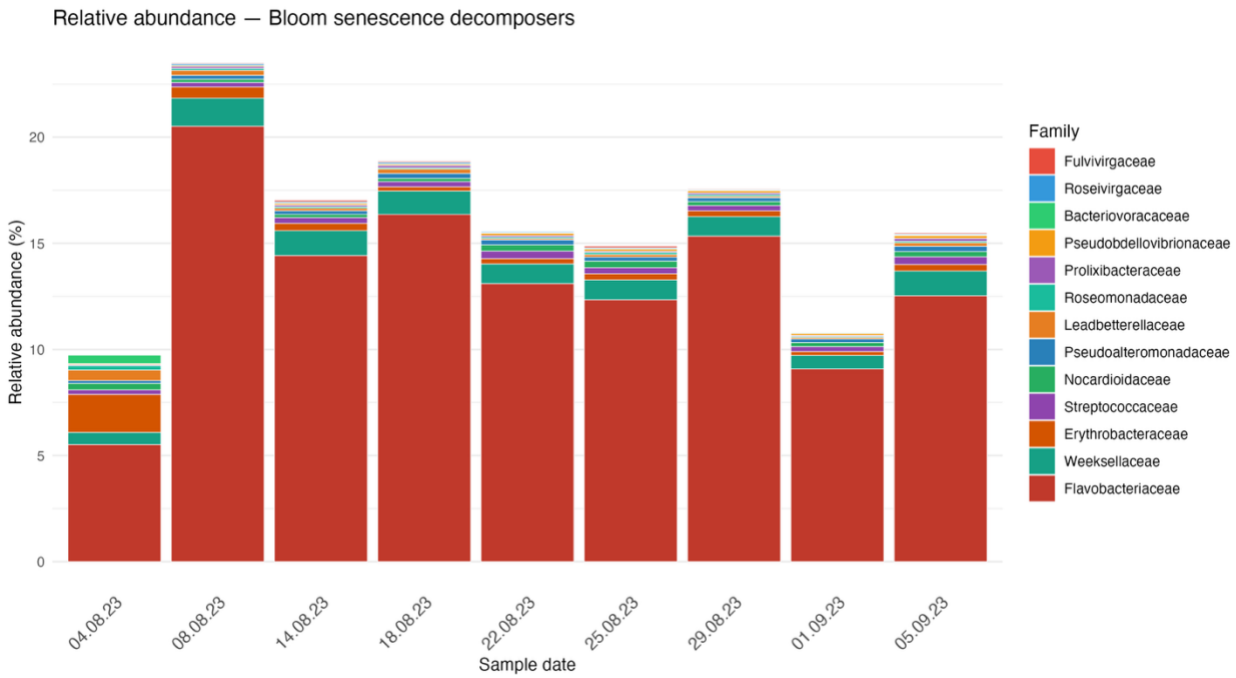

Fig. 7 Family-level relative abundance of the bloom senescence decomposer guild across ten Simnas Fish Hatchery pond sampling timepoints.

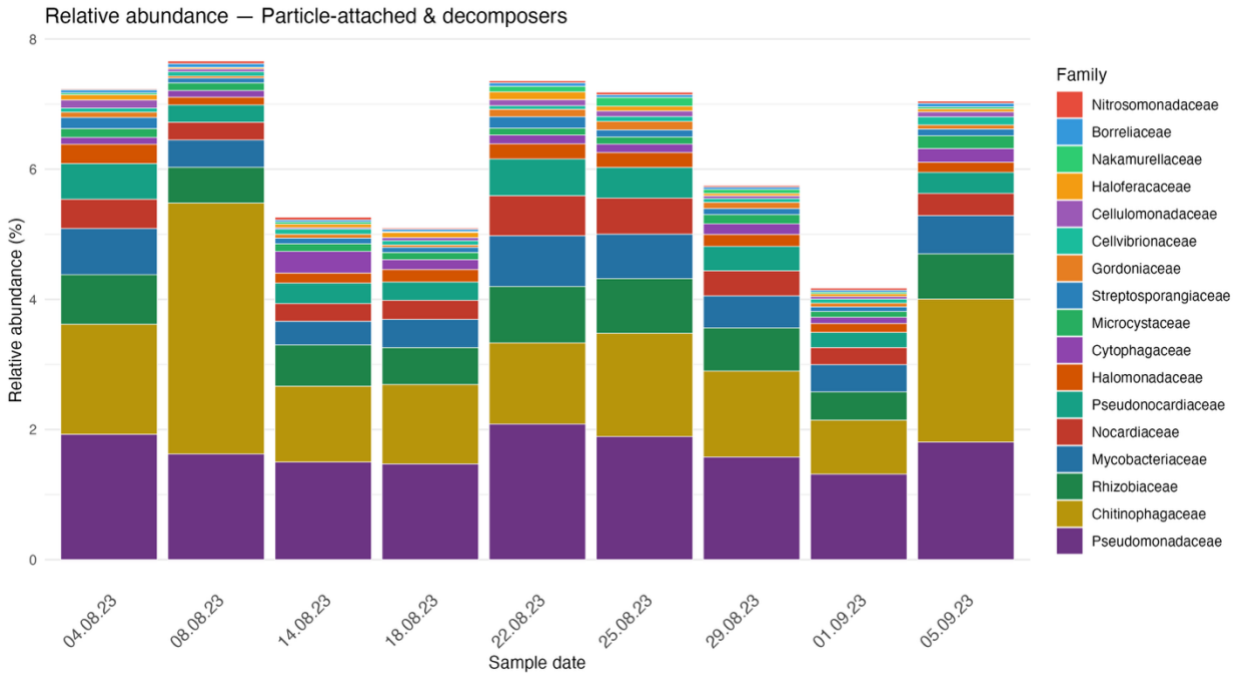

Fig. 8 Family-level relative abundance of the particle-attached and decomposers guild across ten Simnas Fish Hatchery pond sampling timepoints.

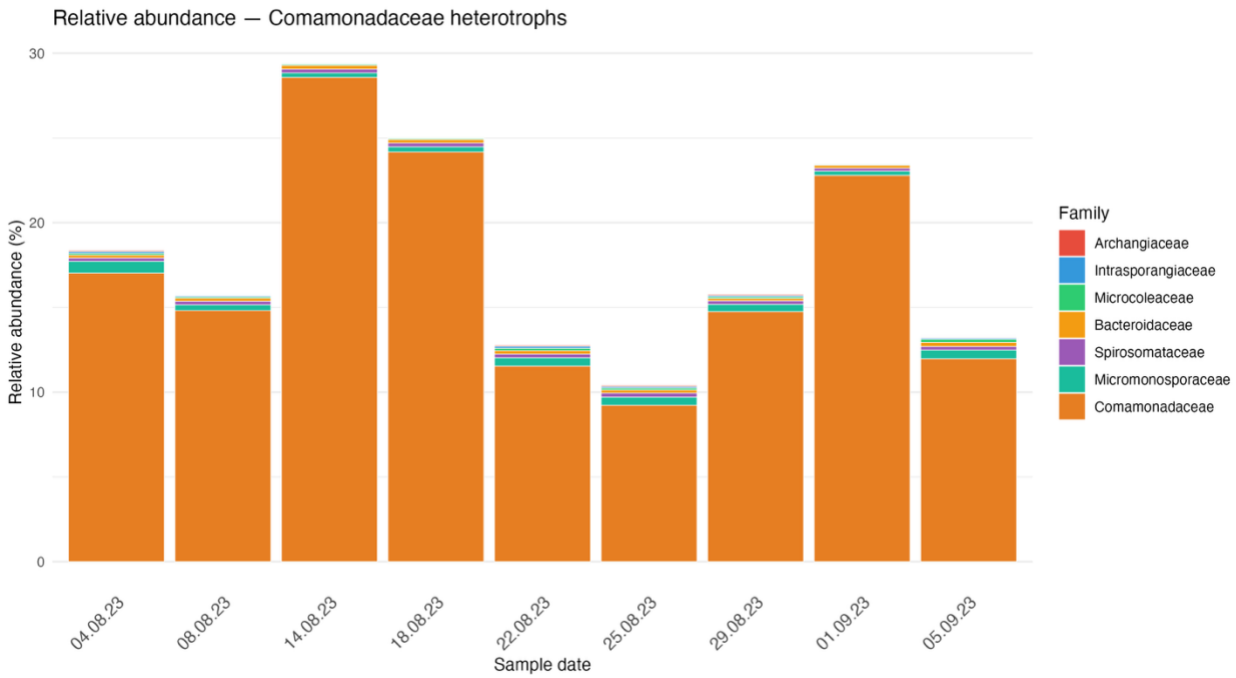

Fig. 9 Family-level relative abundance of the Comamonadaceae heterotrophs guild across ten Simnas Fish Hatchery pond sampling timepoints.

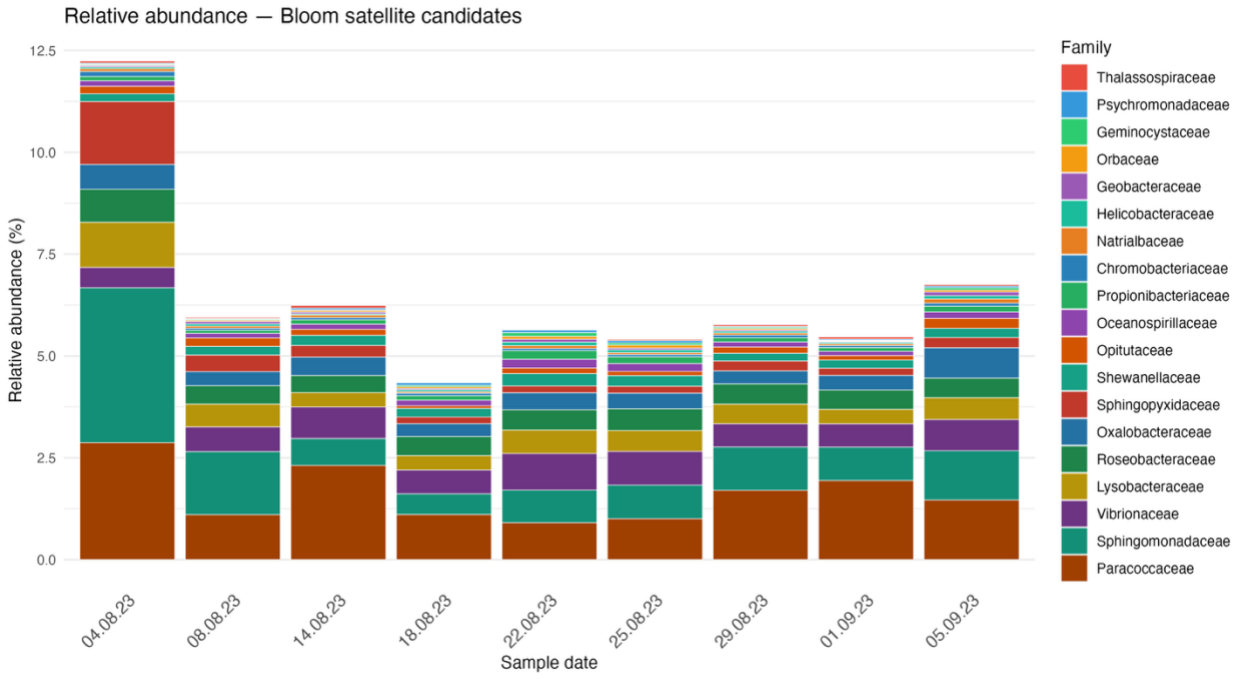

Fig. 10 Family-level relative abundance of the bloom satellite candidates guild across ten Simnas Fish Hatchery pond sampling timepoints.

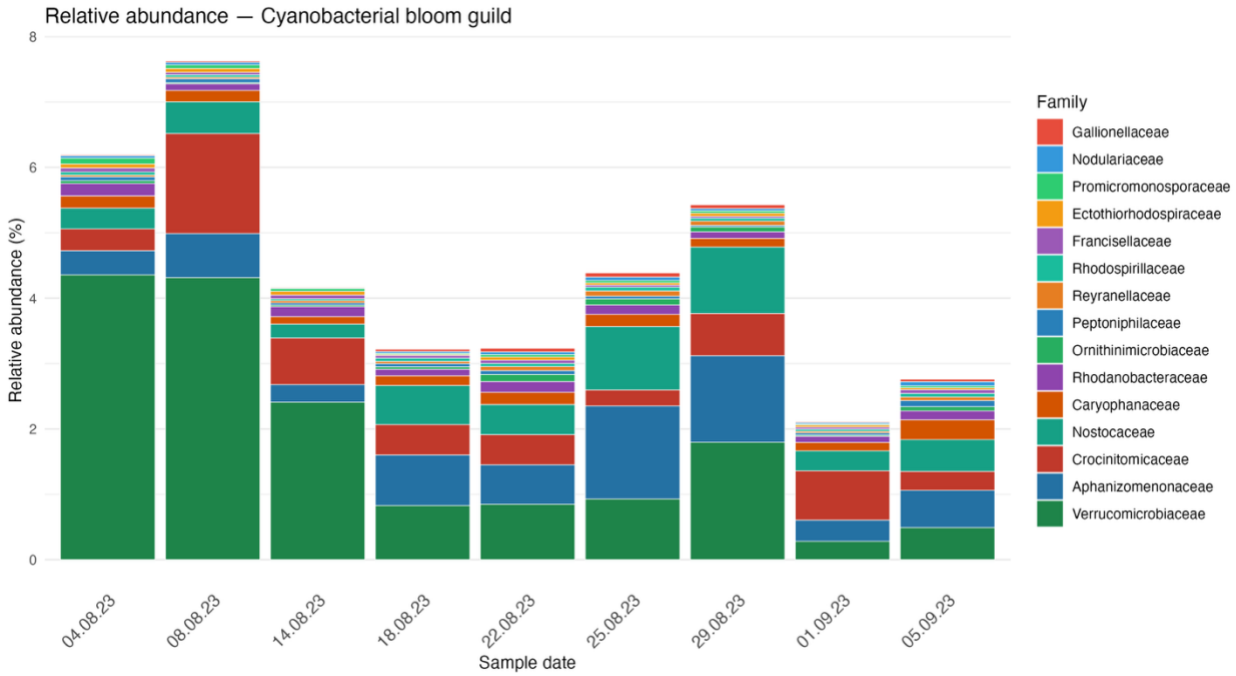

Fig. 11 Family-level relative abundance of the Cyanobacterial bloom guild across ten Simnas Fish Hatchery pond sampling timepoints.

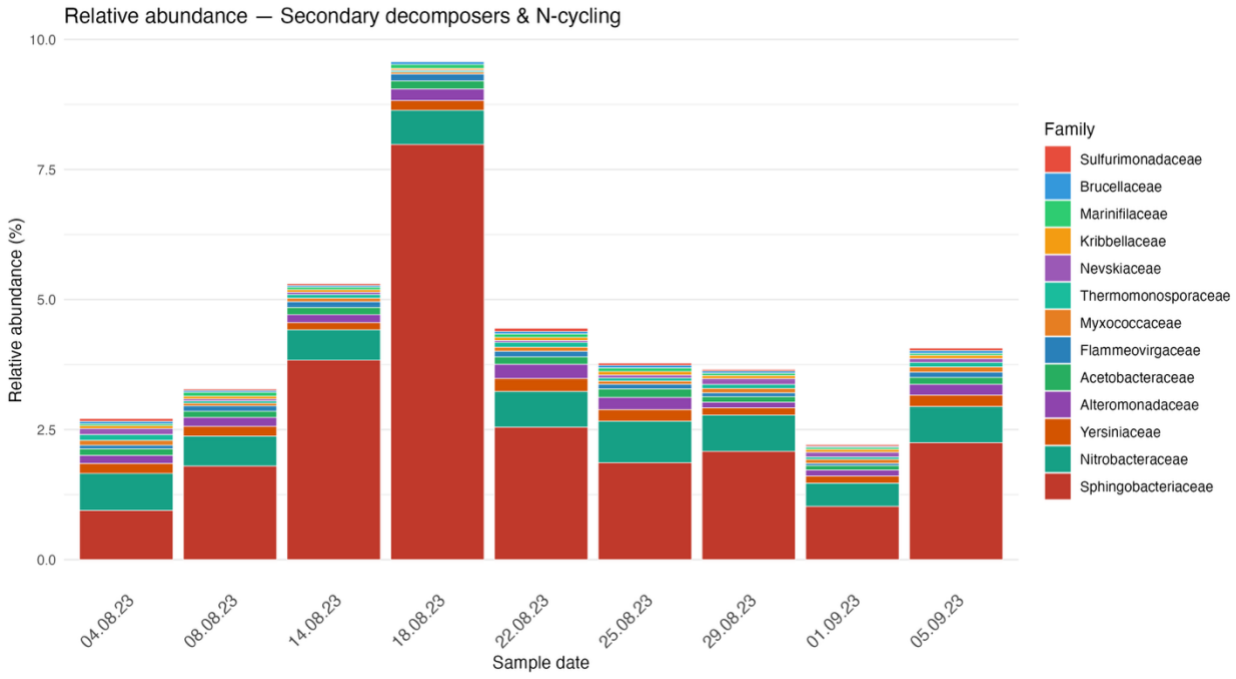

Fig. 12 Family-level relative abundance of the Secondary decomposers & N-cycling guild across ten Simnas Fish Hatchery pond sampling timepoints.

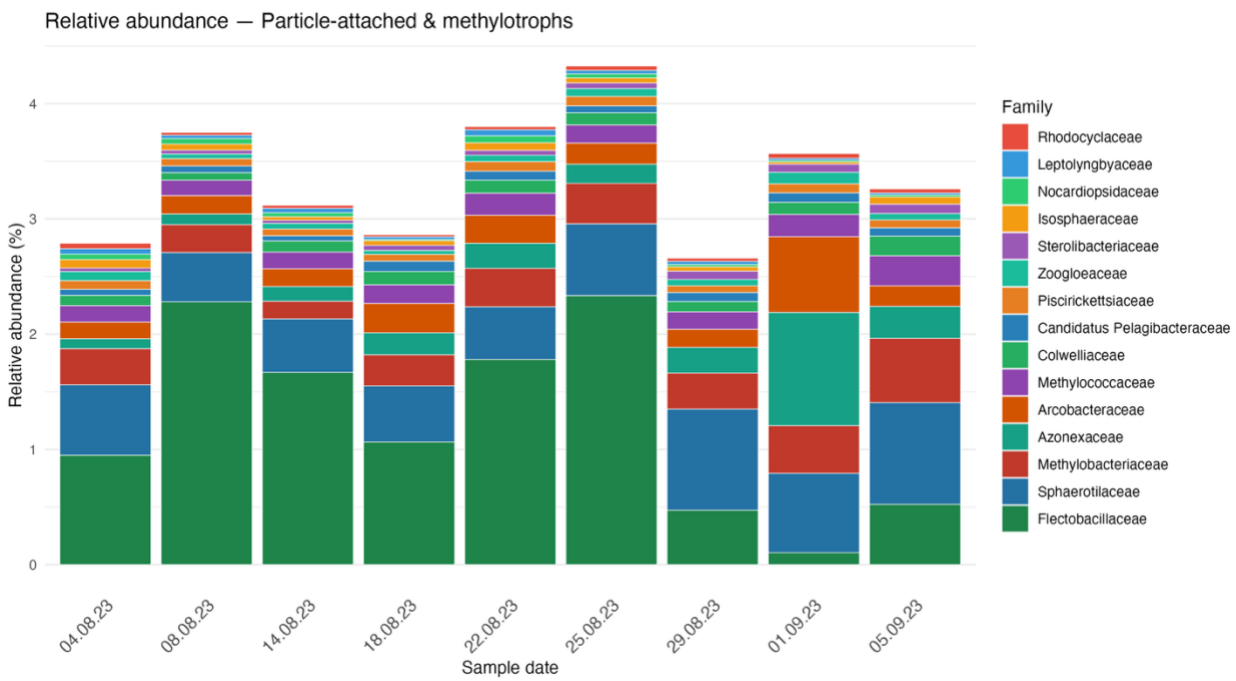

Fig. 13 Family-level relative abundance of the particle-attached and methylotrophs guild across ten Simnas Fish Hatchery pond sampling timepoints.