Advancing Long-Read Metagenomics: Computational Tools for Real-Time Monitoring and Biotechnological Applications

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Timo Niklas Lucas
aus Esslingen am Neckar

Tübingen
2025

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	10.10.2025
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter/-in:	Prof. Dr. Daniel Huson
2. Berichterstatter/-in:	Prof. Dr. Nadine Ziemert

License

Published articles included in this dissertation:

1. Spirito, C.M., Lucas, T.N., Patz, S., Jeon, B.S., Werner, J.J., Trondsen, L.H., Guzman, J.J., Huson, D.H., Angenent, L.T. (2024). *Variability in n-caprylate and n-caproate producing microbiomes in reactors with in-line product extraction.* mSystems.
DOI: https://doi.org/10.1128/msystems.00416-24

2. Lucas, T.N., Biehain, U., Gautam, A., Gemeinhardt, K., Lass, T., Konzalla, S., Ley, R.E., Angenent, L.T., Huson, D.H. (2025). *MMonitor for Real-Time Monitoring of Microbial Communities Using Long Reads.* Cell Reports Methods.
DOI: https://doi.org/10.1016/j.crmeth.2025.101266

3. Gemeinhardt, K., Jeon, B.S., Ntihuga, J.N., Wang, H., Schlaiß, C., Lucas, T.N., Bessarab, I., Nalpas, N., Zhou, N., Usack, J.G., Huson, D.H., Williams, R.B.H., Maček, B., Aristilde, L., Angenent, L.T. (2025). *Toward industrial C8 production: oxygen intrusion drives renewable n-caprylate production from ethanol and acetate via intermediate metabolite production.* Green Chemistry.
DOI: https://doi.org/10.1039/D5GC00411J

Abstract

Metagenomics has expanded our ability to study complex microbial communities, yet its practical application faces barriers including the requirement of programming expertise, and complex command-line workflows that hinder rapid analysis and interpretation. These challenges can affect collaborative research between wet-lab scientists and bioinformaticians, often creating bottlenecks in time-sensitive studies.

This dissertation addresses these challenges by describing accessible computational tools specifically optimized for long-read sequencing data and real-time monitoring of microbial communities. The research includes an investigation of chain-elongating microbiomes in bioreactors, revealing key microbial players and metabolic pathways involved in medium-chain carboxylate production. Through metagenomic analyses, the collaborative work identifies the role of oxygen in n-caprylate production and the complex interactions between aerobic and anaerobic species.

The dissertation introduces MMonitor, a novel software platform that combines computational pipelines with interactive visualization for real-time analysis of metagenomic data from Oxford Nanopore Technologies sequencing. MMonitor's capabilities are demonstrated through applications in tracking bioreactor microbiomes, as well as in contamination control for artificial intelligence training data. Additional methodological contributions include QuickBinDM, which accelerates long-read metagenomic binning through pre-screening approaches, and GeneGone, a web application for the validation of gene deletion experiments.

Together, these tools and analyses advance our understanding of complex microbial communities while making metagenomic analysis more accessible to researchers. The work has implications for bioinformaticians, biologists and researchers from related fields interested in metagenomics and the tracking of microbial communities.

Contents

List of Figures

List of Tables

List of Abbreviations

ACAT	Acetyl-CoA C-acyltransferase
ACD	Acyl-CoA dehydrogenase
AI	Artificial Intelligence
ANI	Average Nucleotide Identity
API	Application Programming Interface
AWS	Amazon Web Services
BAM	Binary Alignment Map
BLAST	Basic Local Alignment Search Tool
BMC	Bacterial Microcompartment
BUSCO	Benchmarking Universal Single-Copy Orthologs
BWT	Burrows-Wheeler Transformation
cDNA	complementary DNA
CDSs	Coding Sequences
CID	Collision-Induced Dissociation
CLI	Command Line Interface
CoAT	Acetyl-CoA transferase
COD	Chemical Oxygen Demand
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
DBG	de Bruijn Graph
DNA	Deoxyribonucleic acid
EBI	European Bioinformatics Institute
ECH	Enoyl-CoA dehydratase
EM	Expectation Maximization
ESI	Electrospray ionization
ETU	Ethanol-utilizing microcompartment

EUT	Ethanolamine utilizing microcompartment
FAB	Fatty Acid Biosynthesis
FM	Ferragina-Manzini (index)
GPU	Graphics Processing Unit
GRM	Glycyl radical enzyme containing microcompartment
GTDB	Genome Taxonomy Database
GUI	Graphical User Interface
HAD	3-hydroxy-acyl-CoA dehydrogenase
HBD	3-hydroxy-butyryl-CoA dehydrogenase
HMM	Hidden Markov Model
HRT	Hydraulic Retention Time
HTTPS	Hypertext Transfer Protocol Secure
ITS	Internal Transcribed Spacer
KEGG	Kyoto Encyclopedia of Genes and Genomes
LCA	Lowest Common Ancestor
LC-MS	Liquid Chromatography-Mass Spectrometry
MAGs	Metagenome-assembled Genomes
MCCs	Medium-chain carboxylates
MIMAG	Minimum Information about a Metagenome-assembled Genome
NCBI	National Center for Biotechnology Information
NGS	Next-Generation Sequencing
OLC	Overlap-Layout-Consensus
OLR	Organic Loading Rate
ONT	Oxford Nanopore Technologies
ORF	Open Reading Frame
OS	Operating System

OTU	Operational Taxonomic Unit
PATRIC	Pathosystems Resource Integration Center
PCoA	Principal Coordinate Analysis
PCR	Polymerase Chain Reaction
PDU	Propanediol utilizing microcompartment
QC	Quality Control
QIITA	Quantitative Insights Into Microbial Ecology
qPCR	Quantitative Polymerase Chain Reaction
QV	Quality Value
RAM	Random Access Memory
RAST	Rapid Annotation using Subsystems Technology
RBOX	Reverse β-oxidation
RNF	Rnf respiratory complex
rRNA	ribosomal Ribonucleic acid
SMS	Single-Molecule Sequencing
SRA	Sequence Read Archive
SSD	Solid State Drive
TE	Thioesterase
TFA	Trifluoroacetic acid
WGS	Whole Genome Sequencing

Chapter 1

Introduction

Motivation and Scope

Metagenomic analysis through DNA sequencing is a powerful tool for studying microbial communities, especially for microbes that are difficult to culture. The rapid advancement of sequencing technologies, particularly long-read platforms, has created new opportunities for detailed analysis of microbial communities but also new challenges. This dissertation describes the development and application of computational methods for analyzing metagenomic data, with an emphasis on real-time monitoring of microbial communities using long-read sequencing and the analysis of complex bioreactor microbiomes using both established and our newly developed methods.

Dissertation Outline

This dissertation is organized into several interconnected chapters:

- Chapter 2 provides a background on environmental biotechnology, microbial communities in bioreactors, and computational tools for metagenomic analysis.
- Chapter 3 presents a study where we performed a detailed metagenomic analysis of samples from n-caprylate-producing bioreactors, revealing key microorganisms and metabolic pathways.
- Chapter 4 introduces MMonitor, a novel software tool for real-time monitoring of microbial communities using long-read sequencing that was actively developed in collaboration with environmental biotechnology researchers.
- Chapter 5 describes additional software tools and collaborative projects. In this

chapter, we describe how we extended our analyses of bioreactor communities to further study n-caprylate production and how we applied MMonitor for different use cases. Those include the analysis of bioreactors mimicking conditions of the human nasal cavity and contamination detection for isolate sequencing data used for training an AI model. We also describe the development of other tools, including a pipeline for faster frame-shift corrected binning of metagenomic contigs and a tool for validating gene deletions made by gene editing experiments.

- Chapter 6 summarizes the work and discusses its broader implications, strengths and weaknesses, and outlines possible future research directions.

Together, these chapters represent contributions to the field of computational metagenomics, with practical applications in environmental biotechnology and related fields.

Chapter 2

Background and Literature Review

Here we introduce basic concepts and methods potentially helpful for understanding the work presented in this thesis. We start with a general introduction to bioreactors and methods for studying their microbial communities, followed by an in-depth overview of computational tools for metagenomic analysis. Furthermore, specific background related to published work can be found in the introduction section of Chapter 3 and Chapter 4.

Microbial Ecology in Engineered Environments

Bioreactors as Model Systems for Microbial Communities

Metagenomics is the study of genetic material from microbial communities in environmental samples. It has many use cases, one of which is the study of bioreactors. Bioreactors are artificial, controlled environments where complex microbial communities grow under controlled conditions to metabolize substrates into valuable products [1]. In industry, they are used for the production of chemicals, food additives, and other valuable products [2]. They often grow on waste products and have important roles in the circular economy. They act as an environmentally friendly alternative to conventional chemical production [1]. In research, bioreactors are used to cultivate and study microbial communities, often with the goal of understanding how certain products are produced within this complex metabolic network and steering the microbial production pathways, and isolating specific microbes for further study [3, 4].

Characterizing Bioreactor Microbiomes: Methodological Approaches

Amplicon sequencing targeting the 16S rRNA or 18S rRNA genes provides taxonomic profiling of bacterial and eukaryotic communities, offering insights into community composition and diversity [5]. Shotgun metagenomics goes beyond taxonomy to reveal the functional potential of the community by sequencing all genomic material present [5]. This approach enables the reconstruction of genomes, and thus metabolic pathways, and the identification of genes involved in substrate utilization and product formation.

For understanding and quantifying active processes, metatranscriptomics captures gene expression patterns by sequencing community RNA, revealing which organisms and pathways are actively functioning under specific conditions by identifying which genes are being over- or under-expressed [6]. Complementary techniques, such as metaproteomics and metabolomics, provide further functional insights by identifying proteins being expressed and metabolites being produced or consumed [6]. At the cellular level, flow cytometry and electron microscopy techniques allow for the visualization and quantification of different cell populations, their physical characteristics, and spatial organization within the bioreactor environment [7, 3]. Quantitative PCR (qPCR) represents another tool for monitoring specific microbial populations within bioreactors. This technique allows for the targeted quantification of specific taxa or functional genes by measuring the amplification of DNA [8].

Challenges in Bioreactor Microbiome Analysis

Despite these advanced methodologies, several challenges persist in bioreactor microbiome analysis. The high complexity of mixed microbial communities, often containing hundreds to thousands of species with varying abundances, makes comprehensive characterization difficult [9]. Bioreactor communities also exhibit temporal dynamics, requiring frequent sampling to capture community shifts in response to changes in operational parameters [9]. A significant challenge remains in establishing clear links between community structure and functional performance, understanding which organisms contribute to specific metabolic processes, and how their interactions affect overall bioreactor performance [10]. Additionally, distinguishing between active and dormant populations is crucial, as the mere presence of an organism does not necessarily indicate its contribution to observed functions [11].

Microbial Chain Elongation: Ecology and Key Players

Chain elongation bioreactors harbor specialized microbial communities that convert short-chain fatty acids and alcohols into medium-chain carboxylates (MCCs). Recent studies suggest that these communities utilize the reverse β-oxidation (RBOX) pathway, adding two carbons to the chain of the substrate with each cycle [4]. These communities typically feature syntrophic relationships between different groups of microorganisms, including primary fermenters, chain elongators, and often methanogens or other hydrogen-consuming organisms that maintain favorable thermodynamic conditions.

Key players in chain elongation communities often include members of the genus *Clostridium*, particularly *C. kluyveri*, which has been extensively studied for its ability to elongate acetate using ethanol as an electron donor [12]. Other important taxa include various members of the Firmicutes phylum, such as *Megasphaera*, *Caproiciproducens*, and *Ruminococcaceae*, which have been implicated in the production of caproate, caprylate, and other MCCs in different bioreactor configurations [12]. Recent metagenomic studies have also highlighted the importance of *Oscillibacter* species in chain elongation bioreactors, particularly those producing n-caprylate (C8) [12]. *Oscillibacter* is a genus of anaerobic bacteria belonging to the Firmicutes phylum and has been consistently observed in bioreactors achieving high n-caprylate production rates.

The composition and dynamics of chain-elongating communities are strongly influenced by operational parameters such as pH, temperature, organic loading rate, and hydraulic retention time. These parameters can be manipulated to selectively enrich for specific microbe groups and optimize the production of target products. The complex interplay between community structure and reactor performance makes metagenomic analysis particularly valuable for understanding and optimizing these systems if the underlying pathways are known [13, 14, 15].

Further background related to chain-elongating communities and the RBOX pathway can be found at the beginning of Chapter 3.

Computational Metagenomics: Tools and Techniques

The analysis of metagenomic data requires specialized computational tools to extract meaningful biological insights from large, complex datasets. This section reviews the current landscape of computational approaches used in metagenomic analysis, focusing on methods that have been used in the work presented in this thesis, together with similar methods to complement the big picture.

Data Preprocessing and Quality Control

Preprocessing of metagenomic data is often the first step of metagenomic analysis pipelines. It may include quality control with tools like FastQC [16] or fastp [17], read filtering by length or quality, or trimming with tools like Filtlong [18] or seqkit [19], adapter removal with cutadapt [20], error correction with tools like Pilon [21] or Medaka [22], and host DNA removal using popular aligners like Bowtie2 [23] or BWA [24]. Quality control can be used to determine which samples or reads to use for further downstream analyses, and it is often a good idea to work with a more curated, smaller set of reads with higher quality for most analyses to save computational resources and improve the quality of the results.

Taxonomic Profiling Strategies

Taxonomic profilers are computational tools used in metagenomics analysis that classify reads into taxonomic categories. The input is usually sequencing data, and the output is a taxonomic profile, a report of the taxonomic composition of the sample. A variety of classification algorithms and strategies have been developed depending on the data input type. Typical approaches assign reads into taxonomic groups through read mapping by alignment of reads, often using k-mers and similar concepts like minimizers [25]. Afterwards, often a consensus step is applied to reduce false positives for reads with ambiguous alignments, like the lowest common ancestor (LCA) algorithm or the expectation maximization (EM) algorithm [26, 27].

Alignment-Based Methods

A straightforward approach to taxonomic profiling is aligning every read against a large reference database like NCBI-nr or GTDB and then counting the number of reads that map to each taxon. Alignment-based profilers are very accurate and sensitive, able to find distant relatives of query sequences in the reference database, but they are also computationally expensive and may require a lot of memory depending on the size of the reference database. Popular alignment-based tools that can be used for metagenomic profiling include DIAMOND+MEGAN [28, 26] and Kaiju [29] using translated protein alignments or MetaPhlAn [30] using marker gene alignments.

K-mer-Based Classification

K-mer-based classification approaches are a popular alternative to alignment-based methods. They compute k-mers by splitting the query and reference sequences into smaller units and store those in some sort of index structure, like a hash table or FM

index. Instead of comparing the entire read to the reference database, only the k-mers are compared to the reference index, and each read is classified based on the k-mers it contains. Popular k-mer-based tools that can be used for metagenomic profiling include Kraken2 [31] and Centrifuge [32]. K-mer-based methods typically run a few orders of magnitude faster than alignment-based methods but are less accurate and more sensitive to sequencing errors [33].

Classifiers for Long-Read Data

With the advent of single molecule sequencing (SMS) technologies, new taxonomic profilers optimized for noisy long reads have been developed. Traditional k-mer-based taxonomic classifiers, which rely on exact k-mer matching, tend to suffer from reduced performance on long-read data due to higher error rates [33]. Even a small number of sequencing errors can significantly distort the set of k-mers in a read, making exact matches to reference databases unreliable.

To address this, several improvements have been proposed. A workaround is to use error correction tools to first correct the long reads and then use a k-mer-based profiler. Other approaches include using minimizers instead of standard k-mers [25], which are less affected by sequencing errors. Syncmers offer another alternative by selecting a more consistent subset of k-mers [34]. More recently, strobemers, which are non-contiguous k-mer-like constructs, have shown improved sensitivity and specificity over traditional minimizers by linking spaced k-mers, thereby providing better alignment seeds under noisy conditions [35]. These methods preserve locality while reducing the chance that a sequencing error disrupts the match entirely. Although not yet widely integrated into metagenomic classifiers, recent tools and studies have demonstrated their utility in improving read classification under noisy conditions, suggesting that future long-read taxonomic profilers could benefit from adopting these more error-tolerant data structures [35].

Nowadays, taxonomic profilers for long reads often use a combination of such strategies. BugSeq [36] and Emu [27] both employ Minimap2 [25], which uses minimizers to identify potential matches before extending them into alignments. DIAMOND+MEGAN-LR [37] uses frame-shift aware translated protein alignments, followed by a weighted LCA algorithm to generate taxonomic profiles. Recent benchmark studies have evaluated these tools systematically. For example, Portik et al. showed that alignment-based methods like BugSeq and DIAMOND+MEGAN-LR generally outperform k-mer-based methods in terms of precision and recall on long-read datasets [33]. Similarly, Marić et al. found that while k-mer tools like Kraken2 offer fast classification, alignment-based tools provide superior accuracy in identifying low-abundance species [38].

As aligning against huge databases is computationally expensive, there is still room for improvement for further algorithmic development in long-read taxonomic profiling with the goal of creating highly accurate taxonomic profilers that are also fast enough to assign reads in real time.

Metagenomic Assembly

While many metagenomics tools rely only on raw metagenomic sequencing reads, assembling reads to create metagenome-assembled genomes (MAGs) is a requirement for building high-quality reference genomes that can be used for functional annotation and other subsequent analyses. Metagenome assemblers make use of the fact that sequencing errors appear randomly, and so by assembling many reads spanning the same region, the errors can be corrected, yielding fewer, but longer and more accurate contigs. It has been shown that when the assembly completeness is high, using MAGs can outperform raw reads in tasks like taxonomic profiling and functional annotation [39].

Metagenomic assemblers are often extensions of algorithms developed for genome assembly that have been adapted to the uneven coverage of reads in metagenomic data [40]. In the past, genome assemblers were roughly grouped into two categories: overlap-layout-consensus (OLC) and de Bruijn graph (DBG) methods. OLC methods usually first find overlaps between reads, then use those overlaps to construct a layout graph, and from that infer a consensus sequence [41]. DBG methods first split reads into smaller k-mers and k-1-mers and then use those to build a graph where the nodes are k-1-mers and the edges are directed k-mers. Traversing the graph by visiting each edge once can reconstruct the original genome sequence [42].

Over the years, different metagenome assemblers have been developed for different sequencing technologies, utilizing different algorithms and strategies, each with its own strengths and weaknesses. Next we will introduce and discuss some popular (meta)genome assembly tools.

Short-Read Assembly Approaches

Omega [43] is a metagenome assembler for short reads that uses an OLC-based approach, which struggled with distinguishing chimeric contigs from real ones and had issues with large input sizes [44]. MetaVelvet is a DBG-based assembler that uses the coverage of individual k-mers to differentiate between genomes [44, 45]. Other k-mer-based approaches like IDBA-UD and MEGAHIT tried to solve the problem of selecting the correct k-mer size for building the DBG by using a combination of increasing k-mer sizes to build and refine the DBG iteratively [44, 46, 47]. Another popular DBG

assembler is metaSPAdes [48], which constructs a DBG with different k-mer sizes and incorporates different graph simplifications while preserving strain variation.

Recent benchmark studies have evaluated short-read metagenomic assemblers. Goussarov et al. compared multiple tools and found that MEGAHIT recovered a slightly higher fraction of genomes than metaSPAdes, but produced more fragmented assemblies and more chimeric contigs. metaSPAdes yielded more contiguous assemblies and fewer chimeric contigs at the cost of higher runtime [49].

Long-Read Assembly Approaches

Long-read assemblers rely on SMS technologies like Oxford Nanopore and PacBio. By taking advantage of the longer read lengths, they generate longer contigs, often spanning entire microbial genomes, and are able to resolve long repeats. The high error rate of SMS reads, however, poses challenges for most short-read DBG-based assemblers, as most k-mers are not present in multiple reads [50]. Initially, short-read assemblers were used for assembling SMS data, but recently many methods built with SMS data in mind have been developed. These include Canu [51], Flye [50], miniasm [52], Raven [53], wtdbg2 [54], and more, of which not all have been adapted for metagenomic data. Next we provide more details on some of the most popular long-read assemblers and discuss their different strategies for handling noisy metagenomic SMS data.

Canu is a long-read assembler built for SMS data that uses the OLC approach. It uses a tf-idf weighted MinHash approach to find overlaps between reads, assigning higher weights to rare k-mers and lower weights to repetitive k-mers, reducing noise from repeats. Canu builds overlap graphs with a modified greedy best overlap graph (BOG) algorithm [55], using only the best overlaps for each read and filtering out overlaps produced by repeats. Reads are realigned to the layout graph to create a consensus sequence. Canu builds reliable assemblies even with high error rate input sequences and low read depths, but has problems with circularization and slow runtimes compared to other long-read assemblers [56]. Canu can be used for assembling metagenomic data, but it was not built specifically for it and can struggle with complex communities.

Hifiasm-meta is an assembler built for highly accurate long-read data generated by PacBio sequencers. It is based on the hifiasm [57] assembler, which uses minimizers to quickly find overlaps and then uses a string graph to build highly accurate assemblies. Hifiasm-meta contains modifications to better handle metagenomic data, like a prefiltering step that reduces coverage of highly abundant species without removing low-abundance ones, or graph construction algorithms that protect reads coming from low-coverage genomes [58]. Hifiasm-meta is able to create strain-resolved metage-

nomic assemblies, which is good for complex communities with closely related strains. However, string graphs scale poorly with a high number of reads [59] and thus can be computationally expensive for large metagenomic datasets.

metaFlye [60] is a long-read assembler for metagenomic long-read data from both PacBio and ONT and is an extension of the original Flye assembler [50]. The original Flye assembler uses a repeat graph, similar to a DBG, but instead of using k-mers, it builds a repeat graph based on error-prone read overlaps, making it robust to sequencing errors. For this, it first chains reads into longer disjointigs and then builds a repeat graph from their overlaps. It then aligns reads to the repeat graph to untangle the graph and applies multiple simplifications to the graph to first resolve repeats bridged by reads and then those that are not bridged. Flye is a reliable assembler for long-read data that generates fewer indel errors in the assemblies compared to other methods; however, it has high memory usage [56]. metaFlye provides adapted algorithms for handling the assembly of reads with highly uneven coverage distributions and multiple closely related strains [60]. To achieve this, it combines global k-mer counting with analyzing local k-mer distributions, an algorithm for detecting repeat edges that is robust to non-uniform distribution of read coverages, and specialized graph simplification algorithms for resolving strain-induced variations in the repeat graph, generating contigs that are less contiguous but better suited for strain analysis [60].

Benchmarking efforts have also focused on long-read assemblers. In a recent study by Goussarov et al., metaFlye achieved higher completeness compared to Canu, was significantly faster and required less memory. metaFlye also showed better robustness to low coverage, while Canu provided more contiguous assemblies at high coverage [49].

Hybrid Assembly Strategies

Since short and long-read data have different properties, hybrid assembly approaches have been developed to combine the strengths of both. Some examples of hybrid assembly approaches are: DBG2OLC [61], hybridSPAdes [62], MaSuRCA [63], OPERA-LG [64], OPERA-MS [65], and Unicycler [66].

DBG2OLC, as the name suggests, first builds contigs from short reads using a DBG and then aligns the contigs to the long reads, efficiently generating an overlap graph. Like other OLC assemblers, it then constructs a draft assembly by extracting linear paths from the overlap graph and then the final assembly with a consensus step. Using a similar approach, hybridSPAdes first builds a DBG from short reads, but then aligns the long reads to the graph, closing gaps in the graph using the consensus of the long reads and resolving repeats by passing the long-read paths to the exSPANDER module of SPAdes [62].

MaSuRCA does not use SMS reads but instead uses a combination of short Illumina and slightly longer 454 and Sanger sequencing reads. It first transforms the Illumina reads into longer consensus reads that the authors call super reads [63]. It then uses those super reads along with the 454, Sanger, and mate-pair information to create the final assembly.

OPERA-LG is technically a genome scaffolder that operates on pre-assembled contigs generated by short-read assemblers. Used in conjunction with long reads that are converted into synthetic mate-pair links, it then builds a scaffold graph where the edges are the links and the nodes are the contigs. OPERA-LG uses an exact algorithm to find the scaffold arrangement that minimizes discordant links, handling repeats using coverage information [64].

OPERA-MS extends OPERA-LG from isolate genome scaffolding to metagenomic assembly. While OPERA-LG uses long reads as synthetic mate-pairs to scaffold preassembled contigs, OPERA-MS incorporates short-read metagenomic assemblers like MEGAHIT or metaSPAdes to build contigs and adds a clustering step that separates species and strains based on contig coverage and graph connectivity. Each cluster is then scaffolded using OPERA-LG, enabling strain-resolved hybrid assemblies in complex communities [65].

Highly accurate long-read sequencing, such as PacBio HiFi, has significantly reduced the need for hybrid assembly, particularly for bacterial isolates, by enabling nearly complete chromosome assemblies from long reads alone [56, 67]. However, hybrid methods remain valuable in metagenomics and complex microbial communities, where short-read data can complement long-read assemblies by providing higher depth, correcting errors, and resolving strain-level diversity [65].

Hybrid assembly approaches have also been benchmarked in complex microbial communities. A study by Tao et al. found that a hybrid metaSPAdes assembly combining Illumina and HiFi reads recovered the most medium- and high-quality MAGs among all tested strategies. Pure HiFi assemblies yielded longer contigs and better rRNA completeness, but hybrid methods gave the best MAG recovery overall. The study also reported that each approach had different strengths and recommended evaluating a combination of hybrid, short and long read approaches [68].

Metagenome Binning and Refinement

Metagenomic binning is the process of grouping assembled contigs into bins that represent individual genomes, known as MAGs. Modern binning approaches typically use a combination of sequence composition (e.g., GC content, tetranucleotide frequencies)

and coverage information across samples to group contigs likely originating from the same genome [69].

Several popular binning tools have emerged, including MetaBAT2, CONCOCT, and MaxBin2, which use different statistical approaches to cluster contigs [70, 71, 72]. MetaBAT2 employs a modified k-medoid clustering algorithm that considers both tetranucleotide frequencies and co-abundance patterns across samples [70]. CONCOCT uses Gaussian mixture models to cluster contigs based on sequence composition and coverage across multiple samples [71]. MaxBin2 implements an expectation-maximization algorithm to iteratively refine bins [72].

Recent developments in binning technology have focused on improving accuracy and automation. Tools like BusyBee Web provide web-based interfaces for differential composition-based binning [73], while others like VAMB use variational autoencoders to learn optimal representations of contig features for clustering [74]. Some newer approaches also incorporate long-read sequencing data to improve binning accuracy, particularly for closely related strains.

Recent evaluations of binning tools on real and synthetic datasets show that multi-sample binning can dramatically improve MAG recovery. Han et al. found that tools like MetaBAT2, VAMB, and MetaDecoder were highly computationally efficient, and that refinement pipelines such as MetaWRAP further improved quality. The same study showed that binning across multiple samples increased high-quality bin recovery by up to 233% compared to single-sample approaches [75].

Wang et al. benchmarked binning performance on virome data and showed that traditional tools like CONCOCT produced bins with higher contamination, while MetaBAT2 and VAMB yielded higher purity and better strain resolution [76].

Despite these advances, binning remains challenging for complex communities with many closely related species or strains, as these can have similar sequence compositions and coverage patterns. Additionally, incomplete assembly, particularly of low-abundance organisms, can lead to fragmented or missing bins. To address these limitations, ensemble approaches that combine multiple binning tools, followed by refinement steps, are often used in practice.

Consensus and Deduplication Methods

Several ensemble approaches have been developed to leverage the strengths of different binning tools. DAS Tool combines predictions from multiple binners, selecting the highest quality bins while reducing redundancy [77]. MetaWRAP provides a pipeline that integrates multiple binning tools and includes bin refinement modules to improve MAG

quality [78]. These ensemble methods have been shown to recover more high-quality MAGs compared to individual binning tools, particularly from complex communities where different binners may perform better on different subsets of the data.

Quality assessment tools can identify redundant or highly similar MAGs generated from different assemblies or binning methods of the same sample. Tools have emerged to deduplicate and consolidate these redundant MAGs into consensus genomes. dRep uses average nucleotide identity (ANI) to cluster similar genomes and selects the highest-quality representative from each cluster [79]. This helps reduce redundancy while retaining the best-quality MAG for each distinct genome.

These deduplication and consensus approaches are particularly valuable when working with multiple assembly strategies or binning tools, as they help consolidate the results into a non-redundant, high-quality set of MAGs that better represent the true genomic diversity of the community. The consensus approach can also help compensate for the different biases and limitations of individual assembly or binning methods.

Assembly Quality Assessment and Improvement

Assessing the quality of assemblies can be important to validate the assembly and to check its feasibility for downstream analyses. It is especially important for metagenomic assemblies, where the complex community can lead to contamination of MAGs with other closely related strains and where the quality of each MAG is additionally influenced by the binning method used. Different approaches have emerged for this purpose: tools that report statistics about the assembly and compare assemblies to a reference genome, like Quast and MetaQuast, and those that identify marker genes or other features like k-mers in the genome bins, such as BUSCO, CheckM, CheckM2, and Merqury [80, 81, 69, 82].

BUSCO searches for a set of universal single-copy orthologs to estimate completeness, reporting how many are found as complete, fragmented, or missing [80]. CheckM uses lineage-specific marker genes to estimate both completeness and contamination of microbial genomes based on expected gene presence [81]. CheckM2, a deep learning-based successor, improves accuracy by using machine learning models trained on large genome datasets to better estimate quality for diverse taxa [69]. Merqury evaluates assembly quality without a reference by comparing k-mers from the original read set to those found in the assembly, estimating completeness, base-level accuracy (QV), and phasing [82].

Bandage is a software tool that lets users load assembly graphs obtained by different assemblers and visualizes their topologies. It efficiently lays out the graphs, showing

connections between contigs and other useful information like contig coverage and contig lengths. Bandage can be useful for manually checking assembly graphs for large metagenomic assemblies and deciding if they are good for downstream analyses or if they should rather be corrected or recomputed with different parameters. Figure 2.1 shows an example of a metagenomic assembly graph generated by a short-read assembler and visualized with Bandage, illustrating the complex connectivity through sequencing errors, repeated regions, strain variation, or incomplete sequencing. If the coverage was very high, sometimes subsampling the reads can help to resolve the graph better, or if it's too low then sequencing longer for more coverage can help. Generally, discarding short and low-quality reads improves assembly contiguity, but even if the graphs are not fully resolved, it can still be used to recover high-quality MAGs when combined with assembly polishing and metagenome binning.

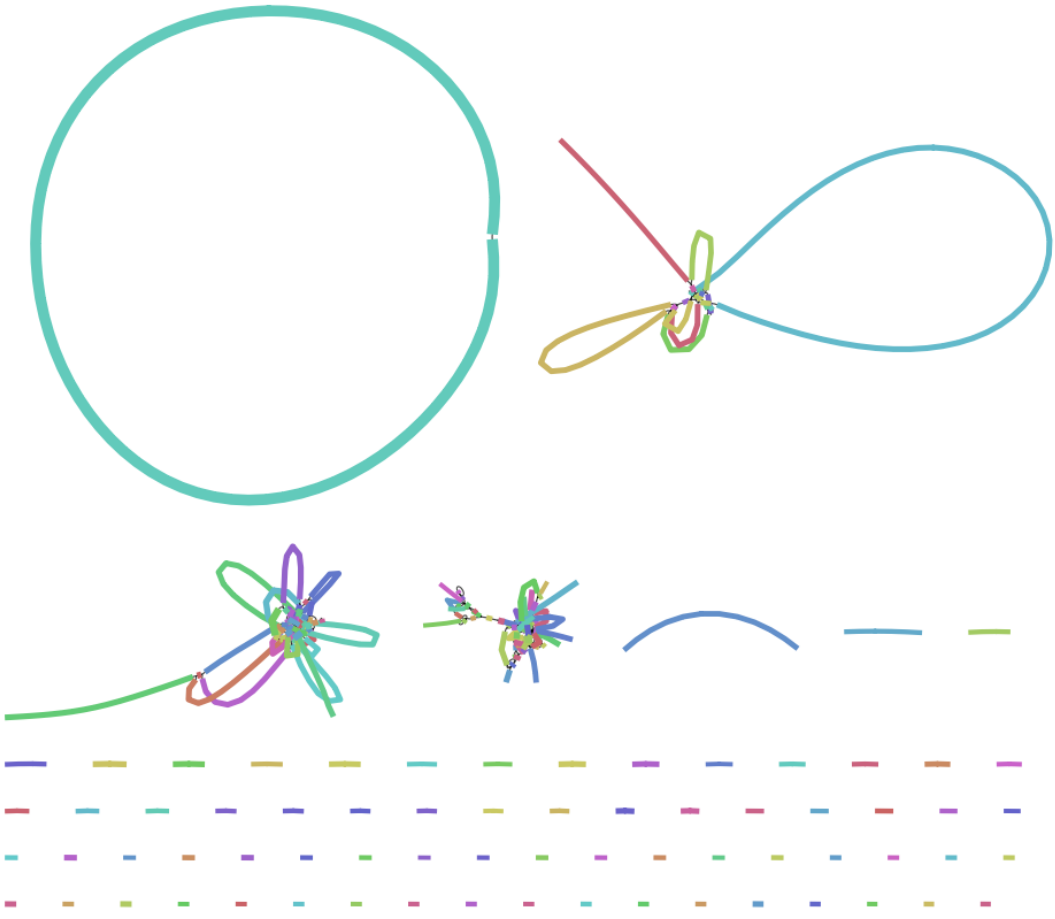

Figure 2.1: Visualization of a metagenomic assembly graph using Bandage. Each node represents a contig, and edges represent connections inferred during assembly. Tangled structures in the graph indicate the presence of similar, but not identical variants of the same genome introduced through unresolved repeats, strain variation, or sequencing errors in the metagenomic sample.

Assembly Quality Improvement

After assembly, various approaches can be employed to improve the quality of metagenomic assemblies. These include polishing, error correction, and consensus generation from multiple assemblies.

Bandage can be used to refine a metagenomic assembly by annotating contigs with metadata and then manually choosing paths through ambiguous regions of the assembly graph. Inspecting the assembly graph can help identify loops in the assembly graph, which, together with other information provided like GC content and coverage, can be used to find errors or strain variation.

Assembly polishing tools correct errors in draft assemblies using the original sequencing reads. For short-read polishing, tools like Pilon and Polypolish align reads back to the assembly to identify and correct errors [21, 83]. Long-read polishing tools such as Medaka and Racon use different algorithms optimized for the error profiles of long-read technologies [84]. These polishing steps are particularly important for long-read assemblies, which typically have higher error rates in their initial drafts.

Error correction tools originally developed for isolate genomes can also be adapted for metagenomic assemblies. Canu includes sophisticated error correction algorithms that can be applied to long reads before assembly or to correct existing assemblies [51].

Consensus approaches generate improved assemblies by combining multiple independent assemblies of the same data. Trycycler creates a consensus assembly from multiple assemblies of the same genome, reconciling differences and producing a more accurate final result [85]. This approach is especially useful for circular bacterial genomes but can be adapted for metagenomic bins. Similarly, Autocycler automates this process, making it more accessible for high-throughput applications [86].

These consensus approaches are particularly valuable for complex communities where different assembly algorithms may perform better on different subsets of the data. The combination of polishing, error correction, and consensus generation can significantly improve the quality of metagenomic assemblies, resulting in more accurate gene predictions and functional annotations in downstream analyses. This is especially important for recovering high-quality MAGs that can serve as reference genomes for previously uncultured organisms.

Functional Annotation Methods

Through metagenome assembly and binning, one can obtain a set of MAGs, each representing a distinct genome through a set of long nucleotide sequences. In order

to assess the metabolic potential of communities, one has to first predict the genes from those genomes through functional annotation. People have developed a variety of annotation tools for this task; here, we will discuss two popular ones: Prokka and Bakta.

Prokka first predicts the coding sequences (CDSs) using another tool called Prodigal, then searches those against pre-built protein databases using HMMs with HMMER and the remaining unassigned using BLAST. Additionally, it identifies non-coding sequences like rRNA, tRNA, and more using Barrnap and Aragorn. The functional annotations are then created by getting the gene names from the best-matching hits.

Prokka had some drawbacks, like its dependency on pre-built databases that may have poor performance for unknown MAGs. Additionally, it may have issues with small proteins and cannot always assign database cross-references correctly. Bakta was developed to address those limitations. It uses a larger taxon-independent reference database that is searched using fast hashing. As a fallback, it uses DIAMOND instead of BLAST. Small proteins are predicted by a special small open reading frame detection algorithm, and it additionally annotates CRISPR arrays in the genomes. As Bakta uses a database that contains way more sequences than Prokka's database, it can provide more and better annotations than Prokka, reducing the number of proteins marked as hypothetical proteins [87]. While still reasonably fast, its runtime is slightly slower and it requires more computational resources; however, this trade-off is usually worth it considering the better annotations. An example of a circular genome annotation generated by Bakta for an *Acetobacter* MAG is shown in Figure 2.2. The fact that we can annotate so much of the chromosome highlights how effective long-read data is for automated bacterial genome annotation.

Current Frontiers and Challenges in Computational Metagenomics

Despite significant progress in computational metagenomics, several challenges remain: Reference databases are continuing to grow in size, and reproducible and reliable metagenomics workflows are still challenging to implement, especially for beginners, as standard procedures are not yet well established and there is no one tool that can do everything and answer all research questions. Computational resources needed for metagenome analysis can be high and require bioinformatics expertise, limiting accessibility for non-computational researchers. This highlights the need for continued advancement of metagenomics tools to enable better and more accessible analyses of microbial communities in the future.

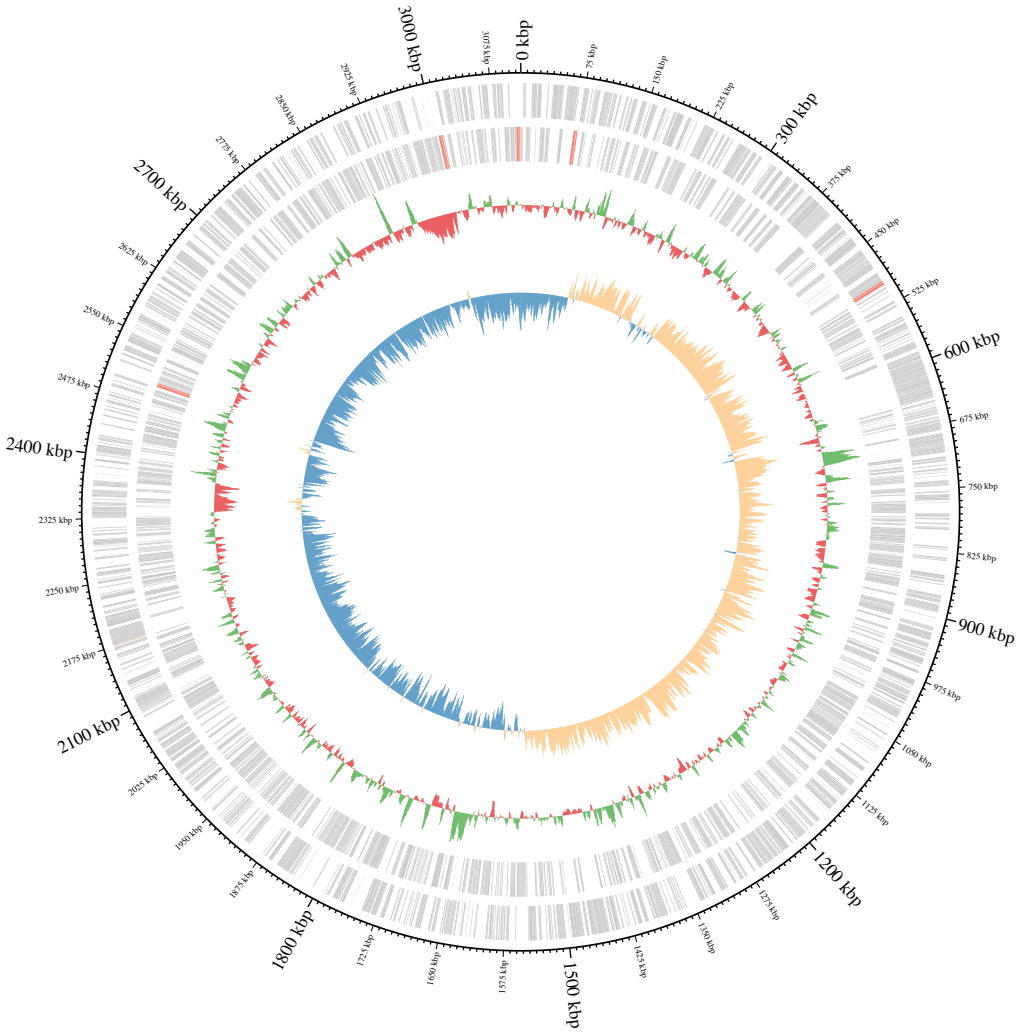

Figure 2.2: Circular genome visualization of a metagenome-assembled genome (MAG) obtained from a bioreactor sequenced with long-read technology, annotated using Bakta. The figure displays coding sequences, GC content, GC skew, and gene predictions around the bacterial genome.

Chapter 3

Metagenomics analyses of n-caprylate and n-caproate producing microbiomes

This chapter is based on a manuscript written by multiple authors. Author contributions are detailed in the table below.

Title of paper:	Variability in n-caprylate and n-caproate producing microbiomes in reactors with in-line product extraction
Status in publication process:	Published in mSystems (DOI: https://doi.org/10.1128/msystems.00416-24)

Author	Author position	Scientific ideas (%)	Data generation (%)	Analysis & interpretation (%)	Paper writing & editing (%)
Catherine M. Spirito	First/Co-first	25	45	20	35
Timo N. Lucas	Co-first	25	25	40	35
Sascha Patz	Third	6	0	5	0
Byoung Seung Jeon	Fourth	6	0	5	2.5
Jeffrey J. Werner	Fifth	6	10	5	5
Lauren H. Trondsen	Sixth	6	10	5	2.5
Juan J. Guzman	Seventh	6	10	5	2.5
Daniel H. Huson	Eighth	10	0	5	5
Largus T. Angenent	Corresponding	10	0	10	12.5

Catherine M. Spirito[1,2], Timo N. Lucas[3], Sascha Patz[3], Byoung Seung Jeon[4], Jeffrey J. Werner[5], Lauren H. Trondsen[1], Juan J. Guzman[1], Daniel H. Huson[3], Largus T. Angenent[1,4,6,7,8]

Catherine M. Spirito and Timo N. Lucas contributed equally to this work.

[1]Department of Biological and Environmental Engineering, Cornell University, Riley-Robb Hall, Ithaca, New York, USA
[2]Office of Undergraduate Research, University of Maryland, College Park, Maryland, USA
[3]Institute for Bioinformatics and Medical Informatics, University of Tübingen, Tübingen, Germany
[4]Department of Geosciences, University of Tübingen, Tübingen, Germany
[5]Chemistry Department, SUNY-Cortland, Bowers Hall, Cortland, New York, USA
[6]AG Angenent, Max Planck Institute for Biology Tübingen, Tübingen, Germany
[7]Department of Biological and Chemical Engineering, Aarhus University, Aarhus, Denmark
[8]The Novo Nordisk Foundation CO2 Research Center (CORC), Aarhus University, Aarhus, Denmark

Abstract

Medium-chain carboxylates (MCCs) are used in various industrial applications. These chemicals are typically extracted from palm oil, which is deemed not sustainable. Recent research has focused on microbial chain elongation using reactors to produce MCCs, such as n-caproate (C6) and n-caprylate (C8), from organic substrates such as wastes. Even though the production of n-caproate is relatively well-characterized, bacteria and metabolic pathways that are responsible for n-caprylate production are not. Here, three 5 L reactors with continuous membrane-based liquid-liquid extraction (i.e., pertraction) were fed ethanol and acetate and operated for an operating period of 234 days with different operating conditions. Metagenomic and metaproteomic analyses were employed. *n*-Caprylate production rates and reactor microbiomes differed between reactors even when operated similarly due to differences in H_2 and O_2 between the reactors. The complete reverse β-oxidation (RBOX) pathway was present and expressed by several bacterial species in the *Clostridia* class. Several *Oscillibacter* spp., including *Oscillibacter valericigenes*, were positively correlated with *n*-caprylate production rates, while *Clostridium kluyveri* was positively correlated with *n*-caproate production. *Pseudoclavibacter caeni*, which is a strictly aerobic bacterium, was abundant across all the operating periods, regardless of *n*-caprylate production

rates. This study provides insight into microbiota that are associated with *n*-caprylate production in open-culture reactors and provides ideas for further work.

Importance

Microbial chain elongation pathways in open-culture biotechnology systems can be utilized to convert organic waste and industrial side streams into valuable industrial chemicals.

Here, we investigated the microbiota and metabolic pathways that produce medium-chain carboxylates (MCCs), including *n*-caproate (C6) and *n*-caprylate (C8), in reactors with in-line product extraction. Although the reactors in this study were operated similarly, different microbial communities dominated and were responsible for chain elongation.

We found that different microbiota were responsible for *n*-caproate or *n*-caprylate production, and this can inform engineers on how to operate the systems better. We also observed which changes in operating conditions steered the production toward and away from *n*-caprylate, but more work is necessary to ascertain a mechanistic understanding that could be predictive. This study provides pertinent research questions for future work.

Introduction

Medium-chain carboxylates (MCCs), such as *n*-caproate and *n*-caprylate, are utilized in a variety of industrial and agricultural applications, including as biofuel precursors, anti-corrosion agents, plasticizers, personal care products, feed additives, and antimicrobials [88]. MCCs are typically produced as a byproduct of palm oil refining [89].

Recent research has focused on producing MCCs in open-culture reactors from organic substrates, including wastes, as part of a circular economy. MCCs have a relatively low solubility in water in their undissociated form. Therefore, MCCs can be extracted from aqueous broths via techniques, such as in-line product extraction, to address MCC toxicity issues and to increase volumetric production rates [90, 91].

Laboratory studies demonstrated the efficient production of MCCs by anaerobic fermenter microbiomes at rates comparable to methane production by anaerobic digester microbiomes [91, 92]. MCCs are produced via pure and open cultures from a variety of substrates, including synthetic substrates [88] utilizing ethanol [90, 93, 94] or lactic acid [95, 96] as the electron donor [89], organic wastes, and industrial side streams [90, 97, 98, 99, 100, 101, 102, 103].

MCCs are often produced via the reverse β-oxidation (RBOX) pathway in which ethanol, lactic acid, or another electron donor is oxidized to acetyl-CoA. Short-chain carboxylates, such as acetate and *n*-butyrate, are then chain elongated to longer-chain carboxylates, such as *n*-caproate (six-carbon chain) and *n*-caprylate (eight-carbon chain) [88, 104, 105, 106] (Figure 3.1). Chain elongation is a cyclic process in which acetyl-CoA enters the cycle and is condensed with an acyl-CoA by acetyl-CoA C-acyltransferase (Thiolase II) (ACAT) to form an acyl-CoA that is two carbon atoms longer than its substrate. The product of this reaction is further reduced by 3-hydroxy-acyl-CoA dehydrogenase (HAD) or 3-hydroxy-butyryl-CoA dehydrogenase (HBD) (Figure 3.1). This product is then dehydrated to 2-enoyl-CoA by enoyl-CoA dehydratase (ECH) and further reduced by an acyl-CoA dehydrogenase (ACD) or butyryl-CoA dehydrogenase to form an elongated acyl-CoA (Figure 3.1). Finally, terminal enzymes such as acetyl-CoA transferase (CoAT) or thioesterase (TE) act to remove coenzyme A from the terminal acyl-CoA molecule and release the corresponding acid (Figure 3.1). Energy is conserved during the RBOX pathway via flavin-based electron bifurcation and the Rnf respiratory complex (RNF) in which an electron-bifurcating acyl-CoA dehydrogenase complex utilizes two electron-transfer flavoproteins (Figure 3.1) [107].

Prior research primarily focused on RBOX as the pathway for MCC production [91, 94, 108, 109, 110]. Recent studies have suggested that the fatty acid biosynthesis (FAB) pathway may also play a role [111, 112]. However, it should be noted that all bacteria use the anabolic FAB pathway to build their phospholipid membranes. In addition, FAB is an anabolic process, it consumes energy rather than producing energy that is necessary for bacterial growth in anaerobic conditions without inorganic electron acceptors for respiration.

Previously, both pure- and open-culture studies identified multiple bacterial strains that produce *n*-caproate. Known *n*-caproate-producing bacteria primarily belong to the phylum Firmicutes, except for *Rhodospirillum rubrum* [113]. Within the phylum Firmicutes, *n*-caproate-producing bacterial strains have been isolated and identified from six genera: *Caproicibacter*, *Caproiciproducens*, *Clostridium* [114, 115, 116], *Eubacterium* [117, 118], *Megasphaera* [119], and *Pseudoramibacter* [109]. In open-culture reactor studies, certain bacteria have been associated with high *n*-caprylate production rates, including *Burkholderia* spp., *Clostridium* group IV spp., *Desulfosporosinus meridiei*, *Oscillospira* spp., Rhodocyclaceae K82 spp., unknown Ruminococcaceae, and *Sphingobacterium multiform* [90, 91, 98]. These studies were based on 16S rRNA gene sequencing data.

Few bacterial isolates have been shown to produce *n*-caprylate. This is attributed to the microbial toxicity of *n*-caprylate and the lack of measurement of *n*-caprylate in prior studies. In 1967, *Ramibacterium alactolyticum*, renamed *Pseudoramibacter alactolyti-*

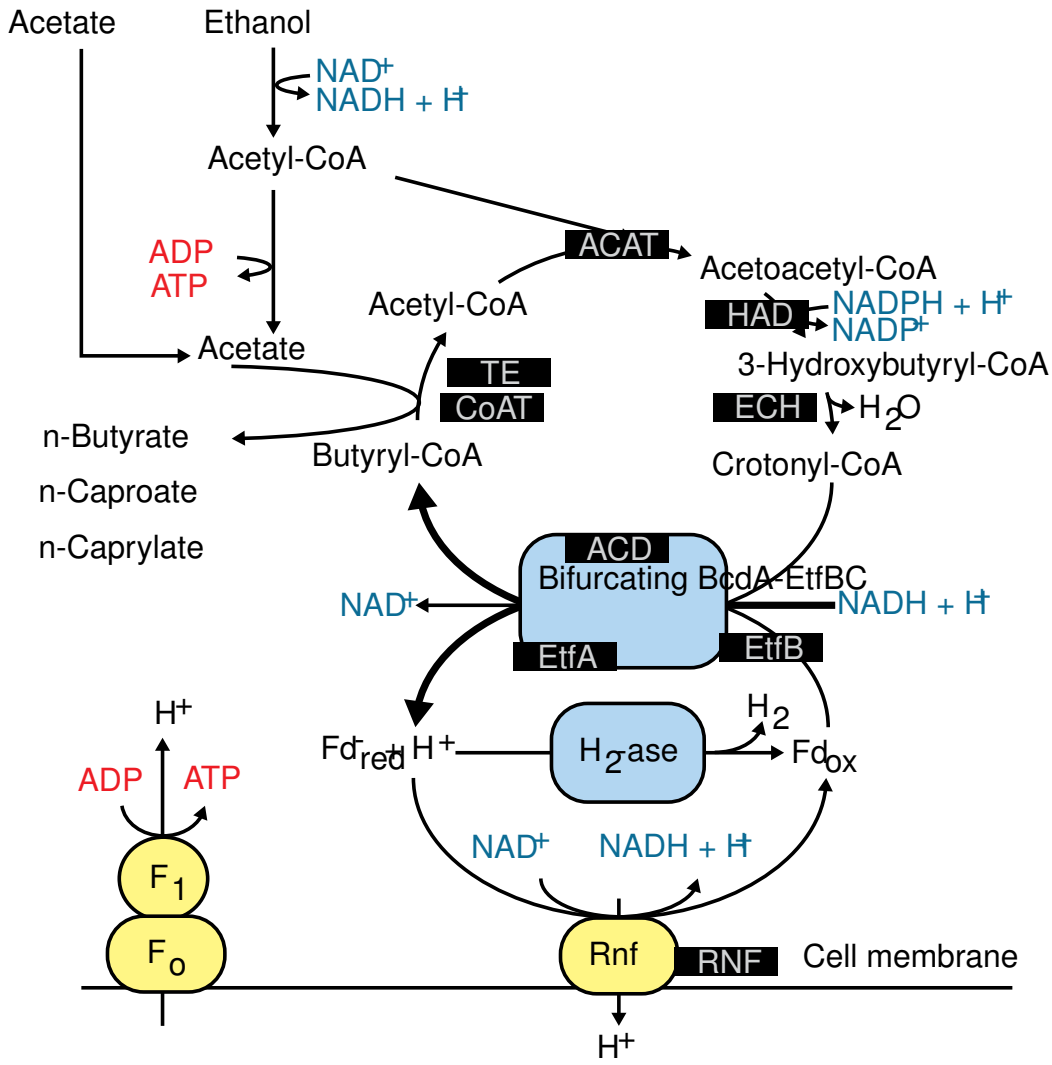

Figure 3.1: The RBOX pathway investigated in this study. The enzymes we examined in this study are highlighted in black boxes. The figure was modified with permission from Angenent et al. RBOX pathway enzymes are: ACAT: acetyl-CoA C-acyltransferase (Thiolase II); HAD: 3-hydroxy-acyl-CoA dehydrogenase; ECH: enoyl-CoA dehydratase; ACD: acyl-CoA dehydrogenase; EtfA/B: electron-transfer-flavoprotein subunit A/B; CoAT: acetyl CoA-transferase; TE: thioesterase; RNF: Rnf respiratory complex.

cus, was shown to produce *n*-caproate and *n*-caprylate from glucose [120]. A previous pure-culture reactor study observed the production of relatively low concentrations of *n*-caprylate by *Clostridium kluyveri*, which is a well-known chain-elongating bacterium, in a reactor fed a 10:1 molar ratio of ethanol and acetate (i.e., syngas effluent), operated at pH 7 and with an in-line membrane-based liquid/liquid extraction (i.e., pertraction) system to reduce the toxicity [97]. At lower pH levels, a lower rate of *n*-caprylate production was observed. For the open-culture studies, a shotgun metagenomics study found an uncultured Clostridiales order bacterium, *Candidatus Weimeria bifida*, gen. nov., sp. nov., which could produce *n*-caprylate from xylose [108, 109]. Research is needed to understand essential players and metabolic pathways in these reactors that optimize *n*-caprylate production.

Here, we investigated the role of the RBOX pathway in producing *n*-caproate and *n*-caprylate. Our original objective was to build and operate three equal stainless steel reactor systems to prevent O_2 intrusion as much as possible. As an independent operating unit, we planned different H_2 concentrations for each system. However, we were unsuccessful in: (i) preventing O_2 intrusion; and (ii) utilizing H_2 as an independent parameter. Regardless, we obtained pertinent data by operating the three 5 L open-culture reactors with in-line product extraction throughout 234 days. We employed Illumina 16S rRNA gene sequencing, shotgun metagenomics, and metaproteomics to characterize microbiomes. The reactors were fed ethanol and acetate and produced *n*-caproate and *n*-caprylate. Even though the reactors were all provided the same substrates and produced MCCs, their microbiota differed. Several bacterial species belonging to the class Clostridia, including *Oscillibacter valericigenes*, expressed the majority of the RBOX pathway. *Oscillibacter* spp. members were found to be positively correlated with *n*-caprylate production rates. The aerobic bacterium *Pseudoclavibacter caeni* was one of the abundant bacteria in the reactor samples, regardless of *n*-caprylate production rates. *P. caeni* may have acted as an O_2 scavenger in the system or provided other unknown roles for producing *n*-caprylate.

Results

The performance of the three reactors differed despite similar operating conditions

We operated three stainless-steel, continuously stirred reactors with a 5 L working volume and in-line product extraction at mesophilic conditions and a pH of 5.5 (Table 3.1; The effluent carboxylate and ethanol concentrations during the 75-day reactor startup period can be found in Fig. S2. After the 75-day start-up period and at the start of

Period I, we mixed the microbiota from all three reactors and then operated the three reactors similarly throughout Period I (without sparging). Regardless, the performance of the three reactors was not similar and varied during this period.

Reactors 1 and 2 achieved promising and similar overall medium-chain production rates (Figure 3.2), but Reactor 3 performed poorly during Period 1 with a lower *n*-caprylate production rate compared to Reactors 1 and 2 (Figure 3.2). Reactor 1 exhibited a higher selectivity (i.e., desired products compared to the substrate) for *n*-caprylate production compared to Reactor 2.

The maximum average volumetric *n*-caprylate production rate was $1.1 \times 10^2 \pm 7.1$ mmol C L^{-1} d^{-1} (0.080 ± 0.005 g L^{-1} h^{-1}) during Period 1D for Reactor 1 (3.2B). Small differences in operating conditions, potentially due to differences in the tightness of the reactor seals resulting in different H_2 and O_2 exchange conditions, seem to have had an amplified impact on chain elongation. This was different from the anaerobic digestion of animal waste, for which we found that four similar reactor operating conditions resulted in almost identical performances after an operating period of 1 year [121].

Table 3.1: Operating conditions of reactors R1-R3 across different periods. HRT: hydraulic retention time; OLR: organic loading rate.

Reactor	Period	Days	HRT (d)	OLR (mmol C L^{-1} d^{-1})	Gas flow rate (L d^{-1})	Sparged with
	1	75–142	8.9 ± 0.3	$1.4 \times 10^2 \pm 7.9$	0	No gas
R1	2	143–184	9.6 ± 0.3	$1.2 \times 10^2 \pm 6.7$	3.38 ± 4.86	N_2 off/on
	3	185–234	9.0 ± 0.3	$1.4 \times 10^2 \pm 6.5$	24.7 ± 13.7	N_2
	1	75–142	8.9 ± 0.5	$1.4 \times 10^2 \pm 8.4$	0	No gas
R2	2	143–184	9.0 ± 0.3	$1.3 \times 10^2 \pm 8.8$	2.45 ± 3.53	N_2, H_2 off/on
	3	185–234	8.9 ± 0.3	$1.4 \times 10^2 \pm 8.2$	6.14 ± 7.00	N_2, H_2
	1	75–142	7.8 ± 0.3	$1.6 \times 10^2 \pm 7.6$	0.50 ± 0.31	No gas
R3	2	143–184	8.4 ± 0.5	$1.3 \times 10^2 \pm 13$	5.02 ± 5.64	N_2 off/on
	3	185–234	7.0 ± 0.6	$1.8 \times 10^2 \pm 15$	10.2 ± 4.83	N_2

Our results show that the H_2 partial pressure is a sensitive parameter to the *n*-caprylate performance, amplifying minor differences in operating conditions. During Period 1, gas in the headspace of Reactor 3 contained $31\% \pm 9.6\%$ H_2 (by volume), whereas H_2 was $9.9\% \pm 5.2\%$ and $1.8\% \pm 1.9\%$ of total gas for Reactors 1 and 2, respectively (Table S1). The reducing equivalents Fd_{red} and NADH produced by the RBOX pathway can reduce the H^+ produced by the pathway to H_2 (Figure 3.1). The reactor tightness and material diffusiveness may influence the H_2 partial pressures because H_2, as the smallest molecule, may easily diffuse out of the system, while other gases would not. We built almost the entire reactor setup out of stainless steel to minimize H_2 diffusion

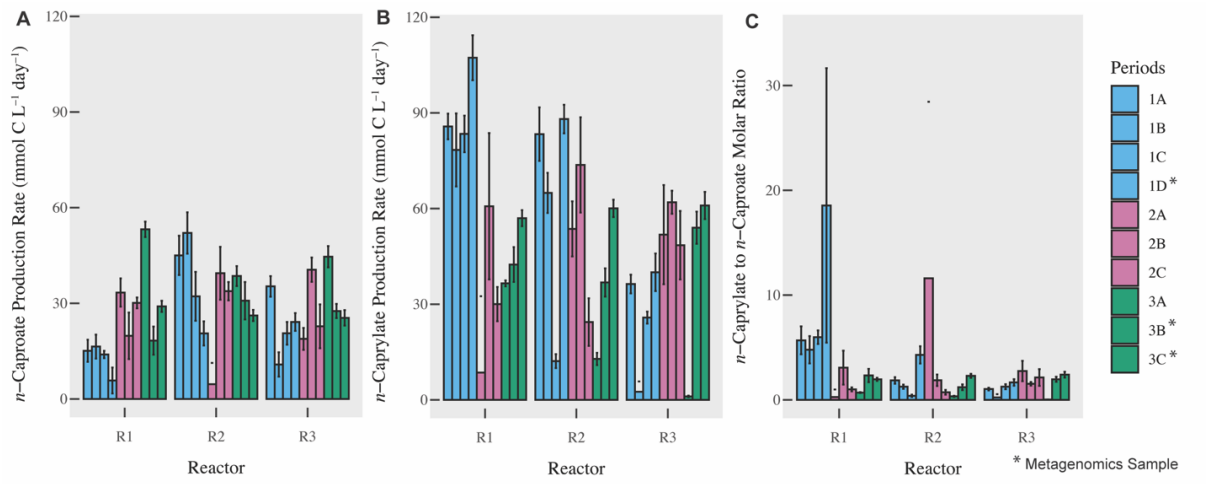

Figure 3.2: Fig 2 *n*-Caproate (A) and *n*-caprylate (B) total production rates (mmol C L^{-1} d^{-1}) and the molar ratio of *n*-caprylate to *n*-caproate (C) in three reactors across three main periods (and 10 periods). Period divisions are explained in the methods. Error bars indicate the standard error for the measurements. *Legend indicates periods (Periods 1D, 3B, and 3C) in which biomass samples were collected from reactors for shotgun metagenomic analysis. R1-3 are Reactors 1-3.

through plastic tubing and connections. However, our results show that we could not prevent H_2 diffusion out of the system, which included a gas recirculation pump and some tubing lines not made of stainless steel. We note that we did not directly measure the reactor tightness and material diffusivity in this study.

The H_2 partial pressure can negatively affect chain elongation by reducing chain-elongating rates and changing the product spectra [106, 122]. High H_2 partial pressures can lower the ethanol oxidation rate to acetate and prevent the Rnf complex from functioning properly within the RBOX pathway, resulting in lower ATP production through substrate-level phosphorylation and membrane-based phosphorylation, respectively [106]. This would lower the growth rate, further slowing the development of an active microbiota. With the relatively high H_2 partial pressures for Reactor 3 during Period 1 compared to Reactors 1 and 2, a significant fraction of ethanol that we fed to Reactor 3 was not converted and left in the effluent, which resulted in a higher average effluent ethanol concentration for Reactor 3 compared to Reactors 1 and 2 ($1.7 \times 10^2 \pm 9.7$ mM vs 47 ± 3.9 mM and 29 ± 4.3 mM, respectively) (Fig. S2; Table S2). Because our reactors were continuously stirred systems, concentrations measured in the effluent were approximately equal to what the reactor microbiomes observed.

To test whether a lower H_2 partial pressure would improve *n*-caprylate production rates, we sparged N_2 gas into Reactor 3 to reduce the percentage of H_2 in the headspace (Table 3.1; Table S1). The sparging decreased the H_2 in the headspace from 31% ± 9.6% (by volume) during Period 1 to 20% ± 14% during Period 2 to 7.3% ± 4.6%

during Period 3 (Table S1), resulting in increased volumetric *n*-caprylate production rates for Reactor 3 during Periods 2 and 3 (Figure 3.3B). Into Reactor 2, we sparged N_2 and H_2 gas. As expected, when H_2 partial pressures increased during Periods 2 and 3 (Table S1), *n*-caprylate productivity decreased for Reactor 2 (Figure 3.3B). However, we observed that the effect of H_2 on *n*-caprylate production was not uniform in all reactors. When the amount of H_2 in the headspace decreased due to N_2 sparging into Reactor 1, we observed decreased *n*-caprylate production rates during Periods 2 and 3 (Figure 3.3B; Table S1). However, sparging with N_2 to remove H_2 may have also removed O_2, which could have an unknown effect. Gas sparging itself was another introduced variable in the experiment that may have decreased biomass growth and *n*-caprylate production for Reactor 1 during Periods 2 and 3. We also noted differences in the acetate, *n*-butyrate, *n*-caproate, and *n*-caprylate concentrations in the effluent of our reactors (Table S2; Fig. S2A through C). Thus, our system was not predictive because we did not fully understand how the environmental conditions in the reactor affect the microbial pathways in the complex microbiota.

Bacterial species abundance correlated with *n*-caprylate production rates

We analyzed the reactor microbiome via 16S rRNA gene sequencing and shotgun metagenomics. Overall, we observed similar trends in the dominance of certain bacterial species during high and low *n*-caprylate production periods in both data sets. We noticed some differences between the sequencing methods, which we attributed to differences in how the data were analyzed and how taxonomy was assigned (see Materials and Methods).

The 16S rRNA gene sequencing data set was derived from approximately weekly biomass samples, which we collected from the reactors throughout the operating period. The shotgun metagenomics data set was smaller and was derived from nine biomass samples, which we collected from the three reactors at three-time points during the operating period (during Periods 1D, 3B, and 3C, as indicated in Figure 3.3). The shotgun metagenomic data set resulted in 477,902,544 reads. We assembled 32 draft genomes from this data, 25 of which were high-quality (>90% completion, <5% contamination) [123], as detailed in Table 3.2. For metagenome-assembled genomes (MAGs) with high contamination, we observed multiple instances of single-copy genes, likely from the same or closely related strains, indicating contamination.

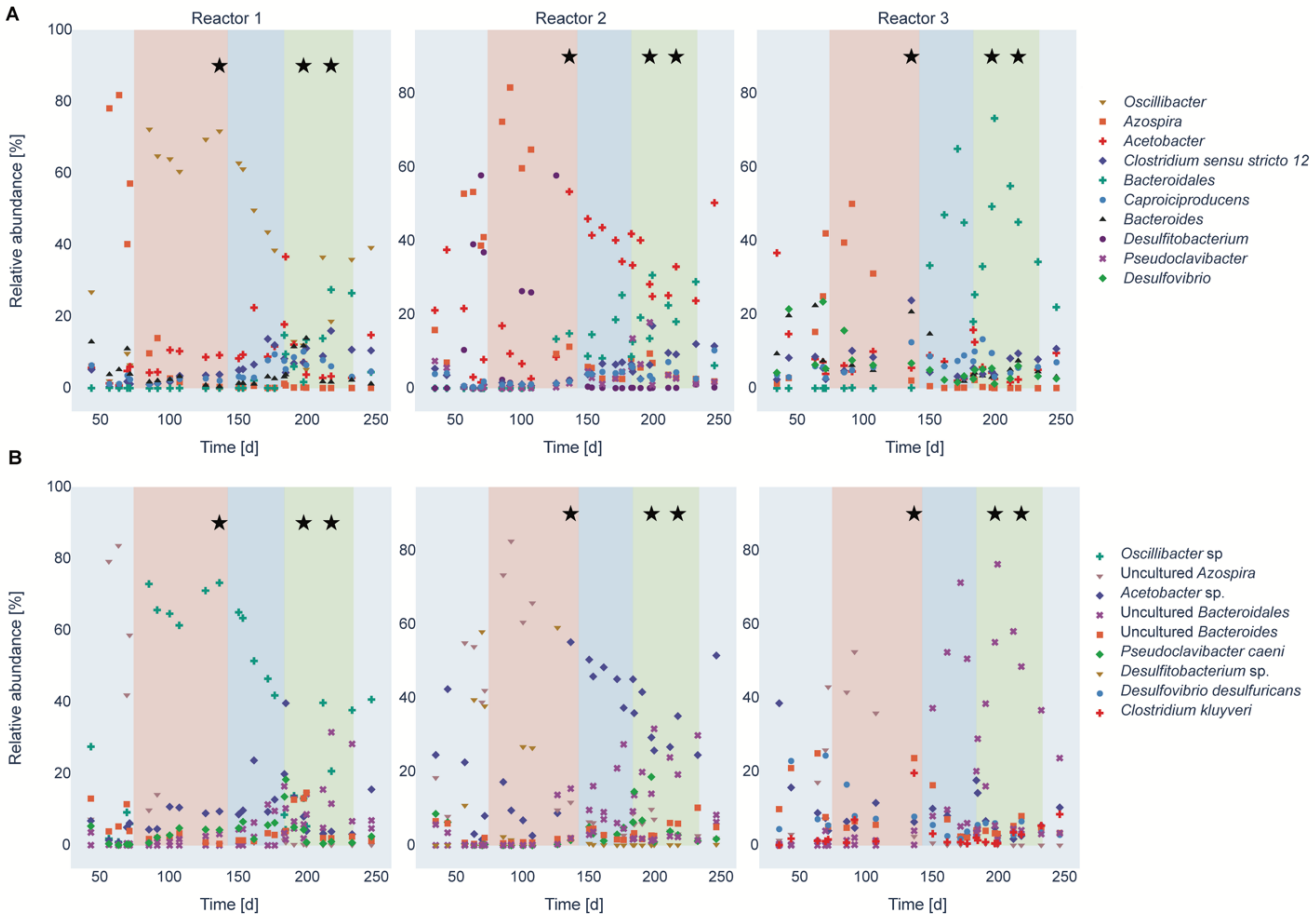

Figure 3.3: Relative abundance of the top seven most dominant taxa of each reactor based on the Illumina 16S rRNA gene sequencing results on the genus level (A) and the species level (B) throughout the operating time. The first 75 days of the operating period were the startup period (light blue). The salmon, blue, and green shadings indicate Periods 1, 2, and 3, respectively; the stars indicate the metagenomic sampling time points.

Table 3.2: MAG identifiers, GTDB-tk taxonomic assignments, and corresponding NCBI names. Species names are shown, if unavailable, the genus name is displayed. Completeness (Comp) and contamination (Cont) were computed using CheckM [81]. The final column indicates the reactor and sampling period. MAGs are ordered by their percentage of completeness.

ID	GTDB	NCBI	Comp (%)	Cont (%)	Reactor/ period
1	*Acetobacter* sp012517935	*Acetobacter* sp.	100	0.75	R3/1D
2	*Methanobacterium_C*	*Methanobacterium*	100	0.8	R1/3C
3	JAAYAE01	Acholeplasmataceae bacterium	99.8	0.9	R2/3B
4	*Intestinimonas*	*Intestinimonas*	98.99	0	R3/3C
5	*Desulfitobacterium*	*Desulfitobacterium*	98.85	2.87	R2/1D
6	*Pseudoclavibacter caeni*	*Pseudoclavibacter caeni*	98.84	0.58	R2/1D
7	*Cellulomonas*	*Cellulomonas*	98.65	0.72	R3/3C
8	JAAYUD01 sp012517855	*Bacteroidales bacterium*	98.49	0	R3/3C
9	JAEWCM01	Bacillota bacterium	98.47	1.32	R1/3B
10	*Bacteroides*	*Bacteroides*	98.12	3.35	R1/3B
11	*Bacteroides* sp900766195	Uncultured *Bacteroides* sp.	98.12	2.79	R3/1D
12	*Clostridium* AM	*Clostridium drakei*	98.09	4.16	R2/3B
13	*Fimivivens*	*Bacillota bacterium*	97.93	0.71	R3/3C
14	*Prevotella*	*Prevotella*	97.80	0.79	R3/3C
15	JAAYSF01	Veillonellaceae bacterium	97.12	4.86	R3/1D
16	*Desulfovibrio legallii*	*Desulfovibrio*	97.04	0	R3/3C
17	*Caproicibacterium* sp002411615	Ruminococcaceae bacterium UBA5397	96.17	0.36	R3/1D
18	*Latilactobacillus fuchuensis*	*Latilactobacillus fuchuensis* DSM 14340 = JCM 11249	95.55	0	R1/3C
19	*Levilactobacillus brevis*	*Levilactobacillus brevis* ATCC 14869 = DSM 20054	94.84	0	R1/3C
20	*Clostridium* B	*Clostridium kluyveri* DSM 555	94.72	0.69	R3/1D
21	*Oscillibacter*	Clostridia bacterium	94.15	2.85	R1/3B
22	*Vescimonas*	Clostridiales bacterium	93.62	2.01	R2/1D

Table 3.2 – continued from previous page

ID	GTDB	NCBI	Comp (%)	Cont (%)	Reactor/ period
23	Oscillospiraceae UBA2922	Clostridiales bacterium UBA2922	92.84	3.36	R3/1D
24	*Pauljensenia*	Actinomyces	92.04	3.79	R3/1D
25	*Clostridium AV fermenticellae*	*Clostridium fermenticellae*	90.01	15.84	R1/1D
26	*Acetobacter fabarum*	*Acetobacter fabarum*	89.38	1.74	R3/1D
27	*Azospira*	*Azospira*	89.34	0.36	R2/1D
28	Clostridiaceae	Clostridiaceae	85.88	0.84	R3/3C
29	*Pygmaiobacter*	Bacillota bacterium	83.09	1.38	R1/1D
30	*Azospira inquinata*	*Azospira inquinata*	76.76	0.84	R2/3B
31	*Onthomonas*	*Clostridiales bacterium*	75.88	2.35	R2/3C
32	*Bulleidia*	*Solobacterium*	75.45	2.22	R1/3B

Two aerobic bacteria, *P. caeni,* and an unknown *Acetobacter* sp., were present in some reactors at relatively high abundances. For Reactors 1 and 2, *P. caeni* was an abundant bacterium. Still, its abundance did not correlate to *n*-caprylate production rates (r = 0.01, P = 0.98; Figure 3.4). For Reactor 2, *Acetobacter* sp. bacteria were dominant during Periods 1 and 2 and declined in abundance during Period 3 when *n*-caprylate production rates decreased (Figure 3.3 and 3.4). The presence of these bacteria shows that O_2 was introduced into the reactors due to an unknown location in the reactor setup.

Certain bacterial species dominated the reactors during periods of relatively low *n*-caprylate production but higher *n*-caproate production rates (Period 3B for Reactors 1 and 2 and Period 1D for Reactor 3; Figure 3.2, Figure 3.3, Figure 3.4). Based on the shotgun metagenomics data, the abundance of *C. kluyveri* was negatively correlated to *n*-caprylate production rates. However, the correlation was not significant (Figure 3.4; r = -0.49, P = 0.18). No correlation was observed between *C. kluyveri* relative abundance and production rates in the 16S rRNA gene sequencing data (Figure 3.3B). For Reactor 1 during Period 3B, Bacteroidaceae bacterium HV4-6-C5C (217,314 aligned reads) and *P. caeni* (210,785 aligned reads) dominated the reactor. For Reactor 2 during Period 3B, *P. caeni* (255,949 aligned reads) and, to a lesser extent, *C. kluyveri* (43,228 aligned reads) dominated the reactor. For Reactor 3 during Period 1D, *C. kluyveri* (180,084

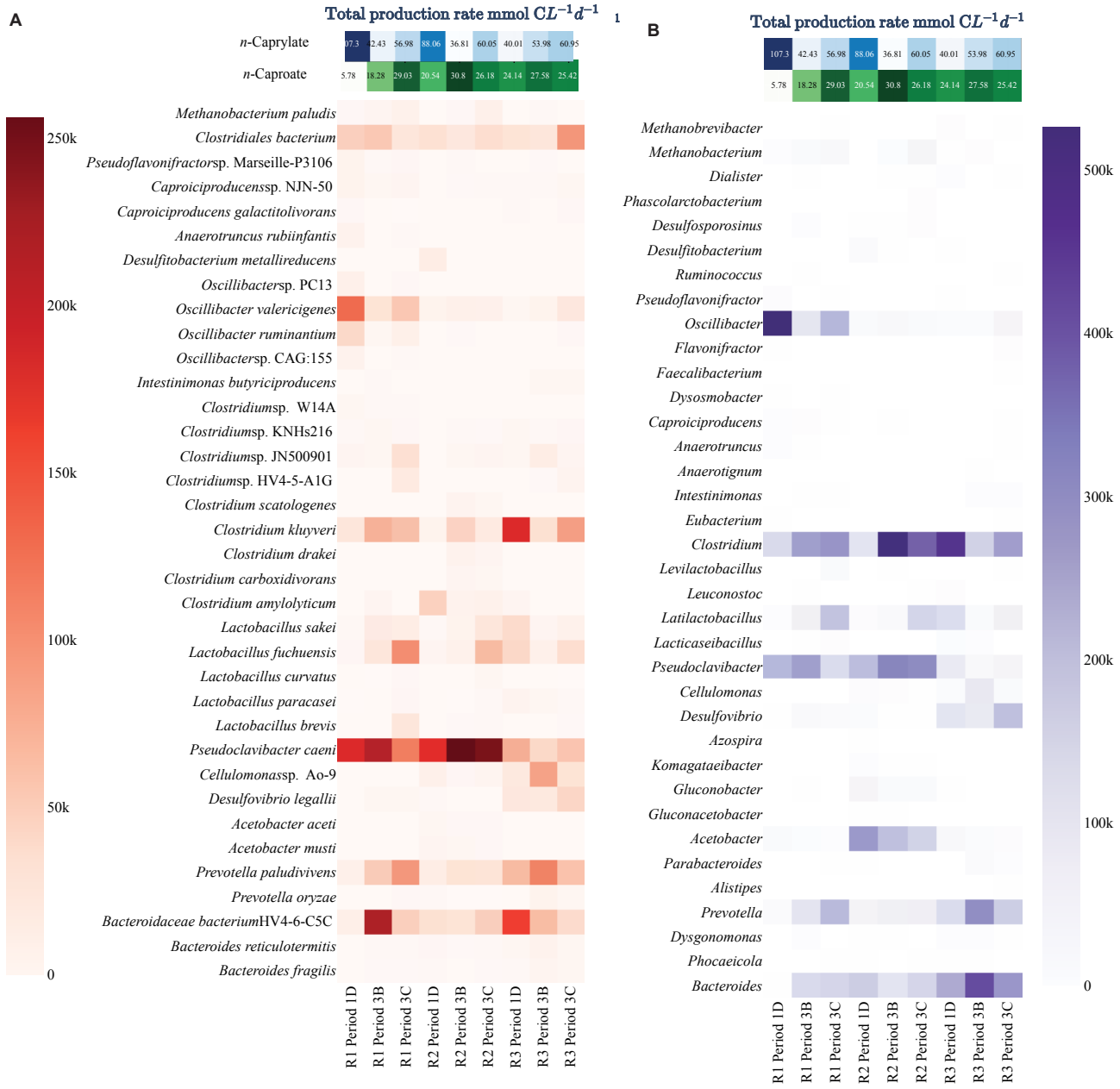

Figure 3.4: The most abundant species (A) and genus (B) in the three bioreactors based on shotgun metagenome analysis. After normalizing the read count for sample size, the heatmaps show the number of reads aligned to each taxon across sampling points. Only taxa with more than 12k reads are displayed. The top of the plots shows the n-caprylate (blue) and n-caproate (green) production rates.

reads) and Bacteroidaceae bacterium HV4-6-C5C (157,401 reads) dominated the reactor (Figure 3.4). Across all reactors, the abundance of Bacteroidaceae bacterium HV4-6-C5C was negatively correlated to *n*-caprylate production rates, though the correlation was not significant (r = -0.54, P = 0.133; Figure 3.4). Based on the 16S rRNA gene sequencing data, an *Azospira* sp. was dominant in all the reactors prior to and at the start of Period 1 and was positively correlated to *n*-caproate production rates (r = 0.41, P = 0.00034; Figure 3.3).

Bacteria with the RBOX pathway

We investigated which metagenomes in our reactors had a complete or nearly complete RBOX pathway (Fig. 5). Specifically, we looked for nine enzymes involved in the RBOX pathway in our metagenomics and proteomics data: acetate CoA-transferase (CoAT), 3-hydroxy-acyl-CoA dehydrogenase (HAD), enoyl-CoA dehydratase (ECH), acyl-CoA dehydrogenase (ACD), electron-transfer-flavoprotein subunit A/B (EtfA/B), acetyl-coenzyme A acetyltransferase (ACAT), thioesterase (TE), and Rnf respiratory complex (RNF) (Figure 3.1).

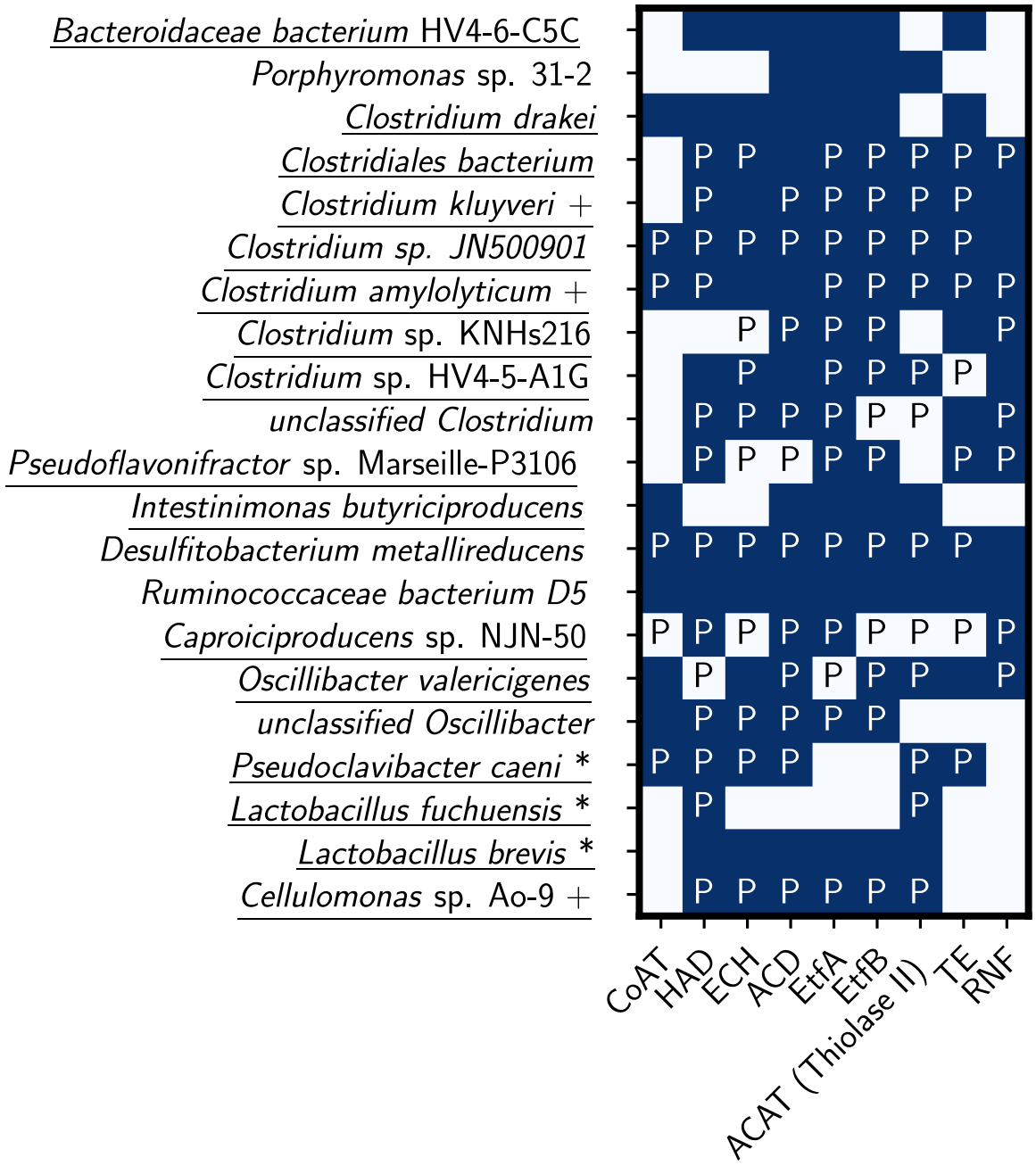

Figure 3.5: Absence or presence of nine enzymes involved in the RBOX pathway (acronyms described previously) in reactor *de-novo* assembled metagenomes and proteomes as monitored by shotgun metagenomics analysis and proteomics. MAG taxonomy was assigned using GTDB-Tk (r220) [124]. Enzyme acronyms were described in Figure 3.1. A blue box denotes the presence in the metagenome, the letter P the presence in the metaproteome, and a white box without a P the absence in both the metagenome and metaproteome. MAGs identified to species level are depicted. * = >90% complete and <5% contaminated (determined with CheckM). + = >80% complete and <10% contamination. More abundant species above the 12k read threshold are underlined.

For both the 16S rRNA gene sequencing data and the shotgun metagenomics data, certain bacterial species within the genus *Oscillibacter* dominated when *n*-caprylate production rates were higher for Reactor 1 during Period 1 and decreased in abundance during later periods when production rates decreased (Figure 3.3B and Figure 3.4). The unknown *Oscillibacter* sp. bacteria that was dominant in the 16S rRNA gene sequencing data had a 95.9% ID to an *O. valericigenes* Sjm18-20 strain [125]. In the shotgun metagenomics analysis, *O. valericigenes* was one of the dominant bacteria in Reactor 1 during Period 1 (126,083 aligned reads, 3.4B). Based on the shotgun metagenomics analysis, *O. valericigenes* abundance positively correlated with *n*-caprylate production rates (Pearson correlation coefficient, r = 0.68, $P = 0.0439$). Several other *Oscillibacter* spp. were also positively correlated with *n*-caprylate production rates, *Oscillibacter* sp. CAG:155 (r = 0.71, $P = 0.0321$), *Oscillibacter ruminantium* (r = 0.65, $P = 0.058$), *Oscillibacter* sp. 1-3 (r = 0.72, $P = 0.0287$), *Oscillibacter* sp. NSJ-62 (r = 0.65, $P = 0.058$), and *Oscillibacter* sp. PC13 (r = 0.68, $P = 0.0439$) (Figure 3.4). Based on the 16S rRNA gene sequencing data, an *Oscillibacter* sp. OTU, an uncultured *Oscillibacter* OTU, and an *O. valericigenes* OTU were positively correlated with *n*-caprylate production rates (r = 0.38, 0.27, and 0.15, and $P = 0.001$, 0.021, 0.208, respectively; Figure 3.3).

Bacterial microcompartments present in reactor microbiomes

The metagenomic analysis revealed the presence of specific bacterial microcompartments. These specialized compartments encapsulate metabolic pathways, enhancing metabolic efficiency and specificity. It is known that the chain-elongator *C. kluyveri* and other Clostridia have microcompartments to protect their intracellular milieu against unstable or toxic chemical intermediates [126]. Notably, an ethanolamine utilizing microcompartment (EUT2B), two propanediol utilizing microcompartments (PDU1D and PDU1C), and a glycyl radical enzyme containing microcompartment (GRM1A) were the dominant microcompartments observed (Fig. S3). These microcompartments are metabolosomes that are usually expressed only when their substrate is present [127].

The ethanol-utilizing microcompartment (ETU) metabolizes ethanolamine, which is a product of the breakdown of phosphatidylethanolamine, to ethanol, acetyl-CoA, and acetyl-phosphate and protects the rest of the bacterial cell from the intermediate acetaldehyde [128]. We also observed the presence of ETU in all time points studied (Fig. S3). This bacterial microcompartment has only been reported in the bacterium *C. kluyveri* [129, 130]. We observed the ETU microcompartment as expected in *C. kluyveri*. We also observed the ETU microcompartment in other bacteria classified to the level *Clostridium* spp. or Clostridiaceae (Fig. S3), though we note that these could be *C. kluyveri* species that cannot be classified to the species level. We also observed

the ETU microcompartment in proteins that had no hit in the taxonomic database (NAs in Fig. S3). *O. valericigenes* Sjm18-20, which dominated Reactor 1 during periods of high *n*-caprylate production, was not found to have ETU microcompartments, though, it did contain EUT2B, PDU1D, and GRM1A, GRM3A, and GRM3C microcompartments.

Discussion

For our open-culture reactors, different microbial communities were correlated with periods of high *n*-caproate or *n*-caprylate production (Figure 3.4). The known chain elongator *C. kluyveri* and a primary fermenter Bacteroidaceae bacterium HV4-6-C5C had higher relative abundances during periods of high *n*-caproate production and decreased in abundance during periods of high *n*-caprylate production (Figures 3.3 and 3.4). *C. kluyveri* produces *n*-caproate from ethanol and short-chain carboxylates (acetate or *n*-butyrate), but there is limited evidence of its ability to produce *n*-caprylate [97]. In particular, the ethanol-utilizing microcompartment, ETU, associated with *C. kluyveri*, underscores its potential role in the efficient conversion of ethanol to acetaldehyde, which is a pivotal step in the production of *n*-caproate and *n*-caprylate. Different *Oscillibacter* species, which include *O. valericigenes* ($r = 0.68$, $P = 0.0439$), were positively correlated to periods of high *n*-caprylate output in the reactors (Figures 3.3 and 3.4). Indeed, *O. valericigenes* included microcompartments to possibly protect themselves during chain elongation. This finding is consistent with prior studies for which members of the Ruminococcaceae family (to which *Oscillibacter* belongs) were isolated from reactors producing *n*-caproate from lactate [91, 131] and Illumina 16S rRNA gene sequencing studies for which Ruminococcaceae members were associated with MCC production in reactors [94, 95, 98].

The unplanned presence of O_2 in our reactors created a niche for aerobic bacteria, such as *P. caeni* and *Acetobacter* species, to survive and become abundant in the reactors (Figure 3.4). The abundance of these aerobic bacteria was not correlated to *n*-caprylate production rates (Figures 3.3 and 3.4). As a result of our inability to build a reactor system that prevented O_2 inclusion, a major caveat existed in our quest to study different H_2 partial pressures on the RBOX. Using gas sparging to remove or add H_2 also removed O_2, which was a sensitive parameter. Even though we could not satisfy our experimental design with the independent parameter H_2, this study provides information on which to base future research, as discussed below.

Aerobic or facultative anaerobic microbes must have quickly consumed the O_2 in our reactors because strict anaerobic microbes, such as methanogens and other obligate anaerobes, were also present in our continuously stirred reactor systems

(Figure 3.4). Prior studies observed aerobes, such as *Acetobacter* [90, 95], and facultative anaerobes, such as *Lactobacillus* [95, 132], in chain elongation reactors. Previous studies from our lab had not found *P. caeni* in similar chain-elongating reactors [90, 91], though the aerobe *Acetobacter* was observed [90]. *P. caeni* was isolated from sewage sludge in 2012 [133], but the *P. caeni* assembly was only added to the NCBI nr database in 2019 (ASM883112v1). *P. caeni* could have been present in previous reactor studies but not detected due to its absence from existing databases. A previous study from 2016 found a phylotype that matches *P. caeni* in batch experiments utilizing biomass from a chain-elongating reactor fed a variety of substrates and found its occurrence did not correspond to chain elongation activity [134].

From our metagenomic and metaproteomic analyses, we conclude that the RBOX pathway was active in our reactors (Figure 3.5). Some abundant bacteria did not have the complete RBOX pathway (Figure 3.5), which may indicate: (i) that our methods did not always identify all genes or proteins; or (ii) that chain elongators live in syntropy with each other to produce the medium-chain carboxylic acids. We observed that RBOX was affected by the partial pressures of H_2 in the headspace of the reactors, which follows the current understanding of chain elongation.

The presence and distribution of specific bacterial microcompartments in the dominant bacteria (Fig. S3) could influence this observed metabolic pathway, reflecting the potential versatility introduced by these microcompartments. We should note that we only have evidence that genes for the bacterial microcompartments were present in the genomes, not that they were expressed. As expected, *C. kluyveri* (*Clostridium* B) had most of the RBOX pathway in the metagenome and proteome (Figure 3.5). Some *Clostridia* class members had the complete RBOX pathway in their metagenome and proteome, including an *Oscillibacter* species (Figure 3.5). We also note that some bacteria found in our study, specifically *A. inquinata* and JAAYAE01 (Acholeplasmataceae bacterium) had the complete RBOX pathway (Figure 3.5), but are not known chain elongators. Previous researchers have also noted the presence of the RBOX pathway in bacteria that are not known chain elongators [12].

Our study provides insight into the bacteria producing *n*-caproate and *n*-caprylate from ethanol and acetate via RBOX. We identified potential candidates for *n*-caprylate production in our reactors. Future studies should try to isolate and sequence *n*-caprylate-producing bacteria. In addition, future studies should also investigate whether a relative lack of diversity in *n*-caprylate-producing reactors affects the stability of these systems. The potential influence of bacterial microcompartments on this metabolic pathway, as observed in our shotgun metagenomic analysis, underscores the need to consider their role in future studies. Future research should also investigate the role of microaerobic

conditions in these reactors because we observed that O_2 is a sensitive parameter, but we do not know why.

Materials and Methods

Continuously fed reactor system

We designed, self-built, and operated three grade-316 stainless steel reactors with a 5.5 L total volume (5 L working volume and 0.5 L headspace volume) (parts utilized in the building system are detailed in Table S3). We maintained the reactor pH at ~5.5 (via periodic additions of 0.5 M HCl) and the temperature at 30 °C ± 1.0 °C. The reactors were continuously mixed via a peristaltic pump (Cole Parmer, Part No. 7520-10), which recirculated the reactor broth at a rate of ~40 mL min^{-1} by removing broth from the top of the reactor liquid level and returning it to the reactor base (internal recycle line; Fig. S1). We continuously fed the reactors with a modified-based media that was previously described [91, 135] and supplemented with ethanol and acetate.

After a 75-day startup period, we mixed broth from all reactors to ensure similar microbiota in each reactor before a restart. We operated the reactors as replicates in which we kept organic loading rates and hydraulic retention time (HRT) at 1.5 × 10^2 ± 4.6 mM C L^{-1} d^{-1} and 8.5 ± 0.2 days, respectively, for a period of 68 days (Period 1 of study—Days 75 to 142; see Table 3.1). This organic loading rate was lower than in prior studies in our lab with ethanol and acetate-fed reactors [90, 91]. Still, the different reactor designs should be noted (i.e., continuously mixed reactors in this study vs upflow anaerobic filters in the prior studies). The solids retention time was not measured in our reactors. During Periods 1 to 3 of the study, the molar ratio of ethanol to acetate was maintained at 10:1 in the substrate, and the ethanol concentration was ~600 mM.

Reactors were inoculated with 10% by volume (~500 mL) of reactor broth from a reactor that was fed semi-continuously (~once every 2 days) with ethanol-rich yeast fermentation beer and operated as an anaerobic sequencing batch reactor for an operating period of approximately 5 years prior to the time we collected the inoculum [94, 103]. For in-line product extraction, we used a setup such as the one previously described by Agler et al. [94]. Detailed information on the reactor setup can be found in the Supporting Information. During Periods 2 and 3, gases (N_2 and H_2) were sparged into the bottom port of the reactors (Table 3.1).

Experimental periods for reactors

The primary study periods were Periods 1 (Days 75 to 142), Period 2 (Days 143 to 184), and Period 3 (Days 185 to 234), and were divided into periods 1A—Days 75 to 91, 1B—Days 92 to 113, 1C—Days 114 to 125, 1D—Days 126 to 142, 2A—Days 143 to 151, 2B—Days 152 to 162, 2C—Days 163 to 184, 3A—Days 185 to 194, 3B—Days 195 to 206, and 3C—Days 207 to 222 (Table 3.1). Prior to Period 1, there was a 75-day startup period for the reactors in which the organic loading rate was incrementally increased to the target loading rate of ~1.4 × 10^2 mM C L^{-1} d^{-1} at an HRT of ~9 days. At the start of Period 1 (Day 75), biomass from all three reactors was combined, mixed, and redistributed. During Period 1, operating conditions (i.e., temperature, pH, product extraction) were kept the same. During Period 2, gas sparging of N_2 gas was tested out (i.e., gas sparging was off and on irregularly between Days 143 to 184) (Table 3.1). At the start of Period 3 (Day 185), biomass was again mixed and redistributed. During Period 3, we sparged Reactor 1 and Reactor 3 continuously with N_2 gas, while we sparged Reactor 2 continuously with a mixture of H_2 and N_2 gas (Table 3.1). Although we did not measure the gas flow rate that we sparged into the reactors during Period 3, sparging rates are assumed to be equal to the measured exit gas flow rates reported due to low gas production rates that we observed during Period 1 without sparging (Table 3.1). Throughout Periods 1 to 3, we aimed for similar organic loading rates to all reactors. Relatively small differences in organic loading rates (Table 3.1) can be attributed to minor differences in the influent flow rate and prepared influent composition supplied to the three reactors.

Liquid and gas analysis

We collected liquid samples from reactor broth and alkaline extraction solution to measure carboxylate and ethanol concentrations. The 2 mL samples of reactor broth were collected from a port in the reactor system recycle line. In contrast, we collected the alkaline extraction solution samples from a ~3 L well-mixed glass reservoir from which the extraction solution was re-circulated. Samples were stored frozen at -20°C prior to analysis. Gas chromatography systems were used to determine carboxylate and ethanol concentrations, as has been described by Usack et al. [136]. We collected gas samples from the gas exit lines of the reactors. CO_2, CH_4, and H_2 concentrations (>0.2% by volume) were measured using a gas chromatography system, which has been described previously [136]. A reduction gas detector was used to measure H_2 gas concentrations <0.2%, which has been described by Kucek et al. [91].

Calculations and statistical analysis of operating data

We calculated the carboxylate production rates as average values for each operating period. We summed the average effluent production rates per liter of the reactor (mmol C L^{-1} d^{-1}) and the average transfer rates via product extraction (mmol C L^{-1} d^{-1}) to yield total production rates per liter of the reactor (mmol C L^{-1} d^{-1}). We calculated the average effluent production rates by dividing the average carboxylate concentration per period by the average HRT. We calculated the average HRT per period based on the average effluent flow rate per period, which was determined gravimetrically. We calculated the average transfer rates by plotting the increasing concentrations of individual carboxylates in alkaline extraction solution vs time. We used least squares methods to determine the slope and the sample standard deviation (LINEST function, Microsoft Excel). We divided the slope by the reactor working volume (5 L) to obtain an average transfer rate per period. RStudio v.1.0.136 [137] was used to run data analysis in R. Concentrations, rates, ratios, and efficiencies are reported as mean value ± standard error in the paper unless noted otherwise.

16S rRNA gene sequencing analysis

We took close-to-weekly biomass samples for Illumina 16S rRNA gene sequencing analysis from the internal recycle line of the reactors. Approximately 10 mL of reactor broth was collected with a 60 mL plastic syringe and distributed into 2 mL Eppendorf tubes. We centrifuged the tubes at 16,873 × *g* for 4 min and discarded the supernatant. Finally, we stored the pelleted biomass samples at -80 °C.

According to the manufacturer's protocol, we extracted genomic DNA using the PowerSoil-htp 96 Well Soil DNA Isolation kit (MO BIO Laboratories Inc., Carlsbad, CA, USA). PCR amplification with 515-forward and 806-reverse Golay barcoded primers targeting the V4 region of the 16S rRNA gene of the extracted DNA was described previously [132] with the following exceptions: Mag-Bind RxnPure Plus magnetic beads solution (Omega Biotek, Norcross, GA, USA) was used instead of Mag-Bind E-Z Pure, and 50 ng DNA per sample was pooled instead of 100 ng. Duplicate PCR reactions of each DNA extract were performed and pooled prior to sequencing. Samples were sent for paired-end sequencing (2 × 250 bp) on the Illumina MiSeq platform (Illumina, San Diego, CA, USA) at the Cornell University Biotechnology Resource Center (Ithaca, NY, USA). We analyzed the resulting 16S rRNA gene sequencing reads using QIIME 2 2017.3 [138] and the Silva database release 138.1 [139]. Finally, we investigated the correlation of the relative abundance for OTUs with *n*-caproate and *n*-caprylate production rates using the scipy-stats package pearsonr [140].

Shotgun metagenomic analysis

We collected biomass samples for shotgun metagenomic analysis approximately weekly from internal liquid-recycle lines of the reactors, which were utilized to mix the reactor liquid. Samples were centrifuged, supernatant was discarded, and biomass was stored at -80℃. Genomic DNA was extracted using the PowerSoil DNA Isolation kit (MO BIO Laboratories Inc.). We used a modified protocol, which has been described by Kucek et al. [91]. After quantifying the extracted DNA, we selected nine samples for shotgun metagenomics sequencing (three samples for each reactor during Periods 1C, 3B, and 3C). For Period 1C, we selected one sample from Reactors 1 and 2 on Day 137 and a pooled sample from Reactor 3 from Days 137, 151, 154, and 162. For Period 3B, we selected a pooled sample from each reactor on Days 198 and 200. For Period 3C, we selected one sample from each reactor on Day 218. Pooled samples were utilized if the concentration of the genomic DNA extracted was low on a single day. The nine selected DNA samples were barcoded and sequenced on two lanes (100 bp per read; single-direction reads) using the Illumina HiSeq platform at the JP Sulzberger Genome Center at Columbia University (New York, NY, USA). We merged the replicates of samples.

Shotgun metagenomics read quality was checked using FastQC version 0.11.9 [16]. The sequence quality scores and histograms passed standard test criteria for all samples (lower quartile for every base above 10 and median above 25). To trim low-quality regions and remove low-quality reads, Trimmomatic [141] version 0.39. was applied on all samples providing the parameters -phred33; LEADING:3; TRAILING:3 as well as SLIDINGWINDOW::4:15 and MINLEN:36. Trimmed reads were aligned to NCBI-nr database (Feb 2021) using DIAMOND [142] version 2.0.7 in blastx mode. The following parameters were used: --outfmt 100 -c1 -b12 -p 32 --top 10 -e 0.001. Resulting alignments were meganized for further analysis using daa-meganizer, which is a tool that is included in MEGAN6 [143]. DIAMOND output files were loaded into MEGAN6, were normalized by sample size, and read counts were extracted for each MAG. Heatmaps were created using a Python script, only displaying MAGs with more than 12 k aligned reads (Figure 3.4). The correlation of read counts with *n*-caproate and *n*-caprylate production rates was investigated using the Pearson correlation coefficient.

De-novo assembly

We performed de-novo assembly for each set of quality filtered reads using MEGAHIT [47].

The samples from Periods 1D and 3C for Reactor 1 failed the per-base-sequence-

content test, while samples from: (i) Periods 1D and 3C for Reactor 1; (ii) Period 1C for Reactor 2; and (iii) Period 3B for Reactor 3 all failed the per-sequence-GC-content test. Sequence duplication level was high in all samples except for the sample from Period 3C from Reactor 2. This problem was introduced when merging two replicates for each sample and is an artifact. Overall, the per-sequence quality scores were sufficient.

Metaproteomic analysis

For metaproteomic sampling, approximately 200 mL of reactor broth was collected from the internal recycle lines of the reactors and distributed into four 50 mL centrifuge tubes. After centrifugation for 10 min at 8,000 × *g* (at 4℃), the supernatant was discarded. Pellets were resuspended in a tris buffer solution and redistributed to 2 mL Eppendorf tubes. We spun the tubes for 4 min at 16,873 × *g* and discarded the supernatant. Next, we stored the pelleted samples at -20℃.

Protein samples were extracted from reactor cell pellets (~100 µL bulk volume) using a gel-free, precipitation-free method to avoid loss of hydrophobic proteins. Cell pellets were suspended in 500 µL 50 mM Tris buffer (pH 8.0) and flash frozen 3× with liquid N_2 as an initial lysis step. 0.1% SDS, 10 mM NaCl, 0.02 M TCEP, and 2 M urea were added to lyse the sample by ultrasonication on the ice at 60% amplitude for 5 min total pulse time, vortexed, and centrifuged 10 min at 12,000 × *g*. Half of the supernatant (~250 µL) was removed and saved. To attempt to desorb more hydrophobic proteins from the pellet, 250 µL of acetonitrile was added to the cell pellet and the remaining supernatant. This was then vortexed and pelleted, while supernatant from this step was removed and re-combined with the first 250 µL of supernatant. The volume of the combined supernatant was decreased to approximately 400 µL via speed vac. We discarded the insoluble pellet. Total protein estimates were measured using the Bradford assay. Protein samples were reduced with an additional 0.05 M TCEP in 0.1 M ammonium bicarbonate at 35℃ for 1 h, alkylated with 40 mM iodoacetamide at room temperature for 30 min, and digested with Pierce Trypsin Protease MS-Grade at an estimated 1:20 trypsin:protein mass ratio for 12 h at 35℃ with 1 mM $CaCl_2$. Sample protein precipitation was avoided during digestion by diluting trypsin protease in 0.1 M ammonium bicarbonate buffer containing 0.02% SDS and 10% acetonitrile before combining with the protein sample. To quench digestion, samples were acidified to a pH of 3.5 with formic acid, acetonitrile was removed via speed-vac, acidified again to pH 3.5, and stored at -20℃. Tryptic peptides were purified using 1 mL Supelclean ENVI-18 SPE tubes and dissolved in 0.1% TFA/0.5% acetonitrile for analysis by liquid chromatography-mass spectrometry (LC-MS).

LC-MS was performed using a Thermo Fisher UltiMate 3000 LC and LTQ-XL mass

spectrometer with a standard ESI source. Microflow chromatography was performed on an Acclaim PepMap 100 column (1 mm × 15 cm; 3 um) at 40 µL/min using a 125 min gradient from 100% water (1% formic acid) to 40% acetonitrile. We operated the LTQ-XL in a 3× double play mode with a 10 s dynamic exclusion time and CID activation. The resulting peptides were compared to a decoy search. Peptides were thrown out based on a probabilistic filter. Proteins were kept if they had at least two unique peptides IDed with high confidence. The resulting 341 protein sequences were aligned against NCBI-nr (Feb 2021) using DIAMOND blastp version 2.0.7 for taxonomic assignment. To check for proteins involved in RBOX, we searched for these proteins using the previously described HMM models.

Supplemental material

Supplemental material related to this study can be found in the original publication and is not included in this thesis.

Data availability

16S rRNA gene sequences are available at EBI (https://www.ebi.ac.uk/) under accession number ERP024135. Sequences and study metadata are publicly available in QIITA (https://qiita.ucsd.edu/) under study number 11227. Shotgun metagenomics data is available at SRA (https://www.ncbi.nlm.nih.gov/sra) under the accession PRJNA824684. Metagenomics and metaproteomics data analysis code is available on GitHub (https://github.com/lucass122/caprylate_reactor_paper).

Acknowledgments

The authors would like to acknowledge Chase Brett and Doug Caveney for their help with constructing the reactor and Alex Marzelli and Dr. Jiajie Xu for assistance with reactor maintenance (all from Cornell University). We acknowledge funding from the U.S. EPA STAR grant fellowship, the U.S. Army Research Laboratory, and the U.S. Army Research Office under contract/grant number W911NF-12-1-0555. We also acknowledge funding from the Alexander von Humboldt Foundation in the framework of the Alexander von Humboldt Professorship to L.T.A., the Novo Nordisk Foundation CO_2 Research Center with grant number NNF21SA0072700 to L.T.A., the Deutsche Forschungsgemeinschaft under Germany's Excellence Strategy (EXC 2124–390838134) to L.T.A. and D.H. A special thanks go to the Reinhard Frank Stiftung to support the exchanges

between the University of Maryland and the University of Tübingen.

Chapter 4

MMonitor for Real-Time Monitoring of Microbial Communities Using Long Reads

This chapter is based on a manuscript written by multiple authors. Author contributions are detailed in the table below.

Title of paper:	MMonitor for Real-Time Monitoring of Microbial Communities Using Long Reads
Status in publication process:	Published in Cell Reports Methods (DOI: https://doi.org/10.1016/j.crmeth.2025.101266)

Author	Author position	Scientific ideas (%)	Data generation (%)	Analysis & interpretation (%)	Paper writing & editing (%)
Timo N. Lucas	First	50	20	60	70
Ulrike Biehain	Second	10	70	20	10
Anupam Gautam	Third	5	0	0	5
Kurt Gemeinhardt	Fourth	0	10	0	0
Tobias Lass	Fifth	5	0	0	0
Simon Konzalla	Sixth	5	0	0	0
Ruth E. Ley	Seventh	5	0	0	0
Largus T. Angenent	Eighth	10	0	10	7.5
Daniel H. Huson	Corresponding	10	0	10	7.5

Timo N. Lucas[1,2], Ulrike Biehain[2,3,7], Anupam Gautam[1,2,7], Kurt Gemeinhardt[3], Tobias Lass[1], Simon Konzalla[1], Ruth E. Ley[2,7], Largus T. Angenent[2,3,4,5,6], Daniel H. Huson[1,2,*]

[1]Institute for Biomedical Informatics (IBMI), University of Tübingen, Sand 14, 72076 Tübingen, Germany
[2]Cluster of Excellence – Controlling Microbes to Fight Infections, University of Tübingen, Auf der Morgenstelle 28, 72074 Tübingen, Germany
[3]Environmental Biotechnology Group, Department of Geosciences, University of Tübingen, Schnarrenbergstraße 94–96, 72076 Tübingen, Germany
[4]Department of Biological and Chemical Engineering, Aarhus University, Gustav Wieds Vej 10D, 8000 Aarhus C, Denmark
[5]The Novo Nordisk Foundation CO_2 Research Center (CORC), Aarhus University, Gustav Wieds Vej 10C, 8000 Aarhus C, Denmark
[6]Ag Angenent, Max Planck Institute for Biology, 72076 Tübingen, Germany
[7]Department of Microbiome Science, Max Planck Institute for Biology, 72076 Tübingen, Germany
[*]Address correspondence to Daniel H. Huson, daniel.huson@uni-tuebingen.de

SUMMARY

Real-time monitoring of microbial communities offers valuable insights into microbial dynamics across diverse environments. However, many existing metagenome analysis tools require advanced computational expertise and are not designed for monitoring. We present MMonitor, an open-source software platform for real-time analysis and visualization of metagenomic Oxford Nanopore Technologies (ONT) sequencing data. MMonitor includes two components: a desktop application for running bioinformatics pipelines through a graphical user interface (GUI) or command-line interface (CLI), and a web-based dashboard for interactive result inspection. The dashboard provides taxonomic composition over time, quality scores, diversity indices, and taxonomy–metadata correlations. Integrated pipelines enable automated de-novo assembly and reconstruction of metagenome-assembled genomes (MAGs). To validate MMonitor, we tracked human gut microbial populations in three bioreactors using 16S rRNA gene sequencing and applied it to whole-genome sequencing (WGS) data to generate high-quality annotated MAGs. We compare MMonitor with other real-time metagenomic tools, outlining their strengths and limitations.

KEYWORDS

Real-time metagenomics, Nanopore sequencing, Bioinformatics software, Microbiome, Data visualization, Biotechnology

Motivation

Metagenome monitoring is essential for understanding and managing microbial communities in dynamic environments. However, existing tools for metagenome analysis often lack key features needed for effective real-time monitoring — including immediate processing of nanopore data, intuitive user interfaces, simple deployment, and clear visual summaries. Many workflows require advanced computational skills, manual integration of multiple software packages, or substantial hardware resources, which can delay responses in biologically active systems such as clinical diagnostics, industrial bioprocesses, and environmental surveillance.

MMonitor addresses these limitations by providing real-time analysis of nanopore sequencing data within an accessible, easy-to-configure platform. It integrates taxonomic and functional profiling with genome assembly, binning, and annotation, while delivering informative dashboard visualizations for rapid interpretation. By automating standard bioinformatics steps and combining them into a cohesive system, MMonitor enables faster and more reliable metagenome tracking for users with diverse expertise across health, biotechnology, agriculture, and environmental applications.

Introduction

Monitoring metagenomes through long-read sequencing

Metagenomics has advanced our understanding of microbial diversity and function in diverse environments [144]. Sequencing and bioinformatic analysis are applied to bioactive systems such as anaerobic digesters, where they support process optimization, or human microbiomes, where they provide insights into host–microbiome interactions and health [145, 146, 147]. Metagenomic monitoring also links microbial dynamics to biogeochemical processes in larger ecosystems such as oceans and soils [148, 149].

Traditional assays (optical, fluorescence-based, or qPCR) sensitively quantify specific taxa, but are limited to predefined targets and cannot capture overall community dynamics. Long-read sequencing technologies, such as ONT platforms, overcome these limitations. Compared to short-read sequencing, long reads provide higher taxonomic

resolution and simplify the assembly of complete genomes from metagenomes or isolates, without requiring complementary short-read data [150, 151]. High-quality genomes are essential for detecting structural variations [152] and improving functional annotations [153].

Analyzing long-read data poses unique computational challenges. Algorithms developed for short reads often struggle with higher error rates and longer sequence lengths. Efficient methods are essential for real-time monitoring, particularly in time-sensitive contexts such as pathogen detection, where rapid analysis can aid in interventions. Several algorithms for long-read analysis have emerged: MetaMaps [154] and Emu [155] were designed specifically for this purpose, while Centrifuge [156] and Kraken2 [157], though originally developed for short reads, have also proven useful for long-read metagenomics [158].

Nanopore sequencing has already been applied to diverse real-time monitoring tasks, including tracking resistance genes [159], identifying outbreaks [160], and pathogen surveillance in insects [161] and freshwater systems [162]. Real-time analysis is particularly valuable in industrial settings, where undetected contamination can cause costly production losses. The ONT MinION, with its portability and rapid turnaround, has even been deployed outside conventional laboratories, including on the International Space Station [163] and on ocean research vessels [164].

Several tools already implement real-time monitoring, including minoTour [165], NanopoReaTA [166], MARTi [167], and Nanometa Live [168], but they face challenges in scalability, ease of use, and flexibility. Many focus heavily on taxonomic profiling without a real-time component, while others lack features such as time-series visualization or effective sample management. More comprehensive pipelines, such as HUMAnN [169] or QIIME [170], can generate detailed reports but often require command-line expertise, limiting accessibility. Online platforms such as MG-RAST [171], BusyBee [172], and BugSeq [173] provide user-friendly interfaces, but do not process continuous sequencing streams. Without frequent database updates, sensitivity also declines due to outdated references.

With these limitations in mind, we developed MMonitor, a software platform optimized for real-time microbial monitoring based on ONT long-read sequencing.

DESIGN

We designed MMonitor considering the following:

- Intuitive user interfaces (graphical & command line) that appeal to researchers of

varying levels of computational skill.

- Ability for time-series monitoring to track taxonomic shifts as new reads arrive.
- Customizable pipelines for 16S rRNA gene and WGS based taxonomic and functional analysis including metagenome assembly.
- Up-to-date databases with the ability to update automatically.
- Easy operation and installation.
- Useable on high-end consumer-grade hardware.
- Remote data access from any device or local processing.

A comparative analysis of our method with other tools in the real-time metagenomics space (e.g. minoTour [165], NanopoReaTA [166], MARTi [167], Nanometa Live [168], MAIRA [174], EPI2ME [175], BoardION [176], and CRuMPIT [177]) can be found in the Results section.

RESULTS

Metagenome monitoring software

Here we illustrate the use of MMonitor, an open-source software for automated monitoring through the real-time analysis of metagenomic nanopore sequencing data. The software consists of two components: a desktop application for running analyses and a web-based application for interactive inspection of results. The desktop part integrates established bioinformatics pipelines that can be run through a graphical user interface (GUI) or a command line interface (CLI), enabling analysis on local or remote systems. The software leverages the real-time characteristics of nanopore data, allowing automated processing of metagenome reads as the sequencer continuously generates data.

The web application features a visualization dashboard that provides dynamic insights into taxonomic composition over time and the functional potential of metagenomes. It includes key metrics such as quality scores, diversity indices, and taxonomy–metadata correlations, with options to export results for further analysis. MMonitor was developed and tested in collaboration with environmental biotechnologists to ensure its applicability for on-site metagenome tracking.

In the following sections, we describe how the software was used to track microbial communities in multiple bioreactors over a month using 16S rRNA gene sequencing

data. We then illustrate how MMonitor processes whole-genome sequencing (WGS) data to generate annotated, high-quality metagenome-assembled genomes (MAGs) for functional insights. Finally, we compare MMonitor with other existing tools for metagenome monitoring.

Data sequenced at regular intervals (e.g., weekly samples in our study, though the frequency depends on the application) are automatically analyzed by MMonitor Desktop, and results are sent to the dashboard. Data can be accessed locally or through a web browser. MMonitor offers two pipelines (Figure 4.1B): one for taxonomic analysis and one for functional analysis. The taxonomic pipeline processes both 16S rRNA gene sequencing and WGS reads, whereas the functional pipeline requires WGS data.

16S rRNA gene sequencing provides rapid taxonomic profiles, and multiplexing allows simultaneous analysis of multiple samples. In contrast, WGS offers deeper taxonomic resolution and strain-level insights but requires more time and computational resources. WGS also enables *de novo* assembly for full genome insights. We recommend 16S rRNA gene sequencing for general community overviews and WGS at key intervals for in-depth analysis. Details of the pipelines are provided in the Methods section.

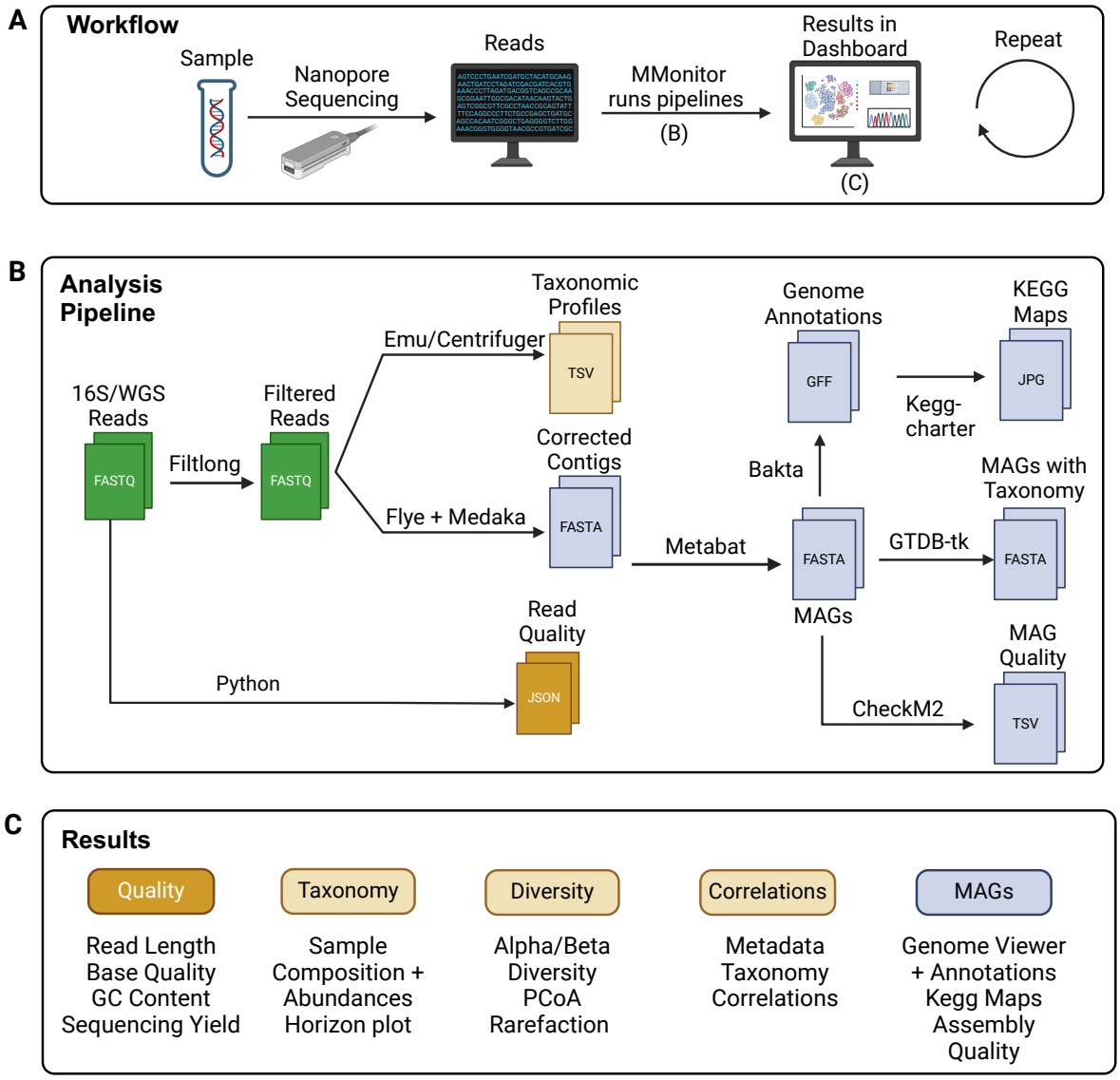

Figure 4.1: Overview of the MMonitor workflow and outputs. (A) Typical lab workflow using MMonitor for real-time metagenome tracking. (B) Computational steps in the taxonomic and functional pipelines. (C) Examples of dashboard outputs (time-series taxonomy, QC, diversity, and correlations). Created with BioRender.com.

16S rRNA gene time series in three BES reactors

This section reports nanopore 16S rRNA gene time series data from three bioelectrochemical systems (BESR1–BESR3). In total, 66 amplicon libraries collected between July and August 2023 were analyzed with MMonitor. For clarity, BESR1–BESR3 are hereafter referred to simply as R1–R3. Unless noted otherwise, all mentions of R1–R3 refer exclusively to the 16S datasets from these BES reactors. Genome-resolved WGS results from a separate reactor are described in a later section and are based on a single sample.

The bioelectrochemical systems originated from a parallel project designed to test whether removing H_2 with electrodes could steer anaerobic fermentation pathways. The reactors achieved stable H_2 removal and produced a 16S rRNA gene sequencing time series. We primarily used these data to validate MMonitor: frequent sampling under a defined perturbation provided an excellent test case for real-time tracking, horizon plots, and diversity analyses.

Reads were quality-filtered using the 16S defaults (removing reads < 1000 bp, > 2000 bp, or with PHRED < 10). The remaining reads were used for downstream taxonomic profiling and diversity analyses.

Here, we analyze the taxonomy of the three monitored bioreactors at different taxonomic ranks (Figure 4.2). Across the three BES reactors (R1, R2, and R3), microbial communities were consistently dominated by the phylum Bacillota, with smaller contributions from Proteobacteria, Actinobacteria, and Bacteroidetes. Over the course of monitoring, MMonitor showed that community composition (Figure 4.2A) was relatively stable at higher taxonomic ranks, with Bacillota dominating throughout. At finer resolutions, fluctuations became more apparent.

At the genus level, taxa such as *Escherichia*, *Streptococcus*, *Lactobacillus*, *Blautia*, *Anaerostipes*, and *Faecalibacterium* dominated alongside other minor genera (Figure 4.2B). In summary, periodic variations were observed at finer taxonomic ranks, while higher ranks showed a more stable equilibrium dominated by Bacillota and Proteobacteria.

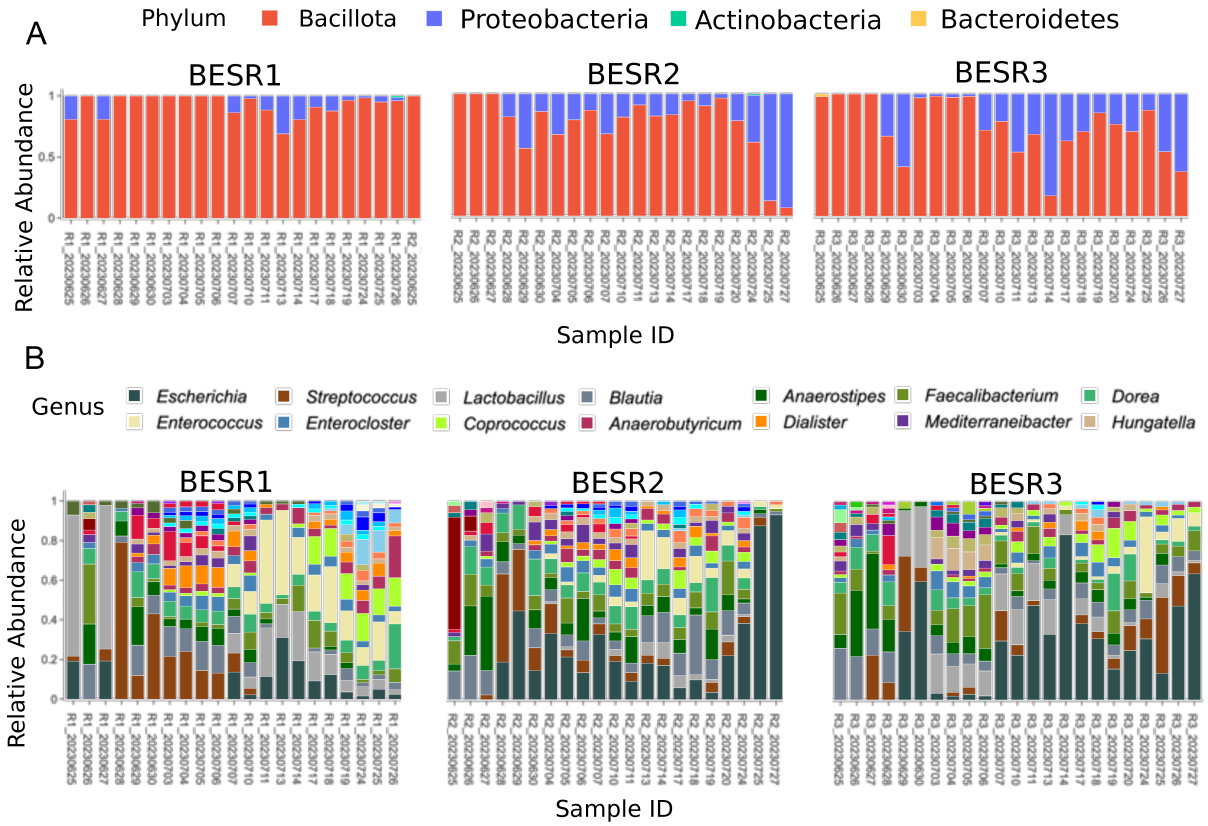

Figure 4.2: Community composition across three BES reactors. Legends display only the 14 most abundant taxa. (A) Stacked bar plots of relative abundance at the *phylum* level for R1–R3. (B) Stacked bar plots at the *genus* level for R1–R3. Created with BioRender.com.

In reactor R1, the five most abundant species were *Enterococcus faecalis*, *Streptococcus salivarius*, *Escherichia coli*, *Faecalibacterium prausnitzii*, and *Anaerostipes hadrus*. Species-level changes were more pronounced: while daily variation was often minor, over the course of weeks the community was unstable, with some taxa emerging and others disappearing (Figure 4.3A).

These dynamics are visualized in a horizon plot (Figure 4.3B), which shows relative abundance changes of the 20 most abundant species in R1 over time. Each row represents a species: red bands indicate increases compared to the mean abundance, and blue bands indicate decreases. For example, *S. salivarius* and *Dorea longicatena* increased early, then declined, while *E. coli* and *E. faecalis* only appeared after 10 July, resulting in continuous blue bands before that date followed by red increases afterward. Such patterns highlight potential key time points, e.g., a marked community shift around 11 July, or earlier changes on 25 June affecting taxa such as *Lactobacillus delbrueckii*, *F. prausnitzii*, and *A. hadrus*.

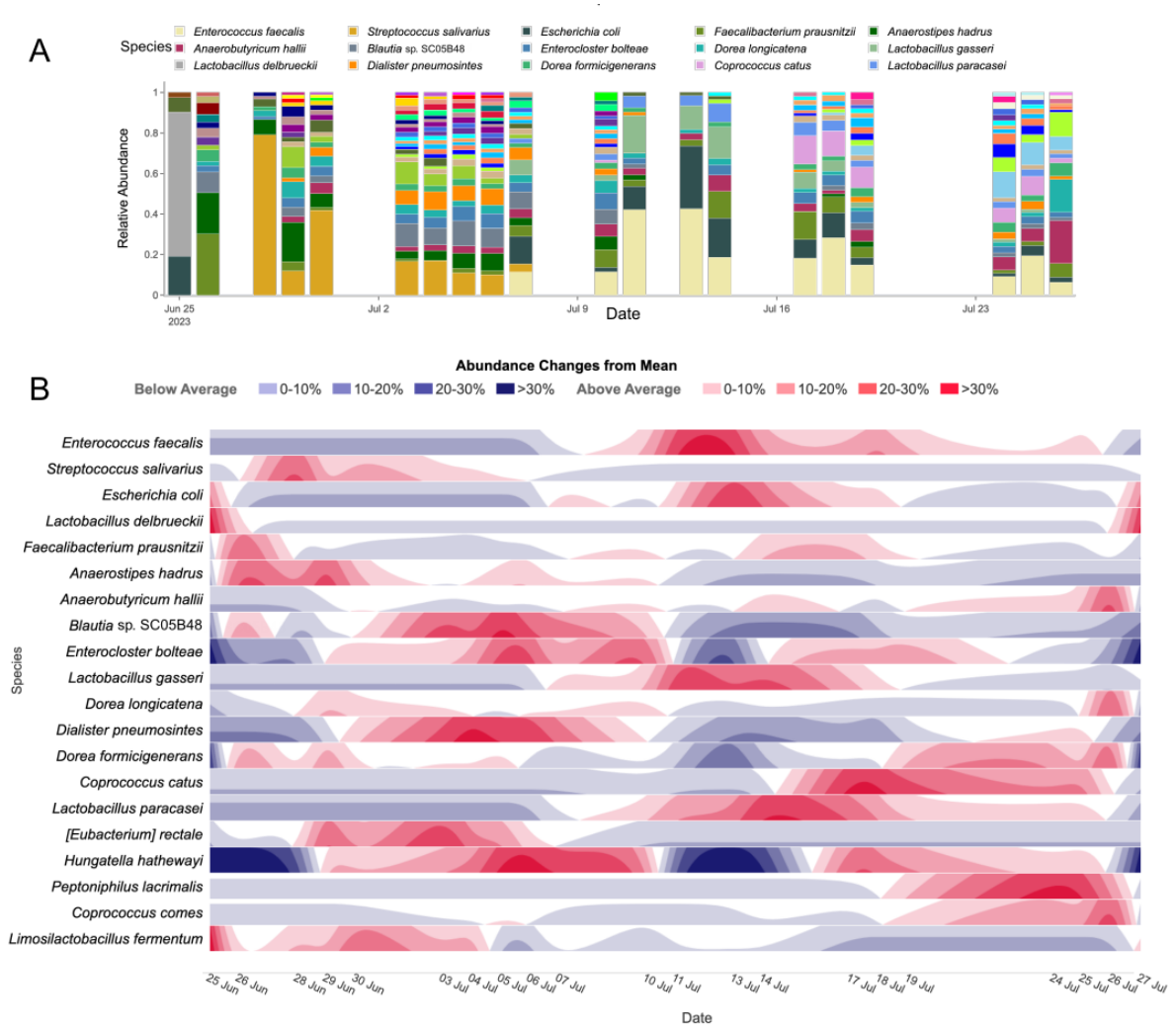

Figure 4.3: Species-level dynamics in reactor R1. (A) Stacked bar plot of the most abundant species over time. (B) Horizon plot showing deviations from mean abundance for the 20 most abundant species (red: above mean; blue: below mean). Created with BioRender.com.

Community shifts in the BES reactors were also described in detail in a previous dissertation based on the same dataset [178]. That study reported marked changes at 201 h, 354 h, and 509 h, with differences between replicates and two enrichment phases in R1 (0–226 h and 251–559 h). These shifts were associated with metabolite concentrations (e.g., acetate, propionate, n-valerate) and biofilm growth at electrodes, which likely influenced hydrogen availability. The exact causes of the observed species turnovers remain unresolved. As the goal of the present study was to demonstrate MMonitor's ability to capture such dynamics in real time, we focused on showcasing monitoring capabilities rather than providing a full biological interpretation.

Analysis based on de novo assembled MAGs and functional analysis

To demonstrate additional use cases of our software, we also applied it to WGS data. This is particularly important, as 16S rRNA gene sequences alone are often unable to differentiate between functionally distinct microbial taxa. Closely related species or strains with nearly identical 16S rRNA genes may possess different genomic content and functional capabilities [179]. WGS data also enables the tracking of other phylogenetic and functional marker genes beyond 16S rRNA, which is particularly valuable in complex biological environments where distinct taxa may fulfill similar ecological roles.

MMonitor successfully generated metagenome-assembled genomes (MAGs) and provided insight into the functional potential of individual microbes. From a nanopore WGS dataset, we obtained seven high-quality MAGs, which met the Minimum Information about a Metagenome-Assembled Genome (MIMAG) standards [123] as assessed by CheckM2 [180].

Taxonomic annotation using GTDB-Tk revealed that six MAGs belonged to Bacteria and one to Archaea (Table 4.1). Two MAGs were classified down to the species level: *Pseudoclavibacter_A caeni* and *Methanobacterium_C congolense*, with ANI values of 99.28% and 97.88%, respectively, indicating high sequence similarity to known reference genomes. The remaining MAGs were classified at the genus level, including *Dysosmobacter*, *Bulleidia*, and *Acetobacter*, suggesting they are novel species within these known genera. In particular, two MAGs (*JAEXAI01* and *JAUZPN01*) lacked close reference genomes and were assigned new placeholders at the genus level, indicating that they may represent novel genera (Table 4.1). MMonitor also annotated the MAGs and mapped the annotations to the KEGG database to create metabolic maps that show the metabolic potential of a MAG.

Table 4.1: High-quality MAGs recovered from nanopore WGS. MAGs meeting MIMAG "high-quality" criteria (completeness $\geq 90\%$, contamination $< 5\%$), with GTDB taxonomic annotation and the closest reference genome (ANI, %).

MAG ID	Completeness (%)	Contamination (%)	Genome size (bp)	GTDB classification	Closest reference genome (ANI %)
bin.2	94.63	0.13	801 950.00	*Pseudoclavibacter_A caeni*	GCF_008831125.1 (99.28%)
bin.5	95.15	1.04	2 551 952.00	*Methanobacterium_C congolense*	GCF_900095295.1 (97.88%)
bin.6	95.58	0.05	6 495 533.00	*Dysosmobacter* sp.	No close reference
bin.11	96.77	0.53	2 816 210.00	*Bulleidia* sp.	No close reference
bin.12	91.25	0.30	1 765 253.00	*JAEXAI01* sp.	No close reference (putative novel lineage)
bin.13	99.96	1.05	3 110 456.00	*Acetobacter* sp.	No close reference
bin.17	91.03	2.30	4 688 061.00	*JAUZPN01* sp.	No close reference (putative novel lineage)

Finally, to benchmark the performance of MMonitor's WGS pipeline for taxonomic abundance estimation, we analyzed the ONT Q20 Zymo WGS mock community, following the preprocessing described in [181]. Figure 4.4 shows MMonitor-estimated abundances versus the theoretical composition, generated with the Python notebook provided in [181].

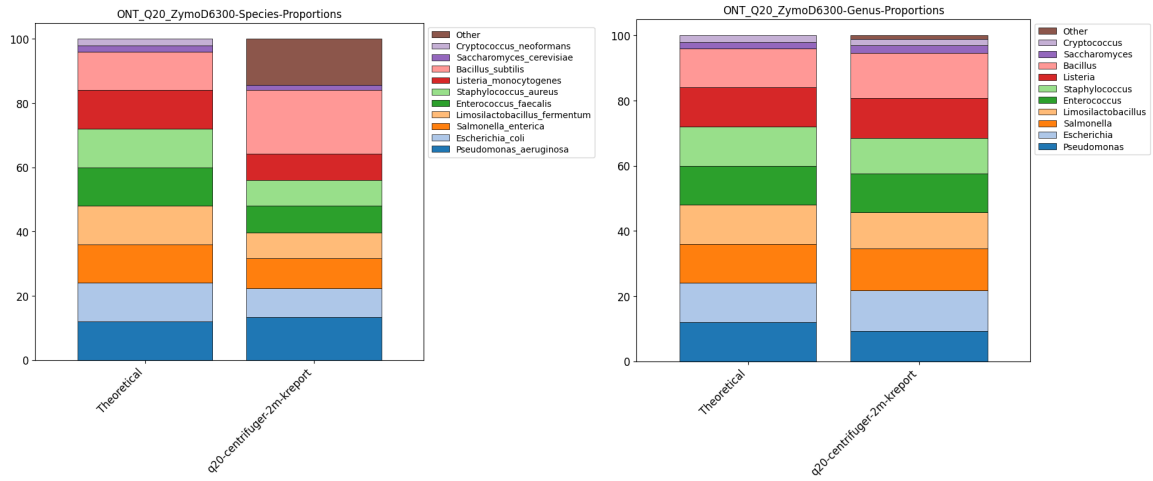

Figure 4.4: Taxonomic classification accuracy of MMonitor on the Zymo Q20 ONT mock community, see Figure 1. Stacked bar plots show the relative abundances of the most abundant taxa compared to the theoretical composition of the dataset. (A) Species-level classification. (B) Genus-level classification. The category *Other* includes low-abundance taxa and unclassified reads below the top ten most abundant groups.

Modified versions of popular algorithms for more efficient real-time monitoring

To optimize MMonitor for real-time tracking, we have made minor adjustments to taxonomic profilers without affecting result quality. Emu now optionally saves reference indices on the hard drive, avoiding recomputation during sequencing runs and improving runtime, especially with SSDs. For high-performance systems, users can increase the batch size when Emu uses minimap2, reducing input/output operations at the cost of more memory use. Centrifuger is modified to hold indices in memory and process new sequences immediately, lowering runtime for small batches where index loading is a bottleneck, enhancing efficiency in real-time scenarios. These changes do not alter the behavior of the core algorithm and can be disabled if necessary. We also introduced updated reference indices that can be downloaded from MMonitor for users who need current databases, along with an option to retrieve sequences and rebuild the indices of chosen domains.

Feature comparison with similar software

To position MMonitor within the landscape of other metagenomics tools, we compared its features to those of existing software. The primary focus of this comparison was on tools that support real-time data processing, including minoTour [165], Nanometa Live [168], MARTi [167], NanopoReaTA [166], EPI2ME [175], boardION [176] and CRuMPIT [177] (see Table 4.2). We examined these tools according to different criteria that influence the usability of a software for real-time monitoring of metagenomes.

Table 4.2: Feature Comparison of Real-Time Metagenome Tools. Nine tools compared across fifteen criteria. "Yes" = fully met, "Partly" = partially met, "No" = not met.

Feature	MMonitor	MinoTour	Nano poReaTA	MARTi	Nanometa Live	MAIRA	EPI2ME	boardION	CRuMPIT
Time-Series Visualization	Yes	Yes	Yes	Yes	No	No	No	Yes	No
Sample Management	Yes	Yes	No	Partly	No	Partly	Partly	No	No
Processing Speed	Fast	Fast	Fast	Mod	Fast	Fast	Fast	Mod	Mod
Sequence Quality Control	Yes	Yes	Yes	Yes	Yes	Partly	Yes	Yes	Yes
Statistical Analysis	Yes	Yes	Partly	Yes	Partly	Partly	Partly	Partly	No
Assembly Support	Yes	No	No	No	No	No	No	No	No
User Interface	Web, GUI, CLI	Web	Web, CLI	GUI	GUI	GUI, CLI	GUI	GUI	CLI
Data Accessibility	Web	Web	No	Web	No	No	Web	No	No
Customizable Databases	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes
Automated Reporting	Yes	Yes	Partly	Partly	Partly	Yes	Yes	No	No
Scalability	High	High	Mod	High	Mod	High	High	Mod	High
Open Source	Yes	Yes	Yes	Yes	Yes	Yes	Partly	Yes	Yes
Cost	Free	Free	Free	Free	Free	Free	Free	Free	Free
Ease of Installation	Easy	Mod	Easy	Easy	Easy	Easy	Easy	Mod	Mod
Real-Time Analysis	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Mod = Moderate. "No" under Data Accessibility indicates local-only use; "Web" denotes remote access.

Detailed feature comparison

All tools in Table 4.2 support real-time processing of nanopore data, but they differ in algorithms, downstream analyses, and usability. Several tools (e.g., MinoTour, MARTi, NanopoReaTA, Nanometa Live) provide real-time taxonomic profiling but lack *de novo* metagenome assembly. Of the compared tools, only MMonitor implements a complete metagenome assembly pipeline with the ability to generate MAGs from long-read data alone. EPI2ME includes an assembly workflow for single isolates, but not for community metagenomes.

MinoTour. A web-based, Django-backed LIMS for ONT devices that provides real-time metrics, run tracking, and taxonomic profiling during sequencing [165]. It integrates adaptive sequencing workflows (e.g., ARTIC) and can be self-hosted. MinoTour does not implement downstream metagenomic analyses (assembly, diversity/correlation time-series) and requires users to deploy their own instance, which may limit accessibility for non-technical users.

NanopoReaTA. A real-time analysis toolbox focused on RNA/cDNA data (differential

transcriptomics) from ONT [166]. Distributed via Docker for portability; command-line operation and Linux requirement can be barriers. Useful for transcription monitoring in real time, but it does not target multi-timepoint microbiome tracking or metagenome assembly.

MARTi. An open-source, browser-based platform for real-time analysis and visualisation of nanopore metagenomic data [167]. Provides dynamic community composition and AMR gene reporting with documented installation and a public demo. Well suited to rapid clinical or research use; does not focus on long-term time-series management or correlation/diversity dashboards. Currently, MARTi lacks support for metagenome assembly and full-genome functional metagenomics, as it is limited to reporting AMR and taxonomic profiles. Its utility strongly depends on up-to-date reference databases, but it does not provide automatic database updates. Users must rely on the developers to periodically create and release new versions.

Nanometa Live. A user-friendly application for real-time metagenomic analysis and pathogen identification that streams ONT reads into taxonomic profiling (Kraken2) with a GUI [168]. Can run offline once installed. It does not include genome assembly/annotation workflows or time-series/diversity/correlation dashboards; focus is on immediate species identification and run-time visualisation.

MAIRA. A standalone, Java-based program for interactive taxonomic and functional analysis of long-read metagenomes on a laptop [174]. Performs fast, online genus-level analysis with on-demand species-level and functional screens (e.g., AMR/virulence). Runs locally (no web access) and targets per-run real-time use rather than multi-timepoint tracking or dashboards.

EPI2ME. A collection of ONT-supported workflows (EPI2ME Labs) for taxonomic profiling of amplicons and metagenomes, AMR screening, and more, runnable on desktop or cloud [175]. An isolate-genome assembly mode is available, but community metagenome assembly is not the target [175]. Workflows are open source, whereas the platform itself is not; multi-sample time-series dashboards are not provided (advanced analyses typically require exporting results).

boardION. An interactive web application for real-time evaluation of ONT sequencing runs [176]. It monitors instrument performance and run QC; it does not perform taxonomic profiling, assembly, or community analyses.

CRuMPIT. A real-time analytical pathway for clinical metagenomics (pathogen detection) using nanopore data [177]. Emphasises rapid identification rather than community-ecology features (e.g., time-series dashboards, diversity/correlation analyses).

Taken together, these comparisons highlight missing features across existing tools (e.g., time-series dashboards, sample management, configurable databases, and assembly for long reads). To address these limitations, we developed MMonitor.

MMonitor. Uniquely among the tools compared, MMonitor integrates full metagenome-assembly pipelines that generate high-quality MAGs and perform annotation-based functional analysis, enabling users to link taxonomic and functional profiles directly from long-read data. It is also the only tool that supports automatic retrieval and updating of reference databases, ensuring analyses always use current data without manual intervention. Beyond these capabilities, MMonitor is designed for accessibility: it ships as prebuilt binaries with minimal setup, is fully open source, and provides both a GUI and CLI with support for remote computation. Samples are organised for easy comparisons, and custom visualisations (e.g., horizon plots) help reveal temporal trends. Pipelines scale efficiently to large datasets, and the dashboard supports interactive exploration across many samples, including very large datasets through filtering and subsetting options. The software runs on multiple platforms with bundled dependencies, includes a Docker container for easy dashboard deployment, and offers an offline mode for secure local inspection of results.

Feature comparison with less direct competitors to MMonitor

Table 4.3 compares MMonitor with indirect competitors that emphasize broad metagenome analysis but lack real-time monitoring. Tools such as MG-RAST, MEGAN, and Anvi'o provide powerful statistical and functional analyses for large studies, yet are not suited to time-sensitive workflows.

Table 4.3: Comparison of MMonitor with indirect competitors that emphasize broad metagenome analysis rather than real-time monitoring. Cells indicate support level (Yes, Limited, No). Abbreviations: GUI, graphical user interface; CLI, command-line interface; N/A, not applicable.

Feature	MMonitor	MG-RAST	MEGAN	Anvi'o	QIIME 2	MetaPhlAn 4	Nephele	PATRIC	MGnify
Time-Series Visualization	Yes	Limited	Limited	Limited	Yes	Limited	Limited	Yes	Limited
Sample Management	Yes	Limited	No	Yes	Limited	No	Limited	Yes	Yes
Speed	Fast	Moderate	Moderate	Variable	Variable	Fast	Moderate	Moderate	Moderate
Sequence Quality Control	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
Statistical Analysis Methods	Yes	Limited	Yes	Yes	Yes	Limited	Yes	Yes	Yes
User Interface	GUI & CLI	Web-based	GUI & CLI	GUI & CLI	CLI	CLI	Web-based	Web-based & CLI	Web-based
Data Accessibility	Web, API	Web	No	No	No	No	Limited	Web	Web
Customizable Databases	Yes	No	Yes	Yes	Yes	Limited	Limited	Yes	Limited
Automated Reporting	Yes	Yes	Limited	Limited	No	Limited	Yes	Yes	Yes
Scalability	High	Moderate	Moderate	Variable	Variable	High	High	High	High
Open Source	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Cost	Free	Free	Free (Academic)	Free	Free	Free	Free	Free	Free
Ease of Installation	Easy	N/A	Moderate	Moderate	Moderate	Easy	Easy	Easy to Moderate	Easy
Real-Time Analysis	Yes	No	No	No	No	No	No	No	No

MG-RAST is an online platform for automated phylogenetic and functional analysis of metagenomic data, tailored to short-read datasets analyzed after server upload. It generates standard visualizations (e.g., bar charts for taxonomy) alongside advanced views such as PCoA, rarefaction, and KEGG maps, and provides metrics like richness and evenness. Maintenance status and reference updates can affect detection performance. In our hands, short-read data from a reactor metagenome yielded genus-level assignments only, limiting its utility for precise monitoring. Although MG-RAST includes sample management, it lacks efficient multisample comparison and time-series analysis. As a web service optimized for short reads, it is unsuitable for real-time monitoring and requires data uploads with typical queue times.

MEGAN [182] is a desktop application with a GUI for taxonomic and functional analysis. It implements numerous algorithms and visualizations and supports multi-sample comparisons. However, it is not optimized for real-time usage or time-series visualization. MEGAN typically requires prior alignment against a protein reference database, which is efficient for large batches but less suitable for frequent, small increments common in monitoring.

Anvi'o [183] is an open-source platform for metagenomic analysis, including binning, pangenomics, and visualization. It supports customizable databases and advanced

statistics but uses a command-line interface with a steep learning curve and is not tailored for real-time monitoring or nanopore data.

QIIME 2 is an extensible microbiome analysis package focused on interactive analysis and visualization. It provides pipelines for taxonomic classification and diversity analyses with a plugin architecture, but is primarily designed for amplicon (16S rRNA gene) data and is not optimized for long-read or real-time sequencing; the command-line interface may limit accessibility.

Nephele [184] is a cloud-based microbiome platform with user-friendly web interfaces integrating tools such as QIIME, mothur, bioBakery, and a5-miseq. It supports amplicon and shotgun metagenomes and leverages AWS for scalable compute, reducing local burden. Reproducibility is emphasized by tracking inputs, parameters, and VM images.

PATRIC [185] is a comprehensive bacterial bioinformatics resource focused on human pathogens. It integrates genomes using RAST and provides comparative genomics tools (genome browsing, protein family sorting, pathways) plus community-derived data (disease information, experiments, literature) for infectious disease research.

MGnify [186] (formerly EBI Metagenomics) provides assembly, analysis, and storage for microbiome sequences. It offers taxonomic and functional annotations across diverse datasets and supports metagenomic and metatranscriptomic data, including long reads. Standardized and versioned pipelines ensure consistent, reproducible analyses.

Taxonomic classification of WGS MOCK dataset

MMonitor's pipelines employ previously described methods that have been benchmarked by their authors; therefore, we did not perform extensive re-benchmarking here. Because MMonitor uses updated indices for taxonomic classification, we evaluated it on the ONT Q20 Zymo WGS mock community, following the preprocessing described in [181]. Figure S1 shows MMonitor-estimated abundances versus the theoretical composition, generated with the Python notebook provided by [181].

MMonitor identified all genera, with only slight deviations from theoretical abundances. At the species level, MMonitor correctly classified the bacterial genomes but did not assign reads to *Cryptococcus neoformans* (yeast); instead, reads were assigned to a closely related *Cryptococcus* species. Species-level abundances were less accurate than genus-level abundances due to additional false positives summarized as "Other" (Figure S1). Overall, MMonitor captured the taxonomic composition of the mock community and aligned closely with theoretical expectations.

Discussion

Recent advances in sequencing technologies, along with their increasing application in clinical, environmental, and industrial microbiology, highlight the need for accessible and optimized tools for metagenome monitoring. However, the costs of sequencing and downstream analysis remain a practical limitation. These expenses vary depending on the sequencing platform, library preparation strategy, and computational setup, and include not only sequencing itself but also consumables, reagents, and computational infrastructure. Nanopore sequencing, though generally more expensive than short-read approaches, enables species- and strain-level resolution as well as real-time analysis, but at the cost of higher error rates. Real-time local analysis reduces infrastructure needs and allows on-site sequencing, but consumables such as flow cells, reagents, and computational capacity must also be considered. Alternatives to WGS-based monitoring include targeted molecular assays, such as 16S rRNA gene profiling with meta-barcoding or hybridization probe capture. These approaches can quantify specific microbes with high sensitivity; however, they are restricted to targets with high similarity and may not capture the overall composition of the community.

Possible applications for monitoring

MMonitor has a diverse range of applications, not limited to certain environments or species. It can be applied to any nanopore sequencing data, from metagenomes or isolates, and supports both whole-genome sequencing and targeted sequencing. In environmental biotechnology, it can be used to monitor microbial populations in bioreactors, aiding optimization of production processes for biofuels or bioremediation.

Metagenomic monitoring can uncover novel functional genes and enzymes, including those from uncultivable microorganisms, that can be harnessed to improve biomass conversion and biofuel production [187]. It can also track shifts in microbial community composition and activity, providing information needed to optimize and control microbiome behaviour for greater yield, stability, and a broader product range in large-scale biomanufacturing systems, as emphasized in [188]. In clinical microbiology, real-time monitoring of microbial communities can assist in tracking pathogen emergence and antibiotic resistance patterns. The ability to correlate microbial abundances with environmental or clinical metadata enhances the utility of MMonitor in these contexts.

Although metagenome monitoring has not yet been standardized as part of routine production workflows, several large-scale applications already demonstrate its feasibility. For example, shotgun metagenomics has been used to characterize resistomes in wastewater treatment plants [189], to profile microbial dynamics across sewer net-

works [190], and to monitor bioleaching microbial communities via high-throughput sequencing [191]. These studies show that, although sequencing turnaround currently limits fully real-time feedback, metagenome monitoring is already integrated into industrial workflows and is expected to expand as sequencing costs and analysis latency decrease.

Relevance of reactor systems to this study

MMonitor was validated in two complementary contexts. BES reactors provided short, controlled enrichments with frequent 16S sampling, enabling rapid taxonomic tracking and diversity analyses. In contrast, AF and UASB reactors from a separate MCFA project supplied long-running WGS datasets to demonstrate genome assembly, binning, and functional monitoring.

Together, these systems highlight the dual use cases of MMonitor: (i) rapid 16S-based tracking (BES) and (ii) genome-resolved functional monitoring in continuous bioprocesses (AF/UASB). The BES itself was only indirectly related to MMonitor, but its time series offered a stringent test case for the 16S pipeline.

Automation and predictive analytics

Another development avenue is the automation of sampling and library preparation. Manual collection, DNA extraction, and library construction remain time-consuming and variable. Many wastewater facilities already use autosamplers as part of monitoring workflows [192], and open-source automation platforms such as Opentrons and VolTRAX are emerging to standardize metagenomic library preparation [193].

In addition, MMonitor's database structure provides opportunities for predictive analyses. By continuously storing and organizing metagenome data, the system could support downstream applications such as statistical modeling, machine learning, or other AI-based approaches to detect patterns and anticipate community dynamics. This would extend MMonitor beyond monitoring toward predictive microbiome analytics.

Conclusion

In summary, MMonitor addresses current limitations in real-time metagenome analysis by providing a fully integrated solution that combines rapid taxonomic profiling, de novo assembly, and intuitive time-series visualization. By automating data analysis and key steps such as database updates and sample management, MMonitor enables users to capture dynamic shifts in microbial populations. Validation on bioreactor

samples demonstrated its reliability for both 16S rRNA gene and WGS data, and the comparison with existing tools highlights complementary strengths across the field. As an open-source project, MMonitor welcomes contributions from the scientific community, fostering continuous improvement and broader adoption across microbiome research. Looking ahead, integration of automated sampling and preparation workflows could further enhance its utility by enabling near-continuous monitoring in industrial and clinical settings.

Resource availability

Data availability

This study did not generate new unique reagents. Raw 16S rRNA gene sequencing and WGS data have been deposited at NCBI under BioProject **PRJNA1216071** (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1216071). Reactor designs and cultivation protocols are described in detail in this study and referenced publications.

Code availability

- The MMonitor source code (version 0.2.1), built binaries, and usage instructions are available on GitHub (https://github.com/lucast122/mmonitor) and archived on Zenodo under DOI: 10.5281/zenodo.17376966.
- Analysis scripts, configuration files, and manuscript source are available at https://github.com/lucast122/mmonitor-paper.
- Database-building scripts are included in the MMonitor repository. Required databases can be downloaded through the graphical user interface or referenced directly via the repository documentation.
- A hosted instance of the MMonitor dashboard is available at https://www.mmonitor.org.

Any additional information required to re-analyze the data reported in this paper is available from the lead contact upon request.

Limitations of the study

Despite its strengths, MMonitor has certain constraints and areas for improvement.

- High sequencing loads: Very frequent updates (e.g., from multiple sequencers) may overburden consumer-grade hardware, as current optimizations target moderate analysis intervals.

- OS restrictions: Although the desktop app runs on multiple platforms, the underlying bioinformatics pipelines require a Unix-based environment. Full native support for Windows is not yet available.
- Short-read integration: MMonitor is optimized for nanopore data. While the WGS pipeline uses a genome database compatible with both long- and short-read data, short-read workflows have not been extensively tested. Future releases will include full short-read support (16S rRNA gene profiling and assembly). Other long-read technologies should also work but remain untested.
- Database coverage: Default references focus on bacteria, archaea, and fungi. Viral or other specialized targets may require manual database creation and indexing.

Planned improvements include:

- Containerization: Docker or Singularity support to simplify cross-platform usage, including Windows.
- Expanded functional analyses: Integration of antibiotic resistance gene monitoring and methylation detection (via Dorado) for pathogen surveillance and microbial regulation studies.
- Broader input support: Incorporation of transcriptomic, proteomic, or metabolomic data, and compatibility with additional sequencing technologies.
- Performance optimization: GPU acceleration, improved I/O handling, and optimized reference databases for faster, large-scale analyses.

STAR Methods

Experimental model and subject details

Microbial samples Two distinct reactor systems were studied. For 16S rRNA gene sequencing experiments, three bioelectrochemical system (BES) reactors were inoculated with fresh human fecal samples collected with informed consent and processed anaerobically (see Method details). For WGS experiments, two continuous bioreactors (anaerobic filter and UASB) were inoculated with a chain-elongating microbiome derived from a previously operated CSTR and maintained over a period of ~1000 days. This experiment was approved by the Ethics Committee of the Medical Faculty of the University of Tübingen (Project number: 456/2023A). Anonymized stool samples were collected with the permission of all human subjects.

Method details

Analysis pipeline

MMonitor was written in Python (v3.11). The desktop application manages data input and pipeline execution, while the computations are performed by external bioinformatics tools (Figure 4.1B). Benchmarks and methods for these tools are available in their respective publications.

All input reads are first processed by Filtlong [18] (v0.2.1) for quality filtering. Parameters can be set in the Analysis Configuration tab of MMonitor Desktop. By default, WGS reads shorter than 2000 bp and 16S rRNA gene reads shorter than 1000 bp or longer than 2000 bp are discarded, as well as reads with PHRED scores below 10. Only reads that pass this filtering are used for downstream analysis. Basic quality statistics (quality scores, read lengths, and base counts) are computed directly from input reads using Biopython SeqIO [194].

For taxonomic analysis of nanopore WGS data, MMonitor uses Centrifuger [195], which applies the Burrows–Wheeler transform (BWT) and Ferragina–Manzini (FM) index [196, 197]. Centrifuger is similar to Centrifuge [156], but employs run-block compression to reduce memory requirements with only a modest increase in runtime. For nanopore 16S rRNA gene profiling, MMonitor relies on Emu [155], which combines minimap2 [198] alignments with an expectation–maximization algorithm [199] to refine relative abundances.

For functional analysis, WGS reads are required. The assembly pipeline follows previous work [200]. Reads are assembled into contigs using Flye [60] with the `-meta` flag and, by default, `-nano-raw`. Users can specify `-nano-hq` if basecalling was performed with a sup model. The contigs are polished using Medaka (v2.0.1) [201]. Medaka attempts to auto-detect the correct model, but if data were basecalled with an old model or are not in pod5 format, users must supply the correct model manually (see documentation: https://github.com/nanoporetech/medaka).

The consensus assembly is binned with MetaBAT2 [70], and resulting bins are annotated taxonomically using GTDB-tk (v2.4.0) [124]. MAG quality is assessed with CheckM2 [180]; bins with completeness above 80% and contamination below 10% are retained. Bakta (v1.9.4) [87] is then used for functional annotation, which is mapped to KEGG pathways using keggcharter (v1.0.2) [202, 203].

In addition to the GUI, MMonitor can be run in command-line mode, enabling remote execution on other systems (e.g., via SSH), as described in the project README.

Webserver and Dashboard The web server backend was designed with the Django

framework [204] and the dashboard was implemented using a combination of Plotly Dash and JavaScript [205]. Docker was used to bundle the dependencies of the server and can be used to deploy it. We offer a default online version of the MMonitor server that all users can access. However, if you set up your own MMonitor server, you gain full control over the database and user management, allowing you to customize access and data handling according to your needs.

In addition to visualizations, the dashboard also provides several statistical methods. Normalized counts c_n (4.1) are calculated from raw counts c_r, the number of aligned bases b, and a scaling factor f:

$$c_n = \frac{c_r}{b} \times f \tag{4.1}$$

The scaling factor is the average number of aligned bases across all samples and ensures that raw and normalized counts have a comparable magnitude.

Alpha and beta diversity are calculated using `scikit-bio` [206] (v0.5.9) based on normalized counts. For alpha diversity, the Shannon (4.2) and Simpson (4.3) indices are used, where S is the number of unique taxa and p_i is the relative abundance of taxon i:

$$H = -\sum_{i=1}^{S} p_i \log_2 p_i \tag{4.2}$$

$$D = 1 - \sum_{i=1}^{S} p_i^2 \tag{4.3}$$

Beta diversity is assessed using the Bray–Curtis distance (4.4) between all samples, where u_i and v_i are the proportions of taxon i in samples u and v:

$$d(u, v) = \frac{\sum_i |u_i - v_i|}{\sum_i |u_i + v_i|} \tag{4.4}$$

Principal coordinate analysis (PCoA) is performed on Bray–Curtis distances using `scikit-bio` ordination functions. Horizon plots were implemented with the JavaScript D3 plotting library for each taxon, displaying differences in abundance relative to the previous sample.

Taxonomy–metadata correlations are computed with `pandas` and `scikit-bio` using Pearson (4.5), Kendall (4.6), or Spearman (4.7) correlation coefficients. Here, cov is the covariance, X is a series of metadata values, Y is the series of taxon counts, C and D are the numbers of concordant and discordant pairs, and $R(X)$ is the rank of variable X [207, 208].

$$\rho(X, Y) = \frac{\mathrm{cov}(X, Y)}{\sigma_X \, \sigma_Y} \tag{4.5}$$

$$\tau = \frac{C - D}{C + D} \tag{4.6}$$

$$r_s = \rho_{R(X),R(Y)} = \frac{\mathrm{cov}(R(X), R(Y))}{\sigma_{R(X)}\,\sigma_{R(Y)}} \tag{4.7}$$

Reference database creation MMonitor utilizes customized databases for taxonomic classification. The Database Manager window can be used to change profiler indices or to download and build fresh versions on demand. To save build time for users, we provide pre-built indices online (see data availability section). To create the Centrifuger index, we downloaded all complete genomes of bacteria, archaea, and fungi from the NCBI RefSeq database [209] (as of October 2024) and concatenated them into a single file. We also downloaded the taxonomy tree and taxonomy mapping file from the NCBI taxonomy dump. The centrifuger index was then built using the centrifuger-build command with default parameters, providing the concatenated genomes, taxonomy tree, and mapping file previously downloaded.
For the Emu database we followed the authors' in their methods section Emu 16S database. All 16S rRNA gene sequences for bacteria and archaea from the NCBI targeted loci directory and the rrnDB-5.9 [210] sequences were concatenated into a single file. Then we downloaded the NCBI taxdump and accession-to-taxid mapping file and the mapping file for rrnDB-5.9. We used Emu's build-database and Centrifuger's internal script for building the databases. We did not include the index files in the MMonitor binaries due to their large file size and instead uploaded them to a public repository. If no index is found or provided by the user during run-time, MMonitor will notify the user and try to download it. The scripts for downloading and building the new databases are included in MMonitor and can be run on demand from the Database Manager of the desktop app. You can find the source code and custom index files in the Data Availability section.

While the 16S rRNA and WGS pipelines in MMonitor use different tools and databases, the WGS database is built from complete reference genomes (by default all RefSeq bacterial, archaeal, and fungal genomes). This design enables detection of any marker genes present in these genomes, making the same database creation scripts applicable to marker gene sequencing data, provided the target genes are part of the reference genomes. In future versions, we plan to add an option for users to supply a FASTA file with specific marker gene sequences, allowing the creation of dedicated marker-gene databases for faster querying and targeted monitoring.

16S rRNA gene nanopore sequencing For 16S rRNA gene amplicon sequencing, we used the 16S rRNA gene Barcoding Kit 1-24 (SQK-16S024, Oxford Nanopore Technologies) and an R9.4.1 flow cell (FLO-MIN106). The 16S rRNA gene Barcoding

Kit enables rapid and full-length 16S rRNA gene sequencing for organism identification by using universal primers 27F (5′–AGAGTTTGATCMTGGCTCAG–3′) and 1492R (5′–CGGTTACCTTGTTACGACTT–3′).

For 16S rRNA gene amplification via PCR, we mixed 10 µL of 10 ng DNA with 25 µL of LongAmp® Hot Start Taq 2× Master Mix (New England Biolabs), 10 µL of an individual 16S rRNA gene barcode, and 5 µL of nuclease-free water. PCR cycles were: 1 min at 95 °C; 25 cycles of 20 s at 95 °C, 30 s at 51 °C, 2 min at 65 °C; and a final elongation of 5 min at 65 °C.

Afterward, bead cleaning, preparation of the cleaned-up DNA library with rapid adapter and pooling, and loading of the R9.4.1 flow cell were performed following the manufacturer's protocol. The sequencing time depended on the number of barcodes pooled in one library. We decided to sequence for 2 h per barcode, resulting in reasonable coverage of the 16S rRNA gene for improved classification results. MinKNOW software version 23.11.4 was used for data generation.

Bioelectrochemical system setup (16S rRNA gene sequencing experiments) The BES consisted of two glass chambers (working and counter) separated by a Nafion 117 ion-exchange membrane (Sigma), ensuring that the human fecal sample contacted only the working electrode (E_{we}). A three-electrode configuration (E_{we}, E_{ce}, E_{ref}) was used, with the working electrode poised at +350 mV vs. Ag/AgCl for H_2 removal by oxidation. Both E_{we} and E_{ce} were carbon cloth electrodes (area: 122 cm^2). E_{we} was spray-coated on both sides with platinated carbon (10%) at a loading of 4 mg cm^{-2}. The reference electrode (E_{ref}) was Ag/AgCl (silver wire chloridized, embedded in 3 M KCl and 0.7 g agarose).

Prior to assembly, Nafion 117 membranes were activated in 1 M sulfuric acid for 24 h and equilibrated for 24 h in a salt solution (85.5 mM NaCl, 16.9 mM $Na_2HPO_4{\cdot}2H_2O$, 41.9 mM $NaHCO_3$). Reactors (including the membrane-separated electrodes) were assembled in a sandwich configuration and sterilized at 121 °C for 20 min under aerobic conditions. After autoclaving, chambers were filled with anaerobic, reduced medium and sparged with N_2:CO_2 prior to inoculation. The working volume was 21 mL in both chambers. The geometry minimized the microbe–electrode distance to $\sim$1 mm.

DNA extraction for BES (16S) For DNA extraction, 500 µL of BES liquid was centrifuged (19,000 $\times g$, 5 min, 4 °C) to pellet the cells. DNA was isolated with the AllPrep® PowerFecal Pro DNA/RNA Kit (QIAGEN) using a FastPrep-24™ 5G bead-beater (2 cycles at 6.0 m s^{-1} for 40 s with 30 s rest). We followed the manufacturer's instructions with one exception: <450 µL (instead of 300 µL) supernatant was taken in pretreatment step 6 (after adding CD2), following the vendor's guidance for higher volumes. DNA

quantity was measured with the Qubit™ Flex Fluorometer (Invitrogen) and DNA quality with the NanoPhotometer™ N60/N50 (Implen).

Continuous bioreactors with pertraction system (WGS experiments) Two independently operated, continuously fed upflow bioreactors with integrated liquid–liquid extraction (pertraction) were run in parallel for ∼1019 days. Each bioreactor was a double-walled glass column (height 125 cm) maintained at 30°C and pH 5.5. One vessel contained wheel-shaped plastic packing (anaerobic filter, AF reactor); the second was an upflow anaerobic sludge blanket (UASB) reactor without packing to promote granule formation. Fermentation broth was recirculated through two hollow-fiber membrane contactors: medium-chain carboxylates were extracted into a hydrophobic solvent and subsequently transferred into an alkaline extraction solution. An in-line 0.8 L filter module protected the membranes during early operation.

Medium and inoculum: Both reactors were inoculated with 2 mL glycerol stock (0.04% v/v) from a chain-elongating microbiome previously maintained in a CSTR on ethanol/acetate feed. The defined medium contained ethanol (13.42 g L^{-1}; ∼600 mM C) and acetate (3.12 g L^{-1}; ∼100 mM C) at a 6:1 molar ratio, plus bicarbonate and nutrients.

Operating periods: Over 1019 days, four operating periods (I–IV) were defined by construction and oxygen-management changes: limiting oxygen intrusion (airlocks, N_2 sparging, metal tubing), adjusting reducing agents (L-cysteine, sodium sulfide), and later introducing controlled O_2 supply. Gas additions included N_2/H_2 (95/5% v/v; 0.3 ± 0.1 mL h^{-1}) and air (11.8–17.1 mL h^{-1}) during late operation. These interventions altered the redox balance and *n*-caprylate productivity.

Sampling: Biomass for metagenomics was collected from reactor columns and the filter module at Days 9, 283, 397/455, and 701.

Quantification and statistical analysis

All statistical procedures, including normalized counts, alpha and beta diversity, and taxonomy–metadata associations, are implemented in the MMonitor dashboard and described in detail in the Webserver and Dashboard section (Eqs. 4.1–4.7).

Additional resources

Installation instructions, database management, and remote execution guidelines are available in the MMonitor README at www.github.com/lucast122/mmonitor.

Author Contributions

T.N.L. conceived the main scientific ideas, wrote most of the software, led data analysis and interpretation, and wrote the majority of the manuscript. U.B. generated most of the experimental data and contributed to analysis and writing. A.G. contributed to scientific ideas, analysis, and writing. K.G. assisted with data generation. T.L. and S.K. contributed to scientific ideas and code. R.E.L. provided scientific input. L.T.A. contributed to scientific ideas, interpretation and manuscript writing. D.H.H. contributed to scientific ideas, interpretation, and manuscript writing, and acted as corresponding author. All authors provided feedback and helped improve the manuscript.

ACKNOWLEDGMENTS

The authors thank the Angenent Lab and the Ley Lab for their sequencing data, suggestions, and tool testing. We also acknowledge members of the Algorithms in Bioinformatics labs at the University of Tübingen. Special thanks to Byoung Seung Jeon for building and sequencing the reactor that provided the WGS data. The development of MMonitor was funded through the Deutsche Forschungsgemeinschaft under Germany's Excellence Strategy (EXC 2124 – 390838134) to REL, LTA and, DH. Additional parts of the work were funded by the Alexander von Humboldt Foundation in the framework of the Alexander von Humboldt Professorship to LTA and by the Novo Nordisk Foundation CO2 Research Center (CORC) with grant number NNF21SA0072700 to LTA. AG used the de.NBI Cloud, provided by the German Network for Bioinformatics Infrastructure (de.NBI) and ELIXIR-DE (Forschungszentrum Jülich; projects W-de.NBI-001, W-de.NBI-004, W-de.NBI-008, W-de.NBI-010, W-de.NBI-013, W-de.NBI-014, W-de.NBI-016, W-de.NBI-022), and the High Performance and Cloud Computing Group at the Zentrum für Datenverarbeitung of the University of Tübingen, the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant no INST 37/1159-1 FUGG, to carry out computational analyses for this work.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, ChatGPT-5 (OpenAI) was used to rephrase text, identify typographical and formatting errors, and assist in literature searches. Portions of the Django server (API and request handling), and GUI (webpage html, css & user configuration window) boilerplate code were generated with the assistance of Claude

Sonnet 3.7 (Anthropic) to speed up development of routine components. All content was carefully reviewed and edited by the author, who takes full responsibility for the final manuscript.

Declaration of Interests

The authors declare no conflict of interest.

Chapter 5

Further Collaborative Projects

This chapter includes studies that were performed in collaboration with other researchers. Each section describes a separate study, with the author contributions detailed in a table. The last section describes a study for which no manuscript was written yet, but which is in preparation. For this section the author contributions are detailed in text form at the beginning of the section.

Metagenome Analysis of Chain-Elongating Bioreactor Reveals Caprylate Production Mechanism

This section includes parts from a manuscript written by multiple authors. Author contributions are detailed in the table below. Timo N. Lucas performed metagenomics analysis, helped with data analysis, interpretation and writing.

Title of paper:	Toward industrial C8 production: oxygen intrusion drives renewable n-caprylate production from ethanol and acetate *via* intermediate metabolite production
Status in publication process:	Published in Green Chemistry (DOI: https://doi.org/10.1039/D5GC00411J)

Author	Author position	Scientific ideas (%)	Data generation (%)	Analysis & interpretation (%)	Paper writing (%)
Kurt Gemeinhardt	First	50	35	65	65
Byoung Seung Jeon	Second	20	20	0	0
Jean Nepomuscene Ntihuga	Third	0	10	0	0
Han Wang	Fourth	0	0	0	0
Caroline Schlaiß	Fifth	0	5	0	5
Timo N. Lucas	Sixth	0	15	10	5
Irina Bessarab	Seventh	0	0	0	0
Nicolas Nalpas	Eighth	0	10	10	5
Nanqing Zhou	Ninth	0	5	0	5
Joseph G. Usack	Tenth	0	0	0	0
Daniel H. Huson	Eleventh	0	0	0	5
Rohan B. H. Williams	Twelfth	0	0	0	0
Boris Maček	Thirteenth	0	0	0	0
Ludmilla Aristilde	Fourteenth	10	0	5	5
Largus T. Angenent	Corresponding	20	0	10	5

Toward industrial C8 production: oxygen intrusion drives renewable n-caprylate production from ethanol and acetate *via* intermediate metabolite production

Kurt Gemeinhardt[a], Byoung Seung Jeon[ab], Jean Nepomuscene Ntihuga[a], Han

Wang[a], Caroline Schlaiß[ac], Timo N. Lucas[d], Irina Bessarab[e], Nicolas Nalpas[f], Nanqing Zhou[g], Joseph G. Usack[ah], Daniel H. Huson[d], Rohan B. H. Williams[i], Boris Maček[f], Ludmilla Aristilde[ag], Largus T. Angenent[ajk,*]

[a]Environmental Biotechnology Group, Department of Geosciences, University of Tübingen, Schnarrenbergstraße 94-96, 72076 Tübingen, Germany. E-mail: l.angenent@uni-tuebingen.de
[b]Biomaterials and Processing Center, Korea Institute of Ceramic Engineering and Technology, 202 Osongsaengmyeong 1-ro, 28160 Osong, Republic of Korea
[c]AG Angenent, Max Planck Institute for Biology, Max Planck Ring 5, 72076 Tübingen, Germany
[d]Algorithms in Bioinformatics, Department of Computer Science, University of Tübingen, Sand 14, 72076 Tübingen, Germany
[e]Integrative Analysis Unit, Singapore Centre for Environmental Life Sciences Engineering, National University of Singapore, 28 Medical Drive, 117456 Singapore, Singapore
[f]Proteome Center Tübingen, University of Tübingen, Auf der Morgenstelle 15, 72076 Tübingen, Germany
[g]Department of Civil and Environmental Engineering, Northwestern University, 2145 Sheridan Road, Evanston, 60208 Illinois, USA
[h]Department of Food Science and Technology, University of Georgia, 100 Cedar Street, Athens, 30602 Georgia, USA
[i]Department of Biological and Chemical Engineering, Aarhus University, Gustav Wieds vej 10D, 8000 Aarhus C, Denmark
[j]The Novo Nordisk Foundation CO_2 Research Center (CORC), Aarhus University, Gustav Wieds vej 10C, 8000 Aarhus C, Denmark
[k]Cluster of Excellence – Controlling Microbes to Fight Infections, University of Tübingen, Auf der Morgenstelle 28, 72074 Tübingen, Germany
[*]Correspondence: l.angenent@uni-tuebingen.de

Abstract

Previous bioreactor studies achieved high volumetric n-caprylate (i.e., n-octanoate) production rates and selectivities from ethanol and acetate with chain-elongating microbiomes. However, the metabolic pathways from the substrates to n-caprylate synthesis were unclear. We operated two n-caprylate-producing upflow bioreactors with a synthetic medium to study the underlying metabolic pathways. The operating period exceeded 2.5 years, with a peak volumetric n-caprylate production rate of 190 ± 8.4 mmol C L^{-1} d^{-1} (0.14 g L^{-1} h^{-1}). We identified oxygen availability as a critical perfor-

mance parameter, facilitating intermediate metabolite production from ethanol. Bottle experiments in the presence and absence of oxygen with ^{13}C-labeled ethanol suggest acetyl-coenzyme A-derived production of n-butyrate (i.e., n-butanoate), n-caproate (i.e., n-hexanoate), and n-caprylate. Here, we postulate a trophic hierarchy within the bioreactor microbiomes based on metagenomics, metaproteomics, and metabolomics data, as well as experiments with a *Clostridium kluyveri* isolate. First, the aerobic bacterium *Pseudoclavibacter caeni* and the facultative anaerobic fungus *Cyberlindnera jadinii* converted part of the ethanol pool into the intermediate metabolites succinate, lactate, and pyroglutamate. Second, the strictly anaerobic *C. kluyveri* elongated acetate with the residual ethanol to n-butyrate. Third, *Caproicibacter fermentans* and *Oscillibacter valericigenes* elongated n-butyrate with the intermediate metabolites to n-caproate and then to n-caprylate. Among the carbon chain-elongating pathways of carboxylates, the tricarboxylic acid cycle and the reverse β-oxidation pathways showed a positive correlation with n-caprylate production. The results of this study inspire the realization of a chain-elongating production platform with separately controlled aerobic and anaerobic stages to produce n-caprylate renewably as an attractive chemical from ethanol and acetate as substrates.

Study Background

In this joint collaborative study, the results of the previous bioreactor study described in Chapter 3 were extended by running two additional bioreactors with similar chain-elongating microbial communities. The previous study determined the microbial community and possible metabolic pathways for the production of n-caprylate from ethanol and acetate. However, the exact mechanisms of how the microbial community achieved the high production rates and selectivities were not fully understood. The role of aerobic organisms in the bioreactor, like *Pseudoclavibacter caeni*, which had high abundances in the previous study, was not clear, especially since most known chain elongators are strictly anaerobic. The main goal of this study was to enhance our knowledge of how exactly those communities produce n-caprylate by integrating metagenomics and metaproteomics data with the results of various experiments conducted in the lab.

Contributions to Study

Computational analyses, including shotgun metagenomic and parts of the metaproteomic analysis, identifying key microbial players and metabolic pathways.

Methods

Non-computational methods are described in [211]. After sequencing, basecalling was performed using Guppy basecaller v6.4.6. The resulting FASTQ files were aligned against the NCBI-nr database [209] (downloaded February 2023) using DIAMOND BLASTX v2.1.6 [28] with the optional parameters `-c 1 -b 12 -outfmt 100 -long-reads -p 64`. The resulting DAA files were `meganized` using the daa-meganizer included in MEGAN [143] v6.24.20 with the parameters `-lcp 51 -lg` and the newest mapping file available from February 2022. The meganized files were loaded into MEGAN in comparison mode, and the aligned bases were normalized for sample size. The microbial abundances (aligned bases) for all taxa in the samples were exported and visualized accordingly. To obtain references for the metaproteomic analysis, the metagenome sequences were assembled using Flye [60] (v2.9.2), the resulting contigs were corrected with Medaka [22] (v1.11.3) and annotated using Prokka [212] (v1.14.6). To increase the annotation rate, proteins not classified by Prokka were aligned against the NCBI-nr (February 2023) using DIAMOND BLASTP (v2.1.6) with the parameters `-c 1 -b 12 -outfmt 100 -p 64`, and the taxonomy of the best hits was extracted using the daa2info tool of MEGAN6.

Key Results

The study found a diverse microbial community in the bioreactors, consisting of Bacteria, Archaea, and Fungi, with a mixture of aerobic and anaerobic organisms, as shown in Figure 5.1. The metagenomic WGS data showed four microorganisms that were abundant throughout reactor operation with a significant positive correlation to n-caprylate production: *P. caeni*, *C. fermentans*, *O. valericigenes*, and *M. aggregans*. There were also two microorganisms that showed a significant negative correlation to n-caprylate production: *C. kluyveri* and a fungal species of the genus *Mortierella*. The metaproteomic data showed that the RBOX pathway was the most abundant pathway in the bioreactors, that the TCA cycle was most positively correlated with n-caprylate production, and supported the hypothesis that the fatty acid biosynthesis pathway was not related to n-caprylate production. By integrating metagenomics and metaproteomics data with further experiments, we proposed that a trophic hierarchy exists in the bioreactors, where aerobic organisms such as *Pseudoclavibacter caeni* and *Cyberlindnera jadinii* first produce intermediate metabolites, including succinate, lactate, and pyroglutamate, and then the anaerobic bacteria *O. valericigenes* and *C. fermentans* used those metabolites to produce n-caprylate from n-butyrate.

In conclusion, the metagenomic analysis resolved the taxonomy of the microbial communities and was used to generate de novo assemblies that were annotated to generate

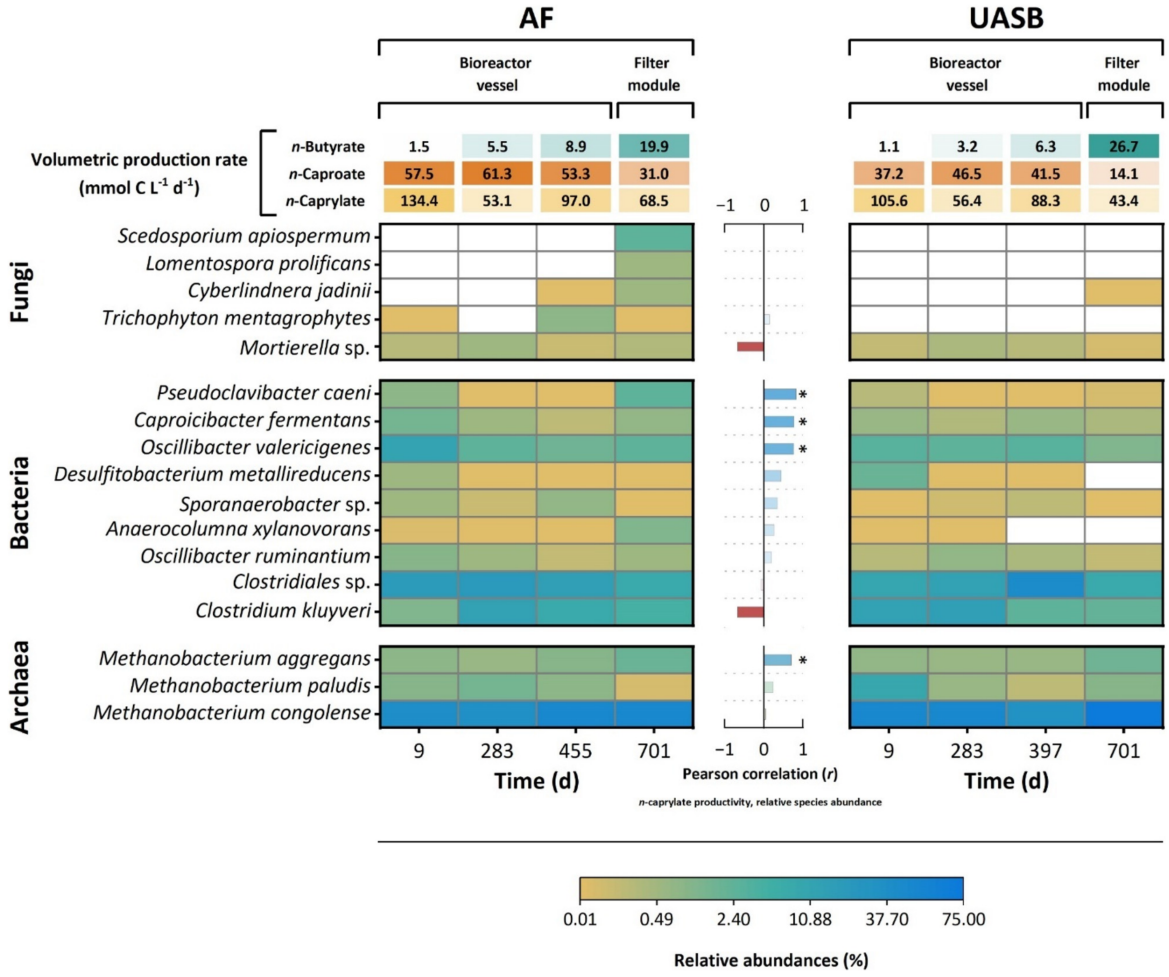

Figure 5.1: The volumetric production rates of n-butyrate, n-caproate, and n-caprylate of the bioreactors during the sampling days are shown at the top. Pearson correlation values between volumetric n-caprylate production rate and relative species abundance are displayed as a bar graph between the heatmaps. Light blue indicates positive correlation, white indicates no correlation, and red indicates negative correlation. Critical values are marked with an asterisk. Illustration by Kurt Gemeinhardt et al. [211].

proteins. These annotations were used in conjunction with mass spectrometry data to generate a metaproteomic dataset. Further experiments were conducted to propose a model for n-caprylate production, which was validated by the metagenomic and metaproteomics data. The computational analyses of the omics data were a key contribution to the study, as they provided the basis for the proposed model and allowed for testing the model, increasing the confidence in the results.

Application of MMonitor: Monitoring nasal microbiome dynamics

This section includes parts from a manuscript written by multiple authors. Author contributions are detailed in the table below. Timo N. Lucas performed software development, data processing, analysis and helped with data interpretation and writing.

Title of paper:	Development of a continuous bioreactor to maintain stable nasal microbiomes from swab specimens and synthetic communities
Status in publication process:	In preparation

Author	Author position	Scientific ideas (%)	Data generation (%)	Analysis & interpretation (%)	Paper writing (%)
Soyoung Ham	First	25	25	20	25
Marcelo Navarro-Diaz	Second	5	10	15	10
Laura Camus	Third	5	10	15	10
Timo N. Lucas	Fourth	5	10	10	10
Hannes Link	Fifth	10	5	5	5
Daniel Petras	Sixth	10	5	5	5
Simon Heilbronner	Seventh	10	5	5	5
Daniel H. Huson	Eighth	5	5	5	5
Largus T. Angenent	Corresponding	25	25	20	25

Development of a continuous microbiome bioreactor system for culturing stable nasal communities

Soyoung Ham[1,2], Marcelo Navarro-Diaz[2,3,4], Laura Camus[2,3], Timo N. Lucas[2,5], Hannes Link[2,3,4], Daniel Petras[2,6], Simon Heilbronner[2,3,7,8], Daniel H. Huson[2,5], and Largus T. Angenent[1,2,9,10,11,*]

[1]Department of Geosciences, University of Tübingen, 72076 Tübingen, Germany
[2]Cluster of Excellence - Controlling Microbes to Fight Infections, University of Tübingen, 72076 Tübingen, Germany
[3]Interfaculty Institute of Microbiology and Infection Medicine, University of Tübingen, 72076 Tübingen, Germany
[4]The M3 Research Center, University of Tübingen, 72076 Tübingen, Germany

[5]Institute for Bioinformatics and Medical Informatics, University of Tübingen, 72076 Tübingen, Germany
[6]Department of Biochemistry, University of California Riverside, CA 92507, USA
[7]Department of Microbiology, University of München, 82152 München, Germany
[8]German Centre for Infection Research, 72076 Tübingen, Germany
[9]AG Angenent, Max Planck Institute for Biology, 72076 Tübingen, Germany
[10]Department of Biological and Chemical Engineering, Aarhus University, 8000 Aarhus C, Denmark
[11]The Novo Nordisk Foundation CO_2 Research Center, Aarhus University, 8000 Aarhus C, Denmark
[*]Correspondence: largus.angenent@uni-tuebingen.de

Abstract

In this joint collaborative study, we initially designed and operated a nasal microbiome bioreactor. We inoculated nasal swabs collected from healthy volunteers into the bioreactor and changed various operational parameters (i.e., operation mode, dilution rate, temperature, pH, and medium). We determined an optimal operating condition with stable microbiome growth and complex communities consisting of main nasal bacteria: continuous mode, 1 d^{-1}, 30℃, pH 6.5, and SNM 3. The optimized bioreactor showed high reproducibility and resilience through replicates and pH perturbation, in terms of optical density, copy numbers of total bacteria and *S. aureus*, and bacterial community. Moreover, all bioreactor microbiomes inoculated with nasal swabs from six volunteers reached the stationary phase and maintained their stabilization over a long time. Statistically significant metabolites also showed stabilized metabolic profiles similar to the bacterial communities.

Study Background

The nasal microbiome represents a complex and specialized microbial ecosystem that serves as the first line of defense against respiratory pathogens. This community plays a crucial role in human health by preventing colonization by opportunistic pathogens, particularly *Staphylococcus aureus*, which is a leading cause of hospital-acquired infections. Current models for studying nasal microbiomes include cell cultures, animal models, and *ex vivo* tissue samples, but these approaches often lack stability over time or fail to capture the true complexity of the human nasal microbiome.

This study aimed to develop and characterize the first continuous bioreactor system specifically designed to cultivate stable nasal microbial communities over extended periods. Such a system would provide a valuable research platform for investigating

colonization resistance mechanisms, studying microbe-microbe interactions in a controlled environment, and testing potential decolonization strategies for pathogens like *S. aureus*. The bioreactor system was designed to mimic the physiological conditions of the human nasal cavity while allowing precise control of environmental parameters, including temperature, pH, and nutrient availability.

Nasal swabs from six healthy volunteers were used to inoculate individual bioreactors, which were then operated in continuous mode for over 60 days. Throughout the operational period, the system was monitored for community stability and composition using multiple analytical approaches, including nanopore sequencing of the 16S rRNA gene and qPCR. The previously described MMonitor software was used to analyze the temporal dynamics of the bioreactor communities to validate the stability of the communities over time.

Methods

The methods relevant to this dissertation focus on the application of MMonitor to the 16S rRNA gene sequencing data. Other methods related to the study are described in the main text [213]. After finishing sequencing, the barcoded data was loaded into MMonitor (v0.2.0) using the multi-input CSV to specify sample metadata and resolve the barcodes. The taxonomy-16S pipeline was then run to generate the taxonomic profiles for each sample using Emu [27] as the profiler with the standard Emu database from 2023. Filtering for host DNA was not necessary, as the used database does not contain human sequences and thus, by default, discards reads that would align to the human genome. The results were first controlled and analyzed using the MMonitor dashboard, and then raw and normalized counts and abundances, as well as diversity metrics of the samples, were exported using the MMonitor dashboard export functions of the taxonomy and diversity tabs to be used for visualization using custom scripts.

Key Results

Figure 5.2 shows the bacterial composition of the bioreactors over time. The analysis revealed that the nasal microbiomes differed significantly between volunteers and comprised a mix of different microbes from different genera, including *Staphylococcus*, *Cutibacterium*, *Corynebacterium*, *Salmonella*, *Bacillus*, *Escherichia*, *Citrobacter*, and more. All bioreactor communities reached a stable state after a certain time of operation, some faster than others; e.g., the community from volunteer 6 took around 12 days to reach a stable state, while the community from volunteer 5 was already dominated by *Proteus mirabilis* after 2 days of operation and only changed slightly after that. Metabolomic data showed similar profiles for the bioreactors as the metagenomic

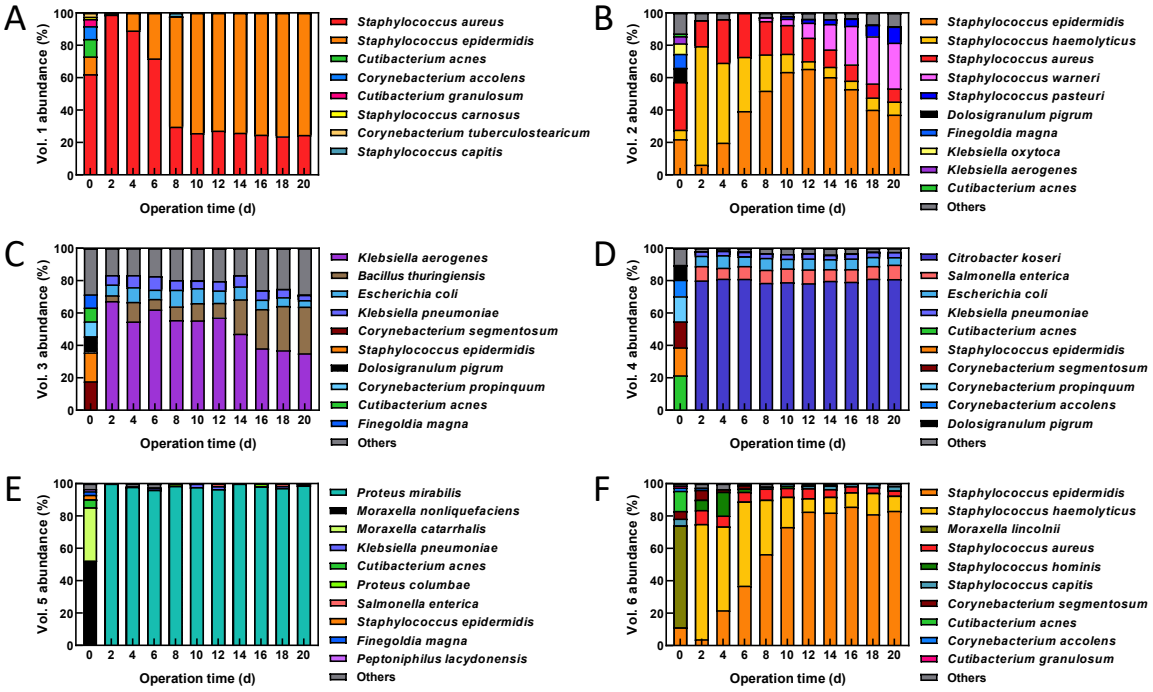

Figure 5.2: Bacterial composition of bioreactors inoculated with nasal swabs from different volunteers throughout operation time. Relative abundances of bioreactor communities from volunteers (A) 1, (B) 2, (C) 3, (D) 4, (E) 5, and (F) 6. The composition of the initial inoculum is shown at operation time 0. Figure created by Soyoung Ham [213].

data. The study demonstrated how bioreactors can be used to model stable nasal microbiomes and that repeated 16S rRNA gene sequencing enables MMonitor to assess the stability of the communities over time. The underlying questions also led to the development of new features in MMonitor, such as the export functions for the taxonomic and diversity tabs.

Application of MMonitor: Quality Control for AI Training Data

This section includes parts from a manuscript written by multiple authors. Author contributions are detailed in the table below. Timo N. Lucas helped with data processing of raw data and writing.

Title of paper:	NanoGraph: Mapping Nanopore Squiggles to Graphs Enables Accurate Taxonomic Assignment
Status in publication process:	Submitted in 2025

Author	Author position	Scientific ideas (%)	Data generation (%)	Analysis & interpretation (%)	Paper writing (%)
Wenhuan Zeng	First/Corresponding	20	10	50	60
Dorian Rollin	Second	20	10	0	0
Yihua Liu	Third	0	20	0	30
Timo N. Lucas	Fourth	20	20	0	0
Anupam Gautam	Fifth	20	10	10	0
Kurt Gemeinhardt	Sixth	0	10	0	0
Largus T. Angenent	Seventh	0	10	0	0
Ruth E. Ley	Eighth	0	10	0	0
Daniel H. Huson	Corresponding	20	0	40	10

NanoGraph: Mapping Nanopore Squiggles to Graphs Enables Accurate Taxonomic Assignment

Wenhuan Zeng[1,*], Dorian Rollin[2], Yihua Liu[3], Timo N. Lucas[1], Anupam Gautam[1], Kurt Gemeinhardt[4], Largus T. Angenent[4], Ruth E. Ley[3], Daniel H. Huson[1,*]

[1]Algorithms in Bioinformatics, Institute for Bioinformatics and Medical Informatics, University of Tübingen, Sand 14, 72076 Tübingen, Germany

[2]Department of Bioinformatics, University Claude Bernard Lyon 1, 43 Bd du 11 Novembre 1918, 69100, Villeurbanne, France

[3]Max Planck Institute for Biology Tübingen, Max-Planck-Ring 5, 72076, Tübingen, Germany

[4]Environmental Biotechnology Group, Department of Geosciences, University of Tübingen, Schnarrenbergstraße 94-96, 72076, Tübingen, Germany

[*]Corresponding authors: wenhuan.zeng@uni-tuebingen.de and daniel.huson@uni-

tuebingen.de

Abstract

Nanopore sequencing technology offers long sequencing reads and real-time analysis capabilities, making it a powerful tool for addressing diverse questions in the life sciences. This technology detects electronic raw signals from samples, which are converted into nucleotide sequences (A, T, G, and C) through a process known as basecalling. These sequences can subsequently be used for various types of analysis. To enhance the efficiency of taxonomic classification in Nanopore sequencing and explore the challenges of applying deep learning algorithms to ultra-long sequences, we developed NanoGraph, which is a graph-based deep learning framework designed to classify samples based on their taxonomic lineages. NanoGraph processes raw signals (of substantial length) by transforming them into topological graph structures. We evaluated NanoGraph's performance using a customized simulated dataset and benchmarked it against a previous study on public datasets, demonstrating superior results. Additionally, we assessed its practical usability after fine-tuning the trained model on real raw signal datasets generated in our wet lab. In summary, NanoGraph provides a robust and effective approach for taxonomic classification of Nanopore-sequenced samples. Python source code related to this study is freely available at: https://github.com/husonlab/nano-graph.

Project Background

This joint collaborative project aimed to develop a novel taxonomic classification method using raw nanopore electrical signals ("squiggles") instead of basecalled sequences. The rationale behind this was that basecallers are developed continuously, which leads to many different versions of basecallers being available, each potentially with different biases. The approach of NanoGraph promises faster, basecaller-independent analysis but requires high-purity bacterial isolate data for training the underlying graph neural network models. A possible use case could be in monitoring food or water samples for known pathogens and other areas where speed is critical. Contamination in training data could severely impair model accuracy. Therefore, rigorous quality control of isolates prior to signal extraction was essential.

Methods

Methods relevant for this dissertation focus on the generation of the training dataset for NanoGraph. Other methods can be found in the main text [214].

We constructed the datasets by simulating raw signals from nucleotide sequences sequenced by ONT using DeepSimulator [215], thus avoiding the expense of downloading large FAST5 files. The classification task focuses on distinguishing samples from five bioreactor-associated species: *Methanothermobacter thermautotrophicus* (M. thermautotrophicus), *Methanobrevibacter smithii* (M. smithii), *Clostridium kluyveri* (C. kluyveri), *Clostridium ljungdahlii* (C. ljungdahlii), and *Pseudoclavibacter caeni* (P. caeni). Genome FASTA files for these species were obtained from the NCBI RefSeq database [216] with the following assembly IDs: GCF_027554905.1, GCF_000016525.1, GCF_000010265.1, GCF_000143685.1, and GCF_008831125.1, each containing 10,000 reads. Using DeepSimulator V1.5 with default parameters, we generated FAST5 files for each genome. We used one FAST5 file per read, obtaining an output of 10,000 files per species. Each binary FAST5 file was then converted into a sequence of discrete numbers separated by commas. As a result, the 10,000 FAST5 files for each species were transformed into 10,000 rows, each representing a sequence of discrete numbers. This process yielded an initial dataset of 50,000 rows across the five species. The initial dataset was further divided into a training set and an independent test set in an 8:2 ratio.

The prepared libraries were loaded onto R10.4.1 flow cells (Oxford Nanopore Technologies) and sequenced using the MinION Mk1B (Oxford Nanopore Technologies) controlled by MinKNOW (23.11.4, Oxford Nanopore Technologies). Sequencing was performed with simultaneous super-accurate basecalling at 400 bases per second, using a minimum Q score of 10 and a minimum read length of 200 bp. The purity of the lab cultures was confirmed using the basecalled sequences *via* MMonitor's taxonomic WGS pipeline, using Centrifuger [195] as profiler and an index built from all RefSeq genomes downloaded in December 2024 (https://www.mmonitor.org). The POD5 files from pure cultures were used for model training. Data processing: The POD5 files were converted to FAST5 files and processed using the same methods previously described for handling FAST5 files from simulated raw signals. These files were subsequently mapped to graph data for use in the NanoGraph framework. Overall, we obtained 21,000 samples in the processed raw signal datasets, with 7,000 samples per species. We further split it into training and test datasets with a ratio of 8:2; the training dataset contains 16,800 samples, 5,600 samples per species, and the test dataset contains 4,200 samples, 1,400 samples per species.

Key Results

Figure 5.3 shows how NanoGraph was trained on the signals simulated from the NCBI RefSeq sequences from the five species and evaluated on real signals from the wet

lab. The wet lab signals were basecalled using Dorado (v0.5.2), and the resulting reads were fed into MMonitor to screen for contamination, keeping only sequences that were classified as the target species. Evaluation of NanoGraph and comparison with a previous study on a public dataset showed that NanoGraph can accurately predict the presence of the target species from the raw signals. The study also demonstrated MMonitor's versatility as a quality control tool when pure isolate data is required for a study.

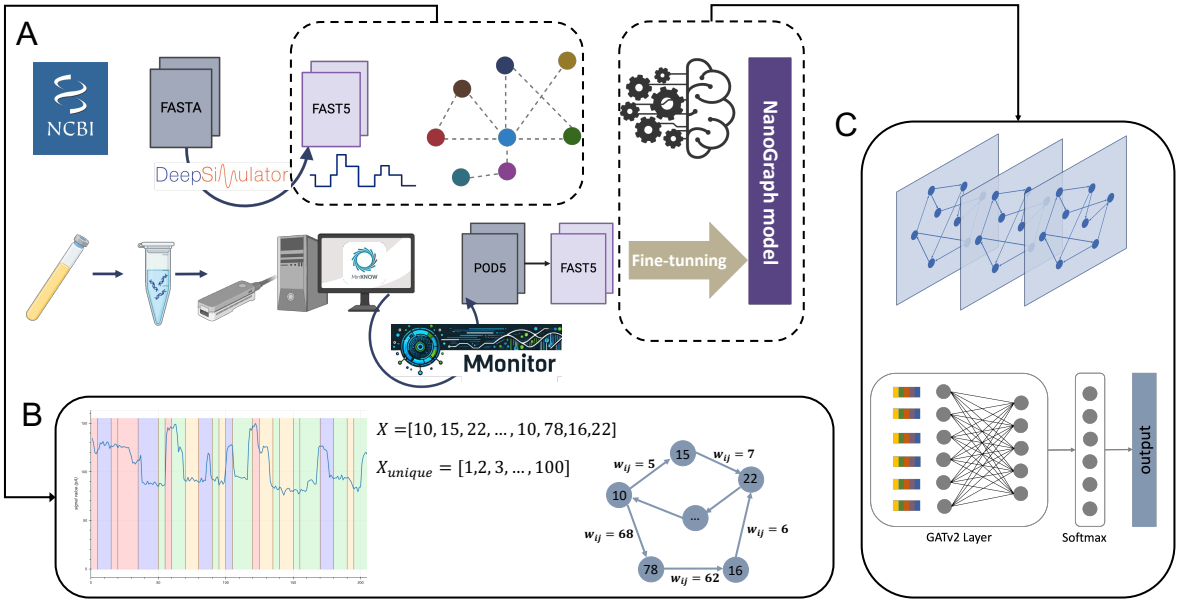

Figure 5.3: NanoGraph: A Graph-Based Framework for Nanopore Raw Signal Classification. (A) NanoGraph is trained on simulated datasets, where raw signals were generated from NCBI RefSeq sequences across five species. Its performance was compared with that of a previous study using a public dataset and further evaluated on a real signal dataset collected in the wet lab through transfer learning. (B) Each Nanopore raw signal is represented as a weighted directed graph, where nodes correspond to unique numbers in the signal, directed edges indicate the order of numbers, and edge weights represent differences between neighboring numbers. (C) The NanoGraph model utilizes GATv2 layers to capture complex hidden patterns in graphs. Once trained, NanoGraph can accurately classify raw signal data derived from two or more species. Figure created by Wenhuan Zeng et al. [214].

GeneGone: Web Application for CRISPR Deletion Validation

This section includes parts of an ongoing study performed by multiple authors. Author contributions are detailed in the table below. Timo N. Lucas performed software development, data analysis and helped with data interpretation and writing.

Title of paper:	A host-directed virulence factor of Clostridium perfringens is modulated by gut commensal strains
Status in publication process:	In preparation

Author	Author position	Scientific ideas (%)	Data generation (%)	Analysis & interpretation (%)	Paper writing (%)
Julia Schumacher	First	35	70	60	40
Paolo Stincone	Second	0	15	5	5
Johanna Rapp	Third	0	10	5	5
Timo Niklas Lucas	Fourth	5	0	10	5
Carlos Llaca-Bautista	Fifth	0	5	0	5
Lisa Maier	Sixth	15	0	0	5
Hannes Link	Seventh	5	0	0	5
Daniel Petras	Eighth	5	0	0	5
Bastian Molitor	Corresponding	35	0	20	25

Molecular responses of Clostridium perfringens to the presence of gut commensal bacteria

Julia Schumacher[1,2], Paolo Stincone[2,3], Johanna Rapp[2,4], Timo N. Lucas[2,5], Carlos Llaca-Bautista[1], Lisa Maier[2,6,7], Hannes Link[2,4], Daniel Petras[2,3,8], Bastian Molitor[1,2,9,*]

[1]Environmental Biotechnology Group, Department of Geosciences, University of Tübingen, Germany
[2]Cluster of Excellence EXC 2124: Controlling Microbes to Fight Infection (CMFI), University of Tübingen, Germany
[3]Organismic Interactions, Interfaculty Institute of Microbiology and Infection Medicine, Tübingen, Germany
[4]Bacterial Metabolomics, Interfaculty Institute of Microbiology and Infection Medicine,

University of Tübingen, Germany
[5]Algorithms in Bioinformatics, Department of Computer Science, University of Tübingen, Germany
[6]Interfaculty Institute for Microbiology and Infection Medicine Tübingen, University of Tübingen, Germany
[7]M3-Research Center for Malignome, Metabolome and Microbiome, University of Tübingen, Germany
[8]Department of Biochemistry, University of California Riverside, USA
[9]Microbial Metabolic Biochemistry, Institute of Biochemistry, University of Leipzig, Germany
[*]Correspondence: bastian.molitor@uni-leipzig.de

Abstract

Clostridium perfringens is a common member of the human gut microbiome but can induce diseases in the host under certain conditions. *C. perfringens*-associated diarrhea is often self-regulating but severe cases and other *C. perfringens*-induced infections are typically treated with a combination of antibiotics and surgical debridement. The overuse of antibiotics in our society is becoming problematic as antibiotic resistance traits are increasing, leading to a growing number of antibiotic resistance-associated deaths. Furthermore, not only the increasing resistance traits are problematic, but also the fact that antimicrobial drugs are often not pathogen-specific and target commensal bacteria in addition. This can induce dysbiosis and diseases. Microbiome modulation offers a promising approach to overcome these challenges. Understanding the interspecies interactions between microbes within a microbiota holds significant importance in elucidating how a microbiome can be modulated precisely and effectively to benefit the host's well-being. Here, we combined qPCR, proteome, and metabolome analyses to investigate the interactions between the pathobiont *C. perfringens* and human gut commensals on physiological and molecular levels. Genetic modification of *C. perfringens* enabled deeper insight into the function of proteins that we identified as relevant in co-cultures via the proteome analysis. We found that several commensal species inhibit *C. perfringens* growth and were involved in regulating its energy metabolism. Especially, four Bacteroidaceae strains not only modulated *C. perfringens* carbon metabolism but also its toxin production. This understanding serves as a starting point for the development of new therapeutic strategies that can make use of the potential of the microbiome to confer health benefits to the host.

Project Summary

CRISPR-Cas9 gene editing enables targeted gene deletions and validating the success of these deletions through sequencing is a common step in the gene editing workflow. The validation, however, requires analysis of sequencing data, by using multiple bioinformatics command line tools and scripting. As we didn't find any existing tools optimized for validating these deletions automatically, we developed GeneGone in a joint collaboration, an easy-to-use web application for deletion validation based on read coverage analysis. The tool allowed our collaborators to validate the deletion of a gene in a *Clostridium perfringens* mutant strain, as part of a larger project and its quick implementation using the Streamlit framework demonstrated its usefulness for rapid prototyping and deployment of bioinformatics tools in a lab setting.

Key Results

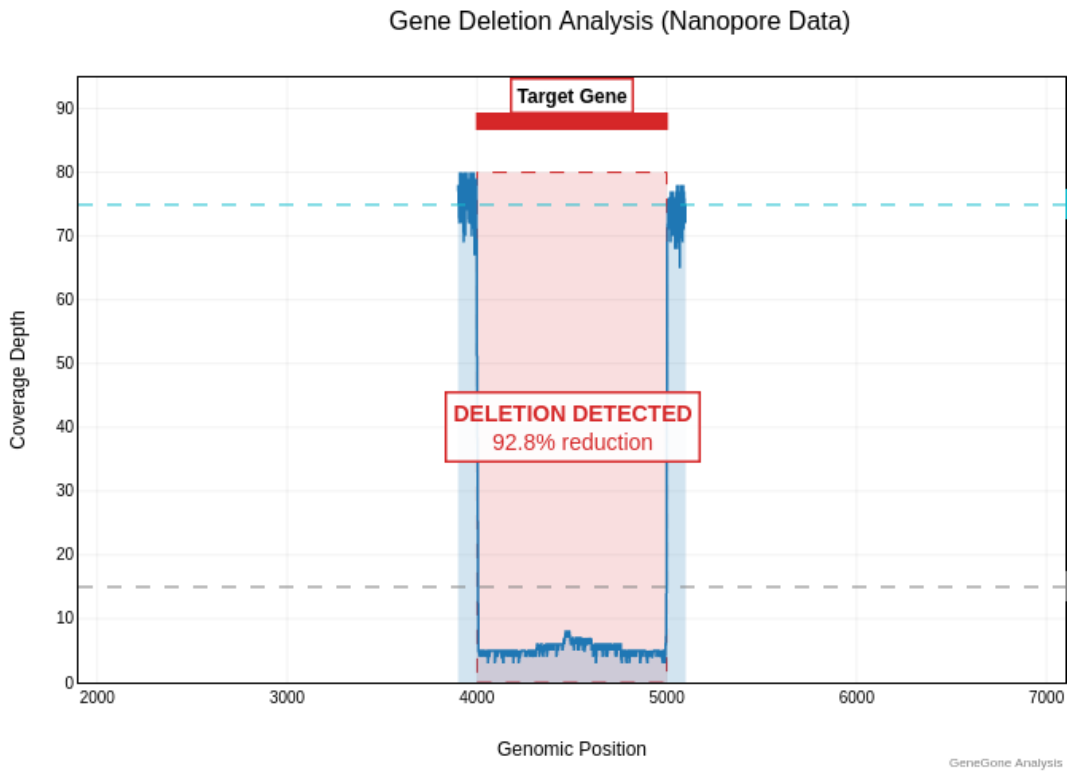

Figure 5.4: Deletion validation through coverage analysis. The plot shows the coverage of the aligned reads across the target gene area. A strong drop in coverage at the target gene region indicates a successful deletion.

GeneGone provides a minimalistic solution for deletion validation by requiring the user to upload a reference genome of the wild type strain and a fasta file containing reads of the targeted genome, as well as the approximate position of the edited gene in the target genome. Through alignment of the reads to the reference and coverage analysis

it is able to detect if the deletion was successful at the targeted area. If the deletion was successful, a strong reduction in the coverage at this area can be observed and the tool notifies the user by providing visualization of the coverage together with summary statistics. We tested the tool first on simulated data and then on a real nanopore dataset of two *Clostridium perfringens* wild type and mutant strains, that we sequenced using the ONT MinION platform. Figure 5.4 illustrates how GeneGone detects a successful deletion using example test data (not the actual experimental data). In this example, a gene at position 4000-5000 was deleted, as indicated by a 92.8 percent reduction in coverage at this region compared to the flanking regions.

Additionally, we aligned the full mutant genome against the full wild-type genome and the gene sequence that was deleted against both the wild-type and the mutant genomes. The results showed that the deleted gene had no alignment in the mutant genome and a perfect match in the wild-type genome at position 2,595,336. The genome-genome alignment showed an almost perfect match between the two genomes, only broken once by a gap caused by the successful deletion, indicating that no other detectable edits were made to the genome.

Methods

The *Clostridium perfringens* wild type and mutant genomes were sequenced using nanopore sequencing on the ONT MinION platform. The wild type and mutant genomes were assembled using Flye [60] (v2.9.1) and the deletion was validated using GeneGone. GeneGone was written in Python and uses the Streamlit framework for the web interface. Behind the scenes, GeneGone runs Minimap2 [25] (v2.28-r1209) for aligning long pacbio and nanopore reads, or BWA [24] (v0.7.17) for aligning short illumina reads, and samtools [217] (v1.17) for calculating the coverage.

To confirm that the gene deletion had taken place, the nucleotide sequence of the deleted gene was aligned against the corrected assembly of the previously sequenced mutant genome using Minimap2 (v2.28-r1209) with the -x map-ont preset (optimized for short sequence alignment). The same gene sequence was also aligned to the wild-type genome using the same Minimap2 options. The full mutant genome was aligned to the full wild-type chromosome using Minimap2 with the -x asm5 preset, which is suitable for aligning similar genome assemblies.

GeneGone is free open source software and the code and documentation are available on GitHub at https://github.com/lucast122/GeneGone.

Comparison of Python web frameworks in a lab setting

In Chapter 4 we introduced MMonitor, which includes a web application based on the Django framework. Django is a powerful framework, well suited for building scalable and feature-rich web applications. It has integrated support for databases, APIs, user authentication, and more, and allows to build apps following a Model-View-Controller architecture. This allows developers to create maintainable and scalable web applications, but setup can be time consuming as it requires a lot of boilerplate code. For building a feature that lets users upload data in Django one first needs to write URL routing (HTTP endpoint), a view function (backend logic) and a template (frontend), define a model to store the data, setup a webserver for deployment, and more. This can be unnecessary for small applications that only require a simple interface to upload data, run a few tools and visualize the results as it often is with specialized bioinformatics workflows.

The Streamlit framework was originally developed for sharing machine learning applications, but given Python's rich bioinformatics ecosystem, it is also useful for providing bioinformatics tools to anyone with a web browser, only writing Python code. It comes with an integrated webserver which is useful for rapid deployment of applications to the local network or the web in a matter of minutes. Applications can also be deployed to the free community cloud service with a click of a button if they're on a public Github repository. This allows collaborators to test tools and provide feedback almost immediately, resulting in a quicker implementation of all required features. Streamlit allows to write high level Python code to generate web applications, by using template components, that can be declared as Python variables. Eliminating the need for time-consuming front and backend development, Streamlit is a good choice for building small, specialized applications quickly in Python, without excessive boilerplate-code or requiring writing any HTML, CSS or JavaScript. To build the data upload feature described above in Streamlit it is enough to include a file uploader component that saves the uploaded files to a dataframe.

This is especially useful in lab settings, where biological questions need to be answered quickly and in a reproducible way, allowing for future repetition of an analysis with new data. By building and deploying tools in this way, programmers and users can work more closely together, better bridging the gap between code and biology. Researchers without programming experience can interact with intuitive interfaces, while developers receive fast feedback on usability and functionality. As a result this can help research institutions to rapidly transform analysis scripts or notebooks into shareable, interactive tools that enhance collaboration, reproducibility, and accessibility of computational analyses.

QuickBinDM: Accelerating Long-Read Metagenomic Binning

This section is based on the preliminary results of an ongoing study, which has not been published yet. A manuscript is in preparation, but requires further work. Sequencing was performed by Kurt Gemeinhardt, software development, data analysis and writing was done by Timo N. Lucas.

Project Summary

Metagenomic binning, the process of reconstructing individual genomes (MAGs) from mixed microbial sequence data, is crucial for linking function to taxonomy. Traditional methods face challenges with ONT data as most binners do not account for frameshift errors, which are common in assemblies from long reads due to sequencing errors. DIAMOND+MEGAN provides frameshift correction, but can be slow for large datasets, as it requires alignment of the contigs against a complete protein reference database like NCBI-nr. This makes binning large metagenomes a computationally demanding task, to be performed only on high performance systems. However, access to those systems is not always possible for all researchers and some projects may require frequent metagenome assembly and binning, e.g. in the context of monitoring. The goal of this project was to develop a pipeline that could be used for binning metagenomic contigs, while being fast and requiring fewer resources. For this, we developed QuickBinDM, a pipeline that combines DIAMOND [28]+MEGAN [143] for taxonomic metagenome binning and frameshift correction, with fast ANI-based pre-screening using skANI [218]. This filtering reduces the number of proteins that need to be compared with the contigs, dramatically reducing the time it takes to bin the contigs. We used QuickBinDM to bin a bioreactor metagenome sequencing using nanopore data, with the pre-screening and reduced protein database, QuickBinDM reduced the time it took for binning a set of metagenome contigs from several hours to several minutes compared to the normal DIAMOND+MEGAN-LR binning pipeline, while creating the same amount of high quality MAGs.

Key Results

We applied QuickBinDM to a bioreactor metagenome sequencing from the study described at the beginning of this chapter once with filtering and the dynamically created protein database, and once without filtering and using the complete protein database. The dataset contained 2689 contigs with an N50 of 35,445 bp and a total length of approximately 54 mbp.

Table 5.1: Overall comparison of two QuickBinDM runs on the Reactor_2_Filter assembly. The "Dynamic DB" run used a Skani-filtered reference subset, whereas the "Full-nr" run searched the complete NCBI-nr database downloaded on February 14th, 2023. High quality is defined as completeness $\geq 90\,\%$ and contamination $\leq 5\,\%$, high completeness is defined as completeness $\geq 90\,\%$, medium completeness is defined as completeness $\geq 70\,\%$ and low completeness is defined as completeness $< 70\,\%$, regardless of contamination. Comp. = Completeness, Cont. = Contamination.

Metric	**Dynamic DB**	**Full NCBI-nr**
Runtime	17.5 min	65.8 h
MEGAN bins (#)	70	244
Contigs binned	2 566	2 497
Bases binned	54.29 Mbp	53.05 Mbp
High-quality MAGs	1	1
High-completeness ($\geq$90 %)	5	3
Medium-completeness (70-90 %)	2	6
Low-completeness (<70 %)	63	235
Bins with Cont. $\leq$5 % (top 10)	2	2
Mean contamination (top 10)	31.1 %	17.9 %
Largest bin size	11.65 Mbp	5.17 Mbp
Best bin (Comp. / Cont.)	99.97 % / 4.45 %	99.94 % / 0.58 %

Table 5.1 shows a comparison of the two runs. Without filtering, the QuickBinDM behaves like the normal DIAMOND+MEGAN binning, where the user has to provide the database for DIAMOND. Using the full NCBI-nr the pipeline took 65.85 hours to bin the dataset, while using the filtering step the full pipeline took 17.47 minutes, a speedup of 226x. It is important to note that we ran both tools on a machine with a 32 core CPU and 512GB of RAM, so the speedup may be even higher on machines with fewer resources, as limited memory could be a bottleneck for the full nr run, requiring more IO operations to load the reference database during alignment. The run using the filtered database created 70 bins, while the full-nr run created 244 bins, however most of those 244 bins had low completeness and the binning run with the full database used fewer contigs and bases for the binning overall. Both runs created one high quality MAG, the filtered run created more high complete bins, but the high complete bins had higher contamination, sometimes exceeding 100%, indicating strain mixing. The full nr run had a lower mean contamination of 17.9%, compared to the filtered run with 31.1%.

Looking at the top 10 bins in Table 5.2, we can see that the top bins are similar between the two runs, but the filtered run has more high complete bins, while the full nr run has less contamination in the top bins. The filtered run captured a complete genome of the order Eubacteriales (synonym for Clostridiales), that the full run did not capture, but the reduced database run also assigned fewer species names to the bins. In summary, the filtered database results are comparable to the full nr run, but the contamination of the

Table 5.2: Top ten bins from each run ranked by completeness (CheckM2). Values are completeness / contamination (%).

Rank	Dynamic DB	Rank	Full NCBI–nr
1	*Eubacteriales* 100 / 102.3	1	*M. congolense* 99.9 / 10.1
2	*M. congolense* 99.9 / 10.4	2	*M. paludis* 99.9 / 0.6
3	*Methanobacterium* 99.9 / 4.5	3	*uncl. Clostridium* 91.1 / 14.7
4	*Clostridium* 99.8 / 26.3	4	*Eggerthellaceae* 89.8 / 23.3
5	*Oscillospiraceae* 98.4 / 102.3	5	*Oscillibacter* 82.7 / 52.6
6	*Ethanoligenens h.* 76.9 / 24.1	6	*C. kluyveri* 80.5 / 5.4
7	*C. kluyveri* 75.8 / 13.5	7	*Clostridiales bacterium* 78.4 / 0.3
8	*Firmicutes* 63.2 / 18.0	8	*Firmicutes* 77.2 / 21.8
9	*C. tetani* 52.3 / 5.8	9	*Eubacteriales* 74.9 / 37.1
10	*Caproicibacterium a.* 35.6 / 3.7	10	*Oscillospiraceae* 57.7 / 13.1

bins could be a problem for some use cases.

As QuickBinDM uses RefSeq genomes for the database, there is a lack of proteins from different strains of the same species, which can lead to over-binning of closely related species. This could be mitigated by using a different database, or by including a step that adds additional strain-specific proteins to the database after the filtering step. Overall, the results show that using a filter before binning can yield reasonable results and drastically reduce the runtime of the binning process, while still allowing for frameshift correction unlike other binning tools with comparable speed. More thorough benchmarking is needed to evaluate the performance of QuickBinDM on a wider range of datasets including an in-depth analysis of the quality of the bins and the use of mock communities to better evaluate the accuracy of the binning, but our proof of concept shows that ANI pre-screening can be a good addition to the binning process, especially if computational resources are limited. QuickBinDM is available as a tool on GitHub at https://github.com/lucast122/QuickBinDM.

Methods

The methods related to sequencing and assembly can be found in the published study [211]. We evaluated the performance of QuickBinDM using the Filter2 dataset of said study only. QuickBinDM requires skANI, DIAMOND, MEGAN and CheckM2 to be installed on the machine running the pipeline. For testing we used skANI [218] (v0.2.2) for the ANI pre-screening, DIAMOND [28] (v2.1.11.165) for the protein alignment, MEGAN [143] (v6.25.10) for binning and frameshift correction of the contigs and CheckM2 [69] (v1.1.0) for the quality assessment of the bins. For running the pipeline with the full nr database we used the default parameters, but provided the full nr database to DIAMOND with the parameter '–use-db' as well as the MEGAN mapping

file with '–megan-map'. For running the pipeline with the filtered database we used a 95% ANI cutoff with '-a 95', the top 10 hits per contig based on the ANI score with '–max-hits-per-contig 10' and a minimum query coverage of 60% with '-q 60'. As a reference for the ANI screening step we used all bacterial, archaeal, and fungal RefSeq genomes from NCBI downloaded on November 16th, 2023.

Chapter 6

Discussion and Future Directions

Synthesis of Key Findings and Contributions

Here, we summarize the key findings and contributions from the research presented in this dissertation and discuss potential implications for the field and future work. First, we summarized existing literature about biotechnology and metagenomics and presented the current state-of-the-art in metagenomic analysis using short- and long-read sequencing technologies.

Then, we presented extensive analyses of metagenomic data that revealed complex microbial communities and provided insight into the metabolic processes involved in microbial chain elongation of medium-chain fatty acids. We demonstrated how to apply metagenomics to study individual microbes' roles in the context of reverse β-oxidation, gaining insights that can now be used to build industrial platforms for sustainable medium-chain fatty acid production.

Furthermore, our metagenome analyses provided the basis for multiple short- and long-read metagenomic pipelines, which we applied to a variety of bioreactor communities, resulting in two published papers that expanded our knowledge of chain-elongating microbiomes. We uploaded our datasets to public repositories, making them available to other researchers.

Extending the long-read pipelines into MMonitor, a tool for real-time metagenomic analysis, we created a new platform, allowing researchers to study the complex temporal dynamics of microbial communities in real-time while automating complex analysis pipelines and downstream analyses. Beyond the original MMonitor publication, we further demonstrated its versatility by applying it to human nasal microbiomes and showed that it can be used to detect contamination in bacterial isolates.

Furthermore, our work contributed to the development of methods for faster binning of metagenomic genomes with the DIAMOND+MEGAN pipeline, a tool for validating deletion mutants after CRISPR editing, and a machine learning-based method for predicting the presence of microbes in a sample from raw nanopore signals.

Insights into chain-elongating microbial communities

Our computational analysis of bioreactor communities showed how short- and long-read metagenomic data can be used to characterize microbial communities both taxonomically and functionally. In the past, broad usage of short-read sequencing technologies gave rise to a large number of algorithms and mature pipelines for taxonomic and functional analysis. In Chapter 3, we presented how short-read data can be used to classify microbial communities at the species level. Correlating the taxonomic composition with metabolic processes provided insight into key species involved in chain elongation of medium-chain fatty acids. Further analysis of key genes involved in reverse β-oxidation and fatty acid biosynthesis from the metagenomes and metaproteomes helped to characterize the genetic potential of individual microbes in the context of chain elongation, enabling us to validate hypotheses about the roles of individual microbes in the bioreactor communities.

In a follow-up study, we used long-read nanopore sequencing to generate high-quality genomes from both metagenomes and bacterial isolates, showing improvements in the quality and contiguity of the genomes compared to the study that relied on short-read data, resulting in more complete annotations and complementing the mass spectrometry data for the metaproteomic analysis.

While both short- and long-read data can be used to generate annotations and call variants, short-read assemblies were often highly fragmented and contained large numbers of contigs. In contrast, long-read assemblies often yielded complete genomes with a single contig. However, higher contiguity comes with the drawback of increased sequencing error rates of the reads. This can possibly introduce frameshift errors in the annotations and other methods that rely on translated sequences, reducing their accuracy. To mitigate frameshift errors, one could use MEGAN's built-in frameshift correction; however, this can be computationally intensive for large metagenomes. To reduce this computational burden, we developed QuickBinDM, which achieves higher speed by using skANI pre-screening to reduce the search space, ultimately trading off some accuracy for speed when binning metagenome assemblies. There also exist other sequencing error and frameshift correction methods that can be used to improve the sequencing quality, and new basecalling models are being developed to further reduce the error rate, so frameshift errors might become less of an issue in the future

when using long-read sequencing.

While our analyses highlighted the importance of rigorous quality control and error correction when dealing with long-read data, they also showed that long-read sequencing can yield high-quality genomes without needing additional short-read data. This is demonstrated by the fact that our long-read assemblies yielded several high-quality MAGs, using only assembly, correction, and binning. While not necessary for our use case, those assemblies could have been even further improved by combining multiple assemblies and binning procedures into a consensus assembly.

Our investigations into chain-elongating microbiomes revealed complex microbial communities with distinct ecological roles in medium-chain fatty acid production. Through metagenomic and metaproteomic analyses of multiple reactor systems over extended operating periods, we identified key microbial players and metabolic pathways that drive n-caprylate production.

A significant finding from our studies was the critical role of oxygen in driving n-caprylate production. While chain elongation is traditionally considered a strictly anaerobic process, our work demonstrated that controlled oxygen availability creates a beneficial trophic hierarchy that enhances n-caprylate yields. This was evident in both our three-reactor study, where oxygen intrusion affected performance, and in our follow-up investigation, where we confirmed oxygen's role in n-caprylate production through stable-isotope tracing experiments.

Taxonomic analyses consistently identified specific microbial players across different reactor systems. The aerobic bacterium *Pseudoclavibacter caeni* showed a strong positive correlation with n-caprylate production rates, despite its inability to produce medium-chain fatty acids directly. We discovered that *P. caeni* performs crucial pre-conversion of ethanol into intermediate metabolites, including succinate, lactate, and pyroglutamate. These intermediates serve as more effective substrates for chain elongation than ethanol alone, especially at high ethanol concentrations that can inhibit anaerobic processes. The extensive analyses of the meta-omics data, combined with experimental efforts, ultimately led to the proposal of a new model to explain n-caprylate production in these microbial communities. This model can now be tested and further evaluated in future studies, hopefully bringing us closer to the goal of efficient and renewable industrial-scale medium-chain fatty acid production.

Development of MMonitor and its Applications

In Chapter 4, we presented MMonitor, a tool for real-time metagenomic analysis, and we want to highlight some of the key features and discuss the design choices we made.

To make use of the real-time data, we designed the software to be able to handle incoming sequencing data in batches, iteratively updating taxonomic analyses as more sequencing data is generated. We achieved this by integrating fast taxonomic profilers that we adapted for the purpose of monitoring.

A common issue with metagenomics tools is that they can be complicated to install or use for people without bioinformatics expertise. We tackled this by providing a simple graphical user interface and packaging the software so it can be run locally without installation. We also integrated a command-line interface for more experienced users that also allows use on remote systems like high-performance clusters. We decided on a client-server architecture, where the client executes analysis pipelines and the server displays the results in a web interface. This design allows researchers to locally start their analysis and remotely view the results without the need for time-consuming downstream analysis. However, if desired, the software can run fully locally without sending data to an external server when using the offline mode. This might be important for sensitive data or when there is limited access to the internet.

The client provides multiple pipelines that work out-of-the-box to analyze long-read metagenomic or isolate data, but the user still has control over the parameters of the pipelines using a configuration window.

The server automates common downstream analysis steps and generates high-quality visualizations that can be exported, e.g., for use in publications. The visualizations are interactive, and besides the common plots like stacked bar charts and heatmaps, we also included plots useful for time-series analysis. For example, the horizon charts provide a compact overview of the taxonomic composition over time, and the diversity app informs the user about changes in diversity over time. We did not want to bind the user to our downstream analyses, so we designed the dashboard to allow the export of the raw data used for visualization. We applied this feature to the nasal microbiome study, where our collaborators used the dashboard to obtain the raw data to extend the analysis with their own scripts.

Our hosted instance of the web server encrypts all data via HTTPS and makes use of Nginx and Gunicorn to serve the web application in a secure and scalable way. If needed, the user can also self-host the web server on their own infrastructure, giving them full control over their data without relying on the offline mode.

As MMonitor depends on other bioinformatics methods that themselves depend on external databases, we integrated the ability to update the databases used for taxonomic profiling if desired by the user. The user can either provide their own indices for Centrifuger or Emu or use the database manager to download the latest sequences

from NCBI, building new indices from scratch. This feature is important to make the software easier to maintain as databases are updated and allows the user to benefit from the ever-growing amount of sequencing data available.

Beyond industrial applications, we have demonstrated MMonitor's versatility in clinical microbiology and quality control workflows. Our collaboration on the nasal microbiome bioreactor system showcases MMonitor's ability to track complex microbial dynamics in a controlled environment mimicking the human nasal cavity. Using 16S rRNA gene sequencing, we achieved strain-level resolution, enabling detailed monitoring of community shifts in order to assess when a community reached a stable state. This application is particularly valuable for studying colonization resistance against pathogens like *Staphylococcus aureus*, which has implications for developing decolonization strategies in clinical settings.

Additionally, we demonstrated MMonitor's utility for contamination detection in bacterial isolates, where rapid and accurate identification of contaminants can serve as a quality control measure. By using nanopore sequencing data, MMonitor can detect and quantify unexpected organisms in supposedly pure cultures within minutes, serving as a quality control tool for both research and industrial microbiology. These diverse applications highlight MMonitor's adaptability beyond its original design, on which we will continue to work. This capability proved especially valuable in our work on the NanoGraph project, where ensuring the purity of bacterial isolates was essential for generating high-quality data to test the machine learning model.

Advancements in Metagenomic Methodologies

Our work also contributed to methodological advancements beyond the MMonitor platform itself. The development of the DIAMOND+MEGAN-LR pipeline, optimized within QuickBinDM, demonstrated significant acceleration for taxonomic binning from long reads by incorporating skANI pre-screening. This addressed a key bottleneck in our long-read analysis workflows that required metagenomic assembly and binning. Furthermore, the GeneGone web application provides an accessible tool for validating CRISPR-Cas9 gene deletions using sequencing coverage analysis, simplifying a common experimental workflow for molecular biologists.

Future Research Directions and Conclusions

Our research has progressed along two paths: developing advanced computational tools for metagenomic analysis and applying these tools to understand complex microbial

communities across different environments. We have demonstrated this approach in studying chain elongation processes and, more recently, in a nasal microbiome study. These interconnected research areas represent significant opportunities for future work, with advancements in one area directly benefiting the other. Real-time metagenomic monitoring enables deeper insights into microbial dynamics, while the specific requirements of analyzing specialized microbiomes inform the development of more targeted analytical tools. Below, we explore future directions for these research areas and their potential synergies.

Our results demonstrated that computational analysis of metagenomic data provides valuable insights into microbial communities. We successfully automated metagenomics analysis to examine microbial communities in real-time and developed a foundational tool that researchers can utilize and extend. Nevertheless, numerous challenges and opportunities remain for future research in metagenomics, as answering important biological questions remains complex, with no universal pipeline applicable to all scenarios.

The rapidly evolving nature of bioinformatics and metagenomics presents ongoing challenges. New databases are continuously created while existing ones undergo frequent updates, requiring continuous software development to keep up. Beyond software considerations, hardware developments, such as sequencing technologies and computing infrastructure, continue to evolve and must be accounted for in future research. To keep pace with this rapidly changing field, continuous updates to existing methods and the development of new approaches are essential. Researchers could test new clustered databases for taxonomic profiling and other approaches, like AI-based methods.

Given these constant changes, the integration of community-maintained pipelines and databases represents a promising direction. Combining these resources with rigorous testing protocols could facilitate the development of more maintainable software solutions. We implemented standard pipelines and gave users the option to customize them, though the software still needs further development, especially for more specialized use cases. We will expand on the configurability of the graphical interface to reach a similar level of customization that is currently available by manually using command-line tools. It is important to note, however, that automatic analysis will never reach the quality of manual, human expert analysis, and automated approaches should not be seen as a replacement for human expertise but rather as convenient tools to accelerate routine tasks. MMonitor's automatic database update functionality represents progress in this direction, though new databases require thorough testing and validation before integration. We anticipate exploring alternative databases for taxonomic profiling and investigating AI-based methodologies in future work.

While our taxonomic profiling pipelines operate effectively on consumer-grade hardware, metagenome assembly and binning remain computationally intensive processes. Since basecalling already requires substantial GPU resources, utilizing these to accelerate pipelines represents another potential improvement. As long-read sequencing technology inevitably advances, new opportunities for metagenomic analysis and monitoring will emerge, which we look forward to implementing in future updates. Given that benchmarking studies showed that long-read profilers based on alignment are very accurate, it may be useful to make them faster, e.g., by including pre-screening as we did in QuickBinDM. Given that k-mer-based methods are very fast but struggle with high error rates, researchers could try to use concepts like strobemers or syncmers for taxonomic profiling in a similar way that minimap uses minimizers. An algorithm that utilizes the speed of k-mer-like methods, but optimized for erroneous reads, to find candidate matches in combination with accurate alignments could be a good compromise between speed and accuracy.

Metagenomics has advanced rapidly in recent years, enabling real-time tracking of microbiomes through DNA sequencing. However, much work remains to be done before we have tools that can accurately and consistently estimate taxonomic composition. In addition, there is a strong need for improved functional analysis and interpretation tools to illuminate the microbial dark matter and the complex metabolic networks within microbial communities across the planet.

Despite the enormous volume of metagenomic data now available, we are still only scratching the surface of the biological insights it can offer. The number of microbial genomes being sequenced far exceeds our current capacity in terms of both tools and researchers to analyze and interpret them. Future research must focus not only on improving computational methods, but also on building platforms that help scientists make sense of this rapidly expanding microbial world.

We believe that this work contributed to advancing our understanding of microbial communities and our ability to monitor them in real-time. The tools and methodologies described here provide a foundation for future research in computational metagenomics, especially when the accurate measurement of temporal microbial dynamics is necessary. With a strong focus on real-world applications due to the collaborative nature of the projects, we hope that the tools and concepts evaluated here will be useful to other researchers.

Acknowledgments

I would like to express my heartfelt gratitude to everyone who made this journey possible and supported me along the way.

First and foremost, I want to thank my supervisor, Prof. Dr. Daniel H. Huson, for giving me this opportunity. Daniel, your patience with my endless questions, your thoughtful feedback, and our many insightful discussions have really helped me along the way. Thank you for being such a supportive and understanding supervisor, even when things didn't go according to plan.

I'm deeply grateful to my second supervisor, Prof. Dr. Largus T. Angenent, for welcoming me into the lab and teaching me so much about the experimental side of science. Your enthusiasm for research is contagious, and your feedback has been invaluable throughout this work.

A huge thank you to the former lab members who took me under their wing in the beginning: Sascha Patz and Dr. Caner Bagci, who taught me the ropes and helped me navigate the world of bioinformatics. Thanks also to Dr. Benjamin Albrecht, Dr. Monika Zeller, Dr. Xi Chen and Dr. Wenhuan Zeng, for all the knowledge you shared with me.

To my current lab family Dr. Anupam Gautam, Banu Cetinkaya, and Julia Fischer. You've made this journey so much more enjoyable. Thank you for all the coffee breaks, the brainstorming sessions, and for creating such a wonderful working atmosphere.

I've been incredibly fortunate to work with amazing collaborators around the world. Dr. Ulrike Biehain, our collaboration has been fantastic, and I've learned so much from our discussions and your feedback. Thanks also to Prof. Dr. Rohan B. H. Williams, Dr. Irina Bessarab, and Dr. Byoung Seung Jeon for bringing your expertise to our joint projects.

A special thanks to Prof. Dr. Catherine Spirito for the great collaboration that opened up new research directions for me.

I'm grateful to Kurt Gemeinhardt, Dr. Soyoung Ham, Dr. Julia Schumacher, and Caroline

Schlaiß for the wonderful collaboration and for sharing your data with me. Working with you has been both productive and fun.

Thank you to Simon Konzalla, Tobias Lass, and Emilio Diebold for their contributions to the software projects as part of your thesis work. Also thank you to all the other students who contributed through their work.

To all the anonymous reviewers and editors who helped improve our work, thank you for your time and constructive feedback.

Finally, I want to thank my wife, family and friends for putting up with me during the stressful times and celebrating with me during the good ones. This wouldn't have been possible without your support.

This work was funded through the Deutsche Forschungsgemeinschaft under Germany's Excellence Strategy (EXC 2124 – 390838134).

Bibliography

[1] Wan Abd Al Qadr Imad Wan-Mohtar, Zul Ilham, and Neil J. Rowan. “Editorial: The value of microbial bioreactors to meet challenges in the circular bioeconomy”. In: *Frontiers in Bioengineering and Biotechnology* 11 (2023), p. 960420. DOI: 10.3389/fbioe.2023.960420.

[2] Kalbm Kodithuwakku, Mahinda Senevirathne, and Se-Kwon Kim. “Application of Bioreactor Technology for the Food Industry”. In: *Bioreactor Technology in Food Processing*. CRC Press, pp. 183–211.

[3] Hiroyuki Imachi et al. “Cultivable microbial community in 2-km-deep, 20-million-year-old subseafloor coalbeds through 1000 days anaerobic bioreactor cultivation”. In: *Scientific Reports* 9 (2019), p. 2305. DOI: 10.1038/s41598-019-38754-w.

[4] Largus T. Angenent et al. “Chain elongation with reactor microbiomes: Open-culture biotechnology to produce biochemicals”. In: *Environmental Science & Technology* 50.6 (2016), pp. 2796–2810. DOI: 10.1021/acs.est.5b04847.

[5] Richa Bharti and Dominik G. Grimm. “Current challenges and best-practice protocols for microbiome analysis”. In: *Briefings in Bioinformatics* 22.1 (2021), pp. 178–193. DOI: 10.1093/bib/bbz155.

[6] Muzaffer Arıkan and Thilo Muth. “Integrated multi-omics analyses of microbial communities: a review of the current state and future directions”. In: *Molecular Omics* 19.5 (2023), pp. 607–623. DOI: 10.1039/D3MO00089C.

[7] Peter Rubbens and Ruben Props. “Computational analysis of microbial flow cytometry data”. In: *mSystems* 6.1 (2021), e00895–20. DOI: 10.1128/mSystems.00895-20.

[8] Victoria de la Sovera et al. “Microbial community evolution in a lab-scale reactor operated to obtain biomass for biochemical methane potential assays”. In: *Applied Microbiology and Biotechnology* 108.1 (2024), p. 519. DOI: 10.1007/s00253-024-13305-0.

[9] Stefanie Widder et al. “Challenges in microbial ecology: building predictive understanding of community function and dynamics”. In: *The ISME Journal* 10 (2016), pp. 2557–2568. DOI: 10.1038/ismej.2016.45.

[10] Stilianos Louca et al. “Function and functional redundancy in microbial systems”. In: *Nature Ecology & Evolution* 2 (2018), pp. 936–943. DOI: 10.1038/s41559-018-0519-1.

[11] Jay T. Lennon et al. “Principles of seed banks and the emergence of complexity from dormancy”. In: *Nature Communications* 12 (2021), p. 4807. DOI: 10.1038/s41467-021-25137-5.

[12] Sofia Esquivel-Elizondo et al. “The isolate Caproiciproducens sp. 7D4C2 produces n-caproate at mildly acidic conditions from hexoses: genome and rBOX comparison with related strains and chain-elongating bacteria”. In: *Frontiers in microbiology* 11 (2021), p. 594524.

[13] Pieter Candry et al. “Enrichment and characterisation of ethanol chain elongating communities from natural and engineered environments”. In: *Scientific reports* 10.1 (2020), p. 3682.

[14] Jose Antonio Magdalena, Silvia Greses, and Cristina González-Fernández. “Impact of organic loading rate in volatile fatty acids production and population dynamics using microalgae biomass as substrate”. In: *Scientific reports* 9.1 (2019), p. 18374.

[15] MJ Gonçalves, C González-Fernández, and Silvia Greses. “Long hydraulic retention time mediates stable volatile fatty acids production against slight pH oscillations”. In: *Waste Management* 176 (2024), pp. 140–148.

[16] S. Andrews. *FastQC: a quality control tool for high throughput sequence data*. en. Cambridge, United Kingdom: Babraham Bioinformatics, Babraham Institute, 2010.

[17] Shifu Chen et al. “fastp: an ultra-fast all-in-one FASTQ preprocessor”. In: *Bioinformatics* 34.17 (2018), pp. i884–i890. DOI: 10.1093/bioinformatics/bty560.

[18] Ryan Wick. *Filtlong: quality filtering tool for long reads*. https://github.com/rrwick/Filtlong. 2019.

[19] Wei Shen et al. “SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation”. In: *PLOS ONE* 11.10 (2016), e0163962. DOI: 10.1371/journal.pone.0163962.

[20] Marcel Martin. “Cutadapt removes adapter sequences from high-throughput sequencing reads”. In: *EMBnet.journal* 17.1 (2011), pp. 10–12. DOI: 10.14806/ej.17.1.200.

[21] Bruce J. Walker et al. “Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement”. In: *PLoS ONE* 9.11 (2014), e112963. DOI: 10.1371/journal.pone.0112963.

[22] Oxford Nanopore Technologies. *Medaka: sequence correction using neural networks (software)*. https://github.com/nanoporetech/medaka. 2019.

[23] Ben Langmead and Steven L. Salzberg. “Fast gapped-read alignment with Bowtie 2”. In: *Nature Methods* 9.4 (2012), pp. 357–359. DOI: 10.1038/nmeth.1923.

[24] Heng Li and Richard Durbin. “Fast and accurate short read alignment with Burrows–Wheeler transform”. In: *Bioinformatics* 25.14 (2009), pp. 1754–1760. DOI: 10.1093/bioinformatics/btp324.

[25] Heng Li. “Minimap2: pairwise alignment for nucleotide sequences”. In: *Bioinformatics* 34.18 (2018), pp. 3094–3100. DOI: 10.1093/bioinformatics/bty191.

[26] Daniel H. Huson et al. “MEGAN analysis of metagenomic data”. In: *Genome Research* 17.3 (2007), pp. 377–386. DOI: 10.1101/gr.5969107.

[27] Kristen D. Curry et al. “Emu: species-level microbial community profiling of full-length 16S rRNA Oxford Nanopore sequencing data”. In: *Nature Methods* 19.7 (2022), pp. 845–853. DOI: 10.1038/s41592-022-01520-4.

[28] Benjamin Buchfink, Chao Xie, and Daniel H. Huson. “Fast and sensitive protein alignment using DIAMOND”. In: *Nature Methods* 12.1 (2015), pp. 59–60. DOI: 10.1038/nmeth.3176.

[29] Philipp Menzel, Kim Lee Ng, and Anders Krogh. “Fast and sensitive taxonomic classification for metagenomics with Kaiju”. In: *Nature Communications* 7 (2016), p. 11257. DOI: 10.1038/ncomms11257.

[30] Nicola Segata et al. “Metagenomic microbial community profiling using unique clade-specific marker genes”. In: *Nature Methods* 9.8 (2012), pp. 811–814. DOI: 10.1038/nmeth.2066.

[31] Derrick E. Wood, Jennifer Lu, and Ben Langmead. “Improved metagenomic analysis with Kraken 2”. In: *Genome Biology* 20.1 (2019), p. 257. DOI: 10.1186/s13059-019-1891-0.

[32] Daehwan Kim et al. “Centrifuge: rapid and sensitive classification of metagenomic sequences”. In: *Genome Research* 26.12 (2016), pp. 1721–1729. DOI: 10.1101/gr.210641.116.

[33] Daniel M. Portik, C. Titus Brown, and N. Tessa Pierce-Ward. “Evaluation of taxonomic classification and profiling methods for long-read shotgun metagenomic sequencing datasets”. In: *BMC Bioinformatics* 23.1 (2022), p. 541. DOI: 10.1186/s12859-022-05103-0.

[34] Robert Edgar. “Syncmers are more sensitive than minimizers for selecting conserved k-mers in biological sequences”. In: *PeerJ* 9 (2021), e10805.

[35] Kristoffer Sahlin. “Effective sequence similarity detection with strobemers”. In: *Genome Research* 31.11 (2021), pp. 2080–2094. DOI: 10.1101/gr.275648.121.

[36] Jeremy Fan, Steven Huang, and Samuel D. Chorlton. “BugSeq: a highly accurate cloud platform for long-read metagenomic analyses”. In: *BMC Bioinformatics* 22.1 (2021), p. 160. DOI: 10.1186/s12859-021-04089-5.

[37] Daniel H. Huson et al. “MEGAN-LR: new algorithms allow accurate binning and interactive exploration of metagenomic long reads and contigs”. In: *Biology Direct* 13.6 (2018), p. 6. DOI: 10.1186/s13062-018-0208-7.

[38] Josip Marić et al. “Comparative analysis of metagenomic classifiers for long-read sequencing datasets”. In: *BMC Bioinformatics* 25.1 (2024), p. 15. DOI: 10.1186/s12859-024-05634-8.

[39] Javier Tamames, Marta Cobo-Simón, and Fernando Puente-Sánchez. “Assessing the performance of different approaches for functional and taxonomic annotation of metagenomes”. In: *BMC genomics* 20 (2019), pp. 1–16.

[40] Shaopeng Liu et al. “Analysis of metagenomic data”. In: *Nature Reviews Methods Primers* 5.1 (2025), p. 5.

[41] Zhenyu Li et al. “Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph”. In: *Briefings in functional genomics* 11.1 (2012), pp. 25–37.

[42] Phillip EC Compeau, Pavel A Pevzner, and Glenn Tesler. “Why are de Bruijn graphs useful for genome assembly?” In: *Nature biotechnology* 29.11 (2011), p. 987.

[43] Bahlul Haider et al. “Omega: an overlap-graph de novo assembler for metagenomics”. In: *Bioinformatics* 30.19 (2014), pp. 2717–2722.

[44] Chao Yang et al. “A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data”. In: *Computational and structural biotechnology journal* 19 (2021), pp. 6301–6314.

[45] Toshiaki Namiki et al. “MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads”. In: *Proceedings of the 2nd ACM conference on bioinformatics, computational biology and biomedicine*. 2011, pp. 116–124.

[46] Yu Peng et al. “IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth”. In: *Bioinformatics* 28.11 (2012), pp. 1420–1428.

[47] D. Li et al. “MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph”. en. In: *Bioinformatics* 31 (2015), pp. 1674–1676.

[48] Sergey Nurk et al. “metaSPAdes: a new versatile metagenomic assembler”. In: *Genome research* 27.5 (2017), pp. 824–834.

[49] Gleb Goussarov et al. “Benchmarking short-, long- and hybrid-read assemblers for metagenome sequencing of complex microbial communities”. In: *Microbiology (Reading, Engl.)* 170.6 (2024), p. 001469. DOI: 10.1099/mic.0.001469.

[50] Mikhail Kolmogorov et al. “Assembly of long, error-prone reads using repeat graphs”. In: *Nature biotechnology* 37.5 (2019), pp. 540–546.

[51] Sergey Koren et al. “Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation”. In: *Genome research* 27.5 (2017), pp. 722–736.

[52] Heng Li. “Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences”. In: *Bioinformatics* 32.14 (2016), pp. 2103–2110.

[53] Robert Vaser and Mile Šikić. “Time-and memory-efficient genome assembly with Raven”. In: *Nature Computational Science* 1.5 (2021), pp. 332–336.

[54] Jue Ruan and Heng Li. “Fast and accurate long-read assembly with wtdbg2”. In: *Nature methods* 17.2 (2020), pp. 155–158.

[55] Jason R Miller et al. “Aggressive assembly of pyrosequencing reads with mates”. In: *Bioinformatics* 24.24 (2008), pp. 2818–2824.

[56] Ryan R Wick and Kathryn E Holt. “Benchmarking of long-read assemblers for prokaryote whole genome sequencing”. In: *F1000Research* 8 (2021), p. 2138.

[57] Haoyu Cheng et al. “Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm”. In: *Nature methods* 18.2 (2021), pp. 170–175.

[58] Xiaowen Feng et al. “Metagenome assembly of high-fidelity long reads with hifiasm-meta”. In: *Nature methods* 19.6 (2022), pp. 671–674.

[59] Gaëtan Benoit et al. “High-quality metagenome assembly from long accurate reads with metaMDBG”. In: *Nature Biotechnology* 42.9 (2024), pp. 1378–1383.

[60] Mikhail Kolmogorov et al. “metaFlye: scalable long-read metagenome assembly using repeat graphs”. In: *Nature Methods* 17.11 (Nov. 2020), pp. 1103–1110. ISSN: 15487105. DOI: 10.1038/s41592-020-00971-x.

[61] Chengxi Ye et al. “DBG2OLC: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies”. In: *Scientific reports* 6.1 (2016), p. 31900.

[62] Dmitry Antipov et al. “hybridSPAdes: an algorithm for hybrid assembly of short and long reads”. In: *Bioinformatics* 32.7 (2016), pp. 1009–1015.

[63] Aleksey V Zimin et al. “The MaSuRCA genome assembler”. In: *Bioinformatics* 29.21 (2013), pp. 2669–2677.

[64] Song Gao et al. “OPERA-LG: efficient and exact scaffolding of large, repeat-rich eukaryotic genomes with performance guarantees”. In: *Genome biology* 17 (2016), pp. 1–16.

[65] Denis Bertrand et al. “Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes”. In: *Nature biotechnology* 37.8 (2019), pp. 937–944.

[66] Ryan R Wick et al. “Unicycler: resolving bacterial genome assemblies from short and long sequencing reads”. In: *PLoS computational biology* 13.6 (2017), e1005595.

[67] Ting Hon et al. “Highly accurate long-read HiFi sequencing data for five complex genomes”. In: *Scientific data* 7.1 (2020), p. 399.

[68] Ye Tao et al. “Improved Assembly of Metagenome-Assembled Genomes and Viruses in Tibetan Saline Lake Sediment by HiFi Metagenomic Sequencing”. In: *Microbiology Spectrum* 11.1 (2023), e03328–22. DOI: 10.1128/spectrum.03328-22.

[69] Alex Chklovski et al. “CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning”. In: *Nature Methods* 20.8 (2023), pp. 1203–1212. DOI: 10.1038/s41592-023-01940-w.

[70] Dongwan D Kang et al. “MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies”. In: *PeerJ* 7 (2019), e7359.

[71] Johannes Alneberg et al. “Binning metagenomic contigs by coverage and composition”. In: *Nature methods* 11.11 (2014), pp. 1144–1146.

[72] Yu-Wei Wu, Blake A Simmons, and Steven W Singer. “MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets”. In: *Bioinformatics* 32.4 (2016), pp. 605–607.

[73] Cedric C Laczny et al. “BusyBee Web: metagenomic data analysis by bootstrapped supervised binning and annotation”. In: *Nucleic acids research* 45.W1 (2017), W171–W179.

[74] Jakob Nybo Nissen et al. “Improved metagenome binning and assembly using deep variational autoencoders”. In: *Nature biotechnology* 39.5 (2021), pp. 555–560.

[75] Haitao Han, Ziye Wang, and Shanfeng Zhu. “Benchmarking metagenomic binning tools on real datasets across sequencing platforms and binning modes”. In: *Nature Communications* 16 (2025), p. 2865. DOI: 10.1038/s41467-025-57957-6.

[76] Huarui Wang et al. “Complementary insights into gut viral genomes: a comparative benchmark of short- and long-read metagenomes using diverse assemblers and binners”. In: *Microbiome* 12.1 (2024), p. 260. DOI: 10.1186/s40168-024-01981-z.

[77] Christian MK Sieber et al. “Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy”. In: *Nature microbiology* 3.7 (2018), pp. 836–843.

[78] Gherman V Uritskiy, Jocelyne DiRuggiero, and James Taylor. “MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis”. In: *Microbiome* 6 (2018), pp. 1–13.

[79] Matthew R Olm et al. “dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication”. In: *The ISME journal* 11.12 (2017), pp. 2864–2868.

[80] Felipe A Simão et al. “BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs”. In: *Bioinformatics* 31.19 (2015), pp. 3210–3212.

[81] D.H. Parks et al. “CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes”. en. In: *Genome research* 25 (2015), pp. 1043–1055.

[82] Arang Rhie et al. “Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies”. In: *Genome biology* 21 (2020), pp. 1–27.

[83] Ryan R Wick and Kathryn E Holt. “Polypolish: short-read polishing of long-read bacterial genome assemblies”. In: *PLoS computational biology* 18.1 (2022), e1009802.

[84] Robert Vaser et al. “Fast and accurate de novo genome assembly from long uncorrected reads”. In: *Genome research* 27.5 (2017), pp. 737–746.

[85] Ryan R Wick et al. “Trycycler: consensus long-read assemblies for bacterial genomes”. In: *Genome biology* 22 (2021), pp. 1–17.

[86] Ryan R Wick, Benjamin P Howden, and Timothy P Stinear. “Autocycler: long-read consensus assembly for bacterial genomes”. In: *bioRxiv* (2025), pp. 2025–05.

[87] Oliver Schwengers et al. “Bakta: Rapid and standardized annotation of bacterial genomes via alignment-free sequence identification”. In: *Microbial Genomics* 7.11 (2021). ISSN: 20575858. DOI: 10.1099/MGEN.0.000685.

[88] P. Stamatopoulou et al. “Fermentation of organic residues to beneficial chemicals: A review of medium-chain fatty acid production”. en. In: *Processes* 8.1571 (2020).

[89] A. Mancini et al. “Biological and nutritional properties of palm oil and palmitic acid: Effects on health”. en. In: *Molecules* 20 (2015), pp. 17339–17361.

[90] C.M. Spirito, A.M. Marzilli, and L.T. Angenent. “Higher substrate ratios of ethanol to acetate steered chain elongation toward n-caprylate in a bioreactor with product extraction”. en. In: *Environmental Science & Technology* 52 (2018), pp. 13438–13447.

[91] L.A. Kucek, C.M. Spirito, and L.T. Angenent. “High n-caprylate productivities and specificities from dilute ethanol and acetate: chain elongation with microbiomes to upgrade products from syngas fermentation”. en. In: *Energy & Environmental Science* 9 (2016), pp. 3482–3494.

[92] T.I.M. Grootscholten et al. “Chain elongation of acetate and ethanol in an upflow anaerobic filter for high rate MCFA production”. en. In: *Bioresource Technology* 135 (2013), pp. 440–445.

[93] K.J. Steinbusch et al. “Biological formation of n-caproate and n-caprylate from acetate: fuel and chemicals from low grade biomass”. en. In: *Energy & Environmental Science* 4 (2011), pp. 216–224.

[94] Matthew T. Agler et al. “Chain elongation with reactor microbiomes: Upgrading dilute ethanol to medium-chain carboxylates”. In: *Energy and Environmental Science* 5.8 (2012), pp. 8189–8192. ISSN: 17545706. DOI: 10.1039/c2ee22101b.

[95] L.A. Kucek, M. Nguyen, and L.T. Angenent. “Conversion of l-lactate into n-caproate by a continuously fed reactor microbiome”. en. In: *Water Research* 93 (2016), pp. 163–171.

[96] X.Y. Zhu et al. “The synthesis of n-caproate from lactate: a new efficient process for medium-chain carboxylates production”. en. In: *Scientific Reports* 5.14360 (2015).

[97] S. Gildemyn et al. “Upgrading syngas fermentation effluent using Clostridium kluyveri in a continuous fermentation”. en. In: *Biotechnology for Biofuels* 10.83 (2017).

[98] S.J. Andersen et al. “A Clostridium group IV species dominates and suppresses a mixed culture fermentation by tolerance to medium chain fatty acids products”. en. In: *Frontiers in Bioengineering and Biotechnology* 5.8 (2017).

[99] J.M. Carvajal-Arroyo et al. en. Granular fermentation enables high rate caproic acid production from solid-free thin stillage. Green Chemistry 21:1330-1339. 2019.

[100] J.M. Carvajal-Arroyo et al. “Production and extraction of medium chain carboxylic acids at a semi-pilot scale”. en. In: *Chemical Engineering Journal* 416.127886 (2021).

[101] A. Duber et al. “Exploiting the real wastewater potential for resource recovery – n-caproate production from acid whey”. en. In: *Green Chemistry* 20 (2018), pp. 3790–3803.

[102] J. Xu et al. “In-line and selective phase separation of medium-chain carboxylic acids using membrane electrolysis”. en. In: *Chemical Communications* 51 (2015), pp. 6847–6850.

[103] S. Ge et al. “Long-term n-caproic acid production from yeast-fermentation beer in an anaerobic bioreactor with continuous product extraction”. en. In: *Environ Sci Technol* 49 (2015), pp. 8012–8021.

[104] C.M. Spirito et al. “Chain elongation in anaerobic reactor microbiomes to recover resources from waste”. en. In: *Current Opinion in Biotechnology* 27 (2014), pp. 115–122.

[105] Cavalcante WdA et al. “Anaerobic fermentation for n-caproic acid production: A review”. en. In: *Process Biochemistry* 54 (2017), pp. 106–119.

[106] L.T. Angenent et al. “Chain elongation with reactor microbiomes: Open-culture biotechnology to produce biochemicals”. en. In: *Environmental Science & Technology* 50 (2016), pp. 2796–2810.

[107] W. Buckel and R.K. Thauer. “Energy conservation via electron bifurcating ferredoxin reduction and proton/Na+ translocating ferredoxin oxidation”. en. In: *Biochimica et Biophysica Acta (BBA) - Bioenergetics* 1827 (2013), pp. 94–113.

[108] M.J. Scarborough et al. en. Diagnosing and predicting mixed-culture fermentations with unicellular and guild-based metabolic models. mSystems 5:e00755-20. 2020.

[109] M.J. Scarborough et al. *Medium-chain fatty acid synthesis by "Candidatus Weimeria bifida*. en. gen. nov., sp. nov., and "Candidatus Pseudoramibacter fermentans" sp. nov. Applied and Environmental Microbiology 86:e02242-19. 2020.

[110] M.J. Scarborough et al. en. Metatranscriptomic and thermodynamic insights into medium-chain fatty acid production using an anaerobic microbiome. mSystems 3:e00221-18. 2018.

[111] S.-L. Wu et al. "Unveiling the mechanisms of medium-chain fatty acid production from waste activated sludge alkaline fermentation liquor through physiological, thermodynamic and metagenomic investigations". en. In: *Water Research* 169.115218 (2020).

[112] W. Han et al. "Metabolic Interactions of a Chain Elongation Microbiome". en. In: *Applied and Environmental Microbiology* 84:e01614-18 (2018).

[113] H. Gest. "A serendipic legacy: Erwin Esmarch's isolation of the first photosynthetic bacterium in pure culture". en. In: *Photosynth Res* 46 (1995), pp. 473–8.

[114] X. Zhu et al. "Production of high-concentration n-caproic acid from lactate through fermentation using a newly isolated Ruminococcaceae bacterium CPB6". en. In: *Biotechnology for Biofuels* 10.102 (2017).

[115] B.T. Bornstein and H.A. Barker. "The energy metabolism of Clostridium kluyveri and the synthesis of fatty acids". en. In: *Journal of Biological Chemistry* 172 (1948), pp. 659–669.

[116] H.A. Barker, M.D. Kamen, and B.T. Bornstein. "The synthesis of n-butyric and n-caproic acids from ethanol and acetic acid by Clostridium kluyveri". en. In: *Proceedings of the National Academy of Sciences*. Vol. 31. 1945, pp. 373–381.

[117] R.J. Wallace et al. "Eubacterium pyruvativorans sp. nov., a novel non-saccharolytic anaerobe from the rumen that ferments pyruvate and amino acids, forms caproate and utilizes acetate and propionate". en. In: *International Journal of Systematic and Evolutionary Microbiology* 53 (2003), pp. 965–970.

[118] R.J. Wallace et al. "Metabolic properties of Eubacterium pyruvativorans, a ruminal 'hyper-ammonia-producing' anaerobe with metabolic properties analogous to those of Clostridium kluyveri". lt. In: *Microbiology* 150 (2004), pp. 2921–2930.

[119] B.S. Jeon et al. “Production of medium-chain carboxylic acids by Megasphaera sp. MH with supplemental electron acceptors”. en. In: *Biotechnology for Biofuels* 9.129 (2016).

[120] L.V. Holdeman, E.P. Cato, and W.E.C. Moore. “Amended description of Ramibacterium alactolyticum Prévot and Taffanel with proposal of a neotype strain1”. en. In: *International Journal of Systematic and Evolutionary Microbiology* 17 (1967), pp. 323–341.

[121] J.J. Werner et al. “Microbial Community Dynamics and Stability during an Ammonia-Induced Shift to Syntrophic Acetate Oxidation”. en. In: *Applied and Environmental Microbiology* 80 (2014), pp. 3375–3383.

[122] Maximilienne Toetie Allaart et al. “Physiological and stoichiometric characterization of ethanol-based chain elongation in the absence of short-chain carboxylic acids”. In: *Scientific Reports* 13.1 (2023), p. 17370.

[123] Robert M. Bowers et al. *Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea*. Aug. 2017. DOI: 10.1038/nbt.3893.

[124] Pierre Alain Chaumeil et al. “GTDB-Tk: A toolkit to classify genomes with the genome taxonomy database”. In: *Bioinformatics* 36.6 (2020), pp. 1925–1927. ISSN: 14602059. DOI: 10.1093/bioinformatics/btz848.

[125] Y. Katano et al. *Complete genome sequence of Oscillibacter valericigenes Sjm18-20(T*. nl. =NBRC 101213(T)). Standards in Genomic Sciences 6:406-414. 2012.

[126] Alfred M Spormann. *Principles of microbial metabolism and metabolic ecology*. Springer, 2023.

[127] C.A. Kerfeld et al. “Bacterial microcompartments”. en. In: *Nature Reviews Microbiology* 16 (2018), pp. 277–290.

[128] M. Held, M.B. Quin, and C. Schmidt-Dannert. “Eut bacterial microcompartments: insights into their function, structure, and bioengineering applications”. en. In: *J Mol Microbiol Biotechnol* 23 (2013), pp. 308–20.

[129] D. Heldt et al. “Structure of a trimeric bacterial microcompartment shell protein, EtuB, associated with ethanol utilization in Clostridium kluyveri”. en. In: *Biochemical Journal* 423 (2009), pp. 199–207.

[130] H. Seedorf et al. “The genome of Clostridium kluyveri, a strict anaerobe with unique metabolic features”. en. In: *Proceedings of the National Academy of Sciences*. Vol. 105. 2008, pp. 2128–2133.

[131] J. Kim et al. “Chain elongation with reactor microbiomes: Upgrading dilute ethanol to medium-chain carboxylates”. In: *Energy and Environmental Science* 5.8 (2015), pp. 8189–8192. DOI: 10.1039/c2ee22101b.

[132] L. Regueiro et al. “Comparing the inhibitory thresholds of dairy manure co-digesters after prolonged acclimation periods: Part 2 – correlations between microbiomes and environment”. en. In: *Water Research* 87 (2015), pp. 458–466.

[133] S. Srinivasan et al. “Pseudoclavibacter caeni sp. nov., isolated from sludge of a sewage disposal plant”. en. In: *International Journal of Systematic and Evolutionary Microbiology* 62 (2012), pp. 786–790.

[134] M. Coma et al. “Product diversity linked to substrate usage in chain elongation by mixed-culture fermentation”. en. In: *Environmental Science & Technology* 50 (2016), pp. 6467–6476.

[135] D. Vasudevan, H. Richter, and L.T. Angenent. “Upgrading dilute ethanol from syngas fermentation to n-caproate with reactor microbiomes”. en. In: *Bioresource Technology* 151 (2014), pp. 378–382.

[136] J.G. Usack and L.T. Angenent. “Comparing the inhibitory thresholds of dairy manure co-digesters after prolonged acclimation periods: Part 1 – Performance and operating limits”. en. In: *Water Research* 87 (2015), pp. 446–457.

[137] RStudio Team. *RStudio: Integrated development environment for R, RStudio*. it. Boston, MA: Inc, 2016. URL: http://www.rstudio.com/..

[138] E. Bolyen et al. en. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature Biotechnology 37:852-857. 2019.

[139] Christian Quast et al. “The SILVA ribosomal RNA gene database project: improved data processing and web-based tools”. In: *Nucleic Acids Research* 41.D1 (2012), pp. D590–D596.

[140] P. Virtanen et al. “SciPy 1.0: fundamental algorithms for scientific computing in Python”. la. In: *Nature Methods* 17 (2020), pp. 261–272.

[141] A.M. Bolger, M. Lohse, and B. Usadel. “Trimmomatic: a flexible trimmer for Illumina sequence data”. en. In: *Bioinformatics* 30 (2014), pp. 2114–2120.

[142] B. Buchfink, C. Xie, and D.H. Huson. “Fast and sensitive protein alignment using DIAMOND”. en. In: *Nature methods* 12.59 (2015).

[143] D.H. Huson et al. “MEGAN community edition-interactive exploration and analysis of large-scale microbiome sequencing data”. en. In: *PLoS computational biology* 12 (2016).

[144] Nguyen Nhat Nam et al. “Metagenomics: an effective approach for exploring microbial diversity and functions”. In: *Foods* 12.11 (2023), p. 2140. DOI: 10.3390/foods12112140.

[145] Le Zhang et al. “Bioinformatics analysis of metagenomics data of biogas-producing microbial communities in anaerobic digesters: A review”. In: *Renewable and Sustainable Energy Reviews* 100 (Feb. 2019), pp. 110–126. ISSN: 1364-0321. DOI: 10.1016/J.RSER.2018.10.021.

[146] Kyungjin Cho et al. “Microbial community shifts in a farm-scale anaerobic digester treating swine waste: correlations between bacteria communities associated with hydrogenotrophic methanogens and environmental conditions”. In: *Science of The Total Environment* 601–602 (2017), pp. 167–176. DOI: 10.1016/j.scitotenv.2017.05.188.

[147] Muneer Ahmad Malla et al. “Exploring the human microbiome: the potential future role of next-generation sequencing in disease diagnosis and treatment”. In: *Frontiers in Immunology* 9 (2019), p. 2868. DOI: 10.3389/fimmu.2018.02868.

[148] Hans-Peter Grossart et al. “Linking metagenomics to aquatic microbial ecology and biogeochemical cycles”. In: *Limnology and Oceanography* 65 (2020), S2–S20. DOI: 10.1002/lno.11382.

[149] A. F. Vieira, M. Moura, and L. Silva. “Soil metagenomics in grasslands and forests – A review and bibliometric analysis”. In: *Applied Soil Ecology* 167 (Nov. 2021), p. 104047. ISSN: 0929-1393. DOI: 10.1016/J.APSOIL.2021.104047.

[150] Atsufumi Ohta et al. “Using nanopore sequencing to identify fungi from clinical samples with high phylogenetic resolution”. In: *Scientific Reports* 13.1 (Dec. 2023). ISSN: 20452322. DOI: 10.1038/s41598-023-37016-0.

[151] Mantas Sereika et al. “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”. In: *Nature Methods* 19.7 (July 2022), pp. 823–826. ISSN: 15487105. DOI: 10.1038/s41592-022-01539-7.

[152] Henry C.M. Leung et al. “Detecting structural variations with precise breakpoints using low-depth WGS data from a single oxford nanopore MinION flowcell”. In: *Scientific Reports* 12.1 (Dec. 2022). ISSN: 20452322. DOI: 10.1038/s41598-022-08576-4.

[153] Seung Chul Shin et al. “Nanopore sequencing reads improve assembly and gene annotation of the Parochlus steinenii genome”. In: *Scientific Reports* 9.1 (Dec. 2019). ISSN: 20452322. DOI: 10.1038/s41598-019-41549-8.

[154] Alexander T. Dilthey et al. “Strain-level metagenomic assignment and compositional estimation for long reads with MetaMaps”. In: *Nature Communications* 10.1 (Dec. 2019). ISSN: 20411723. DOI: 10.1038/s41467-019-10934-2.

[155] Kristen D. Curry et al. “Emu: species-level microbial community profiling of full-length 16S rRNA Oxford Nanopore sequencing data”. In: *Nature Methods* 19.7 (July 2022), pp. 845–853. ISSN: 15487105. DOI: 10.1038/s41592-022-01520-4.

[156] Daehwan Kim et al. “Centrifuge: Rapid and sensitive classification of metagenomic sequences”. In: *Genome Research* 26.12 (Dec. 2016), pp. 1721–1729. ISSN: 15495469. DOI: 10.1101/gr.210641.116.

[157] Derrick E Wood, Jennifer Lu, and Ben Langmead. “Improved metagenomic analysis with Kraken 2”. In: *Genome biology* 20.1 (2019), p. 257.

[158] Josip Marić et al. “Comparative analysis of metagenomic classifiers for long-read sequencing datasets”. In: *BMC Bioinformatics* 25.1 (Dec. 2024). ISSN: 14712105. DOI: 10.1186/s12859-024-05634-8.

[159] Ela Sauerborn et al. “Detection of hidden antibiotic resistance through real-time genomics”. In: *Nature Communications* 15.1 (Dec. 2024). ISSN: 20411723. DOI: 10.1038/s41467-024-49851-4.

[160] Thomas Hoenen et al. “Nanopore sequencing as a rapidly deployable Ebola outbreak tool”. In: *Emerging Infectious Diseases* 22.2 (Feb. 2016), pp. 331–334. ISSN: 10806059. DOI: 10.3201/eid2202.151796.

[161] Fang Shiang Lim et al. “Advancing pathogen surveillance by nanopore sequencing and genotype characterization of *Acheta domesticus* densovirus in mass-reared house crickets”. In: *Scientific Reports* 14.1 (2024), p. 8525. DOI: 10.1038/s41598-024-58768-3.

[162] Lara Urban et al. “Freshwater monitoring by nanopore sequencing”. In: *eLife* 10 (Jan. 2021), pp. 1–27. ISSN: 2050084X. DOI: 10.7554/eLife.61504.

[163] Sarah L. Castro-Wallace et al. “Nanopore DNA Sequencing and Genome Assembly on the International Space Station”. In: *Scientific Reports* 7.1 (Dec. 2017). ISSN: 20452322. DOI: 10.1038/s41598-017-18364-0.

[164] David Werner et al. “MinION Nanopore sequencing accelerates progress towards ubiquitous genetics in water research”. In: *Water* 14.16 (2022), p. 2491. DOI: 10.3390/w14162491.

[165] Rory Munro et al. “minoTour, real-time monitoring and analysis for nanopore sequencers”. In: *Bioinformatics* 38.4 (2022), pp. 1133–1135. ISSN: 14602059. DOI: 10.1093/bioinformatics/btab780.

[166] Anna Wierczeiko et al. “Gene expression”. In: (). DOI: 10.5281/zenodo.8099825. URL: https://doi.org/10.5281/zenodo.8099825..

[167] Ned Peel et al. “MARTi: a real-time analysis and visualisation tool for nanopore metagenomics”. In: *bioRxiv* (2025), pp. 2025–02. DOI: 10.1101/2025.02.14.638261.

[168] Kristofer Sandås et al. “Nanometa Live: a user-friendly application for real-time metagenomic data analysis and pathogen identification”. In: *Bioinformatics* 40.3 (2024), btae108. DOI: 10.1093/bioinformatics/btae108.

[169] Francesco Beghini et al. “Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3”. In: *eLife* 10 (2021), e65088. DOI: 10.7554/eLife.65088.

[170] J Gregory Caporaso et al. “QIIME allows analysis of high-throughput community sequencing data”. In: *Nature Methods* 7.5 (2010), pp. 335–336. DOI: 10.1038/nmeth.f.303.

[171] Kevin P Keegan, Elizabeth M Glass, and Folker Meyer. “MG-RAST, a metagenomics service for analysis of microbial community structure and function”. In: *Microbial Environmental Genomics (MEG)*. Vol. 1399. Methods in Molecular Biology. Springer, 2016, pp. 207–233. DOI: 10.1007/978-1-4939-3369-3_13.

[172] Georges P. Schmartz et al. “BusyBee Web: towards comprehensive and differential composition-based metagenomic binning”. In: *Nucleic Acids Research* 50.W1 (July 2022), W132–W137. ISSN: 13624962. DOI: 10.1093/nar/gkac298.

[173] Jeremy Fan, Steven Huang, and Samuel D. Chorlton. “BugSeq: a highly accurate cloud platform for long-read metagenomic analyses”. In: *BMC Bioinformatics* 22.1 (Dec. 2021). ISSN: 14712105. DOI: 10.1186/s12859-021-04089-5.

[174] Benjamin Albrecht, Caner Bağcı, and Daniel H Huson. “MAIRA—real-time taxonomic and functional analysis of long reads on a laptop”. In: *BMC Bioinformatics* 21.Suppl 13 (2020), p. 390. DOI: 10.1186/s12859-020-03684-2.

[175] Oxford Nanopore Technologies. *EPI2ME Labs*.

[176] Aimeric Bruno, Jean Marc Aury, and Stefan Engelen. “BoardION: real-time monitoring of Oxford Nanopore sequencing instruments”. In: *BMC Bioinformatics* 22.1 (Dec. 2021). ISSN: 14712105. DOI: 10.1186/s12859-021-04161-0.

[177] Nicholas D Sanderson et al. “Real-time analysis of nanopore-based metagenomic sequencing from infected orthopaedic devices”. In: *BMC Genomics* 19.1 (2018), p. 714. DOI: 10.1186/s12864-018-5094-y.

[178] Ulrike Biehain. “Human Gut Microbes Under Power-Development of a Bioelectrochemical System to Uncouple and Interrogate H2-Syntrophic Partners in the Human Gut Microbiota”. PhD thesis. Dissertation, Tübingen, Universität Tübingen, 2024. DOI: 10.15496/publikation-95989.

[179] Mohamed S. Donia and Michael A. Fischbach. “Small molecules from the human microbiota”. In: *Science* 349.6246 (July 2015). ISSN: 10959203. DOI: 10.1126/science.1254766.

[180] Alex Chklovski et al. “CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning”. In: *Nature Methods* 20.8 (2023), pp. 1203–1212. DOI: 10.1038/s41592-023-01940-w.

[181] Daniel M. Portik, C. Titus Brown, and N. Tessa Pierce-Ward. “Evaluation of taxonomic classification and profiling methods for long-read shotgun metagenomic sequencing datasets”. In: *BMC Bioinformatics* 23.1 (Dec. 2022). ISSN: 14712105. DOI: 10.1186/s12859-022-05103-0.

[182] Daniel H. Huson et al. “MEGAN-LR: New algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs”. In: *Biology Direct* 13.1 (Apr. 2018). ISSN: 17456150. DOI: 10.1186/s13062-018-0208-7.

[183] A Murat Eren et al. “Anvi’o: an advanced analysis and visualization platform for ‘omics data”. In: *PeerJ* 3 (2015), e1319. DOI: 10.7717/peerj.1319.

[184] Nick Weber et al. “Nephele: A cloud platform for simplified, standardized and reproducible microbiome data analysis”. In: *Bioinformatics* 34.8 (Apr. 2018), pp. 1411–1413. ISSN: 14602059. DOI: 10.1093/bioinformatics/btx617.

[185] Joseph J. Gillespie et al. “Patric: The comprehensive bacterial bioinformatics resource with a focus on human pathogenic species”. In: *Infection and Immunity* 79.11 (Nov. 2011), pp. 4286–4298. ISSN: 00199567. DOI: 10.1128/IAI.00207-11.

[186] Lorna Richardson et al. “MGnify: the microbiome sequence data analysis resource in 2023”. In: *Nucleic Acids Research* 51.1 D (Jan. 2023), pp. D753–D759. ISSN: 13624962. DOI: 10.1093/nar/gkac1080.

[187] Km Sartaj et al. *Unravelling Metagenomics Approach for Microbial Biofuel Production*. Nov. 2022. DOI: 10.3390/genes13111942.

[188] Matthew James Scarborough et al. *Microbiomes for sustainable biomanufacturing*. Feb. 2022. DOI: 10.1016/j.mib.2021.09.015.

[189] You Che et al. “Mobile antibiotic resistome in wastewater treatment plants revealed by Nanopore metagenomic sequencing”. In: *Microbiome* 7.1 (Mar. 2019). ISSN: 20492618. DOI: 10.1186/s40168-019-0663-0.

[190] Noah Fierer et al. “A Metagenomic Investigation of Spatial and Temporal Changes in Sewage Microbiomes across a University Campus”. In: *mSystems* 7.5 (Oct. 2022). ISSN: 23795077. DOI: 10.1128/msystems.00651-22.

[191] Shuang Zhou et al. *Assessment of bioleaching microbial community structure and function based on next-generation sequencing technologies*. Dec. 2018. DOI: 10.3390/min8120596.

[192] Michael D. Parkins et al. *Wastewater-based surveillance as a tool for public health action: SARS-CoV-2 and beyond*. Mar. 2024. DOI: 10.1128/cmr.00103-22.

[193] Karrie K.K. Ko, Kern Rei Chng, and Niranjan Nagarajan. “Metagenomics-enabled microbial surveillance”. In: *Nature Microbiology* 7.4 (Apr. 2022), pp. 486–496. ISSN: 20585276. DOI: 10.1038/s41564-022-01089-w.

[194] Peter J.A. Cock et al. “Biopython: Freely available Python tools for computational molecular biology and bioinformatics”. In: *Bioinformatics* 25.11 (June 2009), pp. 1422–1423. ISSN: 13674803. DOI: 10.1093/bioinformatics/btp163.

[195] Li Song and Ben Langmead. “Centrifuger: lossless compression of microbial genomes for efficient and accurate metagenomic sequence classification”. In: *Genome Biology* 25.1 (Dec. 2024). ISSN: 1474760X. DOI: 10.1186/s13059-024-03244-4.

[196] Giovanni Manzini. “An analysis of the Burrows—Wheeler transform”. In: *Journal of the ACM* 48.3 (2001), pp. 407–430. DOI: 10.1145/382780.382782.

[197] Paolo Ferragina et al. “An Alphabet-Friendly FM-Index”. In: *String Processing and Information Retrieval (SPIRE 2004), LNCS 3246*. Springer, 2004, pp. 150–160. DOI: 10.1007/978-3-540-30213-1_23.

[198] Heng Li. “Minimap2: Pairwise alignment for nucleotide sequences”. In: *Bioinformatics* 34.18 (Sept. 2018), pp. 3094–3100. ISSN: 14602059. DOI: 10.1093/bioinformatics/bty191.

[199] T K Moon. “The expectation-maximization algorithm”. In: *IEEE Signal Processing Magazine* 13.6 (1996), pp. 47–60. DOI: 10.1109/79.543975.

[200] Krithika Arumugam et al. “Annotated bacterial chromosomes from frame-shift-corrected long-read metagenomic data”. In: *Microbiome* 7.1 (Apr. 2019). ISSN: 20492618. DOI: 10.1186/s40168-019-0665-y.

[201] Oxford Nanopore Technologies. *Medaka: sequence consensus polishing for nanopore data*. Version v2.0.1. Accessed: 2025-09-10. 2025. URL: https://github.com/nanoporetech/medaka/releases/tag/v2.0.1.

[202] Minoru Kanehisa et al. “KEGG for taxonomy-based analysis of pathways and genomes”. In: *Nucleic Acids Research* 51.D1 (Jan. 2023), pp. D587–D592. ISSN: 13624962. DOI: 10.1093/nar/gkac963.

[203] João C. Sequeira et al. “UPIMAPI, reCOGnizer and KEGGCharter: Bioinformatics tools for functional annotation and visualization of (meta)-omics datasets”. In: *Computational and Structural Biotechnology Journal* 20 (Jan. 2022), pp. 1798–1810. ISSN: 20010370. DOI: 10.1016/j.csbj.2022.03.042.

[204] Django Software Foundation. *django*. Jan. 2024.

[205] Shammamah Hossain. *Visualization of Bioinformatics Data with Dash Bio*. Tech. rep. 2019. URL: https://dash.plot.ly/dash-bio..

[206] Matthew Aton et al. “Scikit-bio: a fundamental Python library for biological omic data analysis”. In: *Nature Methods* (2025), pp. 1–3.

[207] The pandas development team. *pandas-dev/pandas: Pandas*. Feb. 2020. DOI: 10.5281/zenodo.3509134. URL: https://doi.org/10.5281/zenodo.3509134.

[208] Wes Mckinney. *Data Structures for Statistical Computing in Python*. Tech. rep. 2010.

[209] Eric W. Sayers et al. “Database resources of the national center for biotechnology information”. In: *Nucleic Acids Research* 50.D1 (Jan. 2022), pp. D20–D26. ISSN: 13624962. DOI: 10.1093/nar/gkab1112.

[210] Steven F Stoddard et al. “rrn DB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development”. In: *Nucleic acids research* 43.D1 (2015), pp. D593–D598.

[211] Kurt Gemeinhardt et al. “Toward industrial C8 production: oxygen intrusion drives renewable n-caprylate production from ethanol and acetate via intermediate metabolite production”. In: *Green Chemistry* 27.11 (2025), pp. 2931–2949.

[212] T. Seemann. “Prokka: rapid prokaryotic genome annotation”. en. In: *Bioinformatics* 30 (2014), pp. 2068–2069.

[213] Soyoung Ham et al. *Development of a continuous microbiome bioreactor system for culturing stable nasal communities*. Manuscript in preparation. Unpublished manuscript, in preparation. 2025.

[214] Wenhuan Zeng et al. *NanoGraph: Mapping Nanopore Squiggles to Graphs Enables Accurate Taxonomic Assignment*. Manuscript in review. Unpublished manuscript, in review. 2025.

[215] Yu Li et al. “DeepSimulator: a deep simulator for Nanopore sequencing”. In: *Bioinformatics* 34.17 (2018), pp. 2899–2908.

[216] Nuala A O'Leary et al. "Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation". In: *Nucleic acids research* 44.D1 (2016), pp. D733–D745.

[217] Petr Danecek et al. "Twelve years of SAMtools and BCFtools". In: *GigaScience* 10.2 (2021), giab008.

[218] Jim Shaw and Yun William Yu. "Fast and robust metagenomic sequence comparison through sparse chaining with skani". In: *Nature Methods* 20 (2023), pp. 1661–1665.